HANDBOOK OF WEATHER, CLIMATE, AND WATER

ATMOSPHERIC CHEMISTRY, HYDROLOGY, AND SOCIETAL IMPACTS

THOMAS D. POTTER

BRADLEY R. COLMAN

HANDBOOK OF WEATHER, CLIMATE, AND WATER

Atmospheric Chemistry, Hydrology, and Societal Impacts

Edited by

THOMAS D. POTTER

BRADLEY R. COLMAN

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Potter, Thomas D. 1929–
Handbook of weather, climate, and water: chemistry, hydrology, and societal impacts
Thomas D. Potter & Bradley R. Colman.
p. cm.
ISBN 0-471-21489-2 (cloth : alk paper)
1. Atmospheric physics. 2. Atmospheric chemistry. 3. Hydrology. I. Colman, Bradley R. II. Title

QC861.3 .P67 2002
551.6–dc21

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

DEDICATION AND ACKNOWLEDGMENTS

Many people have assisted in the production of this Handbook—the Contributing Editors, the Authors, our editors at Wiley, friends too numerous to mention, and our families who supported us during the long process of completing this work. Professor Peter Shaffer, University of Washington, is owed deep appreciation for his untiring generosity in sharing his experience and talent to solve many problems associated with this large project. They all deserve much credit for their contributions and we want to express our deep thanks to all of them.

Finally, we want to dedicate this work to Tom Lockhart, the Contributing Editor of the Measurements part of the Handbook. Tom passed away in early 2001 and we regret that he will not be able to see the results of his efforts and those of his colleagues in final form.

Tom Potter and Brad Colman

PREFACE

The *Handbook of Weather, Climate, and Water* provides an authoritative report at the start of the 21st Century on the state of scientific knowledge in these exciting and important earth sciences. Weather, climate, and water affect every person on earth every day in some way. These effects range from disasters like killer storms and floods, to large economic effects on energy or agriculture, to health effects such as asthma or heat stress, to daily weather changes that affect air travel, construction, fishing fleets, farmers, and mothers selecting the clothes their children will wear that day, to countless other subjects.

During the past two decades a series of environmental events involving weather, climate, and water around the globe have been highly publicized in the press: the Ozone Hole, Acid Rain, Global Climate Change, El Ninos, major floods in Bangladesh, droughts in the Sahara, and severe storms such as hurricane Andrew in Florida and the F5 tornado in Oklahoma. These events have generated much public interest and controversy regarding the appropriate public policies to deal with them. Such decisions depend critically upon scientific knowledge in the fields of weather, climate, and water.

One of two major purposes of the Handbook is to provide an up-to-date accounting of the sciences that underlie these important societal issues, so that both citizens and decision makers can understand the scientific foundation critical to the process of making informed decisions. To achieve this goal, we commissioned overview chapters on the eight major topics that comprise the Handbook: Atmospheric Dynamics, Climate System, Physical Meteorology, Weather Systems, Measurements, Atmospheric Chemistry, Hydrology, and Societal Impacts. Each of the sections was organized by a distinguished scientist who is a leading authority within that major field. In addition to writing an overview chapter, this scientist served as the Contributing Editor for that section of the Handbook. Each Contributing Editor selected both the topics and authors of the individual chapters, thus ensuring that the most important material has been included. The chapter authors are themselves leading experts in their specialty. These overview chapters present, in terms understandable to everyone, the basic scientific information needed to appreciate the major environmental issues listed above.

The second major purpose of the Handbook is to provide a comprehensive reference volume for scientists who are specialists in the atmospheric and hydrologic areas. In addition, scientists from closely related disciplines and others who wish to get an authoritative scientific accounting of these fields should find this work to be of great value. The 95 professional-level chapters are the first comprehensive and integrated survey of these sciences in over 50 years, the last being completed in 1951 when the American Meteorological Society published the Compendium of Meteorology.

The *Handbook of Weather, Climate and Water* is organized into two volumes containing eight major sections that encompass the fundamentals and critical topic areas across the atmospheric and hydrologic sciences. This volume contains sections on the highly important topics of Atmospheric Chemistry, Hydrology, and Societal Impacts. The section on Atmospheric Chemistry contains thorough descriptions of the major biogeochemical cycles (carbon, oxygen, nitrogen and sulfur) that describe how chemical elements and compounds are transferred between the atmosphere, oceans, land and the biosphere, and their relationship to important environmental issues such as global climate change, the ozone hole, acid rain, and air pollution. The Hydrology section includes in-depth discussions of all parts of the hydrologic cycle (rain, snow, evaporation, runoff, ground water, and soil moisture), plus chapters on floods, remote sensing and GIS in hydrology, and stochastic processes in hydrology. Societal Impacts has chapters on the social effects of all of the major environmental issues.

To better protect against weather, climate, and water hazards, as well as to promote the positive benefits of utilizing more accurate information about these natural events, society needs improved predictions of them. To achieve this, scientists must have a better understanding of the entire atmospheric and hydrologic system. Major advances have been made during the past 50 years to better understand the complex sciences involved. These scientific advances, together with vastly improved technologies such as Doppler radar, new satellite capabilities, numerical methods and computing, have resulted in greatly improved prediction capabilities over the past decade. Major storms are rarely missed nowadays because of the capability of numerical weather-prediction models to more effectively use the data from satellites, radars and surface observations, and weather forecasters' improved understanding of threatening weather systems. Improvements in predictions are ongoing. The public can now rely on the accuracy of forecasts out to about five days, when only a decade or so ago forecasts were accurate to only a day or two. Similarly, large advances have been made in understanding the climate system during the past 20 years. Climate forecasts out to a year are now made routinely and users in many fields find economic advantages in these climate outlooks even with the current marginal accuracies, which no doubt will improve as advances in our understanding of the Climate System occur in future years.

Tom Potter and Brad Colman

Color images from this volume are available at ftp://ftp.wiley.com/public/sci_tech_med/weather/.

CONTRIBUTORS

RICHARD ARIMOTO, Carlsbad Environmental Monitoring and Research Center, New Mexico State University, 1400 University Drive, Carlsbad, NM 88220-3575

ROGER C. BALES, University of Arizona, Department of Hydrology and Water Resources, Tucson, AZ 85721-0011

ABDELLATIF BENCHERIFA, Techlink, Advanced Technology Park, 900 Technology Boulevard, Snite A, Bozeman, MT 59718-6857

MICHELE M. BETSILL, Colorado State University, Department of Political Science, Fort Collins, CO 80523-1782

KEITH BEVEN, Lancaster University, Department of Environmental Science, Lancaster LA1 4YQ, United Kingdom

J. D. BRADSHAW, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

KENNETH BROAD, International Research Institute for Climate Prediction 61, Route 9W, Monell Building Palisades, NY 10964-8000

PAOLO BURLANDO, Institute of Hydromechanics and Water Resouces, ETH-Hönggerberg, Zurich, Switzerland

STANLEY A. CHANGNON, Illinois State Water and Changnon Climatol, 801 Buckthorn, Mahomet, IL 61853

G. CHEN, Georgia Tech, 22 Bobby Dodd Way NW, Room 205A, Atlanta, GA 30332

M. CHIN, National Aeronautics and Space Administration, Goddard Space Flight Center, Mail Code 916, Greenbelt, MD 20771

DON CLINE, National Weather Service, NOAA, National Operational Hydrologic Remote Sensing Center, Chanhassen, MN 55317-8582

STEWART J. COHEN, University of British Columbia, Sustainable Development Research Institute, 2029 West Mall, Vancouver, British Columbia, V6T 1Z2 Canada

J. H. CRAWFORD, NASA Langley Research Center, Mail Code 483, Hampton. VA 23681-0001

D. D. DAVIS, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

THOMAS E. DOWNING, University of Oxford, Environmental Change Unit, Oxford, United Kingdom

MARY W. DOWNTON, National Center for Atmospheric Research, Box 3000, Boulder, CO 80301

EDWIN T. ENGMAN, NASA Goddard Space Flight Center, Laboratory for Hydrospheric Processes, Hydrological Sciences Branch, Code 974 Greenbelt, MD 20771

JACK FISHMAN, NASA Langley Research Center, Atmospheric Sciences Research, Hampton, VA 23681-2219

D. L. FREAD, National Weather Service, Office of Hydrology, 622 Stone Road, Westminster, MD 21158

MICHAEL H. GLANTZ, ESIG/NCAR, 3450 Mitchell Lane, Boulder, CO 80301

WILLIAM B GRANT, NASA Langley Research Center, Atmospheric Sciences Research, MS 401A, Hampton, VA 23681

WILLIAM L. GROSE, NASA Langley Research Center

EVE GRUNTFEST, University of Colorado, Department of Geography and Environmental Studies, 1420 Austin Bluffs Parkway, P.O. Box 7150, Colorado Springs, CO 80933-7150

HOSHIN GUPTA, University of Arizona, Department of Hydrology, Tucson, AZ 85721-0011

R. L HEATHCOTE, Finders University, Adelaide, Australia

PAUL R. HOUSER, NASA-GSFC, Hydrological Sciences Branch, Greenbelt, MD 20771

DANIEL J. JACOB, Harvard University, Department of Earth and Planetary Sciences, 29 Oxford Street, Cambridge, MA 02138

Steven Jennings, University of Colorado at Colorado Springs, 1420 Austin Bluffs Parkway, P.O. Box 7150, Colorado Springs, CO 80933-7150

Jack A. Kaye, NASA Langley Research Center, Office of Earth Sciences, Washington, DC 20546-0001

M. A. K. Khalil, Portland State University, Department of Physics, P.O. Box 751 Portland, OR 97207-0751

William P. Kustas, USDA Agricultural Research Service, Hydrology Laboratory, Beltsville, MD 20750

Donald H. Lenschow, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000

Marcy E. Litvak, University of California-Irvine, Earth System Science, Irvine, CA 92697-3100

S. C. Liu, Georgia Tech, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

William C. Malm, Colorado State University, Foothills Campus, Cooperative Institute for Research in the Atmosphere, Fort Collins, CO 80523-1375

Nandish Mattikalli, Cambridge Research Associates, 1430 Spring Hill Road, Suite 200, McLean, VA 22102

Paulette Middleton, Creator, President Panorama Pathways, http://PanoramaPathways.net; Rand Environment, Environmental Science and Policy Center, 2385 Panorama Avenue, Boulder, CO 80304

Kathleen A. Miller, National Center for Atmospheric Research, Boulder, CO

Mario J. Molina, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307

M. Susan Moran, Southwest Watershed Research Center, USDA Agricultural Research Service, 2000 East Allen Road, Tucson, AZ 85719

Neville Nicholls, Bureau of Meteorology, GPO Box 1289K, Melbourne, Victoria 3001 Australia

John M. Norman, University of Wisconsin, Department of Soil Science, 1525 Observatory Drive, Madison, WI 53706-1299

Paul Novelli, NOAA Climate Monitoring and Diagnostics Laboratory, Environmental Research Laboratories, 325 Broadway, Boulder, CO 80303-3328

Kenneth E. Pickering, University of Maryland, Department of Meteorology, 3433 Computer and Space Sciences Building, College Park, MD 20742-2425

Roger A. Pielke, Jr., University of Colorado/CIRES, Campus Box 488, Boulder, CO 80309-0488

Roger S. Pulwarty, U.S. Department of Commerce, NOAA Office of Global Programs, 1100 Wayne Avenue, Suite 1210 Silver Springs, MD 20910

Jorge A. Ramírez, Colorado State University, Department of Civil Engineering, Fort Collins, CO 80523

José D. Salas, Colorado State University, Department of Civil Engineering, Fort Collins, CO 80523

Stephen H. Schneider, Stanford University, Department of Biological Sciences and Institute of Stanford, CA

Stephen E. Schwartz, Brookhaven National Laboratory, ASD, Blgd. 815E, P.O. Box 5000, Upton, NY 11973-5000

Roger A. Sedjo, Resources for the Future, 1616P Street NW, Washington, DC 20036

John H. Seinfeld, California Institute of Technology, Pasadena, CA 91125

M. J. Shearer, Portland State University, Department of Physics, P.O. Box 751, Portland, OR 97207-0751

Sanford Sillman, University of Michigan, Atmospheric, Oceanic, and Space Sciences, 2455 Hayward, Ann Arbor, MI 48109-2143

James A. Smith, Princeton University, Department of Civil Engineering and Operations R, Princeton, NJ 08544

Soroosh Sorooshian, University of Arizona, Department of Hydrology and Water Resources, College of Engineering and Mines, Tucson, AZ 85721-0011

Yolande Stowell, Environmental Resources Mangement, 8 Cavendish Square, London, W1M OER UK

Will Swearingen, 900 Technology Boulevard, Suite A, Bozeman, MT 59718-6857

Anne M. Thompson, NASA-Goddard Space Flight Center, Laboratory for Atmospheres, Greenbelt, MD 20771

Juan B. Valdés, University of Arizona, Department of Civil Engineering, Tucson, AZ 85721

Coleen Vogel, University of the Witwatersrand, School of Geography, Archaeology and Environment, Private Bag 3, Johannesburg, South Africa

Chris Walcek, University of Albany, Atmospheric Sciences Research Center, 251 Fuller Road, Albany, NY 12203

M. L Wesely, Argonne National Laboratory, Environmental Research Division, 9700 South Cass Ave, 203 ER, Argonne, IL 60439

MARTHA P. L. WHITAKER, University of Arizona, Department of Hydrology and Water Resources, Tuscon, AZ 85271-0011

DONALD A. WILHITE, University of Nebraska, Lincoln, NB 68583-0749

WILLIAM W.-G. YEH, University of California, Department of Civil and Environmental Engineering, 5731/5732 Boelter Hall, Box 951593, Los Angeles, CA 90095-1593

IGOR S. ZONN, ESIG/NCAR, P.O. Box 3000, Boulder, CO 80307

CONTENTS

SECTION 2 HYDROLOGY

Contributing Editor: Soroosh Sorooshian

SECTION 3 SOCIETAL IMPACTS

Contributing Editor: Michael H. Glantz

SECTION 1

ATMOSPHERIC CHEMISTRY

CHAPTER 1

OVERVIEW: ATMOSPHERIC CHEMISTRY

JACK FISHMAN

The study of atmospheric chemistry focuses on how chemical constituents cycle through the atmosphere. Excluding water vapor (which can account for as much as 2 to 3% of the volume of the atmosphere under extremely moist conditions), more than 99.9% of the remaining dry atmosphere is comprised of nitrogen (78.1%), oxygen, (20.9%), and argon (0.93%). Unlike the study of conventional meteorology, where the atmosphere is generally treated as a bulk medium, atmospheric chemistry focuses on each individual constituent (commonly referred to as trace gases) and the chemical reactions that take place among them.

When discussing atmospheric chemistry, it is perhaps most convenient to separate the discussion into two distinct chemical regimes: the stratosphere and the troposphere. In the stratosphere, the most important trace gas is ozone, O_3, whereas in the troposphere, it can be argued that one of the most important trace gases is carbon dioxide, CO_2. Both of these trace gases are intimately tied to the issue of global change as measurements over the past several decades confirm that stratospheric ozone is decreasing and that carbon dioxide is increasing. Ozone in the stratosphere is vital for shielding the biosphere from harmful ultraviolet radiation; a decrease in the amount of ozone in the stratosphere will result in damage to biota at the ground. On the other hand, carbon dioxide is an important trace gas (second in importance to water vapor) that keeps infrared radiation within the lower atmosphere, and it is generally agreed that an increase in CO_2 may have important climatic implications and lead to global warming.

The source of energy that drives the chemical processes in the atmosphere is the same source that drives Earth's weather engine, namely the sun. Furthermore, the high-energy ultraviolet radiation emitted by the sun initiates a series of reactions in

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

the upper atmosphere as these high-energy photons break the stable molecules, N_2 and O_2, apart into their atomic components. This high energy not only is capable of breaking these very strong molecular bonds apart, but it is also capable of stripping away electrons creating a source of ions in the atmosphere above $\sim 50\,\text{km}$. This region of the atmosphere is called the ionosphere, and its chemistry will not be discussed in this section. For more information about the chemistry of the ionosphere, mesosphere, and thermosphere, see Brasseur and Solomon's (1986) *Aeronomy of the Middle Atmosphere*, Chapter 6 and various sections in Chapter 5. These ions and atoms can feed some of the chemical cycles that take place in the stratosphere, such as supplying reactive nitrogen species (e.g., see Fig. 1).

From an atmospheric chemistry point of view, important cycles take place in both the stratosphere and the troposphere; this section will concentrate on the chemistry taking place in these regions of the atmosphere. To a certain extent, the chemistry of the stratosphere is somewhat less complex than the chemistry in the troposphere because only large-scale meteorological processes are present at these high altitudes; smaller scale processes such as precipitation can be generally neglected. Also important is the fact that the sources of trace species in the stratosphere are not determined from small-scale sources and can thus can be quantified using a simplified methodology.

In the stratosphere, observing and gaining an understanding of how the distribution of ozone evolved was the primary research emphasis from the 1930s through the 1960s. Understanding how its abundance and distribution has been perturbed by anthropogenic inputs has been the focus of intense research efforts since the 1970s.

1 STRATOSPHERIC CHEMISTRY: UNDERSTANDING THE OZONE LAYER

Ozone was discovered in 1839 by the German scientist Christian Frederich Schönbein at the University of Basel in Switzerland. Because of its pungent odor, its name was taken from the Greek word *ozein*, meaning "odor." Schönbein's research, subsequent to his discovery, focused on verifying his hypothesis that ozone was a natural trace constituent of the atmosphere. As a result of interest in the late nineteenth century, there are a surprisingly large number of ambient measurements during that time.

The primary study of ozone focused on the chemistry of the stratosphere when it was hypothesized and then verified that most of Earth's ozone was located at an altitude of 20 to 50 km (also called the ozonosphere) high above Earth's surface. The British physicist Sir Sidney Chapman put forth the premise that sufficiently intense ultraviolet radiation [at wavelengths (λ); $\lambda < 242$ nm) breaks apart molecular oxygen into two oxygen atoms. This reaction is commonly written:

$$O_2 + h\nu \rightarrow O + O \qquad \lambda < 242\ \text{nm} \tag{1}$$

where $h\nu$ is the standard notation for a photon.

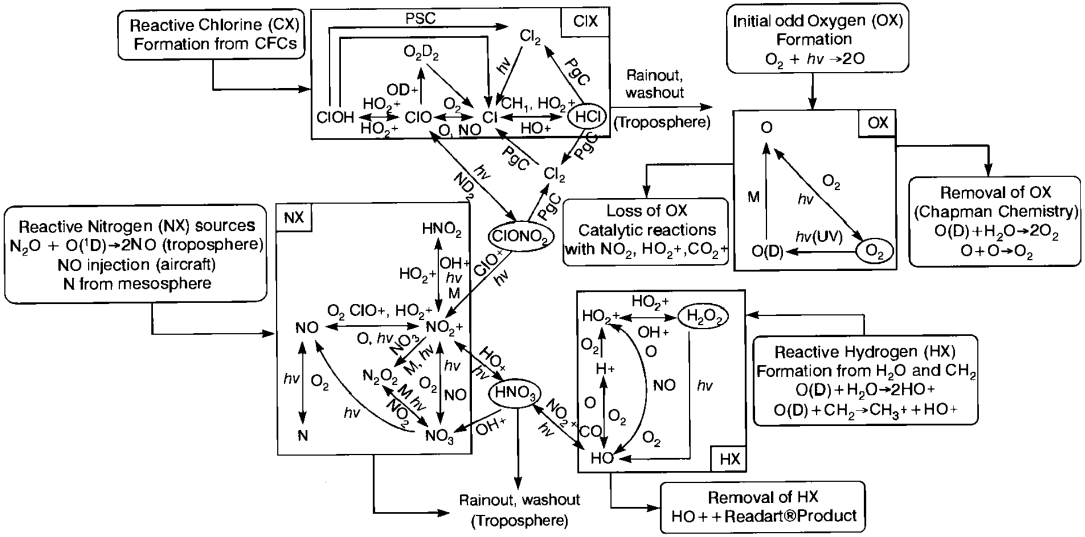


Figure 1 Simplified schematic diagram showing the interaction of the chemical families (large boxes) in the stratosphere. The round-cornered boxes define the sources and sinks for the chemical families. Reservoir (longer-lived) trace species are enclosed by ovals. Nitric acid (HNO_3) is a reservoir species for both reactive hydrogen (HX) and reactive (NX) families; chlorine nitrate ($ClONO_2$) is a reservoir species for the reactive chlorine (ClX) and NX families. The chemistry has been simplified by omitting bromine and iodine chemistry from the figure.

As the air becomes denser at lower altitudes in the stratosphere, most of this high-energy radiation is absorbed, and the oxygen molecules can no longer be broken apart. At these altitudes, the oxygen atoms will efficiently combine with the oxygen molecules and the formation of ozone occurs through the reaction:

$$O + O_2 + M \rightarrow O_3 + M \qquad (2)$$

where M is a nonreactive third body that absorbs any excess collisional energy that may be present. Thus, there is a preferred region in the atmosphere where sufficient ultraviolet energy is concurrently present with the proper amount of molecular density to create ozone, and the altitude region at which these processes are most prevalent is commonly referred to as the ozone layer.

Ozone can also be photolyzed in the atmosphere by weaker ultraviolet radiation ($\lambda < 320$ nm) to give back molecular and atomic oxygen:

$$O_3 + hv \rightarrow O_2 + O(^1D) \qquad \lambda < 320\ \text{nm} \qquad (3)$$

and also by visible radiation ($\lambda < 600$ nm) to yield atomic oxygen in its ground state, $O(^3P)$, rather than the more energetic $O(^1D)$ state; furthermore ozone can react with atomic oxygen (in either its ground or excited state) to give two molecules of oxygen:

$$O_3 + O \rightarrow 2O_2 \qquad (4)$$

To complete the possible reactions in a "pure oxygen" atmosphere, two atoms of oxygen can combine in a three-body reaction to give molecular oxygen back to the system:

$$O + O + M \rightarrow O_2 + M \qquad (5)$$

The set of five reactions involving only the various states of oxygen in the stratosphere are commonly referred to as "Chapman chemistry" and did a remarkable job of describing qualitatively why the ozone layer existed where it did. The speeds at which the five reactions took place in the atmosphere were measured independently in the laboratory and are called reaction rate constants (denoted k_4 for reaction 4, k_5 for reaction 5, etc.). Reaction rate constants are often temperature and pressure dependent. The rates of photolysis are noted by the letter j (e.g., j_3 for photolytic reaction 3, etc.) and are primarily dependent on the cross section of the individual molecule as a function of wavelength (those that have weaker bonds and can be broken apart more easily have larger cross sections) and the number of incident photons at those wavelengths (commonly called the photon flux).

As the field of chemistry progressed, other reactions were measured in the laboratory that were also believed to occur in the atmosphere with sufficient speed that they were eventually hypothesized to take an active role in the destruction and formation of ozone. These reactions dealt with derivatives of various forms of hydrogen in the stratosphere. The chemistry of the stratosphere was modified accordingly to account for this new "wet photochemistry," which involved reactions being measured in the laboratory, was in the 1950s and 1960s. The rationale behind this new chemistry was that atomic oxygen, $O(^1D)$, could react with water vapor to form the hydroxyl radical, OH:

$$H_2O + O(^1D) \rightarrow 2OH \tag{6}$$

Another important source of reactive hydrogen in the stratosphere is degradation of methane CH_4, by $O(^1D)$. Regardless of the initial source of the OH radical, it could then react with ozone to form another radical, HO_2, the hydroperoxy radical, which can lead to a catalytic cycle that becomes an efficient mechanism by which ozone can be removed from the atmosphere:

$$\begin{array}{cr} OH + O_3 \rightarrow HO_2 + O_2 & (7) \\ \underline{HO_2 + O_3 \rightarrow OH + 2O_2} & (8) \\ 2O_3 \rightarrow 3O_2 \quad \text{(net cycle)} & \end{array}$$

The above reactions helped to explain some of the observed differences between the measurements that were routinely made in the 1950s and 1960s and the calculated distribution of ozone determined from an oxygen-only atmosphere.

The next major modification to atmospheric chemistry came about from the inclusion of nitrogen chemistry into the reaction scheme of the stratosphere. Nitrous oxide, N_2O, was known to be a natural trace gas in the troposphere which did not have any identifiable removable mechanisms in lower atmosphere. Consequently, it could drift to the stratosphere where it was eventually attacked by the $O(^1D)$ atom to form nitric oxide, NO:

$$N_2O + O(^1D) \rightarrow 2NO \tag{9}$$

With the presence of NO in the stratosphere, another catalytic cycle of ozone destruction could occur through the following reaction sequence:

$$\begin{array}{rlr} & NO + O_3 \rightarrow NO_2 + O_2 & (10) \\ \text{followed by} & \underline{NO_2 + O \rightarrow NO + O_2} & (11) \\ \text{Net cycle} & O_3 + O \rightarrow 2O_2 & \end{array}$$

The importance of reactive nitrogen chemistry in the stratosphere was independently brought to light circa 1970 by Paul Crutzen, a recent Ph.D. in meteorology at the time from the University of Stockholm, and Harold Johnston, a chemistry professor at the University of California.

These catalytic ozone destruction cycles involving nitrogen and hydrogen species were the impetus behind the Climatic Impact Assessment Program (CIAP) of the

1970s, which became the rationale for determining the potential damage to the ozone layer that might result from flying a fleet of supersonic transport (SST) planes in the lower stratosphere. These planes would emit NO and H_2O directly into the stratosphere, and a confederation of U.S. federal agencies was charged with the task of determining how the ozone layer would be harmed by such a fleet. Although economic considerations eventually lay behind the decision for the United States not to pursue the development of a commercial fleet of SSTs, the environmental debate that developed during the early 1970s also contributed to the decision not to pursue the building of this new type of airplane.

But the environmental concern became even more of a reason to spend an increasing amount of money on stratospheric chemistry when Ralph Cicerone and Richard Stolarski, both at the University of Michigan in the early 1970s, introduced the possibility that chlorine chemistry might also provide another important means by which stratospheric ozone might be destroyed:

$$\begin{array}{rll} & Cl + O_3 \rightarrow ClO + O_2 & (12) \\ \text{followed by} & \underline{ClO + O \rightarrow Cl + O_2} & (13) \\ \text{Net cycle:} & O_3 + O \rightarrow 2O_2 & \end{array}$$

Shortly after the chlorine cycle was identified as a potential mechanism for stratospheric ozone destruction, Mario Molina and F. Sherwood Rowland, both chemists at the University of California at Irvine, proposed that a group of anthropogenic chlorine-containing compounds could provide the source of significant amounts of chlorine in the stratosphere (Molina and Rowland, 1974). These compounds, known as chlorofluoro-carbons ($CFCl_3$ and CF_2Cl_2) were used primarily in air-conditioning systems and as propellants for aerosol spray cans that proliferated the use of these compounds in the 1960s. These substances had no known removal mechanism in the troposphere, and Molina and Rowland hypothesized that their only eventual sink would be drifting to the upper stratosphere where they would be destroyed by high-energy ultraviolet radiation resulting in the release of their reactive chlorine atoms into the chemistry of the stratosphere. Figure 1 shows the chemical reactions within each reactive family [e.g., the reactive nitrogen family (NX) the reactive hydrogen family (HX), etc.] and also how each of these individual chemical cycles would influence stratospheric ozone chemistry. The circled trace gas in each box in Figure 1 is the longest-lived species for that particular reactive group. Chlorine nitrate ($ClONO_2$) and nitric acid (HNO_3) are long-lived trace gases that serve as reservoirs of more than one reactive family.

As predicted, the buildup in chlorine led to a "thinning" of the ozone layer. Not predicted by the atmospheric chemists, however, was that the depletion of ozone intensified in the Antarctic stratosphere because of the unique meteorological conditions there. Stratospheric dynamics are such that an enhanced circulation develops during austral winter, which severely inhibits meridional heat exchange (unlike the Northern Hemisphere, where the position of major mountain ranges closer to the pole results in a more favorable situation for heat from middle and low latitudes to be transported poleward). Thus, temperatures in the Antarctic lower stratosphere reach temperatures that are cold enough to allow for the formation of *polar stratospheric clouds* (PSCs) that provide ice surfaces that greatly perturb stratospheric chemistry by turning the long-lived (and relatively nonreactive) chlorine-containing compounds (chlorine nitrate, $ClONO_2$, and hydrochloric acid, HCl) into chlorine atoms, thereby greatly enhancing the destructive power of the reactive chlorine. Some of the main reactions that are influenced by PSCs are also shown in Figure 1 within the ClX box in the upper left of the figure. The net result has been the formation of the *ozone hole* whereby more than two-thirds of the normal amount of stratospheric ozone can be destroyed within a period few weeks as the austral winter ends (see Chapter 21, "Stratospheric Ozone Observations"). This phenomenon was first identified from ozonesonde measurements made by Joe Farman of the British Antarctic Survey in the early 1980s (Farman et al. 1985). By the early 1990s, more than 80% of the chlorine in the atmosphere was determined to be of anthropogenic origin (see Figure 2).

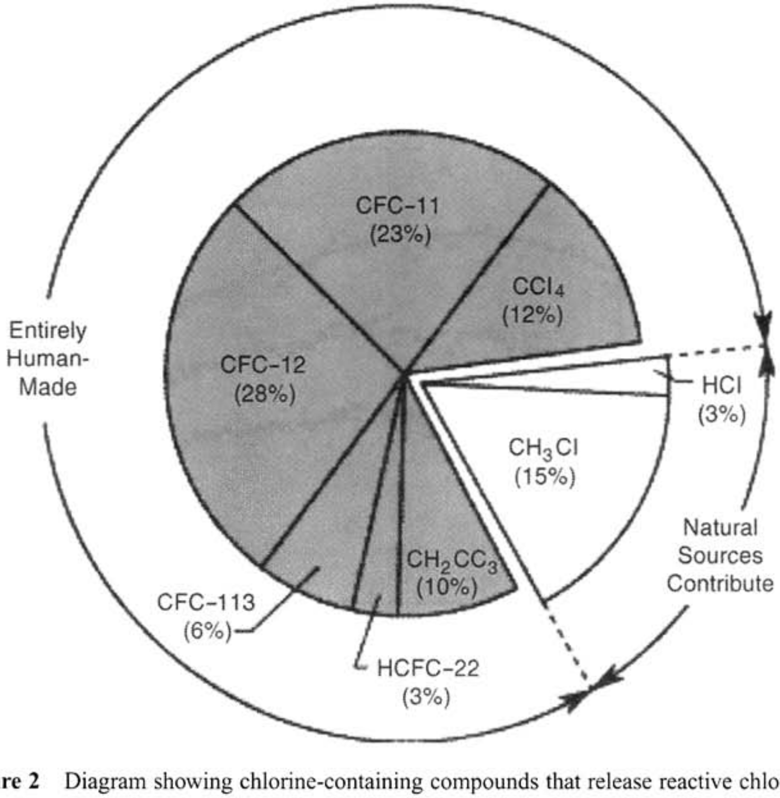


Figure 2 Diagram showing chlorine-containing compounds that release reactive chlorine to the stratosphere and the percentage that each contributes to stratospheric-reactive chlorine family (1994 estimates). Compounds that are completely produced by humans are shaded. Approximately 82% of the chlorine present in the stratosphere is of anthropogenic origin.

The environmental problem of stratospheric ozone depletion was successfully addressed by an international treaty in 1987 referred to as the Montreal Protocol, whereby a plan was set forth to phase out and eventually eliminate the manufacture and use of primary ozone-depleting chlorinated compounds (see Albritton et al., 1999). Figure 3 shows how the amount of man-made chlorine has decreased as a result of the effort to minimize the destruction of the ozone layer. As the amount of chlorine goes down in the stratosphere, model predictions suggest that the ozone hole should return to its pre-1980s level by the second or third decade of the new millennium. As a result of their important work on understanding the ozone layer and the chemical processes that drive the formation and destruction of ozone in both the stratosphere and the troposphere, Paul Crutzen, Mario Molina, and F. Sherwood Rowland were awarded the Nobel Prize for chemistry in 1995, the first time that atmospheric chemists received this coveted award. Whereas Rowland and Molina were both trained as chemists, Crutzen is the first meteorologist to receive the Nobel Prize.

Despite the complexity of Figure 1, it has been simplified by excluding the chemistry of two other halogen compounds: bromine and iodine. Reactive family chemistry of these halogens is similar to that shown for the reactive chlorine family.

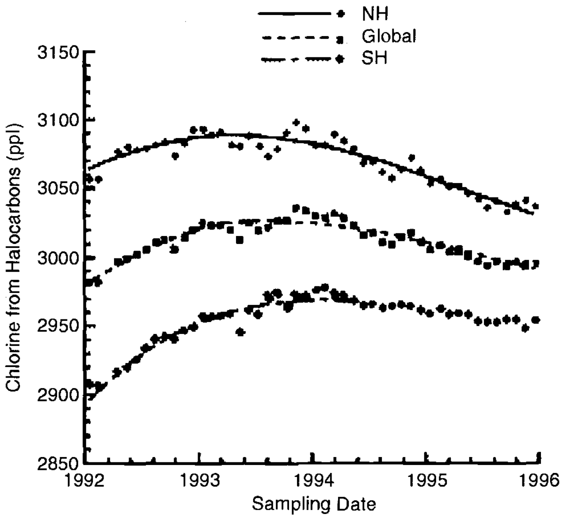


Figure 3 Monthly hemispheric and global tropospheric chlorine content measured between 1992 and 1996. Because of the international effort to curb human-produced chlorinated gases outlined in the Montreal Protocol of 1987 and subsequent amendments to that original agreement, the amount of chlorine in the troposphere showed an annual global decrease for the first time in the early 1990s. It is expected that stratospheric chlorine content will decline in the early 2000s, at which point the amount of ozone in the stratosphere should end its long-term declining trend.

Anthropogenic bromine compounds (called halons) are used as fumigants in agriculture and as a fire retardant on clothing. The most abundant iodine compound is methyl iodide, CH_3I, and has been observed in the atmosphere as a biomass burning product.

2 TROPOSPHERIC CHEMISTRY: A COMPLEX INTERACTION OF BIOGEOCHEMICAL CYCLES

Somewhat analogous to the chemical cycles just described for the stratosphere, atmospheric chemistry in the troposphere can also be viewed as a complex interaction of chemical cycles. These cycles, however, involve direct interaction with the biosphere and are also greatly complicated by the presence of the more complicated meteorology found only in the lower atmosphere. The biosphere serves as the exclusive source of carbon, and the carbon cycle, in turn, has important linkages that transcend land, ocean, and air.

The dominant form of carbon in the atmosphere is its completely oxidized state, carbon dioxide, which is present in the atmosphere at concentrations of ~ 360 ppmv (parts per million, by volume). Between the time carbon is stored in the biosphere and eventually bccomes CO_2, many interesting chemical transformations and interactions take place. The second most abundant carbon-containing trace gas is methane, with an atmospheric concentration of ~ 1.8 ppmv, the only other atmospheric trace gas that exists in the concentrations of more than 1 ppmv, even in regions far removed from its sources. From an atmospheric chemistry point of view, methane plays an important role because its oxidation is closely linked to other carbon-containing trace compounds such as carbon monoxide (CO) and formaldehyde (CH_2O). The role of carbon dioxide, on the other hand, is not directly tied to chemical reactions taking place in the atmosphere, but rather to assessing how natural and anthropogenic processes contribute to increasing the global carbon burden.

The carbon cycle that couples the atmosphere with the biosphere is one of several important biogeochemical cycles in the Earth system. These cycles describe how specific elements and compounds are transferred between the principal global reservoirs—the atmosphere, land, oceans, and biosphere. Chemical and physical processes and transformations determine the partitioning of a material among the reservoirs. For example, for a fixed amount of carbon in these reservoirs, a certain fraction is found in the atmosphere, another fraction in the oceans, and so on; these fractions depend on the way the carbon is transferred between the reservoirs, the sizes of the reservoirs, and other factors. The amount to be found in the atmosphere, where it may have the most substantial effect on climate, is thus determined by the carbon's overall biogeochemical cycle.

The size of these cycles is enormous. For example, the carbon cycle transfers more than 10^{15} g of carbon per year from the atmosphere to other reservoirs. The terrestrial biosphere is dominated by trees and other vegetation; the amount of land biomass in animal form is much smaller, by approximately a factor of 100. Living biomass on land is also ~100 times as massive as the total living biomass in the oceans. Dead biomass is even more plentiful than living biomass. Inanimate organic matter accumulates as plant litter in forests, building up as a layer of humic soil or peat. Eventually, the organic debris is consumed by microorganisms as it decays, or it may be burned; in either case, carbon dioxide is put into the atmosphere. The details of these cycles are complex, and many of the chapters in this section focus on specific key trace gases that are parts of these cycles. Other chapters focus on specific processes that are responsible for the conversion of trace gases to other species that may or may not remain in gaseous form. These processes, however, may provide the dominant vehicle through which various elements are transferred between the various reservoirs within the biogeochemical cycle (e.g., sulfur being removed from the atmosphere and transferred to the land through the formation of acid rain).

In addition to carbon dioxide and methane, the important members of the carbon cycle are carbon monoxide (CO), and a host of nonmethane hydrocarbons [also called volatile organic compounds (VOCs)] consisting of more than one carbon atom and a number of hydrogen atoms (C_nH_m, where n and m are integers). Other important cycles in Earth's atmosphere include nitrogen, oxygen, and sulfur. In the nitrogen (N) cycle, the key compounds are nitrogen (N_2), nitrous oxide (N_2O), nitrogen oxides (NO_x), consisting primarily of nitric oxide (NO) and nitrogen dioxide (NO_2), nitric acid (HNO_3), and the nitrate ion $(NO_3)^-$. Both molecular nitrogen and nitrous oxide are key elements of the biogeochemical nitrogen cycle, although neither is involved in any chemical reactions in the troposphere.

The oxygen cycle in the troposphere is comprised nearly exclusively of molecular oxygen (O_2), and ozone (O_3); note that atomic oxygen is not an important player in the troposphere despite its importance on the stratosphere. The primary players in the sulfur cycle are sulfur dioxide (SO_2), carbonyl sulfide (COS), hydrogen sulfide (H_2S), dimethyl sulfide $(CH_3)_2S$, sulfuric acid (H_2SO_4), and the sulfate ion $(SO_4)^{-2}$.

This section on atmospheric chemistry will include separate articles on reactive nitrogen species, tropospheric ozone, and atmospheric sulfur, as well as the carbon compounds carbon monoxide and methane and nonmethane hydrocarbons. A broad discussion will now be presented on the carbon, nitrogen, and oxygen cycles so that the reader will be better able to envision how the individual trace gases and processes fit into the "big picture."

3 GLOBAL CARBON CYCLE

Measurements of the concentration of air trapped in Antarctic ice cores indicate that over the last 200,000 years, atmospheric concentrations of CO_2 have fluctuated between 200 and 280 ppmv, until the last century. Data for the period AD 1000 and 1800 indicate that the concentration was quite stable, averaging 280 ppmv and varying over that period by only about 10 ppmv, indicating that the CO_2 cycle was in equilibrium in the centuries prior to the industrial revolution. Over the past 200 years, however, the concentration has increased from 280 to more than

360 ppmv; this increase is attributed primarily to burning of fossil fuels (primarily at northern middle latitudes) and tropical biomass burning. Through an examination of the isotopic composition of the CO_2 in the ice core samples, it can likewise be shown that nearly all of this increase is a result of fossil fuel combustion. Carbon has two isotopes, one with a molecular weight of 12 and the other with a molecular weight of 13. Naturally occurring carbon dioxide that has been put into the atmosphere through photosynthesis will be comprised of $\sim 1\%$ of the heavier ^{13}C; CO_2 that has been put into the atmosphere from fossil fuel burning will be slightly depleted in ^{13}C. Analysis of the isotopic $^{13}CO_2$-to-$^{12}CO_2$ ratio from air trapped in ice cores indicates that a smaller fraction of the CO_2 released to the atmosphere before the industrial revolution came from fossil fuel sources when compared to the modern-day ratios.

Currently, approximately 6 GtC (1 GtC = 10^{12} kg of carbon) are released to the atmosphere as a result of fossil fuel combustion. Between about 1850 and the early 1970s, the release of CO_2 increased exponentially at a relatively constant rate of 4.3% per year (Fig. 4). Since the oil crisis of 1973, the concerted efforts to reduce energy consumption have slowed the trend, but, nonetheless, an upward trend still continues, especially since the late 1980s. The cumulative production of CO_2 from fossil fuel is estimated to be 225 GtC, or ~30% of the current amount of CO_2 in the atmosphere. The result of this increase to the atmosphere is reflected in both ambient

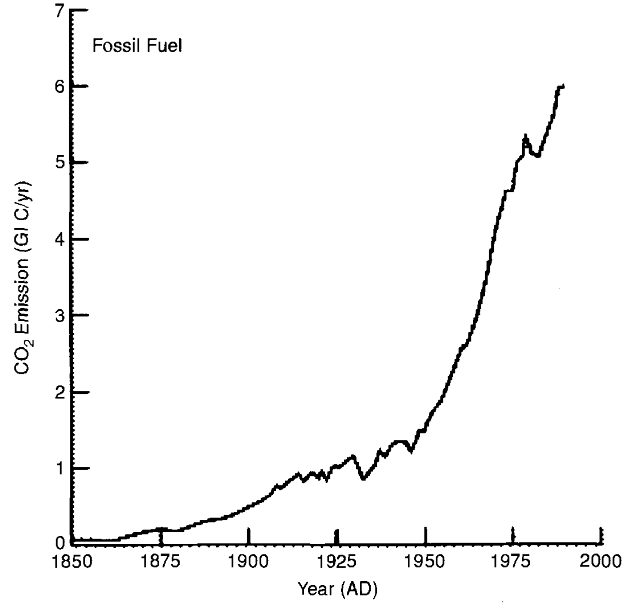
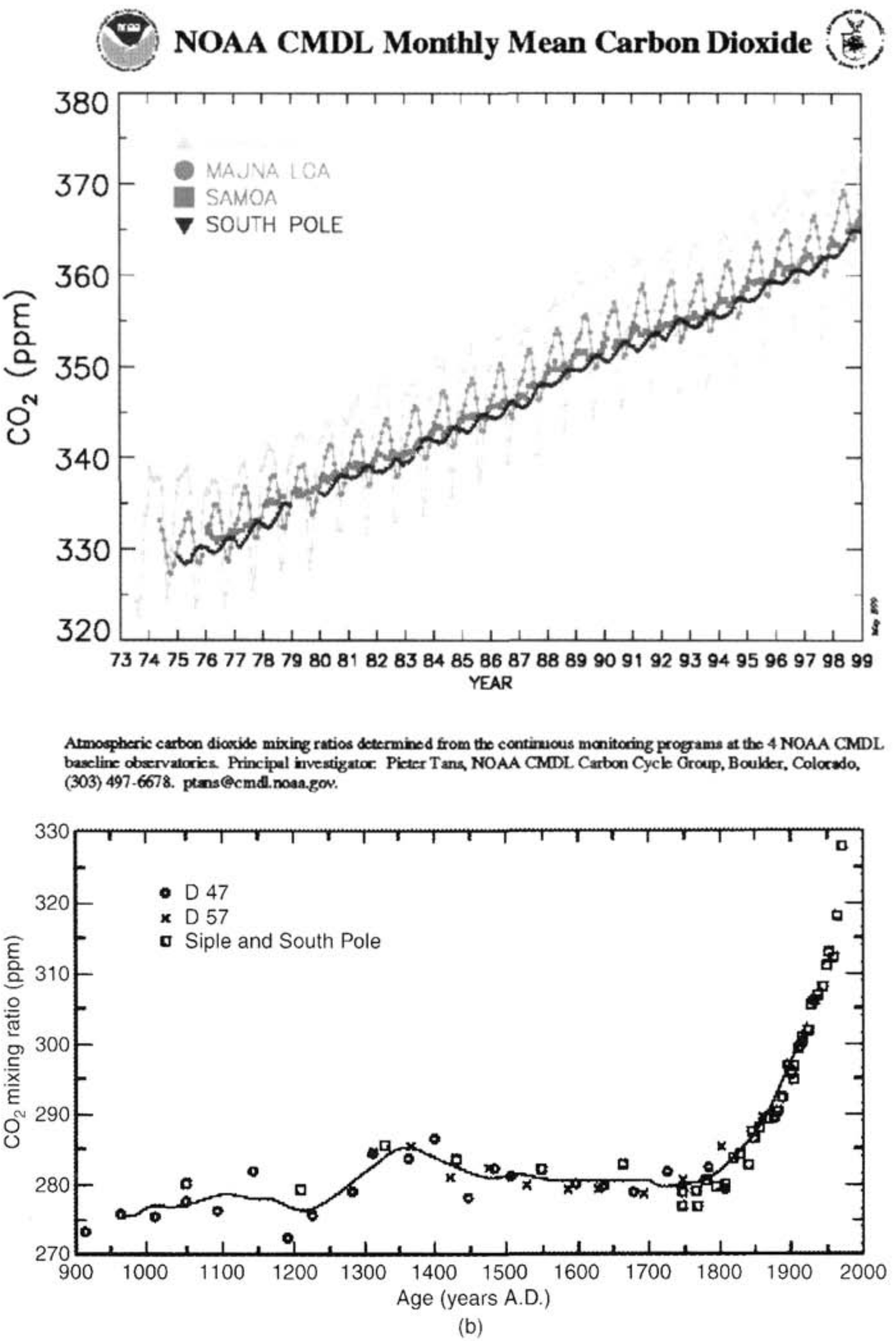


Figure 4 Estimates of CO_2 released to the atmosphere from fossil fuel combustion from 1850 to the present.



Atmospheric carbon dioxide mixing ratios determined from the continuous monitoring programs at the 4 NOAA CMDL baseline observatories. Principal investigator: Pieter Tans, NOAA CMDL Carbon Cycle Group, Boulder, Colorado, (303) 497-6678. ptans@cmdl.noaa.gov.

Figure 5 (see color insert) (*a*) Monthly concentrations of CO_2 measured from gas samples at four monitoring sites operated by NOAA's Climate Monitoring and Diagnostics Laboratory from the early 1970s; (*b*) CO_2 concentrations determined from ice core samples estimated to go back ~1000 years. See ftp site for color image.

measurements from several monitoring sites of NOAA's Climate Monitoring and Diagnostics Laboratory (see Fig. 5a) and from ice core data (Fig. 5b). The current rate of CO_2 increase is ~1.8 ppmv per year.

4 GLOBAL CARBON BUDGET

Balancing the global carbon budget is a challenging effort, but large strides have been made in recent years. One of the largest problems is the difficulty of measuring carbon fluxes over large scales as well as accurately modeling atmospheric and oceanic transport of carbon species. The global carbon cycle is shown schematically in Figure 6. In this figure, the numbers in the reservoirs are given in GtC and the fluxes in GtC per year. This figure shows that the *gross* exchange flux between the ocean and the atmosphere (on the order of 90 GtC/yr), and between the terrestrial biosphere and the atmosphere (on the order of 60 GtC/yr) are at least an order of magnitude larger than the CO_2 emissions from fossil fuel burning (5.5 GtC/yr) and from deforestation (net of 1.6 GtC/yr). On the other hand, the total anthropogenic input of CO_2 to the atmosphere (7.1 GtC/yr) is significant compared to the *net* exchange fluxes between the three carbon reservoirs. In particular, the net CO_2 flux into the terrestrial biosphere is the subject of an ongoing debate. Instead of attempting to deal with the complete carbon budget, it is easier to limit this discus-

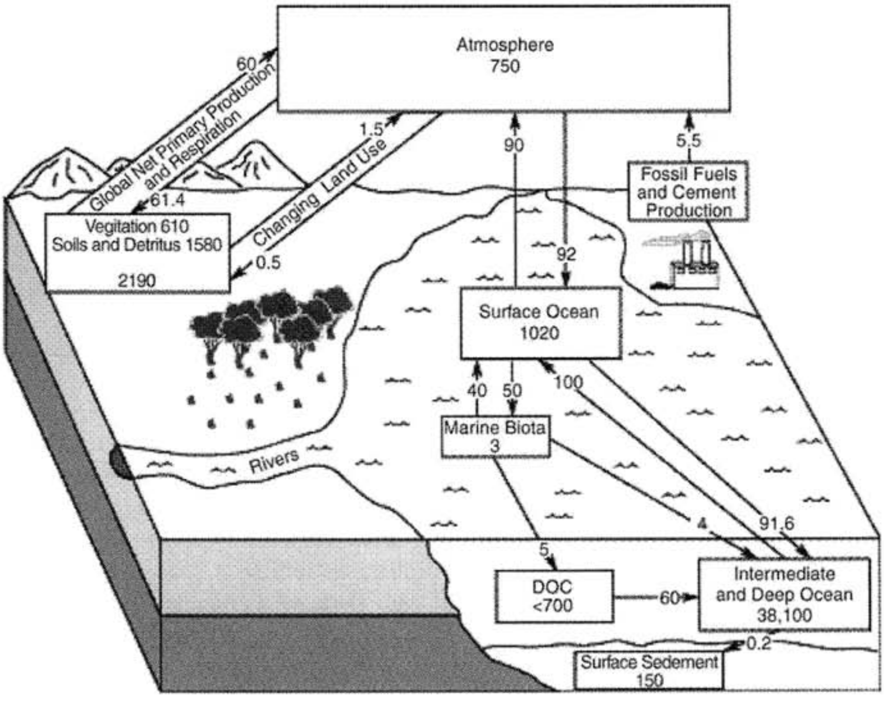


Figure 6 Schematic diagram showing the size of the carbon reservoirs and the amount of annual exchange between them.

sion to the *perturbation budget* for CO_2 (i.e., what happens to the CO_2 injected into the atmosphere by the combination of fossil fuel and biomass combustion and land-use change).

The single value in the global cycle that is known most accurately is the change of atmospheric CO_2 concentration, which has been measured continuously and averaged over a number of sites; it corresponds to 3.3±0.2 GtC/yr. Fossil fuel combustion can also be estimated fairly accurately (through a knowledge of global coal and petroleum production) at a value of 5.5±0.5 GtC/yr. Estimates of CO_2 contributions resulting from deforestation, primarily in tropics, are quite uncertain, ranging from 0.6 to 2.5 GtC/yr. Using an average value of 1.6±1.0 GtC/yr, the Intergovernmental Panel on Climate Change estimates the average sources of anthropogenic CO_2 to the atmosphere as 7.1±1.1 GtC/yr. Since 3.2±0.2 GtC/yr accumulate in the atmosphere, the remaining 3.9 GtC/yr must be reabsorbed either by the oceans or by the terrestrial biosphere. Current models calculate an oceanic uptake of 2.0±0.8, leaving an imbalance (or "missing sink") of 1.4±1.5 GtC/yr. A considerable amount of research in recent years has been directed toward partitioning this missing sink into ocean and land components.

Research on many of the individual scientific issues is currently being conducted by numerous scientists throughout the world and the scientists involved in this research transcend a number of disciplines such as atmospheric science, ecology, microbiology, and others. These scientists are brought together to foster research related to the issues involving global change through the International Geosphere–Biosphere Program (IGBP), under the sponsorship of the International Council of Scientific Unions. IGBP has several "core" programs and one of them is the International Global Atmospheric Chemistry (IGAC), which focuses on tropospheric chemistry. Much of the discussion in the remainder of this chapter and in the other chapters in this section discuss results and research that are linked to the goals of IGBP.

5 ATMOSPHERIC CHEMISTRY WITHIN GLOBAL CARBON CYCLE

Despite its dominance as a member of the global carbon cycle, carbon dioxide does not play any significant role in atmospheric chemistry. The most abundant carbon species that is an active player in atmospheric chemistry is methane (CH_4), which reacts in the troposphere with the hydroxyl radical (OH) to initiate a series of reactions that were shown to play important roles in the global cycles of a number of atmospheric trace species. The seminal work in tropospheric chemistry was published by Hiram Levy in 1971. Levy showed that the hydroxyl radical should exist in sufficient quantities in the troposphere to initiate a sequence of photochemical reactions that both produce and destroy a number of important tropospheric trace gases.

The initial formation of OH comes from the photolysis products of ozone in the troposphere:

$$O_3 + h\nu \rightarrow O_2 + O(^1D) \qquad \lambda < 320\ \text{nm} \qquad (3)$$

followed by its reaction with water vapor:

$$H_2O + O(^1D) \rightarrow 2OH \qquad (6)$$

Hydroxyl then reacts with methane to form a host of products:

$$CH_4 + OH \rightarrow CH_3 + H_2O \qquad (14)$$
$$CH_3 + O_2 + M \rightarrow CH_3O_2 + M \qquad (15)$$
$$CH_3O_2 + NO \rightarrow CH_3O + NO_2 \qquad (16)$$
$$CH_3O + O_2 \rightarrow CH_2O + HO_2 \qquad (17)$$
$$CH_2O + h\nu \rightarrow CO + H_2 \quad \lambda < 360\,\text{nm} \qquad (18)$$

This sequence, commonly referred to as methane oxidation was hypothesized to be a major source of formaldehyde (CH_2O) and carbon monoxide (CO). In addition, other radicals such as CH_3O_2 (methyl peroxy) and HO_2 (hydroperoxy) were formed and became important factors in the tropospheric ozone budget. Once formaldehyde (CH_2O) was formed, it could photolyze in an alternate pathway to produce even more reactive radicals:

$$CH_2O + h\nu \rightarrow HCO + H \quad \lambda < 360\,\text{nm} \qquad (19)$$

followed by

$$H + O_2 + M \rightarrow HO_2 + M \qquad (20)$$
$$HCO + O_2 + M \rightarrow HO_2 + CO + M \qquad (21)$$

Note that either photolysis sequence of CH_2O results in the formation of carbon monoxide.

Like CO_2, methane also absorbs infrared radiation and contributes to global warming. Ice core data show that atmospheric CH_4 concentrations remained relatively constant at about half of its present form for thousands of years before beginning to increase about 200 years ago from ~0.65 ppmv to ~1.8 ppmv (Fig. 7). Methane's atmospheric lifetime is ~8 years and its dominant removal mechanism is oxidation by OH.

Methane is produced in oxygen-deficient environments of Earth's surface (swamps, lakes, rice paddies, tundra, boreal marshes, etc.). Methane production in soils and oceans is the end product of a variety of reductive pathways during the decomposition of organic matter. Methane is also released by cattle, termites, and perhaps other insects, whereas coal mining, natural gas losses, and solid-waste burning are important anthropogenic sources. Large amounts of methane are also produced by biomass burning. The global methane budget is given in Table 1.

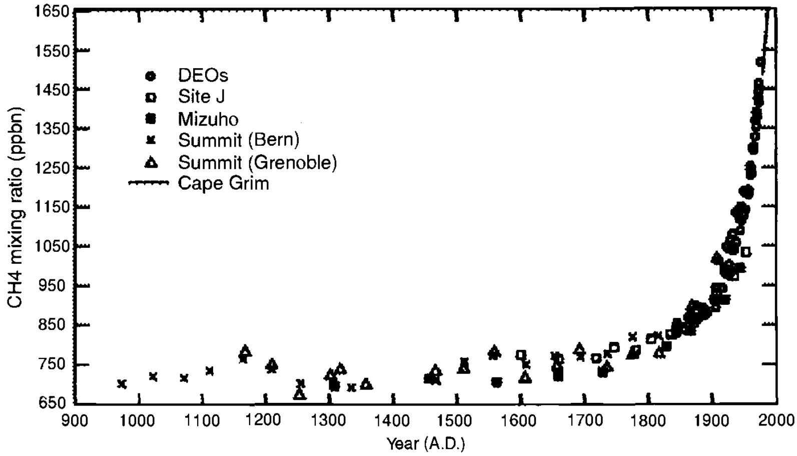


Figure 7 Ice for samples of atmospheric methane showing only a slight increase from 950 to 1800, but a sharp rise in concentration over the last 150 years.

One of the important products of methane oxidation is CO, which can also be oxidized by OH to form CO_2:

$$CO + OH \rightarrow CO_2 + H \tag{22}$$

Thus, CH_4, CO, and CO_2 are linked together through a series of oxidation processes that take place in the atmosphere; all forms of carbon emitted to the atmosphere eventually become CO_2.

As can be seen from the above discussion, carbon monoxide is caught in the middle as an intermediate oxidation product. One of the fundamental questions in tropospheric chemistry is to determine how much CO is emitted directly to the atmosphere, relative to the amount that is produced in situ through CH_4 oxidation. The primary and only significant sink for CO is removal by OH, which leads to an atmospheric residence on the order of 1 to 2 months. Thus, CO can be used as a useful tracer for atmospheric transport processes that take place on times scales of several days to a week or so.

6 CARBON MONOXIDE, NITROGEN OXIDES, AND OXIDIZING CAPACITY OF TROPOSPHERE

In some of the chapters that follow, there are detailed discussions on many individual trace gases such as CO, the oxides of nitrogen, and tropospheric ozone. Additionally, these three trace gases are of particular interest as they interact to a major degree to determine the oxidizing capacity of the troposphere. One of the most important

TABLE 1 Estimated Sources and Sinks of Methane

Sources	Magnitude ($Tg\ CH_4$ per year)	Range ($Tg\ CH_4$ per year)
Natural		
Wetlands	115	100–200
Termites	20	10–50
Ocean	10	5–20
Freshwater	5	1–25
CH_4 hydrate	5	0–5
Anthropogenic		
Coal mining, natural gas and petroleum industry	100	70–120
Rice paddies	60	20–150
Enteric fermentation	80	65–100
Animal wastes	25	20–30
Domestic sewage treatment	25	?
Landfills	30	20–70
Biomass burning	40	20–80
Total sources	515	
Sinks		
Atmospheric removal	470	450–520
Removal by soils	30	15–45
Atmospheric increase	32	28–37
Total sinks	532	

Source: Watson et al., 1992.

cycles that comes into play in the troposphere is the formation of ozone from carbon monoxide oxidation:

$$CO + OH \rightarrow CO_2 + H \tag{22}$$

$$H + O_2 + M \rightarrow HO_2 + M \tag{20}$$

$$HO_2 + NO \rightarrow OH + NO_2 \tag{23}$$

$$NO_2 + h\nu \rightarrow NO + O \qquad \lambda < 420\ \text{nm} \tag{24}$$

$$\underline{O + O_2 + M \rightarrow O_3 + M} \tag{25}$$

$$CO + 2O_2 + h\nu \rightarrow CO_2 + O_3 \qquad \text{(net cycle)}$$

This relatively simple catalytic cycle shows how the global budgets of CO and ozone in the troposphere are intertwined if there is a sufficient amount of NO and NO_2 present in the atmosphere. Since it has already been demonstrated that the CO and CH_4 budgets are linked, one of the driving questions in atmospheric chemistry is the determination of what percentage of all these trace gases is natural, what fraction is anthropogenic, how have these budgets been perturbed over the past decades and centuries, and, lastly, how much will these budgets change in the coming decades. Through the complex interactions of these trace gases, the oxidizing capacity of the troposphere can be determined and possibly even predicted.

Earth is the only planet in this solar system where there is an oxidizing atmosphere, and although nearly all carbon emitted to the atmosphere eventually ends up as completely oxidized CO_2, and all hydrogen-containing trace species end up as H_2O, some interesting and often complex chemistry takes place along these oxidation pathways. Sulfur and nitrogen also eventually become oxidized, and the final products result in the formation of acids that, being soluble, contribute to the formation of acid rain. In recent years, chemical analysis of rain, fog, clouds, and dew have shown that the aqueous chemistry is as equally challenging and even more complicated than gas-phase chemistry. A complete understanding of aqueous-phase chemistry is still evolving, but many of the basic principles have become fairly well established and are described in more detail in some of the chapters in this section.

The most important species in clouds and precipitation is the hydrogen ion, whose concentrations can be indicated by specifying solution acidity, or pH. The presence of atmospheric CO_2 assures that nearly all atmospheric water droplets will be acidic; natural and anthropogenic nitrogen and sulfur increase the acidity (i.e., lower the pH value) to at least pH 5.0. Many urban areas experience pH levels nearer 4.0. Cloud and fog droplets are nearly always more acidic than rain, apparently because smaller cloud drop sizes inhibit dilution of the acidic constituents. In some fogs, the pH of the droplets has been measured as low as 1.7.

Organic compounds in the atmosphere also contribute to cloud acidity. Formaldehyde (CH_2O), often found in high concentrations where urban pollution is present, is a key tropospheric species with sufficiently high solubility that it can affect the acidity of rain. In remote regions, forests are known to emit large quantities of isoprene (C_5H_8), which can react with OH or O_3 to form more complex aldehydes (RCHO), and also resulting in the measurement of acid rain in the range of pH 5.0. Other carbon-, nitrogen-, and sulfur-containing acids also exist and are discussed in the chapters on acid rain, reactive nitrogen species, and sulfur species. The microphysics by which these trace gases are converted to both gaseous and aqueous forms of these acids is also treated explicitly in several chapters in this section.

Returning to the central theme of this overview, a considerable amount of research has been conducted over the past several decades to determine the budgets of carbon monoxide, methane, nitrogen oxides, and tropospheric ozone and to determine how these budgets are affected by human activity. Because these species do interact with each other, a series of different conclusions has been reached regarding human influence on tropospheric chemistry cycles. In the early 1970s, a series of research studies extrapolated Levy's initial hypothesis to postulate that many important chemical processes in the unpolluted atmosphere were dominated by (natural) methane chemistry. Since those initial studies, however, the atmospheric chemistry community has come to recognize that the natural background atmosphere and the concentrations of naturally occurring trace species such as methane, tropospheric ozone, and carbon monoxide have increased dramatically, at rates comparable to, and, in most cases, even more than, carbon dioxide, the hallmark of proof of the concept of global change.

7 ATMOSPHERIC CHEMISTRY AND GLOBAL WARMING

The links of trace gas chemistry to climate change extend beyond the observed increase in CO_2 concentrations. Increases in CH_4, tropospheric O_3, and anthropogenically produced concentrations of the chlorofluorocarbons (CFCs) also contribute to global warming (see Fig. 8). If only the direct radiative effects of the trace gas increases are considered, 62% of the increase would be due to CO_2, 20% to CH_4, 4% to N_2O, and 14% to the CFCs. If chemical feedbacks are considered, the global warming due to changes with respect to changes in ozone concentrations (in both the stratosphere and troposphere) result in changes in ozone being as important as the changes in methane. With the international cooperation now in place to phase out the production and use of CFCs, future scenarios indicate that tropospheric ozone increases will replace methane as the second most important trace gas that contributes to the greenhouse effect (see Houghton et al., 1990; Houghton et al., 1996).

Furthermore, because tropospheric ozone is a relatively short-lived gas, especially when compared to other gases that contribute to global warming, increases in its concentration may have regional and seasonal effects that must be accounted for properly when temperature perturbations are being computed. Another consideration for regional climate change is the presence of particulates that are produced by fossil fuel combustion and biomass burning. Particles screen some of the incoming solar radiation and therefore result in a regional cooling effect. Studies to date show that both tropospheric ozone and man-made aerosols must be included in any scenarios attempting to simulate global climate and that the regional effects caused by these two constituents are comparable to the global perturbation of the longer-lived trace gases.

Lastly, the future buildup of methane and hydrogenated CFCs, the replacement gases to the CFCs, which can be removed from the atmosphere by OH oxidation, is complicated by the potential change in the oxidizing capacity of the troposphere. As O_3 increases in the troposphere, it is likely that the global abundance of OH will also increase, since the primary formation mechanism of OH in the troposphere is initiated by the photolysis of O_3 and the subsequent reaction of the excited oxygen atom with water vapor. If more OH is present, the removal of CH_4 and the replacements for the CFCs, which are removed in the troposphere by OH, becomes more efficient and thus their rate of increase is slowed.

8 STRATOSPHERE–TROPOSPHERE CHEMICAL AND CLIMATE INTERACTION

Although it is convenient to describe the chemistry of the stratosphere and the chemistry of the troposphere separately, both chemical and meteorological processes in one domain play an important role on the chemistry in the other. As previously

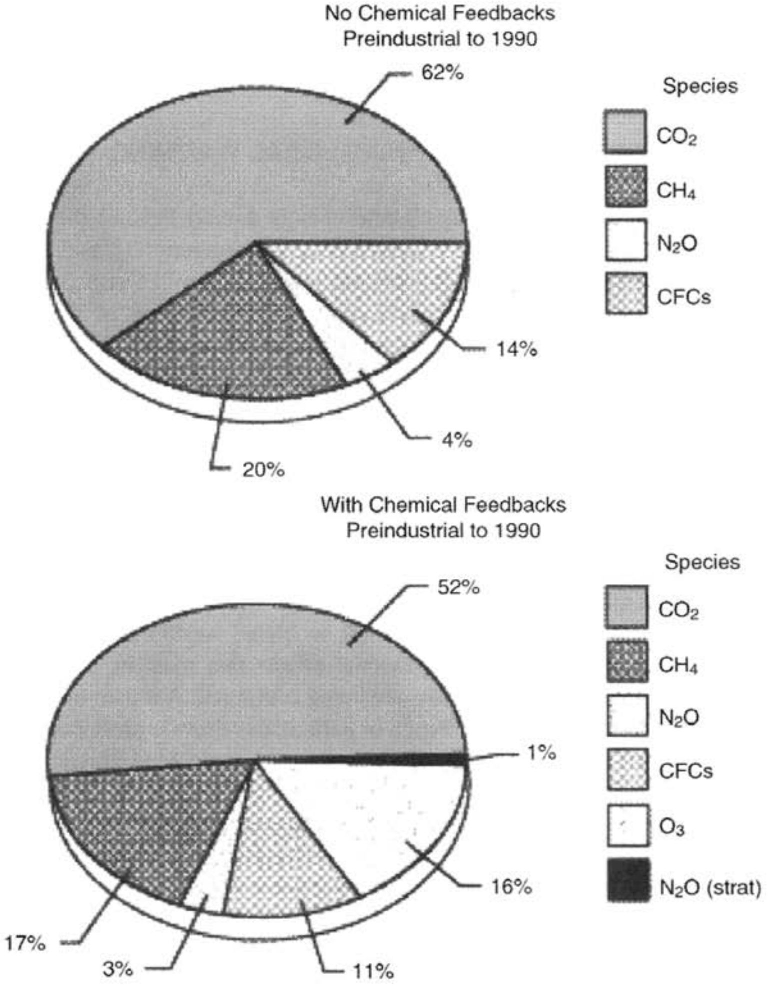


Figure 8 Contributions of various trace gases to global warming. Top graph shows the calculations using a model that does not include contributions from tropospheric ozone increases or from feedback relating to photochemical processes; bottom chart shows contributions using a model that includes photochemical feedback mechanisms. See ftp site for color image.

mentioned, one important aspect of tropospheric chemistry that impacts stratospheric chemistry is the removal of ozone-depleting chlorine in the troposphere before reaching the stratosphere.

Ozone depletion potentials (ODPs) provide a relative measure of the expected impact on ozone per unit mass emission of a gas compared to that expected from the same mass emission of CFC-11 integrated over time. Their primary purpose is for comparison of relative impacts of different gases upon ozone (e.g., for evaluating the relative effects of choices among CFC substitutes upon ozone). The two factors that contribute to how effectively an anthropogenic compound depletes ozone is how much of it reaches the stratosphere and how much chlorine each molecule contains. CFC-11 (trichlorofluoromethane) contains three chlorine atoms and is not removed in the troposphere through chemical reactions. The primary replacements for many of these compounds contain a hydrogen atom, which can be attacked by OH in the troposphere. The lifetimes of these compounds are also presented in Table 2, which was compiled as part of the International Assessment of Ozone Depletion in 1994 (Albritton et al., 1995). Some of the compounds, such as HFC-134a, the primary choice for refrigerant in many automobile air-conditioning systems, contain no chlorine, and thus, virtually no chance of contributing to ozone depletion.

In addition to ODPs being established, the 1994 ozone assessment also produced a series of global warming potentials (GWPs) to provide a simple representation of the relative radiative forcing resulting from a unit mass emission of a greenhouse gas compared to a reference compound. Because of its central role in concerns about climate change, carbon dioxide has generally been used as the reference gas. The values presented in Table 2 are calculations based on a time horizon of 20 years. Evaluations of GWPs must also take into consideration the radiative property of the atmosphere at some future point, including the concentration of CO_2, and other major climate altering compounds such as nitrous oxide and methane. In the 1994

TABLE 2 Estimated Lifetime, Ozone Depletion Potential (ODP), and Global Warming Potential (GWP) for Various Anthropogenic Trace Gases

Trace Gas	Chemical Formula	Lifetime (years)	ODP	GWP
CFC-11	$CFCl_3$	50	1.0	5000
CFC-12	CF_2Cl_2	102	0.82	7900
CFC-113	$C_2F_3Cl_3$	85	0.90	5000
CFC-114	$C_2F_4Cl_2$	300	0.85	6900
CFC-115	C_2F_5Cl	1700	0.40	6200
Carbon tetrachloride	CCl_4	42	1.20	2000
Methyl chloroform	CH_3CCl_3	5.4	0.12	360
HCFC-22	CF_2HCl	13.3	0.04	4300
HCFC-123	$C_2F_3HCl_2$	1.4	0.014	300
HCFC-124	C_2F_4HCl	5.9	0.03	1500
HCFC-141b	$C_2F_3H_3Cl$	9.4	0.10	1800
HCFC-142b	$C_2F_3H_3Cl$	19.5	0.05	4200
HCFC-225ca	$C_3F_5HCl_2$	2.5	0.02	550
HCFC-225cb	$C_3F_5HCl_2$	6.6	0.02	1700
HCFC-134a	CH_2FCF_3	14	$< 1.5 \times 10^{-5}$	3300
HCFC-23	CHF_3	250	$< 4 \times 10^{-4}$	9200
HCFC-125	C_3HF_5	36	$< 3 \times 10^{-5}$	4800
Methyl bromide	CH_3Br	1.3	0.64	6200
Halon-1301	CF_3Br	65	12	
Halon-1211	CF_2HBr	20	5.1	

ozone assessment and the Intergovernmental Panel on Climate Change (IPCC) (Houghton et al., 1996) report, there are also GWPs calculated with time horizons of 100 and 500 years, and such calculations include even more uncertainty than the values presented here because of the assumed scenarios for emissions so far into the future. Thus, although some of the replacement compounds for the CFCs have a negligible impact on the ozone layer, they will make important contributions to the overall greenhouse effect caused by the emission of anthropogenic chemicals released to the atmosphere.

9 STRATOSPHERE–TROPOSPHERE EXCHANGE

The exchange of mass between the stratosphere and troposphere is important to the chemistry of both regions as it brings chemical species with sources in the troposphere (such as CFCs) into the stratosphere, while species with stratospheric origin (such as ozone) can be brought into the troposphere. Thus, the transport can be important for driving the chemistry in both regions. Analogous to the boundary layer being isolated from the free troposphere because of the presence of a substantial inversion, the troposphere is isolated from the stratosphere by the high static stability of the stratosphere. Similarly, just as the boundary layer is turbulent and well mixed compared to the free troposphere, the troposphere is relatively well mixed vertically and horizontally compared to the stratosphere. The mixing time in the troposphere (on the order of months within each hemisphere; on the order of a year between the hemispheres) is much shorter than the time required to exchange the mass of the entire troposphere with the stratosphere (on the order of 18 years, although due to the difference in mass the entire stratosphere mixes with the troposphere every 2 years).

The simplest way to visualize a model for stratosphere–troposphere exchange is to consider bulk exchange between the two domains accomplished by uniform rising motion across the tropical tropopause, poleward drift in the stratosphere, and by continuity of mass, a return flow into the troposphere at middle and high latitudes. Such a circulation was first proposed in the 1940s by Brewer to explain the observed low water vapor mixing ratios in the stratosphere. The only place near the tropopause where the temperature is low enough to accompany such low values of relative humidity is in the tropics, where the tropopause is high and cold. Dobson pointed out that poleward and downward advection of this type of mean circulation was consistent with the observed high concentration of ozone in the lower polar stratosphere, far from the region of photochemical production. Although the Brewer–Dobson model does not provide a complete description of the exchange process, it is believed to be essentially correct; see Holton et al. (1995).

The Brewer–Dobson circulation cell is now known to be predominantly wave driven. The morphology of stratospheric wave forcing indicates that upward movement of air into the stratosphere occurs in the tropics and downward movement of air into the troposphere occurs preferentially in winter in middle and high latitudes. Net cooling is required to transport air from the stratosphere into the troposphere whereas net diabatic heating is required to transport air from the troposphere into the stratosphere. Extensive measurements during the STEP (Stratosphere–Troposphere Exchange Project) in the 1980s showed that specific vigorous convective events were primarily responsible for transporting tropical air into the stratosphere. Tropospheric air can be either mixed directly into the stratosphere when the cumulo-nimbus towers overshoot, mixed across the tropopause by turbulent motion, or moved upward due to radiative heating of cloud tops. The dehydration occurs because some or all of the condensed ice particles are returned to the troposphere by sedimentation while the dry air remains in the stratosphere. Soluble chemical species will be found in the ice particles rather than in the dry air surrounding them, so there may be a greater resistance to cross-tropopause transport of soluble compounds. A schematic diagram illustrating the general concept of the circulation between the troposphere and stratosphere is shown in Figure 9.

Mass flow from the stratosphere to the troposphere tends to be concentrated in dynamical events known as tropopause folds, in which the tropopause on the poleward side of the jet stream is distorted during the development of large-scale weather

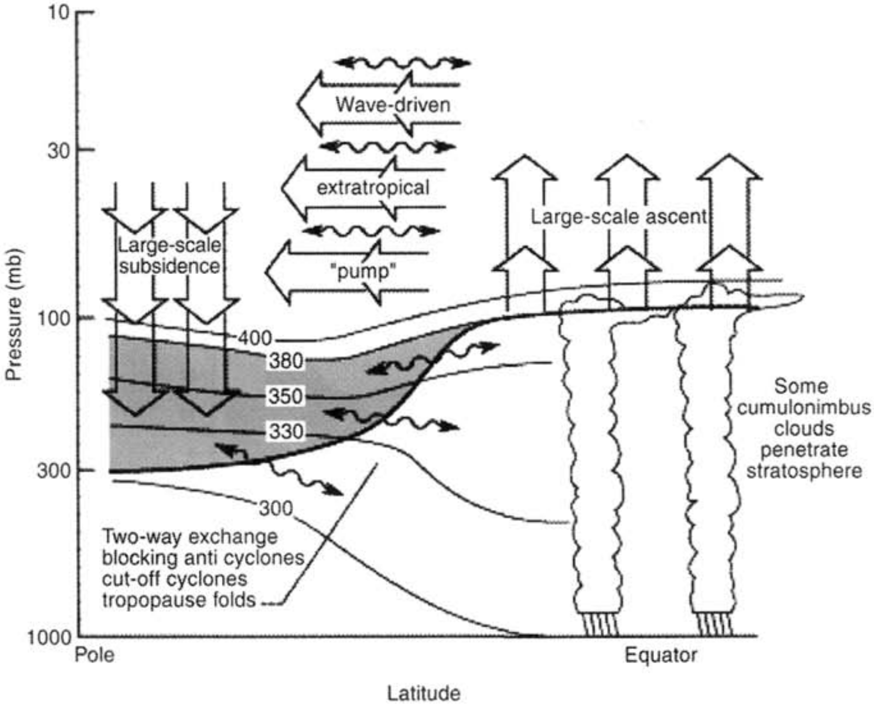


Figure 9 Schematic diagram showing the large-scale dynamical aspects of stratosphere–troposphere exchange. The wiggly double-headed arrows denote meridional transport by large-scale eddy processes. The broad arrows show transport by the global-scale circulation, which is the primary exchange mechanism that moves air across isentropic surfaces. (Reprinted with permission from Holton et al., 1995.)

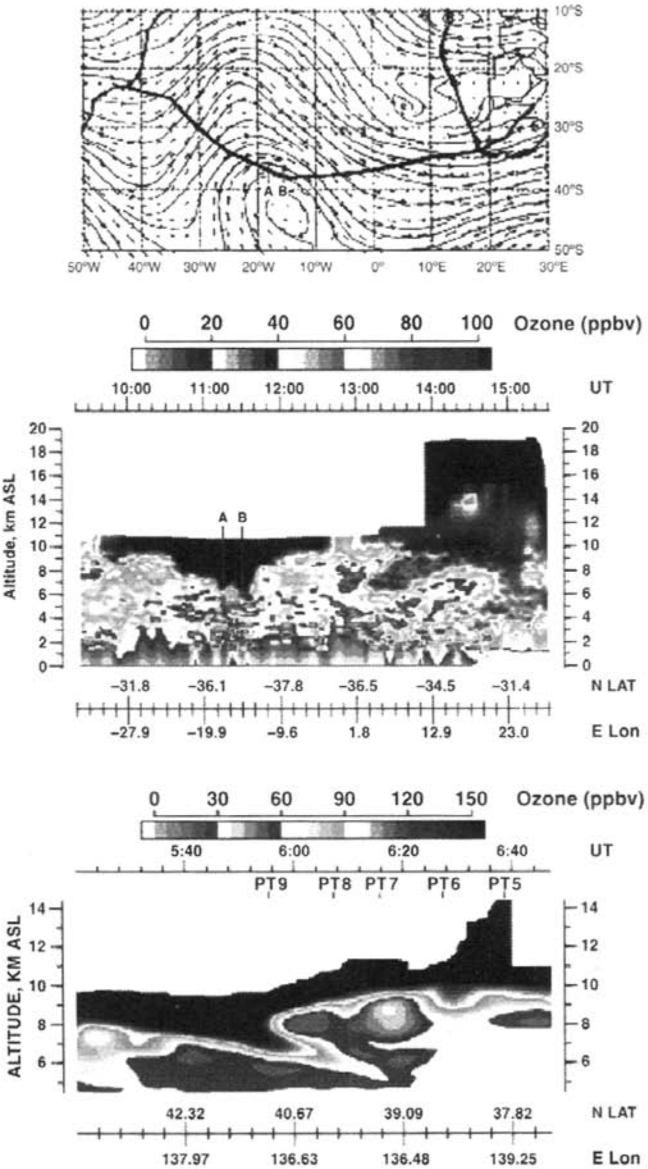


Figure 10 (see color insert) Three-panel figure showing evidence of ozone input from the stratosphere into the troposphere in both hemispheres. The top panel shows the flight path (heavy line) of a DC-8 airplane on October 3, 1992, from South America to Africa that intersected a trough protruding from higher latitudes. Points A and B on that flight path show high concentrations of ozone being transported to altitudes below 6 km in the middle panel; the data depicted in this panel were obtained from a differential absorption lidar system that measured ozone below the 11-km flight level of the DC-8. The lowest panel shows a similar feature for a flight on March 11, 1994, in the Northern Hemisphere. As the airplane flies from north to south in this panel, note the higher tropopause height south of the fold. See ftp site for color image.

systems. Large amounts of stratospheric air extend into the troposphere and much of that air becomes trapped in, and eventually mixed with, the troposphere. An example of stratospheric air coming into the troposphere during a tropopause fold is shown in Figure 10. This figure illustrates the intrusion of stratospheric tracers into the troposphere using a differential absorption laser radar (lidar) instrument that measures ozone below and above it as it flies in an airplane at a cruising altitude of 11 km. The top panel shows the flight path of the airplane (heavy line) and the geopotential height distribution at 200 hPa. This flight path between South America and Africa was part of a field mission in October 1992. Points A and B refer to the location of the two "tongues" of stratospheric air that have descended into the troposphere as the flight path intersected a trough from southern middle latitudes. The middle panel of Figure 10 shows the descent of ozone from the stratosphere (brown areas, >100 ppbv) in conjunction with the tropopause fold (see ftp site for color image). At these points, stratospheric air, as marked by the high concentrations of ozone, has descended to altitudes as low as 6 km. As the flight continues to the east, and as measurements from the upward-looking lidar system became available, the tropopause is located at ~15 km. The higher concentrations of ozone in the middle and upper troposphere (denoted by the orange colors) were formed in situ from widespread biomass burning taking place at this time of the year. The bottom panel is from a flight in March 1994 and perhaps better illustrates the distribution of ozone during a folding event. Note how much higher the tropopause is south of the fold (later in the flight), than at higher latitudes in the beginning of the flight (~8 km at the beginning of the flight), consistent with the schematic shown in Figure 9.

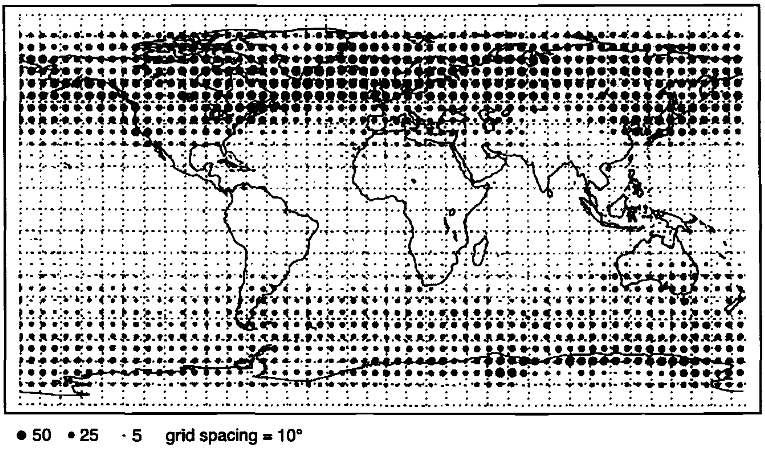


Figure 11 Annual mean distribution of global tropopause folding activity obtained from meteorological analysis over a 10-year period, 1984–1993; the size of the dots denotes the activity corresponding to bringing air from the stratosphere into the troposphere. (Reprinted with permission from Beekman et al., 1997.)

Figure 11 shows the climatological location of stratospheric intrusions weighted by the intensity of the tropopause event to derive a depiction of how much stratospheric air enters the troposphere. The data have been obtained from European Center for Medium-Range Weather Forecasting (ECMWF) data using an identification scheme relating potential vorticity to the exchange of air between the stratosphere and troposphere. This analysis, published in 1997, agrees with previous studies suggesting that considerably more exchange takes place in the Northern Hemisphere relative to the Southern Hemisphere and that the flux in the NH is 6×10^{10} molecules O_3/cm^2 s. This value is in agreement with a number of previous studies since the 1970s that have estimated a cross-tropopause flux using both general circulation models and observations calculating amounts of between 4 and 8×10^{10} molecules O_3/cm^2 s for the NH. With respect to the global tropospheric ozone budget, this "natural" flux of ozone transported would account for only a relatively small fraction of the ozone now commonly measured near Earth's surface, implying that much of the ozone present in the lower atmosphere would not be there without anthropogenic input. The chapter on tropospheric ozone will discuss the *tropospheric ozone* budget in more detail.

REFERENCES

Albritton, D. L., R. T. Watson, and P. J. Aucamp, *Scientific Assessment of Ozone Depletion: 1994*, World Meteorological Organization, Geneva, 1995.

Albritton, D. L., P. J. Aucamp, G. Megie, and R. T. Watson, *Scientific Assessment of Ozone Depletion: 1998*, World Meteorological Organization, Geneva, 1999.

Beekman, M., et al., Regional and global tropopause fold occurrence and related ozone flux across the tropopause, *J. Atmos. Chem.*, *28*, 29–44, 1997.

Brasseur, G., and S. Solomon, *Aeronomy of the Middle Atmosphere* (2nd ed.), Reidel, Dordrecht, 1986.

Farman, J. C., B. G. Gardiner, and J. D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction, *Nature*, **315**, 207–210, 1985.

Holton, J. R., A. R. Douglass, P. H. Haynes, M. E. McIntyre, R. B. Rood, and L. Pfister, Stratosphere-troposphere exchange, *Rev. Geophys.*, *33*, 403–439, 1995.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums (Eds.), Intergovernmental panel on climate change, in *Climate Change, The IPCC Scientific Assessment*, Cambridge University Press, Cambridge, 1990.

Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A Kattenberg, and K. Maskell (Eds.), Intergovernmental panel on climate change, in *Climate Change (1995): The Science of Climate Change*, Cambridge University Press, Cambridge, 1996.

Levy II, H., Normal atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, **173**, 141–143, 1971.

Molina, M. J., and F. S. Rowland, Stratospheric sink for chlorofluoromethanes: chlorine catalyzed destruction of ozone, *Nature*, **249**, 810–814, 1974.

Watson, R. T., L. G. Meiro Filho, E. Sanhueza, and A. Jenetos, Greenhouse gases: Sources and sinks, in J. T. Houghton, B. A. Callander, and S. K. Varney (Eds.), *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment*, Cambridge University Press, Cambridge, 1992, pp.1–40.

CHAPTER 2

OXIDIZING POWER OF ATMOSPHERE

DANIEL J. JACOB

1 INTRODUCTION

The atmosphere is an oxidizing medium. Many environmentally important trace gases are removed from the atmosphere by oxidation, including methane and other organic compounds, carbon monoxide, nitrogen oxides, and sulfur gases (Table 1). Understanding the processes and rates by which species are oxidized in the atmosphere, i.e., the oxidizing power of the atmosphere, is crucial to our knowledge of atmospheric composition. Changes in the oxidizing power of the atmosphere would have a wide range of implications for air pollution, aerosol formation, greenhouse radiative forcing, and stratospheric ozone depletion (Thompson, 1992).

The most abundant oxidants in Earth's atmosphere are O_2 and O_3. They have large bond energies and are hence relatively unreactive. With a few exceptions, oxidation of nonradical atmospheric species by O_2 or O_3 is negligibly slow. Photochemical modeling of stratospheric chemistry in the 1950s first implicated the strong radical oxidants O and OH, generated from photolysis of O_3 and H_2O, in the oxidation of CO and CH_4 (Bates and Witherspoon, 1952). The importance of photochemically generated radicals in the chain oxidation of hydrocarbons leading to urban O_3 smog was also recognized in the 1950s (Leighton, 1961). Smog models of that time hypothesized that O atoms produced in urban air from the photolysis of NO_2 and O_3 would provide the main pathway for hydrocarbon oxidation (Altshuller and Bufalini, 1965, 1971). This mechanism was thought unimportant outside of urban areas because of low O_3 and NO_2 concentrations, and transport to the stratosphere was viewed as necessary for oxidation of CO, CH_4, and other gases present in the global troposphere (Cadle and Allen, 1970). Long atmospheric lifetimes for these gases were implied because of the 10-year residence time of air in the troposphere.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

TABLE 1 Atmospheric Lifetimes of Selected Species

Species	Lifetime[a]	Reference
CH_3CCl_3	4.8 yr (*5.7 yr*)	WMO (1999)
CH_4	8.4 yr (*8.9 yr*)	WMO (1999)
CHF_2Cl	11.8 yr (*12.3 yr*)	WMO (1999)
CH_3Br	0.7 yr (*1.7 yr*)	WMO (1999)
Isoprene[b]	~ 1 h (~ *1 h*)	Jacob et al. (1989)
CO	2 mo (*2 mo*)	Logan et al. (1981)
NO_x ($NO + NO_2$)	~ 1 d (~ *1 d*)[c]	Dentener and Crutzen (1993)
SO_2	~ 1 d (*2 wks*)[d]	Chin et al. (1996)
$(CH_3)_2S$	~ 1 d (~ *1 d*)	Chin et al. (1996)

[a] The atmospheric lifetime of a species is defined as the average time that a molecule of the species remains in the atmosphere before it is removed by one of its sinks. It can be calculated as the atmospheric mass of the species divided by the species loss rate. The first number given for each entry in the column is the mean atmospheric lifetime, and the second number in parentheses is the mean atmospheric lifetime against oxidation by OH.

[b] $CH = C(CH_3)-CH = CH_2$, a major hydrocarbon emitted by vegetation.

[c] Loss of NO_x in summer and in the tropics is mostly by reaction of NO_2 with OH; loss in winter at extratropical latitudes is mostly by a nonphotochemical pathway involving formation of N_2O_5 and hydrolysis to HNO_3. The sum of these two processes results in a lifetime of NO_x of the order of a day.

[d] The principal SO_2 sinks are deposition and in-cloud oxidation by $H_2O_2(aq)$.

This view of a chemically inert troposphere was first challenged by Weinstock (1969) who found from ^{14}CO measurements that the atmospheric lifetime of CO is only ~0.1 years, requiring a dominant sink in the troposphere. Levy (1971) then presented photochemical model calculations for the unpolluted troposphere showing that high concentrations of OH could be generated from photolysis of O_3 in the presence of water vapor and account for the missing sink of CO in the Weinstock (1969) analysis. Further work in the early 1970s confirmed the importance of tropospheric oxidation by OH as the main sink of CO and CH_4 (McConnell et al., 1971; Weinstock and Niki, 1972; Levy et al., 1973) and further showed that OH, not O, is the main oxidant of hydrocarbons in urban air (Heicklen, 1971; Kerr et al., 1972; Demerjian et al., 1974). Considerable evidence over the past three decades supports the view that tropospheric OH is the main oxidant for nonradical species in the atmosphere.

Indirect estimates of global mean OH concentrations have been made since the 1970s using a number of proxies, the most useful of which has been CH_3CCl_3, a long-lived gas emitted by industry and removed from the atmosphere by oxidation by OH (Lovelock, 1977; Singh, 1977). The most recent analyses of CH_3CCl_3 data, based on observations at a worldwide network of sites (Prinn et al., 1995), imply a global mean OH concentration in the troposphere of $(1.1 \pm 0.1) \times 10^6$ molecules/cm^3 (Krol et al., 1998; Spivakovsky et al., 2000). Techniques for direct measurement of tropospheric OH were first developed in the 1970s but suffered from interferences or poor sensitivity. Only in the 1990s have reliable techniques been developed and successfully intercompared (special issue of *Journal of the Atmospheric Sciences*, October 1995; Crosley, 1997). Direct measurements provide the means to test our understanding of the local processes controlling OH concentrations (e.g., McKeen et al., 1997; Jaeglé et al., 1997, 2000; Frost et al., 1999). By simulating these processes in global models, one can assess the sensitivity of the oxidizing power of the atmosphere to different anthropogenic perturbations (Wang and Jacob, 1998).

This chapter reviews current understanding of the factors controlling abundances and long-term trends of OH. It also briefly reviews (Section 3) other atmospheric oxidants that are important in certain environments or for certain nonradical moleculcs. It docs not cover the oxidation of short-lived radical species, which often involves reaction with O_2 or O_3 (Atkinson, 1990). It does not cover either oxidation in the stratosphere, whose importance as a sink for species emitted at the surface is limited by the long time for transfer of air from the troposphere to the stratosphere.

2 HYDROXYL RADICAL OH

Processes Controlling OH Concentrations

A detailed and still fairly current discussion of OH chemistry in the troposphere is given by Logan et al. (1981). The primary source of OH is the photolysis of O_3 to produce an excited state of atomic oxygen, $O(^1D)$, which then reacts with water vapor:

$$O_3 + h\nu \rightarrow O_2 + O(^1D) \qquad (\lambda < 340\,\text{nm}) \tag{R1}$$

$$O(^1D) + M \rightarrow O(^3P) + M \tag{R2}$$

$$O(^3P) + O_2 + M \rightarrow O_3 + M \tag{R3}$$

$$H_2O + O(^1D) \rightarrow 2OH \tag{R4}$$

Here M is an inert molecule (N_2 or O_2). Only $\sim 1\%$ of the $O(^1D)$ atoms produced by (R1) react with H_2O; most are deactivated to the ground-state $O(^3P)$ and recombine with O_2 to return O_3. Photolysis of O_3 to $O(^1D)$ in the troposphere is determined by a narrow band of radiation in the 290- to 330-nm range, reflecting the combined wavelength dependences of the actinic flux, O_3 absorption cross section, and $O(^1D)$ quantum yield (Fig. 1). Radiation in this wavelength range is strongly absorbed by overhead O_3, and hence the production of $O(^1D)$ is strongly dependent on the thickness of the stratospheric O_3 layer (Madronich and Granier, 1992).

The OH radical is consumed on a time scale of ~ 1 s by oxidation of a large number of reduced atmospheric species. Its main sinks in the troposphere are CO and CH_4. Nonmethane hydrocarbons (NMHCs) are also important sinks in the lower troposphere over continents. Oxidation of CO or hydrocarbons by OH propagates a

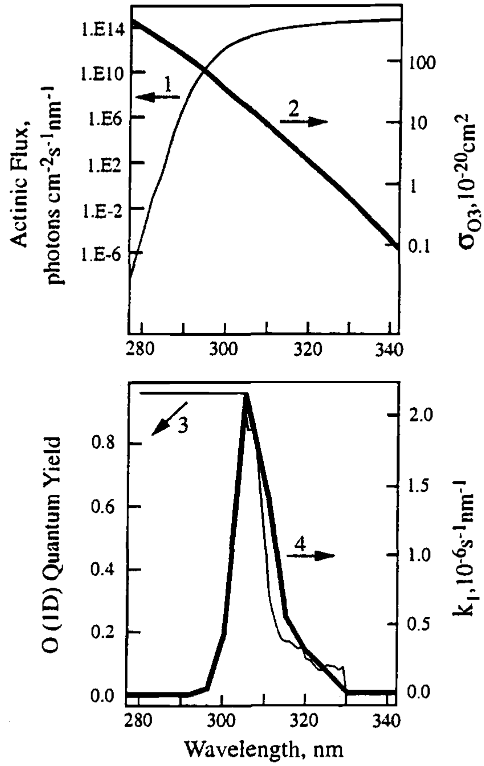


Figure 1 Computation of the rate constant k_1 of reaction (R1) as the integral over all wavelengths of the actinic flux of solar radiation (1) times the absorption cross-section σ_{O3} of ozone (2) and times the O(ID) quantum yield (3). From Jacob (1999).

radical reaction chain initiated by the generation of OH radicals from (R4). The simplest case is oxidation of CO:

$$CO + OH \rightarrow CO_2 + H \tag{R5}$$

$$H + O_2 + M \rightarrow HO_2 + M \tag{R6}$$

The HO_2 radicals may self-react to produce H_2O_2 (hydrogen peroxide):

$$HO_2 + HO_2 \rightarrow H_2O_2 + O_2 \tag{R7}$$

or they may regenerate OH by reaction with NO or O_3:

$$HO_2 + NO \rightarrow OH + NO_2 \tag{R8}$$

$$HO_2 + O_3 \rightarrow OH + 2O_2 \tag{R9}$$

Hydrogen peroxide produced by (R7) is removed from the atmosphere by deposition. It may also photolyze, regenerating OH,

$$H_2O_2 + h\nu \rightarrow 2OH \qquad (R10)$$

or react itself with OH:

$$H_2O_2 + OH \rightarrow H_2O + HO_2 \qquad (R11)$$

The same type of chain mechanism applies to the oxidation of hydrocarbons, but the complexity increases rapidly as the size of the hydrocarbon molecule increases. The mechanism for CH_4 is described here. It begins by

$$CH_4 + OH \rightarrow CH_3 + H_2O \qquad (R12)$$

$$CH_3 + O_2 + M \rightarrow CH_3O_2 + M \qquad (R13)$$

The CH_3O_2 molecule (methylperoxy radical) is analogous to HO_2. Its dominant sinks in the atmosphere are reactions with HO_2 and NO:

$$CH_3O_2 + HO_2 \rightarrow CH_3OOH + O_2 \qquad (R14)$$

$$CH_3O_2 + NO \rightarrow CH_3O + NO_2 \qquad (R15)$$

Similarly to H_2O_2, methylhydroperoxide (CH_3OOH) may either react with OH or photolyze:

$$CH_3OOH + OH \rightarrow CH_2O + OH + H_2O \qquad (R16)$$

$$CH_3OOH + OH \rightarrow CH_3O_2 + H_2O \qquad (R17)$$

$$CH_3OOH + h\nu \rightarrow CH_3O + OH \qquad (R18)$$

The methoxy radical CH_3O produced by (R15) and (R18) reacts rapidly with O_2:

$$CH_3O + O_2 \rightarrow CH_2O + HO_2 \qquad (R19)$$

Formaldehyde produced by (R16) and (R19) may either react with OH or photolyze (two photolysis branches):

$$CH_2O + OH \longrightarrow CHO + H_2O \qquad (R20)$$

$$CH_2O + h\nu \xrightarrow{O_2} CHO + HO_2 \qquad (R21)$$

$$CH_2O + h\nu \longrightarrow CO + H_2 \qquad (R22)$$

Reactions (R20) and (R21) produce the CHO radical, which reacts rapidly with O_2 to yield CO:

$$CHO + O_2 \rightarrow CO + HO_2 \qquad \text{(R23)}$$

In this overall sequence the C(−IV) atom in CH_4 is gradually oxidized to C(−II) in CH_3OOH, C(0) in CH_2O, C(+II) in CO, and C(+IV) in CO_2 (highest oxidation state for carbon).

The regeneration of OH radicals by (R8) plays a critical role in maintaining OH concentrations in the troposphere. The main sink for NO_2 produced by (R8) and (R15) is photolysis, regenerating NO and producing O_3:

$$NO_2 + h\nu \rightarrow NO + O(^3P) \qquad \text{(R24)}$$

$$O(^3P) + O_2 + M \rightarrow O_3 + M \qquad \text{(R3)}$$

This O_3 may then photolyze to yield additional OH by (R1) + (R4). Although reaction (R9) also recycles OH, it consumes in the process an O_3 molecule that could have otherwise photolyzed to produce OH. Therefore, it is not effective for maintaining OH concentrations.

Figure 2 illustrates how tropospheric OH is controlled by chemical cycling of the hydrogen oxide family ($HO_x \equiv$ OH+ peroxy radicals) and the nitrogen oxide family ($NO_x \equiv NO + NO_2$), for the simple case of CO oxidation. The schematic for hydrocarbon oxidation is similar, except that photolysis of carbonyl compounds as in reaction (R21) provides an additional (generally minor) source of HO_x. The dominant sink for the HO_x family is usually the formation of peroxides. As discussed

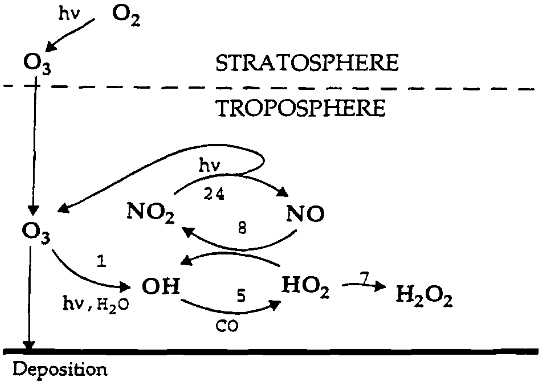


Figure 2 Simplified schematic of O_3–HO_x–NO_x–CO chemistry in the troposphere.

previously, these peroxides may photolyze to recycle HO_x; alternatively, they may deposit or react with OH, providing a terminal sink for HO_x. Sources of NO_x in the troposphere include combustion, microbial activity in soils, and lightning. Sources of CO and hydrocarbons include combustion, industrial processes, soils, and vegetation.

An analytical expression for the dependence of OH concentrations on chemical variables can be obtained from the simplified O_3–HO_x–NO_x–CO system by assuming chemical steady state for the short-lived species $O(^1D)$, H, OH, and also for the chemical family HO_x. The lifetime of HO_x against formation of peroxides is of the order of minutes, so that the steady-state assumption is appropriate. The production rate P_{HO_x} of HO_x from reaction (R4) is given by

$$P_{HO_x} = 2k_4[O(^1D)][H_2O] \equiv 2\frac{k_1k_4}{k_2[M]}[O_3][H_2O] \qquad (1)$$

where k_i is the rate constant for reaction i. In writing Eq. (1) we have used the approximation (R2)≫(R4) to simplify the denominator. Steady state for OH is defined by

$$P_{HO_x} + k_8[HO_2][NO] = k_5[CO][OH] \qquad (2)$$

Loss of HO_x in this system is by (R7). Steady state for HO_x is therefore defined by

$$P_{HO_x} = 2k_7[HO_2]^2 \qquad (3)$$

from which we derive the following expression for the OH concentration:

$$[OH] = \frac{P_{HO_x} + k_8\sqrt{\frac{P_{HO_x}}{2k_7}}[NO]}{k_5[CO]} \qquad (4)$$

We see from Eq. (4) together with Eq. (1) that OH concentrations depend negatively on CO and positively on water vapor, O_3, and NO. The dependence on hydrocarbons is more complicated (as hydrocarbons provide both sinks of OH and sources of HO_x) but is generally negative, similar to CO.

One important caveat to this simplified representation of OH chemistry must be made for high-NO_x environments. When NO_x concentrations exceed a few parts per billion by volume (ppbv), as in urban air, oxidation of NO_2 by OH can become the dominant sink for HO_x:

$$NO_2 + OH + M \rightarrow HNO_3 + M \qquad (R25)$$

Under these conditions, OH concentrations decrease with increasing NO_x (as may be derived by repeating the steady-state calculation above) and increase with increasing hydrocarbons. This situation is commonly denoted the NO_x-saturated (or hydrocarbon-limited) regime, as opposed to the NO_x-limited regime normally encountered in the troposphere.

A second caveat applies to the upper troposphere where water vapor concentrations are low (~100 ppmv). Under these conditions, reaction (R4) may be less important as a primary source of HO_x than photolysis of acetone originating from the biosphere (Singh et al., 1995) or convective injection of peroxides and aldehydes produced in the lower troposphere (Jaeglé et al., 1997; Prather and Jacob, 1997; Müller and Brasseur, 1999). Reaction of OH with HO_2 provides in general the dominant HO_x sink in the upper troposphere, which yields a square root rather than linear dependence of OH concentrations on NO.

Figure 3 shows zonal mean global distributions of OH concentrations computed with a global three-dimensional model of tropospheric O_3–NO_x–hydrocarbon chemistry (Wang et al., 1998b). The highest concentrations (averaging over 2×10^6 molecules/cm^3) are in the tropical middle troposphere, reflecting a combination of high ultraviolet (UV) and high humidity. The large seasonal variation at midlatitudes follows UV radiation. Concentrations tend to be higher in the Northern than in the Southern Hemisphere, reflecting higher NO_x concentrations.

Global Mean OH Concentration

The short lifetime of OH implies that its concentration is highly variable. Deriving the atmospheric lifetimes of gases removed by oxidation by OH requires an estimate of OH concentrations averaged appropriately over time and space. Mass-balance arguments for proxy species with known sources can assist for this purpose. The most successful application, first proposed by Singh (1977) and Lovelock (1977), has been the use of the industrial solvent CH_3CCl_3 to estimate the global mean OH concentration. The source of CH_3CCl_3 is exclusely anthropogenic, and its historical trend is well known from industrial data. Production of CH_3CCl_3 has been banned since 1996 as part of the Montreal Protocol. The dominant sink of CH_3CCl_3 is oxidation by OH in the troposphere (photolysis in the stratosphere and uptake by the oceans are small additional sinks). Tropospheric mixing ratios of CH_3CCl_3 are relatively uniform, so that a mass-balance analysis for CH_3CCl_3 yields a global mean OH concentration weighted by atmospheric mass and by the temperature dependence of the $CH_3CCl_3 + OH$ reaction. The global mean OH concentration obtained in this manner can then be used to infer the lifetimes of other long-lived gases removed by reaction with OH, such as CH_4 and hydrogenated halocarbons (HCFCs) (Prather and Spivakovsky, 1990).

The most recent use of CH_3CCl_3 observations to constrain the global mean OH concentration has been by Krol et al. (1998) and Spivakovsky et al. (2000). These authors derive a CH_3CCl_3 lifetime of 5.5 years in the troposphere against oxidation by OH, corresponding to a global mean OH concentration of $(1.1 \pm 0.2) \times 10^6$ molecules/cm^3. Spivakovsky et al. (2000) point out that the magnitude of the CH_3CCl_3 interhemispheric gradient implies that the difference between the mean OH concentrations in the Northern and Southern Hemispheres is no more than 50%.

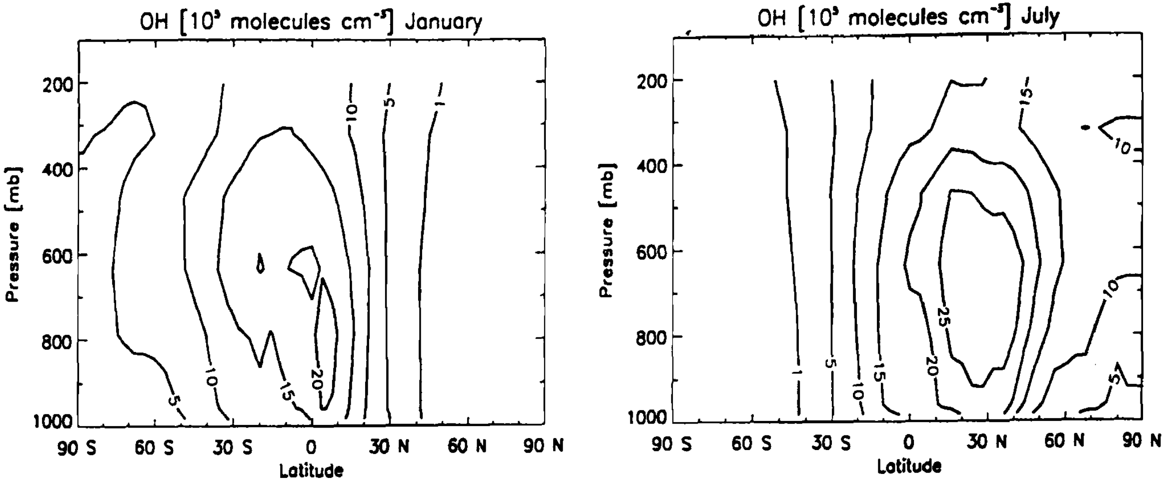


Figure 3 Longitudinally averaged monthly mean OH concentrations.

Mass-balance arguments for other chemical tracers oxidized by OH including ^{14}CO, CHF_2Cl, CH_2Cl_2, and hydrocarbons have been used to confirm the above estimate of the global mean OH concentration and to provide additional constraints on the geographical and seasonal distribution of OH (Volz et al., 1981; Mak et al., 1992; Goldstein et al., 1995; Spivakovsky et al., 2000).

Simulation of the CH_3CCl_3 lifetime has long been a standard test for evaluating the global mean OH concentration computed in tropospheric chemistry models, starting from the work of Crutzen and Fishman (1977). In these models, the OH concentrations are computed from a global simulation of O_3–NO_x–CO–hydrocarbon chemistry that treats emissions, transport, chemistry, and deposition in a self-consistent way (e.g., Wang et al., 1998a). The current generation of models reproduces the atmospheric lifetime of CH_3CCl_3 to within typically 25%.

Measurements of OH Concentrations and Comparisons to Models

The past few years have seen the development of a number of methods for direct measurement of tropospheric OH (special issue of *Journal of Atmospheric Science*, October 1995). Two of these methods, a long-path absorption (LPA) instrument (Mount, 1992) and a chemical ionization mass spectrometry (CIMS) instrument (Eisele and Tanner, 1991) were intercompared formally at a mountain site in Colorado during the Tropospheric OH Photochemistry Experiment (TOHPE). Under well-mixed atmospheric conditions where the local OH measurement from CIMS could be compared to the long-path average from LPA, the intercomparison demonstrated a good correlation between the two instruments down to concentrations of less than 1×10^6 molecules/cm^3, with no significant bias (Crosley, 1997).

A number of ancillary chemical measurements were made during TOHPE that McKeen et al. (1997) used to compare the observed OH concentrations to values computed from a standard photochemical model. The model overestimated OH concentrations by a factor of 1.3 on average. It captured 48% of the variance in the CIMS instrument, although much of that variance was driven by the diurnal cycle. It was not correlated with the LPA instrument, which may reflect the nonlocal nature of the latter measurement.

The model overestimate of OH in TOHPE is consistent with other model measurement comparisons conducted at continental sites (Poppe et al., 1995; Thompson, 1995; George et al., 1999). As discussed by McKeen et al. (1997), possible causes include inadequate model representation of hydrocarbon chemistry or of uptake of HO_x by aerosols. Eisele et al. (1996) conducted a model-measurement comparison using the CIMS instrument at Mauna Loa Observatory, Hawaii (3.4 km altitude); they found good agreement when subsiding motions brought free tropospheric air to the site but a factor of 2 model overestimate under upslope flow, supporting the view that biogenic hydrocarbons may provide important sinks for OH. Frost et al. (1999) found a median model overestimate of 32% in simulation of aircraft observations for clean marine air.

An important aspect of these model-measurement comparisons has been to examine the ability of models to reproduce the dependence of OH concentrations on chemical and meteorological variables. Poppe et al. (1995) found that their model could capture successfully the observed correlations of OH concentrations with UV intensity, temperature, humidity, and CO concentration. Measurements in TOHPE showed OH concentrations increasing with increasing NO_x up to about 2 ppbv NO_x and then decreasing, consistent with model calculations of NO_x versus hydrocarbon-limited chemistry (Eisele et al., 1997; McKeen et al., 1997).

Aircraft measurements of OH and HO_2 concentrations in the upper troposphere have been reported by Brune et al. (1998, 1999) and Wennberg et al. (1998). The measured OH/HO_2 ratios and their variances agree with model values to within the uncertainties of the relevant rate constants, implying a good understanding of the cycling of HO_x (Jaeglé et al., 2000). The observed HO_x concentrations are often several times lower than would be predicted solely from the $O(^1D) + H_2O$ source (R4) and support the presence of other primary HO_x sources in the upper troposphere including acetone, peroxides, and aldehydes.

Long-Term Trends in Atmospheric OH

Assessing human influence on the oxidizing power of the atmosphere is intricate. On the one hand, anthropogenic emissions of CO and hydrocarbon emissions act to deplete OH; on the other hand, anthropogenic emissions of NO_x and the thinning of the stratospheric O_3 layer act to boost OH. Human-induced changes in Earth's climate (temperature, cloudiness, circulation) add to the complication. Large regional differences may be expected in the response of OH to human activity, depending on the relative importance and coupling of the above factors.

A number of global tropospheric chemistry model studies, reviewed by Thompson (1992), have examined the changes in OH concentrations since preindustrial times as driven by trends in emissions of CO, hydrocarbons, and NO_x. These studies report 10 to 30% decrease in the global mean OH concentration from preindustrial times to today, a relatively small effect considering that emissions of CO, CH_4, and NO_x increased severalfold over that period (Table 2). The global three-dimensional model study of Wang and Jacob (1998) indicate a 9% decrease in the global mean

TABLE 2 Comparison of Present and Preindustrial Atmospheres[a]

	Emission					
	CH_4 (Tg CH_4/yr)	Nonmethane Hydrocarbons (Tg C/yr)	CO (Tg CO/yr)	NO_x (Tg N/yr)	O_3 Source[b] (Tg O_3/yr)	[OH][c] (molecules/cm^3)
Preindustrial	160	610[d]	50	9	2300	1.15×10^6
Present	460	710	1040	42	4500	1.04×10^6

[a] Global data from the three-dimensional model study of Wang and Jacob (1998).

[b] Tropospheric O_3 source including transport from the stratosphere (400 Tg O_3/yr in both preindustrial and present cases) and chemical production within the troposphere.

[c] Global mean tropospheric concentration weighted by atmospheric mass.

[d] Biogenic isoprene and acetone.

OH concentration since preindustrial times and suggests that the OH trend should follow roughly the trend of the $S_{NO}/S_C^{3/2}$ ratio, where S_{NO} is the global source of NO and S_C is the global source of CO and hydrocarbons; the parallel changes in S_{NO} and S_C over the past century would thus have had nearly cancelling effects on OH concentrations. This study points out that estimates of past and future trends in OH are highly sensitive to assumed trends in tropical biomass burning because NO_x emitted in the tropics is particularly efficient for generating O_3 and OH.

Observational constraints on long-term OH trends are largely limited to the CH_3CCl_3 record since 1978. An analysis of this record by Krol et al. (1998) indicates a 0.5%/yr increase in global mean OH concentrations over the period 1978 to 1993. This result is consistent with radiative transfer model calculations by Madronich and Granier (1992), which indicate a 0.4%/yr increase in OH concentrations over the 1979 to 1989 decade as a result of stratospheric O_3 depletion.

Estimates of OH trends since preindustrial and glacial times have been made using polar ice core records of CH_2O and H_2O_2. Interpretation of these records is complicated by postdepositional exchange with the atmosphere and reactions within the ice (Neftel et al., 1995). Also, since CH_2O and H_2O_2 have atmospheric lifetimes of about a day, they can only diagnose trends in polar OH, which may be different from global tropospheric trends. Analysis of the CH_2O/CH_4 ratio in a Greenland ice core (Staffelbach et al., 1991) suggests that OH concentrations were 30% higher in the preindustrial atmosphere than today, and 2 to 4 times lower in the last glacial maximum (LGM) than today. Such depletion of OH in the LGM is not consistent with results from tropospheric chemistry models, which indicate higher OH concentrations in glacial than interglacial periods due to lower emissions of CH_4 (Thompson, 1992). Staffelbach et al. (1991) suggested that a thicker stratospheric O_3 layer could be responsible for low OH levels during glacial periods.

Data for H_2O_2 in Greenland ice going back to A.D. 1300 show constant concentrations until about 1970, and a doubling of concentrations since then (Sigg and Neftel, 1991; Anklin and Bales, 1997). Although the rise in H_2O_2 would imply a rise in HO_x, the CH_3CCl_3 record shows no large trends in global mean OH concentrations during that same period.

3 OTHER ATMOSPHERIC OXIDANTS

Other atmospheric oxidants besides OH may also be important in some environments and for some species. They are reviewed briefly below.

Nitrate Radical

The nitrate radical (NO_3) is a strong radical oxidant formed in the oxidation of NO_2 by O_3:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \qquad \text{(R26)}$$

A detailed review of its atmospheric chemistry is given by Wayne (1991). During the daytime, NO_3 photolyzes on a time scale of 1 min to return NO_2:

$$NO_3 + h\nu \rightarrow NO_2 + O \qquad (R27)$$

At night the lifetime of NO_3 is much longer. In high-NO_x regions such as the eastern United States, NO_3 accumulates to concentrations of 10 to 100 parts per trillion by volume (pptv) during the nighttime hours (Wayne, 1991). At these concentrations, NO_3 can provide an important sink for some unsaturated hydrocarbons including isoprene and terpenes (O_3 is also an important oxidant for these compounds). Measurements in relatively polluted marine air over the North Sea indicate a mean nighttime NO_3 concentration of about 10 pptv; at this concentration, NO_3 represents a major sink for biogenic dimethylsulfide (Carslaw et al., 1997). Nighttime accumulation of NO_3 is in general limited by equilibrium with N_2O_5, followed by hydrolysis of N_2O_5 in aerosols:

$$NO_3 + NO_2 + M \longrightarrow N_2O_5 + M \qquad (R28)$$

$$N_2O_5 \xrightarrow{\text{heat}} NO_3 + NO_2 \qquad (R29)$$

$$N_2O_5 + H_2O \xrightarrow{\text{aerosol}} 2HNO_3 \qquad (R30)$$

At low temperatures ($T < 280$ K) the NO_3/N_2O_5 equilibrium is shifted far to the right; thus NO_3 is important only in the warm lower troposphere.

Halogen Radical Oxidants

There has been longstanding interest in the possible role of halogen radicals as tropospheric oxidants (Singh and Kasting, 1988; Chatfield and Crutzen, 1990). The best evidence so far comes from measurements of alkanes and acetylene in Arctic surface air (Jobson et al., 1994), which indicate a sink in April (polar sunrise) consistent with oxidation by Cl atoms present at a concentration of $\sim 1 \times 10^4$ atoms/cm^3. The data also suggest the presence of Br atoms to oxidize acetylene. The source of the halogen oxidants is not well established but likely involves chemical production from sea salt accumulated on the ice over the polar night (Impey et al., 1999).

Generation of halogen oxidants from sea salt would be of little interest for global tropospheric chemistry if it were confined to Arctic sunrise. However, measurements of hydrocarbons and nonradical Cl species in the marine boundary layer (MBL) at midlatitudes and in the tropics suggest that Cl atoms may be present at least occasionally at concentrations in the range 10^4 to 10^5 atoms/cm^3 (Keene et al., 1990, 1996; Pszenny et al., 1993; Singh et al., 1996; Spicer et al., 1998). At such concentrations, oxidation by Cl atoms would provide a major sink for dimethylsulfide and alkanes in the MBL. Even less is known about Br radical chemistry in the MBL, although Toumi (1994) has suggested that BrO could provide an important oxidant for dimethylsulfide. Field measurements of the halogen radicals and their reservoirs HOCl and HOBr are needed.

Cloud and Aerosol Oxidants

Water-soluble atmospheric species incorporated in cloud droplets and aqueous aerosols may dissociate into ions, and the resulting aqueous-phase redox chemistry provides yet another pathway for oxidation of species in the atmosphere. The importance of this pathway has been established for SO_2, which dissociates in water to HSO_3^- and SO_3^{2-} ($pK_{a1} = 1.9$, $pK_{a2} = 7.2$). Rapid oxidation of SO_2 by H_2O_2 in cloud was first suggested by Penkett et al. (1979):

$$SO_2(g) \Leftrightarrow SO_2 \cdot H_2O \tag{R31}$$

$$SO_2 \cdot H_2O \Leftrightarrow HSO_3^- + H^+ \tag{R32}$$

$$H_2O_2(g) \Leftrightarrow H_2O_2(aq) \tag{R33}$$

$$HSO_3^- + H_2O_2(aq) + H^+ \rightarrow SO_4^{2-} + 2H^+ + H_2O \tag{R34}$$

Aircraft measurements by Daum et al. (1984) demonstrated that the reaction is sufficiently fast to titrate either SO_2 or H_2O_2 in cloud (whichever is limiting). It is now well accepted that this mechanism dominates over gas-phase oxidation by OH as a sink for SO_2 in the atmosphere (Chin et al., 1996).

Additional nonradical oxidants may also be important for oxidation of SO_2 in clouds and aqueous aerosols, but their importance is not as well verified as for H_2O_2. At high pH values (pH > 5), $O_3(aq)$ reacts rapidly with SO_3^{2-}:

$$HSO_3^- \Leftrightarrow SO_3^{2-} + H^+ \tag{R35}$$

$$SO_3^{2-} + O_3(aq) \rightarrow SO_4^{2-} + O_2 \tag{R36}$$

This mechanism, taking place in alkaline sea salt aerosols, could represent a major sink for SO_2 in the marine boundary layer (Chameides and Stelson, 1992). Additional SO_2 oxidants in sea salt aerosol may include HOCl and HOBr produced by halogen radical chemistry (Vogt et al., 1996). In polluted clouds, aqueous-phase autoxidation catalyzed by Fe(III) could provide the dominant SO_2 sink (Jacob and Hoffmann, 1983).

ACKNOWLEDGMENT

I thank Clarisa M. Spivakovsky for valuable comments.

REFERENCES

Altshuller, A. P., and J. Bufalini, Photochemical aspects of air pollution: A review, *Photochem. Photobiol.*, *4*, 97–146, 1965.

Altshuller, A. P., and J. Bufalini, Photochemical aspects of air pollution: A review, *Environ. Sci. Technol.*, *5*, 39–62, 1971.

Anklin, M., and R. C. Bales, Recent increases in H_2O_2 concentrations at Summit, Greenland, *J. Geophys. Res.*, *102*, 19099–19104, 1997.

Atkinson, R. A., Gas-phase tropospheric chemistry of organic compounds: A review, *Atmos. Environ.*, *24*, 1–42, 1990.

Bates, D. R., and A. Witherspoon, The photochemistry of some minor constituents of the earth's atmosphere (CO_2, CO, CH_4, N_2O), *Mon. Not. Roy. Astron. Soc.*, *112*, 101, 1952.

Brune, W. H., et al., Airborne in-situ OH and HO_2 observations in the cloud-free troposphere and lower stratosphere during SUCCESS, *Geophys. Res. Lett.*, *25*, 1701–1704, 1998.

Brune, W. H., et al., OH and HO_2 chemistry in the North Atlantic free troposphere, *Geophys. Res. Lett.*, *26*, 3077–3080, 1999.

Cadle, R. D., and E. R. Allen, Atmospheric photochemistry, *Science*, *167*, 243–249, 1970.

Carslaw, N., L. J. Carpenter, J. M. C. Plane, B. J. Allan, R. A. Burgess, K. C. Clemitshaw, H. Coe, and S. A. Penkett, Simultaneous observations of nitrate and peroxy radicals in the marine boundary layer, *J. Geophys. Res.*, *102*, 18917–18933, 1997.

Chameides, W. L., and A. W. Stelson, Aqueous-phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Chatfield, R. C., and P. J. Crutzen, Are there interactions of iodine and sulfur species in marine air photochemistry? *J. Geophys. Res.*, *95*, 22319–22342, 1990.

Chin, M., D. J. Jacob, G. M. Gardner, M. S. Foreman-Fowler, and P. A. Spiro, A global three-dimensional model of tropospheric sulfate, *J. Geophys. Res.*, *101*, 18667–18690, 1996.

Crosley, D. R., The measurement of OH and HO_2 in the troposphere, *J. Atmos. Sci.*, *52*, 3299–3314, 1995.

Crosley, D. R., 1993 Trospospheric OH Experiment: A summary and perspective, *J. Geophys. Res.*, *102*, 6495–6510, 1997.

Crutzen, P. J., and J. Fishman, Average concentrations of OH in the troposphere, and the budgets of CH_4, CO, H_2 and CH_3CCl_3, *Geophys. Res. Lett.*, *4*, 321–324, 1977.

Daum, P. H., S. E. Schwartz, and L. Newman, Acidic and related constituents in liquid-water clouds, *J. Geophys. Res.*, *89*, 1447–1458, 1984.

Demerjian, K. L., J. A. Kerr, and J. G. Calvert, The mechanism of photochemical smog formation, *Adv. Environ. Sci. Technol.*, *4*, 1–262, 1974.

DeMore, W. B., S. P. Sander, D. M. Golden, R. F. Hampson, M. J. Kurylo, C. J. Howard, A. R. Ravishankara, C. E. Kolb, and M. J. Molina, Chemical kinetics and photochemical data for use in stratospheric modeling, *JPL Publication 97–4*, Pasadena, CA, 1997.

Dentener, F. J., and P. J. Crutzen, Reaction of N_2O_5 on tropospheric aerosols: Impact on the global distributions of NO_x, O_3, and OH, *J. Geophys. Res.*, *98*, 7149–7163, 1993.

Eisele, F. L., G. H. Mount, D. Tanner, A. Jefferson, R. Shetter, J. W. Harder, and E. J. Williams, Understanding the production and interconversion of the hydroxyl radical during the Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6457–6465, 1997.

Eisele, F. L., and D. J. Tanner, Ion-assisted tropospheric OH measurements, *J. Geophys. Res.*, *96*, 9295–9308, 1991.

Eisele, F. L., D. J. Tanner, C. A. Cantrell, and J. G. Calvert, Measurements and steady state calculations of OH concentrations at Mauna Loa Observatory, *J. Geophys. Res.*, *101*, 14665–14679, 1996.

Frost, G. J., et al., Photochemical modeling of OH levels during the first aerosol characterization experiment (ACE 1), *J. Geophys. Res.*, *104*, 16041–16052, 1999.

George, L. A., T. M. Hard, and R. J. O'Brien, Measurement of free radicals OH and HO_2 in Los Angeles smog, *J. Geophys. Res.*, *104*, 11643–11655, 1999.

Goldstein, A. H., S. C. Wofsy, and C. M. Spivakovsky, Seasonal variations of nonmethane hydrocarbons in rural New England: Constraints on OH concentrations in northern midlatitudes, *J. Geophys. Res.*, *100*, 21023–21033, 1995.

Heicklen, J., Discussion of "Hydrocarbon reactivities and nitric oxide conversion" by E. R. Stephens, in C. S. Tuesday (Ed.), *Chemical Reactions in Urban Atmospheres*, Elsevier, 1971, pp. 55–59.

Impey, G. A., C. M. Mihele, P. B. Shepson, D. R. Hastie, K. G. Anlauf, and L. A. Barrie, Measurements of photolyzable halogen compounds and bromine radicals during the Polar Sunrise Experiment 1997, *J. Atmos. Chem.*, *34*, 21–37, 1999.

Jacob, D. J., *Introduction to Atmospheric Chemistry*, Princeton University Press, Princeton, NJ, 1999.

Jacob, D. J., and M. R. Hoffmann, A dynamic model for the production of H^+, NO_3^-, and SO_4^{2-} in urban fog, *J. Geophys. Res.*, *88*, 6611–6621, 1983.

Jacob, D. J., S. Sillman, J. A. Logan, and S. C. Wofsy, Least-independent-variables method for simulation of tropospheric ozone, *J. Geophys. Res.*, *94*, 8497–8509, 1989.

Jaeglé, L., et al., Observed OH and HO_2 in the upper troposphere suggest a major source from convective injection of peroxides, *Geophys. Res. Lett.*, *24*, 3181–3184, 1997.

Jaeglé, L., et al., Photochemistry of HO_x in the upper troposphere at northern midlatitudes, *J. Geophys. Res.*, *105*, 3877–3892, 2000.

Jobson, B. T., H. Niki, Y. Yokouchi, J. Bottenheim, F. Hopper, and R. Leaitch, Measurements of C_2-C_6 hydrocarbons during the Polar Sunrise 1992 Experiment: Evidence for Cl atom and Br atom chemistry, *J. Geophys. Res.*, *99*, 25355–25368, 1994.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz-Tokos, H. Sievering, and J. F. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Keene, W. L., D. J. Jacob, and S.-M. Fan, Reactive chlorine: A potential sink for dimethylsulfide and hydrocarbons in the marine boundary layer, *Atmos. Environ.*, *30*, i–iii, 1996.

Kerr, J. A., J. G. Calvert, and K. L. Demerjian, The mechanism of photochemical smog formation, *Chem. Brit.*, *8*, 252–257, 1972.

Krol, M., P. J. van Leeuwen, and J. Lelieveld, Global OH trend inferred from methylchloroform measurements, *J. Geophys. Res.*, *103*, 10697–10711, 1998.

Leighton, P. A., *Photochemistry of Air Pollution*, Academic, New York, 1961.

Levy, H., Normal atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, *173*, 141–143, 1971.

Levy, H., Tropospheric budgets for methane, carbon monoxide, and related species, *J. Geophys. Res.*, *78*, 5325–5332, 1973.

Logan, J. A., M. J. Prather, S. C. Wofsy, and M. B. McElroy, Tropospheric chemistry: A global perspective, *J. Geophys. Res.*, *86*, 7210–7254, 1981.

Lovelock, J. E., Methyl chloroform in the troposphere as an indicator of OH radical abundance, *Nature*, *267*, *32*, 1977.

Madronich, S., and C. Granier, Impact of recent total ozone changes on tropospheric ozone photodissociation, hydroxyl radicals, and methane trends, *Geophys. Res. Lett.*, *19*, 465–467, 1992.

Mak, J. E., C. A. M. Brenninkmeijer, and M. R. Manning, Evidence for a missing carbon monoxide sink based on tropospheric measurements of ^{14}CO, *Geophys. Res. Lett.*, *19*, 1467–1470, 1992.

McConnell, J. C., M. B. McElroy, and S. C. Wofsy, Natural sources of atmospheric CO, *Nature*, *233*, 187–188, 1971.

McKeen, S. A., et al., Photochemical modeling of hydroxyl and its relationship to other species during the Troposheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6467–6493, 1997.

Mount, G., The measurement of tropospheric OH by long-path absorption, 1. Instrumentation, *J. Geophys. Res.*, *97*, 2427–2444, 1992.

Müller, J.-F., and G. Brasseur, Sources of upper tropospheric HO_x: A three-dimensioinal study, *J. Geophys. Res.*, *104*, 1705–1715, 1999.

Neftel, A., R. C. Bales, and D. J. Jacob, H_2O_2 and HCHO in polar snow and their relation to atmospheric chemistry, in R. Delmas (Ed.), *Ice Core Studies of Global Biogeochemical Cycles*, Springer-Verlag, Berlin, 1995, pp. 249–264.

Penkett, S. A., B. M. Jones, K. A. Brice, and A. E. Eggleton, The importance of atmospheric ozone and hydrogen peroxide in oxidizing sulfur dioxide in cloud and rainwater, *Atmos. Environ.*, *13*, 123–137, 1979.

Poppe, D., J. Zimmermann, and H. P. Dorn, Field data and model calculations for the hydroxyl radical, *J. Atmos. Sci.*, *52*, 3402–3407, 1995.

Prather, M. J., and C. M. Spivakovsky, Tropospheric OH and the lifetimes of hydrochlorofluorocarbons, *J. Geophys. Res.*, *95*, 18723–18729, 1990.

Prather, M. J., and D. J. Jacob, A persistent imbalance in HO_x and NO_x photochemistry in the upper troposphere driven by deep tropical convection, *Geophys. Res. Lett.*, *24*, 3189–3192, 1997.

Prinn, R. G., R. F. Weiss, B. R. Miller, J. Huang, F. N. Alyea, D. M. Cunnold, P. J. Fraser, D. E. Hartley, and P. G. Simmonds, Atmospheric trends and lifetime of CH_3CCl_3 and global OH concentrations, *Science*, *269*, 187–192, 1995.

Pszenny, A. A. P., W. C. Keene, D. Jacob, S. Fan, J. R. Maben, M. P. Zetwo, M. Springer-Young, and J. N. Galloway, Evidence of inorganic chlorine gases other than hydrogen chloride in marine surface air, *Geophys. Res. Lett.*, *20*, 699–702, 1993.

Sigg, A., and Neftel, Evidence for a 50% increase in H_2O_2 over the past 200 years from a Greenland ice core, *Nature*, *351*, 557–559, 1991.

Singh, H. B., Atmospheric halocarbons: Evidence in favor of reduced average hydroxyl radical concentration in the troposphere, *Geophys. Res. Lett.*, *4*, 101–104, 1977.

Singh, H. B., M. Kanakidou, P. J. Crutzen, and D. J. Jacob, High concentrations and photochemical fate of oxygenated hydrocarbons in the global troposphere, *Nature*, *378*, 50–54, 1995.

Singh, H. B., and J. F. Kasting, Chlorine-hydrocarbon photochemistry in the marine troposphere and lower stratosphere, *J. Atmos. Chem.*, *7*, 261–286, 1988.

Singh, H. B., et al., Low ozone in the marine boundary layer of the tropical Pacific Ocean: Photochemical loss, chlorine atoms, and entrainment, *J. Geophys. Res.*, *101*, 1907–1917, 1996.

Spicer, C. W., E. G. Chapman, B. J. Finlayson-Pitts, R. A. Plastridge, J. M. Hubbe, J. D. Fast, and C. M. Berkowitz, First observations of Cl_2 and Br_2 in the marine troposphere, *Nature*, *394*, 353–356, 1998.

Spivakovsky, C. M., et al., Three-dimensional climatological distribution of tropospheric OH: Update and evaluation, *J. Geophys. Res.*, *105*, 8931–8980, 2000.

Staffelbach, T., A. Neftel, B., Stauffer, and D. J. Jacob, Formaldehyde in polar ice cores: A possibility to characterize the atmospheric sink of methane in the past? *Nature*, *349*, 603–605, 1991.

Thompson, A. M., The oxidizing capacity of the earth's atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.

Thompson, A. M., Measuring and modeling the tropospheric hydroxyl radical (OH), *J. Atmos. Sci.*, *52*, 3315–3327, 1995.

Toumi, R., BrO as a sink for dimethylsulfide in the marine atmosphere, *Geophys. Res. Lett.*, *21*, 117–120, 1994.

Vogt, R., P. J. Crutzen, and R. Sander, A mechanism for halogen release from sea-salt aerosol in the remote marine boundary layer, *Nature*, *383*, 327–330, 1996.

Volz, A., D. E. Kley, and R. G. Derwent, Seasonal and latitudinal variation of ^{14}CO and the tropospheric concentrations of OH radicals, *J. Geophys. Res.*, *86*, 5163–5171, 1981.

Wang, Y., and D. J. Jacob, Anthropogenic forcing on tropospheric ozone and OH since preindustrial times, *J. Geophys. Res.*, *103*, 31123–31135, 1998.

Wang, Y., D. J. Jacob, and J. A. Logan, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 1. Model formulation, *J. Geophys. Res.*, *103*, 10713–10726, 1998a.

Wang, Y., J. A. Logan, and D. J. Jacob, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 2. Model evaluation and global ozone budget, *J. Geophys. Res.*, *103*, 10727–10756, 1998b.

Wayne, R. P. (Ed.), The nitrate radical: Physics, chemistry, and the atmosphere, *Atmos. Environ.*, *25*, 1–203, 1991.

Weinstock, B., Carbon monoxide: Residence time in the atmosphere, *Science*, *166*, 224–225, 1969.

Weinstock, B., and H. Niki, Carbon monoxide balance in nature, *Science*, *176*, 290–292, 1972.

Wennberg, P. O., et al., HO_x, NO_x, and the production of ozone in the upper troposphere, *Science*, *279*, 49–53, 1998.

World Meteorological Association (WMO), *Scientific Assessment of ozone depletion: 1998*, WMO, Geneva, Switzerland, 1999.

CHAPTER 3

TROPOSPHERIC OZONE

JACK FISHMAN

1 INTRODUCTION

Ozone is the triatomic form of oxygen, O_3, and is generally regarded as the most important species that determines the oxidizing capacity of the troposphere. The word *ozone* comes from the Greek word *ozein*, which means "to smell." Probably the name of this gas originated from early laboratory studies when ozone was first discovered because of its distinctive acrid odor. The German scientist Christian Friedrich Schönbein is credited with ozone's discovery in 1839, while he was a professor at the University of Basel in Switzerland.

One of the goals of Schönbein's research was to show that ozone is a permanent and natural component of the atmosphere. He devised a method to measure ozone in the atmosphere that was capable of measuring very low levels simply and easily. The method used soon became known as *Schönbein paper* and involved the simple process of saturating a strip of paper with potassium iodide (KI) and then allowing it to dry. In the presence of ozone, the potassium iodide oxidized and is converted to potassium iodate (KIO_3). In the process of this conversion the paper changes color to various hues of blue. More ozone present in the atmosphere resulted in the paper becoming a deeper shade of blue. Schönbein calibrated the amount of color change into a measurement standard called *Schönbein* units, which allowed scientists to put out a new piece of Schönbein paper each day and measure the relative amount of ozone in the atmosphere.

Although the methods of measurement have been modified over the years, scientists continued to use KI to measure ozone for more than a century. One modification involved pumping ambient air through a KI solution and measuring the amount of iodide being converted to iodate since an electrical current is created in the solution

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

as the conversion takes place. The reaction took place within a matter of seconds, and the amount of electric current was easily quantifiable. This method, known as the *wet method*, was the predominant way ozone was measured until the 1960s, when other methods using newer optical technology became available and increased the accuracy of the measurements. One problem with the wet method was that other chemicals in the atmosphere interfered with the chemical reaction. The most common of these interfering trace gases is sulfur dioxide (SO_2), a pollutant that is primarily a by-product of coal combustion.

In the early part of the twentieth century, ground-based and balloon-borne measurements discovered that most of the atmosphere's ozone is located in the stratosphere with highest concentrations located between 15 and 30 km. For a long time, it was believed that tropospheric ozone originated from the stratosphere and that most of it was destroyed by contact with Earth's surface. Ozone was known to be produced by the photodissociation of molecular oxygen, O_2, a process that can only occur at wavelengths shorter than 242 nm. The atomic oxygen formed as a product of this photodissociation would then recombine with another oxygen molecule to make ozone. Because such short-wavelength radiation is present only in the stratosphere, no tropospheric ozone production is possible by this mechanism. In the 1940s, however, it became obvious that production of ozone was also taking place in the troposphere. The overall reaction mechanism was eventually identified by Arie Haagen-Smit of the California Institute of Technology located in highly polluted southern California. The smog chemistry hypothesized by Haagen-Smit was still thought to be a relatively small source on the global scale since ~90% of the ozone was located in the stratosphere, creating a ubiquitous source of tropospheric ozone as stratosphere air was transported into the troposphere. It was not until the 1970s that this viewpoint was challenged when Paul Crutzen (Crutzen, 1974) and other scientists at the time showed that consideration of "smog chemistry" in the background troposphere could produce a sizable source of tropospheric ozone and must be included in the global tropospheric ozone budget. Crutzen's pioneering work on tropospheric ozone was noted when he received the Nobel Prize for Chemistry in 1995.

2 CHEMISTRY OF TROPOSPHERIC OZONE FORMATION

Photodissociation of NO_2 by (visible) sunlight is the only significant anthropogenic source of O_3 in the troposphere

$$NO_2 + h\nu \rightarrow NO + O \qquad (\lambda < 420\ \text{nm}) \tag{1}$$

immediately followed by

$$O_2 + O + M \rightarrow O_3 + M \tag{2}$$

where the M in reaction (2) represents any nonreactive molecule that absorbs some of the excess energy of the intermediate product formed in the reaction (2).

The atmospheric oxidation of a hydrocarbon, RH, is initiated by reaction with the hydroxyl radical (OH):

$$RH + OH \rightarrow R + H_2O \qquad (3)$$

where RH can be any molecule containing a hydrogen and a carbon, including methane, CH_4, or any nonmethane hydrocarbon consisting of more than one carbon atom. The product is another radical, denoted R, and water vapor. The radical quickly combines with an oxygen molecule in a three-body reaction:

$$R + O_2 + M \rightarrow RO_2 + M \qquad (4)$$

to form another oxygenated radical, RO_2, called a peroxy radical. The peroxy radicals are the key for converting NO to NO_2:

$$RO_2 + NO \rightarrow RO + NO_2 \qquad (5)$$

In addition, RO attaches to an oxygen molecule to form another peroxy radical:

$$RO + O_2 \rightarrow HO_2 + R'CHO \qquad (6)$$

where R'CHO is an aldehyde (and noting that R′ is a shorter chained carbon radical than R). The HO_2 likewise reacts with NO to form another NO_2 molecule:

$$HO_2 + NO \rightarrow OH + NO_2 \qquad (7)$$

where the two NO_2 molecules photolyze and eventually produce ozone:

$$NO_2 + h\nu \rightarrow NO + O \quad \text{(2 times)} \qquad (1)$$

$$O_2 + O + M \rightarrow O_3 + M \quad \text{(2 times)} \qquad (2)$$

Net: $$RH + 4O_2 + 2h\nu \rightarrow R'CHO + 2H_2O + 2O_3.$$

Additional ozone molecules can also be produced through the oxidation of R'CHO. In this reaction sequence, it is important to note that the nitrogen oxide emitted as a pollutant is still available to make more ozone. If NO were not present in the atmosphere, ozone would not be formed. In fact, the presence of many nonmethane hydrocarbons, by themselves, would result in a destruction of ozone since they, or some of their daughters of the oxidation process, could react with any ozone present in the atmosphere.

On the other hand, if only nitrogen oxides and ozone were present in the atmosphere, an equilibrium would quickly be established since O_3 reacts quickly with NO:

$$NO + O_3 \rightarrow NO_2 + O_2 \qquad (8)$$

and the ratio among NO, NO_2, and O_3 is quickly established by the rates of the reactions among these species:

$$[NO]/[NO_2] = j_1/[O_3]k_8$$

where the brackets denote the concentration of a particular species, j_1 is the rate of photolysis of NO_2, and k_8 is the rate of reaction (8); the relationship among these three gases defined by this ratio is often referred to as the photostationary state and has had an important implication for understanding the formation of ozone near urban areas and subsequent strategies developed for the reduction of ozone concentrations.

3 GLOBAL DISTRIBUTION OF TROPOSPHERIC OZONE

The distribution of tropospheric ozone can be determined from the analyses of satellite data sets obtained independently from two different instruments: The Total Ozone Mapping Spectrometer (TOMS) and the Stratospheric Aerosol and Gas Experiment (SAGE). Between October 1978 and May 1993, TOMS functioned on the *Nimbus 7* satellite and provided daily maps of the distribution of total ozone. Additional TOMS were launched in 1991 (on the Russian *Meteor* satellite) and two in 1996 (see Chapter 21; "Stratospheric Ozone Observations"). The National Aeronautical and Space Administration's (NASA's) Earth Observing System (EOS) now is operational and total ozone will be measured as part of EOS. Total ozone is defined as the integrated amount of ozone between the surface and the top of the atmosphere. A unit of measure for total ozone is a quantity known as the Dobson unit (DU), where $1\ DU = 2.69 \times 10^{16}$ molecules O_3/cm^2. If this amount of ozone were brought down to standard atmospheric temperature and pressure, the depth of this column would be 1 mm. Thus, another common measure of column ozone is mm-atm, where a mm-atm is equivalent to 1 DU. A typical amount of total ozone found in the atmosphere is 300 DU, and approximately 90% of this ozone is located in the stratosphere.

At middle and high latitudes, the distribution of total ozone is primarily governed by the prevailing large-scale circulation patterns. These patterns can vary substantially on a daily basis, and intense gradients of total ozone have been observed with differences of 200 DU at locations less than a few thousand kilometers apart. At these higher latitudes, total ozone amounts can range between $\sim$225 and $\sim$500 DU. Only recently have values as low as 100 DU been observed during austral spring in conjunction with the Antarctic ozone hole.

At lower latitudes, however, the total ozone distribution patterns exhibit much smaller gradients than at middle and high latitudes. The intense gradients of as much as 200 DU found at the higher latitudes are replaced by much more subtle gradients of no more than 20 to 30 DU. Because the primary intent of the measurement of total ozone was to study the distribution of stratospheric ozone, very little research was conducted using the information provided by TOMS in the tropics. Subsequently, however, it has been shown that the variations in total ozone at low latitudes were primarily the result of variability of ozone in the troposphere even though only ~10% of the total ozone was in the troposphere.

The use of TOMS for tropospheric studies has taken a substantive step further when data from SAGE were used to derive the amount of ozone in the stratosphere (Fishman et al., 1990). Ozone measurements from the SAGE instruments (SAGE was launched in February 1979 and operated through November 1981; SAGE II was launched in November 1984 and is still operating) provide the vertical distribution of ozone in the stratosphere. From these profiles, the amount of ozone in the stratosphere can be integrated and then subtracted from the co-located total ozone amount derived independently from the TOMS on the same day.

The distribution of the integrated amount of tropospheric ozone as a function of season is shown in Figure 1 (Fishman et al., 2002). These seasonal depictions show that there is considerably more ozone in the Northern Hemisphere than in the Southern Hemisphere, especially during the summer. During most of the seasons, distinct plumes that seem to result from pollution originating in North America, Asia, Africa, and Europe can be observed. In the three northern continents, the plumes originate over the eastern portions of each landmass and are transported by the prevailing westerly winds for several thousand kilometers. At low latitudes, the highest concentrations of pollution are off the west coast of Africa and is most pronounced during austral spring (September–November). At these latitudes, the prevailing low-level winds are trade winds (easterlies), which would carry the emissions from central and western Africa to the eastern tropical South Atlantic Ocean. The prevailing upper level winds are westerlies, so any ozone that gets to altitudes of ~5 km or higher are transported long distances to the east. Evidence of the long-range transport of emissions from biomass burning in Africa and South America to Australia is evident in long-term Australian data sets of not only ozone but also carbon monoxide and elemental carbon, two other products of widespread burning.

4 TROPOSPHERIC OZONE TRENDS IN NONURBAN TROPOSPHERE

The global distribution of tropospheric ozone shown in Figure 1 illustrates its wide range (more than a factor of 3) of abundance. Therefore, unlike trace gases such as chlorofluorocarbons, nitrous oxide, or carbon dioxide, which exhibit very small spatial gradients, an assessment of the *global* rate of increase of tropospheric ozone is difficult to determine from measurements at only a few locations. Outside of urban areas, only a few stations around the world have continuous long-term measurements of tropospheric ozone. Among these stations are the ones set up by the U.S. National Oceanographic and Atmospheric Administration (NOAA), which has maintained a carefully calibrated monitoring program at a number of stations around the world since the early 1970s (Oltmans and Levy, 1994). The monthly mean concentrations from Barrow and Mauna Loa are shown on the left side of Figure 2*a*. The linear least-squares fit illustrating the trend between 1973 and 1992 for these two data sets is also plotted on these figures. Even though both of these stations show a significant increase over this period, the measurements at Barrow

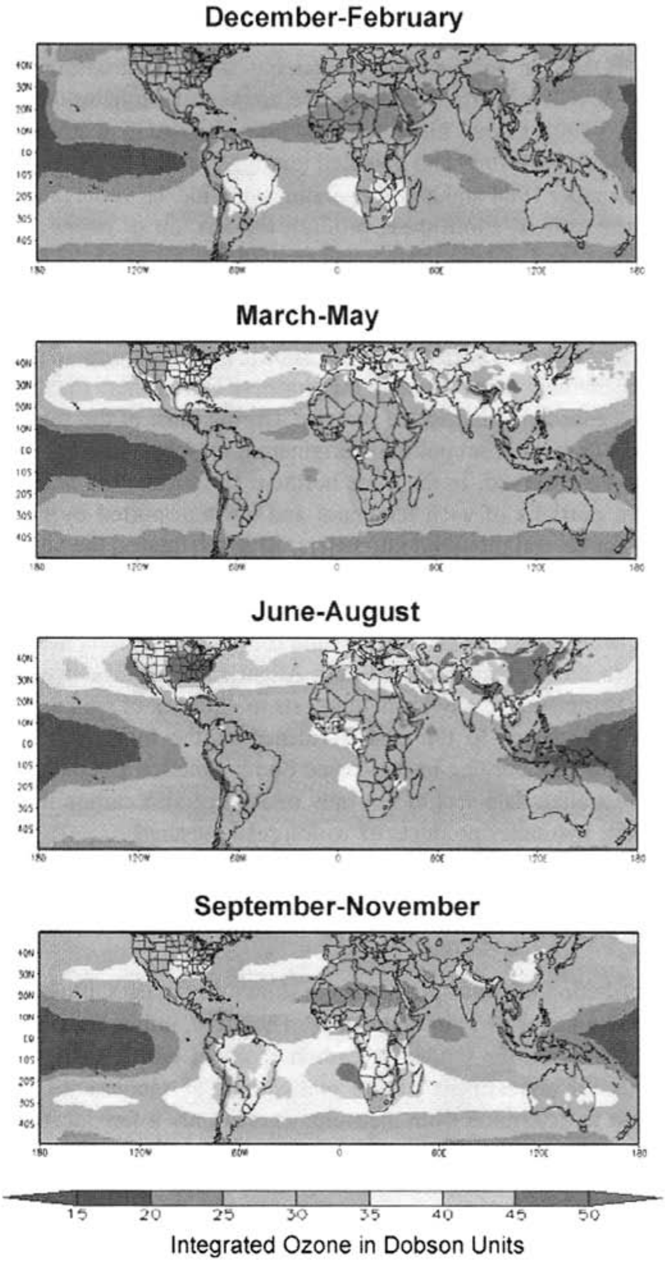


Figure 1 (see color insert) Climatological distribution of tropospheric ozone derived from satellite measurements between 1979 and 2000 (from Fishman et al., 2002). Units of contours and Dobson Units (DU). Regions greater than 40 DU have been shaded. See ftp site for color image.

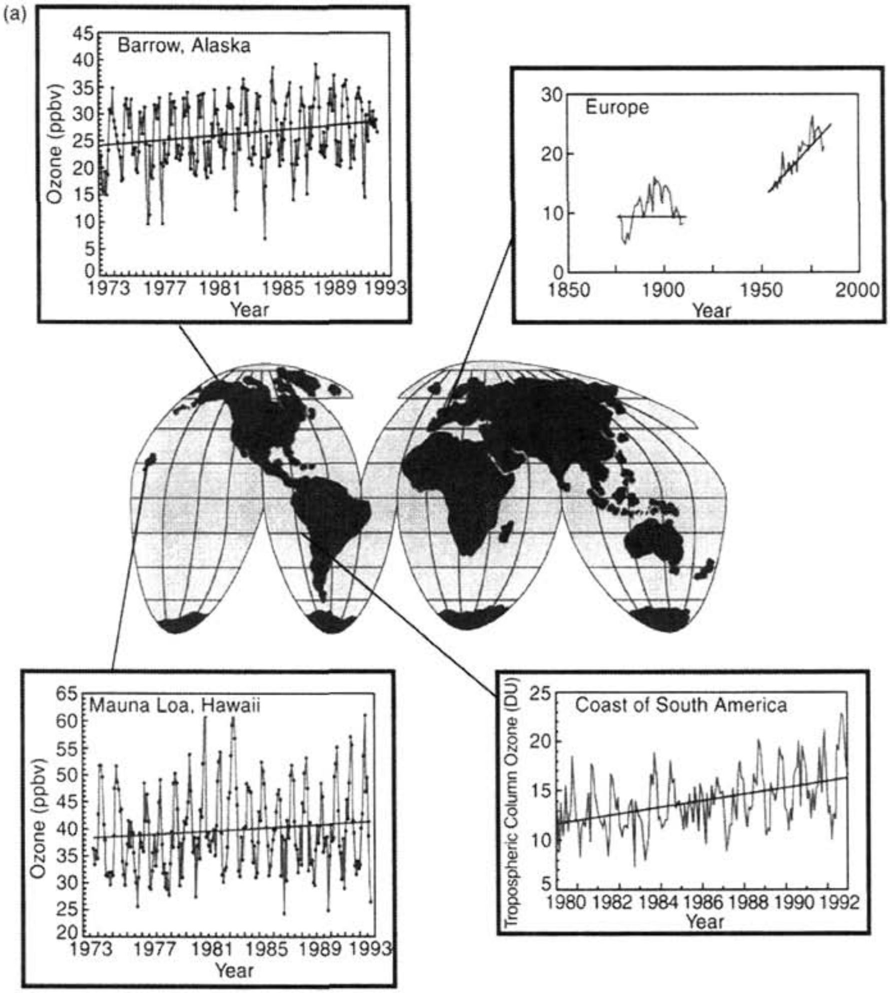


Figure 2 (*a*) (Upper left): monthly mean surface ozone at Barrow and the linear trend for the entire data record. (Lower left): monthly mean surface ozone at Mauna Loa with the linear trend. (Upper right) annual mean ozone concentrations at Montsouris Observatory outside Paris (1876–1910) and Arkona, East Germany (1956–1984). The average ozone concentration at the beginning of the twentieth century near Paris was less than 10 ppb whereas in 1985 the typical ground-level concentrations in central Europe is approaching 30 ppb, implying an increase of about 200% during the century (from Volz and Kley, 1988). (Lower right): ozone trend off coast of South America determined from analysis of satellite measurements of total ozone (from Jiang and Yung, 1996). See ftp site for color image.

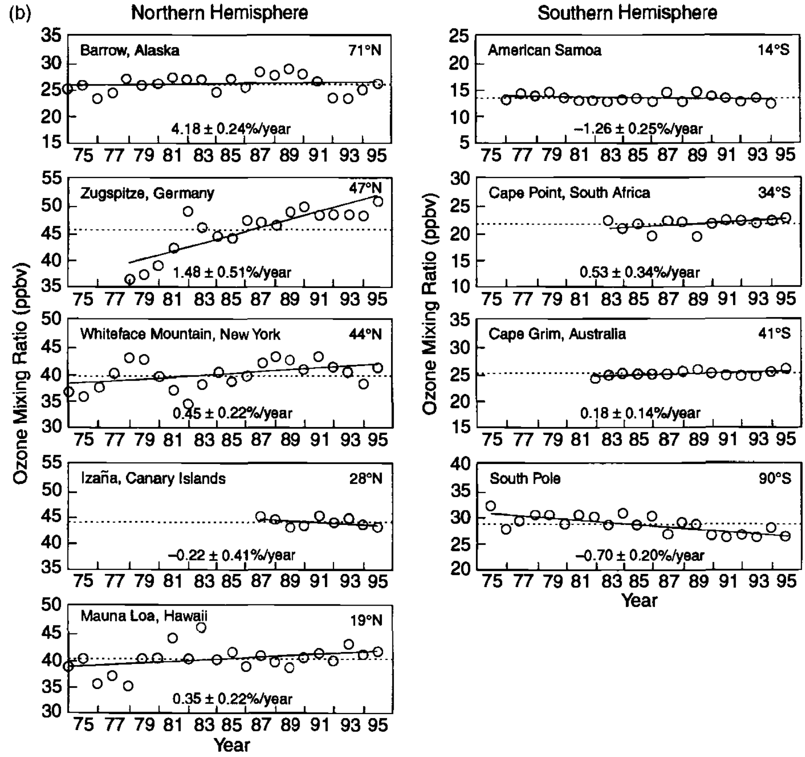


Figure 2 (*b*) Annual average ozone mixing ratios (ppbv) for surface ozone measuring sites. The dashed line is the long-term average. The solid line is the linear least-squares fit to the plotted values. The linear trend and 95% confidence in percent per year is given with each location (from Oltmans et al., 1998).

show that the long-term trend has a strong seasonal dependence; the increase during the summer is 1.73% per year whereas there is almost no trend (−0.07% per year) during the winter. Figure 2*b* summarizes a number of long-term measurements from the NOAA network as well as a few other stations where comparable data exists in the background atmosphere. Curiously, some stations such as American Samoa near the equator show a slight negative trend, whereas a significant negative trend exists at South Pole. The reason for these trend differences at these remote sites is not clear and is currently being studied.

Modern studies have reexamined the Schönbein paper ozone measurements from the late nineteenth century and early twentieth century to determine tropospheric ozone trends over longer time periods. These studies have carefully examined calibration procedures used last century and have determined that a significant increase in tropospheric ozone has occurred over the past century.

More than three decades of measurements using Schönbein's technique were obtained at the Montsouris Observatory outside Paris. The instrument used at this meteorological station was recalibrated and the observations were converted to standard units of measurement consistent with modern measurements. The results from this data set are compared with modern observations obtained in Germany and depicted in the upper right panel of Figure 2*a*. This and other analyses strongly suggest that ozone at the surface has risen from ~10 ppbv to more than 30 ppbv in nonurban Europe and the eastern United States. Although ozone at the surface has likely increased significantly on the time scales of years and decades since the inception of the industrial era, tropospheric measurements above the surface are extremely scarce and difficult to interpret because of the different methods of measurement used since the 1960s. Most of the measurements are from ozonesondes (an ozone sensor placed on a balloon), but several types of sensors have been used and each type is susceptible to interference from other trace gases in the atmosphere. Despite the uncertainty in the measurements, it is generally believed that ozone has increased throughout the entire troposphere since the 1960s, when ozonesonde measurements started on a fairly regular basis.

5 GLOBAL TROPOSPHERIC OZONE BUDGET

The components of the global tropospheric ozone budget can be broken into four general categories: transport from the stratosphere, destruction at Earth's surface, photochemical destruction, and in situ photochemical production. The primary mechanism by which ozone is transported from the stratosphere into the troposphere is through meteorological events referred to as stratospheric intrusions. These events occur in conjunction with the movement of air associated with rapid changes in the intensity and position of the jet stream, the fast-moving westerly river of air that often delineates the position of strong frontal boundaries at middle latitudes. Under these conditions, the tropopause (i.e., the boundary between the troposphere and the stratosphere) often becomes contorted and its position becomes difficult to define and often takes on a "folded" depiction (see Chapter 1, "Overview: Atmospheric Chemistry). Because of this, stratospheric intrusions are also synonymous with tropopause folding events.

The topic of stratosphere–troposphere exchange was an intense research area in the 1960s and early 1970s because of the concern of transport of radioactive debris created by atmospheric nuclear bomb testing from the stratosphere into the lower atmosphere and eventually its deposition to plants, animals and human populations. During this time, the North American Ozonesonde Network was established for the primary purpose of understanding how stratospheric air was transported into the troposphere. From these data, it is generally thought that ~10% of the stratosphere is exchanged annually with the troposphere. From these estimates, the global source of tropospheric ozone from the stratosphere, which was assumed the primary *natural* source of tropospheric ozone could be computed (e.g., Danielsen and Mohnen, 1977).

The other primary component of the global budget of tropospheric ozone is its sink, or how it is destroyed once it is in the troposphere. The early measurements of ozone's vertical distribution always showed that lowest concentrations were near Earth's surface, implying a sink for ozone as it came in contact with the ground. These measurements generally showed much sharper vertical gradients over land and vegetated surfaces than over water and ice surfaces. Thus, one way to determine this deposition sink globally was to make a series of field measurements over a representative sample of surfaces and extrapolate these measurements to the rest of the world. Using this methodology, the globally averaged destruction rate of tropospheric ozone generally converged to a value near 8 to 10×10^{10} molecules O_3/cm^2 s. The accuracy of these estimates was claimed to be ~30%. These calculations were consistent with the few attempts to extrapolate the global input from the stratosphere resulting from stratosphere–troposphere exchange studies, which indicated that a global average of $\sim 8 \times 10^{10}$ molecules O_3/cm^2 s came from the stratosphere. Thus, up until the early 1970s, it was generally believed that the tropospheric ozone budget was balanced by the natural input from the stratosphere and the destruction at Earth's surface (Fabian and Junge, 1970). The potential impact of local-scale photochemical generation (as was known at the time for areas such as southern California) was believed to be insignificant.

A series of studies published shortly thereafter challenged this assumption and proposed that a natural source of tropospheric ozone of comparable magnitude to that of input from the stratosphere existed in the background atmosphere as a result of methane oxidation. For the first time, the paradigm of the tropospheric ozone budget was challenged resulting in a lively debate in the scientific literature in the middle and late 1970s (Chameides and Walker, 1973; Fabian, 1973; Fishman and Crutzen, 1978). These theoretical studies primarily concentrated on the generation of ozone from the oxidation of methane and carbon monoxide, the two most abundant trace gases that could lead to the photochemical formation of tropospheric ozone.

Another important component of the tropospheric ozone budget is its photochemical destruction. As ozone enters from the stratosphere, for example, it is photolyzed at shorter wavelengths to produce an excited state of atomic oxygen, $O(^1D)$, rather than its ground state, $O(^3P)$:

$$O_3 + h\nu \rightarrow O(^1D) + O_2 \qquad (\lambda < 320\ \text{nm}) \tag{9}$$

Once $O(^1D)$ is formed, it can react with water vapor to generate OH:

$$O(^1D) + H_2O \rightarrow 2OH \tag{10}$$

In turn, OH can react with ozone to form the hydroperoxy radical, which can set up a catalytic cycle of ozone destruction, analogous to what happens in the stratosphere:

$$\begin{array}{rll} & OH + O_3 \rightarrow HO_2 + O_2 & (11) \\ \text{followed by} & \underline{HO_2 + O_3 \rightarrow OH + 2O_2} & (12) \\ & 2O_3 \rightarrow O_2 \quad \text{(net reaction sequence)} & \end{array}$$

Reaction (10) is the primary source of OH in the troposphere, and subsequent reactions with OH are the primary means by which most chemicals released to the atmosphere are oxidized and eventually removed. Whereas photochemistry was first proposed as an important photochemical source of tropospheric ozone in the studies written in the early 1970s, it is important to also note that photochemistry is also the dominant sink and is the primary reason that ozone concentrations are generally very low in the tropical troposphere where both water vapor and incoming solar flux are highest. The key to whether photochemistry is a net source or a net sink for tropospheric ozone is most dependent on how much NO is present.

6 CURRENT UNDERSTANDING OF TROPOSPHERIC OZONE BUDGET

The global distribution of tropospheric ozone presented earlier in this chapter illustrates its heterogeneity and underscores the difficulty of quantifying a global budget using the simplistic assumptions about its vertical distribution that had been employed when budgets neglecting photochemical processes were formulated. It is clear from the depiction in Figure 1 that local-scale photochemical generation of ozone has had a considerable impact on the global distribution as evidenced by the dominant plumes originating over North America, Europe, Asia, and Africa. A proper calculation of the tropospheric ozone budget must quantify these local- and regional-scale processes that feed into the global budget. Studies investigating photochemical processes from industrial emissions of volatile organic compounds and nitrogen oxides on scales of ~1000 km showed that the ozone generated on these scales should at least be comparable to the amount generated in the background through methane and carbon monoxide oxidation. In addition, the data now indicate that large quantities of ozone are generated in the tropics as emissions from widespread vegetation burning are oxidized efficiently in the intense tropical sunshine. Furthermore, some recent analyses of ozonesonde data have concluded that very little (perhaps as small as 5%) ozone near the ground had originated in the stratosphere and only ~25% of the ozone observed at 300 mbar had originated in the stratosphere. This analysis agrees with more recent estimates of stratosphere–troposphere mass exchange suggesting that the amount of ozone from the stratosphere is likely only ~30% of the amount determined from the earlier estimates determined in the 1970s.

Calculations from a general circulation model, which includes a complete set of photochemical reactions, have been used to evaluate the tropospheric ozone budget (Wang et al., 1998). The results from these model calculations are shown in the four seasonal panels in Figure 3. These calculations show how the chemical terms are both considerably larger than the input from the stratosphere and the amount of destruction at the ground. In addition, the amount of ozone produced photochemically is generally greater than the amount destroyed. The largest amount of production is at northern middle latitudes in July. The Southern Hemisphere is also a sizable source in both July and October, when biomass burning is most prevalent in the southern tropics and subtropics.

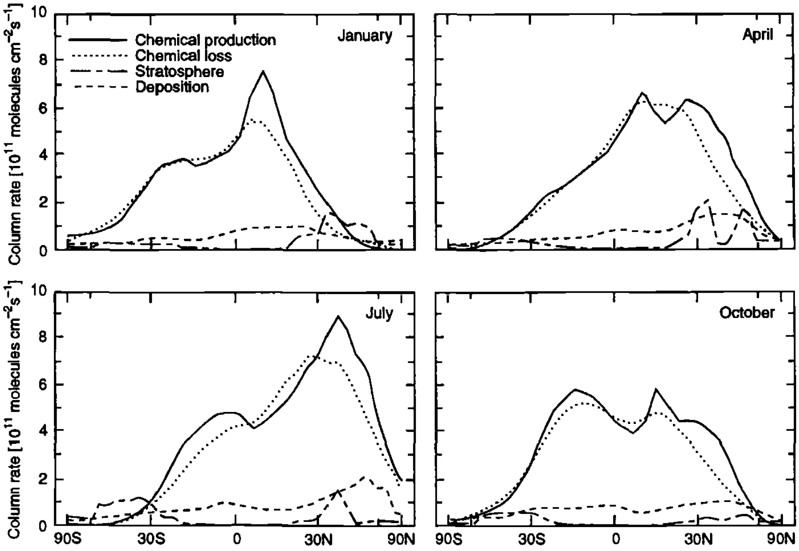


Figure 3 Zonally averaged column budget for tropospheric ozone in different seasons including term from in situ photochemical production and loss, transport from the stratosphere, and deposition. The abscissa scale is linear in sine of latitude (from Wang et al., 1998).

These studies, as well as the documented increase in tropospheric ozone over time scales of decades provide fairly strong evidence that its distribution has changed significantly over the last century and that a large fraction of the tropospheric ozone budget is now likely controlled by anthropogenic pollution from both industrialized and tropical regions of the world. Studies are currently underway to provide more quantitative information, and our understanding of tropospheric ozone will greatly improve as more data are analyzed and more sophisticated global models are developed to study the problem.

REFERENCES

Chameides, W. L., and J. C. G. Walker, A photochemical theory of tropospheric ozone, *J. Geophys. Res.*, *78*, 8751–8760, 1973.

Crutzen, P. J., Photochemical reactions initiated by and influencing ozone in unpolluted tropospheric air, *Tellus*, *26*, 47–57, 1974.

Danielsen, E. F., and V. A. Mohnen, Project dustorm report: Ozone transport, in situ measurements and meteorological analyses of tropopause folding, *J. Geophys. Res.*, *82*, 5867–5877, 1977.

Fabian, P., A theoretical investigation of tropospheric ozone and stratospheric-tropospheric exchange processes, *Pure Appl. Geophys.*, *106–108*, 1044–1057, 1973.

Fabian, P., and C. E. Junge, Global rate of ozone distribution at the earth's surface, *Arch. Meteor. Geophys. Biokl. Ser. A.*, *19*, 161–172, 1970.

Fishman, J., and P. J. Crutzen, The origin of ozone in the troposphere, *Nature*, *274*, 855–858, 1978.

Fishman, J., C. E. Watson, J. C. Larsen, and J. A. Logan, Distribution of tropospheric ozone determined from satellite data, *J. Geophys. Res*, *95*, 3599–3617, 1990.

Fishman, J., A. E. Balok, and F. M. Vukovich, Observing tropospheric trace gases from space: recent advances and future capabilities, *Adv. Space Res.* **29**, 1625–1630, 2002.

Jiang, Y., and Y. L. Yung, Concentrations of tropospheric ozone from 1979 to 1992 over tropical Pacific South America from TOMS data, *Science*, **272**, 745–748, 1996.

Oltmans, S. J., and H. Levy II, Surface ozone measurements from a global network, *Atmos. Environ.*, *28*, 9–24, 1994.

Oltmans, S. J., et al., Trends of tropospheric ozone in the troposphere, *Geophys. Res. Lett.*, *25*, 139–142, 1998.

Volz, A., and D. Kley, Evaluation of the Montsouris series of ozone measurements made in the nineteenth century, *Nature*, *252*, 240–242, 1988.

Wang, Y., D. J. Jacob, and J. A. Logan, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 3. Origin of tropospheric ozone and effects of nonmethane hydrocarbons, *J. Geophys. Res.*, *103*, 10757–10767, 1998.

CHAPTER 4

NITROGEN OXIDES AND OTHER REACTIVE NITROGEN SPECIES

J. H. CRAWFORD, J. D. BRADSHAW, D. D. DAVIS, AND S. C. LIU

1 INTRODUCTION

Nitrogen is most abundant in the atmosphere in its molecular form, N_2, which comprises 78% of Earth's atmosphere. Although virtually inert and of no direct consequence to tropospheric chemistry, this vast reservoir of atmospheric nitrogen enables the existence of trace levels of nitrogen oxides that play a number of critical roles in the chemistry of the atmosphere. Nitrogen oxides, commonly referred to as NO_x, are defined by the sum of the chemical species NO and NO_2. These two atmospheric constituents are grouped for convenience due to their fast photochemical cycling, which brings them into equilibrium generally within a few minutes. The greater family of reactive nitrogen, conventionally denoted by the term NO_y, consists of NO_x as well as a suite of other compounds including NO_3, $N2O_5$, HNO_3, HONO, HO_2NO_2, peroxyacetylnitrate (PAN), and a wide array of other organic nitrogen-containing species. These compounds play important roles in the removal of reactive nitrogen from the atmosphere as well as the transport of reactive nitrogen from source regions to remote areas.

Tropospheric chemical cycles involving NO_x are of fundamental importance to understanding several key atmospheric issues. For instance, the tropospheric ozone abundance is largely regulated by catalytic photochemical cycles involving NO_x, CO, and hydrocarbons that produce ozone (see Chapter 3). NO_x often represents the rate-limiting precursor for ozone production, especially throughout the remote atmosphere. This is due to its short lifetime relative to other precursors. On a regional scale, the role of NO_x in creating high concentrations of ozone detrimental to human health is a major air quality issue in many urban areas. On a global scale, the impact

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

of NO_x on ozone represents an important factor in determining the oxidizing capacity of the atmosphere. NO_x further impacts atmospheric oxidation rates by regulating OH concentrations, especially at high altitudes and latitudes. Since the primary mechanism for removing many pollutant gases from the atmosphere is reaction with OH, NO_x is important to the atmosphere's ability to cleanse itself. This in turn relates to the issue of climate change regarding removal of greenhouse gases such as CH_4. Another important link to the issue of climate change involves the impact of NO_x on ozone production in the upper troposphere where it is most effective as a greenhouse gas.

The tropospheric distribution of NO_x is complicated by a combination of diverse sources. Natural as well as anthropogenic sources exist both at the surface (e.g., soil emissions, biomass burning, and fossil fuel combustion) and in the free troposphere (e.g., lightning, aircraft, and stratosphere–troposphere exchange). Regeneration of NO_x through chemical recycling of various NO_y species represents a secondary source of NO_x in the troposphere. Also, the atmospheric lifetime of NO_x ranges from hours to days depending predominantly on altitude. As a result, NO_x mixing ratios vary from a few parts per trillion in some remote regions to several parts per billion in highly polluted conditions. Given the high variability of NO_x and its importance to several key atmospheric issues, the global NO_x distribution represents a pivotal subject in efforts to fully understand the current state of our atmosphere as well as its future evolution.

2 CHEMICAL TRANSFORMATIONS AND SPECIATION OF REACTIVE NITROGEN

Almost all reactive nitrogen is introduced into the atmosphere as NO, but within minutes, NO reaches equilibrium with NO_2. This NO_x is subsequently transformed into other NO_y species that can be removed, transported, or recycled back to NO_x. A general outline of these transformations is represented in Figure 1, which accompanies the following discussion of the behavior and importance of various NO_y species.

1. *NO_x $(NO + NO_2)$*. During the day, NO and NO_2 experience rapid interconversion via the following simple reaction scheme:

$$NO + O_3 \rightarrow NO_2 + O_2 \qquad (1)$$

$$NO_2 + h\nu \rightarrow NO + O \qquad (2)$$

$$\underline{O + O_2 \rightarrow O_3} \qquad (3)$$

$$\text{Net:} \quad \text{no change}$$

This reaction sequence is a null cycle that serves no purpose photochemically other than to partition NO_x into NO and NO_2. NO may also be converted to NO_2 by hydroperoxy radicals (HO_2) that result from the oxidation of CO as well as organic

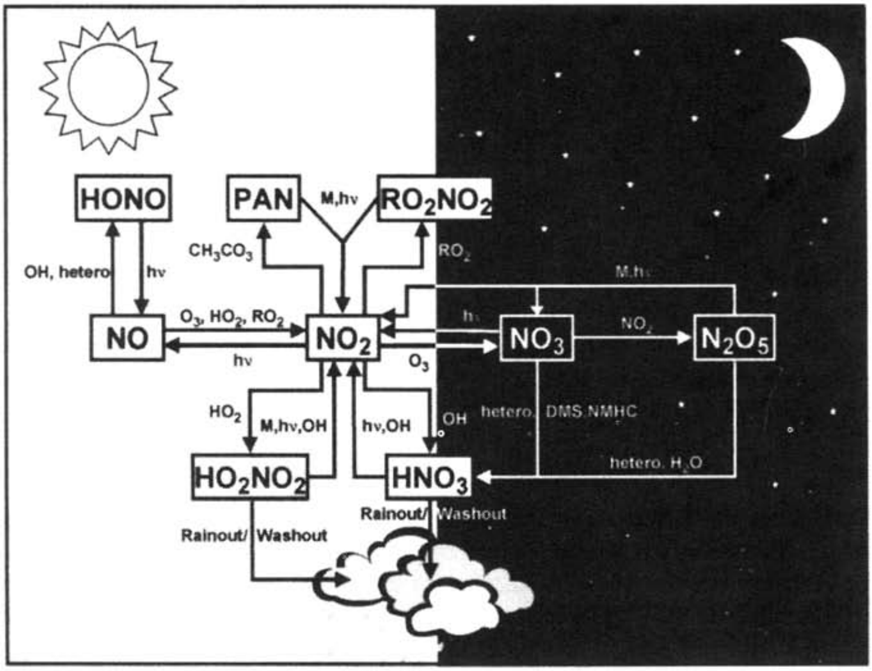


Figure 1 General schematic of reactive nitrogen chemistry in the troposphere arranged to emphasize dominant pathways in the presence of sunlight and in darkness.

peroxy radicals (RO_2, where R denotes a CH_3 or higher organic grouping) that result from the oxidation of hydrocarbons. This cycle of NO–NO_2 interconversion has impacts outside the NO_2–NO system.

$$NO + HO_2 \rightarrow NO_2 + OH \tag{4}$$

$$NO_2 + h\nu \rightarrow NO + O \tag{2}$$

$$O + O_2 \rightarrow O_3 \tag{3}$$

$$\text{Net:} \quad HO_2 + O_2 \rightarrow OH + O_3$$

Here, two key impacts of NO_x interconversion are the formation of O_3 and the regeneration of OH from HO_2.

The conversion of NO_x into other longer-lived NO_y reservoir species is accomplished almost exclusively through reactions involving NO_2 (see Figure 1). Thus, the lifetime of NO_x in the atmosphere relies in part on the partitioning of NO_x between its constituents, NO and NO_2. As shown in Figure 2, the fraction of NO_x existing in the form of NO_2 changes dramatically with altitude. At the surface, NO_x tends to be predominantly in the form of NO_2 since reaction (1) proceeds at a faster rate than reaction (2). Here, NO_x lifetimes are typically one day or less. Reaction (1), however, has a strong temperature dependence and is about 5 times slower at the cold temperatures of the upper troposphere, thus shifting the NO_x equilibrium in favor of NO. To a lesser degree, the increase in reaction (2) with altitude (~50%) also contributes to an NO_x partitioning that favors NO at high altitude. As NO_2 becomes a smaller fraction of NO_x with increasing altitude, the lifetime of NO_x lengthens. In the upper troposphere, NO_x lifetimes can be a few days to a week. The longer lifetime of NO_x at high altitude tends to enhance its per-molecule efficiency in the production of ozone since NO can be cycled through reaction (4) more times before being lost. Although efficiency is increased at high altitude, the ozone production rate per molecule of NO_x is slower owing to the lower abundance of HO_2, which generally decreases with altitude.

2. *NO_3*. The nitrate radical, NO_3, photolyzes within a few seconds in sunlight; thus, it is of negligible importance to the daytime photochemistry of the atmosphere. NO_3 is formed by the reaction of NO_2 with O_3.

$$NO_2 + O_3 \rightarrow NO_3 \tag{5}$$

Overnight at the surface, a significant fraction of NO_2 may be converted by this reaction. The strong temperature dependence of reaction (5), however, slows conversion rates by an order of magnitude for the upper free troposphere, where only a small fraction of NO_2 may be converted overnight. Given its concentration and high

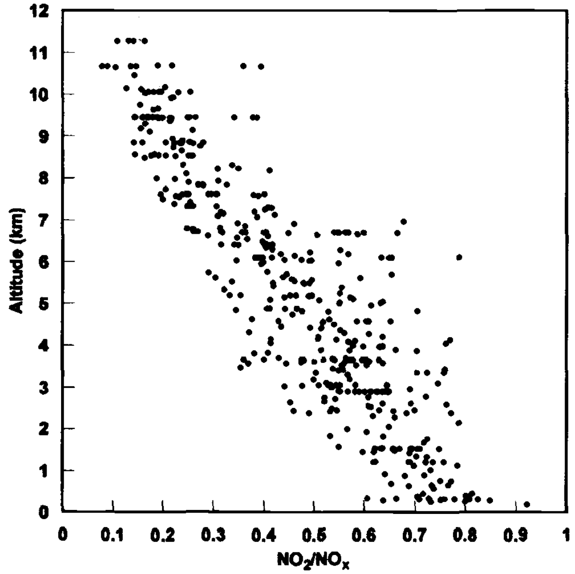


Figure 2 Fraction of NO_x in the form of NO_2 as a function of altitude. Data based on concurrent measurements of NO and NO_2 conducted during NASA's PEM–Tropics A field campaign (Bradshaw et al., 1999).

reactivity, NO_3 can be competitive with OH as an oxidant of dimethylsulfide (DMS) in the marine boundary layer of coastal regions. In general, however, marine NO_x levels are insufficient to support more than a minor role for NO_3 in DMS oxidation. In the continental boundary layer, NO_3 can be important in the oxidation of unsaturated hydrocarbons, e.g., olefins and biogenic hydrocarbons such as isoprene.

3. *N_2O_5*. N_2O_5 is formed during periods of darkness by the reaction of NO_3 and NO_2.

$$NO_3 + NO_2 + M \rightarrow N_2O_5 + M \tag{6}$$

Here, M represents an inert third body, typically N_2 or O_2. N_2O_5 is a thermally labile species; therefore, the equilibrium represented by reaction (6) favors NO_3 near the surface. Larger concentrations of N_2O_5 exist at high altitude where it is favored by colder temperatures, although its concentration is still limited by the slowdown in reaction (5). N_2O_5 also has a lifetime due to photolysis of several hours, but this is not sufficient to prevent significant daytime concentrations in the upper troposphere. N_2O_5 represents a loss of NO_x through its heterogeneous conversion to nitric acid (HNO_3) on aerosol surfaces.

4. *HNO_3*. Nitric acid is believed to be the major reservoir species for NO_x. It is formed primarily by the reaction of NO_2 and OH.

$$NO_2 + OH + M \rightarrow HNO_3 \tag{7}$$

HNO_3 can also result from the heterogeneous reaction of N_2O_5 on aerosols or the reaction of NO_3 with certain hydrocarbons. HNO_3 is efficiently removed from the atmosphere through dry deposition and rainout processes. HNO_3 may also be recycled back into NO_x through either reaction with OH or photolysis with a lifetime of a few weeks. At low altitude these two processes are slow compared to wet removal, but they can be important at high altitude where wet removal is less frequent.

5. *HONO*. Nitrous acid is formed by the gas-phase reaction of NO and OH. It also photolyzes within minutes and thus is a negligibly small component of NO_y. Although the details are not fully understood, evidence exists for nighttime formation of HONO, possibly through heterogeneous processes, based on elevated nighttime observations of HONO. Substantial buildups of HONO may take place in special environments such as overnight in polluted air rich in NO_x or in polar regions with extended periods of darkness. For these special conditions, HONO may for a short time be the dominant source of OH through its rapid photolysis during sunrise periods.

6. *HO_2NO_2*. Pernitric acid is a thermally labile species resulting from the reaction of HO_2 and NO_2. As with N_2O_5, it is favored at the cold temperatures of the upper troposphere. In the upper troposphere, HO_2NO_2 concentrations are limited by reaction with OH and photolysis resulting in a lifetime of only a couple of days, but

concentrations should approach that of NO_x. HO_2NO_2 may also return to NO_x through thermal decomposition in descending air masses.

7. *PAN*. Peroxyacetylnitrate is the most common organic nitrogen-containing species resulting from the reaction of NO_2 with the CH_3CO_3 radical. The CH_3CO_3 radical results from oxidation of a wide range of hydrocarbons, but at high altitude oxidation of acetone appears to be predominantly responsible. While loss in the lower troposphere is dominated by thermal decomposition, loss in the upper troposphere occurs through photolysis with a lifetime of 1 to 2 months. At high altitude, PAN is thermally stable and serves as an effective reservoir for global-scale transport of NO_y. In remote regions, thermal decomposition of PAN in descending air masses can be a dominant source of NO_x. Somewhat analogous to HO_2NO_2, organic peroxy radicals (RO_2) can react with NO_2 to form organic species (RO_2NO_2) with properties similar to those of PAN.

8. CH_3ONO_2. Although not depicted in Figure 1, methyl nitrate represents the most common member of a family of alkyl nitrates that can result from NO_x in the presence of hydrocarbon oxidation. There is also evidence that these species are emitted from the ocean in small amounts. Loss is primarily through photolysis to yield NO_x.

3 SOURCES OF REACTIVE NITROGEN

Unlike most trace species that are emitted only at the surface, NO_x sources exist both at the surface and in the free troposphere. Surface sources include both natural and anthropogenic sources; e.g., soil/microbial emission, fossil fuel combustion, and biomass burning. In the free troposphere, NO_x sources include lightning, aircraft emissions, and stratosphere–troposphere exchange. Table 1 gives estimated source strengths and uncertainties for each of these sources. While there are still substantial uncertainties in these sources, the global NO_x source strength is clearly dominated by surface sources with anthropogenic use of fossil fuels having the greatest contribution. The smaller sources in the free troposphere, however, cannot be trivialized since they are localized in a region of the atmosphere where NO_x lifetimes are maximized. Thus, their proportional impact is greater than their absolute source strengths would imply.

Fossil Fuel Combustion

NO is formed by high-temperature chemical processes during combustion of fossil fuels, both from nitrogen present in fuel and from the oxidation of atmospheric N_2 in the presence of O_2. The distribution for this source is heavily weighted toward the Northern Hemisphere where most of the industrialized world resides. Detailed inventories are available for Canada, the United States, and western Europe describing the spatial patterns of NO_x emissions from fossil fuel combustion and industrial processes [Wagner et al., 1986; Environmental Protection Agency (EPA), 1986;

TABLE 1 Sources of Tropospheric NO_x

Source	Estimated Magnitude and Uncertainty (Tg N/yr)	Principle Location of Emissions
Fossil fuels	22 (13–31)	Midlatitude continental surface (30–60°N)
Biomass burning	7.9 (3–15)	Tropical continental surface
Soil emissions	7.0 (4–12)	Nonpolar continental surface
Lightning	5.0 (2–20)	Tropical/subtropical continental troposphere
Aircraft	0.56 (0.45–1)	NH upper troposphere (30–60°N, 8–13 km)
Strat–Trop exchange	0.64 (0.4–1)	Midlatitude tropopause
Oceans	0.5 (0–1)	Tropical/subtropical upwelling regions
NH_3 oxidation	0.6 (0.3–3)	Free troposphere over industrial regions

Source: Adapted from Lee et al. (1997) and Bradshaw et al. (2000).

Lübkert and Zierock, 1989]. Almost one-half (44%) of the NO_x emissions in the United States are from transportation, 33% from power plants, and 16% from industrial combustion. Similar inventories have been reported for western Europe (Lübkert and DeTilly, 1989) and Asia (Akimoto and Narstu, 1994). Based on 1985 data, approximately 84% of total emissions are accounted for by emissions from North America (28%), Europe (31%), and Asia (31%).

NO_x emissions due to maritime shipping have been estimated to contribute as much as 3 Tg N/yr to the global NO_x budget (Corbett et al., 1999) with half of these emissions occurring in the North Atlantic. While small compared to the overall fossil fuel contribution of 22 Tg N/yr, NO_x emissions from seagoing vessels could prove important over the open ocean in and around shipping lanes far removed from major continental sources (Lawrence and Crutzen, 1999).

Biomass Burning

Biomass burning in tropical and subtropical regions is a significant source of NO_x as well as other chemically and radiatively important species such as CO_2, CO, CH_4, NMHC, N_2O, and aerosols. Recent estimates of NO_x emissions are based on emission factors from laboratory as well as field measurements of burning vegetation coupled with biomass inventories, land-use data, estimates of tropical deforestation, and occurrence of wild fires. Burning in the tropical latitudes is estimated to account for approximately 87% of the global total with Africa, South America, and Asia accounting for approximately 42, 23, and 28%, respectively. Thus far, however, laboratory mass-balance experiments can only account for approximately 30 to 50% of the fuel nitrogen that is released from this burning. Much of the missing nitrogen is thought to be in the form of molecular nitrogen with the remainder possibly representing mineralized ash (10%) and high-molecular-weight compounds containing substantial amounts of nitrogen (Lobert et al., 1991; Yokelson et al., 1996). It is unclear whether these latter compounds might act as a source of NO_y or NO_x to the remote troposphere. Of the known fixed nitrogen compounds, NO_x is the dominant species (54%), having an estimated emission factor of approximately 2.1 g N/kg C (carbon fuel). This is significantly smaller than the emission factors reported for higher temperature fossil fuel combustion sources. Reduced nitrogen compounds such as NH_3 (emission factors of approximately 1.3 g N/kg C) comprise the remaining 46%.

Soil Emissions

Observations have shown that the dominant reactive odd-nitrogen emission from soils to the atmosphere is NO, with lesser emissions of NO_2 and HONO. Studies indicate that a wide range of factors influence the net soil emission of NO_x to the atmosphere. These include climate (through temperature and rainfall), plant growth and decay, the clearing of forests, biomass burning, and fertilization. The three most important variables influencing NO_x emissions are soil temperature, soil moisture content, and soil vegetation cover. NO emission rates have been found to vary almost exponentially with soil temperature, whereas a more linear relationship has been observed with respect to soil nitrate levels. The dependence on soil moisture content appears to be a complex one. Below approximately 15% soil moisture, microbial activity has been found to be primarily water limited and strongly favors nitrification, whereas at higher moisture contents denitrification eventually becomes predominant and NO emissions decrease rapidly. Order of magnitude differences in emission rates also occur between heavily fertilized soils, grasslands, and forested ecosystems (Williams et al., 1992). Large increases in the rates of soil emission have been observed after rain events following long periods of drought and in areas where biomass burning had recently occurred (Neff et al., 1995). Canopy cover has also been shown to be a key factor controlling the net flux of NO_x into the atmosphere, particularly, tropical rain forest canopies, which have been shown to be an effective sink for NO_2 (Jacob and Wofsy, 1990). Agriculture and grass lands account for the bulk of net emissions (41 and 35%, respectively) (Yienger and Levy, 1995). Future changes in soil emissions are expected to be linked to increased use of nitrogen fertilizers and agricultural production.

Lightning

Of all NO_x sources, improving our understanding of lightning is most critical since it has one of the largest uncertainties and represents what appears to be the dominant source of NO_x in the free troposphere. Lightning NO is generated by recombination reactions that occur as this 20,000 K or hotter, 10-MPa pressurized plasma supersonically expands and cools. Quantifying NO_x production from such events is still quite problematic.

Much of the current debate stems directly from estimating the amount of NO_x produced by a "typical" lightning flash. This process has involved combining estimates of the average energy deposited per lightning flash with evaluations of the NO_x produced per joule of energy released and the average global frequency of lightning. Using this approach, the global estimate of 2 Tg N/yr by Kumar et al. (1995) lies at the low end of a clustering of similarly derived estimates based on very low yields of NO_x per lightning flash ($\sim 3.6 \times 10^{25}$ NO_x molecules/flash; e.g., Lawrence et al. (1995) and references therein). At the high end, Liaw et al. (1990) have argued for a source strength as large as 200 Tg N/yr based on a correspondingly larger value for NO_x yield per lightning flash, but this estimate appears unreasonably high based on nitrate deposition records. Many of these treatments have relied on the application of scaling or normalization factors to other investigators' results that may not be valid, and frequently the various forms of lightning have been treated as if they were one type only.

Combining more refined values for the production of NO_x per joule of energy (Goldenbaum and Dickerson, 1993) with updated lightning flash energy values results in a per-flash NO_x yield for negative CG (cloud to ground) lightning of 1 to 2×10^{26} NO/flash. By contrast, for positive CG and IC (intracloud) lightning, the yield is approximately 5×10^{26} and 0.5×10^{26} NO/flash. Furthermore, current evaluations of the global distribution of lightning (Goodman et al., 1988; Christian et al., 1992) in combination with estimated spatial distributions for different types of lightning (Orville, 1994) results in a nominal 100 global flashes/s being proportioned 75% to IC lightning and 25% to CG lightning. Seventy percent of CG lightning is associated with the tropics/subtropics, with 5% positive strokes, and 30% is associated with higher latitudes having 30% positive strokes. Combining these estimates with the midrange value for NO_x production per flash for each lightning type results in a conservative estimate for total NO_x production of 2.5 Tg N/yr by IC lightning, 3 Tg N/yr by negative CG lightning, and 1 Tg N/yr by positive CG lightning.

These global lightning estimates are found to be in generally good agreement with other independent NO_x assessments not dependent on the mechanistic details of lightning. For example, Albritton et al. (1984) constrained the global lightning source strength by examining nitrate deposition records from remote global areas that were expected to be free of impacts from anthropogenic sources. They estimated a lightning source of approximately 8 Tg N/yr (range 2 to 20). More recently, Levy et al. (1996) have used a global chemical transport model in conjunction with remote, upper tropospheric NO measurements to constrain all lightning sources. Their results, which critically depend on their choice of parameterizations for deep convection, indicate a range of values for NO_x production from lightning of 2 to 6 Tg N/yr.

A final issue of importance concerning lightning emissions relates to the altitude distribution of emissions. This is influenced not only by the initial production from lightning but also by subsequent vertical mixing in convective storms. Pickering et al. (1998) have recently produced estimates for the vertical distribution of lightning emissions. Their results show that most lightning NO_x is delivered to the upper troposphere; however, the more vigorous mixing of midlatitude continental storms leads to more downward transport of lightning NO_x than for maritime storms or tropical continental storms. They also showed that the peak NO_x in continental storms occurs at higher altitudes than for maritime storms.

Aircraft Emissions

Subsonic aircraft emissions represent the most quantitatively known direct source of NO_x in the upper troposphere. Aircraft engines use an extremely efficient, high temperature combustion process that primarily produces CO_2, H_2O, and a few percent of other compounds. Like lightning, the initial NO_y content of these emissions consists primarily (>85%) of NO. Estimates of NO_x production are derived from assessments of the emission indices for various engines under different flight conditions and the annual amount of air traffic in terms of the kilograms of fuel consumed. The total strength of this source for both scheduled commercial and nonscheduled (e.g., military and chartered) air traffic in 1992 has been estimated at approximately 0.46 Tg N/yr for all altitudes with approximately 65% of the emissions occurring in the upper troposphere (>8 km), and, of that, approximately 45% has been assessed as occurring between 20° and 45°N with another 36% at latitudes further north (i.e., 55% of total emissions were at altitudes >8 km and latitudes > 20°N). This source of NO_x has also been projected to increase to approximately 1.3 Tg N/yr by the year 2015 (NASA, 1991). Because about 90% of the emissions occur in the free troposphere over the Northern Hemisphere, the impact from this source on the budgets and distributions of NO_x and ozone should be substantially different than in the Southern Hemisphere. In addition, this source may have a larger relative impact on upper tropospheric, winter time Northern Hemispheric NO_x distributions. This reflects the fact that during this period lightning and convection of surface NO_x emissions are both significantly reduced.

Stratosphere–Troposphere Exchange

The source of NO_y in the stratosphere is primarily N_2O oxidation by $O(^1D)$ in the upper stratosphere. Most of the mass that is transported into the troposphere, however, comes from the lower stratosphere. Peak levels of activity for this source occur primarily in the spring. This activity is found to be most vigorous near the subtropical and polar jet streams as well as in mid and high-latitude regions affected by atmospheric overturning associated with largc-scalc low-pressure disturbances and frontal systems. Even though the NO_x flux estimates are small, transported stratospheric odd nitrogen may still be a significant source of free tropospheric NO_x for some remote regions. For example, depending on how much of the NO_x production from lightning is transported to the stratosphere, estimates of the average global flux of NO_y from the stratosphere range from approximately 0.3 to 1 Tg N/yr (Ko et al., 1986; Murphy and Fahey, 1994). These values are close to balancing the stratospheric production of NO_x from N_2O (Kasibhatla et al., 1991).

Although the stratospheric NO_y source seems small compared to boundary layer NO_y sources, it is comparable in magnitude to other free tropospheric sources such as emissions from subsonic aircraft. Unlike subsonic aircraft emissions and lightning, which predominantly release NO directly into the free troposphere, input of NO_y from the middle stratosphere should primarily consist of HNO_3 with only a small, approximately 25% or less (i.e., 0.25 Tg N/yr), contribution from NO_x (Russell et al., 1988; Notholt et al., 1995). However, because removal of NO_y in the upper troposphere is extremely inefficient compared to that in the lower atmosphere, the NO_y from the stratosphere may have a long enough lifetime to impact the distribution of NO_x through recycling reactions, thereby influencing ozone production in the free troposphere.

Oceans

While this source is expected to be small, it could have greater importance for specific regions given its location far from the more dominant continental sources. Zafirou and McFarland (1981) conducted measurements in the equatorial Pacific and found that nitrite photolysis may provide a source of NO from the ocean. Based on air–sea exchange models and the difference between $[NO]_{ocean}$ and $[NO]_{air}$, several investigators derived a source strength of about 0.5 Tg N/yr (Logan, 1983; Liu et al., 1983; Torres and Thompson, 1993). This value is highly uncertain, however, since the data from the equatorial Pacific represent a small sample from a more biologically productive region of the ocean.

Ammonia Oxidation

A poorly quantified, but still potentially important source of NO_x is the oxidation of atmospheric ammonia (NH_3). A major uncertainty regarding this source is the lack of information on the tropospheric distribution of NH_3. NH_3 is initially oxidized by OH to form NH_2. NH_2 may go on to form NO_x through reaction with O_3, however, it may also react with NO_x to form N_2 or N_2O. Based on differences in rate coefficients, NH_3 oxidation should provide a net source of NO_x when ambient NO_x is less than 200 to 500 ppt, a condition that is prevalent in the remote troposphere. Recently, boundary layer NH_3 mixing ratios in the 50 to 900 ppt range (median 250 pptv) have been found over large stretches of the South Pacific and the Southern Ocean (J. Bradshaw, unpublished data). A reasonable estimate for this source is about 0.6 Tg N/yr, assuming a background tropospheric NH_3 value of 150 pptv and a 4-month lifetime for oxidation via OH.

4 TROPOSPHERIC DISTRIBUTION OF REACTIVE NITROGEN

Knowledge concerning the tropospheric distribution of NO_x is critical given its importance to ozone photochemistry. Over much of the remote atmosphere, NO concentrations hover near the critical level necessary for net photochemical production of ozone. This critical NO level varies from as low as about 5 pptv to near 20 pptv depending on ambient conditions (Crawford et al., 1997). As a general rule, observations have shown NO to typically fall below critical levels over remote marine boundary layer environments away from NO_x sources where the lowest ozone values in the atmosphere also occur. By contrast, ozone production in the upper troposphere appears to be ubiquitous given observations of NO consistently above the critical level.

Despite the pivotal role NO_x plays in tropospheric photochemistry, current knowledge of its tropospheric distribution is based on very limited data, especially for remote regions. Reliable methods for measuring NO over the full range of its tropospheric variability (a few pptv to tens of ppbv) have been available since the late 1970s. Even so, field measurements of NO_x from ground, ship, and aircraft platforms provide only limited spatial and temporal coverage. Nevertheless, these observations do reveal some basic features in the global distribution of NO_x (Emmons et al., 1997; Bradshaw et al., 2000). For instance, gradients in NO_x observations are greatest near the surface. NO_x in urban areas is typically in the parts-per-billion range and a few hundred parts-per-trillion are common even in rural areas. For remote oceanic regions, however, NO_x levels are generally less than 50 ppt and often only a few parts per trillion. These trends in NO_x at the surface are consistent with most NO_x sources being land based and the short atmospheric lifetime for NO_x of a day or less. In the upper troposphere, gradients in NO_x are weaker owing to both the longer lifetime for NO_x and generally faster transport. NO_x values in the range of 50 to 200 pptv are often observed; however, observations ranging from only a few pptv to more than a ppbv of NO_x are not uncommon. These extremes are most likely due to the convection of NO_x-poor air in marine environments contrasted by the convection of NO_x-rich polluted air with additional inputs from lightning in continental regions. As a consequence of the difference in NO_x gradients at the surface and high altitude, NO_x is generally observed to increase with altitude over remote locations.

Some of these trends can be seen in data collected from NASA's DC-8 aircraft (see Fig. 3). NO_x in this figure has been estimated from daytime measurements of NO (solar zenith angle $< 70°$) by assuming photochemical equilibrium conditions for NO_2. These data have been taken from the following field campaigns: Pacific Exploratory Mission (PEM)–West A (1991) (Hoell et al., 1996), Transport and Atmospheric Chemistry Near the Equator–Atlantic (TRACE-A, 1992) (Fishman et al., 1996), PEM–Tropics A (1997) (Hoell et al., 1999), and Subsonics Assessment Ozone and Nitrogen Oxides Experiment (SONEX, 1998) (Singh et al., 1999). These campaigns have focused on taking measurements to characterize the remote oceanic troposphere with an emphasis on the upper troposphere. While flown in different years, each campaign was conducted during the fall season (September–November).

Figure 3 shows data for the boundary layer (0 to 1 km), the lower free troposphere (1 to 6 km), and the upper troposphere (6 to 12 km). In general, NO_x over the Atlantic is greater than over the Pacific at all altitudes. This is due to a closer proximity to NO_x sources, which are predominantly land based. The seasonal nature of NO_x sources is also important to the elevated levels of NO_x over the South Atlantic since measurements were taken during the biomass burning season

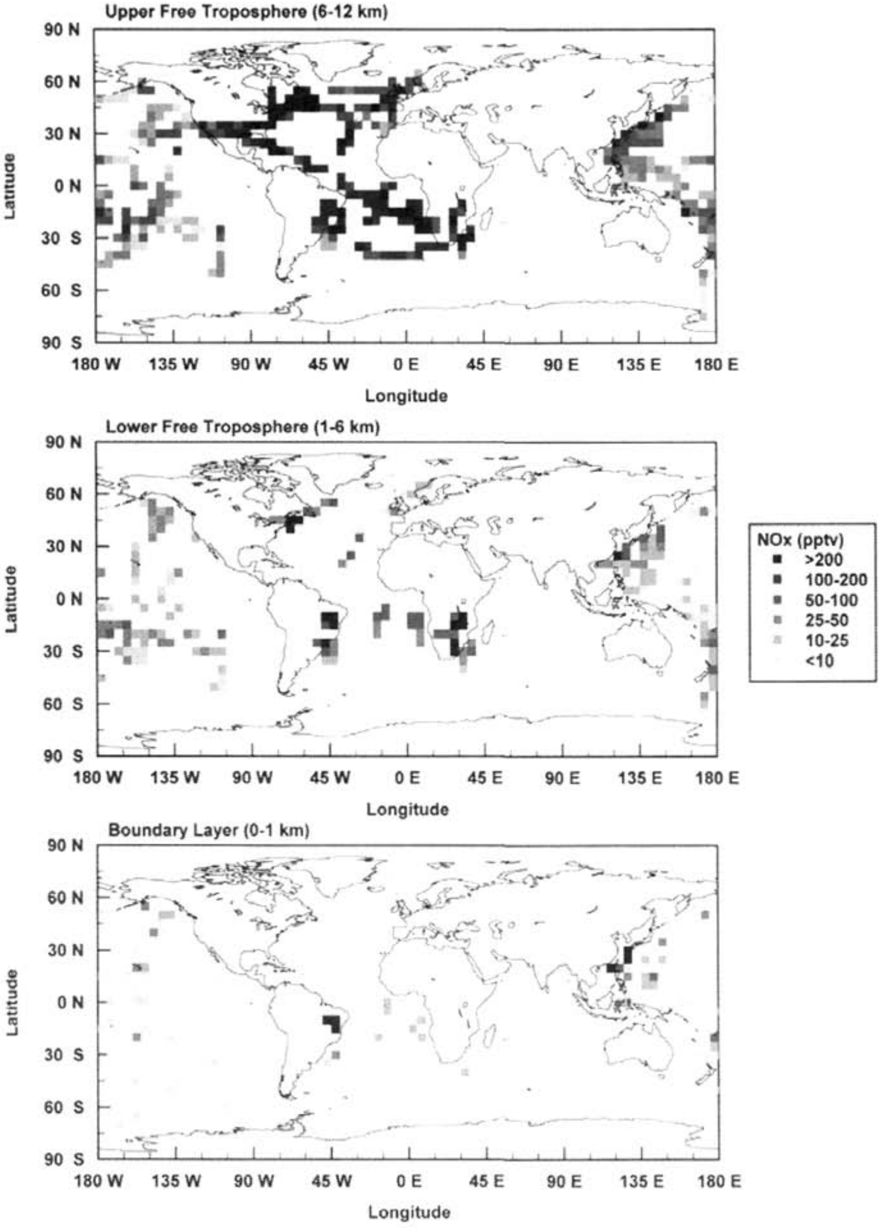


Figure 3 (see color insert) Distribution of NO_x based on measurements taken from NASA's DC-8 aircraft during fall (see text for details). Data are averaged on a 5° × 5° latitude–longitude grid for three altitude ranges. See ftp site for color image.

for South America and Africa. While boundary layer data over and near continental areas are sparse, the strong gradient in surface NO_x is evident in the low values (typically <10 pptv) over the South Pacific. Gradients are weaker at higher altitudes. The increase in NO_x with altitude over remote oceanic regions is also evident for both the South Atlantic and South Pacific data. The decrease in NO_x with altitude over continental areas is less evident since data is sparse, but NO_x values over South America, Southern Africa, and the South China coast do exhibit a decrease with altitude.

Information concerning the distribution of NO_y is even more limited than that for NO_x. The only NO_y species other than NO_x that have been measured with any regularity are HNO_3 and PAN. Total NO_y measurements have been more common, but there are still questions as to what these measurements represent since they often exceed expected values based on the sum of all NO_y constituents (Crosley, 1996). Measurements of HNO_3 in the remote troposphere have consistently fallen well below levels expected based on theory (Liu et al., 1992; Ridley et al., 1998; Schultz et al., 2000). This problem has been particularly troubling since theory predicts HNO_3 to be the dominant NO_y species in the remote troposphere. Heterogeneous mechanisms recycling HNO_3 to NO_x have been hypothesized as a potential solution (Fan et al., 1994; Chatfield, 1994; Hauglustaine et al., 1996; Lary et al., 1997). Possible underestimations in wet removal and partitioning of HNO_3 between gas and aerosol phases have been cited as well (Liu et al., 1992; Wang et al., 1998).

Measurements of PAN support the contention that it plays a strong role in sustaining NO_x in air masses as they are transported away from regions of strong industrial or biomass burning emissions. Decomposition of PAN was found to be adequate to explain NO_x observations at low altitude over eastern Canada (Fan et al., 1994), the South Atlantic (Jacob et al., 1996), and the western, North Pacific (Crawford et al., 1997). For even more remote regions, the decomposition of PAN in descending air masses can also be responsible for sustaining NO_x. This condition was observed by Schultz et al. (1999) over the remote South Pacific.

While global models can be used to estimate the distributions of NO_x and NO_y, the accuracy of these estimates is still very uncertain and are complicated by several factors. First is the level of uncertainty that remains for various NO_x sources, especially natural source strengths as well as spatial distributions. Second, there are still major questions concerning the recycling of NO_x from NO_y reservoir species and the potential role of aerosol in both the removal and recycling of NO_x. Finally, the wide range of photochemical lifetimes for NO_x requires atmospheric models to accurately represent small-scale transport processes (e.g., convective vertical transport of NO_x and wet deposition of HNO_3). The scales for these processes remain significantly smaller than the resolution of current photochemical transport models.

REFERENCES

Akimoto, H., and H. Narstu, Distributions of SO_2, NO_x, and CO_2 emissions from fuel combustion and industrial activities in Asia with 1° x 1° resolution, *Atmos. Environ.*, *28*, 213–225, 1994.

Albritton, D. L., S. C. Liu, and D. Kley, Global nitrate deposition from lightning, in *Proceedings of the Conference on the Environmental Impact of Natural Emissions*, Air Pollution Control Association, Pittsburgh, PA, 1984, pp. 100–112.

Bradshaw, J., et al., Photofragmentation two-photon laser-induced fluorescence detection of NO_2 and NO: Comparison of measurements with model results based on airborne observations during PEM–Tropics A, *Geophys. Res. Lett.*, *26*, 471–474, 1999.

Bradshaw, J., D. Davis, G. Grodzinsky, S. Smyth, R. Newell, S. Sandholm, and S. Liu, Observed distributions of nitrogen oxides in the remote free troposphere from the NASA Global Tropospheric Experiment programs, *Rev. Geophys.*, *38*, 61–116, 2000.

Chatfield, R. B., Anomalous HNO_3/NO_x ratio of remote troposphere air: Conversion of nitric acid to formic acid and NO_x, *Geophys. Res. Lett.*, *21*, 2705–2708, 1994.

Christian, H. J., R. J. Blakeslee, and S. J. Goodman, Lightning imaging sensor (LIS) for the Earth Observing System, NASA Technical Memorandum 4350, Huntsville, AL, February 1992.

Corbett, J. J., P. S. Fischbeck, and S. N. Pandis, Global nitrogen and sulfur inventories for oceangoing ships, *J. Geophys. Res.*, *104*, 3457–3470, 1999.

Crawford, J., et al., An assessment of ozone photochemistry in the extratropical western North Pacific: Impact of continental outflow during the late winter/early spring, *J. Geophys. Res.*, *102*, 28469–28487, 1997.

Crosley, D. R., NO_y blue ribbon panel, *J. Geophys. Res.*, *101*, 2049–2052, 1996.

Emmons, L. K., et al., Climatologies of NO_x and NO_y: A comparison of data and models, *Atmos. Environ.*, *31*, 1851–1904, 1997.

Environmental Protection Agency (EPA), *Development of the 1980 NAPAP Emissions Inventory*, EPA/600/4-85-038, U.S. EPA, Research Triangle Park, NC, 1986, Chapter 4.

Fan, S. M., et al., Origin of tropospheric NO_x over subarctic eastern Canada in summer, *J. Geophys. Res.*, *99*, 16867–16877, 1994.

Fishman, J., J. M. Hoell, Jr., R. D. Bendura, R. J. McNeal, and V. W. J. H. Kirchoff, NASA GTE TRACE A experiment (September–October 1992): Overview, *J. Geophys. Res.*, *101*, 23865–23879, 1996.

Goldenbaum, G. C., and R. R. Dickerson, Nitric oxide production by lightning discharges, *J. Geophys. Res.*, *98*, 18333–18338, 1993.

Goodman, S. J., H. J. Christian, and W. D. Rust, A comparison of the optical pulse characteristics of intracloud and cloud-to-ground lightning as observed above clouds, *J. Appl. Meteorol.*, *27*, 1369–1381, 1988.

Hauglustaine, D. A., B. A. Ridley, S. Solomon, P. G. Hess, and S. Madronich, HNO_3/NO_x ratio in the remote troposphere during MLOPEX 2: Evidence for nitric acid reduction on carbonaceous aerosols? *Geophys. Res. Lett.*, *23*, 2609–2612, 1996.

Hoell, J. M., D. D. Davis, D. J. Jacob, M. O. Rodgers, R. E. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM–Tropics A, August–September 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hoell, J. M., D. D. Davis, S. C. Liu, R. Newell, M. Shipham, H. Akimoto, R. J. McNeal, R. J. Bendura, and J. W. Drewry, Pacific Exploratory Mission—West A (PEM—West A): September–October 1991, *J. Geophys. Res.*, *101*, 1641–1653, 1996.

Jacob, D. J., and S. C. Wofsy, Budgets of reactive nitrogen, hydrocarbons, and ozone over the Amazon forest during the wet season, *J. Geophys. Res.*, *95*, 16737–16754, 1990.

Jacob, D. J., et al., Origin of ozone and NO_x in the tropical troposphere: Photochemical analysis of aircraft observations over the South Atlantic Basin, *J. Geophys. Res.*, *101*, 24235–24250, 1996.

Kasibhatla, P. S., H. Levy II, W. J. Moxim, and W. L. Chameides, The relative impact of stratospheric photochemical production on tropospheric NO_y levels: A model study, *J. Geophys. Res.*, *96*, 18631–18646, 1991.

Ko, M. K. W., M. B. McElroy, D. K. Weisenstein, and N. D. Sze, Lightning: A possible source of stratospheric odd nitrogen, *J. Geophys. Res.*, *91*, 5395–5405, 1986.

Kumar, P. P., G. K. Manohar, and S. S. Kandalgaonkar, Global distribution of nitric oxide produced by lightning and its seasonal variation, *J. Geophys. Res.*, *100*, 11203–11208, 1995.

Lary, D. J., A. M. Lee, R. Toumi, M. J. Newchurch, M. Pirre, and J. B. Renard, Carbon aerosols and atmospheric photochemistry, *J. Geophys. Res.*, *102*, 3671–3682, 1997.

Lawrence, M. G., W. L. Chameides, P. S. Kasibhatla, H. Levy II, and W. Moxim, Lightning and atmospheric chemistry: The rate of atmospheric NO production, in H. Volland (Ed.), *Handbook of Atmospheric Electrodynamics*, Vol. 1, CRC Press, Boca Raton, FL, 1995, pp. 189–202.

Lawrence, M. G., and P. J. Crutzen, Influence of NO_x emissions from ships on tropospheric photochemistry and climate, *Nature*, *402*, 167–170, 1999.

Lee, D. S., et al., Estimations of global NO_x emissions and their uncertainties, *Atmos. Environ.*, *31*, 1735–1749, 1997.

Levy, II, H., W. J. Moxim, and P. S. Kasibhatla, Global 3-dimensional time-dependent lightning source of tropospheric NO_x, *J. Geophys. Res.*, *101*, 22911–22922, 1996.

Liaw, Y. P., D. L. Sisterson, and N. L. Miller, Comparison of field, laboratory, and theoretical estimates of global nitrogen fixation by lightning, *J. Geophys. Res.*, *95*, 22489–22494, 1990.

Liu, S. C., et al., A study of the photochemical and ozone budget during the Mauna Loa Observatory Photochemistry Experiment, *J. Geophys. Res.*, *97*, 10463–10471, 1992.

Liu, S. C., M. McFarland, D. Kley, O. Zafirou, and B. Huebert, Tropospheric NO_x and O_3 budgets in the equatorial Pacific, *J. Geophys. Res.*, *88*, 1360–1368, 1983.

Lobert, J. M., et al., Experimental evaluation of biomass burning emissions: Nitrogen and carbon containing compounds, in J. S. Levine (Ed.), *Global Biomass Burning: Atmospheric Climate and Biospheric Implications*, MIT Press, Cambridge, MA, 1991.

Logan, J. A., Nitrogen oxides in the troposphere: Global and regional budgets, *J. Geophys. Res.*, *88*, 10785–10807, 1983.

Lübkert, B., and S. DeTilly, The OECD—Map emission inventory for SO_2, NO_x and VOC in western Europe, *Atmos. Environ.*, *23*, 3–15, 1989.

Lübkert, B., and K. H. Zierock, European emission inventories—A proposal of international worksharing, *Atmos. Environ.*, *23*, 37–48, 1989.

Murphy, D. M., and D. W. Fahey, An estimate of the flux of stratospheric reactive nitrogen and ozone into the troposphere, *J. Geophys. Res.*, *99*, 5325–5332, 1994.

NASA, High speed research program/atmospheric effects of stratospheric aircraft (HSRP/AESA), Annual Report, 1991.

Neff, J. C., M. Keller, E. A. Holland, A. W. Weitz, and E. Veldkamp, Fluxes of nitric oxide from soils following the clearing and burning of a secondary tropical rain forest, *J. Geophys. Res.*, *100*, 25913–25922, 1995.

Notholt, J., A. Meier, and S. Peil, Total column densities of tropospheric and stratospheric trace gases in the undisturbed arctic summer atmosphere, *J. Atmos. Chem.*, *20*, 311–332, 1995.

Orville, R. E., Cloud-to-ground lightning flash characteristics in the contiguous United States: 1989–1991, *J. Geophys. Res.*, *99*, 10833–10841, 1994.

Pickering, K. E., Y. Wang, W.-K. Tao, C. Price, and J.-F. Müller, Vertical distributions of lightning NO_x for use in regional and global chemical transport models, *J. Geophys. Res.*, *103*, 31203–31216, 1998.

Ridley, B., et al., Measurements of NO_x and PAN and estimates of O_3 production over the seasons during Mauna Loa Observatory Photochemistry Experiment 2, *J. Geophys. Res.*, *103*, 8323–8339, 1998.

Russell III, J. M., et al., Measurements of odd nitrogen compounds in the stratosphere by the ATMOS experiment on Spacelab 3, *J. Geophys. Res.*, *93*, 1718–1736, 1988.

Schultz, M. G., et al., On the origin of tropospheric ozone and NO_x over the tropical South Pacific, *J. Geophys. Res.*, *104*, 5829–5844, 1999.

Schultz, M. G., D. J. Jacob, J. D. Bradshaw, S. T. Sandholm, J. E. Dibb, R. W. Talbot, and H. B. Singh, Chemical NO_x budget in the upper troposphere over the tropical South Pacific, *J. Geophys. Res.*, *105*, 6669–6679, 2000.

Singh, H. B., A. M. Thompson, and H. Schlager, SONEX airborne mission and coordinated POLINAT-2 activity: Overview and accomplishments, *Geophys. Res. Lett.*, *26*, 3053–3056, 1999.

Torres, A. L., and A. M. Thompson, Nitric oxide in the equatorial Pacific boundary layer: SAGA 3 measurements, *J. Geophys. Res.*, *98*, 16949–16954, 1993.

Wagner, J., R. A. Walters, L. J. Maiocco, and D. R. Neal, *Development of the 1980 NAPAP Emissions Inventory*, U.S. Environmental Protection Agency, Washington, DC, 1986.

Wang, Y., J. A. Logan, and D. J. Jacob, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry 2. Model evaluation and global ozone budget, *J. Geophys. Res.*, *103*, 10727–10755, 1998.

Williams, E. J., et al., An intercomparison of five ammonia measurement techniques, *J. Geophys. Res.*, *97*, 11591–11611, 1992.

Yienger, J. J., and H. Levy II, Empirical model of global soil-biogenic NO_x emissions, *J. Geophys. Res.*, *100*, 11447–11464, 1995.

Yokelson, R. J., D. W. T. Griffith, and D. E. Ward, Open-path Fourier transform infrared studies of large-scale laboratory biomass fires, *J. Geophys. Res.*, *101*, 21067–21080, 1996.

Zafirou, O. C., and M. McFarland, Nitric oxide from nitrite photolysis in the central equatorial Pacific, *J. Geophys. Res.*, *86*, 3173–3182, 1981.

CHAPTER 5

CARBON MONOXIDE IN THE ATMOSPHERE

PAUL NOVELLI

Carbon monoxide (CO) is present in trace quantities in the atmosphere. Although first detected in the late 1940s using solar spectroscopic methods,[1,2] few measurements of CO were made during the period between the early 1950s and the mid-1960s. However, as chromatographic and related detection techniques were developed, discrete measurements of CO were made in many locations around the world. These provided considerable insight on global tropospheric distributions; most notable among these was the observation that CO concentrations generally decreased from north to south.[3–5] The significance of CO in atmospheric chemistry was recognized in 1971 when Levy,[6] and McConnell et al.[7] proposed a photochemically driven, radical chain reaction linking the tropospheric cycles of methane (CH_4), CO, nitric oxide and nitrogen dioxide (NO_x), and formaldehyde (CH_2O), with those of the oxidants ozone (O_3), the hydroxyl (OH) and hydroperoxyl (HO_2) radicals. These models describe an atmosphere in which the photolysis of O_3 ($h\nu < 320\,\text{nm}$) leads to the formation of OH, initiating a series of oxidation/reduction reactions that both produce and destroy CO, CH_2O, OH and HO_2.

In much of the background atmosphere the reaction of CO and OH [Eq. (1)] accounts for 90 to 95% of the loss of CO[8] and about 75% of the removal of OH.[9]

$$CO + OH + O_2 \rightarrow CO_2 + HO_2 \qquad (1)$$

While the stoichiometric relationship between CO oxidation and OH loss is dependent upon several possible reaction pathways, the inverse relationship between CO and OH concentrations suggested by Eq. (1) is expected in the background atmosphere.[9,10] Not only does the hydroxyl radical regulate the concentration of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

CO, but oxidation at the expense of OH is also the primary removal pathway for many other reduced gases, several of which are radiatively important [e.g., CH_4, the hydrogenated chlorofluorocarbons (CFCs)]. Therefore, trends in atmospheric CO levels are expected to have an effect on climate through its role in regulating [OH], which in turn affects the levels of several important greenhouse gases.[11]

Carbon monoxide impacts both local and regional air quality through its influence on ozone. In areas of relatively high NO_x levels (>5 to 10 pmol/mol), such as urban areas or air parcels affected by fossil fuel or biomass burning, HO_2 produced through CO oxidation enters into a series of photochemical reactions that produce O_3:

$$NO + HO_2 \rightarrow NO_2 + OH \tag{2}$$

$$NO_2 + h\nu \rightarrow NO + O \tag{3}$$

$$O + O_2 \rightarrow O_3 \tag{4}$$

In the background atmosphere, where $[NO_x]$ is often < 5 pmol/mol, HO_2 produced by the oxidation of CO may destroy O_3.

$$O_3 + HO_2 \rightarrow 2O_2 + OH \tag{5}$$

As a result, the oxidizing capacity of the lower atmosphere is coupled to the concentrations, distributions, and trends of CO.

1 MEASUREMENT TECHNIQUES

Analytical Methods

Measurements of atmospheric CO are conducted using a variety of techniques. Solar spectra recorded at 4.7 μm are used to derive total column abundances and column-averaged mixing ratios.[12,13] Nondispersive infrared radiometry (NDIR)[14,15] and tunable diode laser spectroscopy (TDLS)[16] also make use of CO absorption at 4.7 μm. Both techniques provide a continuous measurement of CO; however, the TDLS provides greater precision (1 ppb) and a higher measurement frequency (10 s), about a factor of 10 greater than NDIR. Gas chromatography (GC), when coupled with a number of different detectors can provide high precision and low detection limits.[17–19] The most common detectors used with GC are flame ionization (with prior conversion of $CO + H_2 \rightarrow CH_4$), electron capture, and hot mercuric oxide reduction. GC techniques can provide a high precision (1 ppb) with a discontinuous measure of CO (frequency on the order of a few minutes).[20]

Calibration

The gas chromatographic methods, NDIR, and some TDLS techniques require calibration against samples with known gas amounts. As far as we are aware, there is no national or commercial laboratory that provides certified CO reference gases at levels found in the background atmosphere. Groups measuring CO must therefore dilute high concentration certified gases to atmospheric levels or obtain standards from other laboratories. Laboratory intercomparisons of the reference gases used by various researchers have shown large differences between groups (up to 25 to 50%).[21–23] CO standards may also be subject to drift over time. Efforts have been made since the early 1990s to compare CO measurements in both the laboratory and the field.[14,21] However, differences between groups still exist, and the integration of data sets requires some prior understanding of how the measurements compare.

2 GLOBAL CO DISTRIBUTIONS

Surface CO

Background Atmosphere. CO varies both temporally and spatially. Figure 1 presents a smoothed representation of the surface distribution of CO in the background marine boundary layer (MBL) as a function of latitude and time. The surface illustrates that CO mixing ratios in both hemispheres exhibit seasonal variation, and

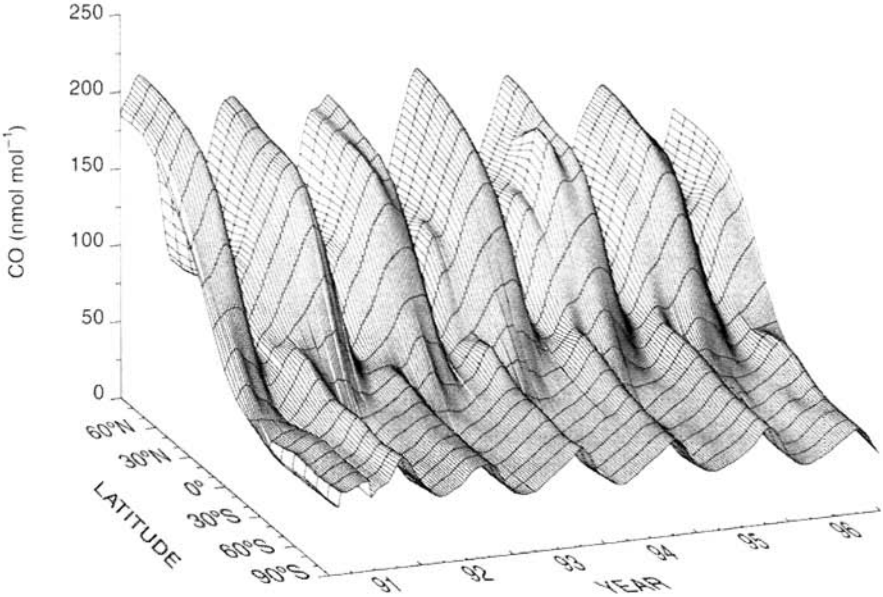


Figure 1 Smooth surface representing the distribution of CO in the marine boundary layer. The surface was created from 38 time series determined from sampling locations in the NOAA/CMDL Cooperative Air Sampling Program. CO mixing ratios were combined in 5° latitude bands, longitudinal differences were averaged, and the combined time series were smoothed in both time and space.[24]

although there are considerable interannual variations, repeatable pattems occur from year to year. Most notable is the seasonal cycle and the interhemispheric gradient. Greatest CO mole fractions in the MBL [200 to 225 nmol CO/mol air (ppb)] are found in the high latitudes of the Northern Hemisphere during late winter/early spring. The high Northern Hemisphere also exhibits the greatest seasonal amplitude (120 to 140 ppb). The imbalance of sources in winter (mostly anthropogenic pollution from the midlatitudes) and sink (when OH levels are lowest) leads to an accumulation of CO in the high Northern Latitudes. Lowest CO mixing ratios in the boundary layer (40 to 50 ppb) are found during the southern summer, where low concentrations are further depressed by reaction with OH. The interhemispheric gradient also exhibits a strong seasonality. The largest difference between the high northern and high southern hemispheres (~150 ppb) occurs in February/March and the minimum difference (10 to 20 ppb) occurs in September/October.[17,18,24]

Polluted Atmosphere. CO levels in urban locations and areas of regional-scale pollution are greater than those found in the background atmosphere, with CO mixing ratios in urban areas often reaching ppm level, orders of magnitude greater than those found in the background troposphere. CO is defined as a criteria species for urban pollution. The lifetime of CO is on the order of several months, and emissions can be transported far from the original source region.[25] Even in areas far distant from CO sources, wide-scale, diffuse pollution may enhance CO levels (up to twice background levels). Air parcels downwind of areas where combustion occurs can also show elevated levels of O_3.[26,27] The enhanced O_3 often reflects its photochemical production [Eqs. (2)–(4)], which is favored in environments having both high CO and NO_x.[28]

Free Troposphere

Near the planet's surface, CO varies with season, proximity to source regions, and latitude. Above the boundary layer, and in the middle and upper troposphere, CO also shows seasonal cycles and spatial distributions. At altitudes higher than a few kilometers, mixing ratios are largely determined by surface source distributions and by vertical and horizontal transport.[29,30] Mixing ratios determined in the free troposphere at mountain observatories in the Northern Hemisphere are typically lower than measurements made at sea level at similar times and latitudes.[18] CO mixing ratios in the free troposphere, studied from aircraft, show interhemispheric differences similar to those at the surface (higher levels in the north compared to the south).[30,31] In the Northern Hemisphere, CO often decreases with height,[31] reflecting the abundance of surface sources, but this may be seasonally dependent.[32] In the Southern Hemisphere, CO may increase with altitude or remain relatively constant,[31] and strong convective transport in the tropics can deliver CO to the middle and upper troposphere. Across the tropopause, CO mixing ratios fall below 50 ppb.[33] Transport from the stratosphere brings air with low CO into the troposphere.

Satellite Measurements

Global distributions of CO in the middle troposphere have been determined by the Measurement of Air Pollution from Satellite (MAPS) instrument. MAPS uses a nadir viewing gas filter correlation radiometry with a maximum signal between 400 and 300 mbar and has been flown aboard the U.S. space shuttle four times between 1981 and 1994. Results obtained in October 1984 and 1994 show very high levels of CO over the southern tropics,[25,34] evidence of the strong effect the transport of emissions from surface biomass burning can have on the middle troposphere.

Future measurements from space promise to provide long-term global coverage of CO distributions in the troposphere. The Measurement of Pollution in the Troposphere (MOPITT) instrument was launched December 1999 aboard the EOS *TERRA* (previously known as the *AM-1*) satellite. MOPITT, like MAPS, is a gas filter radiometer that will determine the column abundance of CO. In addition, MOPITT also will retrieve tropospheric profiles of CO (at 4.7 μm) through pressure and length modulation of the correlation cell. Total column abundances of CO and CH_4 will also be measured (at 2.3 μm).[35] MOPITT is expected to provide nearly continuous monitoring of tropospheric CO for a period of at least 5 years. The EOS *Aura* satellite (formerly denoted as *CHEM-1*) is scheduled for launch in June 2003. The *Aura* payload will include TES (tropospheric emission spectrometer), an infrared imaging Fourier transform spectrometer with high spectral resolution that will determine global distributions of CO (and other radiatively trace gases) in the troposphere and lower stratosphere (*http://aura.nasa.gov/tes*).

3 GLOBAL CO BUDGET

Tropospheric distributions of CO reflect its sources and sink combined with the effects of transport. There are believed to be four major sources of CO (Table 1): fossil fuel combustion and industrial activities, biomass burning, the oxidation of methane, and the oxidation of nonmethane hydrocarbons, primarily isoprene and the monoterpenes.[8,36] Anthropogenic activities are thought to account for about two-

TABLE 1 Estimated Sources and Sinks of CO Typical of Last Decade[36]

Sources	Range (Tg CO/yr)	Sinks	Range (Tg CO/yr)
Industry and transportation	300–500	OH reaction	1400–2600
Biomass burning	300–700	Soil uptake	250–640
Emissions from vegetation	60–160	Loss to the stratosphere	~100
Oceans	20–200		
CH_4 oxidation	400–1000		
NMHC oxidation	200–600		
Total sources	1280–3160	Total sinks	1750–3340

thirds of the total source, and reaction with OH radicals is responsible for much of the loss of CO. Sources are unevenly divided between the hemispheres, with as much as 95% of the fossil fuel source, 63% of the biomass burning source, and 68% of the production from the oxidation of NMHC occurring in the Northern Hemisphere.[8] Three-dimensional global transport models, using sources of these magnitudes, have reproduced the measured surface distributions and seasonal cycles with varying degrees of success.[37–39]

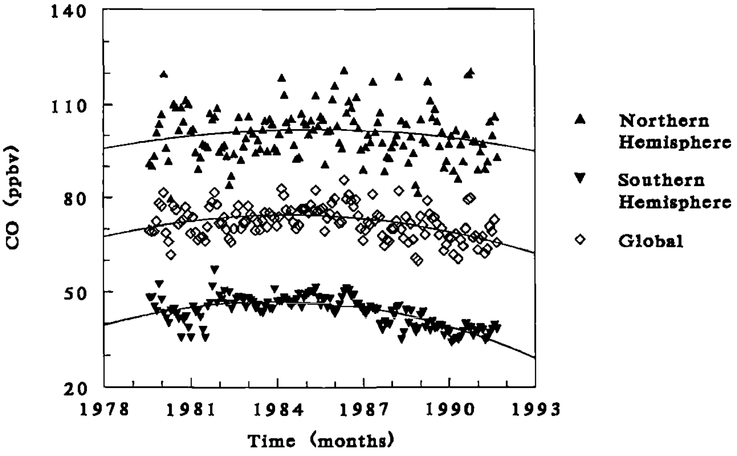
4 TROPOSPHERIC TRENDS

An excess of sources relative to sinks leads to accumulation of gases in the atmosphere. Increasing sources through the industrial era have enhanced atmospheric burdens of CO_2 and CH_4.[36] Similar long-term increases in CO could be expected; however, the few analyses of CO in firn and ice samples have not produced convincing evidence of such a change.[39]

Time series of CO mixing ratios often show periods of increase and decrease.[17,24] Spectroscopic measurements made at Jungfrauhoch, Switzerland, during 1950–1951 and again in 1985–1987 suggested an average rate of increase of ~1% per year in the total column abundance of CO above the European boundary layer.[12] A similar rate of increase was seen in column measurements made over western Russia.[40] Surface measurements made at six sites (evenly distributed between the Northern and Southern Hemispheres) during 1981–1986 suggested a similar rate of increase.[41] In contrast, no significant trend could be identified at Cape Point, South Africa, during the period from the early 1970s through the mid-1980s.[18,42] Trends largely reflect imbalances in its sources and sinks. The reported long-term CO increase in the Northern Hemisphere has been attributed to increasing CO emissions from industrial and transportation-related sources.[12,39,43] However, a quantitative study relating CO emissions and increased atmospheric mixing ratios is still needed.

The long-term increase in CO may have slowed, then reversed in the late 1980s. Khalil and Rasmussen[41,43] present time series determined at six sites beginning 1981 that show an increase in CO over the period 1981–1986, followed by a decrease during 1987–1992 (Fig. 2). The absolute decline in the Northern Hemisphere was about twice that in the south; in the Southern Hemisphere the relative rate of decrease was four times that in the Northern Hemisphere. Novelli et al.[44] reported results from the NOAA/CMDL air sampling network showing a 10% decrease in global average CO mixing ratios during 1992–1993. And while CO declined in both hemispheres, the absolute rate of decrease in the Northern Hemisphere was nearly twice that in the south, while the relative rates were the same (approximately 6 to 7% per year). After 1993, CO levels showed short periods of increase and decrease with some recovery toward pre-1992 levels.[24]

CO time series determined in the north show a significant decrease over the past 10 years; in contrast, a trend in the Southern Hemisphere is more difficult to discern, due in part to the high level of interannual variability. This high variability is likely



Khalil and Rasmussen (1994)

Figure 2 Time series of deseasonalized hemispheric and global mean CO mixing ratios. (Reprinted with permission from Khalil and Rasmussen.[43])

related to yearly variations in emissions from biomass burning. The short-term increases and decreases seen in the recent CO time series[17,24] may be related to interannual variability in biomass burning[43,44] and a short-term increase in OH related to the eruption of Mt. Pinatubo in June 1991.[39,46] Decreased emissions from anthropogenic sources in the Northern Hemisphere have contributed to the observed decrease[44,45].

REFERENCES

1. Migeotte, M., The fundamental band of carbon monoxide at 4. 7 u in the solar spectrum, *Phys. Rev.*, *75*, 1108–1109, 1949.
2. Adel, A., Identification of carbon monoxide in the atmosphere above Flagstaff, Arizona, *J. Astrophys.*, *116*, 442–443, 1952.
3. Robbinson, E. and R. C. Robbins, Atmospheric background concentrations of carbon monoxide, *Ann. New York Acad. Sci.*, *174*, 89–95, 1970.
4. Seiler, W., and C. Junge, Carbon monoxide in the atmosphere, *J. Geophys. Res.*, *75*, 2217–2226, 1970.
5. Seiler, W., and U. Schmidt, New aspects on CO and H_2 cycles in the atmosphere, in N. J. Derco and E. J. Trublar (Eds.), *Proceedings of the International Conference on the Structure, Composition, and General Circulation of the Upper and Lower Atmos. and Possible Anthropogenic Perturbations*, Association of Meteorological and Atmospheric Physics, Toronto, 1974.
6. Levy II, H., Natural atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, *173*, 141–143, 1971.

7. McConnell, J. C., M. B. McElroy, and S. C. Wofsy, Natural sources of atmospheric CO, *Nature*, *233*, 187–188, 1971.

8. Logan, J. A., M. J. Prather, S. C. Wofsy, and M. B. McElroy, Tropospheric chemistry: A global perspective, *J. Geophys. Res.*, *86*, 7210–7254, 1981.

9. Thompson, A. M., The oxidizing capacity of the atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.

10. Sze, N. D., Anthropogenic CO emissions: Implications for the atmospheric CO-OH-CH_4 Cycle, *Science*, *195*, 673–674, 1977.

11. Daniel, J. S and S. Soloman, On the climate forcing of carbon monoxide, *J. Geophys. Res.*, *103*, 13249–13260, 1998 .

12. Zander, R., Ph. Demoulin, D. H. Ehhalt, U. Schmidt, and C. P. Rinsland, Secular increase of the total column abundance of carbon monoxide above central Europe since 1950, *J. Geophys. Res.*, *94*, 11021–11028, 1990.

13. Wallace, L., and W. Livingston, Spectroscopic observations of atmospheric trace gases over Kitt Peak, 2. Nitrous oxide and carbon monoxide from 1979 to 1985, *J. Geophys. Res.*, *95*, 16383–16390, 1990.

14. Doddridge, B. G., R. R. Dickerson, T. G. Spain, S. J. Oltmans, and P. C. Novelli, Measurements of carbon monoxide at Mace Head, Ireland, in ozone in the troposphere and the stratosphere, in R. D. Hudson (Ed.), *Proc. Quad. Ozone Symp.*, *1992*, NASA Conference Publication No. 3266, NASA, Greenbelt, MD, 1994, pp. 134–137.

15. Parrish, D. D., J. S. Holloway, and F. C. Fehsenfeld, Routine, continuous measurement of carbon monoxide with parts per billion precision, *Environ. Sci. Technol.*, *28*, 1615–1618, 1994.

16. Sachse, G. W., G. F. Hill, L. O. Wade and M. G. Perry, Fast-response, high-precision carbon monoxide sensor using a tunable diode laser absorption technique, *J. Geophys. Res.*, *92*, 2071–2081, 1987.

17. Brunke, E.-G., H. E. Scheel, and W. Seiler, Trends of tropospheric CO, N_2O and CH_4 as observed at Cape Point, South Africa, *Atmos. Environ.*, *24A*, 585–595, 1990.

18. Novelli, P. C., L. P. Steele, and P. P. Tans, Mixing ratios of carbon monoxide in the troposphere, *J. Geophys. Res.*, *97*, 20731–20750, 1992.

19. Hurst, D. F., P. S. Bakwin, R. C. Myers, and J. W. Elkins, Behavior of trace gas mixing ratios on a very tall tower in North Carolina, *J. Geophys. Res.*, *102*, 8825–8835, 1997.

20. Novelli, P. C., CO in the atmosphere: Measurement techniques and related issues, *Chemosphere*, *1*, 115–126, 1999.

21. Novelli, P. C., J. W. Elkins, and L. P. Steele, The development and evaluation of a gravimetric reference scale for measurement of atmospheric carbon monoxide, *J. Geophys. Res.*, *96*, 13109–13121, 1991.

22. Weeks, I. A., I. E. Galbally, P. J. Fraser, and G. Matthews, Comparison of the carbon monoxide standards used at Cape Grim and Aspendale, in B. W. Forgan and G. P. Ayers (Eds.), *Baseline Atmospheric Program, 1987*, Australian Government Department of Science and Technology, Canberra, Australia, 1989, pp. 21–25.

23. Novelli, P. C., V. S. Connors, H. G. Reichle, Jr., B. E. Anderson, C. A. M. Brenninkmeijer, E.-G. Brunke, B. G. Doddridge, V. W. J. H. Kirchhoff, J. K. S. Lam, K. A. Masarie, T. Matsou, D. D. Parrish, H. E. Scheel, and L. P. Steele, An internally consistent set of globally distributed atmospheric carbon monoxide mixing ratios developed using results from an intercomparison of measurements, *J. Geophys. Res.*, *103*, 19285–19293, 1998.

24. Novelli, P. C., K. A. Masarie, and P. M. Lang, Distributions and recent trends of carbon monoxide in the troposphere, *J. Geophys. Res.*, *103*, 19015–19033, 1998.
25. Reichle, H. G., Jr., V. S. Connors, J. A. Holland, R. T. Sherrill, H. A. Wallio, J. C. Casas, B. B. Gormsen, and W. Seiler, The distribution of middle tropospheric carbon monoxide during early October 1984, *J. Geophys. Res.*, *95*, 9845–9856, 1990.
26. Fishman, J., K. Fakharuzzaman, B. Cros, and D. Nganga, Identification of widespread pollution in the Southern Hemisphere deduced from satellite analyses, *Science*, *252*, 1693–1696, 1991.
27. Jaffe, D. et al., Transport of Asian air to North America, *Geophys. Res. Lett.*, *26*, 711–714, 1999.
28. Fishman, J. and P. J. Crutzen, The origin of ozone in the troposphere, *Nature*, *272*, 855–858, 1978.
29. Seiler, W., and J. Fishman, The distribution of carbon monoxide and ozone in the free troposphere, *J. Geophys. Res.*, *86*, 7255–7265, 1981.
30. Heidt, L. E., J. P. Krasnec, R. A. Lueb, W. H. Pollock, B. E. Henry, and P. J. Crutzen, Latitudinal distributions of CO and CH_4 over the Pacific, *J. Geophys. Res.*, *85*, 7329–7336, 1980.
31. Marenco, A., M. Macaigne, and S. Prieur, Meridional and vertical CO and CH_4 distributions in the background troposphere (70N–60S; 0–12 k altitude) from the scientific aircraft measurements during the Stratoz III experiment (June 1984), *Atmos. Environ.*, *23*, 185–200, 1989.
32. Yurganov, L. N., D. A. Jaffee, E. Pullman, and P. C. Novelli, Total column and surface densities of atmospheric carbon monoxide in Alaska, 1995, *J. Geophy. Res.*, *103*, 19337–19347, 1998.
33. Seiler, W., and P. Warneck, Decrease of the carbon monoxide mixing ratio at the tropopause, *J. Geophys. Res.*, *77*, 3204–3214, 1972.
34. Connors, V. S., B. B. Gormsen, S. Nolf, and H. G. Reichle, Jr., Spaceborne observations of the global distribution of carbon monoxide in the middle troposphere during April and October 1994, *J. Geophys. Res.*, *104*, 21455–21470, 1999.
35. Drummond, J. R., Measurements of pollution in the troposphere (MOPITT), in J. C. Gille and G. Visconti, (Eds.), *The Use of EOS for Studies of Atmospheric Physics*, North Holland, Amsterdam, 1992, pp. 77–101.
36. Intergovernmental Panel on Climate Change (IPCC), Climate Change 1994: Radiative Forcing of Climate Change, in J. T. Houghton, L. G. M. Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), IPCC, University Press, Cambridge, England, 1995.
37. Allen, D. J., P. Kasibhatla, A. M. Thompson, R. B. Rood, B. G. Doddridge, K. E. Pickering, R. D. Hudson, and S.-J. Lin, Transport-induced interannual variability of carbon monoxide determined using a chemistry and transport model, *J. Geophys. Res.*, *101*, 28655–28669, 1996.
38. Granier, C., J.-F. Muller, S. Madronich, and G. P. Brasseur, Possible causes for the 1990–1993 decrease in the global tropospheric CO abundances: A three-dimensional sensitivity study, *Atmos. Environ.*, *30*, 1673–1682, 1996.
39. Haan, D., and D. Raynaud, Ice core record of CO variations during the last two millennia: Atmospheric implications and chemical interactions within the Greenland ice, *Tellus, 50B*, 253–262, 1998.

40. Yurganov, L. N., E. I. Grechko, and A. V. Dzhola, Zvenigorod carbon monoxide total column time series: 27 years of measurements, *Chemosphere*, *1*, 127–136, 1999.
41. Khalil, M. A. K., and R. A. Rasmussen, Carbon monoxide in the Earth's atmosphere: Indications of a global increase, *Nature*, *332*, 242–245, 1988.
42. Seiler, W., H. Geihl, E.-G. Brunke, and E. Halliday, The seasonality of the CO abundance in the Southern Hemisphere, *Tellus*, *36B*, 219–231, 1984.
43. Khalil, M. A. K., and R. A. Rasmussen, Global decrease in atmospheric carbon monoxide concentration, *Nature*, *370*, 639–641, 1994.
44. Novelli, P. C., K. A. Masarie, P. P. Tans, and P. M. Lang, Recent changes in atmospheric carbon monoxide, *Science*, *263*, 1587–1590, 1994.
45. Bakwin, P. S., P. P. Tans, and P. C. Novelli, Carbon monoxide budget in the Northern Hemisphere, *Geophys. Res. Lett.*, *21*, 433–436, 1994.
46. Bekki, K. S., K. S. Law, and J. A. Pyle, Effect of ozone depletion on atmospheric CH_4 and CO concentrations, *Nature*, *371*, 595–597, 1994.

CHAPTER 6

ATMOSPHERIC METHANE

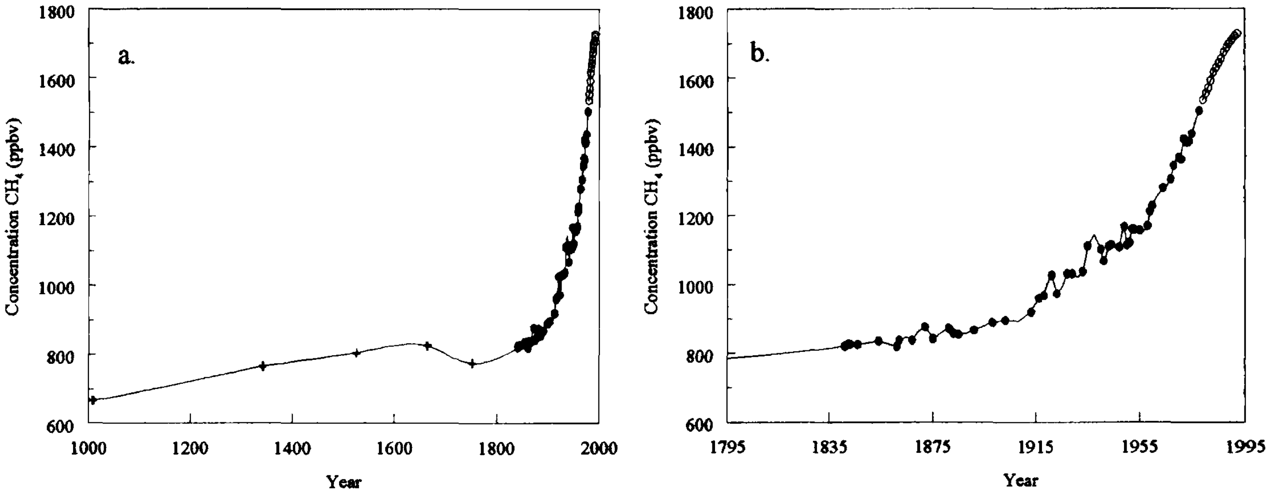
M. A. K. KHALIL AND M. J. SHEARER

1 INTRODUCTION

Methane has been increasing in the atmosphere for about two centuries, resulting in current concentrations that are more than twice the natural levels. Because of these trends, methane is considered to be a potentially important contributor to global warming and other man-made environmental changes that may occur in the future. As a greenhouse gas, every gram of methane released to the atmosphere is about 20 to 60 times as effective as a gram of CO_2, when considered over periods of 20 to 50 years. Moreover, methane has other critical roles in atmospheric chemistry that are also affected by its trends. It exercises a strong influence on the abundance of hydroxyl radicals (OH). These radicals in turn are responsible for removing many man-made and natural gases from the atmosphere. Increasing levels of methane can lead to a lowering of OH levels, that could in turn lead to increases of other gases that may be undesirable. Methane has a complex role in stratospheric chemistry where it is a source of water vapor that tends to deplete the ozone layer but, on the other hand, methane can scavenge chlorine atoms thus protecting the ozone layer from destructive effects of man-made chlorofluorocarbons. In this role high levels of methane are considered desirable. Methane is, therefore, integrally involved in the stability of Earth's environment and, as such, is regarded as one of the important trace gases that are significantly affected by human activities.

We will start by examining the observational data consisting of global distributions and trends. The atmospheric observations are the foundation for most of our current knowledge of the global cycle of methane and our interest in its possible environmental effects. Next we will see how these observations are explained in terms of the processes that produce and destroy methane. Based on the understand-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.


a.
b.
Concentration CH_4 (ppbv)
1800
1600
1400
1200
1000
800
600
1000
1200
1400
1600
1800
2000
1795
1835
1875
1915
1955
1995
Year

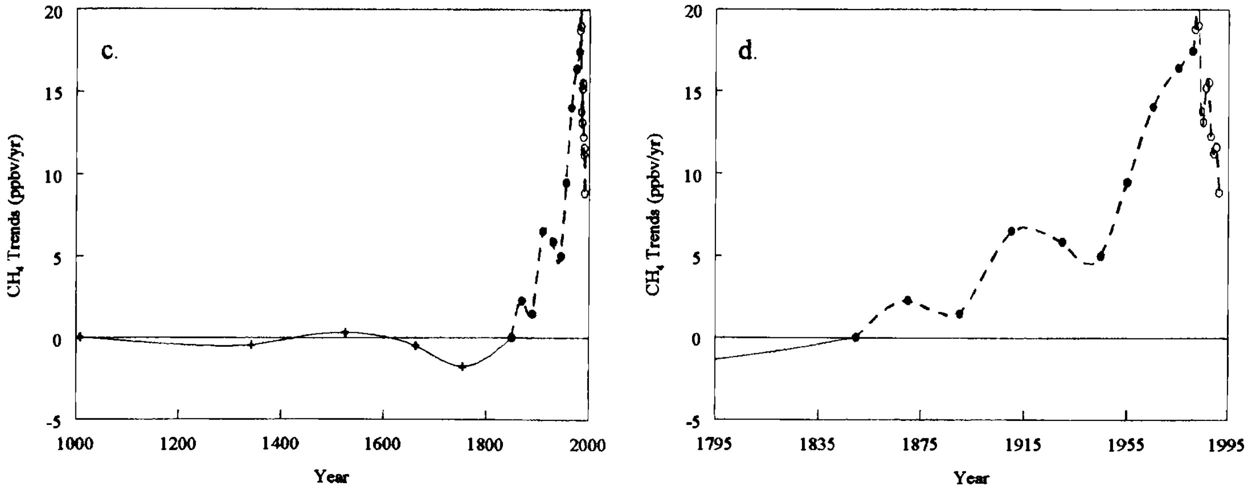


Figure 1 Concentrations of methane (*a*) over the last 1000 years and (*b*) over the last 200 years. Data of Etheridge et al. (1992) (•) and Rasmussen and Khalil (1984) (+) are from ice core samples. Data of Khalil and Rasmussen (○) are global averages from weekly flask samples collected at various latitudes. Trends of methane concentrations (*c*) during the last 1000 years and (*d*) over the last 200 years are calculated from the data shown in (*a*) and (*b*). Rapid increases in methane started only about 200 years ago. These are linear regression estimates of trends over various (nonoverlapping) periods of time between about A.D. 0 and the present. For data between 1840 and 1940, trends were calculated over 20-year periods; for 1940–1980, over 10-year periods; and for 1980–1992 over 2-year periods. The calculated trends were placed at the middle of the time span in each calculation. The earlier data are more sparse. Trends for the period between A.D. 0 and 1800 were calculated for every 10 data points, and the trends are placed at the average time spanned by the 10 data points.

ing gained by such a discussion, we can evaluate plausible expectations for future concentrations and the resulting environmental impact of methane.

2 ATMOSPHERIC OBSERVATIONS

Atmospheric concentrations of methane have been measured systematically for nearly 20 years (Rasmussen and Khalil, 1981; Khalil and Rasmussen, 1983, 1990a; Blake and Rowland, 1988; Steele et al., 1992; Khalil et al., 1993; Dlugokencky et al., 1994). There is an additional record from many independent measurements spanning another 15 years or so back to the early 1960s (Khalil et al., 1989). For earlier times, the ice core record is the only source of information. It extends back over 150,000 years, but for our interest here only the last 1000 years or so are important (Rasmussen and Khalil, 1984; Chappellaz et al., 1990; Etheridge et al., 1992). This record is summarized in Figures 1 and 2. The first panel of these figures shows the time history of methane over the last 1000 years, 100 years, and the most recent decades, and the second panel shows the trends of methane over the same periods.

These data establish two important results: First, that methane concentrations have increased by a factor of about 2.5 over the last 100 to 200 years. And second, that the rates of increase reached peak values during the 1980s, but have been declining since. The rapid increases observed in the 1980s suggested that methane could contribute significantly to global warming in the future. These observations were the compelling reason that drove much of the research on a systematic study of the cycle of atmospheric methane. Later we will return to why these trends are changing.

The most recent decades of data shown in Figure 1 contain many features that reflect the production and destruction processes of methane. There are two salient patterns: The seasonal cycles and the latitudinal concentration gradient. These features are shown more clearly in Figures 3 and 4, respectively.

The data shown are taken at Earth's surface at locations that are far from local sources. As such these concentrations and their patterns represent the large-scale distribution of methane in the atmosphere. In the vertical, up to the tropopause, methane mixing ratios remain nearly the same, implying that the actual concentration of methane in molecules/cm^3 falls off at the same rate as the density of air or approximately 12.5%/km. At higher altitudes, in the stratosphere, the concentrations (molecules/cm^3) fall at an approximate rate of 5%/km as shown in Figure 5.

These observations provide qualitative evidence for several important conclusions regarding the global cycle of methane. Clearly, the methane cycle is out of balance when considered over decadal time scales as evidenced by the generally increasing trends. Moreover, this imbalance arose over the last 100 to 200 years since, before that time, the concentration was unchanging, at least over the previous 1000 years. This would mean that in recent times, more methane is put into the atmosphere than is being removed annually. Second, the latitudinal gradient suggests that the production of methane is considerably higher in the Northern Hemisphere compared with the Southern Hemisphere, if we assume that the destruction processes are similar in

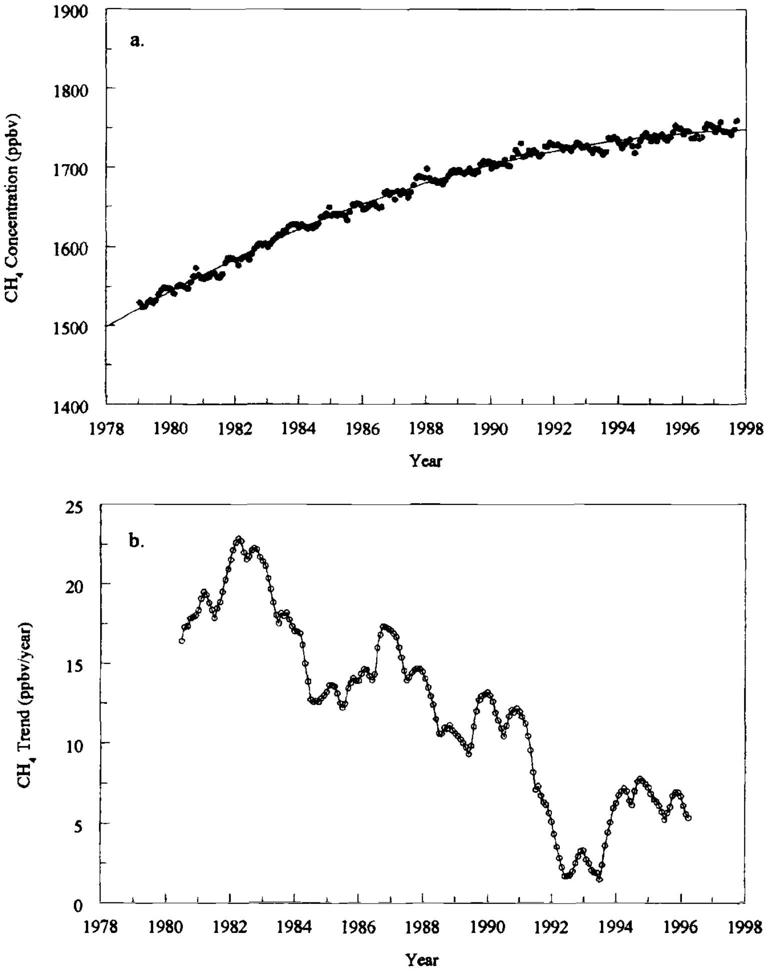


Figure 2 Global average concentration of methane from weekly flask samples collected from six different sites (*a*). The trend of methane (*b*) is calculated by linear regression of 3-year overlapping periods of time, plotted at the center point of the time period.

the two hemispheres. Both these observations, and the timing of the increasing trends, suggest that these changes are caused by human activities. Later, we will look at more direct evidence for this conclusion. Finally, the seasonal cycle suggests that at middle and higher latitudes, the imbalance between production and destruction is greater during winters than summers. This, we will find later, is consistent

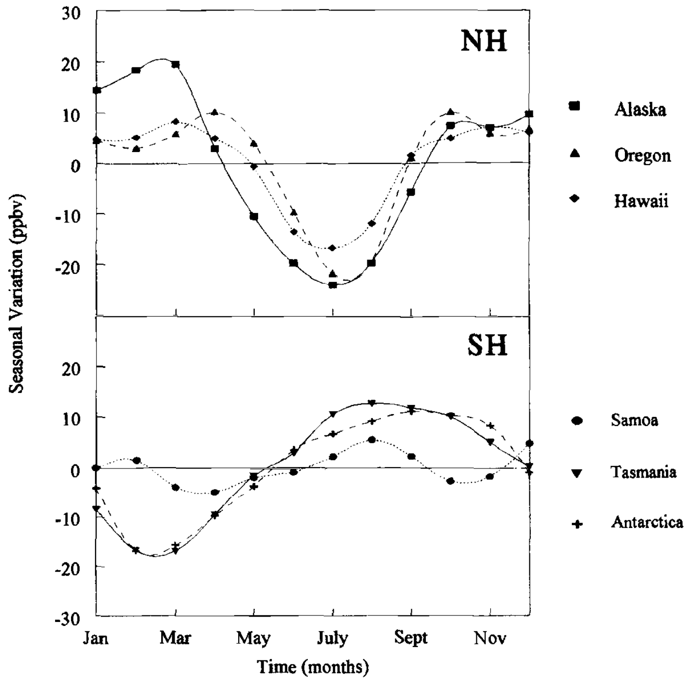


Figure 3 Average seasonal variations of CH_4 at six sites.

with the idea that methane is destroyed by reactions in the atmosphere with hydroxyl radicals that are produced by photochemical processes and hence are most abundant during summers at middle and higher latitudes. While complete objectivity would allow a few alternate explanations, these serve as good working hypotheses. An understanding of the production and destruction processes of methane should explain these observations quantitatively. We will aim toward this goal in the remainder of this chapter.

3 MASS BALANCE

The mass balance of a gas in a hypothetical infinitesimal volume of the atmosphere, in a unit of time, can be expressed as the production less the destruction, added to the net transport of the gas. The net transport can either increase or reduce the concentration within this box during the time of interest. If these three components are perfectly balanced, then the concentration will remain constant; if not, the concentration will change. A fuller treatment of the mass balance requires taking

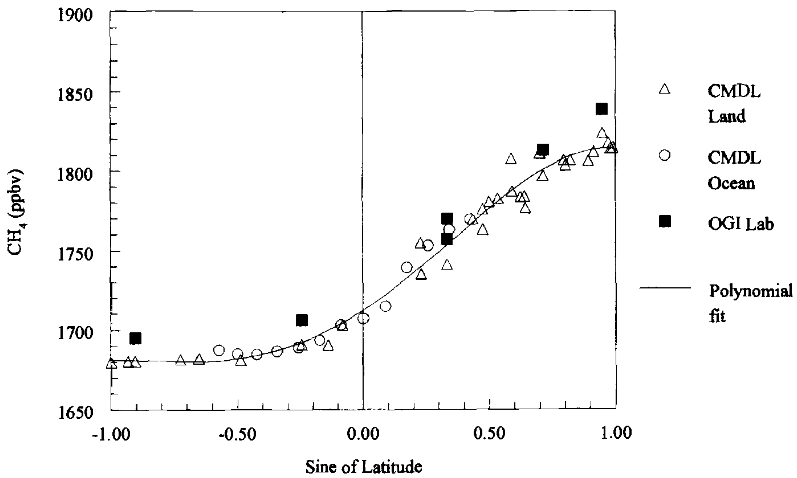


Figure 4 Latitudinal distribution of CH_4. The 1996 NOAA/CMDL data (Dlugokencky et al., 1994) are shown next to the Rasmussen and Khalil flask sampling data from six sites. (Khalil and Rasmussen, 1983) (Calibration difference: $C_{NOAA} = C_{R\&K} - 12\,ppbv$.)

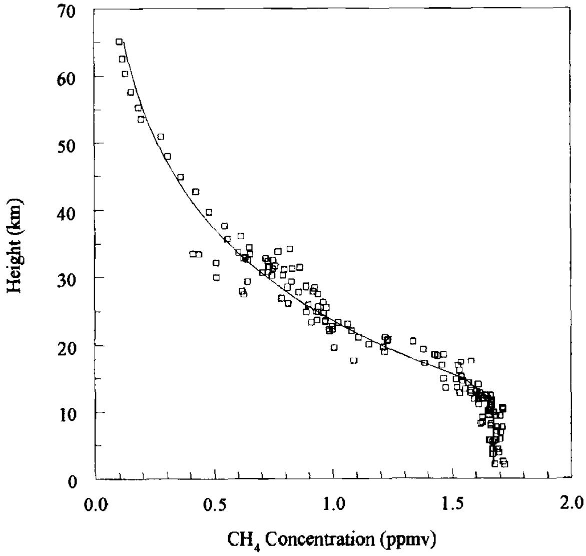


Figure 5 Vertical distribution of CH_4 at 44°N latitude. Data from Fabien et al. (1981), Schmidt et al. (1984, 1987), and Taylor et al. (1989). Concentrations adjusted to base year 1990.

these factors into consideration for each point in the atmosphere and for each unit of time. For our purposes here we will deal with a simplified concept whereby the entire atmosphere is regarded as a single reservoir into which we put methane from its sources and from which methane is removed by a series of chemical and physical processes. Moreover, we take each loss process to be proportional to the amount of methane present at any given time. Considerations of transport of methane are no longer explicitly needed since all methane stays within the atmosphere, and what is moved from one part by the winds, goes to another location, still within the global box, thus not affecting the amount of methane in the global atmosphere. This model can be stated as:

$$\frac{dC}{dt} = S - \frac{C}{\tau} \tag{1}$$

Here C is the amount of methane in Tg (1 Tg $= 10^{12}$ g), S is the emissions from all sources in Tg/yr, and τ is the effective atmospheric lifetime in years. τ is a composite lifetime due to all the processes that remove methane from the atmosphere, or $1/\tau = 1/\tau_1 + 1/\tau_2 + \cdots + 1/\tau_N$ where $\tau_1, \tau_2, \ldots, \tau_N$ represent the lifetimes due to each of N processes. For direct comparisons with measurements C can be converted to ppbv and hence S is expressed in ppbv/yr. The conversion factor is 1 Tg $\approx$ 2.8 ppbv in the global atmosphere.

In the mass-balance equation, we know the solution (C) based on the atmospheric measurements, so our task is to find the two remaining parts of the budget, namely the emissions (S) or the lifetime (τ). For the case of methane, the lifetime is calculated independently, so that the mass-balance equation is essentially a tool for finding the sources that have combined emissions that are consistent with the measured concentrations and calculated loss rates [Eq. (1)]. How Eq. (1) is used will be discussed next; then we will show the recent budgets consistent with the known constraints and the mass balance expressed by Eq. (1).

Since there are many sources, the mass-balance equation by itself is insufficient to constrain how much methane comes from each source. Nonetheless, we can use it to estimate the total annual emissions of methane. Based on a calculated lifetime of 10 years (τ), to be discussed later, and a global burden of 4800 Tg (C) obtained from global measurements (Figs. 1 to 5), and a current rate of change of about 20 Tg/yr (dC/dt; Fig. 1), we see from Eq. (1) that the total worldwide emissions from all sources should be about 500 Tg/yr. This is a useful benchmark.

There are a number of ways to improve on Eq. (1) that will allow us to reduce the uncertainties in the estimate of emissions from individual sources or combinations of sources. We will discuss three approaches here that have been useful in developing better global budgets. One method is to consider the long time series, such as the ice core data over several centuries, and apply Eq. (1) to two different time periods. This method uses the observed changes over long time periods to determine the ratio of anthropogenic to natural emissions. If we assume that several hundred years ago the concentration of methane in the atmosphere was determined entirely by natural processes, we can then estimate the emissions that would be required to satisfy

Eq. (1). This is particularly simple since at that time there are no significant trends, so $S = C/\tau$. Based on the ice core data for C, we can estimate the emissions to be 1700 Tg/yr/10 yr = 170 Tg/yr. We have already done the calculation for recent times suggesting present emissions of 500 Tg/yr. This would imply that there are new sources, presumably due to human activities, amounting to some 330 Tg/yr (Khalil and Rasmussen, 1990b). We have assumed that the lifetime of methane is the same now as it was a century or more ago. There is reason to believe that this is a good approximation, but the matter is open to question.

Another approach is to consider the latitudinal distribution of the various sources, determined by independent data or measurements. Then, a more detailed version of the mass-balance model, which takes into account the budget of methane over small regions of Earth's surface, can be used to determine whether the estimated rate of emissions from the sources is compatible with the measured atmospheric concentrations within each location. This method uses the latitudinal distribution to constrain the strength of the sources (Fung et al., 1991; Brown, 1993; Hein et al., 1997). For instance, we can rule out the oceans as the major source because that would require a more even distribution of methane across the hemispheres than is seen in Figure 4.

A recent approach at constraining the estimates of emissions uses carbon isotopes in methane. Normal measurements of methane cannot distinguish between the molecules of methane that come from one source or another, so only the total amount or concentration C is measured. It has become possible to measure the methane with different isotopes of carbon—specifically $^{12}CH_4$, $^{13}CH_4$, and $^{14}CH_4$ (Tyler, 1986; Stevens and Engelkemeir, 1988; Wahlen et al., 1989; Quay et al., 1991, Lassey et al., 1993). In this case, we can get more information on the global sources (and sinks, or loss processes) of methane since we can now have three equations similar to Eq. (1), one for each isotope. We now have to independently balance three types of methane ($^{12}CH_4$, $^{13}CH_4$, and $^{14}CH_4$) instead of just the sum of all types, using the same sources. There are fewer combinations of sources and emission rates that would balance all types than there are for just the total methane. Recently, stable isotopes of hydrogen in methane ($^{12}CH_3D$) have also been measured (Bergamaschi and Harris, 1995), which would add a fourth type of methane to constrain the sources. This work requires a knowledge of not only the isotopic combination of methane in the atmosphere, but also of the amounts of each type emitted by the sources, and the atmospheric lifetimes of each type of methane. Isotopic measurements hold considerable promise for reducing the uncertainties in the budget of methane.

4 SOURCES AND SINKS

Usually the global emission rate from a source is estimated by using a measured emission factor (grams of CH_4 emitted/day/unit source) and multiplying it by the number of such units in the world (units of source) and the time of year when emissions take place (days/year), which results in the grams of CH_4/year emitted by the source. The complexity of the estimate varies depending on the information available. Once the budget is assembled, it must comply with the constraints discussed earlier.

Over the years many budgets have been proposed. Most of them were not entirely independent of previous estimates but tended to improve the estimates of emissions for one source or another. Two recent budgets are shown in Table 1. Both are "consensus"-type budgets in which several types of estimates by different researchers are put together. The first is from a NATO-sponsored Advanced Research Workshop (Khalil and Shearer, 1993). One of the goals of this project was to improve the budget based on direct measurements of emission factors and data on their global extrapolation. The second budget is from an assessment of the Intergovernmental Panel on Climate Change (IPCC) (Prather et al., 1995). The two budgets show one measure of the level of uncertainty that currently exists in the estimates of emissions from individual sources. Both these budgets are generally consistent with the known constraints, including the total emissions of around 500 Tg/yr discussed earlier. The budgets satisfy the constraints of the ratio of natural to anthropogenic emissions required by the ice core data. The budgets also agree on the major sources.

These budgets, like earlier ones, show that there are a few major sources. The major natural source is the wetlands, as has been known for a long time, since

TABLE 1 Comparison of Two Recent Budgets of Methane Sources

Source	NATO–ARW (1993)	IPCC (1994)
Natural sources (Tg)	Tg	Tg
Wetlands	110	115 (55–150)
Termites	20 (15–35)[a]	20 (10–50)
Open ocean	4	10 (5–50)
Marine sediments	(8–65)	
Geological	10 (1–13)	
Wild fire	2 (2–5)	
Other		15 (10–40)
Natural total	150	160 (110–210)
Anthropogenic source (Tg)	Tg	Tg
Rice agriculture	65 (55–90)	60 (20–100)
Animals	79	85 (65–100)
Manure	15	25 (20–30)
Landfills	22 (11–33)	40 (20–70)
Wastewater treatment	25 (12–38)	25 (15–80)
Biomass burning	50	40 (20–80)
Coal mining	46	30 (15–45)
Natural gas	30 (25–50)	40 (25–50)
Other anthropogenic	13 (7–30)	15 (5–30)
Low-temperature fuels	17	? (1–30)
Anthropogenic total	360	375 (300–450)
Total	510	535 (410–660)

[a] Numbers in parentheses show estimated range of source values.

methane has been called "marsh gas." Other natural sources are generally small but not well constrained. These include termites, oceans, and lakes. Most of the current sources are "anthropogenic." While these emissions are not directly from stacks and other easily identifiable icons of man-made pollution, they are a result of human activities nonetheless. These sources may be classified mostly as agricultural and from use of energy. Of these, rice agriculture, cattle, waste management, biomass burning, coal mining, and use of natural gas are the largest contributors. There are some moderate sized sources of a few teragrams/year that include transportation and fossil fuel combustion. There are perhaps many small sources that together fall within the range of uncertainty of the global emission rate and are therefore not included.

Methane is removed from the atmosphere by a number of processes. The most effective is reaction with hydroxyl radicals, or OH. The main process by which hydroxyl radicals are formed in the atmosphere occurs when sunlight splits an ozone molecule into O_2 and $O(^1D)$, an excited state of the oxygen atom. A few of these $O(^1D)$ atoms react with water vapor (H_2O) to form two OH radicals. OH has a lifetime of a few seconds and is removed mostly by its reaction with methane and CO. In addition to these major processes there are others that contribute to both the formation and destruction of OH radicals in the atmosphere (Thompson, 1992; DeMore et al., 1997). OH radicals are also responsible for removing many other gases from the atmosphere both man-made and natural. For example, many of the recently introduced chemicals (hydrofluorocarbons and hydrochlorofluorocarbons) that replace the chlorofluorocarbons are removed by OH radicals in the lower atmosphere. Methane is not only removed by OH, but there is enough methane in the atmosphere to control the abundance of OH and hence the oxidizing capacity of the atmosphere. Current estimates using photochemical models or proxy data suggest that the average concentration of OH is about 10^6 molecules/cm^3. At any location, OH concentrations vary greatly depending on latitude, altitude, season, and time of day. The rate constant (K) for the reaction of methane with OH is about 2.4×10^{-15} cm^3/molecule/sec 256 K (the average temperature of the atmosphere). Then, according to Eq. (1) the total loss of methane due to reactions with OH should be $C/\tau_{OH} = K_{[OH]}C$ or 400 Tg/yr after appropriate unit conversions. This corresponds to a lifetime of about 12 years due to reactions with OH alone.

Methane is also removed at Earth's surface by deposition and transport into the soils and then utilized by biological processes. This sink is estimated to be about 25 to 30 Tg/yr based on experimental field data. In the stratosphere methane is removed again by reacting with OH and also by other photolytic processes (DeMore et al., 1997). Recently, it has been suggested that there may be significant concentrations of Cl atoms in the marine boundary layer produced by precursors from the oceans. The concentration of these radicals is not known at present, but various estimates put it between 10^3 to 10^6 molecules/cm^3 (Singh and Kasting, 1988; Keene et al., 1990; Graedel and Keene, 1995, and references therein). At the upper limit this would constitute a significant sink for methane since it reacts 17 times faster with Cl atoms than with OH at the temperature in the boundary layer. Depending on how much of the marine atmosphere contains Cl radicals, this sink could be as large as 50 Tg/yr. Although we have not stated the sizable range of uncertainties in the estimates of these smaller sinks, it should be noted that these calculations provide a composite lifetime of about 50 years, which when combined with the lifetime due to OH reactions results in a total global lifetime of about 10 years. This was used to impose the first constraint discussed earlier based on Eq. (1) whereby we calculated the total emissions to be about 500 Tg/yr.

5 PAST AND PRESENT TRENDS

While the budgets discussed so far represent current conditions, we need to know how these emissions have changed over the years before we can match the observed trends of concentrations shown in Figure 1 and represented in Eq. (1). Estimating past emissions, or a time series for each of these sources, is even more difficult than estimating current emission rates. The simplest approach is to assume that the current estimate of the total anthropogenic emissions is proportional to the human population, and the natural emissions have remained the same over the last 100 to 200 years. With these assumptions we can generate the emissions in Eq. (1) for the last century and calculate the expected concentrations. Although these assumptions are rough approximations, the results of this calculation explain the data quite well (Khalil and Rasmussen, 1994).

The assumption that anthropogenic emissions are proportional to human population breaks down in the recent decades. The atmospheric concentrations are not increasing as rapidly as would be expected if the anthropogenic emissions kept pace with the rising population. When we look at the data on the anthropogenic sources such as cattle populations or the area of rice fields, we find that these are not increasing any longer or are increasing very slowly. In the past these sources had been increasing at a rate proportional to human population. It seems then that there is a decoupling of the anthropogenic emissions from the human population. This circumstance makes the use of potential growth of human population an untenable surrogate for future emissions, even though it works well for the past.

An alternate and more detailed approach is to use the available agricultural and energy data to estimate how these emissions may have changed. Fortunately there are good records, going back a hundred years, on the number of cattle in the world and the hectares of rice harvested each year. Similar, but possibly less accurate estimates can also be made for the other anthropogenic sources based on archived records. We estimated the global emissions from the major sources over the last 100 years and calculated the expected concentrations using Eq. (1). The results are shown in Figure 6.

These results show that the available data for the calculation of global emissions over the last 100 years are in fact consistent with the observed concentrations shown in Figure 1. The long-term trends are driven by increases in rice agriculture and domestic cattle and collectively by the other anthropogenic sources. These same sources that led to the major increases of concentration over the last century are now stabilizing and causing the decreasing trends, at least in this model, and a more

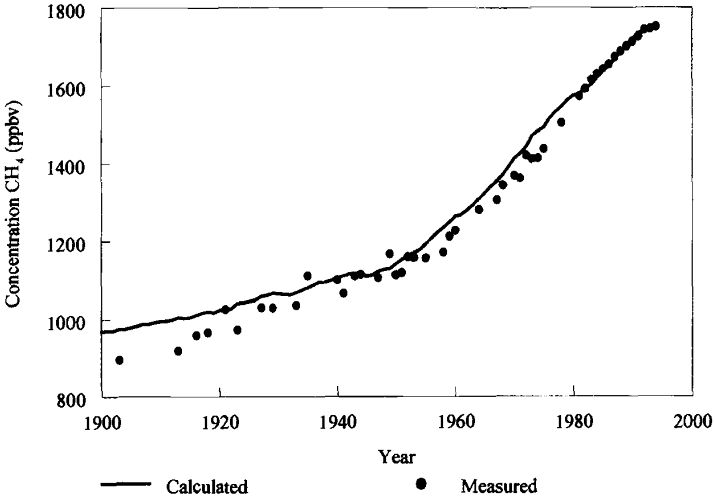


Figure 6 Comparison of the globally averaged calculated concentrations (shown by a smoothed line) with the measured concentrations of methane.

stable concentration of methane at present levels of 1750 ppbv. The results shown in Figure 6 are expanded for the recent decades and the trend both observed and calculated is plotted in Figure 7 (Khalil et al., 1996). These calculations lend support to the idea that the recent slowdown in the trend is caused by a stabilization of the major anthropogenic sources, mainly rice agriculture and domestic cattle.

The trends of methane can also be explained by changing levels of OH. If OH decreases, methane would increase and if it increases methane would decrease. Both these mechanisms have sometimes been discussed as alternatives to the explanation based on changing emission rates (Crutzen and Zimmermann, 1991; Thompson et al., 1993). It is more appropriate to consider this aspect as a contributing factor rather than an alternative explanation. For the long term, it is thought that OH may have decreased, adding to the trend of methane. This decrease of OH may have come about because both CO and CH_4 have increased over the last century, thus increasing the speed of removal of OH leading to lower concentrations. This process may have been partially compensated by increased ozone that can lead to greater production and by other chemical feedbacks. Calculations suggest that the long-term changes of OH are probably small and not sufficient to explain the major part of the observed increase (Khalil and Rasmussen, 1993; Pinto and Khalil, 1991; Lu and Khalil, 1991).

In more recent decades, there has been evidence for the depletion of the ozone layer due to the man-made chlorofluorocarbons. This would cause an increase of ultraviolet (UV) radiation in the troposphere where it would stimulate the splitting of

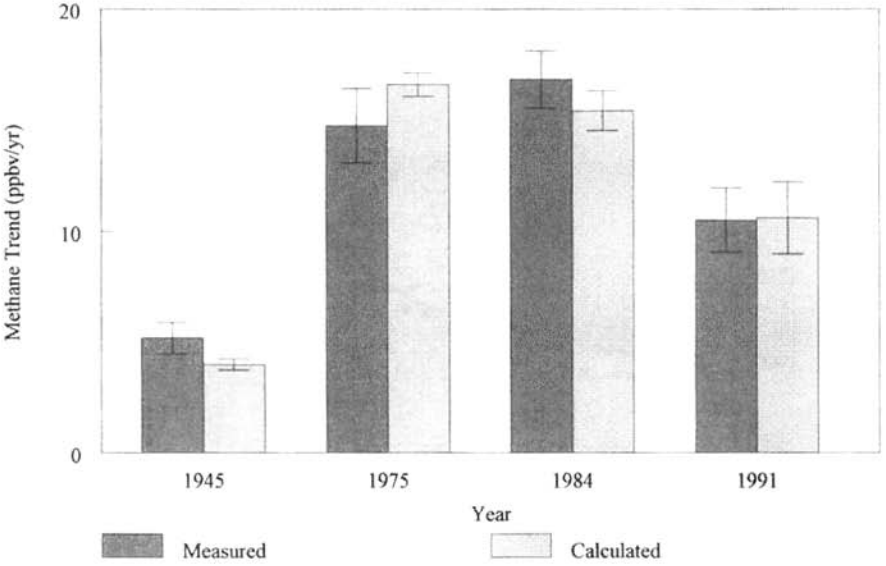


Figure 7 Comparison of trends calculated from measured methane concentrations vs. modeled methane concentrations.

the O_3 molecules discussed earlier into O_2 and $O(^1D)$, thus increasing the production of OH (Madronich and Granier, 1992). The increased OH would cause the trend of methane to slow down. Such a mechanism may contribute to the slowdown of the trend, but there is no experimental evidence that can pin down the magnitude of the OH trend. The increase of OH, if it is occurring, appears to be small and not sufficient to explain the entire observed slowdown (Krol et al., 1998). It will take more work before we can say how much of the current trend is affected by possible changes of OH and how much is from the slowdown of emissions. For the present, however, it seems that the slowdown of emissions can be estimated, and these estimates of changing emissions are sufficient to explain the general pattern of the observations as shown in Figure 7.

6 DISCUSSION AND COMMENTARY

The state of knowledge about the methane cycle is that we have a clear understanding of the global distributions and trends in recent decades and over the last century for which ice core data are used. This is a directly measurable component of the global balance. There are no substantive differences among the various groups who have measured methane in the atmosphere. Field studies of emissions from the various sources are also in broad agreement, and the differences that have been observed are explained by environmental variables. Extrapolation of the field data to global emission rates remains a major source of uncertainty, leaving a sizable uncertainty in the estimates of global emissions from each source. The main features of the trends, both the increases over the last century and the slowdown of the trend in recent times, are consistent with what we know about the change of emissions from anthropogenic sources. There is enough uncertainty that trends caused by changes of OH concentrations can be accommodated.

Although the current understanding of the methane distribution and trends can be explained by the known sources and sinks, the very nature of these explanations clouds our ability to predict future concentrations. We see that the major anthropogenic sources—rice fields, cattle and also biomass burning—are all stabilizing not because of legislated controls, but because there are natural limitations to the growth of these sources. These sources will not keep pace with increasing population as new technologies make it unnecessary to do so. For instance, new high yielding varieties of rice do not require as much land or time in the growing season to produce the same amount of rice as before, thus reducing the emissions of methane per bushel of rice grown. If the anthropogenic sources could be related to population in the future, it would then be easier to predict future emissions under various assumptions of population growth—but this is not possible as we have discussed. Various scenarios that had been hypothesized are no longer likely (Alcamo et al., 1995). Past estimates of the doubling of methane to 3 to 4 ppmv are now unlikely with no known sources that could increase sufficiently to cause such high concentrations. Perhaps the only prediction that can be made is that it is quite unlikely that the concentrations of methane will increase substantially or double in the next decade or two. This is good news for global warming since it is not likely to be as much as previously expected from the increase of methane. As such, methane will continue to play an important role in the global environment, but this role is not likely to increase for years to come.

REFERENCES

Alcamo, J., A. Bouwman, J. Edmonds, A. Grübler, T. Morita, and A. Sugandhy, An evaluation of the IPCC IS92 emission scenarios, in J. T. Houghton, L. G. Meira Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), *Climate Change 1994, Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*, Intergovernmental Panel on Climate Change, Cambridge University Press, Great Britain, 1995.

Bergamaschi, P., and G. W. Harris, Measurement of stable isotope ratios ($^{13}CH_4/^{12}CH_4$; $^{12}CH_3D/^{12}CH_4$) in landfill methane using a tunable diode laser absorption spectrometer, *Global Biogeochem. Cycles*, *9*, 439–447, 1995.

Blake, D. R., and F. S. Rowland, Continuing worldwide increase in tropospheric methane, 1978 to 1987, *Science*, *239*, 1129–1131, 1988.

Brown, M., Deduction of emissions of source gases using an objective inversion algorithm and a chemical transport model, *J. Geophys. Res.*, *98*, 12639–12660, 1993.

Chappellaz, J., J. M. Barnola, D. Raynaud, Y. S. Korotkevich, and C. Lorius, Ice-core record of atmospheric methane over the past 160,000 years, *Nature*, *345*, 127–131, 1990.

Crutzen, P. J., and P. H. Zimmermann, The changing photochemistry of the troposphere, *Tellus*, *43AB*, 136–151, 1991.

DeMore, W. B., S. P. Sander, C. J. Howard, A. R. Ravishankara, D. M. Golden, C. E. Kolb, R. F. Hampson, M. J. Kurylo, and M. J. Molina, *Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling*, JPL Publication 97-4, National Aeronautics and Space Administration Jet Propulsion Laboratory, Pasadena, CA, 1997.

Dlugokencky, E. J., L. P. Steele, P. M. Lang, and K. A. Masarie, The growth rate and distribution of atmospheric methane, *J. Geophys. Res.*, *99*, 17021–17043, 1994.

Etheridge, D. M., G. I. Pearman, and P. J. Fraser, Changes in tropospheric methane between 1841 and 1978 from a high accumulation-rate Antarctic ice core, *Tellus*, *44B*, 282–294, 1992.

Fabian, P., R. Borchers, G. Glentje, W. A. Matthews, W. Seiler, H. Giehl, K. Bunse, F. Müller, U. Schmidt, A. Volz, A. Khedim, and F. J. Johnen, The vertical distribution of stable trace gases at mid-latitudes, *J. Geophys. Res.*, *86*, 5179–5184, 1981.

Fung, I., J. John, J. Lerner, E. Matthews, M. Prather, L. P. Steele, and P. J. Fraser, Three-dimensional model synthesis of the global methane cycle, *J. Geophys. Res.*, *96*, 13033–13065, 1991.

Graedel, T. E., and W. C. Keene, Tropospheric budget of reactive chlorine, *Global Biogeochem. Cycles*, *9*, 47–77, 1995.

Hein, R., P. J. Crutzen, and M. Heimann, An inverse modeling approach to investigate the global atmospheric methane cycle, *Global Biogeochem. Cycles*, *11*, 43–76, 1997.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz-Tokos, H. Sievering, and J. F. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Khalil, M. A. K., and R. A. Rasmussen, Sources, sinks, and seasonal cycles of atmospheric methane, *J. Geophys. Res.*, *88*, 5131–5144, 1983.

Khalil, M. A. K., and R. A. Rasmussen, Atmospheric methane: Recent global trends, *Environ. Sci. Technol.*, *24*, 549–553, 1990a.

Khalil, M. A. K., and R. A. Rasmussen, Constraints on the global sources of methane and an analysis of recent budgets, *Tellus*, *42B*, 229–236, 1990b.

Khalil, M. A. K., and R. A. Rasmussen, Decreasing trend of methane: Unpredictability of future concentrations, *Chemosphere*, *26*, 595–608, 1993.

Khalil, M. A. K., and M. J. Shearer, Sources of methane: An overview, in M. A. K. Khalil (Ed.), *Atmospheric Methane: Sources, Sinks, and Role in Global Change*, NATO ASI Series I: *Global Environmental Change*, Vol. 13, Springer-Verlag, Berlin, 1993.

Khalil, M. A. K., and R. A. Rasmussen, Global emissions of methane during the last several centuries, *Chemosphere*, *29*, 833–842, 1994.

Khalil, M. A. K., R. A. Rasmussen, and F. Moraes, Atmospheric methane at Cape Meares: Analysis of a high resolution data base and its environmental implications, *J. Geophys. Res.*, *98*, 14753–14770, 1993.

Khalil, M. A. K., R. A. Rasmussen, and M. J. Shearer, Trends of atmospheric methane during the 1960s and 1970s, *J. Geophys. Res.*, *94*, 18279–18288, 1989.

Khalil, M. A. K., M. J. Shearer, and R. A. Rasmussen, Atmospheric methane over the last century, *World Resource Rev.*, *8*, 481–492, 1996.

Krol, M., P. J. van Leeuwen, and J. Lelieveld, Global OH trend inferred from methylchloroform measurements, *J. Geophys. Res.*, *103*, 10697–10711, 1998.

Lassey, K. R., D. C. Lowe, C. A. M. Brenninkmeijer, and A. J. Gomez, Atmospheric methane and its carbon isotopes in the Southern Hemisphere: Their time series and an instructive model, *Chemosphere*, *26*, 95–109, 1993.

Lu, Y., and M. A. K. Khalil, Tropospheric OH: Model calculations of spatial, temporal, and secular variations, *Chemosphere*, *23*, 397–444, 1991.

Madronich, S., and C. Granier, Impact of recent total ozone changes on tropospheric ozone photodissociation, hydroxyl radicals, and methane trends, *Geophys Res. Lett.*, *19*, 465–467, 1992.

Pinto, J. P., and M. A. K. Khalil, The stability of tropospheric OH during ice ages, inter-glacial epochs and modern times, *Tellus*, *43B*, 347–352, 1991.

Prather, M., R. Derwent, D. Ehhalt, P. Fraser, E. Sanhueza, and X. Zhou, Other trace gases and atmospheric chemistry, in J. T. Houghton, L. G. Meira Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), *Climate Change 1994, Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*, Intergovernmental Panel on Climate Change, Cambridge University Press, Great Britain, 1995.

Quay, P. D., S. L. King, J. Stutsman, D. O. Wilbur, L. P. Steele, I. Fung, R. H. Gammon, T. A. Brown, G. W. Farwell, P. M Grootes, and F. H. Schmidt, Carbon isotopic composition of atmospheric CH_4: Fossil and biomass burning source strengths, *Global Biogeochem. Cycles*, *5*, 25–47, 1991.

Rasmussen, R. A., and M. A. K. Khalil, Increase in the concentration of atmospheric methane, *Atmos. Environ.*, *15*, 883–886, 1981.

Rasmussen, R. A., and M. A. K. Khalil, Atmospheric methane in the recent and ancient atmospheres: Concentrations, trends, and interhemispheric gradient, *J. Geophys. Res.*, *89*, 11599–11605, 1984.

Schmidt, U., A. Khedim, D. Knapsa, G. Kulessa, and F. J. Johnen, Stratospheric trace gas distributions observed in different seasons, *Adv. Space Res.*, *4*, 131–134, 1984.

Schmidt, U., G. Kulessa, E. Klein, E.-P. Röth, P. Fabian, and R. Borchers, Intercomparison of balloon-borne cryogenic whole air samplers during the MAP/GLOBUS 1983 campaign, *Planet. Space Sci.*, *35*, 647–656, 1987.

Singh, H. B., and J. F. Kasting, Chlorine-hydrocarbon photochemistry in the marine troposphere and lower stratosphere, *J. Atmos. Chem.*, *7*, 261–285, 1988.

Steele, L. P., E. J. Dlugokencky, P. M. Lang, P. P. Tans, R. C. Martin, and K. A. Masarie, Slowing down of the global accumulation of atmospheric methane during the 1980's, *Nature*, *358*, 313–316, 1992.

Stevens, C. M., and A. Engelkemeir, Stable carbon isotope composition of methane from some natural and anthropogenic sources, *J. Geophys. Res.*, *93*, 725–733, 1988.

Taylor, F. W., A. Dudhia, and C. D. Rogers, Proposed reference models for nitrous oxide and methane in the middle atmosphere, in G. M. Keating (Ed.), *Handbook for MAP*, Vol. 31, 1989, pp. 67–79. Middle Atmosphere Program ISCU SCOTEP, U. of Illinois, Urbana , IL, USA.

Thompson, A. M., The oxidizing capacity of the Earth's atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.

Thompson, A. M., J. A. Chappellaz, and I. Y. Fung, The atmospheric CH_4 increase since the Last Glacial Maximum (2) Interactions with oxidants, *Tellus*, *45B*, 242–257, 1993.

Tyler, S. C., Stable carbon isotope ratios in atmospheric methane and some of its sources, *J. Geophys. Res.*, *91*, 13232–13238, 1986.

Wahlen, M., N. Tanaka, R. Henry, B. Deck, J. Zeglen, J. S. Vogel, J. Southon, A. Shemesh, R. Fairbanks, and W. Broecker, Carbon-14 in methane sources and in atmospheric methane: The contribution from fossil carbon, *Science*, *245*, 286–290, 1989.

CHAPTER 7

BIOGENIC NON-METHANE HYDROCARBONS

MARCY E. LITVAK

1 INTRODUCTION

Nonmethane volatile organic compounds (NMVOCs) are emitted from a wide variety of both anthropogenic and biogenic sources. Major anthropogenic sources of NMVOCs include combustion of fossil fuels, solvent evaporation and biomass burning, while direct emissions from plants are the largest biogenic source. Over 90% of the total NMVOCs entering the atmosphere are biogenic (Guenther et al., 1995; Müller, 1992). Recent estimates of the upper limit of global NMVOC emissions from biogenic sources range from 1000 to 1500 Tg C/yr (1 Tg = 10^{12} g), an amount equivalent to the total methane flux from both biogenic and anthropogenic sources (Guenther et al., 1995).

In the atmosphere, NMVOCs are typically very reactive (lifetimes range from minutes to days) and play significant roles in many aspects of atmospheric chemistry. NMVOCs are a key component of the photochemical processes that form ozone and other secondary products in the planetary boundary layer (Fehsenfeld et al., 1992). The other products produced include organic acids, organic nitrates, aerosols, acetone, formaldehyde, and carbon monoxide (Kasting and Singh, 1986; Trainer et al., 1987; Chameides et al., 1988; Jacob and Wofsy, 1988; Andreae et al., 1988; Fehsenfeld et al., 1992). These products are relevant in that they can contribute to both air pollution and climate change. Ozone is not only a potent greenhouse gas but can impact human health and plant productivity. Organic nitrates such as PAN (peroxyacetyl nitrate) are phytotoxic, an important component of urban smog, and also provide a mechanism for transporting reactive nitrogen (NO and NO_2, together referred to as NO_x) over large distances (Sillman and Samson, 1995). Organic

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

aerosol particles scatter light at all visible wavelengths, which creates haze and decreases visibility (Andreae and Crutzen, 1997; Pandis et al., 1991). Finally, NMVOCs and carbon monoxide are considerably more reactive toward the hydroxyl radical (OH) than is methane. Increased levels of CO and NMVOCs therefore can significantly suppress OH concentrations and thus the oxidative capacity of the troposphere, resulting in a longer atmospheric lifetime for methane (Brasseur and Chatfield, 1991; Fehsenfeld et al., 1992).

In many localized areas, biogenic hydrocarbons play a dominant role in generating tropospheric ozone locally and in rural areas downwind from these urban centers. In Atlanta, Georgia, which has a high density of NMVOC-emitting plant species, Geron et al. (1995) estimated that with current NO_x and biogenic hydrocarbon emissions, even if anthropogenic hydrocarbon emissions were reduced to zero, ozone levels would be above the National Ambient Air Quality Standard (NAAQS). In rural areas long distances from polluted plumes, anthropogenic VOCs are so diluted that isoprene and terpenes emitted from vegetation alone are enough to sustain ozone production (Roselle et al., 1991; Hagerman et al., 1997).

The list of NMVOCs emitted from biogenic sources includes well over 1000 compounds. In many ecosystems, isoprene and monoterpenes are the predominant biogenic hydrocarbons emitted, and these compounds account for over half of the total global NMVOC fluxes from biogenic sources (Table 1). However, recent

TABLE 1 Major Biogenic Methane and NMVOC Sources, Source Strength, and Atmospheric Lifetimes[a]

VOC	Primary Natural Sources	Estimated Annual Global Emission (Tg C)	Reactivity in Atmosphere (lifetime in days)
Methane	Wetlands, rice paddies	319–412	4000
Isoprene	Plants	175–503	0.2
Monoterpenes	Plants	127–480	0.1–0.2
Ethene	Plants, soils, oceans	8–25	1.9
Other reactive VOCs (e.g., acetaldehyde, formaldehyde MBO, hexenal family)	Plants	~260	<1
Other less reactive VOCs (e.g., methanol, ethanol, formic acid, acetic acid, acetone)	Plants, soils	~260	>1

[a] Adapted from Fall (1999). Data are derived from Singh and Zimmerman (1992), Conrad (1995), Guenther et al. (1995), Andreae and Crutzen (1997), and Rudolph (1997).

studies have emerged indicating that many other hydrocarbons, particularly oxygenated VOCs, also provide a significant contribution to the total biogenic NMVOC flux (Isidorov et al., 1985; Arey et al., 1991; Winer et al., 1992; König et al., 1995; Helmig et al., 1999). Quantitative measurements of most of these "other" NMVOCs are scarce because of the wide variety of sources and difficulty in reliable identification and quantification of these compounds.

In this chapter, the major classes of hydrocarbons and their oxygenated derivatives emitted to the atmosphere from biogenic sources are reviewed. Information is also provided on ambient mixing ratios, regional and global distribution, and the primary controlling factors over emissions of these classes of biogenic NMVOC's. Of the large group of biogenic NMVOCs that are highly reactive in the atmosphere (have lifetimes of less than one day), this chapter will mainly focus on isoprene, monoterpenes, ethene, propene, butene, acetaldehyde, formaldehyde, 2-methyl-3-buten-2-ol (MBO), and the hexenal family compounds (hexenylacetate, 2-hexenal, 3-hexenol, and hexanal). The nonreactive biogenic NMVOCs (lifetimes of more than one day) covered here include methanol and ethanol, acetone, ethane, and acetic and formic acid. Emissions of alkanes (e.g. ethane, propane, butane) from terrestrial and oceanic natural sources are very low (Lindskog, 1997; Guenther et al., 1994) and are not covered here.

2 BIOGENIC NMVOCs

Isoprene

Isoprene (2-methyl-1,3-butadiene) was first recognized as an emission from plant tissues in the late 1950s (Sanadze, 1991) (Fig. 1). Until recently, it was thought that isoprene was synthesized by the mevalonic acid pathway. It is now known that isoprene is produced in chloroplasts by the glyceraldehyde-3-phosphate pathway, in both an enzyme-dependent (catalyzed by the enzyme isoprene synthase) and nonenzymatic manner (Lichtenthaler et al., 1997). Release to the atmosphere is instantaneous following synthesis, and is the result of simple diffusion of isoprene through cell membranes into the intercellular air spaces and out of pores on the leaf surface, called stomata.

Isoprene emission rates vary among species from 0.1 to 70 μg/g dw h. Not all plants have the ability to produce and emit significant amounts of isoprene. A compilation of species-level isoprene emission screenings from over 800 species of higher plants indicate that, in general, most isoprene emitters are woody deciduous species, although some ferns, vines, and other herbaceous species also emit significant amounts of isoprene (Harley et al., 1999). Phylogenetic patterns are hard to find, as many plant families that contain isoprene emitters, contain nonemitters as well. High isoprene emitting species have been found in the genera *Quercus* (oaks), *Populus* (aspen and poplars), and *Liquidambar* (sweetgum).

Leaf temperature and light intensity are the primary environmental controllers of short-term (hours to days) changes in isoprene production and emission rates from plant foliage (Guenther et al., 1993; Sharkey et al., 1999). Isoprene emissions show

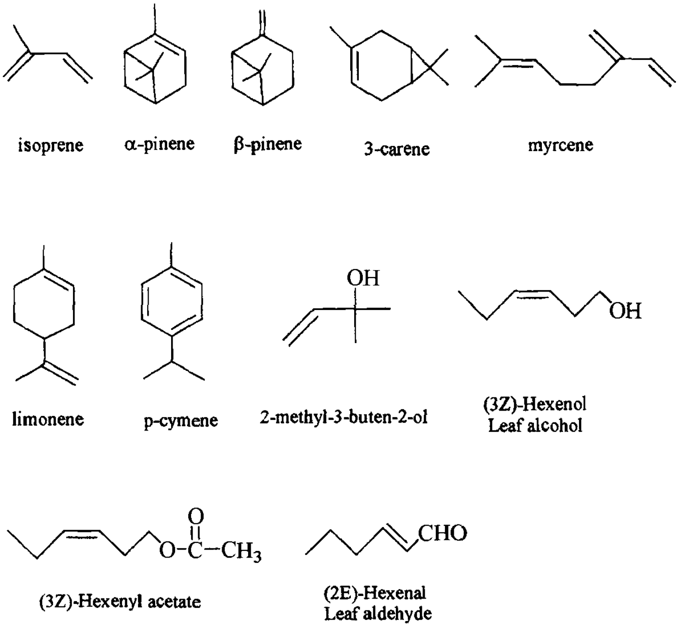


Figure 1 Selected nonmethane hydrocarbons emitted from natural sources.

typical Arrhenius temperature kinetics with species-dependent temperature optima that range from 36 to 40°C (Guenther et al., 1993). Long-term factors that influence isoprene emission rates include light and temperature conditions in which leaves develop, water and nutrient availability, and disease (Monson et al., 1994; Lerdau and Throop, 2000; Anderson et al., 2000; Harley et al., 1994).

Although plants may lose a significant fraction of fixed carbon to isoprene production, it is not known if the production and emission of isoprene serves an adaptive role in plant tissues. One hypothesis is that isoprene protects photosynthetic apparatus against damage from exposure to high temperature and light intensity (Sharkey, 1997). Another possibility is that isoprene scavenges reactive oxidants inside the leaf that can damage plant tissues (Harley et al., 1999).

Over 90% of total isoprene fluxes are from canopy foliage (Guenther, 1999). Emissions from bacteria and fungi in soils and ground cover foliage of mosses and ferns make up the bulk of the remaining natural source strength (9%). Mammals and marine algae and anthropogenic sources (automobile emissions and industrial processes) each contribute less 1% of the global isoprene flux. Although oxidation in the atmosphere is the primary sink for isoprene, microbial consumption in soils is a small net sink as well (Cleveland and Yavitt, 1998).

Typical surface layer mixing ratios of isoprene in the summer range from less than 1 ppbv in a Colorado pine forest (Goldan et al., 1993) to 8 ppbv in the tropical rain forest (Rasmussen and Khalil, 1988) and can be as high as 20 ppbv in rural forests in the southeastern United States (Hagerman et al., 1997). Isoprene ambient concentrations are highest in the summer months and typically show strong diurnal patterns where concentrations sharply increase after sunrise to a maximum in the afternoon and fall to zero at night (Fehsenfeld et al., 1992). This pattern can be explained by the dependence of isoprene emission rates on both temperature and light.

Factors that influence ambient isoprene mixing ratios include emission rates, season, stability of the atmosphere, origin of air masses, and oxidation capacity of the atmosphere (Steinbrecher, 1997). Mixing ratios typically decrease rapidly with altitude since isoprene reacts rapidly in the troposphere with both OH and ozone (e.g., Helmig et al., 1998). Some isoprene has been found in the free troposphere, but only in very low and variable amounts.

Monoterpenes

Monoterpenes ($C_{10}H_{16}$) are a class of structurally diverse compounds produced by over 46 families of flowering plants (e.g., mint, composite, and citrus families), almost all conifers, and some species of liverworts (Banthorpe and Charlwood, 1980; Adam et al., 1996). The accumulation and emission of these compounds directly defends plant tissues against herbivores and pathogens, indirectly defends plant tissues by attracting predators of herbivores, and attracts floral pollinators [reviewed in Langenheim (1994)]. The array of over 1000 different monoterpene structures includes acyclic, monocyclic, and bicyclic forms that can be simple hydrocarbons (e.g., α-pinene, β-pinene, myrcene, δ-3-carene, limonene, *p*-cymene) or oxygenated derivatives (e.g., 1-8 cineole, linalool, camphor) (Fig. 1). The specific monoterpenes produced and emitted from each species is under tight genetic control, and typically only a few monoterpenes dominate the emissions profile of each species.

Monoterpenes, like isoprene, are synthesized in chloroplasts of specialized tissues by the glyceraldehyde-3-phosphate pathway (Lichtenthaler et al., 1997). Two 5C "isoprene" units condense to form a 10C precursor, which is transformed into the myriad of monoterpenes by cyclization reactions catalyzed by the enzymes monoterpene cyclases (Gershenzon and Croteau, 1991). Monoterpenes typically accumulate in storage structures in plant tissues such as glandular trichomes (mints), resin cysts and ducts (conifers), or cavities (eucalypts).

Release to the atmosphere of these stored pools is dependent upon both volatilization and diffusion processes. Plant foliage is the largest source of monoterpene emissions (over 90% of the total global flux) (Guenther, 1999). The remaining fluxes are from woody tissues, buds, cones, and flowers.

In conifers, total foliar monoterpene emission rates vary from 0.01 to 10 μg/g dw h. Emission of monoterpenes is a diffusive process controlled primarily by the influence of needle temperature on monoterpene vapor pressure and monoterpene concentration in the resin ducts and the diffusive resistance of the tissue to volatile losses (Tingey et al., 1991). Other controls over the emission rates of stored pools include leaf age (Lerdau et al., 1997), phenology (Fukui and Doskey, 1998; Cao et al., 1997; Lerdau et al., 1995), herbivory (Litvak and Monson, 1998; Litvak et al., 1999), relative humidity (Dement et al., 1975), foliar moisture (Lamb et al., 1985), and water stress (Yani et al., 1993).

Atmospheric monoterpene concentrations in four rural sites in the southeastern United States ranged from 0.32 to 0.63 ppbv in the summer, and from 0.125 to 0.19 ppbv in the winter (Hagerman et al., 1997). Maximum summer ambient concentrations of total monoterpenes were 0.80 ppbv above a lodgepole pine forest in Colorado (Roberts et al., 1983) and 0.38 above a ponderosa pine plantation in the Sierra Nevada (Lamanna and Goldstein, 1999). Clear diurnal patterns in ambient concentrations of monoterpenes are not as pronounced as those observed for isoprene, but concentrations are often highest at night and lowest during the day (e.g., Lamanna and Goldstein, 1999; Hagerman et al., 1997). Both vertical mixing and chemical loss are important controllers of ambient monoterpene concentrations. In clean air, vertical mixing and dispersion are the most important factors (Hewitt et al., 1995). Because monoterpenes are still emitted at night when atmospheric conditions are relatively stable, concentrations often increase after sunset until the breakdown of these conditions in the morning.

Recent evidence also indicates some species (e.g., Holm oak *Quercus ilex*, Norway spruce *Picea abies*, *Pinus pinea*, and *Acer saccharinum*) produce and emit monoterpenes that do not accumulate in pools (Steinbrecher et al., 1993; Loreto et al., 1996; Staudt et al., 1997). These monoterpenes are emitted at relatively high rates in a light- and temperature-dependent manner very similar to that observed for isoprene and are sensitive to water stress (Bertin and Staudt, 1996), and phenology (Staudt et al., 1997). Many questions remain concerning the physiological and ecological controls over light-dependent production and emission of monoterpenes as well as the specific roles these compounds play in plant tissues.

The aromatic *p*-cymene (1-methyl-4-isopropyl-benzene) is the only volatile arene emitted from vegetation. Trace fluxes of *p*-cymene have been measured from conifers, sage, and eucalyptus but together are equivalent to only 1% of the estimated global monoterpene source strength (Fehsenfeld et al., 1992).

Light Alkenes

Substantial quantities of ethene, propene, and butenes, are released annually from automobiles, industry, and biomass burning (estimated at 10 Tg/yr). However, atmospheric measurements of alkenes made in remote areas that are not impacted by urban or industrial emissions suggest the presence of biogenic sources of light alkenes as well (Lamanna and Goldstein, 1999; Goldstein et al., 1996; Heikes et al., 1996a; Rudolph, 1997).

Emissions from terrestrial ecosystems, particularly plant tissues, make up the bulk of the total global emissions of ethene from natural sources (Sawada and Totsuka, 1986). Ethene functions as a hormone in plant tissues that triggers growth and developmental processes including seed germination, flowering, fruit ripening, senescence, and growth regulation [reviewed in Abeles et al. (1992)]. In addition, ethene is a well-known stress indicator and may play a role in triggering plant defense mechanisms. The amino acid L-methionine is enzymatically converted to ethene in a two-step process involving the intermediate 1-aminocyclo-propane-1-carboxylate (ACC) [reviewed in Fall (1999)]. Production and emission rates of ethene vary with species, tissue type, and phenology and are significantly induced in response to wounding, air pollution, insect and pathogen attack, drought, water-logging, high and low temperatures, and gamma radiation [reviewed in Abeles et al. (1992)]. Global estimates of ethene fluxes from undisturbed canopy foliage are 2 to 4 Tg/yr (Table 1; Rudolph, 1997).

Ethene is also emitted in small quantities from soil microorganisms. Fluxes are correlated with the organic matter content in soil and on a global scale are 2.6 to 3.7 Tg/yr (Rudolph, 1997). Fluxes of ethene, propene, butene, and acetylene have been measured from wetlands but are insignificant on a global scale.

Due to the short lifetimes, measurement difficulty, and wide variety of sources of ethene and propene, ambient concentrations of these compounds are variable. Mean summertime emission rates of ethene, propene, and 1-butene from a deciduous forest in the northeastern United States were 2.6, 1.1, and 0.4×10^{10} molecules/cm^{-2} s, respectively (Goldstein et al., 1996). In this forest, biogenic emissions of propene and 1-butene exceeded the anthropogenic emissions, while biogenic emissions of ethene were equivalent to 50% of emissions from anthropogenic sources. Maximum ambient concentrations above the forest were 0.2, 0.95, and 0.08 ppbv for propene, ethene, and 1-butene, respectively (Goldstein et al., 1996). Lamanna and Goldstein (1999) also observed a local biogenic source for ethene and propene in measurements above a Sierra Nevada ponderosa pine plantation where ambient concentrations of these compounds varied between 0.18 and 0.45 ppbv.

Photochemical degradation of dissolved organic carbon (DOC) released by marine algae results in an estimated global emission rate of 5 Tg/yr for ethene, propene, butenes, and acetylene from ocean surface water [reviewed in Rudolph (1997)]. Fluxes vary seasonally, increase with DOC and light intensity (particularly shorter wavelengths), and depend strongly on DOC, biological activity of the algae, and wind speed (drives the exchange at the air–sea interface) (Ratte et al., 1995). Fluxes are inferred from a combination of atmospheric measurements, seawater measurements, air–sea exchange rates and photochemical models, and encompass large uncertainties. In the remote marine boundary layer and free troposphere over the South Atlantic and western Indian Oceans, ethene and propene concentrations were less than 20 and 6 ppt, respectively (Heikes et al., 1996a).

Alcohols

A C5 alcohol, 2-methyl-3-buten-2-ol (MBO), was recently identified in air samples taken in a Colorado pine forest in concentrations higher than isoprene (up to 3.5 ppbv; Goldan et al., 1993). It is now known that MBO is emitted at relatively high rates from many pine species that grow predominantly in the western United States (up to 70 μg C/g h). Fluxes of MBO, like isoprene, are both light and temperature dependent, suggesting that MBO is emitted immediately following production rather than stored in specialized structures (Harley et al., 1998). Although the production mechanism of MBO in plant tissues is not well known, there is some evidence that it is derived from a 5C precursor of isoprene (Fall, 1999).

Most of the nonreactive other NMVOC flux in Table 1 is contributed by methanol (Guenther et al., 1995). Methanol fluxes measured from leaves are comparable to isoprene and monoterpenes and vary from 0.2 to 40 μg C/h g dry weight (MacDonald and Fall, 1993; Nemecek-Marshall et al., 1995). Emission rates of methanol are highest in young leaves and vary with phenology, leaf damage, and stomatal conductance (Nemecek-Marshall et al., 1995; Fukui and Doskey, 1998). Significant fluxes of methanol and ethanol have also been measured from decaying plant material. Warneke et al. (1999) estimate that globally, emissions from decaying plant material alone could account for 18 to 40 Tg of methanol per year.

Methanol was one of the most abundant VOCs detected above a pine forest canopy in the rural southeastern United States, with summertime mixing ratios of 10 to 20 ppbv (Goldan et al., 1995). Like isoprene, ambient methanol mixing ratios in these studies varied diurnally and peaked in the midafternoon, suggesting that at least in these rural forested areas, methanol was derived primarily from biogenic sources. At a rural site in Colorado, maximum summertime ambient mixing ratios were 6 ppbv (Goldan et al., 1997). Relatively high concentrations of methanol have been detected in the free troposphere, particularly in the northern midlatitudes (0.6 to 0.8 ppbv in northern areas and 0.4 ppbv in southern areas) (Singh et al., 1995). In addition to direct emissions from vegetation, sources of atmospheric methanol include fossil fuel use, biomass burning, and tropospheric production.

Other nonterpenoid alcohols emitted from many agricultural crops, grasses, pastures and forest trees include 3Z-hexenol (leaf alcohol), ethanol, methyl propanol, butanol, and octanol (Isidorov et al., 1985; Arey et al., 1993; Macdonald and Fall, 1993; König et al., 1995; Puxbaum, 1997; Kirstine et al., 1998; Helmig et al., 1999). Production of many of these alcohols, particularly leaf alcohol and ethanol, varies with phenology and is triggered by physical injury and environmental stress (Kirstine et al., 1998; MacDonald et al., 1989). Fluxes of alcohols and other oxygenated VOC's released during the process of crop harvesting may be large enough to have a short-term influence on local air quality (Karl et al., 2001).

Aldehydes and Ketones

Many aldehydes and ketones that are detected in the atmosphere, e.g., acetaldehyde (ethanal), formaldehyde, propanal, butanal, acetone, and butenone, have both anthropogenic and biogenic sources. The dominant sources of these species are fossil fuel combustion, biomass burning, and photochemical oxidation of man-made and natural hydrocarbons, but direct emissions from a variety of forest trees, shrubs, grasses, ferns and mosses occur as well (Isidorov et al., 1985; MacDonald and Fall, 1993; Kotzias et al., 1997; Fukui and Doskey, 1998).

Acetone in plant tissues is produced through fatty acid oxidation [reviewed in Fall (1999)]. Small acetone fluxes have been measured from live plant foliage, decaying vegetation, and seeds and buds of many conifer species, suggesting at least some of the acetone measured in forest canopies and the free troposphere is contributed by natural sources (Fukui and Doskey, 1998; Warneke et al., 1999; Kotzias et al., 1997; MacDonald and Fall, 1993). Warneke et al. (1999) estimated that on a global scale, decaying vegetation emits 6 to 8 Tg of acetone annually.

Acetone is one of the most abundant oxygenated species in the remote atmosphere (Singh et al., 1995). In the free troposphere over the Pacific Ocean, Singh et al. (1995) measured acetone concentrations that range from 0.5 ppbv in the northern latitudes to 0.25 ppbv in the southern latitudes.

Singh et al. (1995) estimated that direct biogenic emissions account for 21% of the total global acetone source. Like methanol, acetone was one of the most abundant VOCs measured above several rural forested areas in Alabama (4 to 7 ppbv; Goldan et al., 1995). Summertime acetone mixing ratios in the Sierra Nevada above a ponderosa pine plantation ranged from 1.5 to 8 ppbv (Lamanna and Goldstein, 1999). At this site, biogenic sources (primarily oxidation of the alcohol MBO) accounted for 45% of the acetone concentrations that exceeded background levels (Goldstein and Schade, 1999). In remote and rural forested regions in Europe, ambient surface acetone concentrations varied between 0.2 and 2.2 ppbv (Solberg et al., 1996). Solberg et al. (1996) observed a strong seasonal dependence of acetone mixing ratios at these sites where summertime maximum acetone concentrations are correlated with high concentrations of biogenic VOC precursors.

The most common aldehydes directly released from the tissues of many plants are 2E-hexenal (also called leaf aldehyde) and other C6 aldehydes from the hexenal family (Hatanaka et al., 1987; Arey et al., 1993; Fukui and Doskey, 1998; Kirstine et al., 1998). In undisturbed tissues, hexenal aldehyde emission rates are small (1.0 to 27 ng/g dw h) (Konig et al., 1995). Hexenals function as antibiotics in plant tissues, however, and emission increases in response to physical wounding, herbivory, and pathogen attack, suggesting that current estimates are low.

Formaldehyde and acetaldehyde are directly emitted from plant foliage at relatively low rates (0.2 to 1 μg/g dw h) (Kesselmeier et al., 1997). In most areas, photochemical oxidation of isoprene and other biogenic VOC precursors emitted from vegetation is a more important biogenic source of these compounds than direct emission (e.g., Fried et al., 1997).

Typical background mixing ratios of formaldehyde are 0.1 to 0.15 ppbv (Heikes et al., 1996b). In a rural site in Colorado, the midday background formaldehyde mixing ratio was 1.17 (Fried et al., 1997). In rural areas in Europe, Solberg et al. (1996) observed a seasonal pattern in formaldehyde ambient mixing ratios similar to acetone, where concentrations are highest in the summer (1.3 to 5.9 ppbv), compared to the rest of the year (0.4 to 2.4 ppbv). Arlander et al. (1990) report a latitudinal distribution of formaldehyde from measurements taken over the Pacific Ocean. Maximum mixing ratios (between 0.6 and 0.8 ppbv) of formaldehyde during this cruise were seen between 20°N and the equator, reflecting the latitudinal distribution of both anthropogenic and biogenic alkene precursors.

Organic Acids

Atmospheric mixing ratios of formic and acetic acid typically range from 0.02 to 1.9 in remote and marine locations to 1 to 16 ppbv in urban polluted areas [reviewed in Khare et al. (1999)]. Sources of these organic acids include fossil fuel combustion, biomass burning, direct emissions from formicine ants, soils and plant foliage, and photochemical production in the atmosphere through isoprene and monoterpene oxidation. Direct emissions of both acids have been measured from the European species *Quercus ilex* and *Pinus pinea*, tropical trees in the Amazon, and savanna soils (Kesselmeier et al., 1997; Talbot et al., 1990; Sanhueza and Andreae, 1991). Though the precise biosynthetic mechanisms of organic acids in plant tissues is unknown, acetic acid is formed through lipid metabolism, and formic acid is a by-product of carbohydrate and C1 metabolism [reviewed in Fall (1999)]. In soils, microbial activity is the likely organic acid source.

Vertical profiles and observed seasonal, diurnal, and latitudinal patterns of organic acid concentrations in both precipitation and gas-phase measurements support a significant biogenic source of these acids in rural midlatitude continental, tropical continental, and marine locations (Khare et al., 1999). For example, mixing ratios over the Amazon Basin, and temperate forests in eastern United States were higher during the growing season and the afternoon than in the winter and at night or in the early morning (Keene and Galloway, 1988; Talbot et al., 1988, 1990). Although photochemical production of precursors emitted from vegetation is considered to be the dominant biogenic source of organic acids, direct emission by soils and vegetation can be important in rural areas in the eastern United States and in the Amazon rainforest (Andreae et al., 1988; Talbot et al., 1990, 1995). Formic and acetic acid budgets in marine atmospheres suggest the presence of a natural source as well (Arlander et al., 1990; Heikes et al., 1996a).

3 REGIONAL AND GLOBAL DISTRIBUTION OF BIOGENIC NMVOC EMISSIONS

To understand the impact biogenic NMVOCs have on tropospheric chemistry, reliable emission estimates at local, regional, and global scales are necessary. Flux measurements made at a variety of scales are the primary means of both developing and evaluating these emission estimates. The techniques used to measure these fluxes are reviewed in Guenther et al. (1996). Enclosure methods are used to estimate fluxes on small scales including from a single leaf, branch, or whole tree. These measurements are a particularly good way to quantify species-specific basal emission rates (or the capacity to emit NMVOCs under a standard set of environmental conditions). Tower-based micrometeorological techniques are used to directly measure canopy-scale fluxes of NMVOCs on diurnal, seasonal, and annual time scales. These techniques include eddy covariance, relaxed eddy accumulation (REA), surface layer gradient, and tracer methods. Finally, sampling systems on tethered balloons and aircraft are used to construct vertical mixing ratio profiles and calculate surface fluxes on scales of tens to hundreds of kilometers using eddy accumulation, REA, mixed-layer mass balance, and mixed-layer gradient methods.

To construct inventories, basal emission rates from a wide range of vegetation classes are modified by instantaneous changes in both temperature and light intensity using algorithms developed by Guenther et al. (1993), multiplied by estimates of foliar density of each vegetation class, and aggregated to give flux estimates on regional and global scales (Lamb et al., 1987; Guenther et al., 1995; Guenther, 1997). Large uncertainties are associated with these inventories, however, due to gaps in our knowledge of (1) the contribution of nonfoliar emissions, (2) physiological and ecological controls over emissions from plants, (3) specific emission factors from a wider variety of plants and ecosystems, particularly of nonterpenoid NMVOCs, and (4) detailed data on coverage of ecosystem type, foliage density, surface temperatures, and radiation properties (Steinbrecher, 1997).

These inventories are useful for identifying where, on a regional basis, biogenic contributions to total NMVOC fluxes are particularly relevant due to vegetation type, foliar density, and ambient temperature patterns. For example, in urban areas such as Los Angeles, biogenic NMVOCs contribute a relatively small fraction to the total VOC emissions (Benjamin et al., 1997). In Atlanta and remote rural areas, however, biogenic sources, during the summer months especially, can dominate the total VOC emission profile (Geron et al., 1995; Hagerman et al., 1997). In North America as a whole, and in Norway, Sweden, and Finland, biogenic emissions of VOCs exceed anthropogenic emissions (Guenther et al., 1995; Simpson et al., 1995). In Italy, biogenic emissions account for 50% of the total VOCs emitted (Simpson et al., 1995).

On a global scale, Guenther et al. (1995) derived estimates for emissions of isoprene (420 Tg C/yr), monoterpenes (130 Tg C/yr), and other reactive VOCs (280 Tg C/yr). As expected due to the influence of light intensity and temperature on emissions, biogenic fluxes show seasonal as well as latitudinal differences (Guenther et al., 1995; Guenther, 1999). Drought deciduous forests and savannas in the tropics contributed half of all the global VOCs from biogenic sources in this estimate. Other woodlands, crops, and shrublands contributed 10 to 20% of these fluxes. Crops, in particular, were high emitters of VOCs other than isoprene and monoterpenes.

Relative to isoprene, trends in the global distribution of monoterpene and other NMVOC fluxes are hard to find. The high reactivity, spatial and temporal variability in source strengths, and uncertainties in reliable identification and quantification of these species have contributed to large variability in observed ambient mixing ratios. Thus, although measurements of these species have been made, considerable work is necessary to truly understand the global distribution of these reactive VOCs in the atmosphere.

4 SUMMARY AND CONCLUSIONS

A variety of nonmethane hydrocarbons are released from natural sources, particularly plant foliage, in quantities sufficient to alter production of tropospheric ozone, organic acids and nitrates, PAN, OH, and CO. Isoprene and monoterpenes together dominate NMVOC fluxes from many species, ecosystems, regions, and on a global scale. Although the mechanisms of production, emission, and degradation in the atmosphere are fairly well known for isoprene and monoterpenes, uncertainties such as why only certain plants produce these compounds and detailed distributions of the vegetation sources remain.

Uncertainties are largest for VOCs other than isoprene and monoterpenes. Nonterpenoid hydrocarbons contribute an estimated 45% of the total biogenic VOC global fluxes. To make more reliable estimates of the source strength and atmospheric impacts of these hydrocarbons, a better understanding of the biological and ecological factors that control spatial and temporal variability in these fluxes is needed.

Current NMVOC inventories rely on empirical models based only on the response of emissions to temperature, light, and foliar density. Given that emissions of isoprene, monoterpenes, and many of the other VOCs are influenced by a whole suite of physiological and ecological factors, using these inventories to predict emissions in response to disturbances, land-use change, and/or climate change is risky. An important aspect of future biogenic VOC research is to incorporate a more mechanistic understanding of VOC production and emission into emission inventories (Monson et al., 1995). In this way inventories will be able to extrapolate emission rates and the impacts of these emissions on atmospheric chemistry across complex ecological gradients in both space and time.

REFERENCES

Abeles, F. B., P. W. Morgan, and M. E. Saltveit, *Ethylene in Plant Biology*, 2nd ed., Academic, New York, 1992.

Adam, K.-P., J. Crock, and R. Croteau, Partial purification and characterization of a monoterpene cyclase, limonene synthase, from the liverwort *Ricciocarpos natans*, *Arch. Biochem. Biophys. 332*, 352–356, 1996.

Anderson, L. J., P. C. Harley, R. K. Monson, and R. B. Jackson, Reduction of isoprene emissions from live oak (Quercus fusiformis) with oak wilt, *Tree Phys., 20*, 1199–1203, 2000.

Andreae, M. O., R. W. Talbot, T. W. Andreae, and R. C. Harriss, Formic and acetic acid over the Central Amazon region, Brazil. 1. Dry season, *J. Geophys. Res.*, *93*, 1616–1624, 1988.

Andreae, M. O., and P. J. Crutzen, Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry, *Science*, *276*, 1052–1058, 1997.

Arey, J., A. M. Winer, R. Atkinson, S. M. Aschmann, W. D. Long, and C. L. Morrison, The emission of (*Z*)-3-hexen-1-ol, (*Z*)-3-hexenylaceteate and other oxygenated hydrocarbons from agricultural plant species, *Atmos. Environ.*, *25A*, 1063–1075, 1993.

Arlander, D. W., D. R. Cronn, J. C. Farmer, F. A. Menzi, and H. H. Westberg, Gaseous oxygenated hydrocarbons in the remote marine troposphere, *J. Geophys. Res.*, *95*, 16391–16403, 1990.

Banthorpe, D., and V. Charlwood, The terpenoids, in E. Bell, and V. Charlwood (Eds.), *Encyclopedia of Plant Physiology*, Springer-Verlag, Berlin, 1980, pp. 185–220.

Benjamin, M. T., M. Sudol, D. Vorsatz, and A. M. Winer, A spatially and temporally resolved biogenic hydrocarbon emissions inventory for the California South coast air basin, *Atmos. Environ.*, *31*, 3087–3100, 1997.

Bertin, N., and M. Staudt, Effect of water stress on monoterpene emissions from young potted Holm oak (*Quercus ilex* L.) trees, *Oecologia*, *107*, 456–462, 1996.

Bonsang, B., and C. Boissard, Global distribution of reactive hydrocarbons in the atmosphere, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 43–97.

Cao, X. L., C. Boissard, A. J. Juan, C. N. Hewitt, and M. Gallagher, Biogenic emissions of volatile organic compounds from gorse (*Ulex europaeus*): Diurnal emission fluxes at Kelling Heath, England, *J. Geophys. Res.*, *102*, 18903–18915, 1997.

Chameides, W. L., R. W. Lindsay, J. Richardson, and C. S. Kiang, The role of biogenic hydrocarbons in urban photochemical smog: Atlanta as a case study, *Science*, *241*, 1–10, 1988.

Cleveland, C. C., and J. B. Yavitt, Microbial consumption of atmopheric isoprene in a temperate forest soil, *Appl. Environ. Microbiol.*, *64*, 172–177, 1998.

Conrad, R., Soil microbial processes and the cycling of atmospheric trace gases, *Phil. Trans. R. Soc. Lond. Ser. A.*, *351*, 219–230, 1995.

Dement, W. A., B. J. Tyson, and H. A. Mooney, Mechanism of monoterpene volatilization in *Salvia mellifera*, *Phytochemistry*, *14*, 2555–2557, 1975.

Fall, R., Biogenic emissions of volatile organic compounds from higher plants, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 43–97.

Fehsenfeld, F., J. Calvert, R. Fall, P. Goldan, A. Guenther, C. N. Hewitt, B. Lamb, S. Liu, M. Trainer, H. Westberg, and P. Zimmerman, Emissions of volatile organic compounds from vegetation and the implications for atmospheric chemistry, *Global Biogeochem. Cycles*, *6*, 389–430, 1992.

Fried, A., S. Mckeen, S. Sewell, J. Harder, B. Henry, P. Goldan, W. Kuster, E. Williams, K. Baumann, R. Shetter, and C. Cantrell, Photochemistry of formaldehyde during the 1993 Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6283–6296, 1997.

Fukui, Y., and P. V. Doskey, Air-surface exchange of nonmethane organic compounds at a grassland site: Seasonal variations and stressed emissions, *J. Geophys. Res.*, *103*, 13153–13168, 1998.

Geron, C., T. Pierce, and A. Guenther, Reassessment of biogenic volatile organic compound emissions in the Atlanta area, *Atmos. Environ.*, *29*, 1569–1578, 1995.

Gershenzon, J., and R. Croteau, Terpenoids, in G. A. Rosenthal and M. R. Gerenbaum (Eds.), *Herbivores. Their Interactions with Secondary Plant Metabolites*, 2nd ed., Vol. 1: *The Chemical Participants*, Academic, San Diego, 1991, pp. 165–219.

Goldan, P. D., W. C. Kuster, and F. C. Fehsenfeld, Nonmethane hydrocarbon measurements during the Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6315–6324, 1997.

Goldan, P. D., W. C. Kuster, F. C. Fehsenfeld, and S. A. Montzka, The observation of a C5 alcohol in a North American pine forest, *Geophys. Res. Lett.*, *20*, 1039–1042, 1993.

Goldan, P. D., W. C. Kuster, F. C. Fehsenfeld, and S. A. Montzka, Hydrocarbon measurements in the southeastern United States: The Rural Oxidants in the Southern Environment (ROSE) Program 1990, 1995.

Goldstein, A. H., S. M. Fan, M. L. Goulden, J. W. Munger, and S. C. Wofsy, Emissions of ethene, propene, and 1-butene by a midlatitude forest, *J. Geophys. Res.*, *101*, 9149–9157, 1996.

Goldstein, A. H., and G. W. Schade, Quantifying the biogenic and anthropogenic contributions to high concentrations of acetone observed in the Sierra Nevada Mountains (CA), paper presented at American Geophysical Union Fall Meeting, San Francisco, CA, December 13–17, 1999.

Graedel, T. E., *Chemical Compounds in the Atmosphere*, Academic, New York, 1979.

Guenther, A., Seasonal and spatial variations in natural volatile organic compound emissions, *Ecol. Appl.*, *7*, 34–45, 1997.

Guenther, A., Modeling biogenic VOC emissions to the atmosphere, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 97–118.

Guenther, A., P. Zimmerman, P. Harley, R. Monson, and R. Fall, Isoprene and monoterpene emission rate variability: Model evaluation and sensitivity analysis, *J. Geophys. Res.*, *98*, 12609–12617, 1993.

Guenther, A., P. Zimmerman, and M. Wildermuth, Natural volatile organic compound emission rate estimates for U.S. woodland landscapes, *Atmos. Environ.*, *28*, 1197–1210, 1994.

Guenther, A., C. N. Hewitt, D. Erickson, R. Fall, C. Geron, T. Gradel, P. Harley, L. Klinger, M. Lerdau, W. A. McKay, T. Pierce, B. Scholes, R. Steinbrecher, R. Tallamraju, J. Taylor, and P. Zimmerman, A global model of natural volatile organic compound emissions, *J. Geophys. Res.*, *100*, 8873–8892, 1995.

Guenther, A., W. Baugh, K. David, G. Hamptom, P. Harley, L. Klinger, P. Zimmerman, E. Allwine, S. Dilts, B. Lamb, H. Westberg, D. Baldocchi, C. Geron, and T. Pierce, Isoprene fluxes measured by enclosure, relaxed eddy accumulation, surface-layer gradient, mixed-layer gradient, and mass balance techniques, *J. Geophys. Res.*, *101*, 18555–18568, 1996.

Haagen-Smit, A. J., Chemistry and physiology of Los Angeles smog, *Ind. Eng. Chem.*, *44*, 1342–1345, 1952.

Hagerman, L. M., V. P. Aneja, and W. A. Lonneman, Characterization of nonmethane hydrocarbons in the rural Southeast United States, *Atmos. Environ.*, *31*, 4017–4038, 1997.

Harley, P. C., M. E. Litvak, T. D. Sharkey, and R. K. Monson, Isoprene emissions from velvelt bean leaves – interactions among nitrogen availability, growth photon flux density and leaf development, *Plant Phys.*, *105*, 279–285, 1994.

Harley, P., V. Fridd-Stroud, J. Greenberg, A. Guenther, and P. Vasconcellos, Emission of 2-methyl-3-buten-2-ol by pines: A potentially large natural source of reactive carbon to the atmosphere, *J. Geophys. Res.*, *103*, 25479–25486, 1998.

Harley, P., R. K. Monson, and M. T. Lerdau, Ecological and evolutionary aspects of isoprene emission from plants, *Oecologia*, *118*, 109–123, 1999.

Hatanaka, A., T. Kajiwara, and J. Sekiya, Biosynthetic pathways for C6-aldehydes formation from linolenic acid in greed leaves, *Chem. Phys. Lipids*, *44*, 341–361, 1987.

Heikes, B., M. Lee, D. Jacob, R. Talbot, J. Bradshaw, H. Singh, D. Blake, B. Anderson, H. Fuelberg, and A. M. Thompson, Ozone, hydroperoxides, oxides of nitrogen, and hydrocarbon budgets in the marine boundary layer over the South Atlantic, *J. Geophys. Res.*, *101*, 24221–24234, 1996a.

Heikes, B., B. McCuly, X. Zhou, Y.-N. Lee, K. I. Mopper, X. Chen, G. Mackay, D. Karecki, H. Schiff, T. Campos, and E. Atlas, Formaldehyde methods comparison in the remote lower troposphere during the Mauna Loa Photochemistry Experiment 2, *101*, 14741–14755, 1996b.

Helmig, D., L. F. Klinger, A. Guenthe, L. Vierling, C. Geron and P. Zimmerman, Biogenic volatile organic compound emissions (BVOCs) I. Identifications from three continental sites in the US, *Chemosphere, 38*, 2163–2187, 1999.

Helmig, D., B. Balsley, K. Davis, L. R. Kuck, M. Jensen, J. Bognar, T. Smith, Jr., R. Vasquez Arrieta, R. Rodriguez, and J. W. Birks, Vertical profiling and determination of landscape fluxes of biogenic nonmethane hydrocarbons within the planetary boundary layer in the Peruvian Amazon, *J. Geophys. Res.*, *103*, 25519–25532, 1998.

Isidorov, V. A., I. G. Zenkevich, and B. V. Ioffe, Volatile organic compounds in the atmosphere of forests, *Atmos. Environ.*, *19*, 1–8, 1985.

Jacob, D. J., and S. C. Wofsy, Photochemistry of biogenic emissions over the Amazon forest, *J. Geophys. Res.*, *93*(D2), 1477–1486, 1988.

Karl, T., A. Guenther, C. Lindinger, A. Jordan, R. Fall, and W. Lindinger. Eddy covariance measurements of oxygenated volatile organic compound fluxes from crop harvesting using a redesigned proton-transfer-reaction mass spectrometer. *J. Geophys. Res.*, *106*, 24157–24167, 2001.

Kasting, J. F., and H. B. Singh, Nonmethane hydrocarbons in the troposphere – impact on the odd hydrogen and odd nitrogen chemistry, *J. Geophys. Res.*, *91*, 3239–3256, 1986.

Keene, W. C., and J. N. Galloway, The biogeochemical cycling of formic and acetic acids through the troposphere: An overview of current understanding, *Tellus*, *40B*, 322–334, 1988.

Kesselmeier, J., K. Bode, U. Hofmann, H. Muller, L. Schaefer, A. Wolf, P. Ciccioli, E. Brancaleoni, A. Cecinato, M. Frattoni, P. Foster, C. Ferrari, V. Jacob, J. L. Fugit, L. Dutaur, V. Simon, and L. Torres, Emission of short chained organic acids, aldehydes and monoterpenes from *Quercus ilex* L. and *Pinus pinea* L. in relation to physiological activities, carbon budget and emission algorithms, *Atmos. Environ.*, *31*(SI), 119–133, 1997.

Khare, P., N. Kumar, K. M. Kumari, and S. S. Srivastava, Atmospheric formic and acetic acids: An overview, *Rev. Geophys.*, *37*, 227–248, 1999.

Kirstine, W., I. Galbally, Y. Yuerong, and M. Hooper, Emissions of volatile organic compounds (primarily oxygenated species) from pasture, *J. Geophys. Res.*, *103*, 10605–10619, 1998.

König, G., M. Brunda, H. Puxbaum, C. N. Hewitt, and S. C. Duckham, Relative contribution of oxygenated hydrocarbons to the total biogenic VOC emissions of selected mid-European agricultural and natural plant species, *Atmos. Environ.*, *29*, 861–874, 1995.

Kotzias, D., C. Konidari, and C. Sparta, Volatile carbonyl compounds of biogenic origin—Emission and concentration in the atmosphere, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 67–78.

Lamanna, M. S., and A. H. Goldstein, In situ measurements of C2–C10 volatile organic compounds above a Sierra Nevada ponderosa pine plantation, *J. Geophys. Res.*, *104*, 21247–21262, 1999.

Lamb, B., A. Guenther, D. Gay, and H. Westberg, A national inventory of biogenic hydrocarbon emissions, *Atmos. Environ.*, *21*, 1695–1705, 1987.

Lamb, B., H. Westberg, and G. Allwine, Biogenic hydrocarbon emissions from deciduous and coniferous trees in the United States, *J. Geophys. Res.*, *90*, 2380–2390, 1985.

Langenheim, J. H., Higher plant terpenoids: A phytocentric overview of their ecological roles, *J. Chem. Ecol.*, *20*, 1223–1280, 1994.

Lerdau, M., and H. L. Throop, Sources of variability in isoprene emission and photosynthesis in two species of tropical wet forest trees, *Biotropica*, *32*, 670–676, 2000.

Lerdau, M., M. Litvak, P. Palmer, and R. Monson, Controls over monoterpene emissions from boreal forest conifers, *Tree Phys.*, *17*, 491–499, 1997.

Lerdau, M., P. Matson, R. Fall, and R. Monson, Ecological controls over monoterpene emission from Douglas-fir *Pseudotusga menziesii*, *Ecology*, *76*, 2640–2647, 1995.

Lichtenthaler, H. K., J. Schwender, A. Disch, and M. Rohmer, Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway, *FEBS Lett.*, *400*, 271–274, 1997.

Lindskog, A., The influence of biosphere on the budgets of VOC: Ethane, propane, *n*-butane and *i*-butane, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 45–52.

Litvak, M. L., S. Madronich, and R. K. Monson, Herbivore-induced monoterpene emissions from coniferous forests: Potential impact on local tropospheric chemistry, *Ecol. Appl.*, *9*, 1147–1159, 1999.

Litvak, M. E., and R. K. Monson, Patterns of constitutive and induced monoterpene production in conifer needles in relation to insect herbivory, *Oecologia*, *118*, 531–540, 1998.

Loreto, F., P. Ciccioli, A. Cecinato, E. Brancaleoni, M. Frattoni, C. Fabozzi, and D. Tricoli, Evidence of the photosynthetic origin of monoterpenes emitted by *Quercus ilex* L. leaves by C^{13} labeling, *Plant Physiol.*, *110*, 1317–1322, 1996.

MacDonald, R. C., and R. Fall, Detection of substantial emissions of methanol from plants to the atmosphere, *Atmos. Environ.*, *27A*, 1709–1713, 1993.

MacDonald, R. C., T. W. Kimmerer, and M. Razzaghi, Aerobic ethanol production by leaves: Evidence for air pollution stress in trees in the Ohio River Valley, USA, *Environ. Pollut.*, *62*, 337–351, 1989.

Monson, R. K., P. C. Harley, M. E. Litvak, M. Wildermuth, A. B. Guenther, P. R. Zimmerman, and R. Fall, Environmental and developmental controls over the seasonal pattern of isoprene emission from aspen leaves. *Oecologia, 99*, 260–270, 1994.

Monson, R. K., M. T. Lerdau, T. D. Sharkey, D. S. Schimel, and R. Fall, Biological aspects of constructing volatile organic compound emission inventories, *Atmos. Env.*, *29*, 2989–3002, 1995.

Müller, J. F., Geographical distribution and seasonal variation of surface emissions and deposition velocities of atmospheric trace gases, *J. Geophys. Res.*, *97*, 3787–3804, 1992.

Nemecek-Marhsall, M., R. C. MacDonald, J. F. Franzen, C. L. Wojciechowski, and R. Fall, Methanol emission from leaves, *Plant Phys.*, *108*, 1359–1368, 1995.

Pandis, S. N., S. E. Paulson, J. H. Seinfeld, and R. C. Flagan, Aerosol formation in the photooxidation of isoprene and *β*-pinene, *Atmos. Environ.*, *Part A*, *26*, 2269–2282, 1991.

Puxbaum, H., Biogenic emissions of alcohols, ester, ether and higher aldehydes, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 79–99.

Rasmussen, R. A., and M. A. Khalil, Isoprene over the Amazon basin, *J. Geophys. Res.*, *93*, 1417–1421, 1988.

Ratte, M., C. Plassdulmer, R. Koppmann, and J. Rudolph, Horizontal and vertical profiles of light hydrocarbons in sea water related to biological, chemical and physical parameters, *Tellus, Series B*, *47*, 607–623, 1995.

Roberts, J. M., F. C. Fehsenfeld, D. L. Albritton, and R. E. Sievers, Measurement of monoterpene hydrocarbons at Niwot Ridge, Colorado, *J. Geophys. Res.*, *88*, 10667–10678, 1983.

Roselle, S. J., T. E. Pierce, and K. L. Schere, The sensitivity of regional ozone modeling to biogenic hydrocarbons, *J. Geophys. Res.*, *96*, 7371–7394, 1991.

Rudolph, J., Biogenic sources of atmospheric alkenes and acetylene, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 53–65.

Sanadze, G. A., Isoprene effect–light dependent emission of isoprene by green parts of plants, in T. D. Sharkey, E. A. Holland, and H. A. Mooney (Eds.), *Trace Gas emissions by Plants*, Academic, San Diego, 1991, pp. 135–152.

Sanhueza, E., and M. O. Andreae, Emission of formic and acetic acids from tropical savanna soils, *Geophys. Res. Lett.*, *18*, 1707–1710, 1991.

Sawada, S., and T. Totsuka, Natural and anthropogenic sources and fate of atmospheric ethylene, *Atmos. Environ.*, *20*, 821–832, 1986.

Sharkey, T. D., Emission of low molecular mass hydrocarbons from plants, *Trends Plant Sci,*, *1*, 78–82, 1996.

Sharkey, T. D., Isoprene production in trees, in H. Rennenberg, W. Eschrich, and H. Ziegler (Eds.), *Trees—Contributions to Modern Tree Physiology*, Backhuys, The Netherlands, 1997, pp. 109–118.

Sharkey, T. D., E. L. Singsaas, M. T. Lerdau, and C. D. Geron, Weather effects on isoprene emission capacity and applications in emissions algorithms, *Ecol. Appl.*, *9*, 1132–1137, 1999.

Sillman, S., and F. J. Samson, Impact of temperature on oxidant photochemistry in urban, polluted rural and remote environments, *J. Geophys. Res.*, *100*, 11497–11508, 1995.

Simpson, D., A. Guenther, C. N. Hewitt, and R. Steinbrecher, Biogenic emissions in Europe. I. Estimates and uncertainties, *J. Geophys. Res.*, *100*, 506–512, 1995.

Singh, H. B., M. Kanakidou, P. J. Crutzen, and D. J. Jacob, High concentrations and photochemical fate of oxygenated hydrocarbons in the global troposphere, *Nature*, *378*, 50–54, 1995.

Singh, H. B., and P. R. Zimmerman, Atmopheric distribution and sources of nonmethane hydrocarbons, in J. O. Nriagu (Ed.), *Gaseous Pollutants: Characterization and Cycling*, Wiley-Interscience, New York, 1992, pp. 177–235.

Solberg, S., C. Dye, N. Schmidbauer, A. Herzog, and R. Gehrig, Carbonyls and nonmethane hydrocarbons at rural European sites from the Mediterranean to the arctic, *J. Atmos. Chem.*, *25*, 33–66, 1996.

Staudt, M., N. Bertin, U. Hansen, G. Seufert, P. Ciccioli, P. Foster, B. Frenzel, J.-L. Fugit, and L. Torres, The BEMA-project: Seasonal and diurnal patterns of monoterpene emissions from *Pinus pinea* (L.) measured under field conditions, *Atmos. Environ.*, *31*, 145–156, 1997.

Steinbrecher, R., Isoprene: Production by plants and ecosystem-level estimates, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 101–114.

Steinbrecher, R., W. Schürmann, A.-M. Schreiner, and H. Ziegler, Terpenoid emissions from common oak (*Quercus robur* L.) and Norway spruce (*Picea abies* L. Karst.), in J. Slanina, G. Angeletti, and S. Beilke (Eds.), *Proceedings of the Joint CEC/BIATEX Workshop on the General Assessment of Biogenic Emissions and Deposition of Nitrogen Compounds, Sulfur Compounds and Oxidants in Europe*, CEC Environ. Res. Progr. Report 47, 1993, pp. 251–261.

Talbot, R. W., B. W. Mosher, B. G. Heikes, D. J. Jacob, J. W. Munger, B. C. Daube, W. C. Keene, J. R. Maben, and R. S. Artz, Carboxylic-acids in the rural continental atmosphere over the eastern United States during the Shenandoah cloud and photochemistry experiment, *J. Geophys. Res., 100*, 9335–9343, 1995.

Talbot, R. W., M. O. Andreae, H. Berresheim, D. J. Jacob, and K. M. Beecher, Sources and sinks of formic, acetic, and pyruvic acids over central Amazonia, 2, Wet season, *J. Geophys. Res.*, *95*, 16799–16811, 1990.

Talbot, R. W., K. M. Beecher, R. C. Harriss, and W. R. Cofer III, Atmospheric geochemistry of formic and acetic acids at a midlatitude temperate site, *J. Geophys. Res.*, *93*, 1638–1652, 1988.

Talbot, R. W., B. W. Mosher, B. G. Heikes, D. J. Jacob, J. W. Munger, B. C. Daube, W. C. Keene, J. R. Maben, and R. S. Artz, Carboxylic acids in the rural continental atmosphere over the eastern United States during the Shenandoah Cloud and Photochemistry Experiment.

Tingey, D. T., D. P. Turner, and J. A. Weber, Factors controlling the emissions of monoterpenes and other volatile organic compounds, in T. D. Sharkey, E. A. Holland, and H. Mooney (Eds.), *Trace Gas Emissions from Plants*, Academic, San Diego, 1991, pp. 93–119.

Trainer, M., E. J. Williams, D. D. Parrish, M. P. Buhr, E. J. Allwine, H. Westberg, F. C. Fehsenfeld, and S. C. Liu, Models and observations of the impact of natural hydrocarbons on rural ozone.

Yani, A., G. Pauly, M. Faye, F. Salin, and M. Gleizes, The effect of a long term water stress on the metabolism and emission of terpenes of the foliage of *Cupressus sempervirens*, *Plant Cell Environ.*, *16*, 975–981, 1993.

Warneke, C., T. Karl, H. Judmaier, A. Hansel, A. Jordan, W. Lindinger, and P. Crutzen, Acetone, methanol and other partially oxidized volatile organic emissions from dead plant matter by abiological processes: Significance for atmospheric HO_x chemistry, *Global Biogeochem. Cycles*, *13*, 9–17, 1999.

Winer, A., J. Arey, R. Atkinson, S. Aschman, W. Long, L. Morrison, and D. Olszyk, Emission rates of organics from vegetation in California's Central Valley, *Atmos. Environ.*, *26A*, 2647–2659, 1992.

CHAPTER 8

ATMOSPHERIC SULFUR

D. D. DAVIS, G. CHEN, AND M. CHIN

1 INTRODUCTION

The focus of this chapter is that of providing the reader with an overview of atmospheric sulfur. It will address the issues of where sulfur comes from, how it is processed, and how it gets returned to the planetary surface. It will also endeavor to show how sulfur, during its atmospheric cycle, plays a significant role in helping to maintain a stable global environment.

Sulfur is an element that is essential to life on this planet. Living organisms at nearly all levels of sophistication ingest sulfur from their environment, mainly in the form of sulfate or amino acid sulfur. But living organisms not only ingest sulfur, they also have a decisive impact on the chemical forms and total burden that is found in the atmosphere. During the process known as *assimilatory sulfate reduction*, microorganisms and plants use sulfate to build sulfur-containing proteins for purposes of storing energy and to support cell growth. During food digestion, animals are able to generate energy from the catabolism of these proteins, breaking them down to their chemical building blocks, the amino acids. The further breakdown of these compounds results in the release of volatile sulfur back to the environment.

Some microorganisms living in anoxic environments, such as tidal flats, obtain energy from using sulfate as an electron acceptor instead of O_2. This process is called *disimilatory sulfate reduction*. Hydrogen sulfide (H_2S) released during this process often combines with iron minerals to form pyrite, FeS, resulting in its incorporation into sediment layers. Alternatively, the H_2S may react with buried organic matter, thus forming a source of sulfur in fossil oil and coal deposits. In general, the turnover of sulfur in dissimilatory processes is several orders of magnitude faster than in assimilatory processes. The biological sulfur cycle is therefore mainly controlled by anaerobic sulfate reducing bacteria. [For further details on the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

assimilatory and dissimilatory biological processes, the reader is referred to reviews by Krouse and McCready (1979) and Andreae and Jaeschke (1992).]

Sulfur, having six valence electrons, has the potential for existing in the atmosphere in a wide range of oxidation states, ranging from -2 to $+6$, with the most common states being -2, $+4$, and $+6$. In most remote global locations, atmospheric sulfur is found at concentration levels of only 1 ppbv (part-per billion by volume) or less. However, for continental regions, particularly those experiencing significant industrial development, concentrations can reach upwards of 100 to 200 ppbv. Similar levels can be found in regions under the influence of active volcanoes. This trend in global sulfur levels is a strong reflection of the distribution of the major sources of sulfur as illustrated in Figure 1. From this abbreviated picture of the atmospheric sulfur cycle, the most critical members of the atmospheric sulfur family are identified as dimethyl sulfide (DMS), sulfur dioxide (SO_2), sulfuric acid (H_2SO_4), and aerosol sulfate ($SO_4{}^{2-}$). The latter species is typically found in the form of condensation nuclei (CN) or cloud condensation nuclei (CCN). Two of the major primary sulfur sources are shown as SO_2, emitted from the burning of large quantities of fossil fuel or, alternatively, from volcanoes, and DMS, which predominantly is released from the world's oceans. This source is obviously dispersed over a much larger global surface area than is primary SO_2, leading to much lower concentration levels of sulfur over remote regions. In the latter context, shown also in Figure 1, is the most recently identified remote sulfur source, emissions from ships (Corbett and Fischbeck, 1997; Corbett et al., 1999). This new source, however, is significantly smaller than the three previously discussed.

Among the central points revealed in Figure 1 is the fact that the atmosphere can be viewed as a large oxidizing chemical reactor in which sulfur, emitted from Earth's surface, enters the atmosphere in a chemically reduced oxidation state (typically -2 and $+4$) is oxidized to the $+6$ state, and then in ionic form (i.e., higher solubility) is returned to the biosphere, thus closing the cycle. The processes responsible for oxidizing sulfur are shown as occurring by both gas phase as well as heterogeneous reactions. Once in the $+6$ oxidation state, this sets the stage for the final contribution from atmospheric sulfur toward maintaining a stable global environment, namely, its impact on the planetary radiation budget. As shown in Figure 1 atmospheric aerosols, which are predominately composed of sulfur, can have a significant impact on the planet's climate via their influence on direct scattering of incoming solar radiation and by their controlling the radiative characteristics and formation rates of clouds (Charlson et al., 1992).

In Section 2 of this chapter, we will expand on the source inventories for DMS and SO_2 as well as present inventories for several less important primary sulfur source species, including those for H_2S, carbony sulfide (OCS), and carbon disulfide (CS_2). Of particular significance will be the hemispheric distributions of these collective sulfur sources and how they manifest themselves in observed concentration levels. Section 3 will also explore in greater detail the oxidation processes responsible for converting the dominant sulfur species (i.e., DMS and SO_2) into forms that result in their removal. Section 4 will combine the source inventory data presented in Section 3 and the chemical transformation information discussed in Section 2 in exploring global distributions of SO_2 and $SO_4{}^{2-}$. Finally, in Section 5,

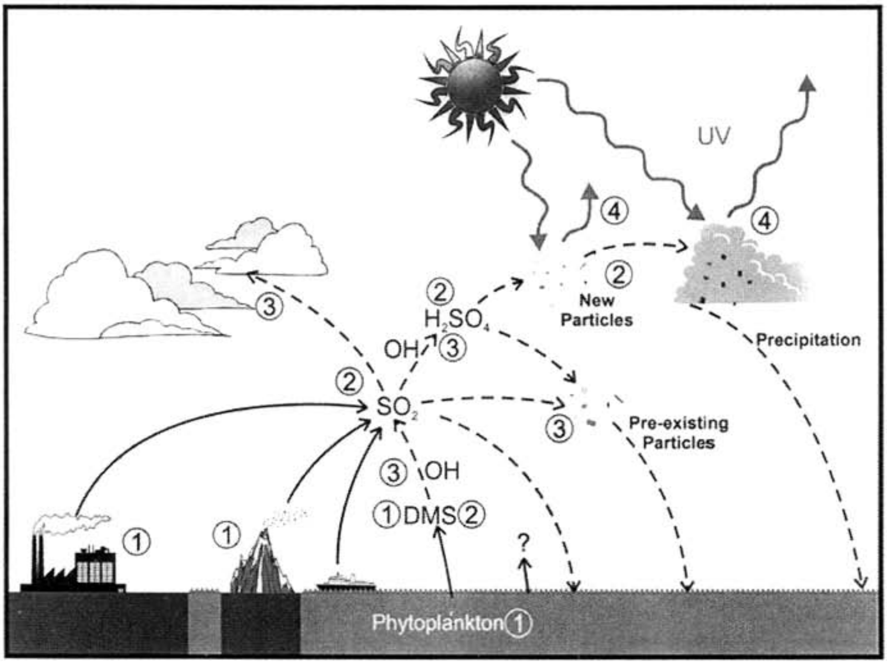


Figure 1 Simplified tropospheric sulfur cycle: ① The three largest documented global sulfur sources: ocean emissions, volcanoes, and fossil fuel burning. ② The most critical species involved in the cycling of sulfur: DMS, SO_2, H_2SO_4, and SO_4^{2-} (as CN, and CCN). ③ The major chemical processes in the cycling of tropospheric sulfur encompass gas-phase and heterogeneous reactions. ④ Among the important environmental impacts of tropospheric sulfur: formation of new particles and promotion of aerosol growth. Both are critical factors in Earth's radiation budget.

we present an overview of sources, sinks, and transformations of sulfur in the stratosphere with a special emphasis on sulfur sources responsible for maintaining the "background" level of stratospheric aerosol.

The authors note that because of the more fundamental chemical nature of Section 3, the discussion in this section is necessarily presented in greater detail than are other sections. The reader may choose, therefore, to by-pass this section. As the text is configured, this can be done without a major loss in grasping the larger global picture of atmospheric sulfur and how this element is critically coupled to the larger planetary environment.

2 CHEMICAL FORMS, SOURCES, AND CONCENTRATION LEVELS

As shown in Figure 2, some 11 different sulfur compounds define over 98% of the sulfur speciation in the atmosphere. Those in which sulfur is found bonded to either

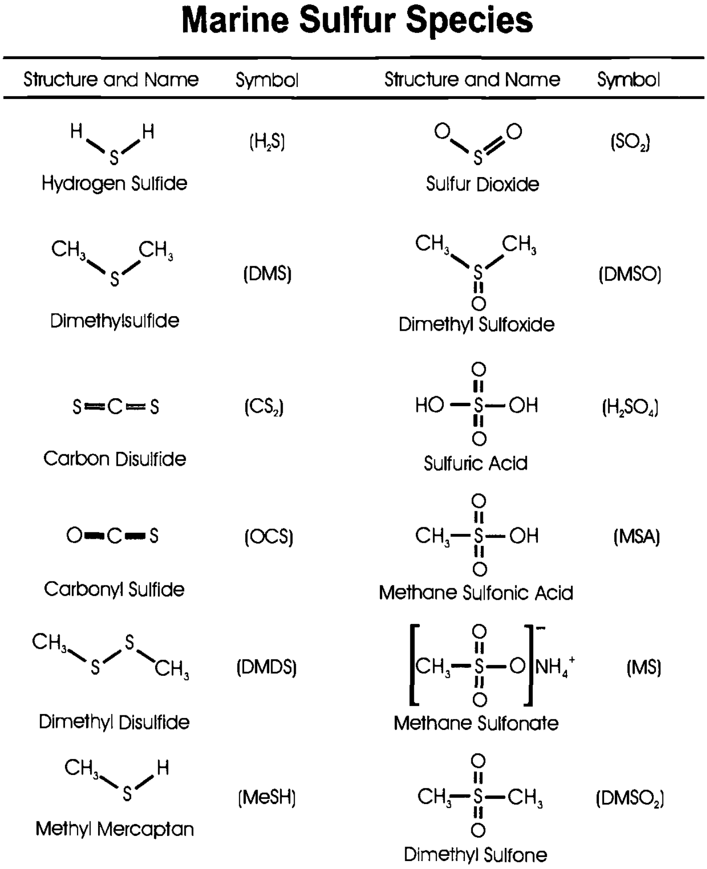


Figure 2 Chemical formulas and simple structures of the most common sulfur species in the troposphere. This list defines ~98% of the sulfur loading of the atmosphere.

hydrogen or carbon are typically in the lowest oxidation state, e.g., −2. As oxygen is sequentially added to sulfur, the oxidation state moves to 0, +4, and +6. Examples of these different states include dimethyl sulfoxide (DMSO), SO_2, and methane sulfonic acid (CH_3SO_3H), respectively.

As noted in Section 1, of those sulfur species shown in Figure 2, SO_2 and DMS are by far the most important primary forms emitted into the atmosphere. This point is further illustrated in Table 1, which provides a compilation of primary sulfur sources. From here it can be seen that of the total global average flux of 128 Tg S/yr, nearly 90% of this is defined by SO_2 and DMS emissions.

TABLE 1 Global Sulfur Emission Inventory

	DMS	SO_2	SO_4^{2-}	Other Reduced Sulfur	Total
Northern Hemisphere					
Combustion of fossil fuel	0.37–0.42	65–90	1.8–2	1.1–1.4	68–94
Oceans	5.8–9.7			0.1–0.4	5.9–10
Volcanoes		2.4–6.6	1.4–2.9	0.47–1.2	4.3–11
Other[a]	0.032–0.6	1.3–1.6	1.3–2.5	0.18–1.6	2.8–6.2
Anthropogenic total	0.37–0.42	65–90	1.9–2.1	1.1–1.5	70–96
Natural total	5.8–10	2.4–6.6	2.6–5.3	0.7–3	12–25
N.H. total	6.2–11	69–98	4.5–7.4	1.8–4.5	81–121
Southern Hemisphere					
Combustion of fossil fuel	0.02	7.1–9.2	0.2	0.12–0.13	7.4–9.6
Oceans	9.2–15			0.04–0.36	9.2–15
Volcanoes		1–2.6	0.6–1.1	0.13–0.43	1.7–4.1
Other[a]	0.021–0.2	1–1.3	0.8–1.6	0.096–0.53	2–3.7
Anthropogenic total	0.02	7.1–9.5	0.24	0.13–0.18	8.5–10.9
Natural total	9.2–15	1–2.6	1.4–2.7	0.26–1.3	12–22
S.H. total	9.2–15	9.1–13	1.6–2.9	0.39–1.5	20–33
Global total	15–26	78–111	6.1–10	2.2–6	102–154

[a] Includes biomass burning.

For DMS, the data indicate that 97% of the global flux of 15 to 26 Tg S/yr results from emissions from the ocean (Berresheim et al., 1995). Of this marine total, 61% is from the SH and 39% from the NH. The next largest contributor is split between wetland sulfur releases and those from anthropogenic/industrial emissions. By contrast, for SO_2 the global flux is largely defined by NH emissions (e.g., ~88%). This reflects the major contribution made from fossil fuel burning in the highly industrialized NH [for details see Spiro et al. (1992) and Hameed and Dignon (1992)]. Anthropogenic emissions from the SH make up still another 9% of the global total for SO_2 with volcanic emissions making up most of the remainder. This means that volcanic emissions define the second largest primary SO_2 global source but comprise, on average, only ~7% of the total. (Note, during years involving major eruptions, this source is substantially larger.) As noted earlier in the text, a very recent addition to the global inventory of SO_2 are emissions from ships. This source is currently estimated to be 2 to 4% of the total. In the case of other reduced sulfur (i.e., H_2S, CS_2, and OCS), the fluxes from the NH and SH are within a factor of 3 of each other and are made up of significant contributions from both natural and anthropogenic sources.

Overall, Table 1 clearly indicates that insofar as gross amounts of sulfur are concerned, fossil fuel combustion in combination with industrial emissions represent the single largest sulfur source in the NH. This is followed by nearly equal contributions from volcanoes and marine emissions. Still smaller emissions can be attributed to biomass burning and wetlands, and to direct release from plants and soils. By contrast, in the SH ocean emissions of sulfur are nearly 2 times larger than those from fossil fuel combustion, the latter being followed by volcanic and ship emissions. [For a more in-depth survey see Bates et al. (1992b).]

A centrally important conclusion that can be extracted from Table 1 is that anthropogenic sources of sulfur have overtaken natural sources in the NH. For example, of the 101 Tg S/yr (on average) released in the NH, nearly 83 Tg S/yr (i.e., 82%) can be assigned to human activities. By contrast, for the SH only 37% can be similarly assigned. The fact that human-related activities are now overshadowing the natural sulfur cycle in the NH raises some serious questions as to what environmental price tag is being paid for such a transgression? Shifts in atmospheric acidity and atmospheric turbidity in the NH have now been documented, and new concerns are being voiced about the impact of elevated sulfur on regional weather patterns and in long-term climate changes [Charlson et al. (1992)]. Thus, sulfur, like other trace chemical substances in our environment, when present in too large amounts has the potential for creating deleterious consequences for humankind.

The global source strength of a sulfur species, in combination with its areal source distribution and atmospheric lifetime, typically define the species concentration level at a given location. This being true, given that several sulfur species have the same integrated global emission fluxes, the species having the more regionally focused source tends to generate the highest concentration levels. On the other hand, the longer the lifetime of a sulfur species, all other things being equal, the higher its concentration and the lower its variability. For example, as shown in Table 2, the very long lived species OCS (i.e., 2 to 4 years) has a global median concentration

TABLE 2 Observed Mixing Ratio of Atmospheric Sulfur Species

Sulfur Species	Background Marine Boundary Layer		Background Continental Boundary Layer		Polluted Continental Boundary Layer	
	Typical Range	Median	Typical Range	Median	Typical Range	Median
H_2S	2–30	10	5–150	60	80–810+	365
DMS	15–300	65	1–20	8	0–10?	< 5?
OCS	400–800	600	300–7000	550	300–1800	545
CS_2	1–35	10	15–50	30	65–370	190
SO_2	20–50	35	20–1000	500	150–6000+	1500
H_2S	7–13	9	1–7	6	IDTA	IDTA
DMS	0–20	2	IDTA	IDTA	IDTA	IDTA
OCS	1–8	4	< 3–18	7	IDTA	IDTA
CS_2	IDTA[a]	IDTA	IDTA	IDTA	IDTA	IDTA
SO_2[a]	10–80	30	60–260	100	IDTA	IDTA

[a] IDTA, Insufficient data to assess.

centered around 500 pptv and varies by no more than a factor of 2 on a regional and global scale. At the opposite end of this scale, Table 2 reveals that SO_2, which has both highly focused continental sources and a relatively short lifetime (i.e., 0.5 to 9 days), displays some of the largest gradients of any sulfur compound with concentrations ranging from 35 to 5000 pptv. Of particular significance is the gradient between background continental regions and remote marine areas where factors of nearly 15 are seen. By contrast, DMS, which has a somewhat similar lifetime to SO_2, typically shows far more modest boundary layer concentration gradients. This is in keeping with DMS having a far less focused source region. Interestingly, due to the combination of its short lifetime, efficiency of vertical mixing, and the absence of high altitude sources, DMS unlike SO_2 displays very significant altitudinal gradients. Similar arguments to those given for OCS, DMS, and SO_2 can be used to explain the concentration levels and gradients observed for other sulfur species.

3 TRANSFORMATIONS

As stated earlier, three of the major players in the atmospheric sulfur cycle are DMS, SO_2, and SO_4^{2-}. Reflecting this conclusion, the present section on transformations will primarily focus on the processes by which DMS and SO_2 undergo further oxidation to reach the final oxidation state of sulfur +6. Of special significance will be the +6 sulfur forms $H_2SO_4(g)$, SO_4^{2-}(non-sea-salt sulfate, NSS), and methane sulfonate (MS).

DMS Oxidation

Some of the earliest studies that attempted to define the oxidation products of DMS were those carried out by Niki et al. (1983), Hatakeyama et al. (1985), and Grosjean (1984) in the early 1980s. These studies can best be labeled as "chamber studies" in that they typically involved filling a large multi-liter vessel with air mixtures containing DMS, NO, and other trace species (i.e., HONO), and then activating the system with solar or artificial radiation to produce OH radicals. There have been many different versions of the chamber-type study [see reviews by Yin et al. (1990), Turnipseed and Ravishankara (1993), and Berresheim et al. (1995)], some starting with sulfur in the form of DMS while others have used intermediate oxidation products like DMSO. The early studies as well as those that have followed have been quite revealing in demonstrating that among the important oxidation products generated from DMS are SO_2 and MSA, with lesser amounts of DMSO and $DMSO_2$. In fact, all of these products have now been directly measured in the atmosphere using modern instrumental techniques.

Although qualitatively revealing, chamber studies have also had their limitations. This reflects the fact that the gas mixtures employed have been significantly different chemically than that which is typically found in a marine boundary layer (MBL) environment. In this case two of the more important species involved have been DMS itself and the radical scavenging species NO. Both have typically been present in chambers studies at concentration levels several orders of magnitude higher than those found in the marine boundary layer. In addition, chamber studies have inherently been flawed due to their inescapably large surface-to-volume ratios (STVR). These have also been several orders of magnitude higher than those found in a marine environment (e.g., as aerosol surface area) and, thus, have led to greatly enhanced heterogeneous wall reactions. Since both concentration levels of DMS and NO as well as STVR factors impact on the DMS oxidation mechanism, not surprisingly, product distributions from individual chamber studies have been found to deviate significantly from study to study. They have thus left unanswered many of the quantitative details of the DMS product distribution within the MBL.

Among the more informative studies that have helped unravel aspects of the DMS oxidation mechanism have been those involving detailed laboratory kinetic investigations. These studies have focused on examining individual elementary reactions, the sum total of which, if available, would serve to define the overall DMS oxidation mechanism. One of the more pivotal of these was a study reported by Hynes et al. (1986). This study revealed that the reaction of OH with DMS proceeds not by a single reaction pathway but rather by two independent channels labeled kinetically as abstraction and addition, i.e., the reactions

$$CH_3SCH_3 + OH \Rightarrow CH_3S(OH)CH_3 \Rightarrow CH_2SCH_3 + H_2O \qquad (1)$$

$$CH_3SCH_3 + OH \Rightarrow CH_3S(OH)CH_3 \qquad (2a)$$

$$CH_3S(OH)CH_3 + O_2 \Rightarrow \text{products} \qquad (2b)$$

As shown in Figure 3, the abstraction channel is nearly temperature independent; whereas, the addition reaction reveals a very significant negative temperature dependence. The crossover point for near equal contributions from both channels is seen as near 285 K. The early thinking on this mechanistic finding was that the OH abstraction channel was the channel that predominantly led to the formation of SO_2, while products such as DMSO, $DMSO_2$, and MSA were believed to be associated with the OH addition channel.

Evidence supporting the above position has included extensive field observations in which the stable end products MS and non-sea-salt sulfate (NSS) were measured and the value of their ratio then examined as a function of the ambient temperature. It was argued that SO_2 could be expected to undergo reasonably fast oxidation in the MBL via heterogeneous processes (see discussion under SO_2 Oxidation), thus forming NSS. On the other hand, MSA(g), formed from the addition channel, would be quickly scavenged by sea-salt aerosol to form MS. If indeed the above processes collectively define the mechanism by which both products are formed, as noted above the measured ratio of MS to NSS could be expected to provide a good chemical reflection of the average temperature at which the DMS oxidation occurred. In fact, extensive field measurements that have evaluated this ratio over a range of latitudes and altitudes have shown that the lowest values (e.g., 0.07) occur at tropical latitudes and that some of the highest values (e.g., ≥ 0.34) tend to be found at much higher latitudes (e.g., Berresheim, 1987; Savoie and Prospero, 1989; Bates et al., 1992a). However, a more limited but still quite significant number of

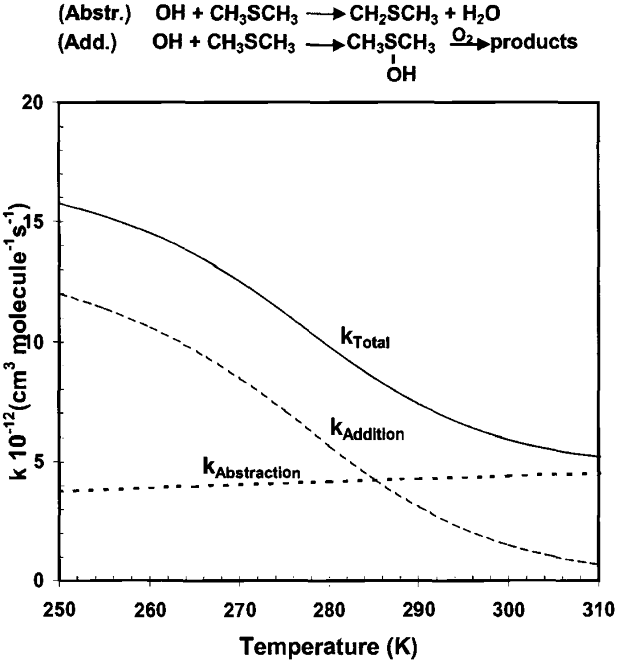


Figure 3 Temperature dependence of the rate coefficients for the OH/DMS addition and abstraction reaction channels as well as the total *k* value (modified from Berresheim et al., 1995).

observations have also been reported that do not follow this simple trend (e.g., Berresheim et al., 1995; Davis et al., 1998). Since these observations appear to be equally valid, they most likely point to a DMS oxidation mechanism that is more complex than originally thought. (Note, the potential importance of the MS/NSS ratio as defined by DMS oxidation rests in the fact that this ratio, if well understood, could be used to apportion the DMS contribution to total NSS. Perhaps, more importantly, it could be used as an indicator of the temperature environment under which the DMS oxidation process took place.)

One rendition of the DMS oxidation mechanism that reflects the thinking that the overall process is actually quite complex is that shown in Figure 4. This mechanism (shown here in abbreviated form) has folded in the most recent results from both field studies as well as laboratory kinetic investigations. Quite significant is the clear indication that not only is the OH abstraction channel a source of SO_2 but that the addition branch, in several different steps, also can form SO_2 as a product. Equally important, the stable product MS is shown as a product of both addition and abstraction channels. Its production efficiency is shown as being even further convoluted as

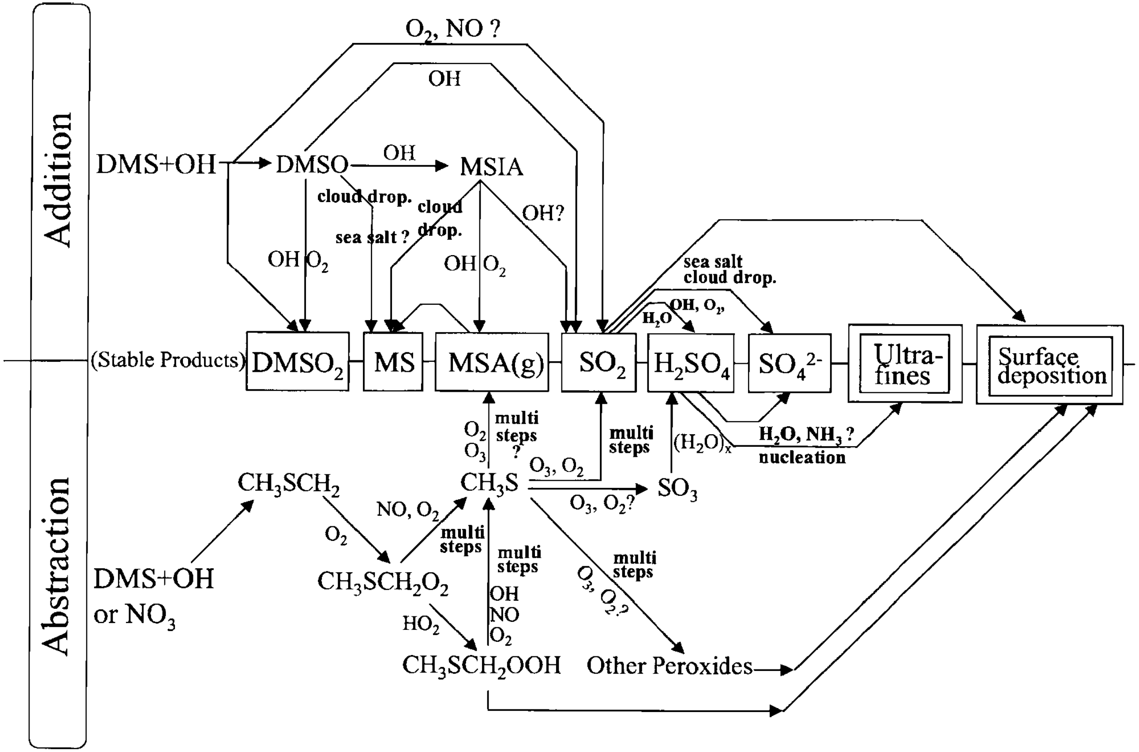


Figure 4 Abbreviated DMS oxidation scheme (modified from Davis et al., 1999).

a result of it being formed through competing gas and heterogeneous processes involving the intermediates DMSO and MSIA. Although the mechanism shown is still speculative (e.g., many of the elementary reactions have not yet been fully characterized), recent sulfur field studies, covering a wide range of latitudes, have provided evidence that strongly supports key aspects of this mechanism.

Unlike some of the earliest sulfur field studies, more recent investigations have reported a significant coupling between DMS and SO_2. Given that both DMS and SO_2 typically have MBL lifetimes of 0.5 to 2 days, if DMS is a significant source of SO_2, one would expect that these two species should be anticorrelated when measured at a rate significantly shorter than their respective lifetimes. Alternatively, if measured at a time resolution significantly longer than their respective lifetimes, one would expect a positive correlation. Thus, depending on the sampling rate, the appearance of either a correlation or anticorrelation in field data would signal there being a significant oxidative pathway from DMS to SO_2. As cited above, in the earliest studies for which simultaneous measurements of DMS and SO_2 were recorded, no relation between these two sulfur species was found. However, with improvements in instrumentation and the more judicious selection of field sites (e.g., free of anthropogenic pollution sources), a quite different picture is now emerging. Among the more significant studies has been that reported by Bandy et al. (1996), which took place in 1994 at Christmas Island. Located at 2°N in the middle of the Pacific Ocean, Christmas Island defines an ideal setting for studying DMS oxidation chemistry. Situated near the middle of the equatorial upwelling, it experiences near year-round elevated levels of DMS with no evidence of significant other sources of SO_2. In addition, being located well within the strong trade wind regime, it typically experiences stable meteorological conditions for several days at a time. Finally, due to the high solar flux and water vapor levels present, it also defines an environment where very high levels of the critical boundary layer oxidizing agent OH can be found. Reflecting these optimum conditions, Bandy and co-workers (1966) reported the first convincing field data showing a strong diel relation between DMS and SO_2. These investigator's high temporal resolution data were recorded over a 9-day time period and revealed a clear and convincing anticorrelation between these two sulfur species. The estimated DMS to SO_2 conversion efficiency was reported as $62 \pm 6\%$.

In an airborne follow-up study at Christmas Island in 1996 [part of the National Aeronautics and Space Administration's (NASA's) GTE PEM–Tropics A program, Hoell et al. (1999)], the sulfur database reported was even more revealing. During this investigation direct observations were recorded of both DMS and SO_2 as well as the oxidizing agent OH. Equally significant was the availability in this new study of meteorological and chemical data as a function of altitude. As shown in Figure 5a, the profiles for DMS, SO_2, and OH make for a very convincing case that DMS is a major source of SO_2 and that DMS oxidation predominantly occurs via OH radicals. An analysis of these new data by Davis et al. (1999) resulted in an overall DMS to SO_2 conversion efficiency of $72 \pm 22\%$, well within the range reported by Bandy et al. (1996). Given that the abstraction channel at the temperatures of Christmas Island (298 K) represented 70% of the total OH/DMS reaction rate, together with the conservative estimate that at least 8% of the product yield from DMS forms species

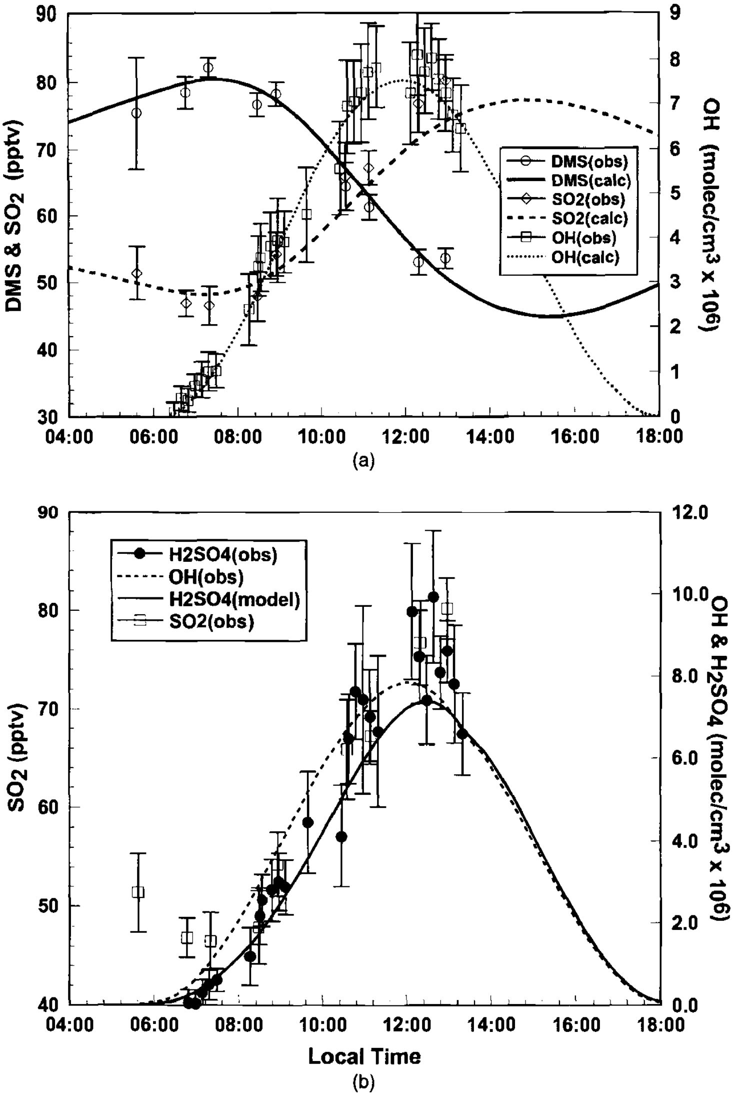


Figure 5 Analysis of airborne sulfur data collected during NASA's PEM–Tropics A program: (*a*) observed and model profiles for DMS, SO_2, and OH for tropical boundary layer conditions; (*b*) plot showing SO_2 conversion to H_2SO_4 via reaction with OH (see e.g. Davis et al, 1999). [Note: DMS and SO_2 observational data are those recorded by Thornton et al. (1999). The OH and H_2SO_4 observational data are those recorded by Mauldin et al. (1999a,b)].

other than SO_2, one can estimate that the likely range for SO_2 formation from the abstraction channel is between 0.6 and 0.9. This suggests that the contribution of SO_2 from the addition channel is probably not much lower than 0.4; however, the issue of SO_2 contributions from the addition channel is more fully and better explored in the text below based on field studies conducted at much lower ambient temperatures.

Shifting to the recent National Science Foundation/National Oceanic and Atmospheric Administration (NSF/NOAA) program ACE-1 [Aerosol Characterization Experiment, Bates et al. (1998)], the opportunity again presented itself to examine DMS oxidation chemistry under remote conditions, but with the field site being defined as the Southern Ocean, just to the south of Tasmania. Thus, it provided an environment having much lower ambient temperatures. Recall that because of the strong negative temperature dependence of the OH/DMS addition channel, the addition channel becomes more prominent under these conditions. In fact, based on the average temperature recorded on the ACE-1 aircraft sampling platform (i.e., 280 K), both channels are estimated to be of near equal importance. Davis (unpublished results) in his analysis of the resulting DMS and SO_2 data has estimated that the overall DMS-to-SO_2 conversion efficiency to be still quite high, e.g., 0.7 to 0.9. Thus, to be consistent with the measured SO_2 levels and the previously cited average efficiency for SO_2 production from the abstraction channel of 0.75, he assigned a range of 0.7 to 0.9 to the addition channel. Although seemly quite high, the latter range is seen as being consistent with the previously discussed SO_2 tropical analysis.

In yet another study at still higher latitudes (i.e., 66°S), the land-based NSF SCATE Antarctic Program (Berresheim and Eisele, 1998), the average temperature recorded was only 273 K. In this case the addition channel is estimated to be 65% of the total OH/DMS kinetic rate. Although SO_2 measurements were not made during this campaign, direct observations of OH, DMS, and H_2SO_4 suggest that the overall SO_2 conversion efficiency required to support the observed H_2SO_4 levels would need to be well above 60%. Thus, these data again point toward an SO_2 efficiency for the addition channel of 50% or higher (Davis et al., 1998).

Quite interestingly, in all of the analyses cited above, involving high conversion efficiencies of DMS to SO_2, the measurements being used have been those reported by Bandy and co-workers (1996). The latter group pioneered the use of the isotopic-dilution gas-chromatographic mass-spectrometric technique (IDGCMS) for both field measurements of DMS and SO_2 (Bandy et al., 1993). On the other hand, in two recent ship-based studies, one in the tropical South Pacific [MAGE (Marine Aerosol and Gases Experiment), Yvon et al. (1996)], the other at high latitudes as part of the ACE-1 program (De Bruyn et al., 1998), a quite different technique was employed in the measurement of SO_2. In both of these cases SO_2 measurements were made using the aqueous-phase fluorescence method (Saltzman et al., 1993). In both studies, evidence was found of diel trends in DMS and SO_2, further confirming the important role of photochemical oxidation of DMS. The reported overall conversion efficiencies for DMS to SO_2, however, were estimated to be significantly different from those cited above. In each of the ship studies, the conversion efficiencies ranged from 30 to 50%, a factor of 1.5 to 2 lower than those already cited. It seems unlikely at this juncture, even though the latter studies were conducted on different platforms, that the results would differ by this amount. Both the average temperature and levels of other critical chemical species appear to have been similar for the tropical studies and the high latitude investigations. Whether the reported difference is due to SO_2 measurement difficulties or to yet unknown factors cannot be determined at this time.

Evidence that oxidizing agents (i.e., NO_3 and Cl) other than OH are important in converting DMS to SO_2 has been primarily based on results from laboratory kinetic studies [see, e.g., Berresheim et al. (1995) and references therein]. In combination with model estimated levels of NO_3 and Cl, the tentative conclusion has been reached that only the NO_3 mechanism is significant and that for most marine areas even this oxidant is unimportant. The exception would be for highly populated coastal influenced regions where major sources of NO_x could be expected. As related to the importance of Cl atom oxidation DMS, even though some recent sulfur field data (e.g., MAGE study) suggest that the impact from Cl might be significant, other results reveal a different picture. For example, the results from the previously discussed field studies at Christmas Island as well as those in the Southern Ocean suggest that Cl atom DMS oxidation is less than 15%. An independent study by Singh et al. (1996), in which C_2Cl_4 budget arguments were used to evaluate the significance of boundary layer Cl atom oxidation, also resulted in a similar conclusion, namely, that the latter chemistry is of negligible importance in remote marine regions.

In addition to the pivotal question related to the oxidative conversion efficiency of DMS to SO_2, significant other DMS oxidation issues also continue to be the subject of continuing research. These include identifying the DMS oxidation intermediate(s) responsible for gas-phase $H_2SO_4(g)$ formation, the elucidation of the pathways by which DMSO, $DMSO_2$, MSA(g), and MS are formed, and the identification of the factors controlling the MS/NSS ratio. In the text that follows these DMS issues are further explored in the context of both recent laboratory kinetic investigations as well as the results from recent marine sulfur field studies.

As related to $H_2SO_4(g)$ formation, a review of the literature points to two major reaction sequences as being of potential importance. These include (3a) to (3d) and (4a) and (3d), e.g.,

$$\text{DMS} + \text{OH multisteps} \Rightarrow SO_2 + \text{other products} \quad (3a)$$

$$SO_2 + OH + M \Rightarrow HSO_3 + M \quad (3b)$$

$$HSO_3 + O_2 \Rightarrow SO_3 + HO_2 \quad (3c)$$

$$SO_3 + (H_2O)x \Rightarrow H_2SO_4(g) + H_2O \quad (3d)$$

$$\text{DMS} + \text{OH multisteps} \Rightarrow SO_3 + \text{other products} \quad (4a)$$

$$SO_3 + (H_2O)x \Rightarrow H_2SO_4(g) + H_2O \quad (3d)$$

Two of the most revealing recent field studies that have examined this issue are the previously discussed PEM–Tropics A Christmas Island airborne study and the ground-based SCATE study in Antarctica. Recall, that during the PEM–Tropics A field investigation direct observations of DMS, OH, SO_2, H_2SO_4, and total aerosol surface area were simultaneously recorded. As shown in Figure 5*b*, using the known rate coefficients for processes (3a) to (3d), together with recently measured aerosol sticking coefficients for H_2SO_4, Davis et al. (1999) concluded that the observed profile for $H_2SO_4(g)$ could be convincingly explained in terms of the observed diel profiles for SO_2 and OH. Since, as previously discussed, the observed SO_2 profile from this field study was also explicable in terms of OH/DMS oxidation, these new results are consistent with the idea that SO_2 is the critical DMS intermediate leading to gas-phase H_2SO_4 formation. Taking a different approach, Jefferson et al. (1998), not having direct observations of SO_2, used the observations of DMS, H_2SO_4, OH, and total aerosol surface area from the SCATE program to evaluate the two quantities k_{OH} [OH][DMS] and $k_{surf}[H_2SO_4]$. It was argued that if the direct formation of SO_3 from DMS were important, given the reasonably short lifetimes for both H_2SO_4 and SO_3 in the Antarctic environment (i.e., < 1 h), one would expect a significant correlation between these two quantities. In fact, the R^2 value was less than 0.2, indicating no relationship. Although still lacking the finality that comes from having a comprehensive set of elementary rate constants for each step in a mechanism, the collective results cited above strongly suggest that SO_2, not SO_3, is the dominant intermediate from the oxidation of DMS that leads to the gas-phase formation of H_2SO_4.

Field observations bearing on the mechanistic details surrounding the formation of DMSO and $DMSO_2$ from DMS oxidation have been limited in number and conflicting in their results. They include results from a sulfur field study near the Washington coast, the previously discussed Antarctic SCATE program, and finally the 1994 Christmas Island study. During the SCATE program, there were ~6 days of near continuous recording of DMSO and $DMSO_2$ (Berresheim et al., 1998). However, in the analysis of these data Davis et al. (1998) could find only 1 day out of the 6 sampled in which it appeared that both DMSO and $DMSO_2$ levels were controlled by local photochemical production. For all other days DMSO and $DMSO_2$ were shown to be controlled by transport processes, wherein large quantities of ocean-released DMS were initially carried aloft into the lower free troposphere, oxidized, and then returned in the form of intermediate as well as +6 oxidation state sulfur. But, on January 19, 1994, it appears that there was a significant break in this cycle in that background levels of both DMSO and $DMSO_2$ were found to be a factor of 10 lower than during the other 5 sampling days. Only on this day, was there any evidence of a diurnal profile for DMSO that tracked the measured ultraviolet (UV) solar irradiance. From their analysis of these data, Davis and co-workers (1998) estimated that the DMSO formation efficiency from the OH/DMS addition channel could range from 0.5 to 1.0, a value well within the limiting value assigned to this branching ratio based on two independent laboratory kinetic investigations. The $DMSO_2$ data, although considerably more noisy, were found to be most consistent with a branching efficiency for DMSO/OH to $DMSO_2$ of ~0.3.

In the field study near the Washington coast (Berresheim et al., 1993), 2 to 3 days of DMS and DMSO data were collected, but with very little ancillary data to facilitate defining the photochemical environment for this investigation. On April 14, however, sunny conditions prevailed nearly all day, and DMS and DMSO were sampled while air was advected in from the Pacific Ocean. For this specific case Berresheim et al. (1995) were able to estimate a branching ratio for the DMS/DMSO addition channel of ~0.5 but with a large uncertainty. This result may again be viewed as in good agreement with the above-cited results by Davis et al. (1998), but both have large uncertainties associated with them. In the 1994 Christmas Island study the measured levels of DMSO were found to be incredibly large relative to the median values of DMS observed, i.e., median DMSO 25 pptv, median DMS 200 pptv. Chen et al. (2000) in their analysis of this data quickly concluded that the two observations were totally irreconcilable in terms of any known DMS oxidation mechanism. This led them to put forward two possible hypotheses: (1) There were possibly unknown difficulties in the measurements of DMSO or (2) yet unknown sources of DMSO may exist. Still more recently NASA's PEM Tropics B Field study, direct airborne measurements of DMS, DMSO, and OH from sunrise to 1pm local time indicated that the highest levels of DMSO were at or near sunrise. Concentrations were found to decrease throughout the remainder of the measurement period (Nowak et al., 2001). The authors have pointed out that the temporal behaviour of DMSO is totally contrary to that expected if DMSO was formed only from the reaction of DMS with OH. Thus, these new data also suggest an additional source of DMSO, one that operates at night as well as possibly during daylight hours. Quite clearly if the latter hypothesis is correct, Table 1, as related to global sources of sulfur, would require further modification.

The issue of how efficiently DMSO might be formed through the OH/DMS addition channel raises the equally important question: Does the further oxidation of this species provide an effective pathway for formation of MSA(g)? This +6 oxidation state sulfur compound is shown in Figure 4 being formed from the oxidation of methane sulfinic acid (MSIA), another intermediate from the OH/DMS addition channel. Recent laboratory kinetic data (Hynes et al., 1996; Urbanski et al., 1998) would seem to support this notion in that they found the OH/DMSO reaction to be very fast and that the reaction appears to lead to near unity yields of the CH_3 radical. The product CH_3 radical is one that would be expected if the initial adduct formed from the reaction of OH with DMSO broke apart to form MSIA. Even so, in both of the airborne field studies previously cited, as well as during project SCATE, direct observations of gas-phase MSA have revealed a very low production efficiency for this species, e.g., typically $\leq 1\%$. The latter result is significant in that it translates to our explaining no more than 2 to 5% of the observed MS aerosol loading from the condensation of MSA(g). This means that both under tropical as well as the low-temperature conditions of the Antarctic, gas-phase production of MSA is not the major source of MS (the latter being a frequently cited measurement in much of the older literature involving DMS field studies).

The above MSA(g) results again focus our attention on the question touched on earlier in the text: How well do we really understand the factors controlling the value of the much cited MS/NSS ratio? Recall earlier (bottom of page 132) in the text we discussed the fact that some observations have shown a trend of increasing values in this ratio with decreasing temperatures (i.e., increasing latitude) but that notable exceptions had also been seen in this trend. One recent explanation for both the observed low yield of MSA(g) and yet significant yields of MS has been that proposed by Jefferson et al. (1998). To explain the MSA(g)/MS SCATE results, these investigators proposed that the rather high median levels of 1 to 2 pptv of DMSO observed on the Palmer Peninsula (see earlier discussion in text related to DMSO formation at Palmer) could only be accounted for if heterogeneous DMSO reactions were occurring on sea salt aerosols. This hypothesis is supported by the fact that in several independent aqueous phase kinetic studies, DMSO has been shown to react in the aqueous phase (in the presence of oxidizing agents such as OH radicals) to form MSIA. This species was observed to subsequently undergo further oxidation to yield MS. Davis et al. (1999) came to a similar conclusion when analyzing the PEM–Tropics A data; however, these investigators noted that both gas-phase DMSO and MSIA(g) would be equally good candidates as a heterogeneous source of MS. In yet another laboratory kinetic study, Lee and Zhou (1994) examined the aqueous-phase reaction of DMS with O_3 as a possible source of aerosol-phase oxidized sulfur. What they found was that because of the very low Henry's law constants for both DMS and O_3, the probability of this aqueous process is rather unfavorable. However, under the most favorable conditions involving heavy clouds and relatively high O_3 levels, it could prove to be a significant source of oxidized sulfur. Collectively, the above findings would seem to suggest that the formation of MS and possibly other sulfur species must be viewed in the context of both gas-phase and heterogeneous chemistry in the atmosphere.

The implications of the above findings are quite significant in that they clearly point to the possibility that the MS/NSS ratio depends not only on the temperature of the environment where DMS oxidation occurs, but is equally, if not more, influenced by the nature of the aerosol environment, e.g., sea salt loading, cloud density, and liquid water content. For very low aerosol loadings a substantial fraction of the DMSO and MSIA(g) from the OH/DMS addition channel would most likely react with OH in the gas phase to produce SO_2. Most of this SO_2 would subsequently be converted into NSS. On the other hand, for very high aerosol loadings nearly all DMSO and MSIA would likely be scavenged, producing MS as a final product. Thus, for a given sampling location where the aerosol loading might vary from day to day, one could expect to find a range of MS/NSS values. In this context, one of the most stable environments in which MS/NSS values would be rather constant would be that defined by the tropical marine BL. Here the abstraction branch would strongly dominate DMS oxidation and the sea-salt aerosol loading would remain both relatively high and reasonably constant. Indeed, some of the most consistent values for the MS/NSS ratio have been those measured in the tropics (e.g., Saltzman et al., 1986; Savoie and Prospero, 1989; Berresheim et al., 1995). However, in spite of what appears to be a reasonably well documented environment, one should not lose sight of our earlier discussion that hinted at the strong possibility that there may be a significant and yet unidentified source of DMSO. If so, considerable rethinking of tropical marine sulfur chemistry may be necessary.

Thus far our DMS discussions have been primarily focused on the marine boundary layer (MBL). It may be asked, therefore, how dramatically does this picture change if the oxidation of DMS were to occur in the free troposphere (e.g., above 2 km)? In fact, as hinted at in our earlier discussions of the temperature dependence of the OH/DMS reaction, quite significant changes can occur. In the free troposphere three major physical changes occur in the environment: the temperature drops (e.g., 6.5°C/km), the pressure drops (i.e., exponentially), and the average aerosol surface area drops by at least one order of magnitude. It is the first and third of these shifts that potentially could have the most significant impact on DMS oxidation chemistry. Recall, that the OH/DMS addition channel has the strongest dependence on temperature (increasing with decreasing temperature), and therefore this channel becomes the dominant one with increasing altitude. On the other hand, laboratory studies suggest that this channel probably also has the greatest diversity in oxidation products. Equally important, the oxidation product distribution from this channel appears to have the greatest dependence on aerosol surface area, i.e., heterogeneous reactions. Thus, speculating on the net effect of these factors might point toward enhanced levels of both $DMSO_2$ and MSA(g). It could also mean that a much larger fraction of the DMSO and MSIA would be oxidized via OH, leading to higher yields of SO_2 or new products like sulfurous acid (H_2SO_3) from this channel. This sequence of reactions, in turn, could lead to the higher yields of $H_2SO_4(g)$ which under the cold temperatures of the upper troposphere could form the basis for new aerosol particle formation as suggested by Clarke (1993). Still another interesting result from this high-altitude DMS chemistry would be its impact on the MS/NSS ratio. For example, with an enhancement in the yield of SO_2 from the addition channel, the value of this ratio might remain reasonably low even though the temperature at which the oxidation occurred was quite low. Suffice it to say, both new laboratory kinetic studies as well as field observations will be required to actually quantify this chemistry.

SO_2 Oxidation

As both a primary source species (e.g., combustion and volcanoes) and as one of the major products from DMS oxidation, the atmospheric fate of SO_2 represents a major component of the atmospheric cycling of sulfur. It is estimated that betwcen 40 and 60% of this SO_2 is directly deposited to either land or ocean surface areas (Berresheim et al., 1995). The remainder is believed converted into sulfur +6, although the detailed mechanisms by which this final oxidation state is reached continues to be the focus of ongoing research. What is now reasonably clear is that there are at least two general pathways by which this is achieved: one involving gas-phase chemistry, the other involving heterogeneous reactions. The gas-phase process is now relatively well understood, involving the ubiquitous oxidizing agent OH, e.g., reactions (3b) to (3d). The first two steps were reasonably well established by the mid 1980s (Finlayson-Pitts and Pitts, 1986); however, only recently have the details of step (3d) been established (Lovejoy et al., 1996). This process has now been shown to involve a quadratic dependence on H_2O. But, considering the amount of H_2O in the atmosphere, this step is rarely if ever the rate-limiting step. In virtually all cases step (3b) is rate limiting.

As related to the gas-phase oxidation of SO_2, the importance of this process must be viewed both from the perspective of converting bulk atmospheric SO_2 to sulfate and from the point of view of its role as a major source of gas-phase H_2SO_4. Current evidence suggests that the gas-phase oxidation of SO_2 is probably no greater than 20% of the total, and in the final analysis it could be no more than 5 to 10% (Lelieveld and Heintzenberg, 1992). On the other hand, the gas-phase production of H_2SO_4 now appears to represent a critical step in the formation of new particles (via heterogeneous nucleation) that ultimately leads to cloud formation (e.g., Kreidenweis and Seinfeld, 1988). Thus, in the absence of this source, it would be difficult to explain how the atmosphere resupplies itself with CCN. CCN are routinely removed by both wet and dry deposition. This suggests then that the gas-phase oxidation of SO_2 is of primary importance in the atmosphere defining the strong link between sulfur emissions and climate effects.

Of the 80 to 90% of the SO_2 that is oxidized by non-gas-phase pathways, both heterogeneous reactions involving cloud droplets as well as sea salt aerosols are considered important. [For highly industrialized regions the influence of soot particles, trace metals such as Fe(+3), Mn(+2), and Cu(+2), and organic carbon reactions must also be included.] In remote marine areas, the heterogeneous oxidation process typically involves several steps. The first of these involves the critical equilibria shown in (5a) and (5b):

$$SO_2(g) + H_2O \rightleftharpoons SO_2(aq) \tag{5a}$$

$$SO_2(g) \rightleftharpoons HSO_3^- \tag{5b}$$

$$HSO_3^- \rightleftharpoons SO_3^{2-} + H^+ \tag{5c}$$

The presence of these equilibrium reactions means that the dominant form of sulfur +4, in the bulk aqueous phase, depends very much on the acidity of the aerosol species. For the most typical range of acidity in the troposphere, the dominant form of sulfur is the bisulfite ion (HSO_3^-). However, because of shifts in the levels of the individual forms of sulfur with changing pH, as well as the dependence of the reaction coefficients on pH, the most important aqueous-phase pathway for oxidation of sulfur +4 can be a strong function of pH, and therefore on the total amount of sulfur converted (e.g., Martin, 1984). This point is illustrated in Figure 6. Here it can be seen that for pH values above 5, the oxidation by O_3 represents the dominant pathway; whereas for pH values less than 5, oxidation via H_2O_2 becomes the major source of sulfur +6. Other investigators (e.g., Chameides and Stelson, 1992; Sievering et al., 1992) have proposed that the sensitivity of the aqueous-phase oxidation of sulfur +4 to percent sulfur converted might be much smaller than originally thought. It has been suggested that this would be particularly true when the aerosol species is sea salt. The above group of investigators have argued that sea salt contains a natural buffering capacity involving the bicarbonate/carbonate system. Thus, seawater aerosol might be able to sustain a high rate of conversion of +4 sulfur to +6 through the O_3 oxidative pathway for extended periods of time.

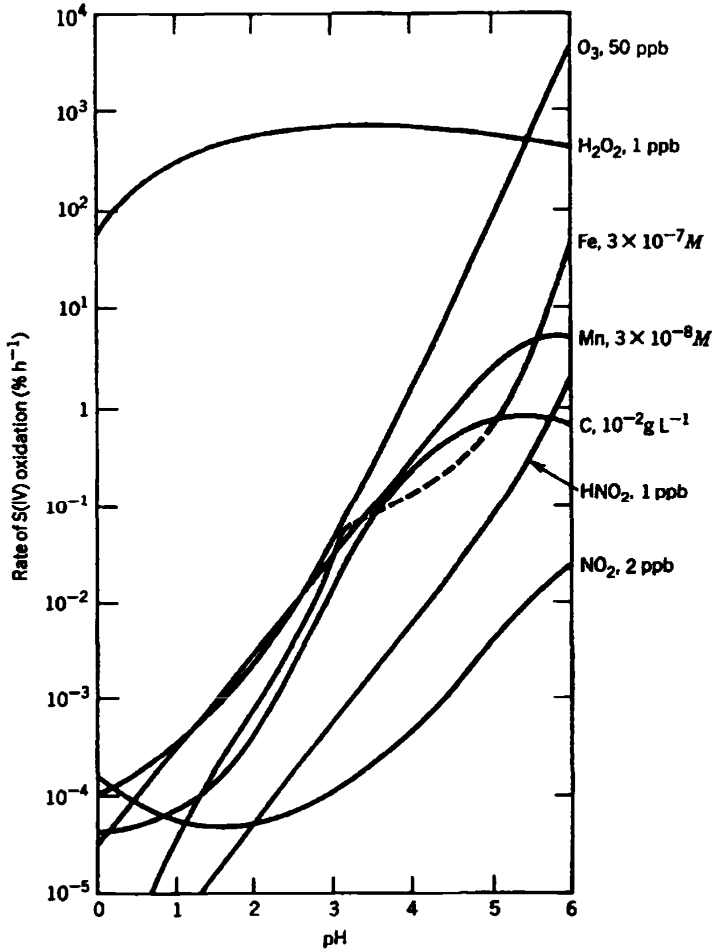


Figure 6 Estimated rates of oxidation of S(IV) in solution and on carbon surfaces as a function of pH (taken from Martin, 1984).

Although the above discussion might be viewed as downplaying the overall role of atmospheric photochemistry in the conversion of sulfur +4 to the +6 state, this clearly is not the case. For example, not only are there potentially other aqueous-phase reactions driven by scavenged gas-phase radicals such as HO_2 (e.g., Chameides and Davis, 1982); but there is also the fact that the critical heterogeneous oxidizing agents are typically O_3 and H_2O_2. Both of these species are themselves predominantly generated in the gas phase via photochemical processes. Thus, both the SO_2 oxidation by OH and that by heterogeneous pathways must be viewed as significantly influenced by local and/or regional photochemistry.

4 GLOBAL DISTRIBUTIONS OF SO_2 AND SULFATE

In Section 1 the point was made that of the stable atmospheric forms of sulfur, SO_2 and SO_4^{2-} were among the more important species that need to be well understood. In this context, a key characteristic that needs to be examined for both species is its atmospheric lifetime. Both tend to be long enough to permit their being transported over near hemispheric scales. Thus, their global distributions represent an important

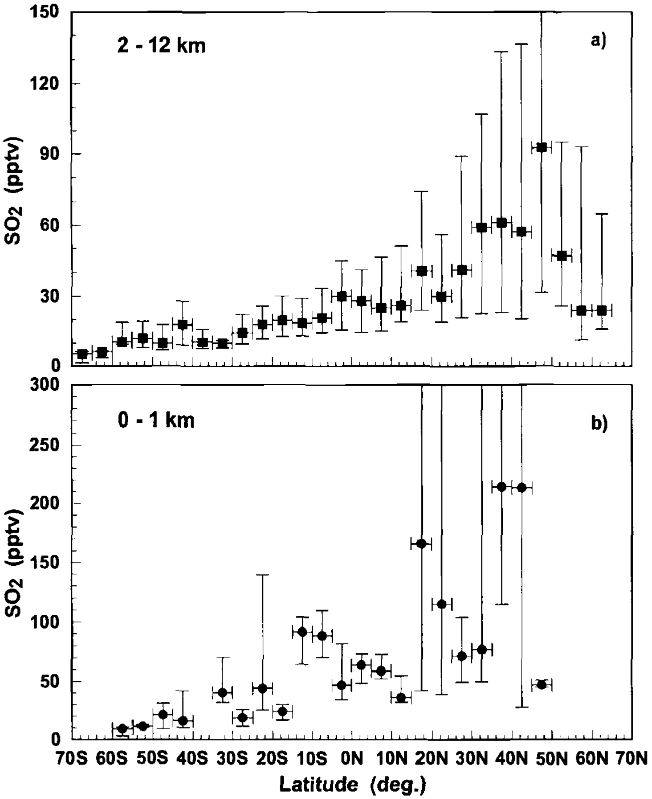


Figure 7 Latitudinal distributions of SO_2 for the altitude ranges of: (*a*) 0 to 1 km and (*b*) 2–12 km (modified from Thornton et al., 1999). The data in these plots have been derived from individual observations that have been binned every 5° of latitude. Vertical bars on each 5° latitude bin signify the one sigma variability in the average value for each bin.

indicator of humankind's influence on the natural sulfur cycle. At present the database for both species is still quite limited with vertically resolved data now being available for no more than 20% of the global atmosphere. Most of the latter data is also limited to no more than one or two seasons of the year with the geographical coverage being confined largely to the North and South Pacific Oceans. Representative of these data are the latitudinal plots of SO_2 shown in Figures 7*a* and 7*b*. For clarification purposes, these data have been binned for the altitude ranges of 0 to 1 and 2 to 12 km. Several points made earlier in the text, concerning anthropogenic effects, are clearly revealed in these plots. For example, the Northern Hemisphere is seen as having an average mixing ratio for SO_2 that is nearly five times higher than that for the Southern Hemisphere. Equally significant is what appears to be direct evidence for the focused release of SO_2 in the highly industrialized midlatitude region of 30 to 55°N.

Thus, given the limitations of current field data, one must turn to models to explore in greater depth the global atmospheric picture of sulfur. In this case the goal is that of gaining further insight into the distributions and variations in sulfur compounds and the processes that regulate their concentration levels. Several global-scale chemistry transport models have been developed in the past 7 years for just this purpose. These models endeavor to place available input sulfur data, as related to sources, sinks, and concentration levels, into a comprehensive global sulfur cycle (e.g., Langner and Rodhe, 1991; Pham et al., 1995; Feichter et al., 1996; Chin et al., 1996, 2000; Chuang et al., 1997; Koch et al., 1999; Barth et al., 2000). All include tropospheric sulfate and its major precursors (i.e., DMS and SO_2) and contain modules designed to handle anthropogenic and natural emissions, chemical transformations, advection/convection, and dry and wet deposition. Illustrative of the output from these models, we show in Figure 8*a* and 8*b* the annually averaged global surface-air distributions for SO_2 and sulfate based on results from Chin et al.'s (2000) model. This model includes sulfur from fossil fuel and biofuel combustion, shipping and aircraft emissions, biomass burning, volcanoes, and biogenic sources. Here it can be seen that the maximum SO_2 concentrations are clearly located at latitudes between 30 and 75°N, corresponding to the major industrial source regions of eastern United States, Europe, and eastern Asia. The levels of SO_2 are seen ranging from 1 to over 10 ppb. Interestingly, significant surface SO_2 concentrations are also shown to be present over southern Africa and Chile, largely reflecting ore smelting operations. The distribution of surface-air sulfate over the continents is found to be very similar to that for SO_2, although the gradients are clearly smaller. These observations reflect the fact that sulfate is the primary product of SO_2 oxidation and that transport as well as dry and wet deposition represent major losses for SO_2 (see discussion later in chapter). The model results also reveal that sulfur concentrations in the Arctic and near coastal regions in the Northern Hemisphere tend to be heavily influenced by anthropogenic releases. Returning to the field data shown in Figure 7*a*, the model values found at these near continental locations appear to be substantial larger than those reported in the latitudinal plots taken from the data of Thornton et al. (1999). Recall, however, that most of these data were recorded over remote regions of the Pacific. The sulfate distribution (i.e.,

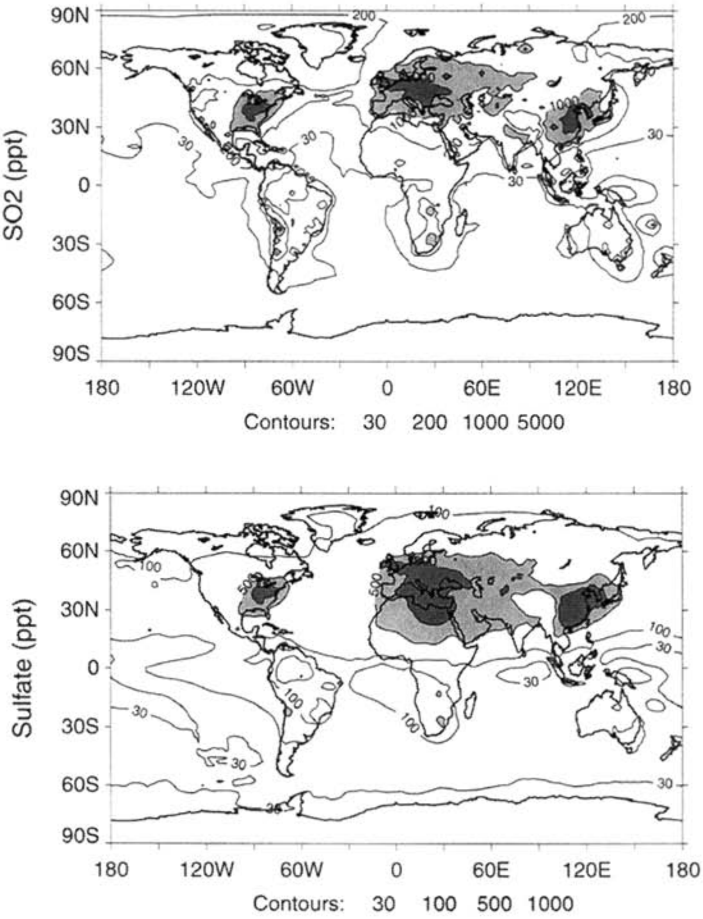


Figure 8 Global sulfur mixing ratios in the lowest 500 m as derived from the global chemistry transport model of Chin et al. (1999): (*a*) SO_2 and (*b*) $SO_4{}^{2-}$.

Figure 8*b*) which also is shown falling off like SO_2 as one moves from continental regions to the open ocean, is similarly in good agreement with observational data when one considers the geographical location of the surface sampling sites [see, e.g., the data of Savoie et al. (1989) over the North Atlantic, that of Savoie and Prospero (1989) over the North Pacific, and that in the Arctic reported by Barrie et al. (1989)].

As seen in the SO_2 data of Thornton et al. (1999) and in the model results shown in Figure 9*a* and 9*b*, the impact from anthropogenic emissions of sulfur is significantly attenuated at altitudes well above the boundary layer. This is due both to a large fraction of the combustion-based SO_2 and sulfate being deposited within the continental boundary layer and also to the more efficient dispersion of these species once at higher altitudes. Chin and Jacob (1996) have estimated that about 40% of the Northern Hemisphere industrial source of SO_2, is transported out as SO_2, or sulfate

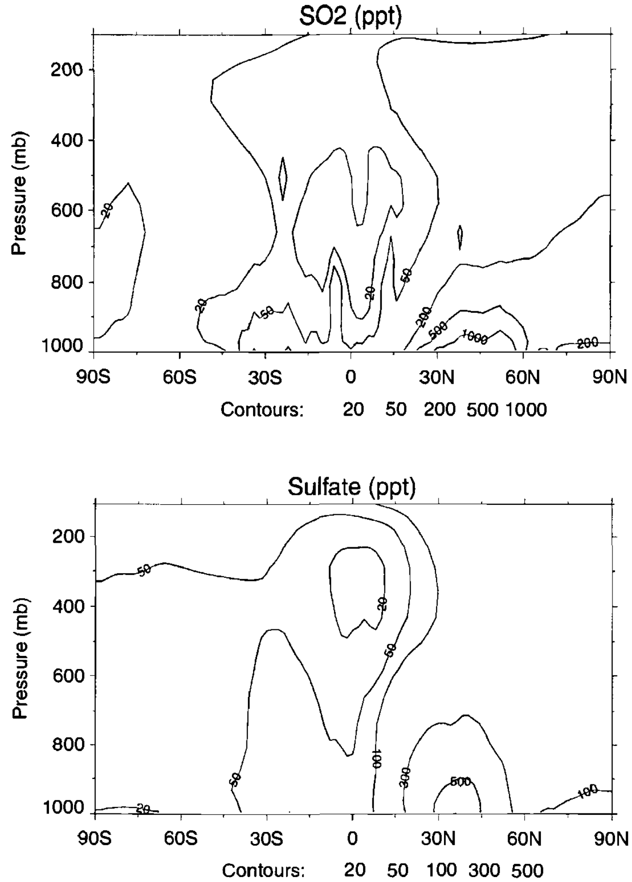


Figure 9 Global altitudinal and latitudinal distribution of the sulfur mixing ratio based on the global chemistry transport model of Chin et al. (1999): (*a*) SO_2 and (*b*) SO_4^{2-}.

to the neighboring oceans and to the free troposphere, while the rest is removed by dry and wet depositions within the source region itself. These authors conclude that dry deposition takes up nearly one third of surface SO_2 emissions directly in the polluted region itself. Thus, although global anthropogenic emissions of SO_2 account for about 70 to 80% of the total emission of sulfate precursors, their contribution to the total sulfate burden is likely to be substantially less.

Yet another interesting result from the model studies is their assessment of the importance of volcanic emissions to the global sulfate burden in the troposphere. Chin and Jacob (1996) have found that this source is a significant contributor to the

TABLE 3 Ranges of Sources, Sinks, Total Mass and Lifetimes of SO_2 and Sulfate from Seven Global Sulfur Models

	Ranges	Median[a]
SO_2		
Total source (Tg S/yr)		95.7
Anthropogenic emission	63.7–92.0	66.5
Biomass burning	2.2–2.9	2.3
Volcanoes	3.4–8.5	5.5
Photochemical production	10.0–24.7	16.9
Total sink (Tg S/yr)		91.2
Gas-phase oxidation	6.1–16.8	9.2
In-cloud oxidation	23.3–55.5	42
Dry deposition	16.0–55.0	35.5
Wet deposition	0–19.9	9.0
Total atmospheric burden (Tg S)	0.2–0.6	0.4
Lifetime (days)	0.6–2.6	1.5
Sulfate		
Total source (Tg S/yr)		50.6
Anthropogenic emission	0–3.5	1.4
Gas-phase production	6.1–16.8	9.2
In-cloud processing	23.3–57.8	40.0
Total sink (Tg S/yr)		50.2
Dry deposition	3.7–17.0	6.7
Wet deposition	34.6–61.0	44.5
Total atmospheric burden (Tg S)	0.3–0.96	0.63
Lifetime (days)	3.9–5.8	4.6

Note: Models are from Langner and Rodhe (1991), Pham et al. (1995), Feichter et al. (1995), Chin et al. (1996), Chuang et al. (1997), Koch et al. (1999), and Barth et al. (1999).

[a] As a result of using median values derived from several different models to define the total source and sink for SO_2 and sulfate, these values do not necessarily balance.

sulfate budget in the middle and upper troposphere. This is due to volcanoes providing direct injection of sulfur gases to upper altitudes where species such as SO_2 typically have lifetimes an order of magnitude longer than that in the boundary layer. The major impact of global volcanic emissions at high altitude is a conclusion also reached by Graf et al. (1998). These investigators found that the global mean radiative forcing by volcanic sulfate aerosols was actually comparable to anthropogenic aerosols.

Table 3 summarizes the global SO_2-sulfate budget results based on the modeling results from several groups (e.g. Langner and Rodhe, 1991; Pham et al., 1995; Feichter et al., 1996; Chin et al., 1996, 2000; Chuang et al., 1997; Koch et al., 1999; Barth et al., 2000). As one might expect, there are a number of differences among these models, especially in their handling of meteorological fields and parameterizations. Even so, all still agree on certain key points. For example, all models assign 70 to 75% of sulfate precursor emissions to anthropogenic activities; 30 to 45% of the primary SO_2 is also estimated to be removed by dry deposition; and finally, it is agreed that concerning the oxidation of SO_2 to sulfate, 65 to 85% of the total is dominated by in-cloud processes. Among the important areas where significant uncertainties still exist is that of fully understanding the levels of SO_2 and sulfate in remote marine regions. As discussed in Section 3, of particular concern is assessing the relative contributions at free tropospheric altitudes of sulfate derived from DMS oxidation versus that from volcanoes and long-range transport of surface-generated continental sources.

5 STRATOSPHERIC SULFUR

The presences of sulfur in the stratosphere in the form of a sulfate aerosol layer, or Junge layer, was first reported in the early 1960s (Junge et al., 1961). Since its discovery, there have been substantial advances in understanding the effects of stratospheric sulfur on climate and atmospheric chemistry. The primary importance of stratospheric sulfur is that it affects Earth's radiative balance. Aerosols can directly scatter incoming solar radiation back to space. This results in a cooling of Earth's surface. By absorbing outgoing infrared radiation, however, they can also cause a warming of the stratosphere. These effects have been observed after major volcanic eruptions (e.g., Labutzke and McCormick, 1992). Stratospheric aerosols can also have an indirect effect on the radiative balance by acting as CCN. For example, they are involved in forming polar stratospheric clouds (PSCs) and possibly in the development of large-scale cirrus clouds. In addition, stratospheric aerosols may play a significant role in stratospheric chemistry by providing surfaces upon which heterogeneous reactions take place. Such reactions appear to be centrally important as a means of modulating stratospheric ozone levels (see, e.g., Hofmann and Solomon, 1989).

The composition of stratospheric aerosols appears to be mainly sulfate (Rosen, 1971). The most likely source of this sulfate is oxidation of SO_2 to form $H_2SO_4(g)$ as discussed above in Section 3. This oxidation step would then be followed by nucleation and condensation. On the basis of numerous observations of stratospheric aerosols over the past 30 years, volcanic eruptions that inject large amounts of SO_2 directly into the stratosphere are now believed to be one of the dominant sources of stratospheric sulfate aerosols. However, because of the presence of a persistent background of aerosol even during periods when no major volcanic eruptions occurred, there has been considerable speculation concerning other possible sources of this aerosol.

The importance of carbonyl sulfide (OCS) as a stratospheric aerosol source was first proposed by Crutzen (1979). As noted in Section 2, carbonyl sulfide is the most abundant sulfur compound in the atmosphere. It is emitted at Earth's surface by natural and anthropogenic sources, and it is also formed by the oxidation of carbon disulfide (CS_2) (Chin and Davis, 1993). Recall, however, that because of its chemical inertness in the troposphere, it is found to have a near uniform mixing ratio (i.e., 500 pptv) throughout this region. Because of this, significant quantities of OCS are transported to the stratosphere where it undergoes photodecomposition and/or oxidation via reactions with $O(^3P)$ atoms and OH radicals. The resulting product SO_2, like that from volcanic injections, is then converted to sulfate aerosol. Early modeling studies supported Crutzen's hypothesis and showed that the flux of OCS into the stratosphere was sufficient to maintain the background sulfate aerosol layer.

Twenty years later, with a far more extensive set of OCS atmospheric observations and with improved laboratory reaction rate data, Chin and Davis (1995) reanalyzed the stratospheric significance of OCS as a source of background sulfate aerosols. They compared the flux of OCS calculated in a one-dimensional model with the flux needed to sustain the background aerosol level. Historically, the background level has been estimated from the ratio of background aerosol mass to aerosol lifetime. Departing from earlier analyses, Chin and Davis (1995) found that OCS could provide only 20 to 50% of the required sulfur. This conclusion was based on two important insights: (1) The so-called background aerosol layer observed during volcanic quiescent periods still contained a significant amount of residual volcanic aerosol, and (2) important sources other than OCS quite likely were also contributing to background sulfate aerosol levels. Although a more recent one-dimensional model study, which included microphysical processes, proposed that a sustainable background sulfate layer could indeed be maintained by OCS oxidation (Zhao et al., 1995), Weisenstein et al. (1997) report results that are much closer to those given earlier by Chin and Davis. Weisenstein et al., using a global two-dimensional model, found that OCS oxidation could only account for half of the background sulfur loading. They also found that convective transport of SO_2 in the tropical troposphere could provide the other half of the background sulfate aerosol. In a still more recent study Mills et al. (1999), based on new measurements by Wilson et al. (1999), have suggested that in addition to SO_2, tropospheric sulfate aerosol at the tropopause could make a significant contribution to the sulfate loading of the stratosphere during quiescent periods. As shown in Figures 9*a* and 9*b*, the concentrations of SO_2 and sulfate at the Northern Hemisphere tropopause can reach 50 and 100 pptv, respectively. Quite clearly, there are still important aspects of the so-called background aerosol layer issue that are still unresolved.

The variability in background sulfate aerosol levels has also drawn considerable attention since human activities may have already perturbed the natural background level. For example, there have been reports published indicating that we could be experiencing as much as a 6 to 8% per year increase in background levels. Although the initial speculation was focused on these increases being tied to anthropogenic emissions of OCS, upon further reflection this explanation has been largely rejected. This follows from the fact that there has been no significant long-term trend in OCS concentrations in the troposphere over the last 20 years. Hofmann (1991) has noted, however, that the increase in background aerosol mass is closely related to increases in sulfur emissions from high-altitude aircraft. Another possible anthropogenic source would involve the direct transport of SO_2 and sulfate from the troposphere, as discussed earlier. The anthropogenic fraction of sulfate in the Northern Hemisphere's upper troposphere can vary from 20% in January to 60 to 80% in July (Chin et al., 2000). Thus, an increase in anthropogenic SO_2 emissions could have made an impact on the stratospheric aerosol level.

A quite different perspective on background stratospheric aerosol trends has been put forward by Chin and Davis (1995). They have raised the question whether one can even reliably define a baseline value for aerosol in an environment that is continually being disturbed by new volcanic injections of sulfur. They point out that there were only 2 years in the 10-year record cited by Sedlacek et al. (1983) and 2 years in the 18-year observations by Hofmann (1990) that could be identified as "volcanic quiescent" periods. In neither case, however, was it possible to convincingly show that the aerosol or sulfate levels observed during these periods were free of any significant volcanic influence. Given the multiyear residence time of volcanic aerosols and the frequency of minor volcanic injections, Chin and Davis (1995) argued that overall there still remains a serious question whether a true background sulfate aerosol level (i.e., one largely uninfluenced by volcanic emissions) has as yet been observed. Thus, critical to any future analyses designed to show the role of tropospheric sulfur compounds in forming stratospheric sulfur aerosol will be the further elucidation of the volcanic component of the so-called background aerosol layer.

REFERENCES

Andreae, M. O., and W. A. Jaeschke, Exchange of sulphur between biosphere and atmosphere over temperate and tropical regions, in R. W. Howarth, J. W. B. Stewart, and M. V. Ivanov (Eds.), *Sulphur Cycling on the Continents: Wetlands, Terrestrial Ecosystems, and Associated Water Bodies*, SCOPE 48, Wiley, Chichester, 1992, pp. 27–61.

Bandy, A. R., D. C. Thomton, B. W. Blomquist, S. Chen, T. P. Wade, J. C. Ianni, G. M. Mitchell, and W. Nadler, Chemistry of dimethyl sulfide in the equatorial Pacific atmosphere, *Geophys. Res. Lett.*, *23*, 741–744, 1996.

Bandy, A. R., D. C. Thomton, and A. R. Driedger III, Airbome measurements of sulfur dioxide, dimethyl sulfide, carbon disulfide, and carbonyl sulfide by isotope dilution gas chromatography/mass spectrometry, *J. Geophys. Res.*, *98*, 23423–23433, 1993.

Barrie, L. A., M. P. Olson, and K. K. Oikawa, The flux of anthropogenic sulphur into the Arctic from mid-latitudes in 1979/80, *Atmos. Environ.*, *18*, 2711–2722, 1989.

Barth, M. C., P. J. Rasch, J. T. Kiehl, C. M. Benkovitz, and S. E. Schwartz, Sulfur chemistry in the NCAR CCM: Description, evaluation, features and sensitivity to aqueous chemistry, *J. Geophys. Res.*, *105*, 1387–1415, 2000.

Bates, T. S., J. A. Calhoun, and P. K. Quinn, Variations in the methanesulfonate to sulfate molar ratio in submicrometer marine aerosol particles over the South Pacific Ocean, *J. Geophys. Res.*, *97*, 9859–9865, 1992a.

Bates, T. S., B. K. Lamb, A. Guenther, J. Dignon, and R. E. Stoiber, Sulfur emissions to the atmosphere from natural sources, *J. Atmos. Chem.*, *14*, 315–337, 1992b.

Bates, T. S., B. J. Huebert, J. L. Gras, F. B. Griffiths, and P. A. Durkee, The international Global Atmospheric Chemistry (IGAC) Project's First Aerosol Characterization Experiment (ACE 1): Overview, *J. Geophys. Res.*, *103*, 16297–16318, 1998.

Berresheim, H., Biogenic sulfur emissions from the Subantarctic and Antarctic oceans, *J. Geophys. Res.*, *92*, 13245–13262, 1987.

Berresheim, H., and F. L. Eisele, Sulfur chemistry in the Antarctic Troposphere Experiment: An overview of project SCATE, *J. Geophys. Res.*, *103*, 1619–1627, 1998.

Berresheim, H., F. L. Eisele, D. J. Tanner, D. S. Covert, L. McInnes, and D. C. Ramsey-Bell, Atmospheric sulfur chemistry and cloud condensation nuclei (CCN) concentrations over the northeastern Pacific coast, *J. Geophys. Res.*, *98*, 12701–112711, 1993.

Berresheim, H., P. Wine, and D. Davis, Sulfur in the atmosphere, in H. B. Singh (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 251–307.

Chameides, W. L., and D. D. Davis, The free radical chemistry of cloud droplets and its impact upon the composition of rain, *J. Geophys. Res.*, *87*, 4863–4877, 1982.

Chameides, W. L., and A. W. Stelson, Aqueous-phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Charlson, R. J., S. E. Schwartz, J. M. Hales, R. D. Cess, J. A. Coakley, J. E. Hansen, and D. J. Hofmann, Climate forcing by anthropogenic aerosols, *Science*, *255*, 423–430, 1992.

Chen, G., D. D. Davis, P. Kasibhatla, A. R. Bandy, D. C. Thornton, B. J. Huebert, A. D. Clarke, and B. Blomquist, A study of DMS oxidation in the tropics: Comparison of Christmas Island field observations of DMS, SO_2, and DMSO with model simulations, *J. Atmos. Chem.*, *37*, 137–160, 2000.

Chin, M., and D. D. Davis, Global sources and sinks of OCS and CS_2 and their distributions, *Global Biogeochem. Cycles*, *7*, 321–337, 1993.

Chin, M., and D. D. Davis, A reanalysis of carbonyl sulfide as a source of stratospheric background sulfur aerosol, *J. Geophys. Res.*, *100*, 8993–9005, 1995.

Chin, M., and D. J. Jacob, Anthropogenic and natural contributions to tropospheric sulfate: A global model analysis, *J. Geophys. Res.*, *101*, 18691–18699, 1996.

Chin, M., D. J. Jacob, G. M. Gardner, M. S. Foreman-Fowler, P. A. Spiro, and D. L. Savoie, A global three-dimensional model of tropospheric sulfate, *J. Geophys. Res.*, *101*, 18667–18690, 1996.

Chin, M., R. Rood, S.-J. Lin, J. F. Muller, and A. Thompson, Atmospheric sulfur cycle simulated in the global model GOCART: Model description and global properties, *J. Geophys. Res.*, *105*, 24671–24687, 2000.

Chuang, C. C., J. E. Penner, K. E. Taylor, A. S. Grossman, and J. J. Walton, An assessment of the radiative effects of anthropogenic sulfate, *J. Geophys. Res.*, *102*, 3761–3778, 1997.

Clarke, A. D., Atmospheric nuclei in the Pacific midtroposphere: Their nature, concentration and evolution, *J. Geophys. Res.*, *98*, 20633–20647, 1993.

Corbett, J. J., and P. S. Fischbeck, Emissions from ship, *Science*, *278*, 823–824, 1997.

Corbett, J. J., P. S. Fischbeck, and S. N. Pandis, Global nitrogen and sulfur inventories for oceangoing ships, *J. Geophys. Res.*, *104*, 457–3470, 1999.

Crutzen, P. J., The possible importance of OCS for the sulfate layer of the stratosphere, *Geophys. Res. Lett.*, *3*, 73–76, 1979.

Davis, D. D., G. Chen, A. Bandy, D. Thornton, F. Eisele, L. Mauldin, D. Tanner, D. Lenschow, H. Fuelberg, B. Huebert, J. Heath, A. Clarke, and D. Blake, Dimethyl sulfide oxidation in the equatorial Pacific: Comparison of model simulations with field observations for DMS, SO_2, H_2SO_4(g), MSA(g), MS, and NSS, *J. Geophys. Res.*, *104*, 5765–5784, 1999.

Davis, D. D., (unpublished results) in text.

Davis, D. D., G. Chen, P. Kasibhatla, A. Jefferson, D. Tanner, F. Eisele, D. Lenschow, W. Neff, and H. Berresheim, DMS oxidation in the Antarctic marine boundary layer I: Comparison of model simulations and field observations for DMS, DMSO, $DMSO_2$, H_2SO_4(g), MSA(g), and MSA(p), *J. Geophys. Res.*, *103*, 1657–1678, 1998.

De Bruyn, W. J., T. S. Bates, J. M. Cainey, and E. S. Saltzman, Shipboard measurements of dimethyl sulfide and SO_2 southwest of Tasmania during the First Aerosol Characterization Experiment (ACE 1), *J. Geophys. Res.*, *103*, 16703–16711, 1998.

Feichter, J., E. Kjellstrom, H. Rodhe, F. Dentener, J. Lelieveld, and G.-J. Roelofs, Simulation of the tropospheric sulfur cycle in a global climate model, *Atmos. Environ.*, *30*, 1693–1708, 1996.

Finlayson-Pitts, B., and J. N. Pitts, *Atmospheric Chemistry: Fundamentals and Experimental Techniques*, Wiley, New York, 1986.

Graf, H.-F., B. Langmann, and J. Feichter, The contribution of Earth degassing to the atmospheric sulfur budget, *Chem. Geol.*, *147*, 131–145, 1998.

Grosjean, D., Photooxidation of methyl sulfide, ethyl sulfide, and methanethiol, *Environ. Sci. Technol.*, *18*, 460–468, 1984.

Hameed, S., and J. Dignon, Global emissions of nitrogen and sulfur oxides in fossil fuel combustion: 1970–1986, *J. Air Waste Mgmt. Assoc.*, *42*, 159–163, 1992.

Hatakeyama, S., K. Izumi, and H. Akimoto, Yield of SO_2 and formation of aerosol in the photo-oxidation of DMS under atmospheric conditions, *Atmos. Environ.*, *19*, 135–141, 1985.

Hoell, Jr., J. M., D. D. Davis, D. J. Jacob, M. O. Rogers, R. B. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM—Tropics A, August–September, 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hofmann, D. J., Increase in the stratospheric background sulfuric acid aerosol mass in the past 10 year, *Science*, *248*, 996–1000, 1990.

Hofmann, D. J., Aircraft sulphur emissions, *Nature*, *349*, 659, 1991.

Hofmann, D. J., and S. Solomon, Ozone destruction through heterogeneous chemistry following the eruption of El Chichon, *J. Geophys. Res.*, *94*, 5029–5041, 1989.

Hynes, A. J., and P. H. Wine, The atmospheric chemistry of dimethylsulfoxide (DMSO) kinetics and mechanism of the OH + DMSO reaction, *J. Atmos. Chem.*, *24*, 23–37, 1996.

Hynes, A. J., P. H. Wine, and D. H. Semmes, Kinetics and mechanism of OH reactions with organic sulfides, *J. Phys. Chem.*, *90*, 4148–4156, 1986.

Jefferson, A., D. J. Tanner, F. L. Eisele, D. D. Davis, G. Chen, J. Crawford, J. W. Huey, A. L. Torres, and H. Berresheim, OH photochemistry and methane sulfonic acid formation in the coastal Antarctic boundary layer, *J. Geophys. Res.*, *103*, 1647–1656, 1998.

Junge, C. E., C. W. Chagnon, and J. E. Manson, Stratospheric aerosols, *J. Meteorol.*, *18*, 81–108, 1961.

Koch, D., D. Jacob, I. Tegen, D. Rind, and M. Chin, Tropospheric sulfur simulation and sulfate direct radiative forcing in the Goddard Institute for Space Studies general circulation model, *J. Geophys. Res.*, *104*, 23799–23822, 1999.

Kreidenweis, S. M., and J. H. Seinfeld, Nucleation of sulfuric acid–water and methanesulfonic acid–water solution particles: Implications for the atmospheric chemistry of organosulfur species, *Atmos. Environ.*, *22*, 283–296, 1988.

Krouse, H. R., and R. G. L. McCready, Reductive reactions in the sulfur cycle, in P. A. Trudinger and D. J. Swaine (Eds.), *Biogeochemical Cycling of Mineral-Forming Elements*, Elsevier, Amsterdam, 1979, pp. 315–368.

Labutzke, K., and M. P. McCormick, Stratospheric temperature increase due to Pinatobo aerosols, *Geophys. Res. Lett.*, *19*, 207–210, 1992.

Langner, J., and H. Rodhe, A global three-dimensional model of the tropospheric sulfur cycle, *J. Atmos. Chem.*, *13*, 225–263, 1991.

Lee, Y.-N., and X. Zhou, Aqueous reaction kinetics of ozone and dimethlysulfide and its atmospheric implications, *J. Geophys. Res.*, *99*, 3597–3605, 1994.

Lelieveld, J., and J. Heintzenberg, Sulfate cooling effect on climate through in-cloud oxidation of anthropogenic SO_2, *Science*, *258*, 117–120, 1992.

Lovejoy, E. R., D. R. Hanson, and L. G. Huey, Kinetics and products of the gas-phase reaction of SO_3 with water, *J. Phys. Chem.*, *100*, 19911–19916, 1996.

Martin, L. R., Kinetic studies of sulfite oxidation in aqueous solution, in J. G. Calvert (Ed.), *SO_2, NO and NO_2 Oxidation Mechanisms: Atmospheric Considerations*, Butterworth, Boston, 1984, pp. 63–100.

Mauldin III, R. L., D. J. Tanner, and F. L. Eisele, Measurements of OH during PEM—Tropics A, *J. Geophys. Res.*, *104*, 5817–5827, 1999a.

Mauldin III, R. L., D. J. Tanner, J. A. Heath, B. J. Huebert, and F. L. Eisele, Observations of H_2SO_4 and MSA during PEM—Tropics, *J. Geophys. Res.*, *104*, 5801–5816, 1999b.

Mills, J. M., O. B. Toon, and S. Solomon, A microphysical analysis of non-volcanic sources of atmospheric sulfate, in *EOS Trans.*, AGU, 1999 San Francisco Fall Meeting, Vol. 80, 1999, p. F169.

Niki, H., P. D. Maker, C. M. Savage, and L. P. Breitenbach, An FTIR study of the mechanism for the reaction HO + CH_3SCH_3, *Int. J. Chem. Kinet.*, *15*, 647–654, 1983.

Nowak, J., D. D. Davis, G. Chen, F. Eisek, D. Tanner, L. Mauldin III, C. Cantrell, E. Koscinch, A. Baudy, D. Thornton, and A. Clarke, *Geophys. Res. Ltt.*, p2201–2204, 2001.

Pham, M., J.-F. Müller, G. P. Brasseur, C. Granier, and G. Mégie, A three-dimensional study of the tropospheric sulfur cycle, *J. Geophys. Res.*, *100*, 26061–26092, 1995.

Rosen, J. M., The boiling point of stratospheric aerosols, *J. Appl. Meteor.*, *10*, 1044–1046, 1971.

Saltzman, E. S., D. L. Saugie, R. G. Zika, and J. M. Prosporo, Methane sulfonic acid and non-sea-salt sulfate in Pacific air: regional and seasonal variations. *J. Atmos. Chem.*, 4, 227–240, 1986.

Saltzman, E. S., S. A. Yvon, and P. A. Matrai, Low-level detection of atmospheric sulfur dioxide measurement using HPLC/fluorescence detection, *J. Atmos. Chem.*, *17*, 73–90, 1993.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-sea-salt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Savoie, D. L., J. M. Prospero, and E. S. Saltzman, Nitrate, non sea-salt sulfate and nitrate in trade wind aerosols at Barbados: Evidence for long range transport, *J. Geophys. Res.*, *94*, 5069–5080, 1989.

Sedlacek, W. A., E. J. Mroz, A. L. Lazrus, and B. W. Gandrud, A decade of stratospheric sulfate measurements compared with observations of volcanic eruptions, *J. Geophys. Res.*, *88*, 3741–3776, 1983.

Sievering, H., J. Boatman, E. Gorman, Y. Kim, L. Anderson, G. Ennis, M. Luria, and S. Pandis, Removal of sulphur from the marine boundary layer by ozone oxidation in sea-salt aerosol, *Nature*, *360*, 571–573 , 1992.

Singh, H. B., A. Thakur, Y. E. Chen, and M. Kanakidou, Tetrachloroethene as an indicator of low Cl atom concentrations in the troposphere, *Geophys. Res. Lett.*, *23*, 1529–1532, 1996.

Spiro, P. A., D. J. Jacob, and J. A. Logan, Global inventory of sulfur inventory of sulfur emissions with 1° × 1° resolution, *J. Geophys. Res.*, *97*, 6023–6036, 1992.

Thornton, D. C., A. R. Bandy, B. W. Bloomquist, A. R. Driedger, and T. P. Wade, Sulfur dioxide distribution over the Pacific Ocean 1991–1996, *J. Geophys. Res.*, *104*, 5845–5854, 1999.

Turnipseed, A. A., and A. R. Ravishankara, The atmospheric oxidation of dimethyl sulfide: Elementary steps in a complex mechanism, in G. Restelli, and G. Angeletti (Eds.), *Dimethylsulphide: Oceans, Atmosphere and Climate*, Kluwer, Dordrecht, 1993, pp. 185–196.

Urbanski, S. P., R. E. Stickel, and P. H. Wine, Mechanistic and kinetic study of the gas-phase reaction of hydroxyl radical with dimethyl sulfoxide, *J. Phys. Chem.*, *102*, 10522–10529, 1998.

Weisenstein, D. K., G. K. Yue, M. K. W. Ko, N. D. Sze, J. M. Rodriguez, and C. J. Scott, A two-dimensional model of sulfur species and aerosols, *J. Geophys. Res.*, *102*, 13019–13035, 1997.

Wilson, J. O., O. A. Brock, and J. J. Jonsson, In situ measurements of aerosol properties in the upper troposphere and lower stratosphere: Is the Pinatubo aerosol still decaying? in *EOS. Trans.*, AGU, 1999 Fall Meeting, Vol. 80, 1999, p. F169.

Yin, F., D. Grossjean, and J. H. Seinfeld, Photooxidation of dimethyl sulfide and dimethyl disulfide, *J. Atmos. Chem.*, *11*, 309–399, 1990.

Yvon, S. A., E. S. Saltzman, D. J. Cooper, T. S. Bates, and A. M. Thompson, Atmospheric sulfur cycling in the tropical Pacific marine boundary layer (12°S, 135°W): A comparison of field data and model results 2. Sulfur dioxide, *J. Geophys. Res.*, *101*, 6911–6918, 1996.

Zhao, J.-X, R. P. Turco, and O. B. Toon, A model simulation of volcanic aerosol evolution in the stratosphere, *J. Geophys. Res.*, *100*, 7315–7328, 1995.

CHAPTER 9

CONVECTIVE TRANSPORT

KENNETH E. PICKERING

1 INTRODUCTION

In the early 1980s it was recognized that observed free tropospheric mixing ratios of some trace gases could not be explained simply by large-scale transport and eddy diffusion. Crutzen and Gidel (1983), Gidel (1983), and Chatfield and Crutzen (1984) hypothesized that convective clouds played an important role in rapid atmospheric vertical transport of trace species and tested parameterizations of convective transport in atmospheric chemical models. At nearly the same time evidence was shown of venting of the boundary layer by shallow fair weather cumulus clouds (e.g., Greenhut et al., 1984; Greenhut, 1986). Field experiments were conducted in 1985 that resulted in verification of the hypothesis that deep convective clouds are instrumental in atmospheric transport of trace constituents (Dickerson et al., 1987; Garstang et al., 1988). Once pollutants are lofted to the middle and upper troposphere, they typically have a much longer chemical lifetime and with the generally stronger winds at these altitudes they can be transported large distances from their source regions. Photochemical reactions occur during this long-range transport. Pickering et al. (1990) demonstrated that venting of boundary layer pollutants by convective clouds (both shallow and deep) causes enhanced ozone production in the free troposphere. Therefore, convection aids in the transformation of local pollution into a contribution to global atmospheric pollution.

Field studies have established that downward transport of larger O_3 and NO_x mixing ratios from the free troposphere to the boundary layer is an important process over the remote oceans (e.g., Piotrowicz et al., 1991), as well as the upward transport of very low O_3 mixing ratios from the boundary layer to the upper troposphere (Kley et al., 1996). Global modeling by Lelieveld and Crutzen (1994) suggests that the downward mixing of O_3 into the boundary layer is the dominant global effect of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

deep convection. Some indications of downward transport of O_3 from higher altitudes (possibly from the stratosphere) in the anvils of thunderstorms have been observed (Dickerson et al., 1987; Poulida et al., 1996; Suhre et al., 1997). Ozone is most effective as a greenhouse gas in the vicinity of the tropopause. Therefore, changes in the vertical profile of O_3 in the upper troposphere caused by deep convection have important radiative forcing implications for climate.

More detailed discussion of observations of convective transport are presented in Section 2. Simulation of convective transport in cloud-resolving models and its parameterization in larger-scale models is discussed in Section 3, as well as implications for O_3 production following convective redistribution.

2 OBSERVATIONS

Venting by Nonprecipitating Cumulus Clouds

Some fraction of shallow fair weather cumulus clouds actively vent boundary layer pollutants to the free troposphere (Stull, 1985). The first airborne observations of this phenomenon were conducted by Greenhut et al. (1984) over a heavily urbanized area, measuring the in-cloud flux of ozone in a relatively largc cumulus cloud. An extension of this work was reported by Greenhut (1986) in which data from over 100 aircraft penetrations of isolated nonprecipitating cumulus clouds over rural and suburban areas were obtained. Ching and Alkezweeny (1986) reported tracer (SF_6) studies associated with nonprecipitating cumulus (fair weather cumulus and cumulus congestus). Their experiments showed that the active cumulus clouds transported mixed layer air upward into the overlying free troposphere and suggested that active cumuli can also induce rapid downward transport from the free troposphere into the mixed layer. A UV-DIAL (ultraviolet differential absorption lidar) provided space-height cross sections of aerosols and ozone over North Carolina in a study of cumulus venting reported by Ching et al. (1988). Data collected on evening flights showed regions of cloud debris containing aerosol and ozone in the lower free troposphere in excess of background, suggesting that significant vertical exchange had taken place during afternoon cumulus cloud activity. Efforts have also been made to estimate the vertical transport by ensembles of nonprecipitating cumuli in regional chemical transport models (e.g., Vukovich and Ching, 1990).

Deep Convection

Midlatitudes. The first unequivocal observations of deep convective transport of boundary layer pollutants to the upper troposphere were documented by Dickerson et al. (1987). Instrumentation aboard three research aircraft measured CO, O_3, NO, NO_x, NO_y, and hydrocarbons in the vicinity of an active mesoscale convective system near the Oklahoma/Arkansas border during the 1985 PRE-STORM experiment. Anvil penetrations about 2 h after maturity found greatly enhanced mixing ratios of all of the aforementioned species compared with outside of the cloud. Among the species measured, CO is the best tracer of upward convective transport because it is produced primarily in the boundary layer and has an atmospheric lifetime much longer than the time scale of a thunderstorm. In the observed storm CO measurements exceeded 160 ppbv as high as 11 km, compared with ~70 ppbv outside of the cloud (Fig. 1*a*). Nonmethane hydrocarbons (NMHC) with moderate lifetimes can also trace convective transport from the boundary layer. Ozone can also be an indicator of convective transport; in the polluted troposphere large ozone values will indicate upward transport from the boundary layer, but in the clean atmosphere such values are indicative of downward transport from the uppermost troposphere or lowermost stratosphere. In this case measured ozone in the upper rear portion of the anvil peaked at 98 ppbv, while boundary layer values were only ~65 ppbv (Fig. 1*b*). It is likely that some higher ozone stratospheric air mixed into the anvil. Because lightning makes major contributions to reactive nitrogen in thunderstorms, NO_x measurements are unsuitable as a convective tracer.

The large amount of vertical trace gas transport noted by Dickerson et al. (1987) cannot, however, be extrapolated to all convective cells. Pickering et al. (1988) reported airborne measurements of trace gases taken in the vicinity of a line of towering cumulus and cumulonimbus clouds that also occurred during PRE-STORM. In this case trace gas mixing ratios in the tops of these clouds were near ambient levels. Meteorological analyses showed that these clouds were located above a cold front that prevented entry of air from the boundary layer directly below or near the clouds. Instead, the air entering these clouds likely originated in the layer immediately above the boundary layer, which was quite clean. Enhanced values of ozone precursor gases were found in the upper troposphere during another PRE-STORM flight conducted in clear air (Pickering et al., 1989). These observations were identified through correlation analysis as indicative of air with a recent boundary layer source and were traced through back trajectory analysis to deep convection that occurred 600 km upstream. Luke et al. (1992) summarized the air chemistry data from all 18 flights during PRE-STORM by categorizing each case according to synoptic flow patterns. Storms in the maritime flow regime transported large amounts of CO, O_3, and NO_y into the upper troposphere, with the midtroposphere remaining relatively clean. During frontal passages a combination of stratiform and convective clouds mixed pollutants more uniformly into the middle and upper levels; high mixing ratios of CO were found at all altitudes.

Other flights in the vicinity of convective storms over the continental United States were reported by Kleinman and Daum (1991), showing a strong decrease of aerosol particles and water vapor with altitude. However, CO and NO_y were more uniformly distributed in the vertical. Plumelike features, attributed to convective outflow, were noted at high altitude in which mixing ratios of boundary layer pollutants increased by 50% or more above background over a distance of several kilometers. Within these features aerosols and water vapor were enhanced over background values, but these soluble substances were always depleted relative to the insoluble species such as CO, suggesting in-cloud removal of the soluble material.

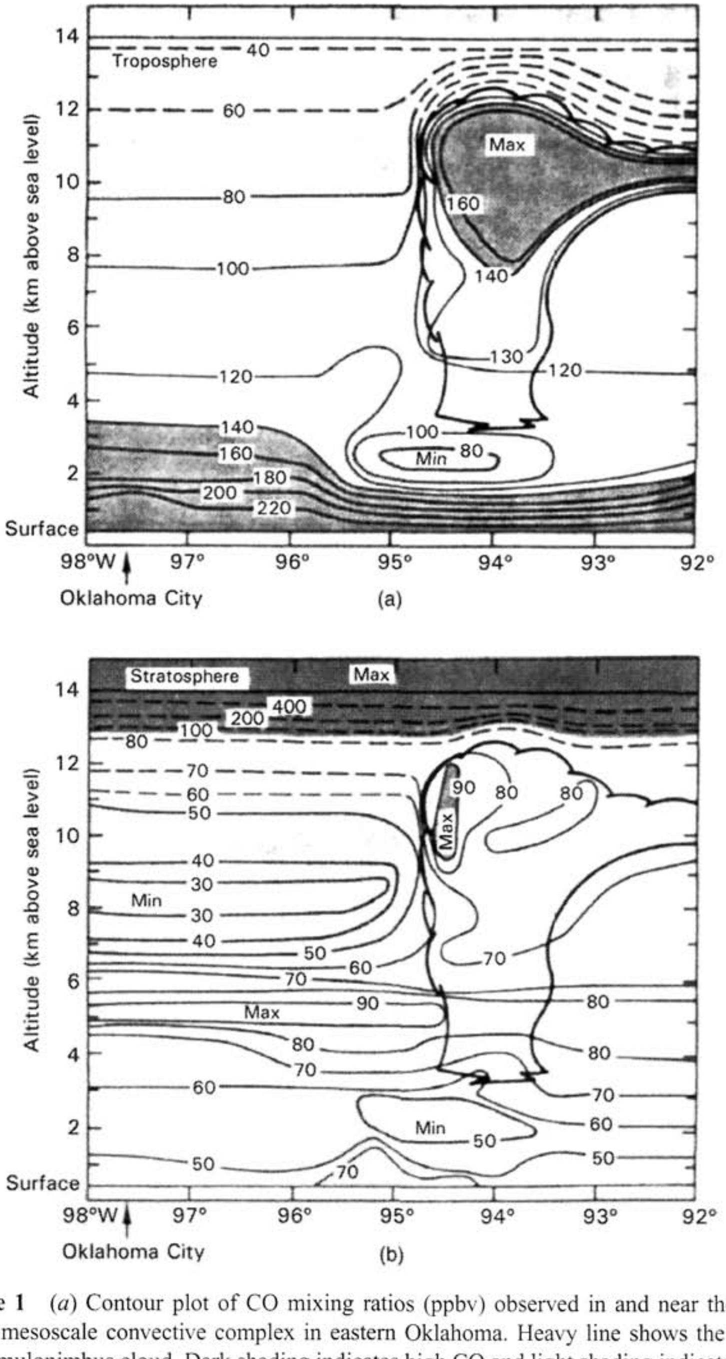


Figure 1 (*a*) Contour plot of CO mixing ratios (ppbv) observed in and near the June 15, 1985, mesoscale convective complex in eastern Oklahoma. Heavy line shows the outline of the cumulonimbus cloud. Dark shading indicates high CO and light shading indicates low CO. Dashed contour lines are plotted according to climatology since no direct measurements were made in that area. (*b*) Same as (*a*) but for ozone (ppbv). From Dickerson et al. (1987).

Poulida et al. (1996) reported observations taken prior to, in, and around a squall line over North Dakota that evolved into a mesoscale convective complex. In this case the anvil extended well into what used to be the stratosphere. Air in the anvil was characterized by low concentrations of O_3 (Fig. 2) and high CO, NO, and NO_y relative to outside the cloud. This layer of tropospheric air lay above a tongue of stratospheric air, indicating that extensive stratosphere–troposphere exchange had occurred. The flux of O_3 into the troposphere and the fluxes of water vapor, CO, NO_y, and hydrocarbons into the stratosphere were estimated for the storm. If only a small fraction of this material from such anvils remained in the stratosphere, it likely dominates the chemistry of the lower stratosphere in this midlatitude region.

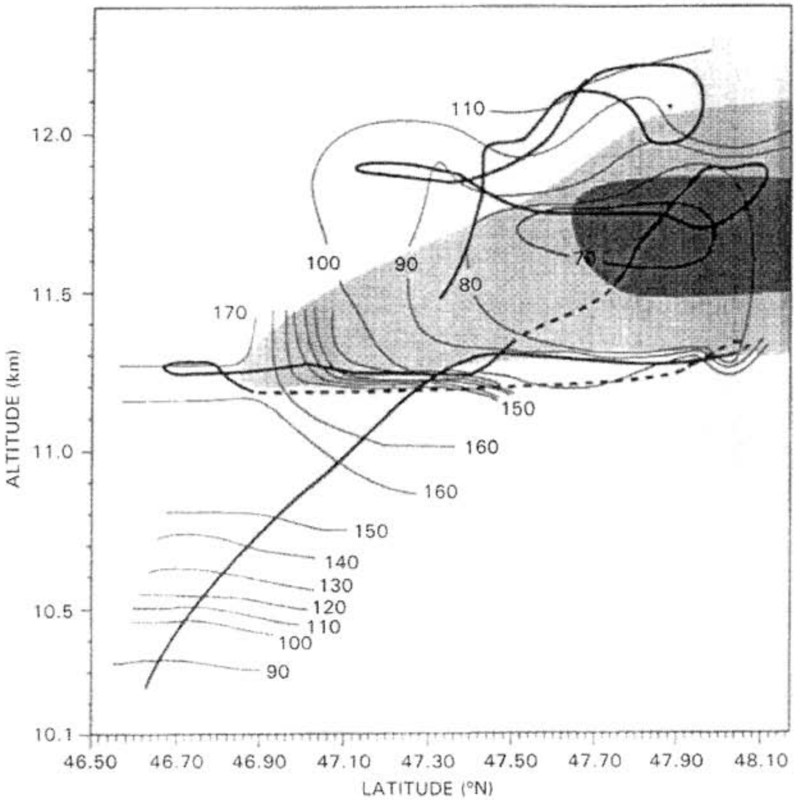


Figure 2 Ozone concentrations in the anvil of a mesoscale convective complex over North Dakota on June 28, 1989. Heavy line indicates flight track projected onto a vertical plane. Thin lines are ozone isopleths every 10 ppbv. Shading shows location of anvil based on aircraft ice particle measurements. Heavier shading indicates greater particle concentrations. From Poulida et al. (1996).

Tropics. Several deep convection experiments with chemical measurements have been conducted in the tropics. Thompson et al. (1997) have summarized many of these results concerning convective transport of trace gases and their consequences for tropospheric ozone production. Garstang et al. (1988) reported measurements taken in front and behind a dry-season squall line over the Amazon rainforest during the NASA ABLE 2A (Amazon Boundary Layer Experiment) project in 1985. The importance of specific processes within the storm (updrafts and downdrafts) as well as the net result of convective transport (atmospheric overturning) were noted. Since the measurements were confined to the lowest 5 km, downward transport of chemical tracers (e.g., ozone) was the most evident feature (Fig. 3). A major emphasis was placed on sampling convective systems during the ABLE 2B wet-season experiment in the same region in 1987. Scala et al. (1990) reported on a locally occurring ABLE 2B convective system, showing that trace gases in the lower troposphere in the wake of the system were well mixed in the vertical. NO measurements behind the storm were greater than ahead of the system, indicating downward transport from above. However, the NO mixing ratios were low enough that ozone production/destruction rates were very small, allowing ozone to be considered a valid tracer of convective

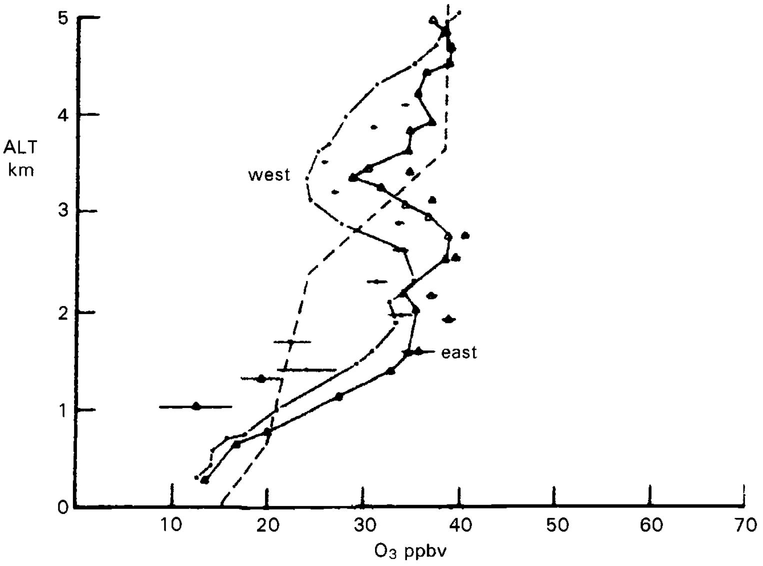


Figure 3 Vertical profiles of ozone concentration along the east and west sides of the August 3, 1985, squall line observed in Brazil in ABLE 2A. The mean profile of ozone from flights in undisturbed weather is shown with a dashed line. Means and standard deviations of ozone from UV-DIAL measurements are shown with symbols and horizontal lines. From Garstang et al. (1988).

transport. Aided by a convective cloud model, Scala et al. (1990) concluded that this system showed that deep undilute convective transport in closed conduits as suggested by Riehl and Simpson (1979) may not necessarily occur in very moist continental tropical systems; the conduits appeared to leak.

Pickering et al. (1996) reported data from a flight of the NASA DC-8 aircraft over Brazil during the TRACE-A (Transport and Atmospheric Chemistry near the Equator—Atlantic) experiment conducted during the biomass burning season of 1992. Outflow from mesoscale convective systems was sampled at 9.5 and 11.3 km showing enhancement of CO mixing ratios typically by a factor of 3 above background (200 to 300 vs. 90 ppbv; see Fig. 4) and significant increases in NO_x and hydrocarbons. Both lightning and transport made important contributions to the enhanced NO_x at cloud outflow levels. Cloud-resolving and regional transport models, a trajectory model, and a photochemical model were used in illustrating the importance of convective events in the ozone budget of the South Atlantic region (see further details in Section 3).

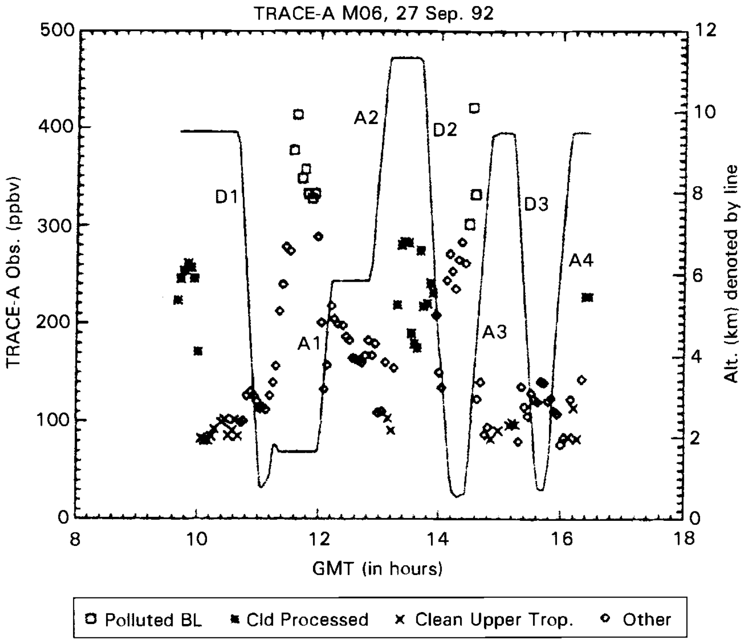


Figure 4 Summary of CO (ppbv) measurements from NASA DC-8 aircraft taken on September 27, 1992, north of Brasilia. Ascents (A) and descents (D) are noted. Three regimes are denoted with indicated symbols and are defined as follows: polluted BL, altitude < 4 km and CO > 300 ppbv; cloud-processed, altitude > 6 km and CO > 150 ppbv; and clean upper troposphere, altitude > 6 km and CO < 120 ppbv. From Pickering et al. (1996).

Over remote marine areas the effects of deep convection on trace gas distributions differ from that over moderately polluted continental regions. Chemical measurements taken by the NASA ER-2 aircraft during the Stratosphere-Troposphere Exchange Project (STEP) off the northern coast of Australia show the influence of very deep convective events. Between 14.5 and 16.5 km on the February 2–3, 1987, flight, perturbations in the chemical profiles were noted that included pronounced maxima in CO, water vapor, CCN and minima of NO_y and ozone (Pickering et al., 1993). Trajectory analysis showed that these air parcels likely were transported from convective cells 800 to 900 km upstream. Very low boundary layer mixing ratios of NO_y and ozone in this remote region were apparently transported upward in the convection. A similar result was noted in CEPEX (Central Equatorial Pacific Experiment; Kley et al., 1996) where a series of ozonesonde ascents showed very low upper tropospheric ozone following deep convection.

Data from convective outflow in the NASA PEM–West A and B experiments (Pacific Exploratory Mission) have been reported by Newell et al. (1996) and by Kawakami et al. (1997). Newell et al. (1996) described sampling of a typhoon in the western Pacific. Boundary layer inflow contained low values of O_3, CO, and hydrocarbons, but high values of dimethylsulfide (DMS). There was no evidence of downward entrainment of stratospheric air into the eye region based on ozone measurements. The DMS data suggested substantial entrainment of boundary layer air into the system, particularly in the eyewall region. Kawakami et al. (1997) reported very low NO_y mixing ratios in the upper troposphere during the February PEM–West B flights between 1°N and 14°N. These measurements were accompanied by very low ozone and large mixing ratios of water vapor and CH_3I, suggesting that the low NO_y values were likely due to convetive transport of tropical marine boundary layer air. Other upper tropospheric measurements showed enhanced NO and high NO_x/NO_y ratios accompanied by low CO, indicative of NO production by lightning.

Danielsen (1993) presented evidence from Darwin, Australia, ER-2 flights in STEP that rapid vertical irreversible transport of lower tropospheric air into the lower tropical statosphere occurs in convective cloud turrets and by large-scale upwelling in tropical cyclones. Suhre et al. (1997) reported O_3 measurements from the tropical Atlantic upper troposphere (10–12 km) taken from commercial aircraft showing mixing ratios of 100 to 500 ppbv at a horizontal scale of 5 to 80 km in the proximity of deep convection. It is hypothesized that there is either direct input of stratospheric O_3 into the anvils of these systems or there is downward convective transport of O_3-rich air that has been transported quasi-isentropically from the extratropical stratosphere.

3 MODELING

Cloud Scale

The Goddard Cumulus Ensemble (GCE) model (Tao and Simpson, 1993) has been used by Pickering et al. (1991, 1992a,b, 1993, 1996), Scala et al. (1990), and

Stenchikov et al. (1996) in the analysis of convective transport of trace gases. The cloud model is nonhydrostatic and contains detailed representation of cloud microphysical processes. Two- and three-dimensional versions of the model have been applied in transport analyses. The initial conditions for the model are usually from a sounding of temperature, water vapor, and winds representative of the region of storm development. Model-generated wind fields can be used to perform air parcel trajectory analyses and tracer advection calculations. Scala et al. (1990) conducted detailed air parcel trajectory analyses for an ABLE 2B storm to investigate flow patterns within the system. In this case the model showed that more than 50% of the air transported to the anvil region originated at or above 6 km, not from the boundary layer via undilute core updrafts. The trajectories also allowed diagnosis of a rotor-type circulation in the low to mid levels of the storm, which was responsible for thorough mixing of the lower troposphere (Fig. 5).

Pickering et al. (1991) used trajectory analyses derived from the GCE model wind fields for the ABLE 2A storm observed by Garstang et al. (1988) to identify air parcels that were undisturbed or modified by the storm (Fig. 6). Tracer transport calculations were performed for CO, O_3, and NO_x, and difference fields showing the changes in mixing ratio of each of these species due to convective transport were computed (Fig. 7). Enhanced values of ozone precursors (NO_x and CO) in a biomass burning haze layer just above the boundary layer were redistributed upward and downward by the storm. Profiles taken from the two-dimensional tracer fields before and after convective transport were used in a one-dimensional photochemical

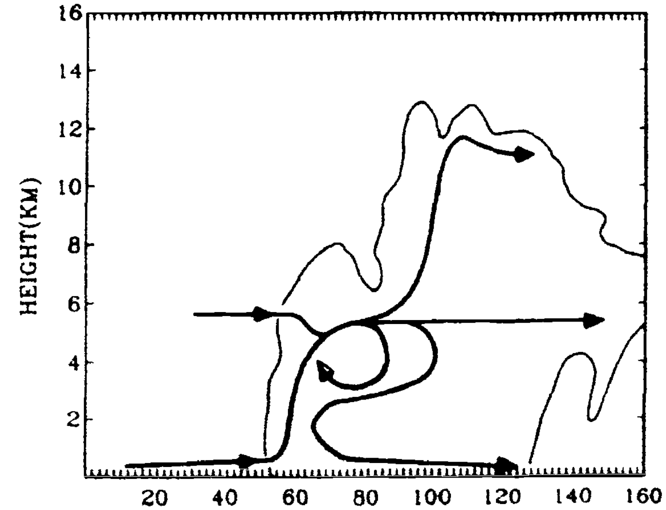


Figure 5 Composite schematic of the predominant transport pathways for the May 6, 1987, ABLE 2B simulated squall convection based on backward and forward trajectory analyses. The model cloud outline at 300 min in the simulation is shown. The horizontal dimension is 80 km. From Scala et al. (1990).

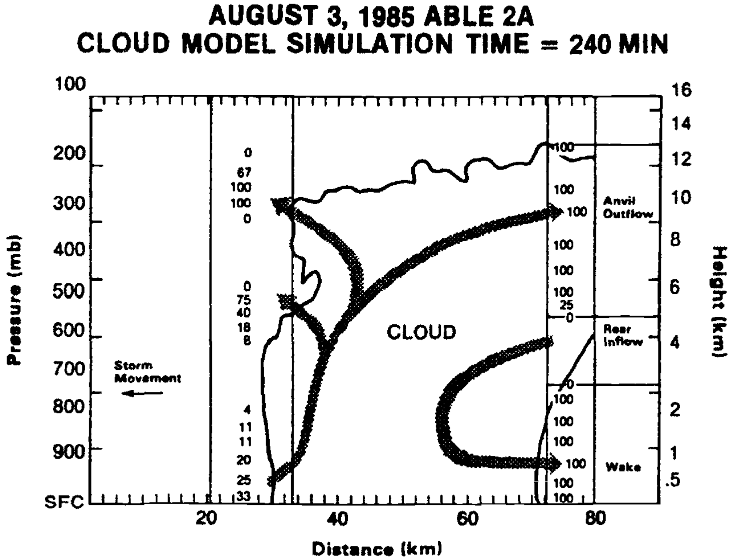


Figure 6 Summary of back trajectories produced by the GCE model for the August 3, 1985, ABLE 2A squall line. Numbers in the vertical column ahead and behind the cloud indicate the percentage of the air at that altitude that is outflow from the cloud. Most of the air pumped out of the boundary layer exits from the anvil (8 to 12 km) and the air in the "wake" has also been processed. Most of the air in the boundary layer ahead of the storm is unperturbed. Arrows indicate main flow paths. From Pickering et al. (1991).

model to estimate ozone production rates. The upward transport of O_3 precursors changed the photochemical tendency of the upper troposphere from that of O_3 destruction to that of production. The same storm dynamics were used in a sensitivity study of convective transport and subsequent free tropospheric O_3 production for conditions of more intense biomass burning pollution (Pickering et al., 1992b). Assuming a pristine middle and upper troposphere prior to convection, enhancements of O_3 production postconvection potentially could be as great as a factor of ~ 50.

Similar methods were used by Pickering et al. (1992a) to examine transport of urban plumes by deep convection. Transport of the Oklahoma City plume by the June 10–11, 1985, PRE-STORM squall line and of the Manaus, Brazil, plume by the April 26, 1987, ABLE 2B squall line were simulated with the two-dimensional GCE model. In the Oklahoma event forward trajectories from the boundary layer at the leading edge of the storm showed that almost 75% of the low-level inflow was transported to altitudes exceeding 8 km. Over 35% of the air parcels reached altitudes over 12 km. For the Amazonian storm, 50% of the trajectories indicated transport to altitudes greater than 12 km. However, nearly 25% of the air parcels indicated air being detrained from the rear of the cloud between 4 and 8 km, and

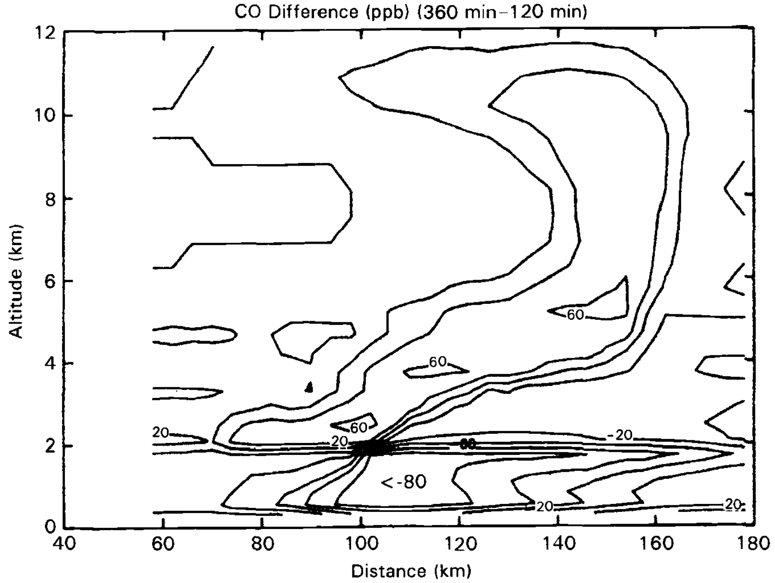


Figure 7 Difference (postconvection minus undisturbed) in model-computed CO tracer concentrations for the August 3, 1995, ABLE 2A squall line. Increases in CO are noted throughout the main updraft region and anvil, and decreases are seen in the downdraft region. From Pickering et al. (1991).

15% became involved in a rotor-type circulation located behind the convective updrafts. In each of these cases tracer transport calculations were performed for CO, NO_x, O_3, and hydrocarbons. The three-dimensional version of the GCE model has also been run for the June 10–11, 1985, PRE-STORM case and for the September 26, 1992, event from TRACE-A. Figure 8 shows the redistributed CO from the rural Oklahoma boundary layer as simulated by the model-generated three-dimensional wind field. Free tropospheric O_3 production enhancement of a factor of 2.5 for Oklahoma rural air and ~4 for the Oklahoma City case were calculated, while with a pristine preconvective upper troposphere an enhancement of a factor of 35 was estimated for the Manaus, Brazil, case.

Stenchikov et al. (1996) used the two-dimensional GCE model to simulate the North Dakota storm observed by Poulida et al. (1996). This storm showed the unusual feature of an anvil formed well within the stratosphere. The increase of CO and water vapor above the altitude of the preconvective tropopause was computed in the model. The total mass of CO across the model domain above this level increased by almost a factor of 2 during the convective event. Downward transport of ozone from the stratosphere was noted in the simulation in the rear anvil. Wang et al. (1995) simulated a tropical convective storm observed during CEPEX using the cloud dynamics and cloud transport models of Wang and Chang (1993).

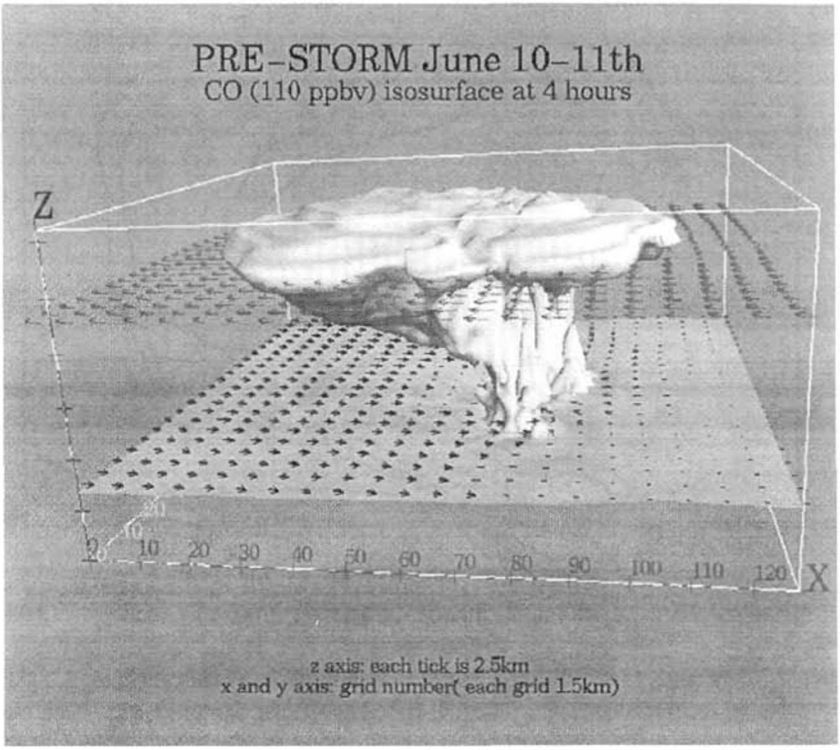

PRE-STORM June 10–11th
CO (110 ppbv) isosurface at 4 hours
Z
0
10
20
30
40
50
60
70
80
90
100
110
120
X
z axis: each tick is 2.5km
x and y axis: grid number(each grid 1.5km)

Figure 8 Isosurface of CO mixing ratio (110 ppbv) computed for the June 10, 1985, PRE-STORM squall line over Oklahoma using the three-dimensional GCE model. Measured rural CO mixing ratios used as initial conditions.

The simulated cloud tower extended into the lower stratosphere and a widespread anvil was produced. Intense mixing of boundary layer air into the cloud resulted in low ozone throughout the tower and the anvil. Stratospheric air with high-ozone mixing ratios was brought into the upper portion of the anvil. The model did not show any significant transport of boundary layer gases into the stratosphere.

Regional

Regional estimates of deep convective transport have been made through use of a traveling one-dimensional model, regional transport models driven by parameterized convective mass fluxes from mesoscale meteorological models, and a statistical-dynamical approach. Chatfield and Delany (1990) simulated convective transport for a hypothetical case over South America during the biomass burning season using a traveling one-dimensional model containing cloud-scale vertical transport and chemistry. They showed that the "mix-then-cook" scenario of rapid vertical transport of ozone precursors in deep convection allowed a more persistent increase in the tropospheric ozone column over a wide region than did the "cook-then-mix" scenario of transport in a fair weather boundary layer for several days prior to venting.

Pickering et al. (1992c) used a combination of deep convective cloud cover statistics from the International Satellite Cloud Climatology Project (ISCCP) and convective transport statistics from GCE model simulations of prototype storms to estimate that between 10 and 40% of CO from biomass burning in the Brazilian state of Rondonia is vented from the boundary layer by deep convection. The statistical-dynamical approach was also used by Thompson et al. (1994) to estimate the convective transport component of the boundary layer CO budget for the central United States for the month of June (Fig. 9). Deep convective venting of the boundary layer dominated other components of the CO budget during early summer, providing a net (upward minus downward) flux of 18.1 x 10^8 kg CO/month to the free troposphere. In this respect the central United States acts as a "chimney" for the country.

Regional chemical transport models (CTMs) have been used for applications such as simulations of photochemical ozone production, acid deposition, and fine particulate matter. Walcek et al. (1990) included a parameterization of cloud-scale

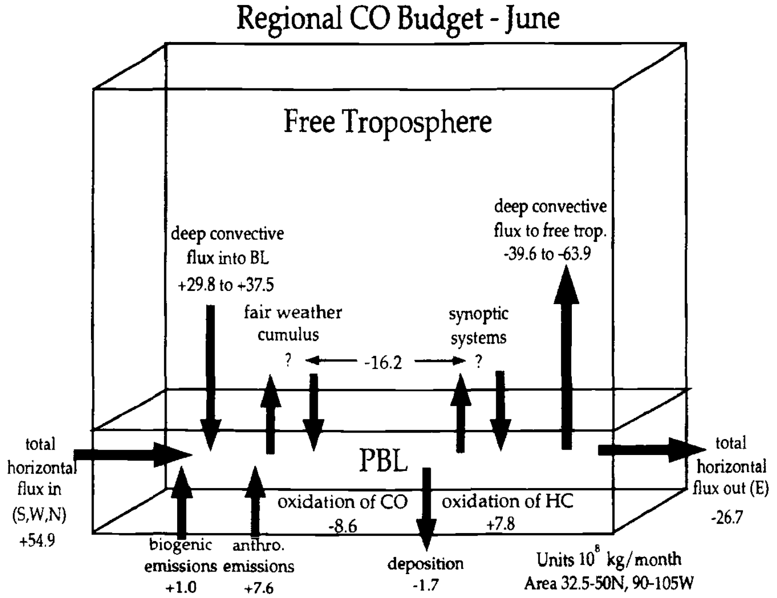


Figure 9 Regional boundary layer CO budget for the central United States (32.5°N to 50°N; 90°N to 105°W). Note magnitudes of upward and downward deep convective transport components. Question marks signify that relative amounts of CO flux due to shallow convection and synoptic-scale systems are unknown. From Thompson et al. (1994).

aqueous chemistry, scavenging, and vertical mixing in the chemistry model of Chang et al. (1987). The vertical distribution of cloud microphysical properties and the amount of subcloud-layer air lifted to each cloud layer are determined using a simple entrainment hypothesis (Walcek and Taylor, 1986). Vertically integrated O_3 formation rates over the northeast United States were enhanced by ~50% when the in-cloud vertical motions were included in the model.

Wang et al. (1996) simulated the September 26–27, 1992, TRACE–A mesoscale convective systems (MCS) and the June 10–11, 1985, PRE-STORM squall line with the NCAR/Penn State Mesoscale Model (MM5; Grell et al., 1994; Dudhia, 1993). Convection is parameterized as a subgrid-scale process in MM5; two convective parameterizations were tested in the Wang et al. (1996) work. These were the Grell (1993) and Kain and Fritsch (1993) schemes. Mass fluxes and detrainment profiles from these schemes were used along with the three-dimensional wind fields in CO tracer transport calculations for the two convective events. The time-evolving tracer fields in the upper troposphere are different in the tropical MCS and the midlatitude squall line. The nearly stationary tropical system produced regions of large upper tropospheric CO that moved very little in the horizontal by the end of the 24-h simulation, whereas enhanced upper tropospheric CO propagates with the relatively fast moving midlatitude squall line. Using a grid size of 25 to 30 km, the parameterized subgrid vertical transport represented 48% (Grell, 1993) and 41% (Kain and Fritsch, 1993) of the total upward transport in the tropical case and 64% (Kain–Fritsch) in the midlatitude case. Pickering et al. (1996) demonstrated that the MM5 convective transport (Fig. 10) reproduced the observed factor of three enhancement of upper tropospheric CO and that over several days downwind transport the enhanced upper tropospheric O_3 precursor mixing ratios allowed O_3 production to proceed at a rate ~4 times faster than would have occurred in undisturbed air. The U.S. Environmental Protection Agency (EPA) has developed a Community Multiscale Air Quality (CMAQ) modeling system that uses MM5 with the Kain–Fritsch convective scheme as the dynamical driver (Ching et al., 1998).

Global

Convective transport in global chemistry and transport models is treated as a subgrid-scale process that is parameterized typically using cloud mass flux information from a general circulation model (GCM) or global data assimilation system. Jacob and Prather (1990) simulated the distribution of radon-222 over North America using a three-dimensional CTM driven with meteorological fields from the NASA Goddard Institute for Space Studies (GISS) GCM II (Hansen et al., 1983), having a horizontal resolution of 4° x 5° and 9 layers in the vertical. Simulation of convective transport in the CTM follows the scheme used in the GCM to transport momentum, sensible heat, and moisture. The model gave a reasonable simulation of radon-222 observations over the United States, but with some significant discrepancies that were traced to problems in the GCM meteorology. Improved simulations of transport have been obtained using a newer convective parameterization of Del Genio and Yao (1988).

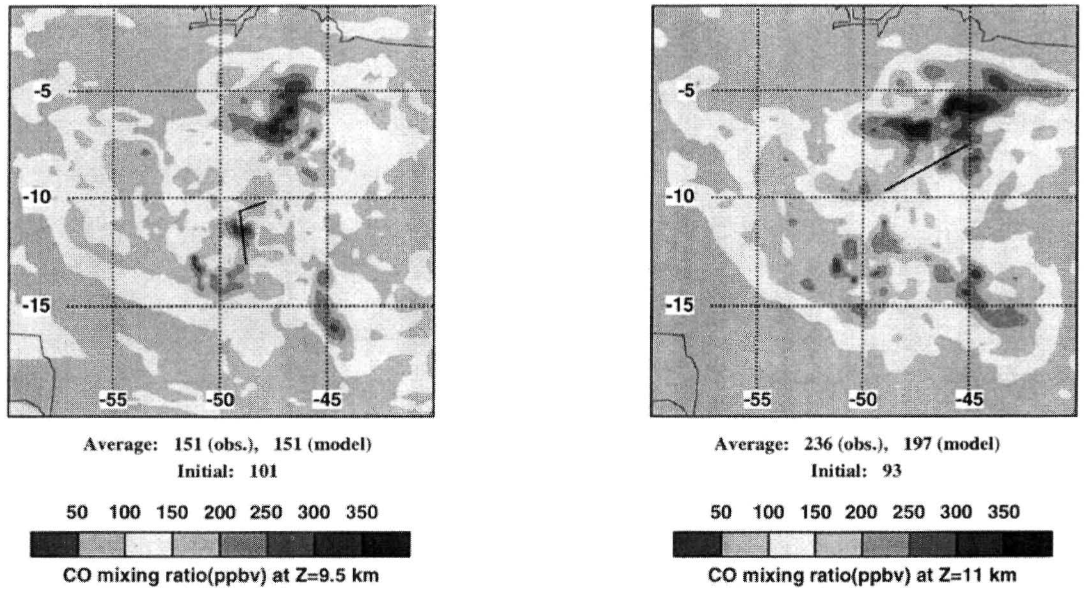

-5
-10
-15
-55
-50
-45
Average: 151 (obs.), 151 (model)
Initial: 101
50 100 150 200 250 300 350
CO mixing ratio(ppbv) at Z=9.5 km
-5
-10
-15
-55
-50
-45
Average: 236 (obs.), 197 (model)
Initial: 93
50 100 150 200 250 300 350
CO mixing ratio(ppbv) at Z=11 km

Figure 10 MM5 simulation result for CO tracer following TRACE-A mesoscale convective events. Shown are CO mixing ratios at 1200 UT September 27, 1992, at altitudes 9.5 and 11 km. Region shown is fine-grid (30-km resolution) domain of MM5 simulation. Includes grid-scale and subgrid transport. From Pickering et al. (1996).

While GCMs can provide data only for a "typical" year, data assimilation systems can provide "real" day-by-day meteorological conditions, such that CTM output can be compared directly with observations of trace gases. The NASA Goddard Earth Observing System Data Assimilation System (GEOS-1 DAS; Schubert et al., 1993) provides archived global data sets for the period 1980–1995, at 2° x 2.5° resolution with 20 layers in the vertical. Convection is parameterized with the relaxed Arakawa–Schubert scheme (Moorthi and Suarez, 1992). Pickering et al. (1995) showed that the cloud mass fluxes from GEOS-1 DAS are reasonable for the June 10–11, 1985, PRE-STORM squall line based on comparisons with the GCE model (cloud-resolving model) simulations of the same storm (Fig. 11). In addition, the GEOS-1 DAS cloud mass fluxes compared favorably with the regional estimates of convective transport for the central United States presented by Thompson et al. (1994). Allen et al. (1996a,b) have used the GEOS-1 DAS data to drive global CTM calculations for radon-222 and for CO. However, Allen et al. (1997) have shown that the GEOS-1 DAS overestimates the amount and frequency of convection in the tropics and underestimates the convective activity over midlatitude marine storm tracks.

Mahowald et al. (1995) investigated the behavior of seven different cumulus parameterization schemes in deriving convective transport from meteorological analysis data sets that did not routinely archive cloud mass fluxes. The derived convective transport was used in a column model and showed that the resulting vertical profile of trace gases was highly sensitive to the parameterization used.

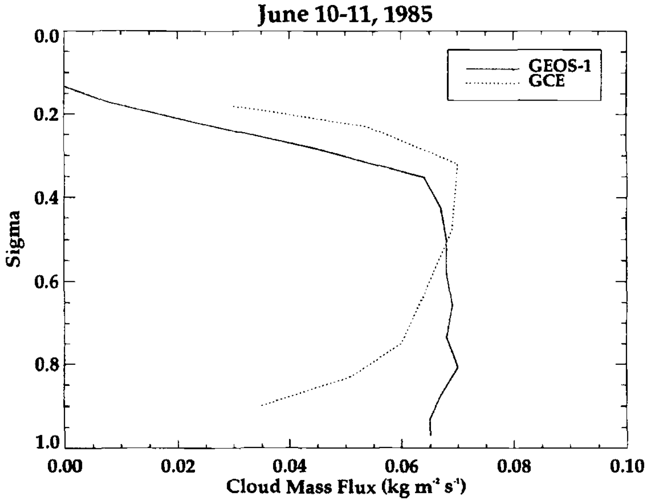


Figure 11 Profiles of cloud mass flux for June 10–11, 1985, PRE-STORM squall line computed by GEOS-1 DAS and by the GCE model. From Pickering et al. (1995).

Rasch et al. (1997) have described use of the output from the NCAR (National Center for Atmospheric Research) Community Climate Model (CCM3) in a chemical transport model. This CTM uses results from the CCM3 convective parameterizations [(Zhang and McFarlane (1995) penetrative convection parameterization and the Hack (1994) scheme for shallow convection)].

4 SUMMARY

Observations and model simulations over the last 15 years have greatly clarified the role of convection in transporting trace constituents in the atmosphere. It is now well established that some nonprecipitating cumulus clouds aid in venting the boundary layer. However, methods to determine the fraction of such clouds that actively transport trace gases to the free troposphere for a region on a given day still require further work. It is also well established that deep convection can transport large quantities of boundary layer gases to the middle and upper troposphere where they have a much longer chemical lifetime and can be transported large distances from their source region. Ozone production in the free troposphere can be enhanced by a factor of 4 or more as a result of deep convection. Downdrafts in convective storms can transport cleaner air from the midtroposphere down to the boundary layer. In remote regions low values of ozone and NO_x can be transported to the upper troposphere, decreasing ozone and ozone production at these altitudes in such regions. In addition, convection induces downward transport of larger O_3 mixing ratios into the remote boundary layer where photochemistry and surface deposition destroy O_3. Storms that reach near or above the preconvective tropopause can induce exchange of trace constituents between stratosphere and troposphere.

Cloud-resolving models are the best tool for detailed studies of convective transport by individual storm systems. Air parcel trajectories and tracer transport calculations using the wind fields from such models are useful for understanding the flow patterns involved in the convective transport process. A cloud model is also useful in evaluating parameterized convective transport in regional or global models. Considerable uncertainty still exists in the output of convective parameterizations concerning the frequency, location, and magnitude of vertical transport, making convective transport one of the largest sources of uncertainty in regional and global CTMs.

REFERENCES

Allen, D. J., P. Kasibhatla, A. M. Thompson, R. B. Rood, B. G. Doddridge, K. E. Pickering, R. D. Hudson, and S.-J. Lin, Transport-induced interannual variability of carbon monoxide determined using a chemistry and transport model, *J. Geophys. Res.*, *101*, 28655–28669, 1996a.

Allen, D. J., R. B. Rood, A. M. Thompson, and R. Hudson, Three dimensional radon 222 calculations using assimilated meteorological data and a convective mixing algorithm, *J. Geophys. Res.*, *101*, 6871–6881, 1996b.

Allen, D. J., K. E. Pickering, and A. Molod, An evaluation of deep convective mixing in the Goddard chemical transport model using ISCCP cloud parameters, *J. Geophys. Res.*, *102*, 25467–25476, 1997.

Chang, J. S., R. A. Brost, I. S. I. Isaksen, S. Madronick, P. Middleton, W. R. Stockwell, C. J. Walcek, A three-dimansional Eulerian acid deposition model: Physical concepts and formulation, *J. Geophys. Res.*, *92*, 14, 681–14, 700, 1987.

Chatfield, R. B., and P. J. Crutzen, Sulfur dioxide in remote oceanic air: Cloud transport of reactive precursors, *J. Geophys. Res.*, *89*, 7111–7132, 1984.

Chatfield R. B., and A. C. Delany, Convection links biomass burning to increased tropical ozone: However, models will tend to overpredict O_3, *J. Geophys. Res.*, *95*, 18473–18488, 1990.

Ching, J. K. S., and A. J. Alkezweeny, Tracer study of vertical exchange by cumulus clouds, *J. Clim. Appl. Meteorol.*, *25*, 1702–1711, 1986.

Ching, J. K. S., S. T. Shipley, and E. V. Browell, Evidence for cloud venting of mixed layer ozone and aerosols, *Atmos. Environ.*, *22*, 225–242, 1988.

Ching, J. K. S., D. W. Byun, J. Young, F. Binkowski, J. Pleim, S. Roselle, J. Godowitch, W. Benjey, and G. Gipson, Science features in Models-3 Community Multiscale Air Quality System, in *Preprints of the Tenth Joint AMS/AWMA Conference on Applications of Air Pollution Meteorology*, Phoenix, AZ, 1998.

Crutzen, P. J., and L. T. Gidel, A two-dimensional photochemical model of the atmosphere, 2. The tropospheric budgets of the anthropogenic chlorocarbons, CO, CH_4, CH_3Cl and the effect of various NO_x sources on tropospheric ozone, *J. Geophys. Res.*, *88*, 6641–6661, 1983.

Danielsen, E. F., In-situ evidence of rapid, vertical, irreversible transport of lower tropospheric air into the lower tropical stratosphere by convective cloud turrets and by large-scale upwelling in tropical cyclones, *J. Geophys. Res.*, *98*, 8665–8681, 1993.

Del Genio, A. D., and M. S. Yao, Sensitivity of a global climate model to the specification of convective updraft and downdraft mass fluxes, *J. Atmos. Sci.*, *45*, 2641–2668, 1988.

Dickerson, R. R., G. J. Huffman, W. T. Luke, L. J. Nunnermacker, K. E. Pickering, A. C. D. Leslie, C. G. Lindsey, W. G. N. Slinn, T. J. Kelly, P. H. Daum, A. C. Delany, J. P. Greenberg, P. R. Zimmerman, J. F. Boatman, J. D. Ray, and D. H. Stedman, Thunderstorms: An important mechanism in the transport of pollutants, *Science*, *235*, 460–465, 1987.

Dudhia, J., A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front, *Monthly Weather Rev.*, *121*, 1493–1513, 1993.

Garstang, M., J. Scala, S. Greco, R. Harriss, S. Beck, E. Browell, G. Sachse, G. Gregory, G. Hill, J. Simpson, W.-K. Tao, and A. Torres, Trace gas exchanges and convective transports over the Amazonian rain forest, *J. Geophys. Res.*, *93*, 1528–1550, 1988.

Gidel, L. T., Cumulus cloud transport of transient tracers, *J. Geophys. Res.*, *88*, 6587–6599, 1983.

Greenhut, G. K., Transport of ozone between boundary layer and cloud layer by cumulus clouds, *J. Geophys. Res.*, *91*, 8613–8622, 1986.

Greenhut, G. K., J. K. S. Ching, R. Pearson, Jr., and T. P. Repoff, Transport of ozone by turbulence and clouds in an urban boundary layer, *J. Geophys. Res.*, *89*, 4757–4766, 1984.

Grell, G. A., Prognostic evaluation of assumptions used by cumulus parameterizations, *Monthly Weather Rev.*, *121*, 764–767, 1993.

Grell, G. A., J. Dudhia, and D. Stauffer, *A Description of the Fifth Generation Penn State/NCAR Mesoscale Model (MM5)*, NCAR/TN-389 + STR, National Center for Atmospheric Research, Boulder, CO, 1994.

Hack, J. J., Parameterization of moist convection in the NCAR Community Climate Model, CCM2, *J. Geophys. Res.*, *99*, 5551–5568, 1994.

Hansen, J., G. Russell, D. Rind, P. Stone, A. Lacis, S. Lebendeff, R. Ruedy, and L. Travis, Efficient three-dimensional global models for climate studies: Models I and II, *Monthly Weather Rev.*, *111*, 609–662, 1983.

Jacob, D. J., and M. J. Prather, Radon-222 as a test of convective transport in a general circulation model, *Tellus*, *42B*, 118–134, 1990.

Kain, J. S., and J. M. Fritsch, Convective parameterization in mesoscale models: The Kain-Fritsch scheme, in K. A. Emanuel and D. J. Raymond (Eds.), *The Representation of Cumulus Convection in Numerical Models*, American Meteorological Society, Boston, MA, 1993.

Kawakami, S., Y. Kondo, M. Koike, H. Nakajima, G. L. Gregory, G. W. Sachse, R. E. Newell, E. V. Browell, D. R. Blake, J. M. Rodriguez, and J. T. Merrill, Impact of lightning and convection on reactive nitrogen in the tropical free troposphere, *J. Geophys. Res.*, *102*, 28367–28384, 1997.

Kleinman, L. I., and P. H. Daum, Vertical distribution of aerosol particles, water vapor, and insoluble trace gases in convectively mixed air, *J. Geophys. Res.*, *96*, 991–1005, 1991.

Kley, D., P. J. Crutzen, H. G. J. Smit, H. Vomel, S. J. Oltmans, H. Grassl, and V. Ramanathan, Observations of near-zero ozone concentrations over the convective Pacific: Effects on air chemistry, *Science*, *274*, 230–233, 1996.

Lelieveld, J., and P. J. Crutzen, Role of deep cloud convection in the ozone budget of the troposphere, *Science*, *264*, 1759–1761, 1994.

Luke, W. T., R. R. Dickerson, W. F. Ryan, K. E. Pickering, and L. J. Nunnermacker, Tropospheric chemistry over the lower Great Plains of the United States 2. Trace gas profiles and distributions, *J. Geophys. Res.*, *97*, 20647–20670, 1992.

Mahowald, N. M., P. J. Rasch, and R. G. Prinn, Cumulus parameterizations in chemical transport models, *J. Geophys. Res.*, *100*, 26173–26190, 1995.

Moorthi, S., and M. J. Suarez, Relaxed Arakawa-Schubert: A parameterization of moist convection for general circulation models, *Monthly Weather Rev.*, *120*, 978–1002, 1992.

Newell, R. E., et al., Atmospheric sampling of Supertyphoon Mireille with NASA DC-8 aircraft on September 27, 1991, during PEM-West A, *J. Geophys. Res.*, *101*, 1853–1871, 1996.

Pickering, K. E., R. R. Dickerson, G. J. Huffman, J. F. Boatman, and A. Schanot, Trace gas transport in the vicinity of frontal convective clouds, *J. Geophys. Res.*, *93*, 759–773, 1988.

Pickering, K. E., R. R. Dickerson, W. T. Luke, and L. J. Nunnermacker, Clear-sky vertical profiles of trace gases as influenced by upstream convective activity, *J. Geophys. Res.*, *94*, 14879–14892, 1989.

Pickering, K. E., A. M. Thompson, R. R. Dickerson, W. T. Luke, and D. P. McNamara, Model calculations of tropospheric ozone production potential following observed convective events, *J. Geophys. Res.*, *95*, 14049–14062, 1990.

Pickering, K. E., A. M. Thompson, J. R. Scala, W.-K. Tao, J. Simpson, and M. Garstang, Photochemical ozone production in tropical squall line convection during NASA Global

Tropospheric Experiment/Amazon Boundary Layer Experiment 2A, *J. Geophys. Res.*, *96*, 3099–3114, 1991.

Pickering, K. E., A. M. Thompson, J. Scala, W.-K. Tao, R. R. Dickerson, and J. Simpson, Free tropospheric ozone production following entrainment of urban plumes into deep convection, *J. Geophys Res.*, *97*, 17985–18000, 1992a.

Pickering, K. E., A. M. Thompson, J. R. Scala, W.-K. Tao, and J. Simpson, Ozone production potential following convective redistribution of biomass emissions, *J. Atmos. Chem.*, *14*, 297–313, 1992b.

Pickering, K. E., A. M. Thompson, W.-K. Tao, and T. L. Kucsera, Upper tropospheric ozone production following mesoscale convection during STEP/EMEX, *J. Geophys. Res.*, *98*, 8737–8749, 1993.

Pickering, K. E., A. M. Thompson, W.-K. Tao, R. B. Rood, D. P. McNamara, and A. M. Molod, Vertical transport by convective clouds: Comparisons of three modeling approaches, *Geophys. Res. Lett.*, *22*, 1089–1092, 1995.

Pickering, K. E., J. R. Scala, A. M. Thompson, W.-K. Tao, and J. Simpson, A regional estimate of the convective transport of CO from biomass burning, *Geophys. Res. Lett.*, *19*, 289–292, 1992c.

Pickering, K. E., A. M. Thompson, Y. Wang, W.-K. Tao, D. P. McNamara, V. W. J. H. Kirchhoff, B. G. Heikes, G. W. Sachse, J. D. Bradshaw, G. L. Gregory, and D. R. Blake, Convective transport of biomass burning emissions over Brazil during TRACE-A, *J. Geophys. Res.*, *101*, 23993–24012, 1996.

Piotrowicz, S. R., H. F. Bezdek, G. R. Harvey, M. Springer-Young, and K. J. Hanson, On the ozone minimum over the equatorial Pacific Ocean, *J. Geophys. Res.*, *96*, 18679–18687, 1991.

Poulida, O., R. R. Dickerson, and A. Heymsfield, Troposphere-stratosphere exchange in a midlatitude mesoscale convective complex: 1. Observations, *J. Geophys. Res.*, *101*, 6823–6836, 1996.

Rasch, P. J., N. M. Mahowald, and B. E. Eaton, Representations of transport, convection, and the hydrologic cycle in chemical transport models: Implications for the modeling of short-lived and soluble species, *J. Geophys. Res.*, *102*, 28127–28152, 1997.

Riehl, H., and J. Simpson, The heat balance of the equatorial zone, revisited, *Beitr. Phys. Atmos.*, *52*, 287–305, 1979.

Scala, J., M. Garstang, W.-K. Tao, K. Pickering, A. Thompson, J. Simpson, V. Kirchhoff, E. Browell, G. Sachse, A. Torres, G. Gregory, R. Rasmussen, and M. Khalil, Cloud draft structure and trace gas transport, *J. Geophys. Res.*, *95*, 17017–17030, 1990.

Schubert, S. D., R. B. Rood, and J. Pfaendtner, An assimilated data set for earth science applications, *Bull. Am. Meteoral. Soc.*, *74*, 2331–2342, 1993.

Stenchikov, G., R. Dickerson, K. Pickering, W. Ellis, B. Doddridge, S. Kondragunta, and O. Poulida, Stratosphere-troposphere exchange in a mid-latitude mesoscale convective complex: Part 2, Numerical simulations, *J. Geophys. Res.*, *101*, 6837–6851, 1996.

Stull, R. B., A fair-weather cumulus cloud classification scheme for mixed layer studies, *J. Clim. Appl. Meteorol.*, *24*, 49–56, 1985.

Suhre, K., J.-P. Cammas, P. Nedelec, R. Rosset, A. Marenco, and H. G. J. Smit, Ozone-rich transients in the upper equatorial Atlantic troposphere, *Nature*, *388*, 661–663, 1997.

Tao, W.-K., and J. Simpson, The Goddard cumulus ensemble model. Part I: Model description, *Terrestr. Atmos. Oceanic Sci.*, *4*, 35–72, 1993.

Thompson, A. M., K. E. Pickering, R. R. Dickerson, W. G. Ellis, Jr., D. J. Jacob, J. R. Scala, W.-K. Tao, D. P. McNamara, and J. Simpson, Convective transport over the Central United States and its role in the regional CO and O_3 budgets, *J. Geophys. Res.*, *99*, 18, 703–18, 711, 1994.

Thompson, A. M., W.-K. Tao, K. E. Pickering, J. R. Scala, and J. Simpson, Tropical deep convection and ozone formation, *Bull. Am. Meteorol. Soc.*, *78*, 1043–1054, 1997.

Vukovich, F. M., and J. K. S. Ching, A semi-impirical approach to estimate vertical transport by nonprecipitating convective clouds on a regional scale, *Atmos. Environ.*, *24A*, 2153–2168, 1990.

Walcek, C. J., W. R. Stockwell, and J. S. Chang, Theoretical estimates of the dynamic, radiative, and chemical effects of clouds on tropospheric trace gases, *Atmos. Res.*, *25*, 53–69, 1990.

Walcek, C. J., and G. R. Taylor, A theoretical method for computing vertical distribution of acidity and sulfate production within cumulus clouds, *J. Atmos. Sci.*, *43*, 339–355, 1986.

Wang, C., and J. S. Chang, A three-dimensional numerical model of cloud dynamics, microphysics, and chemistry 1, Concepts and formulation, *J. Geophys. Res.*, *98*, 14827–14844, 1993.

Wang, C., P. J. Crutzen, V. Ramanathan, and S. F. Williams, The role of a deep convective storm over the tropical Pacific Ocean in the redistribution of atmospheric chemical species, *J. Geophys. Res.*, *100*, 11509–11516, 1995.

Wang, Y., W.-K. Tao, K. E. Pickering, A. M. Thompson, J. S. Kain, R. F. Adler, J. Simpson, P. R. Keehn, and G. S. Lai, Mesoscale model simulations of TRACE-A and PRE-STORM convective systems and associated tracer transport, *J. Geophys. Res.*, *101*, 24013–24027, 1996.

Zhang, G. J., and N. A. McFarlane, Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model, *Atmos. Ocean*, *33*, 407–446, 1995.

CHAPTER 10

BOUNDARY LAYER PROCESSES AND FLUX MEASUREMENTS

DONALD H. LENSCHOW

1 INTRODUCTION

The planetary boundary layer (PBL) is that part of the atmosphere that interacts with Earth's surface on a time scale of about an hour. This rapid interaction is a direct result of turbulence, which is an essential feature of the PBL. The sources of turbulence are wind shear and convection. Defining characteristics of turbulence are its chaotic fluctuations and diffusiveness. That is, trace constituents released into a turbulent fluid are rapidly diffused, and the small-scale patterns of this diffusion cannot be predicted. Because of the randomness and the large range of scales of PBL turbulence, processes in the PBL are often described in terms of statistical averages of fluctuations. This means that most measurements of PBL structure need to be spatially or temporally averaged before they can be quantitatively interpreted.

If generation of turbulence by convection is occurring in the PBL, it is known as an *unstable* or *convective* boundary layer (CBL); if the hydrodynamic stratification of the PBL acts to suppress or dissipate turbulence, it is known as a *stable* boundary layer (SBL). When relative humidity reaches 100% within the PBL, clouds form that can have a dramatic effect on its subsequent evolution.

2 BOUNDARY LAYER EVOLUTION

Over land, the daily solar cycle determines the PBL evolution. In the morning, the sun starts to warm the ground, which has been cooling through the night by infrared radiation. Clear air is nearly transparent to the sun's short-wave (visible) radiation

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

and thus is warmed only slightly by direct solar radiation. Instead, the ground absorbs most of the solar radiation and then warms the air above it mostly through convection, which is the upward movement of buoyant parcels of air warmed by contact with the surface, in combination with compensating downward transfer of cooler, more dense air from above. This process generates turbulence that, in turn, increases the efficiency of transport of atmospheric constituents in the CBL. Efficient mixing means that the *lapse rate*, which is the rate of change of temperature with height, is nearly adiabatic throughout much of the CBL. This means that vertical displacements of an air parcel do not change the buoyancy of the parcel relative to its environment. Figure 1 shows the structure of the CBL using *virtual potential temperature*, which is constant with height in an adiabatic layer, and a scalar variable with a surface source and negligible concentration above the CBL.

The CBL continues to deepen typically at least until early afternoon to perhaps 1 to 3 km. Generally, the *relative humidity* in the upper part of the CBL tends to

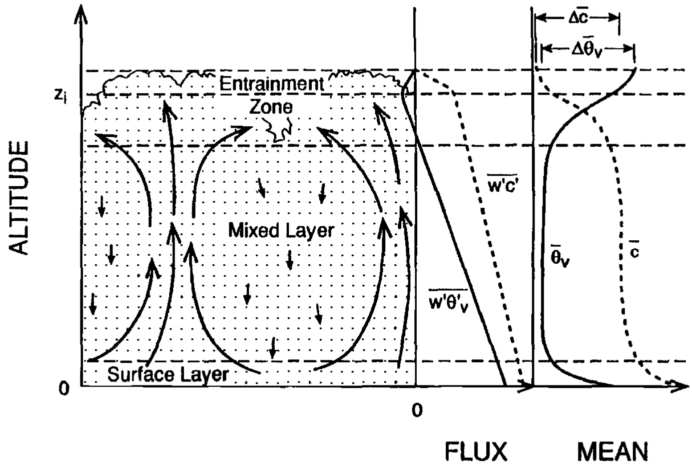


Figure 1 Convective boundary layer. On the left, the sublayers that make up the boundary layer are shown, along with a schematic of flow patterns—upward-moving buoyant thermals forming near the top of the surface layer, extending through the mixed layer, and dissipating in the upper part of the boundary layer. Part of their kinetic energy is dissipated in the entrainment zone by entraining warmer, more buoyant air from above into the boundary layer. The thermals have a smaller total area, and consequently a larger velocity magnitude than the compensating downward-moving air in between the thermals. The flux profiles in the middle panel show the virtual potential temperature (θ_v) flux (which has a universal shape), and the flux of a scalar $\mathscr{C}$ which (in this case) has a source at the surface and whose mean concentration decreases with height throughout and immediately above the boundary layer. The virtual potential temperature can normally be assumed to be a conserved variable in a well-mixed clear boundary layer. The mean virtual potential temperature and scalar concentration profiles, including their jumps across the top of the boundary layer, are shown on the right.

increase through the day from moistening due to surface *evapotranspiration* and turbulent mixing. If the humidity reaches saturation, clouds develop at the top of the CBL.

As the solar heating decreases late in the day, convection disappears and radiative cooling at the surface again dominates over solar warming. Eventually, the ground becomes cooler than the overlying air, which radiatively cools much more slowly than the ground. At this point, the SBL develops, which is considerably shallower (ranging from a few tens to a few hundred meters deep) and less turbulent because buoyancy is now suppressing the turbulence generated by shear. As a result, turbulent transport is less efficient. Above the surface layer, turbulence becomes intermittent and, because of the stable stratification, gravity waves may become important. The top of the SBL is not as well defined as the daytime CBL since the turbulence decreases much more slowly with height. Because of the low turbulence, discrete layers with their own characteristic properties may form and advect at different speeds. After sunrise, the ground begins to warm, and the cycle repeats.

Over the ocean, the large heat capacity and effective heat conductivity keep the ocean temperature nearly constant over the daily cycle; thus the daily cycle is often insignificant in the marine boundary layer (MBL). This means that the MBL typically has much less production of turbulence energy by buoyancy and is usually shallower (perhaps 0.5 to 1.5 km) than the daytime CBL over land. On average, there is less cloudiness at the top of the MBL during the day than at night because absorption of solar short-wave radiation warms the MBL directly and thus reduces the production of turbulence by buoyancy and the degree of mixing.

3 STRUCTURE OF THE BOUNDARY LAYER

Boundary layer processes are dominated by the diffusive character of turbulence. This diffusion can be described as a flux of a constituent, which is the rate of transport of the constituent across a surface per unit time and per unit area. Alternatively, it can be expressed as a constituent density times a velocity. Both velocity components and scalars are defined as sums of a mean and a fluctuation, $\mathcal{U}_i = U_i + u_i$ and $\mathcal{S} = S + s$. The horizontal wind components are commonly defined as $\mathcal{U}_1 \equiv \mathcal{U}$ and $\mathcal{U}_2 \equiv \mathcal{V}$, while the vertical component is $\mathcal{U}_3 \equiv \mathcal{E}$). The average of the fluctuations, $\bar{u}_i = \bar{s} = 0$, where the overbar is the common convention to denote an average of a turbulence variable over a time period or a length long enough to give a stable estimate of its mean.

In a turbulent fluid, the scalar flux is defined as the sum or integral of fluctuations in the velocity normal to a surface times the concurrent fluctuations in constituent density divided by the time period or length over which the sum or integral is calculated. Normally in the PBL, only the vertical component of the flux is of interest. Thus,

$$F_s = \frac{1}{T}\int_0^T ws\,dt = \overline{ws} \tag{1}$$

where w and s are fluctuations in the vertical wind component and a scalar density, respectively, and T is the averaging time period. Equation (1) defines the *eddy correlation* technique for measuring flux.

The PBL can be further broken down into sublayers: The lowest few tens of meters is called the *surface layer.* In this region, which is less than 10% of the depth of the PBL, the fluxes can be considered constant with height, but for variables with large surface fluxes, the mean vertical gradients of these variables are large relative to the rest of the PBL. It is common to relate the flux in the surface layer to a gradient by a diffusivity,

$$F_s = -K_s \frac{\partial S}{\partial z} \tag{2}$$

Since transport by turbulent eddies in the surface layer is roughly about 10^5 times more efficient than transport by molecular diffusion, for scalars we call K_s the eddy diffusivity and for momentum the eddy viscosity. Near the surface, the typical maximum horizontal dimension of the eddies making important contributions to the flux is roughly about 200 times the height above the surface. Since the efficiency of turbulent transport scales with the size of the eddies, the eddy diffusivity increases approximately linearly with height. Equivalently, since the flux is approximately constant in the surface layer, the gradient decreases approximately inversely with height near the surface.

In the CBL, the layer above the surface layer and below the region near the PBL top is called the *mixed layer.* Since this encompasses the bulk of the PBL, the fluxes show considerable variability. Here the gradients are small because the mixing process is efficient. The individual turbulent eddies extend throughout the mixed layer and are called *thermals* (or *plumes*). The typical maximum size of eddies making important contributions to the flux changes more slowly with height than in the surface layer, and is roughly about 40 times the depth of the CBL. The turbulence energy (the sum of the three component velocity variances) reaches a maximum at about one third the height of the mixed layer.

Above the mixed layer is the *entrainment layer.* In this region, a sharp interface typically occurs between the CBL and the overlying nonturbulent *free atmosphere.* This interface is generated by turbulent eddies that protrude into the nonturbulent layer, engulf or capture volumes of nonturbulent air, and fall back into the mixed layer. Smaller-scale turbulence within these larger eddies then commingles this entrained air with CBL air so that it becomes part of the mixed layer. Details of the clear (cloud-free) CBL structure are shown in Figure 1.

If clouds form at the top of the CBL, one of two scenarios occurs: If the density decrease across the top of the CBL is small enough that condensation and mixing processes decrease the density to a value below the overlying air, the cloud can penetrate through the top of the CBL and form a *cumulus.* In that case, the CBL top is approximately at cloud base, and the clouds penetrate into the overlying air until mixing with their environment limits their growth and they lose their buoyancy. This *venting* process injects CBL air into the overlying atmosphere, which is a way to increase humidity above the CBL and introduce trace constituents originating at the surface or within the CBL into the free atmosphere. Compensating downward motion can dilute the concentration of pollutants in the CBL. If the decrease in density across the top of the CBL is large enough that phase changes and mixing of overlying air with CBL air do not decrease the density sufficiently to lower the density below the overlying air, the cloud layer is contained within the CBL and a *stratus* or *stratocumulus* cloud layer may exist. In this case, the CBL maintains a sharp interface. Radiative cooling at cloud top can also generate CBL turbulence and contribute to the efficiency of mixing in the cloud-capped CBL.

In the SBL, the layer above the surface layer is a region of decreasing intensity and increasing intermittency of turbulence. Here the individual turbulent eddies may not extend throughout the SBL, but occur in sublayers that develop locally enhanced shear that may intermittently break down into turbulence. The turbulence is then dissipated with the net result that the gradients in the sublayer are reduced by the transient turbulence event. The process may very well be repeated over time. Since the turbulence is more local in nature, the scales and structure are less well defined than in the CBL, and velocity and scalar variances decrease with height throughout the SBL. Relatively large gradients of both wind and scalars, and multiple sublayers that are only intermittently coupled may exist. The top of the SBL generally does not have a well-defined lid. Because of the intermittency and smaller length scales of the mixing process, the shallower and less well defined structure of the SBL, and the presence of gravity waves that produce velocity fluctuations but no flux, trace constituent fluxes are much more problematic and much less frequently measured in the SBL than the CBL. For these reasons, most of the subsequent discussion deals solely with the CBL.

4 SCALES AND PROCESSES

Wind shear is the rate of change of wind with height,

$$\frac{\partial U}{\partial z}+\frac{\partial V}{\partial z} \tag{3}$$

where U and V are the averaged horizontal wind components. Because of drag induced by Earth's surface, the horizontal wind approaches zero at the surface. Since the eddy viscosity increases approximately linearly with height and the kinematic momentum flux ($\overline{uw}$) is approximately constant [and equal to $(\overline{uw})_0$] in the surface layer, the wind shear decreases roughly inversely with height very near the surface. Further above the surface, the wind shear, as well as scalar gradients in the surface layer depend also on the stability—that is, if the surface buoyancy flux is positive, the magnitudes of the wind shear and scalar gradients decrease with height less rapidly than the inverse of height, and if the buoyancy flux is negative, they decrease more rapidly than the inverse of height. Furthermore, the direction of the mean wind is assumed to be constant with height in the surface layer, so the coordinate system can be defined such that $V = 0$. Thus near the surface, or in a surface layer with zero surface buoyancy flux (a neutrally stratified PBL),

$$\frac{\partial U}{\partial z} = \frac{u_*}{kz} \tag{4}$$

where $u_*^2 = -(\overline{uw})_0$ is the friction velocity and k is the von Kármán constant ($\simeq 0.4$). A typical range of values for u_* over a treeless vegetated surface in moderate winds would be 0.2 to 0.8 m/s. A similar relation holds for scalar quantities in the surface layer,

$$\frac{\partial S}{\partial z} = \frac{S_*}{kz} \tag{5}$$

where $S_* = -F_{s0}/u_*$ and F_{s0} is the surface-layer flux of $\mathscr{S}$.

These equations can be integrated to obtain vertical profiles,

$$U(z_2) - U(z_1) = \frac{u_*}{k} \ln \frac{z_2}{z_1} \tag{6}$$

and

$$S(z_2) - S(z_1) = \frac{S_*}{k} \ln \frac{z_2}{z_1} \tag{7}$$

Since $U \to 0$ at the surface, we define the *roughness length* z_0 as the height at which the extrapolated wind profile goes to zero so that

$$U(z) = \frac{u_*}{k} \ln \frac{z}{z_0} \tag{8}$$

The roughness length is approximately $\frac{1}{30}$ the height of individual surface roughness elements. It ranges from about 10^{-4} m over calm water to about 0.5 m over a forest.

The production of turbulence energy by wind shear is given by the product of mean wind shear and (kinematic) momentum flux,

$$E_u = -\overline{uw}\left(\frac{\partial U}{\partial z} + \frac{\partial V}{\partial z}\right) \tag{9}$$

Near the surface, or in a neutral PBL, inserting (4) into (9) reduces (9) to

$$E_u = \frac{u_*^3}{kz} \tag{10}$$

Production (dissipation) of turbulence by convection can be expressed in terms of a buoyancy flux,

$$F_b = \frac{g}{T}\overline{wT_v} \tag{11}$$

where g is gravity, T is temperature, and T_v is virtual temperature, which includes the density effects of water vapor on the temperature. The parameter g/T is the buoyancy parameter, which is the thermal expansion coefficient of air times the acceleration of gravity. Since the buoyancy flux can also be negative, this term may also act to dissipate turbulence. The total turbulence energy production is the sum of (11) and (9). Averaged over the entire PBL, the energy production must be equal to the turbulence dissipation, which is the loss of turbulence energy due to the viscous forces that occur predominantly at very small scales (i.e., less than a few centimeters). In effect, the viscous forces convert the kinetic energy of turbulence into thermal energy, and thus heat the air (although the temperature increase is insignificant).

Near the surface, the negative ratio of energy production by buoyancy to production by shear is given by

$$-\frac{F_{b0}}{u_*^3/kz} \tag{12}$$

The length at which this ratio is unity, called the *Obukhov length*, is given by

$$L = -\frac{u_*^3}{kF_{b0}} \tag{13}$$

This is a measure of the stability of the surface layer and is used as a scaling height to normalize the observation height. Similarly, velocity, temperature, and scalar variables in the surface layer can be normalized by u_*, $T_* = -(\overline{wT})_0/u_*$, and S_*. In this way, normalized surface layer variables as functions of height can be expressed as universal functions in both the unstably and stably stratified surface layer. This is a powerful technique for relating surface layer measurements to a universal surface layer structure in diabatic (nonzero surface buoyancy flux) PBLs with enough mean wind to generate a well-defined u_*. For example, (4) and (5) can be extended to the diabatic surface layer by including stability functions in the formulations,

$$\frac{\partial U}{\partial z} = \frac{u_*}{kz}\phi_m\left(\frac{z}{L}\right) \tag{14}$$

and

$$\frac{\partial S}{\partial z} = \frac{S_*}{kz}\phi_h\left(\frac{z}{L}\right) \tag{15}$$

where the stability functions ϕ_m and ϕ_h have been obtained empirically from carefully designed field studies. For a neutral PBL, $\phi = 1$; for an unstable PBL, $\phi < 1$; and for a stable PBL, $\phi > 1$. These expressions can be integrated as in (6), (7), and

(8) to relate the fluxes to measurements of velocity and scalar differences at two heights in the surface layer.

A similar procedure is used in the mixed layer, with the scaling height being the depth of the CBL, z_i and the velocity scale being the *Deardorff velocity*,

$$w_* = (F_{b0} z_i)^{1/3} \tag{16}$$

However, in the mixed layer a further complication is that the behavior of mixed-layer variables depends not only on surface fluxes but also on fluxes through the top of the CBL, the *entrainment fluxes*. Therefore, for scalar fluxes in the mixed layer both the surface flux F_{s0} and the entrainment flux F_{szi} need to be incorporated in generalized formulations. For the scalar flux-gradient relationship, this can be expressed as

$$\frac{\partial S}{\partial z} = -\frac{F_{s0}}{w_* z_i} g_0\left(\frac{z}{z_i}\right) - \frac{F_{szi}}{w_* z_i} g_{zi}\left(\frac{z}{z_i}\right) \tag{17}$$

where $g_0(z/z_i)$ and $g_{zi}(z/z_i)$ are the normalized mixed-layer gradient functions. Thus far, these gradient functions have not been measured in the atmosphere; however, they have been estimated from detailed numerical simulations of the CBL.

Usually it is the density of the trace constituent that is measured since most sensors respond to the number of molecules in a particular volume of air. In estimating the flux of a species, we normally calculate the quantity $\overline{ws}$ with the assumption that $W = 0$ at the surface. This is not strictly true even over a horizontally homogeneous surface if the water vapor and temperature fluxes are not zero. This arises from the constraint that the flux that is most realistically zero at the surface is the mass flux of dry air, $\overline{\rho w} = 0$. Intuitively, we can see that in the case of a heated surface, rising parcels of air will be on average warmer and lighter, and consequently contain fewer molecules per unit volume than their surroundings, while descending parcels will be colder and denser, and contain more molecules than their surroundings so that for zero species flux at the surface, $\overline{ws} < 0$. This is known as the *Webb effect*. To obtain the correct flux, it is necessary to correct for $W \neq 0$ by incorporating terms proportional to the fluxes of humidity and temperature. This correction becomes significant if $\overline{ws}/S$ is less than about 0.01 m/s. Alternatively, if instead of measuring the constituent density, we measure its mixing ratio with respect to dry air, there is no Webb correction. In subsequent discussion we disregard this correction, but note that it can be important for surface fluxes of relatively long-lived atmospheric species such as CO_2, CH_4, or N_2O.

5 OBSERVATIONAL TECHNIQUES

Since the boundary layer is a conduit for transport of trace species between the surface and the overlying free troposphere, measuring species fluxes within the PBL is a standard approach for estimating their sources or sinks at the surface, as well as their rates of exchange with the overlying atmosphere. There are many ways to measure fluxes in the PBL. However, the two most widely used platforms are: (1) tower measurements in the surface layer and (2) airplane measurements in the mixed layer. There are, of course, other platforms that are used. For example, in the marine surface layer, ship-mounted instruments are used and in the mixed layer tethered balloons and neutrally buoyant airships have been used. The most direct and fundamental flux measurement technique is the eddy correlation technique [Eq. (1)]. However, this requires fast-response high-resolution measurements of species concentration and vertical air velocity over a time period or distance long enough to obtain a sufficiently accurate average of the turbulent fluctuations. As a rule of thumb, to estimate the flux to about 10% accuracy, one should average over several times the maximum eddy size making significant contributions to the flux. Generally, measuring fluxes from a tower a few meters above the surface for moderate winds requires an averaging time of about 20 min, which is about the same as that required for measuring fluxes from an aircraft in the middle of the mixed layer flying at 100 m/s.

At the other end of the spectrum, the smallest scale eddies that need to be measured to estimate a flux in the surface layer are roughly about $0.5z$. In the mixed layer, the smallest scales are roughly about $0.1z_i$. Therefore, for measurements from a tower at a height of 2 m, and a wind speed of 5 m/s, a frequency response of least 5 Hz is required. In the mixed layer, for an aircraft flying at 100 m/s in the middle of a 1-km-deep CBL, a frequency response of at least 1 Hz is required. If the aircraft is flying lower, say about 30 m, which is in the upper part of the surface layer, a frequency response of at least 7 Hz is required. [To achieve a frequency response of f_c hertz using a sensor with a first-order time response, a sensor time constant of about $1/(6f_c)$ s is required.]

In carrying out flux measurements by eddy correlation, both vertical velocity and species concentrations must be measured concurrently. *Sonic anemometers* are often used for the vertical velocity measurement from towers since they have good velocity resolution, adequate time response, and no moving parts. The air velocity component along the path between two sets of sonic transducers is obtained from the difference in the velocity of sound traveling along the same path in opposite directions. In addition, since the speed of sound is approximately proportional to the square root of virtual temperature, sonic anemometers are usually configured to also measure the virtual temperature, and thus the buoyancy flux can be obtained as well. Three-axis sonic anemometers are available commercially for measuring eddy correlation fluxes in the surface layer.

Measurement of air velocity components from aircraft requires measuring both the velocity of the air with respect to the aircraft and the velocity and angular orientation of the aircraft with respect to Earth. The former is often obtained from pressure measurements on the nose of the aircraft or from a probe mounted on a noseboom ahead of the aircraft. Pressure difference measurements are sensed from sets of ports. A forward-looking and a static pressure port are used to sense the airspeed, and sets of ports at different angles in both the horizontal and vertical plane of the aircraft are used to sense the flow angles of the air. The aircraft orientation and velocity are often obtained from an inertial navigation system (INS) which senses the attitude angles and acceleration of the aircraft. The acceleration components are then integrated to obtain the velocity, and integrated again to obtain the position of the airplane. Often, navigational information from the satellite-based Global Positioning System (GPS) is used to remove drift inherent in the INS due to integration of a bias in the accelerometers. The air velocity is obtained from the difference between the velocity of the air with respect to the airplane, which is rotated by means of the attitude angles to an Earth-based coordinate system, and the velocity of the airplane, which is also measured in an Earth-based coordinate system.

Several techniques have been used to measure species concentration with sufficient resolution and frequency response that direct eddy correlation fluxes can be obtained. Water vapor fluxes have been obtained from both infrared and ultraviolet absorption devices. Fluxes of several other trace gases can also be sensed by infrared absorption, including CO_2, CH_4, and CO. Chemiluminescence is another inherently fast technique useful for ozone, isoprene, and possibly NO, NO_2, and dimethyl sulfide. In this technique, a reactive gas is mixed with the air, which reacts with the species being measured, with the resulting emission of photons detected by a photomultiplier tube. Finally, a tandem mass spectrometer, which ionizes, accelerates, and segregates the target species molecules has been used for measuring fluxes of acetone, ammonia, and formic acid in the surface layer.

Nearly all the techniques listed above (except for some open-path radiation absorption devices) require the air to be drawn into a sensing chamber of some sort. This requires careful consideration of the ducting system to ensure that the flow is fast enough and the ducting short enough that significant attenuation of concentration fluctuations does not occur in the frequency region with significant contributions to the flux. Generally this means that if the duct is longer than a couple of meters, the Reynolds number of the flow in the duct,

$$\mathrm{Re} = \frac{dU_t}{\nu} \tag{18}$$

where d is the tube diameter, U_t the flow velocity in the tube, and ν is the kinematic molecular viscosity ($\simeq 0.15 \times 10^{-4}\,\mathrm{m}^2/\mathrm{s}$ for air at room temperature), must be greater than the critical value for turbulence to exist in the tube; i.e., $\mathrm{Re} > 2300$.

In addition to direct eddy correlation, several other techniques have been used for flux measurement. Most of these alternatives are implemented to relax the high-frequency requirements of direct eddy correlation. Conceptually, perhaps the simplest approach is to make measurements of species concentration less frequently, but grab the sample quickly so as to still retain the required frequency response. By this disjunct sampling technique, a flux can be estimated even if the frequency response of the concentration measurement is reduced by nearly an order of magnitude below what is required for direct eddy correlation. Another approach, called *eddy accumulation*, is to collect the air sample at a rate proportional to the vertical velocity, with the upward-moving air going into one reservoir and downward-moving air into another. The flux is then proportional to the difference in concentration between the two reservoirs. With this approach, there is no longer any requirement for fast-response species measurement. In effect, the requirement for fast response is shifted to the flow control. Disadvantages of this approach are the small concentration difference between the two reservoirs and the requirement for fast-response and accurate flow control.

There are many other techniques for estimating flux, mostly with the objective of reducing the high-frequency response requirement, but, in contrast to the above, these approaches utilize some empirical relationship between the flux and some other variables. One simplification of eddy accumulation, called *relaxed eddy accumulation*, is to collect the air at a constant rate, regardless of the magnitude of the vertical velocity, in either of the two reservoirs depending on the sign of the vertical velocity. The flux then depends, in addition to the concentration difference between the two reservoirs, on the standard deviation of the vertical velocity and a parameter that depends on the vertical velocity distribution.

Measuring the gradient of species concentration either in the surface layer or the mixed layer is also used to estimate the surface flux. In the surface layer, the flux can be estimated from the integral of Eq. (15); i.e., from a difference in concentration between two levels plus the friction velocity, u_*, and a measure of the stability L, which depends on u_* and F_{b0}. Again, this does not require fast-response concentration measurements, but it does require measurement of small differences in concentration, as well as estimates of buoyancy and momentum fluxes.

In the mixed layer, Eq. (17) can similarly be integrated and solved for both the surface and the entrainment fluxes from mean concentration differences. However, since there are now two unknowns, mean concentration must be measured at a minimum of three levels to obtain two concentration differences unless one of the fluxes is estimated by another technique. Typical values of the normalized gradient functions have been estimated from large-eddy numerical simulations of the CBL to be, for $g_0(z/z_i)$ about 13 at $z/z_i = 0.1$ and about 1 at $z/z_i = 0.5$, and for $g_{zi}(z/z_i)$ about 70 at $z/z_i = 0.9$ and about 3 at $z/z_i = 0.5$. Since a typical value for w_* is about 1 m/s, we see that by taking the ratio of (17) to (5) the mixed-layer gradient is roughly about 1% of the surface layer gradient. This is again a reflection of the relative efficiency of transport in the mixed layer compared to the surface layer. This relatively small mixed-layer gradient is offset to a considerable extent by the much larger height differences that can be used in the mixed layer. Nevertheless, the concentration differences obtained by integration of the surface layer gradient formulation (5) can be several times larger than the differences obtained from integration of the mixed-layer gradient formulation (17). Thus far, the mixed-layer gradient technique has been used to estimate surface fluxes of isoprene and dimethyl sulfide, both of which have sources only at the surface and lifetimes of less than a couple of days, which reduces their concentration above the CBL to near zero.

Both surface layer and mixed-layer similarity relationships have also been obtained for scalar variance profiles. These relationships are based on the hypothesis that CBL variance is generated solely by surface and entrainment fluxes. In practice, this may have advantages for measuring flux, particularly in the mixed layer if fast-response scalar measurements are practicable but concurrent vertical velocity measurements are not, since the mean concentration differences can be small. On the other hand, in practice mesoscale variability may contribute to the measured scalar variance and may be hard to estimate or remove from the measured variance.

Other less direct techniques exist for measuring constituent fluxes. One approach is to assume that the transport characteristics of a tracer species in the surface layer are the same as the species under consideration. Then if both eddy correlation fluxes and concentration differences are available for the tracer species, and only difference measurements are available for the species under consideration, the ratio of the unknown flux to the known flux is equal to the ratio of the tracer species difference to the unknown species difference.

Another approach is to use the budget equation of the species to solve for the surface (or entrainment) flux. The budget equation for the mean concentration of a species is given by

$$\frac{\partial S}{\partial t} + U(z)\frac{\partial S}{\partial x} + \frac{\partial F_s}{\partial z} = Q_s \tag{19}$$

where Q_s is the internal (e.g., chemical) source or sink of S, and we have assumed, for simplicity, that $V = W = 0$. This can be integrated, e.g., from the surface up to a height z and solved for the surface flux to obtain

$$F_{s0} = \frac{\partial\langle S\rangle}{\partial t} + z\langle U\rangle\frac{\partial S}{\partial x} + (\overline{ws})_z - z\langle Q_s\rangle \tag{20}$$

where $\langle\ \rangle$ denotes an average over the layer from the surface to height z. This approach has been used by aircraft flying in a Lagrangian flight pattern—i.e., advecting the flight pattern with the PBL mean wind using constant-level balloons as tracers, so that the second term on the right side of (20) is zero—and carrying out a series of flights over a day or more. In this case, the surface flux is obtained from the residual of the time rate of change, the entrainment flux, and the chemical source/sink terms.

REFERENCES

Fowler, D., and J. H. Duyzer, Micrometeorological techniques for the measurement of trace gas exchange, in M. O. Andreae and D. S. Schimel (Eds.), *Exchange of Trace Gases between Terrestrial Ecosystems and the Atmosphere*, Wiley, New York, 1989, 189–207.

Garratt, J. R. *The Atmospheric Boundary Layer*, Cambridge University Press, New York, 1992.

Kaimal, J. C., and J. J. Finnigan, *Atmospheric Boundary Layer Flows: Their Structure and Measurement*, Oxford University Press, New York, 1994.

Lenschow, D. H., Aircraft measurements in the boundary layer, in D. H. Lenschow (Ed.), *Probing the Atmospheric Boundary Layer*, American Meteorological Society, Boston, MA, 1986, pp. 29–55.

Lenschow, D. H., and B. B. Hicks (Eds.), *Global Tropospheric Chemistry: Chemical Fluxes in the Global Atmosphere*, National Center for Atmospheric Research, Boulder, CO, 1989.

Matson, P. A., and R. C. Harriss, *Biogenic Trace Gases: Measuring Emissions from Soil and Water*, Blackwell Science, Cambridge, MA, 1995.

Raupach, M. R., Canopy transport processes, in W. L. Steffen and O. T. Denmead (Eds.), *Flow and Transport in the Natural Environment: Advances and Applications*, Springer-Verlag, Berlin, 1988, 98–127.

Stull, R. B., *An Introduction to Boundary Layer Meteorology*, Kluwer Academic, Dordrecht, The Netherlands, 1988.

Wesely, M. L., Turbulent transport of ozone to surfaces common in the eastern half of the United States, in S. E. Schwartz (Ed.), *Trace Atmospheric Constituents: Properties, Transformations, and Fates*, Wiley, New York, 1983, pp. 346–370.

Wyngaard, J. C., On the maintenance and measurement of scalar fluxes, in T. J. Schmugge and J.-C. André (Eds.), *Land Surface Evaporation: Measurement and Parameterization*, Springer-Verlag, New York, 1991, pp. 199–229.

CHAPTER 11

SOURCES AND COMPOSITION OF AEROSOL PARTICLES

RICHARD ARIMOTO

1 INTRODUCTION

The atmospheric aerosol is a suspension of solid and liquid particles in the air that displays a degree of stability with respect to gravitational settling. Aerosol particles originate from a large number of sources whose influences can change dramatically over time scales of minutes to hours or can remain relatively constant for years. Although the term *aerosol* technically applies to both the solid and liquid particles and the gases in which they are suspended, common usage allows *aerosol* to refer to the particles alone, a practice that will be followed here. Increasing interest in the sources and composition of aerosols has resulted from a growing awareness of their linkages to meteorology, climate, and global change and from a better appreciation of the roles these particles play in biogeochemical cycles.

Aerosols range in size from clusters of molecules (≤ 0.001 μm radius) to ultragiant particles with radii of 100 μm or more. In one commonly used scheme (Junge, 1963), the particle size spectrum for the aerosol is separated into three classes: (1) Aitken particles (≤ 0.1 μm radius); (2) large particles (0.1 to 1.0 μm radius); and giant particles (> 1 μm radius). In another scheme (Whitby, 1978) particles with radii, $r < 0.1$ μm, which are formed by homogeneous condensation, are referred to as the nucleation mode; particles in the 0.1 to 1-μm size range are referred to as the accumulation mode because they are formed from the accumulation of nucleation mode particles and the deposition of gases. Often aerosols < 1 μm radius are simply referred to as fine particles (e.g., Heintzenberg, 1989), while larger aerosols, with a peak in the mass distribution at $r = 2$ to 5 μm, compose the coarse mode.

The various schemes for characterizing aerosol size distributions generally rely on the concept of an aerodynamic equivalent size, which is a normalization based on

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

the behavior of a spherical particle of unit density (1 g/cm^3). Most solid aerosol particles, however, are not spherical and few are of unit density. Some aerosol particles, such as sea salt under low relative humidity, are crystalline while others, including many composed of mineral matter, are angular. Aerosol particles also take the shape of rods or flat plates, and both natural and anthropogenic aerosols can be aggregates of smaller particles, which can be roughly spherical or in some cases can form chains. Biological aerosols, especially spores and pollen, often display complex geometries that have evolved to favor dispersal over long distances.

Populations of aerosol particles can be classified according to various criteria in addition to size: natural versus anthropogenic, organic versus inorganic, internally mixed versus externally mixed, mechanically generated (primary particles) versus products of gaseous reactions (secondary), etc. These various classification schemes often serve specific purposes, and the diversity in the different types of aerosols can hardly be overemphasized. This chapter will serve as an introduction to the sources and composition of the particles that compose the atmospheric aerosol. It will also summarize information regarding the sources and composition of aerosols and introduce some of the ways in which the biogeochemical cycles of aerosols are linked to those of other atmospheric constituents.

2 MECHANICALLY GENERATED AEROSOLS

Particles produced by mechanical processes tend to be larger than those resulting from gas-to-particle conversion. In general particles larger than a micrometer are mechanically formed by processes such as the wind erosion of soils, the bursting of bubbles in seawater, the shedding of plant fragments, etc. Although the relationship between the size of an aerosol particle and length of time it remains suspended in the atmosphere is complex, larger particles generally fall out of suspension more quickly than smaller ones (Fig. 1); hence large mechanically generated particles tend to have comparatively short atmospheric residence times.

Even though most mechanically generated aerosols are removed from the atmosphere close to their sources, some coarse particles remain suspended in the atmosphere for weeks and can travel thousands of kilometers before finally being deposited. While they are in suspension, aerosol particles can react with gases, with hydrometeors, and with other particles. As illustrated below, such reactions link the cycles of various atmospheric constituents in complicated ways.

The strengths of the various aerosol sources can be evaluated in several ways, and one of the most straightforward is to consider the mass of material injected into the atmosphere. As mechanical sources tend to produce physically and aerodynamically large particles, the importance of these sources is most evident when mass fluxes or related characteristics, such as particle volume, are being considered. In contrast, when evaluating source strengths with respect to the numbers of particles produced, the contributions from the mechanical sources tend to be less important compared with those producing numerous small particles via gas-to-particle conversion.

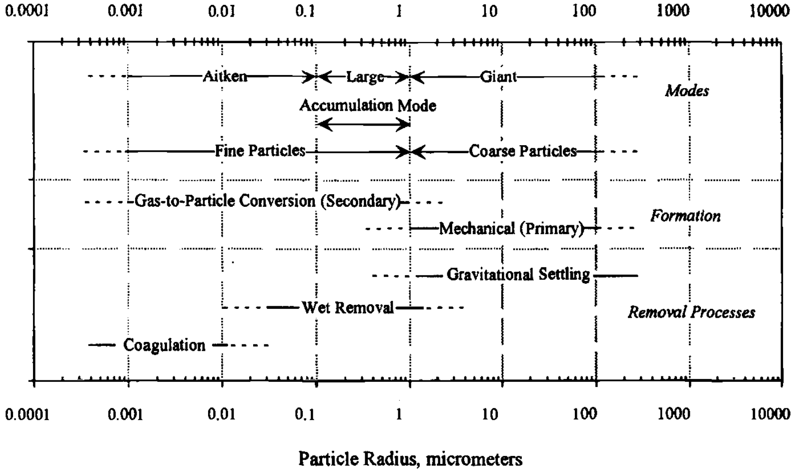


Figure 1 Characteristics of aerosol particles and the processes by which they are removed from the atmosphere.

Mineral Aerosol

The physical and chemical weathering of Earth's continental crust results in the production of mineral aerosol particles, commonly called atmospheric dust; and this represents one of the largest sources on a mass basis for natural particulate material in the atmosphere. Chinese records of dust storms date back thousands of years, and plumes of mineral aerosol over the oceans have been observed by mariners since humans have taken to the sea. Modern technology has shown that dust plumes over the oceans are among the most dramatic features seen in satellite images of aerosol optical depth (Husar et al., 1997).

Worldwide, about a third of Earth's surface can be considered potential sources for dust, but the arid and semiarid lands in Africa and Asia are the largest sources (Fig. 2). Climate clearly affects the amount of atmospheric dust produced. In general more dust is generated as the land becomes drier, but in hyperarid areas deserts can become "blown out" and less important as sources. Drought cycles also are linked to the emissions of desert dust. For example, studies at Barbados, an island in the North Atlantic Ocean, have shown that atmospheric dust concentrations increased during the Sahelian drought of the early 1970s (Prospero and Nees, 1977). Dust loads in the atmosphere also can vary over longer periods of time as a consequence of large-scale changes in climate and circulation. In this context, studies by An et al. (1990) suggest that patterns in dust deposition to the Chinese loess plateau over thousands of years can be linked to variations in the strength of the Asian winter monsoon.

The exact amount of dust injected into the atmosphere remains uncertain, but recent estimates are of the order of ~1500 Tg/yr (Andreae, 1995; Tegen et al.,

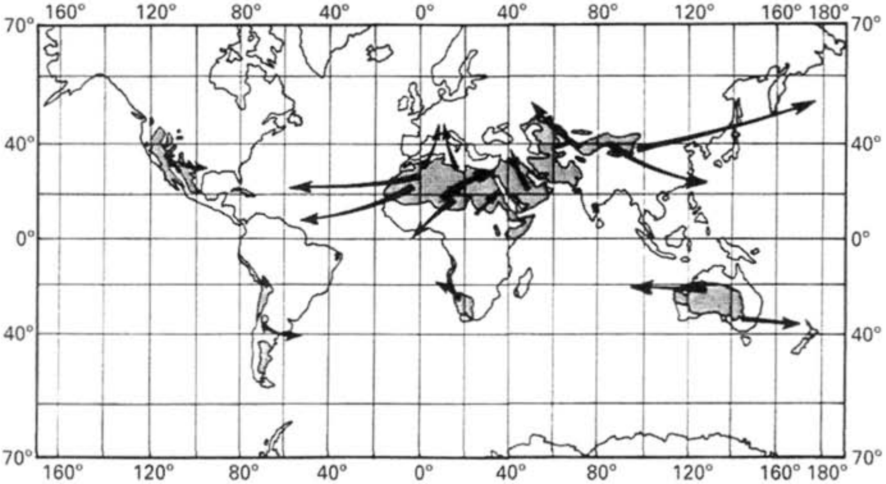


Figure 2 Sources for mineral aerosol (atmospheric dust). (From Péwe, 1981.)

1996), with an uncertainty of perhaps a factor of 2. As human activities have altered the global landscape, a portion of the atmospheric dust load can be considered anthropogenic. For example, modeling studies by Tegen et al. (1996) indicate that 50% ± 20% of the global dust flux may come from disturbed soils. On the other hand, efforts made by humans to reclaim some desert lands (Parungo et al., 1994) may have reduced the strength of natural dust sources.

Mineral particles are formed by a variety of processes, including grinding, weathering, abrasion, etc. (Pye, 1987). Once the particles are formed, the wind deflates and disperses them, but other factors, such as the sizes and shapes of the particles, the roughness of the particle bed, the cohesiveness of the particles, the presence of cementing agents, the extent of vegetative cover, and especially the amount of soil moisture influence the erodibility of the soils. Studies of the dynamics of the dust generation process by Gillette et al., (1974) showed that sandblasting of the soils was the dominant mechanism for mineral aerosol production by wind erosion, and these and other authors have shown that the wind velocity also shapes the size distributions of the suspended dust particles.

The transport of desert dust affects the global cycles of nitrogen, phosphorus, sulfur and various trace elements (Prospero, 1981; Schlesinger et al., 1990). Some areas of Earth benefit from the transport and deposition of mineral dust; for example, the fertility of the Loess Plateau in central China results in large measure from the accumulation of nutrient-rich mineral particles transported through the atmosphere from deserts in northwestern China (Liu et al., 1985). Similarly, dust originating from the Sahara Desert is transported through the atmosphere to the Central Amazon Basin where it supplies critical trace elements (Swap et al., 1992). Other parts of the continents are stripped of nutrients by the combined actions of wind and water erosion, and the economic consequences of erosion are substantial. For example, cost estimates for lost agricultural productivity, damage to waterways and infrastructure, and public health problems due to erosion by wind and water run into the billions of dollars for the United States alone (Pimentel et al., 1995).

The transport and deposition of mineral aerosol affects the cycles of a large number of trace elements in addition to those of N, P, and S. Bulk atmospheric dust particles generally have an elemental composition similar to that of average crustal rock (Rahn, 1976), and the composition of the ambient aerosol often is evaluated through "enrichment factors" (EFs) which are defined as

$$\mathrm{EF(Al, Crust)} = \frac{(\mathrm{X/Al})_{\mathrm{Aerosol}}}{(\mathrm{X/Al})_{\mathrm{Crust}}}$$

where X is any element of interest; Al is aluminum, a commonly used reference element; and the subscripts Aerosol and Crust refer to the aerosol sample of interest and the crustal reference material, respectively. Another commonly used reference element is Si, but Sc or a variety of other elements would serve the purpose equally well.

Weathered crustal material is the presumptive source for any element whose enrichment factor for a given sample approaches unity; those elements with EFs greater than ~5 have significant noncrustal sources. Direct comparisons of elemental ratios in aerosol samples versus crustal rock also show that the atmospheric loadings of mineral dust govern the concentrations of a large number of trace elements in the atmosphere (Table 1). It is important to point out, however, that individual mineral dust particles can have a composition quite different from either the bulk dust or average crustal material (Anderson et al., 1996).

In a study of erodible soils, Schütz and Rahn (1982) showed the concentrations of most elements increased as particle sizes decreased to 20 to 50 μm radius, but the concentrations reached a plateau for particles less than 10 to 20 μm in radius. These authors predicted that some variability in the elemental composition of dust should occur near the desert source areas where a significant fraction of the particles would have radii > 10 μm. More than ~1000 km from the sources, however, the bulk of the particles would be < 10 μm in radius, and therefore these authors concluded that the elemental composition of dust transported long distances would be similar to that of the continental crust.

Dust particles in the atmosphere are far from inert, and reactions occurring on dust particles have significant implications for several important chemical cycles. Direct observations of individual particles showed sulfate coatings were present on $>40\%$ of the mineral dust particles collected over the North Atlantic at 25°N, and nitrate coatings were observed on $>30\%$ of the particles (Parungo et al., 1986). Further evidence for the uptake of gaseous sulfur species on dust from the Asia–Pacific region was obtained through statistical analyses of the elemental composition of aerosols (Winchester and Wang, 1990). Reactions between dust particles and gaseous nitrogen oxides have been reported from laboratory studies (Mamane and Gottlieb, 1992) and from analyses of ambient aerosols (Wu and Okada, 1994). The formation of nitrate on dust particles via heterogeneous reactions constitutes a sink

TABLE 1 Mass Ratios of Crustal Elements to Aluminum for High-Dust Events at Barbados, Bermuda, and Izaña

	Observed			
Element	Barbados	Bermuda	Izaña	Average Crustal Rock[a]
Ba	6.2×10^{-3}	1.2×10^{-2}	9.8×10^{-3}	6.8×10^{-3}
Ca	2.9×10^{-1}	3.3×10^{-1}	3.6×10^{-1}	3.7×10^{-1}
Co	2.4×10^{-4}	2.4×10^{-4}	3.0×10^{-4}	1.2×10^{-4}
Cr	1.1×10^{-3}	1.9×10^{-3}	1.5×10^{-3}	4.4×10^{-4}
Cs	4.7×10^{-5}	6.2×10^{-5}	7.1×10^{-5}	4.6×10^{-5}
Eu	2.2×10^{-5}	2.3×10^{-5}	2.9×10^{-5}	1.1×10^{-5}
Fe	5.1×10^{-1}	6.1×10^{-1}	7.0×10^{-1}	4.4×10^{-1}
Hf	5.2×10^{-5}	5.8×10^{-5}	8.1×10^{-5}	7.2×10^{-5}
Mg	3.7×10^{-1}	3.2×10^{-1}	3.0×10^{-1}	1.6×10^{-1}
Mn	1.1×10^{-2}	9.5×10^{-3}	1.2×10^{-2}	7.5×10^{-3}
Na	1.1×10^{0}	3.3×10^{-1}	1.1×10^{-1}	3.6×10^{-1}
Rb	1.1×10^{-3}	1.7×10^{-3}	1.6×10^{-3}	1.4×10^{-3}
Sb	1.2×10^{-5}	1.3×10^{-5}	1.4×10^{-5}	2.5×10^{-6}
Sc	1.7×10^{-4}	2.0×10^{-4}	2.3×10^{-4}	1.4×10^{-4}
Ta	2.1×10^{-5}	2.1×10^{-5}	2.8×10^{-5}	2.7×10^{-5}
Tb	1.6×10^{-5}	1.5×10^{-5}	1.9×10^{-5}	8.0×10^{-6}
Th	1.6×10^{-4}	2.0×10^{-4}	2.0×10^{-4}	1.3×10^{-4}
V	1.5×10^{-3}	3.5×10^{-3}	1.6×10^{-3}	7.5×10^{-4}
Yb	4.2×10^{-5}	5.0×10^{-5}	5.2×10^{-5}	2.7×10^{-5}

[a] Taylor and McLennan (1985).

for nitrogen oxides, and reactions involving dust, N_2O_5, O_3, and HO_2 radicals may affect the cycles of photochemical oxidants, leading to decreases in tropospheric ozone near dust sources (Dentener et al., 1996).

Mineralogical studies have shown that atmospheric dust consists of silicates (quartz and feldspars); clay minerals (e.g., kaolinite, smectite, illite, mica), carbonates (calcite and dolomite), and sulfur minerals (gypsum and anhydrite) (see Pye, 1987). Mineralogical analyses of aerosol particles by Zhou and Tazaki (1996) have provided independent lines of evidence for the chemical reactivity of atmospheric dust. Their analyses showed S-rich submicrometer particles frequently are found attached to mineral dust particles, and they inferred that H_2SO_4 reacted with calcite during transport to form gypsum.

The selective removal of dust particles as a function of particle size during transport probably has little effect on elemental composition, except perhaps for the rare earths (Sholkovitz et al., 1993), but size fractionation can lead to mineralogical differences among samples (Johnson, 1976). Glaccum and Prospero (1980) similarly suggested that the proportion of quartz particles relative to clay minerals should be low in dusts that have traveled long distances owing to the preferential fallout of the quartz particles, which tend to be aerodynamically large. Even so giant quartz particles ~50 μm radius have been found in the atmosphere over the central North Pacific, thousands of kilometers from their sources (Betzer et al., 1988). An unresolved paradox confronting atmospheric scientists is that the presence of such large particles so far from their sources is difficult, if not impossible, to explain based on our current understanding of transport dynamics.

3 SOURCES PRODUCING PRIMARY AND SECONDARY PARTICLES

Primary Particles from Oceans: Sea Salt Aerosol

Surface winds cause the production of aerosols from the sea as well as from the land. The effects of the wind over the ocean are mediated by breaking waves, bursting bubbles, and to a lesser extent the formation of large spume droplets torn from waves by strong winds. The seasalt injected into the atmosphere is another of the large sources for aerosols on a mass basis, of the order of 1000 to 10,000 Tg/yr (Blanchard, 1983). One reason for the large uncertainty in this estimate is that there is no strict definition of what constitutes a sea salt aerosol particle, i.e., very large particles have such short atmospheric residence times that one might question whether they are truly suspended in the atmosphere.

The production of sea salt particles is mainly due to oceanic whitecaps. Bubbles from the whitecaps burst at the sea surface producing film droplets and jet droplets, which have quite different properties and whose proportions vary as a function of bubble size (Blanchard, 1980). Model estimates by Erickson and Duce (1988) indicate the mass median radii (MMR) for sea salt over the oceans (50% of the sea salt mass occurs on particles smaller than the MMR and 50% on particles larger than the MMR) should range between 3.0 and 7.5 m, which is in good agreement with observations. Both the amount of salt produced and the sizes of the particles vary in response to wind speed, but a consideration of the relative amounts of ocean and land covering Earth's surface together with the atmospheric loadings of dust and sea salt shows the production of sea salt particles may be less efficient than dust.

For many substances, including sulfate, the composition of fresh, bulk, sea salt aerosols is similar to that of seawater, but as noted for mineral aerosol, one must recognize that the elemental composition of individual sea salt particles may be quite different from that of the bulk aerosol. Once the jet and film drops are ejected into the atmosphere, the water in them begins to evaporate, leading to a droplet of high ionic strength, which can undergo repeated cycles of dilution and concentration. Fractionary recrystallization within the evaporating drops can be followed by shattering of the particles, and this process can lead to variations in the content of elements such as Mg, S, K, and Ca among individual sea salt particles (Mouri et al., 1993).

Some substances, including some heavy metals, organic matter, radionuclides, and nutrient species are enriched in the sea salt aerosols as a result of the scavenging of surface-active material as bubbles pass through the water column and rupture the sea surface microlayer (e.g., MacIntyre 1974; Wallace and Duce, 1975). The enrichments of trace elements in experimentally produced sea salt particles can reach several tens of thousands (Weisel et al., 1983), and this enrichment process can lead to a recycling of material between the surface ocean and the atmospheric marine boundary layer.

Sea-salt particles contain variable amounts of organic material, and large salt particles from over the remote oceans typically contain organic carbon with an isotopic composition similar to that of source materials in seawater (Buat-Ménard et al., 1989). The carbon concentrations in sea salt particles (normalized to Na) are several 100-fold higher than that of seawater, presumably as a result of the same physicochemical processes causing the enrichments of inorganic materials (Wallace and Duce, 1975). In some areas of the oceans, terrestrial sources also can contribute significant amounts of carbon to large particles, but those continental sources, whether anthropogenic or natural, are more important for submicrometer marine aerosols.

Although the organic composition of marine aerosols is only partially characterized at best, numerous organic compounds have been detected in marine aerosols. For example, Peltzer and Gagosian (1989) investigated aliphatic hydrocarbons, wax esters, fatty alcohols, sterols, fatty acids, and long-chain unsaturated ketones in aerosols from several sites in the Pacific Ocean. These compounds were used as biomarkers in studies of the sources, transport, and transformation of organic material in the marine atmosphere. Kawamura and Usukura (1993) investigated dicarboxylic acids in the western North Pacific, and they concluded the diacids were mainly from Asia, but some diacids were produced by photochemical reactions in situ.

Sea-salt particles also react with a variety of gaseous components of the marine atmosphere; most notably HNO_3, methanesulfonic acid, and H_2SO_4 are sorbed by liquid sea salt particles and HCl and HF are displaced (Ericksson, 1960; Okada et al., 1978). The modeling of acid displacement reactions is made difficult by the nonideal behavior of the high solute concentrations in the sea salt droplets (Brimblecombe and Clegg, 1988). However, analyses of individual particles from the North Atlantic suggest that Cl loss from sea salt can be accompanied by the formation of $NaNO_3$ (Pósfai et al., 1995). Reactions of sea salt with various N gases, such as NO_2, $ClNO_3$, and N_2O_5, have been observed, and those reactions could lead to the formation of $NaNO_3$ (Schroeder and Urone, 1974; Finlayson-Pitts et al., 1989; Keene et al., 1990). Such reactions also could generate reactive Cl atoms; and analogous to the hydroxyl radical, the atomic chlorine would participate in photochemical reactions with various organic substances.

Other aqueous-phase reactions in sea salt aerosols involve the oxidation of SO_2 by O_3 to non-sea-salt (NSS) sulfate (i.e., the sulfate in excess of what can be attributed to sea salt from unfractionated seawater) (Sievering et al., 1992, 1995; Chameides and Stelson, 1992). Recent analyses by Keene et al. (1998) indicate that the oxidation of SO_2 by ozone in sea salt aerosols is only a minor source for NSS sulfate, and these authors suggested that oxidation of SO_2 primarily occurs via another pathway, possibly involving HOX, where X is chlorine or bromine. As SO_2 originates from anthropogenic as well as natural sources, these reactions not only show how heterogeneous reactions can affect the composition of aerosols but also illustrate how intimately the chemistry of pollutants can be linked with the cycles of natural substances in the atmosphere.

Secondary Particles from Marine Sulfur Compounds

Some sources produce aerosols both by mechanical processes and by gas-to-particle conversion, the oceans being a good example of this. In addition to the mechanically generated sea salt aerosol, the oceans emit gaseous compounds that can be oxidized and eventually produce aerosol particles. The most important chemicals in this case contain sulfur, especially dimethylsulfide (DMS), methanesulfonic acid (MSA), and sulfate. Interest in the marine sulfur cycle increased enormously after a hypothesis was proposed connecting oceanic phytoplankton to aerosols to clouds and hence to climate (Charlson et al., 1987). In this hypothesis, DMS produced by marine phytoplankton evades from the ocean and is oxidized to sulfate aerosol (and MSA), which can act as cloud condensation nuclei (CCN). The number of CCN in the atmosphere affects the reflectivity of clouds (the cloud albedo), and in this way oceanic emissions of reduced sulfur gases are linked to climate. Sulfate aerosols, whether marine derived or originating from continental emissions, also can reflect solar radiation back to space, and in so doing influence weather and climate directly.

There is at present no universally accepted chemical mechanism for the formation of NSS sulfate via DMS oxidation. One of the controversies that arose with respect to the oxidation of DMS was whether sulfur dioxide (SO_2) was an important intermediate in the pathway leading to sulfate aerosol (Bandy et al., 1992; Lin and Chameides, 1993). Recent studies indicate that the pathway from DMS to NSS sulfate does indeed include SO_2 as an intermediate, but our knowledge of the other compounds and reactions in the pathway is far from complete (Keene et al., 1998). Whether from DMS or from pollution sources, SO_2 can be further oxidized both in the gas and liquid phase to H_2SO_4, but these processes are sensitive to gas-phase nitric acid and ammonia concentrations (Clegg and Toumi, 1997), again demonstrating links between the chemistry of aerosols and gaseous species.

Over vast areas of Earth's oceans, from about 30°N to 30°S, the ratio of biogenic NSS sulfate to MSA tends to be constant, so much so that a MSA/NSS sulfate mass ratio of $\sim$18 to 20 has been used as a diagnostic for marine biogenic sulfate (Savoie and Prospero, 1989; Savoie et al., 1994; Arimoto et al., 1996). The relative amounts of MSA and NSS sulfate do change with latitude however; above $\sim$30° (N or S) more of the biogenic sulfur occurs as MSA (Berresheim, 1987; Pszenny et al., 1989; Koga et al., 1991; Bates et al., 1992). One of the central issues of the marine sulfur cycle currently being investigated is the extent to which this latitudinal dependence in MSA/NSS sulfate ratios is driven by the temperature dependencies of various reactions in the DMS oxidation pathways versus the influences of other photochemically active compounds.

In the marine atmosphere gaseous H_2SO_4 either can deposit on existing surfaces, again potentially involving sea salt, or it can form new sulfate particles via homogeneous nucleation. A fundamental issue concerning the marine sulfur cycle has to do with where new aerosol sulfate particles form. The cycling and fate of particles formed in the marine boundary layer (MBL) would be far different from those formed in the free troposphere, and therefore this issue has important implications for the direct and indirect effects of the aerosols on solar radiation. While new particles form in the marine boundary layer under certain conditions (Covert et al., 1992), most of the new particle production evidently occurs in the proximity of clouds (Hegg et al., 1990; Perry and Hobbs, 1994). Recent studies conducted for ACE-1 (aerosol characterization experiment) showed that few new particles formed in the MBL over the Southern Ocean south of Australia (Clarke et al., 1997). Instead layers of new particles, DMS, MSA, and H_2SO_4 were observed in the free troposphere in the outflow regions of clouds at altitudes of several kilometers. Andreae and Crutzen (1997) suggest the DMS–aerosol–climate connection may still pertain because the subsidence of aerosol-laden air from the free troposphere into the MBL can supply particles that are initially too small to act as CCN but through heterogeneous or cloud processes can grow and become CCN.

Volcanoes

Volcanoes are another natural source producing both primary and secondary aerosol particles. One of the distinctive aspects of volcanic emissions is that strong eruptions can inject materials directly into the stratosphere where aerosol-induced effects on the balance of solar radiation and ozone depletion, for example, can persist for years (McCormick et al., 1995). Much of the work on volcanic aerosols has, in fact, focused on the stratosphere, but that topic will not be covered in this chapter.

The amounts of material produced by volcanoes can be quite considerable when they are active, but volcanic eruptions are episodic, and most volcanoes exhibit periods of dormancy following the releases of gases, particles, and lava that constitute the active phase. Explosive volcanoes eject primary particles, mainly silicate dust particles and ash, into the atmosphere; but many of the primary particles are so large that they settle out quickly and close to their source. Andreae (1995) observed that during periods of extreme volcanic activity, as much as 10,000 Tg of dust could be produced per year. An annual flux of that magnitude would be larger than that from any other aerosol source, with the possible exception of sea salt. In less active times, the flux of primary particles from volcanoes, ~4 Tg/yr, would be almost negligible on a global scale. The long-term average production of primary particles from volcanoes has been estimated as 33 Tg/yr (Andreae, 1995).

Volcanoes also release water vapor, CO_2, SO_2, fluorine, and chlorine into the atmosphere (Lambert et al., 1988; Symonds et al., 1988) both from explosive events and during noneruptive activity. Secondary particles, mainly sulfuric acid droplets, can form from these gaseous emissions but, on a mass basis, the production of secondary particles during periods of high volcanic activity is much smaller than that of primary particles. For nonexplosive, basaltic volcanoes and fissures, the masses of primary and secondary particles produced are more nearly comparable, but the combined production of primary and secondary particles from these sources is considerably smaller than during more active periods. More important, the volcanic emissions of sulfur (9.3 to 11.8 Tg S per year in total) are equivalent to 10 to 30% of the total anthropogenic sulfur flux into the atmosphere; this amount is large enough to significantly affect the chemistry of sulfur in the atmosphere (Berresheim et al., 1995).

Volcanoes inject substantial quantities of trace elements into the atmosphere, and as plumes of volcanic material cool, gaseous species condense and attach to particles. As a result the composition of volcanic aerosols typically is quite different from the parent magma. Compared with crustal material, volcanic particles tend to be enriched with relatively volatile elements, including Zn, Cu, Au, Pb, As, Cd, Sb, and Se (Buat-Ménard, 1990). These enrichments vary not only among different volcanoes but also during different eruptive stages for a particular volcano. The elemental composition of the aerosol is particularly sensitive to the amount of nondegassed magmatic material brought to the surface by the volcanic activity. Along this same line, iridium enrichments were found during the eruption of Kilauea (Zoller et al., 1983) but not six other volcanoes, presumably reflecting the different types of magma involved in the eruptions.

Globally volcanoes supply ~50% of the ^{210}Po in the atmosphere (Lambert et al., 1982). This nuclide is the last radioactive daughter in the decay series of the naturally occurring radionuclide ^{238}U. In Antarctica, volcanoes are a particularly important source for volatile radionuclides because snow and ice cover minimizes the impact of many other sources (Polian and Lambert, 1979). These authors found that the $^{210}\mathrm{Po/SO_2}$ ratios in the plume from Mt. Erebus were 30-fold higher than those for areas not affected by volcanoes. The recognition of a strong volcanic source for ^{210}Po enabled Lambert et al. (1988) to estimate trace element fluxes from volcanoes by scaling them relative to ^{210}Po. The uncertainties in these figures are estimated to be a factor of 3, but in general the volcanic inputs of trace elements probably are $<20\%$ of their total atmospheric inputs (Nriagu, 1989). One exception to this is Bi whose volcanic source strength is perhaps 10 times higher than the inputs from either natural or anthropogenic sources (Lee et al., 1986; Lambert et al., 1988).

Biological Aerosols

The winds not only generate aerosols but also scatter preformed particles, including pollen grains, spores, bacteria, viruses, algae, fungi, nematodes, protozoa, and fragments of plant and animal tissues. The concentrations of certain kinds of biological aerosols are monitored for allergy sufferers through the familiar air quality indices of fungal spores and pollen. More generally, however, investigations of biological aerosols have been limited despite their relevance for studies of air quality, climate, chemical cycles, and so forth. Biological aerosols span a large range in size, from radii of $<0.1\ \mu$m for viruses to hundreds of micrometers for large pollen grains and spores. Evolution has shaped certain types of pollen and spores to favor their dispersal through the atmosphere, and thus even though they are geometrically quite large, such particles can be transported over long distances and to great heights. For example, culturable fungi have been recovered from the atmosphere at altitudes between 57 and 77 km (Imshenetsky et al., 1978).

Fungi are among the most abundant of the viable biological aerosols (Duce et al., 1983), and their numbers vary strongly with location, season, meteorology, and diurnal cycle. Worldwide, fungi of the genus *Cladosporium* are the most abundant, and in temperate regions fungal spores from this genus are especially abundant in summer and early fall. Spores and hyphal fragments from other genera of fungi, including *Alternaria*, *Drechslera*, *Epicoccum*, *Aspergillis*, and *Penicillium*, are commonly collected in samples of particulate matter in air. Under some circumstances, such as crop harvesting or mowing, the numbers of airborne fungal spores from local sources can reach impressive numbers, up to 10^9 spores per cubic meter of air (Levetin, 1995). Many species of fungi contain substances called allergans that trigger allergic reactions in humans, and various types of fungi cause respiratory and opportunistic infections. Leathers (1981) reported that each year several hundred thousand persons are infected with airborne, disease-causing fungi in the United States alone. Among the pathogenic fungi, *Coccidiodes imitis*, which is endemic to the southwestern United States and causes "valley fever," presents particularly serious health and economic problems. Each year several hundred persons require hospitalization because of this fungus, which is spread via spores transported through the atmosphere from the desert regions to metropolitan areas.

Bacteria are patchily distributed in the atmosphere, and they are released from both natural and anthropogenic sources by various mechanical processes including wind abrasion, agricultural activities, etc. Typical concentrations of culturable bacteria range from 10 to 1000 colony-forming units per cubic meter of air, but the numbers of bacteria in the atmosphere can reach 10^9 per cubic meter under disturbed conditions (Muilenberg, 1995). Bubbles rising in the oceans scavenge bacteria and viruses from the water column (Blanchard, 1983; Baylor et al., 1977, respectively) in the same way organic carbon and trace elements are scavenged, and bacterial enrichments of several 100-fold have been observed in the aerosol relative to seawater. Airborne bacteria produced in the operation of wastewater treatment plants pose potential health hazards (Hickey and Reist, 1975), but the best known case of a health problem associated with biological aerosols is Legionnaires' disease caused by bacteria (*Legionella pneumophilia*) growing in air-conditioning cooling towers (Dondero et al., 1980).

Pollen grains contain genetic material from male seed plants, and one group of pollen-producing plants has been classified as anemophilous because they entrust pollination of the female flowers to the wind rather than insects (Muilenberg, 1995). Pollen is produced in flowers, and the amount of airborne pollen is governed by the life cycles of plants, which in turn are influenced by extrinsic factors such as temperature, the photocycle, and precipitation. The walls of many pollen grains are composed of a resistant outer layer made of a polymer called sporopollenin and an inner wall of cellulose. Allergans associated with pollen most commonly affect the upper respiratory tract as hay fever and related maladies, but pollen exposure also can lead to asthma. The pollen from anemophilous plants, which are more common in temperate areas and less so in the tropics, tends to be smaller than from the entomophilous (insect-pollinated) plants. Studies of pollen grains in marine sediments have been used in paleoclimate reconstructions, particularly those involving paleowinds (e.g., Hoogheimstra, 1987).

Aerosols from Biomass Burning

The burning of living and dead vegetation (biomass burning) is widespread over Earth, and this is a globally significant source for aerosols and for a variety of radiatively active and chemically reactive trace gases. Most of the biomass burned is caused by human activities as opposed to natural fires, and this mainly occurs in the tropics, involving savannas more than forests (Hao and Liu, 1994). Some burning sources are persistent, but emissions from savanna burning vary biennially because the growth of the savanna vegetation occurs during the wet season, and afterwards as biomass dries, it is burned.

Aerosols produced by biomass burning are composed of some black carbon (soot) but mainly organic carbon with hydrogenated and oxygenated functional groups. The amount of black carbon produced by fires is highly variable, and this is determined by the type of fuel consumed and whether the fire is in the ignition phase, flaming, or smoldering. Moreover, the composition of the aerosols can change quite rapidly—over time scales of seconds to minutes—as the particles are advected away from the fire. The transformation of particles in a smoke plume can lead to internal mixtures, i.e., particles with cores of black carbon and low-volatility organic compounds become coated with outer layers of more volatile organics (Mazurek et al., 1996).

The amount of particulate matter put into the atmosphere by biomass burning is estimated to be ~90 Tg per year (37 Tg/yr of this is from savanna burning), and this amounts to more than 20% of the total suspended particulates from all anthropogenic sources (Andreae et al., 1996). The black carbon from biomass burning (60 Tg/yr) accounts for an even larger fraction, i.e., two-thirds of the global black carbon emissions from all anthropogenic sources. These authors estimate that the number of cloud condensation nuclei generated globally by biomass burning activities is 35×10^{27}, over 10% of which is from savanna fires.

The organic carbon composition of biomass burning aerosols is not fully characterized owing at least in part to the diversity of fuels burned in different geographical regions. However, studies of tropical biomass burning have shown the major organic components are straight-chain, aliphatic, and oxygenated compounds, triterpenoids from plant waxes, resins/gums, and biopolymers (Simoneit et al., 1996). The fatty acid composition of aerosols produced in laboratory and field burns (Ballentine et al., 1996) was found to be dominated by saturated even-chain compounds, reflecting the importance of terrigenous plant waxes.

The polycyclic aromatic hydrocarbons (PAHs) in aerosols from biomass burning include biphenyl, trimethylnapthalenes, phenanthrene, anthracene, methylphenanthrenes, fluoranthene, pyrene, methylpyrenes, chrysene, benzanthracene, benzofluoranthenes, benzo ([e] and [a]) pyrenes, indenopyrene, benzo(ghi)perylene, and coronene (Simoneit et al., 1996; Ballentine et al., 1996). Simoneit et al. (1996)

also reported the occurrence of oxy-PAHs in burning products, including fluorenone, anthra-9,10-quinone, cyclopenta(def)phenanthrene-4-one, benzo[a]fluorene-11-one, benzanthrone, and napthanthrone. Both the PAHs and the oxy-PAHs are produced by incomplete combustion, and the production and transport of PAHs is of particular concern owing to the carcinogenicity of these compounds. Several organic compounds (amyrones, friedeline, aromatic A-noroleananes, syringaldehyde, vanillin, syringic acid, and vanillic acid) have been proposed as tracers of aerosols from biomass burning (Simoneit et al., 1996).

Other proposed tracers of biomass burning include the trace elements potassium and zinc (Andreae, 1983). These elements are enriched in biomass burning aerosols, and the ratios of K and Zn to black carbon (Cb) in aerosols from burning have been found to be quite constant (K/Cb ~1.3; Zn/Cb ~5.4‰, Cachier et al., 1996). Elemental analyses of coarse aerosols from prescribed savanna burns by Maenhaut et al. (1996) showed the fires were a major source for black carbon, P, K, Ca, Mn, Zn, Sr, and I. For fine particles ($r < 1$ μm), the flaming and smoldering phases of the fires were a major source for black carbon, Cl, Br, I, K, Cu, Zn, Rb, Sb, Cs, and Pb; and under flaming conditions also important for Na and S. These authors also found that fires could mobilize significant amounts of mineral dust, presumably through the convection associated with the fires.

Other Pollution-Derived Aerosols

Aerosols, both primary and secondary, are generated by a variety of pollution sources in addition to biomass burning; these include industrial processes, electric utilities, transportation, construction, and other fuel combustion. While large-scale urban air pollution is a consequence of modern industrial and technological development, smoke produced by indoor fires was perhaps the earliest form of air pollution (Brimblecombe, 1995). It is remarkable that the emissions from many anthropogenic sources are known with greater certainty than are those from natural sources. Even so, there are major gaps in our understanding of anthropogenic aerosol sources on a global scale (Graedel et al., 1993). These gaps include limited information on the geographical distribution of sources, inadequate measurements of the sizes of the particles emitted by the various sources, a lack of knowledge concerning the transformations of the particles as they age, and only a recent appreciation of the complex ways in which the entire mix of atmospheric constituents, especially nitrogen oxides, volatile organic compounds (VOCs), and ozone, affect the formation and composition of aerosols (Meng et al., 1997).

Pollution emissions are, of course, subject to many of the same processes discussed above for natural aerosol sources, such as new particle formation via gas-to-particle conversion or the condensation of volatile materials in plumes emitted by high-temperature sources. In addition studies of semivolatile organic compounds, including pollution-derived PAHs, have shown that the partitioning of these compounds between the gas phase and particles is largely determined by their subcooled liquid–vapor pressures (Bidleman and Foreman, 1987). Calculations by these authors based on the approach of Yamasaki et al. (1984) indicated that for typical, urban, suspended-particle loads, 11 to 55% of the total mass of a substance with a vapor pressure of 10^{-6} Torr should be partitioned in the particulate phase. Other factors besides the total suspended particle load that can influence the vapor/particle partitioning of organic substance include relative humidity (Pankow et al., 1993), temperature, and the radiative flux (Kamens et al., 1988). Gas–particle partitioning studies of organics have shown that it is important to determine whether the gas-phase species are adsorbed to a solid particle's surface or absorbed into a liquid phase (Pankow, 1994). This same issue is certainly relevant to the gas–particle partitioning of inorganic species, specifically with respect to wetted aerosols and liquid droplets. Partitioning among the gas, liquid, and solid phases in the atmosphere is especially relevant for the chemistry of sulfur and nitrogen oxides and acid deposition, but further discussion of this topic is beyond the scope of this chapter.

Some quantitative estimates of trace element emissions from anthropogenic sources were produced in the 1970s and 1980s (e.g., Lantzy and Mackenzie, 1979; Nriagu, 1989; Pacyna, 1986; Nriagu and Pacyna, 1988). These emission estimates clearly show that the biogeochemical cycles of a number of trace elements have been severely perturbed by human activities (Table 2). However, of the trace elements it is atmospheric Pb that has been subject to the greatest perturbation, and while anthropogenic Pb has been spread throughout Earth, largely as a result of atmospheric transport (e.g., Murozumi et al., 1969; Patterson, 1987), the concentrations of Pb in the atmosphere have started to decline in response to the phase out of leaded gasolines (Huang et al., 1996).

There are examples of aerosol pollution even more extreme than Pb, and these involve the atmospheric releases of substances that exist purely as a result of human activities. Examples of the substances involved include synthetic organic chemicals, such as polychlorinated biphenyls (PCBs) and various types of pesticides. Radio-

TABLE 2 Percent of Total Atmospheric Emissions from Natural Sources[a]

Element	Percent from Natural Sources
Antimony	41
Arsenic	39
Cadmium	15
Chromium	59
Copper	44
Lead	4
Manganese	89
Mercury	41
Molybdenum	48
Nickel	35
Selenium	58
Vanadium	25
Zinc	34

[a] Data from Nriagu (1989).

active nuclides produced by nuclear weapons testing and by nuclear reactors also have been released into the environment, and these man-made nuclides also make up a component of the contemporary aerosol. The dispersal of these and other pollutant aerosols to the most remote parts of the globe is a measure of the efficiency with which atmospheric transport operates.

4 CONCLUDING REMARKS

Much of the current interest in aerosols focuses on two areas, first the connections between aerosols and climate, and second the links between aerosol pollution and human disease. Interest in the aerosol–climate connection centers on the direct and cloud-mediated effects of aerosols on solar radiation. Concern over the health effects of aerosol particles < 2.5 µm in diameter (the PM-2.5 fraction) is responsible for the recently enacted standards regarding particulate matter (PM) in the United States (Federal Register, 1997). Both of these general areas of interest require accurate information on the formation, composition, chemical reactivity, and transport of aerosol particles. Models and measurements have been used to determine where aerosols are produced and what they are made of, but advances in remote sensing and analytical methods will lead to a more comprehensive picture of the sources, composition, and reactivity of the atmospheric aerosol.

Beyond these concerns, it is now recognized that understanding the chemistry of aerosols is a key to dealing with other atmospheric constituents of immediate concern, including, for example, certain photochemical oxidants (Finlayson-Pitts and Pitts, 1997). Furthermore, understanding the linkages among the chemical cycles of aerosols, VOCs, and NO_x, will be required for the development of effective pollution control strategies (Meng et al., 1997). This newly recognized need for an integrated approach to understanding and controlling air pollution also will lead to improvements in the socioeconomic models that are becoming increasingly important in policy-making decisions.

Climate models are incorporating more and more information on the sources, composition, and fluxes of aerosols. Improved estimates of the quantities of gases and aerosols emitted into the atmosphere from natural and anthropogenic sources are being developed in association with the Global Emissions Inventory Activity (GEIA), a component of the International Global Atmospheric Chemistry Program (IGAC, e.g., Graedel et al., 1993). Given the global dimensions of aerosol pollution problems coupled with the heterogeneity of the aerosol distribution, it is appropriate that coordinated international efforts are being mounted to address both scientific and public health issues.

REFERENCES

An, S., T. S. Liu, Y. C. Lu, S. C. Porter, G. Kukla, W. H. Wu, and Y. M. Hua, The long-term paleomonsoon variation recorded by the loess-paleosol sequence in central China, *Quat. Int.*, *7/8*, 91–95, 1990.

Anderson, J. R., R. R. Buseck, T. L. Patterson, and R. Arimoto, Characterization of the Bermuda tropospheric aerosol by combined individual-particle and bulk-aerosol analysis, *Atmos. Environ.*, *30*, 319–338, 1996.

Andreae, M. O., Soot carbon and excess fine potassium: Long-range transport of combustion derived aerosols, *Science*, *220*, 1148–1151, 1983.

Arimoto, R., R. A. Duce, D. L. Savoie, J. M. Prospero, R. Talbot, J. D. Cullen, U. Tomza, N. F. Lewis, and B. J. Ray, Relationships among aerosol constituents from Asia and the North Pacific during PEM-West A, *J. Geophys. Res.*, *101*, 2011–2023, 1996.

Andreae, M. O., Climatic effects of changing atmospheric aerosol levels, in A. Henderson-Sellers (Ed.), *World Survey of Climatology*, Vol. 16: *Future Climates of the World*, Elsevier, Amsterdam, 1995, pp. 341–392.

Andreae, M. O., E. Atlas, H. Cachier, W. R. Cofer III, G. W. Harris, G. Helas, R. Koppmann, J-P. Lacaux, and D. E. Ward, Trace gas and aerosol emissions from savanna fires, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 1, MIT Press, Cambridge, MA, 1996, pp. 278–295.

Andreae, M. O., and P. J. Crutzen, Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry, *Science*, 276, 1052–1058, 1997.

Ballentine, D. C., S. A. Macko, V. C. Turekian, W. P. Gilhooly, and B. Martincigh, Chemical and isotopic characterization of aerosols collected during sugar cane burning in South Africa, in J. S. Levine (Ed), *Biomass burning and global change*, Vol. 1. MIT Press, Cambridge, MA, 1996, pp. 460–465.

Bandy, A. R., D. L. Scott, B. W. Blomquist, S. H. Chen, and D. C. Thornton, Low yields of SO_2 from dimethyl sulfide oxidation in the marine boundary layer, *Geophys. Res. Lett.*, *19*, 1125–1127, 1992.

Bates, T. S., J. A. Calhoun, and P. K. Quinn, Variations in the methanesulfonate to sulfate molar ratio in submicrometer marine aerosol particles over the South Pacific Ocean, *J. Geophys. Res.*, *97*, 9859–9865, 1992.

Baylor, E. R., V. Peters, and M. B. Baylor, Water-to-air transfer of virus, *Science*, *197*, 763–764, 1977.

Berresheim, H., Biogenic sulfur emissions from the Subantartic and Antartic oceans, *J. Geophys. Res.*, *92*, 13, 245–13, 1987.

Berresheim, H., P. H. Wine, and D. D. Davis, Sulfur in the atmosphere, in H. B. Singh (Ed.), *Composition, Chemistry and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 251–307.

Betzer, P. R., K. L. Carder, R. A. Duce, J. T. Merrill, N. W. Tindale, M. Uematsu, D. K. Costello, R. W. Young, R. A. Breland, R. E. Bernstein, and A. M. Greco, Long-range transport of giant mineral aerosol particles, *Nature*, *336*, 568–571, 1988.

Bidleman, T. F., and W. T. Foreman, Vapor–particle partitioning of semivolatile organic compounds, in R. A. Hites and S. J. Eisenreich (Eds.), *Sources and Fates of Aquatic Pollutants*, American Chemical Society, Washington, DC, 1987, pp. 27–56.

Blanchard, D. C., The production, concentration, and vertical distribution of the sea–salt aerosol, *Ann. N.Y. Acad Sci.*, *338*, 330–347, 1980.

Blanchard, D. C., The production, distribution, and bacterial enrichment of the sea–salt aerosol, in P. S. Liss and W. G. N. Slinn (Eds.), *Air-Sea Exchange of Gases and Particles*, D. Reidel Publishing Company, Dordrecht, Holland, 1983, pp. 299–405.

Brimblecombe, P., History of air pollution, in H. B. Singh (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 1–18.

Brimblecombe, P., and S. L. Clegg, The solubility and behaviour of acid gases in the marine aerosol, *J. Atmos. Chem.*, 7, 1–18, 1988.

Buat-Ménard, P., H. Cachier, and R. Chesselet, Sources of particulate carbon in the marine atmosphere, in J. P. Riley, R. Chester, and R. A. Duce (Eds), *Chem. Oceanogr., Vol. 10*, Academic Press, London, 1989, 252–279.

Buat-Ménard, P., Global source strength and long–range atmospheric transport of trace elements emitted by volcanic activity, in A. H. Knap (Ed.), *The Long-Range Atmospheric Transport of Natural and Contaminant Substances*, Kluwer Academic, Dordrecht, 1990, pp. 163–175.

Cachier, H., C. Liousse, M.-H. Pertuisot, A. Gaudichet, F. Echaler, and J.-P. Lacaux, African fire particulate emission and atmospheric influence, in J. S. Levine (Ed), *Biomass burning and global change*, Vol. 1. MIT Press, Cambridge, MA, 1996, pp. 428–440.

Chameides, W. L., and A. W. Stelson, Aqueous–phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Charlson, R. J., J. E. Lovelock, M. O. Andreae, and S. G. Warren, Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, *Nature 326*, 655–661, 1987.

Clarke, A. D., J. L. Varner, F. Eisele, R. L. Mauldin, D. Tanner, and M. Litchy, Particle production in the remote marine atmosphere: Cloud outflow and subsidence during ACE–1, *J. Geophys Res.*, 103, 16397–16409, 1997.

Clegg, N. A., and R. Toumi, Sensitivity of sulphur dioxide oxidation in sea salt to nitric acid and ammonia gas phase concentrations, *J. Geophys. Res.*, 102, 23241–23249, 1997.

Covert, D. S., V. N. Kapustin, P. K. Quinn, and T. S. Bates, New particle formation in the marine boundary layer, *J. Geophys. Res.*, 97, 20581–20589, 1992.

Dentener, F. J., G. R. Carmichael, Y. Zhang, J. Lelieveld, and P. J. Crutzen, Role of mineral aerosol as a reactive surface in the global troposphere, *J. Geophys. Res.*, *101*, 22869–22889, 1996.

Dondero, Jr., T. J., R. C. Rendtorff, G. F. Mallison, R. M. Weeks, J. S. Levy, E. W. Wong, and W. Schaffner, An outbreak of Legionnaires' disease associated with a contaminated air-conditioning cooling tower, *N. Engl. J. Med*, *302*. 365–370, 1980.

Duce, R. A., V. A. Mohnen, P. R. Zimmerman, D. Grosjean, W. Cautreels, R. Chatfield, R. Jaenicke, J. A. Ogren, E. D. Pellizzari, G. T. Wallace, Organic material in the global troposphere, *Rev. Geophys. Space Phys.*, *21*, 921–952, 1983.

Ericksson, E., The yearly circulation of chloride and sulphur in nature, meteorological, geochemical and pedological implications, Part 2, *Tellus*, *12*, 63–109, 1960.

Erickson, D. J. and R. A. Duce, On the global flux of atmospheric sea salt, *J. Geophys. Res.*, *93*, 14079–14088, 1988.

Federal Register, *62*, 38762–38896, 1997.

Finlayson–Pitts, B. J., M. J. Ezell, and J. N. Pitts, Formation of chemically active chlorine compounds by reactions of atmospheric NaCl particles with gaseous N_2NO_5 and $ClONO_2$, *Nature*, *337*, 241–244, 1989.

Finlayson–Pitts, B. J., and J. N. Pitts, Jr., Tropospheric air pollution: ozone, airborne toxics, polycyclic aromatic hydrocarbons, and particles, *Science*, *276*, 1045–1052, 1997.

Gillette, D. A., I. H. Blifford, and D. W. Fryrear, The influence of wind velocity on the size distributions of aerosols generated by the wind erosion of soils, *J. Geophys. Res.*, *79*, 4068–4075, 1974.

Glaccum, R. A., and J. M. Prospero, Saharan aerosols over the tropical North Atlantic-mineralogy, *Mar. Geol.*, *37*, 295–321, 1980.

Graedel, R. E., T. S. Bates, A. F. Bouwman, D. Cunnold, J. Dignon, I. Fung, D. J. Jacob, B. K. Lamb, J. A. Logan, G. Marland, P. Middleton, J. M. Pacyna, M. Placet, and C. Veldt, A compilation of inventories of emissions to the atmosphere, *Global Biogeochem. Cycles*, *7*, 1–26, 1993.

Hao, W. M., and M. H. Liu, Spatial and temporal distribution of tropical biomass burning, *Global Biogeochem. Cycles*, *8*, 495–503, 1994.

Hegg, D. A., L. F. Radke, and P. V. Hobbs, Particle production associated with marine clouds, *J. Geophys. Res.*, *95*, 13917–13926, 1990.

Heintzenberg, J., Fine particles in the global troposphere - a review, *Tellus*, *41B*, 149–160, 1989.

Hickey, J. L. S., and P. C. Reist, Health significance of airborne microorganisms from wastewater treatment processes, *J. Water Pollut. Control Fed.*, *47*, 2758–2773, 1975.

Hoogheimstra, H., Variations of the NW African trade wind regime during the last 140,000 years: Changes in pollen flux evidenced by marine sediment records, in M. Leinen and M. Sarnthein (Eds.), *Paleoclimatology and Paleometeorology: Modern and Past Patterns of Global Atmospheric Transport*, NATO ASI Series, Kluwer Academic, Dordrecht, 1987, pp. 733–770.

Huang, S., R. Arimoto, and K. Rahn, Changes in atmospheric lead and other pollution-derived trace elements at Bermuda, *J. Geophys. Res.*, *101*, 21033–21040, 1996.

Husar, R. B., J. M. Prospero and L. L. Stowe. Characterization of tropospheric aerosols over the oceans with the NOAA/AVHRR optical thickness operational product. *J. Geophys. Res.* *102*, 16, 16889–16909, 1997.

Imshenetsky, A. A., S. V. Lysenko, G. A. Kazakov, Upper boundary of the biosphere, *Appl. Environ. Microbiol.*, *35*, 1–5, 1978.

Johnson, L. R., Particle-size fractionation of eolian dusts during transport and sampling, *Marine Geo.*, *21*, M17–M21, 1976.

Junge, C. E., *Air Chemistry and Radioactivity*, Academic Press, New York, 1963.

Kamens, R. M., Z. Guo, J. N. Fulcher, and D. A. Bell, Influence of humidity, sunlight, and temperature on the daytime decay of polyaromatic hydrocarbons on atmospheric soot particles, *Environ. Sci. Technol.*, *22*, 103–108, 1988.

Kawamura, K., and K. Usukura, Distributions of low molecular weight dicarboxylic acids in the North Pacific aerosol samples, *J. Oceanogr.*, *49*, 271–283, 1993.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz–Tokos, H. Sievering, and J. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Keene, W. C., R. Sander, A. A. P. Pszenny, R. Vogt, P. J. Crutzen, and J. N. Galloway, Aerosol pH in the marine boundary layer: A review and model evaluation, *J. Aerosol Sci.*, 29, 239–356, 1998.

Koga, S. H. Tanaka, M. Yamato, T. Yamanouchi, F. Nishio and Y. Iwasaka, Methanesulfonic acid and non-sea-salt sulfate over both hemispheric oceans, *J. Meteorol. Soc. Jpn.*, *69*, 1–14, 1991.

Lambert, G., M.-F. Le Cloarec, and M. Pennisi, Volcanic output of SO_2 and trace metals: A new approach, *Geochim. Cosmochim. Acta*, 52, 39–42, 1988.

Lambert, G., G. Polian, J. Sanak, A. Buisson, R. Ardouin, and A. Jegou, Volcanic output of long-lived radon daughters, *J. Geophys. Res.*, *87*, 11103–11108, 1982.

Lantzy, R. L., and F. T. Mackenzie, Global cycles and assessment of man's impact, *Geochim. Cosmochim. Acta*, 43, 511–515, 1979.

Leathers, C. R., Plant components of desert dust in Arizona and their significance for man, in *Geological Society of America Special Paper 186*, Desert Dust: Origin, Characteristics, and Effect on Man, T. L. Péwé (Ed.), 1981, pp. 191–206.

Lee, D. S., J. M. Edmond, and K. W. Bruland, Bismuth in the Atlantic and North Pacific: A natural analogue to plutonium and lead? *Earth Planet. Sci. Lett.*, *76*, 254–2262, 1986.

Levetin, E., Fungi, in H. A. Burge (Ed.), *Bioaerosols*, Lewis Publishers, Boca Raton, 1995, pp. 87–120.

Lin, X., and W. L. Chameides, CCN formation from DMS oxidation without SO_2 acting as an intermediate, *Geophys. Res. Lett.*, *20*, 579–582, 1993.

Liu, T., et al (Thirty-five coauthors), *Loess and the Environment*, China Ocean Press, Beijing, 251 pp, 1985.

MacIntyre, F., and J. W. Winchester, Phosphate ion enrichment in drops from breaking bubbles, *J. Phys. Chem.*, *73*, 2163–2169, 1969.

MacIntyre, F., Chemical fractionation and sea-surface microlayer processes, in E. D. Goldberg (Ed.), *The Sea*, Vol. 5, J. Wiley, New York, 1974, pp. 245–299.

Maenhaut, W., I. Salma, J. Cafmeyer, H. J. Annegarn, and M. O. Andreae, Regional atmospheric aerosol composition and sources in the eastern Transvaal, South Africa, and impact of biomass burning, *J. Geophys. Res.*, *101*, 23631–23650, 1996.

Mamane, Y., and J. Gottlieb, Nitrate formation on sea-salt and mineral particles—a single particle approach, *Atmos. Environ.*, *26A*, 1763–1769, 1992.

Mazurek, M., C. Laterza, L. Newman, P. Daum, W. R. Cofer III, J. S. Levine, and E. L. Winstead, Composition of carbonaceous smoke particles from prescribed burning of a Canadian boreal forest: Organic aerosol characterization by gas chromatography, in S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 2, MIT Press, Cambridge, MA, 1996, pp. 840–847.

McCormick, M. P., L. W. Thomason, and C. R. Trepte, Atmospheric effects of the Mt. Pinatubo eruption, *Nature*, *373*, 399–404, 1995.

Meng, Z., D. Dabdub, and J. H. Seinfeld, Chemical coupling between atmospheric ozone and particulate matter, *Science*, *277*, 116–119, 1997.

Mouri, H., K. Okada, and K. Shigehara, Variation of Mg, S, K and Ca Contents in individual sea-salt particles, *Tellus*, *45B*, 80–85, 1993.

Muilenberg, M. G., The outdoor aerosol, in H. A. Burge, (Ed.), *Bioaerosols*, Lewis Publishers, Boca Raton, 1995, pp. 163–204.

Murozumi, M., T. J. Chow, and C. Patterson, Chemical concentrations of pollutant lead aerosols, terrestrial dusts and sea salts in Greenland and Antarctic snow strata, *Geochim. Cosmochim. Acta*, *33*, 1247–1294, 1969.

Nriagu, J. O., A global assessment of natural sources of atmospheric trace metals, *Nature*, *338*, 47–49, 1989.

Nriagu, J. O., and J. M. Pacyna, Quantitative assessment of worldwide contamination of air, water and soils by trace metals, *Nature*, *333*, 134–139, 1988.

Okada, K., Y. Ishizaka, T. Masuzawa, and K. Isono, Chlorine deficiency in coastal aerosols, *J. Meteorol. Soc. Jpn.*, *56*, 501–507, 1978.

Pacyna, J. M., Atmospheric trace elements from natural and anthropogenic sources, in J. O. Nriagu and C. I. Davidson (Eds.), *Toxic Metals in the Atmosphere*, Wiley, New York, 1986, pp. 33–52.

Pankow, J. F., An absorption model of gas/particle partitioning of organic compounds in the atmosphere, *Atmos. Environ.*, *28*, 185–188, 1994.

Pankow, J. F., J. M. E. Storey, and H. Yamasaki, Effects of relative humidity on gas/particle partitioning of semivolatile organic compounds to urban particulate matter, *Environ. Sci. Technol.*, *27*, 2220–2226, 1993.

Parungo, F., Z. Li, X. Li, D. Yang, and J. Harris, Gobi dust storms and the Great Green Wall, *Geophys. Res. Lett.*, *21*, 999–1002, 1994.

Parungo, F. P., C. T. Nagamoto, and J. M. Harris, Temporal and spatial variations of marine aerosols over the Atlantic Ocean, *Atmos. Res.*, *20*, 23–37, 1986.

Patterson, C., Global pollution measured by lead in mid-ocean sediments, *Nature*, *326*, 244–245, 1987.

Peltzer, E. T., and R. B. Gagosian, Organic geochemistry of aerosols over the Pacific Ocean, in J. P. Riley, R. Chester, and R. A. Duce, (Eds.), *Chemical Oceanography*, Academic, London, 1989, pp. 281–338.

Perry, K. D., and P. V. Hobbs, Further evidence for particle nucleation in clean air adjacent to marine cumulus clouds, *J. Geophys. Res.*, *99*, 22803–22818, 1994.

Péwe, T. L., Desert dust: An overview, in *Geological Society of America Special Paper 186*, 1981, pp. 1–10.

Pimentel, D., C. Harvey, P. Resosudarmo, K. Sinclair, D. Kurz, M. McNair, S. Crist, L. Shpritz, L. Fitton, R. Saffouri, and R. Blair, Environmental and economic costs of soil erosion and conservation benefits, *Science 267*, 1117–1123, 1995.

Polian, G., and G. Lambert, Radon daughters and sulfur output from Erebus Volcano, Antarctica, *J. Volcanol. Geotherm. Res.*, *6*, 125–137, 1979.

Pósfai, M., J. R. Anderson, and P. R. Buseck, Compositional variations of sea-salt-mode aerosol particles from the North Atlantic, *J. Geophys. Res.*, *100*, 23063-23074, 1995.

Prospero, J. M. and R. T. Nees, Dust concentration of the equatorial North Atlantic: possible relationship to the Sahelian drought, *Science*, *196*, 1196–1198, 1977.

Pszenny, A. A. P., A. J. Castell, J. N. Galloway, and R. A. Duce, A study of the sulfur cycle in the Antartic marine boundary layer, *J. Geophys. Res.*, *94*, 9819–9830, 1989.

Pye, K., *Aeolian dust and dust deposits*, Academic Press, London, 1987, 334 pp.

Rahn, K., The chemical composition of the atmospheric aerosol, Tech. Report, Grad. Sch. of Oceanogr., Univ. of Rhode Island, Kingston, 265 pp. 1976.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-seasalt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Savoie, D. L., J. M. Propero, R. Arimoto, and R. A. Duce, Nonsea-salt sulfate and methanesulfonate at American Samoa, *J. Geophys. Res.*, *99*, 3587–3596, 1994.

Schlesinger, W. H., J. F. Reynolds, G. L. Cunningham, L. F. Huenneke, W. M. Jarrell, R. A. Virginia, and W. G. Whitford, Biological feedbacks in global desertification, *Science*, *247*, 1043–1048, 1990.

Schroeder, W. H., and P. Urone, Formation of nitrosyl chloride from sea particles in air, *Environ. Sci. Technol.*, *8*, 756–758, 1974.

Schütz, L., and K. A. Rahn, Trace-element concentrations in erodible soils, *Atmos. Environ.*, *16*, 171–176, 1982.

Sholkovitz, E. R., T. M. Church, and R. Arimoto, Rare earth element composition of rainwater, wet deposition and aerosols, *J. Geophys. Res.*, 98, 20587–20599, 1993.

Sievering, H., J. Boatman, E. Gorman, Y. Kim, L. Anderson, G. Ennis, M. Luria, and S. Pandis, Removal of sulphur from the marine boundary layer by ozone oxidation in sea-salt aerosols, *Nature*, *360*, 571–573, 1992.

Sievering, H., E. Gorman, T. Ley, A. Pszenny, M. Springer-Young, J. Boatman, Y. Kim, C. Nagamoto, and D. Wellman, Ozone oxidation of sulfur in sea-salt aerosol particles during the Azores Marine Aerosol and Gas Exchange experiment, *J. Geophys. Res.*, *100*, 23075–23081, 1995.

Simoneit, B. R. T., M. R. bin Abas, G. R. Cass, W. F. Rogge, M. A. Mazurek, L. J. Standley, and L. M. Hildermann, Natural organic compounds as tracers for biomass combustion in aerosols, in S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 1, MIT Press, Cambridge, MA, 1996, pp. 509–517.

Swap, R., M. Garstang, S. Greco, R. Talbot, and P. Kållberg, Saharan dust in the Amazon Basin, *Tellus*, *44B*, 133–149, 1992.

Symonds, R. B., W. I. Rose, and M. H. Reed, Contribution of Cl- and F-bearing gases to the atmosphere by volcanoes, *Nature*, *334*, 415–418, 1988.

Taylor, S. R. and S. M. McLennan, *The Continental Crust: Its Composition and Evolution*, 312 pp., Blackwells, Oxford, England, 1985.

Tegen, I., A. A. Lacis, and I. Fung, The influence on climate forcing of mineral aerosols from disturbed soils, *Nature*, *380*, 419–423, 1996.

Wallace, G. T., and R. A. Duce, Concentration of particulate trace metals and particulate organic carbon in marine surface waters by a bubble flotation mechanism, *Marine Chem.*, *2*, 157–181, 1975.

Weisel, C. P., R. A. Duce, J. L. Fashing, and R. W. Heaton, Estimates of the transport of trace elements from the ocean to the atmosphere, *J. Geophys. Res.*, *89*, 11607–11618, 1984.

Whitby, K. T., The physical characteristics of sulfur aerosols, *Atmos. Environ.*, *12*, 135–159, 1978.

Winchester, J. W., and M.-X. Wang, Acidic sulfur uptake by alkaline dust in the Asia-Pacific region, in L. Newman, W. Wang, and C. S. Kiang (Eds.), *Proceedings International Conference on Global and Regional Environmental Atmospheric Chemistry*, U.S. Department of Energy, Washington, DC, 1990, pp. 13–23.

Wu, P.-M., and K. Okada, Nature of coarse nitrate particles in the atmosphere—a single particle approach, *Atmos. Environ.*, *28*, 2053–2060, 1994.

Yamasaki, H., K. Kuwata, and Y. Kuge, Determination of vapor pressure of polycyclic aromatic hydrocarbons in the supercooled liquid phase and their adsorption on airborne particulate matter, *Nippon Kagaku Kaishi*, *8*, 1324–1329 (*Chem. Abstr.*, *101*, 156747p), 1984.

Zhou, G., and K. Tazaki, Seasonal variation of gypsum in aerosol and its effect on the acidity of wet precipitation on the Japan Sea side of Japan, *Atmos. Environ.*, *30*, 3301–3308, 1996.

Zoller, W. H., J. R. Parrington, and J. M. P. Kotra, Iridium enrichment in airborne particles from Kilauea Volcano: January 1983, *Science*, *222*, 1118–1121, 1983.

CHAPTER 12

AEROSOLS: FORMATION AND MICROPHYSICS IN THE TROPOSPHERE

JOHN H. SEINFELD

1 INTRODUCTION

Particles in the atmosphere arise from natural sources, such as wind-borne dust, sea spray, and volcanoes, and from anthropogenic activities, such as combustion of fuels. While an aerosol is technically defined as a suspension of fine solid or liquid particles in a gas, common usage refers to the aerosol as the particulate component only. Emitted directly as particles (primary aerosol) or formed in the atmosphere by gas-to-particle conversion processes (secondary aerosol), atmospheric aerosols are generally considered to be the particles that range in size from a few nanometers to tens of micrometers in diameter. Once airborne, particles can change their size and composition by condensation of vapor species or by evaporation, by coagulating with other particles, by chemical reaction, or by activation in the presence of water supersaturation to become fog and cloud droplets. Particles smaller than 1 μm diameter generally have atmospheric concentrations in the range from around tens to thousands per cubic centimeter; those exceeding 1 μm diameter are usually found at concentrations less than 1 per cm^3.

A significant fraction of the tropospheric aerosol is anthropogenic in origin. Chemical components of tropospheric aerosols include sulfate, ammonium, nitrate, sodium, chloride, trace metals, carbonaceous material, crustal elements, and water. The carbonaceous fraction consists of both elemental and organic carbon. Elemental carbon, also called black carbon, graphitic carbon, or soot, is emitted directly into the atmosphere, predominantly from combustion processes. Particulate organic

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

carbon is emitted directly by sources or can result from atmospheric condensation of low-volatility organic gases.

2 PARTICLE SIZE DISTRIBUTION

Size is the most important single characteristic of an aerosol particle. For a spherical particle, diameter (or radius) is the usual reported dimension. When a particle is not spherical, the size can be reported either in terms of a length scale characteristic of its silhouette or of a hypothetical sphere with equivalent dynamic properties, such as settling velocity in air. For example, the *aerodynamic diameter* of a particle represents the diameter of a unit density ($\rho_p = 1$ g/cm^3) sphere having the same terminal settling velocity as the particle sampled, whatever its size, shape, or density.

When particles, at total number concentration N (particles/cm^3), are measured and the number of particles dN having diameters between D_p and $D_p + dD_p$, where dD_p is a small increment of diameter, are counted, the particle size distribution $n(D_p)$ is defined as $n(D_p) = dN/dD_p$ (reciprocal micrometers per cubic centimeters), where D_p is usually measured in micrometers. The integral of the size distribution over all sizes is the total number concentration:

$$N = \int_0^\infty n(D_p)\, dD_p \tag{1}$$

The log-normal distribution is particularly useful for representing aerosol size distributions because it does not allow negative particle sizes,

$$n(D_p) = \frac{N}{\sqrt{2\pi}\ln\sigma_g}\exp\left[-\frac{(\ln D_p - \ln D_g)^2}{2\ln^2\ \sigma_g}\right] \tag{2}$$

where D_g is the geometric mean diameter and σ_g is the geometric standard deviation. These parameters can be determined from discrete particle count data by

$$\ln D_g = \frac{1}{N}\sum_i N_i \ln D_{pi} \tag{3}$$

$$\ln\sigma_g = \left[\frac{1}{N}\sum_i(\ln D_{pi} - \ln D_g)^2\right]^{1/2} \tag{4}$$

3 RESIDENCE TIMES OF PARTICLES IN THE TROPOSPHERE

Particles are eventually removed from the atmosphere by two mechanisms: deposition at Earth's surface, so-called dry deposition, and scavenging by droplets, so-called wet deposition (Seinfeld and Pandis, 1998). Because wet and dry deposition lead to relatively short residence times in the troposphere and because the geographical distribution of particle sources is highly nonuniform, tropospheric aerosols vary widely in concentration and composition over Earth. Whereas atmospheric trace gases have lifetimes ranging from less than a second to a century or more, the residence times of particles in the troposphere vary only from a few days to a few weeks.

The dry deposition flux of particles to the surface, F_d, is assumed to be proportional to the particle concentration at a reference height, C, i.e., $F_d = v_d C$, where the proportionality constant v_d, the deposition velocity, depends on the meteorological state of the atmosphere and the size of the particles. Three processes serve to deliver particles to Earth's surface: gravitational settling, turbulent transport, and Brownian diffusion. Although virtually any atmospheric flow is turbulent, a very thin laminar sublayer exists immediately adjacent to the surface. Turbulence brings particles down to the laminar sublayer, through which Brownian diffusion and settling govern transport. Small particles have a relatively large Brownian diffusivity, so move efficiently through the sublayer, whereas larger particles transfer primarily via inertia or settling. Those in between, in the size range of 0.1 to 1 μm diameter, are deposited about an order of magnitude slower than those at either the small or large extremes because none of the mechanisms is relatively effective in this intermediate size range.

Wet deposition involves the scavenging of particles by droplets and the subsequent removal by precipitation. Scavenging is necessary, but not sufficient, for wet deposition to occur since cloud or rain drops can evaporate, and if this occurs, the scavenged particle is returned to the air mass. As opposed to dry deposition, which operates only at Earth's surface, wet deposition serves to remove particles from the entire air mass. The rate of particle collection by falling drops is proportional to the number of drops, their settling velocity, their cross-sectional area, and a collection efficiency. The efficiency with which a particle is collected by a falling drop depends on the mechanics of particle motion in the vicinity of the drop. As with dry deposition, there is a minimum in the total collection efficiency in the 0.1 to 1 μm size range—small particles diffuse to the drop surface, larger ones collide with it, while in between neither process is very efficient.

Particles that become activated to grow to fog or cloud droplets are termed cloud condensation nuclei (CCN). At a given mass of water-soluble material in the particle, there is a critical value of the ambient water supersaturation, above which the particle undergoes an unstable process of spontaneous water accretion, leading to a cloud droplet (Seinfeld and Pandis, 1998). The critical water supersaturation for activation results from a combination of the curvature increase in and the solute concentration lowering of the water vapor pressure over a droplet. The number of particles that can act as CCN thus depends on the water supersaturation. For marine stratiform clouds, for example, supersaturations are in the range of 0.1 to 0.5%, which corresponds to a minimum CCN particle diameter of 0.05 to 0.14 μm. CCN number concentrations vary from fewer than $100/\text{cm}^3$ in remote marine regions to a few thousand per cubic centimeter in polluted urban areas. Once activated, fog and cloud droplets grow to sizes exceeding 10 μm diameter. Particles that are not activated to form droplets may remain as airborne aerosol or be removed by falling drops.

Aerosol lifetimes in the atmosphere depend primarily on the size of the particle and the height in the atmosphere at which the particle resides. The residence time τ can be viewed as an exponential half-life, the time required for a population of particles of a given size to decay to $1/e$ of its initial concentration.

An empirical expression for atmospheric particle residence time as a function of particle size and altitude that is useful for estimates is (Jaenicke, 1988)

$$\frac{1}{\tau} = \frac{1}{K}\left[\left(\frac{D_p}{D_{\text{max}}}\right)^2 + \left(\frac{D_p}{D_{\text{max}}}\right)^{-2}\right] + \frac{1}{\tau_{\text{wet}}} \tag{5}$$

where $K = 1.28 \times 10^8$ s (constant), $D_{\text{max}} = 0.6$ μm (the diameter of particle with maximum residence time), and τ_{wet} is the lifetime for removal of particles by wet deposition. The first term on the right-hand side of Eq. (5) represents dry removal at Earth's surface; τ_{wet} depends mainly on the altitude in the atmosphere. Roughly three altitude regions can be distinguished by the frequency with which precipitation scavenging occurs:

Height ≤ 1.5 km (lower troposphere)	$\tau_{\text{wet}} \approx 6.9 \times 10^5$ s (8 days)
Middle troposphere to tropopause	$\tau_{\text{wet}} \approx 1.8 \times 10^6$ s (3 weeks)
Tropopause and above	$\tau_{\text{wet}} \approx 1.7 \times 10^7$ s (200 days)

Figure 1 shows atmospheric particle residence time τ as a function of particle radius, $D_p/2$. Dry removal predominates for particle radii either much smaller or larger than 0.3 μm; in the region around 0.3 μm wet scavenging is the most effective removal mechanism.

4 TROPOSPHERIC AEROSOLS

The relatively short residence time of aerosols in the troposphere, usually less than a couple of weeks, results in significant spatial variations in particle concentration, size, and composition. In an effort to categorize tropospheric aerosols, eight approximate classes can be identified: marine, remote continental, nonurban continental, urban, desert, polar, biomass burning, and background (free troposphere) (Heintzenberg, 1989; Fitzgerald, 1991; Jaenicke, 1993).

Marine Aerosol

Particles over the remote oceans are largely of marine origin (Savoie and Prospero, 1989). Marine atmosphere particle concentrations are normally in the range of 100 to 300/cm^3 (Fitzgerald, 1991). Typically the coarse particle mode (diamcters cxcccding 1 μm), comprising 95% of the total mass but only 5 to 10% of the particle number, results from the evaporation of sea spray produced by bursting bubbles or wind-induced wave breaking (Blanchard and Woodcock, 1957; Monahan et al.,

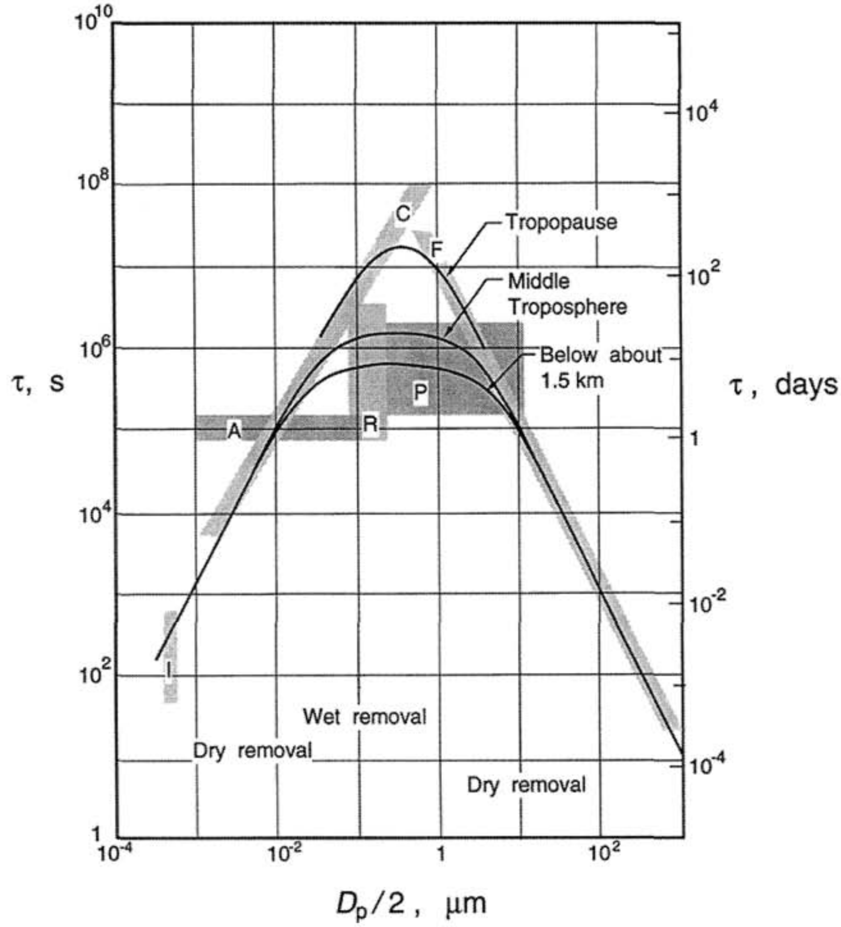


Figure 1 Residence times of tropospheric aerosols as a function of particle radius $D_p/2$. I, small atmospheric ions; A, so-called Aitken particles, radii 0.001 to 0.1 μm, residence time estimated from geographical distributions; C, residence time calculated based on Brownian coagulation of particles; R, radioactivity; P, precipitation removal; F, sedimentation. The three curves shown correspond to the three altitude levels indicated based on Eq. (5). Adapted from Jaenicke (1988).

1983). Typical sea salt aerosol concentrations in the marine boundary layer (MBL) are thought to be around 20 to $30/\text{cm}^3$ (Blanchard and Cipriano, 1987; O'Dowd and Smith, 1993).

Remote Continental Aerosol

Aerosol number concentrations in remote continental regions average around $10^4/\text{cm}^3$, and mass concentrations average about 20 $\mu\text{g}/\text{m}^3$ (Bashurova et al., 1992; Koutsenogii et al., 1993; Koutsenogii and Jaenicke, 1994). A typical remote continental aerosol number distribution has three modes centered at diameters about 0.02, 0.12, and 1.8 μm (Bashurova et al., 1992; Koutsenogii et al., 1993; Koutsenogii and Jaenicke, 1994).

Urban Aerosol

Urban aerosols are strongly anthropogenic in origin, with a definite combustion signature. Number concentrations usually exceed $10^5/\text{cm}^3$, and size distributions typically exhibit three modes: named nuclei, accumulation, and coarse. The constituents of urban aerosol comprise the full spectrum of compounds possible in the atmospheric aerosol: sulfate, nitrate, ammonium, elemental and organic carbon (EC and OC), and crustal compounds (silicon, aluminum, calcium, and iron oxides). These aerosols result from primary emissions (EC, OC, soil material) and gas-to-particle transformation of the oxides of nitrogen, hydrocarbons, ammonia, and SO_2.

Desert Aerosol

Large amounts of dust are emitted to the atmosphere from deserts, especially during high wind periods. Most of these particles are coarse ($D_p > 1$ μm), and many are deposited close to their source; some fraction of the smaller particles can be transported over large distances (Prospero, 1990). For example, dust from the Sahara is regularly detected on Barbados Island across the Atlantic Ocean (Andreae, 1995). The chemical composition of desert aerosol reflects its soil source and is often rich in calcium compounds and other alkaline elements.

Polar Aerosol

Air masses generally remain over polar ice for extended periods of time. Any large particles that are present have sufficient time to deposit out, leaving a monodisperse aerosol with a mean size of about 0.15 μm and a number concentration in the range of 15 to 150 particles/cm^3 (Browell et al., 1992). Suitable meteorological conditions leading to the transport of anthropogenic aerosol to high latitudes in winter and early spring produce so-called arctic haze (Barrie, 1986).

Background Aerosol

The aerosol in the mid and upper troposphere, the so-called free troposphere, is often termed background aerosol. It is well-aged aerosol with a composition and size distribution reflecting the simultaneous effects of gas-to-particle conversion, long-range transport, and removal processes. The number concentration of background aerosol is in the range of $300/\text{cm}^3$ (Raes et al., 2000), and its size distribution is nearly monodisperse with peak diameters in the 0.2 to 0.5 μm range (Leaitch and Isaac, 1991). Regions with the lowest mass concentrations generally exhibit the highest number concentrations (Clarke, 1993), suggesting that nucleation may be a major source of particle number. Volatility measurements of the free tropospheric aerosol suggest a composition dominated by sulfates (Clarke, 1993; Hofmann, 1993)

Biomass Burning Aerosol

Remote biomass burning (forest and savanna fires, agricultural burning, etc.) is a major source of both primary (ash, elemental carbon) and secondary (organic carbon, sulfate, nitrate, and ammonium) aerosol. The chemical composition of the aerosol produced depends on the characteristics of the combustion: hot, flaming fires (e.g., savanna fires) emit mainly EC aerosol, while smoldering fires emit mainly organic particles (Andreae, 1995). The number concentrations of aerosols produced by biomass fires are of the order of tens of thousands of particles per cubic centimeter close to the source and less than $1000/\text{cm}^3$ after a few days of transport. Mass median diameters in fresh fire plumes are typically in the range of 0.1 to 0.3 μm and evolve toward values in the range of 0.2 to 0.4 μm during the first few hours after emission.

Nonurban Continental Aerosol

Nonurban continental aerosol is often acidic as a result of anthropogenic sulfate or nitrate. Typical aerosol number concentrations of nonurban continental aerosol are in the range of $10^3/\text{cm}^3$, with mass concentrations around 30 $\mu\text{g/m}^3$ (Anderson et al., 1993). The aerosol mass distribution usually exhibits a trimodal structure similar to that of urban aerosol.

Figure 2 shows approximate atmospheric aerosol number concentration N and volume concentration V as a function of altitude z. Because of the large variations in number and volume concentrations in the lower troposphere (remote continental, rural continental, urban, marine, polar), the vertical profiles fan out at low altitudes

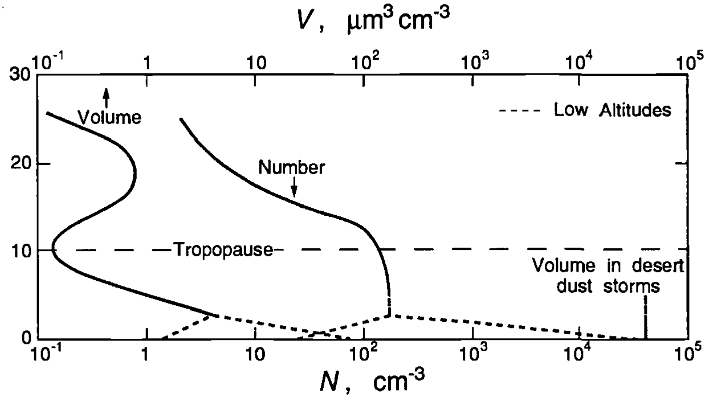


Figure 2 Atmospheric aerosol number N and volume V concentrations as a function of altitude z. Adapted from Jaenicke (1988).

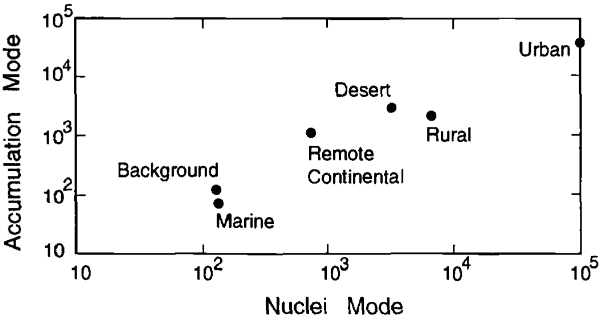


Figure 3 Typical aerosol number concentrations in accumulation and nuclei modes for six classes of global aerosols. Clement and Ford (1997), based on number concentrations reported by Jaenicke (1993). Reprinted from *Journal of Aerosol Science*, Vol. 28, No. 1, C. F. Clement and I. J. Ford, Properties and modelling of global aerosols, 5743-4, Copyright 1997, with permission from Elsevier Science.

(dashed lines). Aerosol number concentration decreases continuously with increasing altitude, whereas volume concentration decreases up to the tropopause (about 10 km) and then increases in the stratospheric aerosol layer, reaching maximum between 15 and 20 km altitude.

Atmospheric aerosol number concentrations can be naturally divided into three groups: nuclei mode particles with diameters ≤ 0.01 μm, accumulation mode particles with diameters 0.01 to 1 μm, and coarse particles with diameters exceeding 1 μm. Figure 3 shows the total number concentrations of the nuclei and accumulation modes for urban, rural, desert, remote continental, marine, and background aerosols. (There is no obvious nucleation mode for the polar aerosol.) Number concentrations of the two modes are seen to be equal over a range of 3 orders of magnitude in different parts of the atmosphere.

5 AEROSOL MICROPHYSICS

The atmosphere subjects aerosol particles to an array of transport and transformation processes that alter their size, number, and composition. Advective and turbulent transport of particles is nearly identical to that of the interstitial gas. Transformation processes include condensation and evaporation, which result from diffusion of vapors between the particle and the interstitial gas, homogeneous nucleation to produce new particles from supersaturated vapors, coagulation, which combines two particles into one by collision and sticking, and chemical reactions occurring in individual particles (Seinfeld and Pandis, 1998). That a major portion of atmospheric aerosol mass is secondary in nature is indicative of the importance of gas-to-particle conversion.

The aqueous phase of atmospheric aerosols contains primarily strong electrolytes such as sodium chloride, nitric and sulfuric acids, and ammonium. At relative humidities much below saturation, the vast majority of water in the atmosphere is in the vapor phase, and therefore any liquid water associated with aerosol particles is too small to affect the ambient relative humidity. For relative humidities below saturation, water is in equilibrium between the vapor and aqueous phases because the characteristic time for water equilibration is relatively short compared to all other processes taking place. Other volatile aerosol species may or may not be in equilibrium depending on their equilibration characteristic time (Seinfeld and Pandis, 1998).

Much nucleation research relevant to the atmosphere has been focused, via measurement and theory, on the binary nucleation of sulfuric acid and water (Seinfeld and Pandis, 1998). Although a classical theory of binary nucleation of sulfuric acid and water exists (Jaecker-Voirol and Mirabel, 1989), substantial uncertainty still remains as to how accurately this classical theory represents the actual nucleation process. Measured nucleation rates can differ from theoretically predicted values by several orders of magnitude. From the point of view of atmospheric applications, significant nucleation rates can be defined as those exceeding 1 nucleus/cm^3 s.

Coagulation is the process whereby two particles collide and stick to form a single particle. Atmospheric processes that may lead to particle collisions include Brownian motion, turbulent shear, and differential settling. The latter two can be shown to be much less effective in this regard than Brownian motion (Wexler et al., 1994). In atmospheric aerosol dynamics, coagulation does not play a significant role unless number concentrations are relatively high and/or residence times are relatively long.

6 CONCLUSION

Despite significant progress in our understanding of the global aerosol system over the past two decades, our knowledge of the sources and dynamics of atmospheric aerosols remains limited. Difficulties associated with measurement of the aerosol size/composition distribution, combined with the significant spatial and temporal variability of tropospheric aerosol, have resulted in only scattered knowledge of its global distribution. Aerosols in remote locations and in the middle and upper troposphere have received relatively little attention, and their size/composition distribution remains largely unexplored. Most of the existing measurements are ground based, and, consequently, there is a general lack of information on the vertical aerosol distribution. Moreover, few measurements of the chemical composition of the smallest atmospheric particles, e.g., smaller than 50 nm diameter, are available.

REFERENCES

Anderson, B. E., G. L. Gregory, J. D. W. Barrick, J. E. Collins, G. W. Sachse, D. Bagwell, M. C. Shipham, J. D. Bradshaw, and S. T. Sandholm, The impact of U.S. continental outflow on ozone and aerosol distributions over the westem Atlantic, *J. Geophys. Res.*, *98*, 23477–23489, 1993.

Andreae, M. O., Climate effects of changing atmospheric aerosol levels, in A. Henderson-Sellers (Ed.), *World Survey of Climatology*, Elsevier, Amsterdam, 1995, pp. 347–398.

Barrie, L. A., Arctic air pollution: An overview of current knowledge, *Atmos. Environ.*, *29*, 643–663, 1986.

Bashurova, V. S., V. Dreiling, T. V. Hodger, R. Jaenicke, K. P. Koutsenogii, P. K. Koutsenogii, M. Kraemer, V. I. Makarov, V. A. Obolkin, V. L. Potjomkin, and A. Y. Pusep, Measurements of atmospheric condensation nuclei size distributions in Siberia, *J. Aerosol Sci.*, *23*, 191–199, 1992.

Blanchard, D. C., and R. J. Cipriano, Biological regulation of climate, *Nature*, *330*, 526, 1987.

Blanchard, D. C., and A. H. Woodcock, Bubble formation and modification in the sea and its meteorological significance, *Tellus*, *9*, 145–158, 1957.

Browell, E. V., C. F. Butler, S. A. Kooi, M. A. Fenn, R. C. Harriss, and G. L. Gregory, Large–scale variability of ozone and aerosols in the summertime Arctic and sub-Arctic atmosphere, *J. Geophys Res.*, *97*, 16433–16450, 1992.

Clarke, A. D., Atmospheric nuclei in the remote free troposphere, *J. Atmos. Chem.*, *14*, 479–488, 1993.

Clement, C. F., and I. J. Ford, Properties and modeling of global aerosols, *J. Aerosol Sci.*, *28*, (Suppl. 1), 5743–5744, 1997.

Fitzgerald, J. W., Marine aerosols: A review, *Atmos. Environ.*, *25A*, 533–545, 1991.

Heintzenberg, J., Fine particles in the troposphere, a review, *Tellus*, *41B*, 149–160, 1989.

Hofmann, D. J., Twenty years of balloon-borne tropospheric aerosol measurements at Laramie, Wyoming, *J. Geophys. Res.*, *98*, 12753–12766, 1993.

Jaecker-Voirol, A., and P. Mirabel, Heteromolecular nucleation in the sulfuric acid-water system, *Atmos. Environ.*, *23*, 2053–2057, 1989.

Jaenicke, R., Aerosol physics and chemistry, in G. Fischer (Ed.), *Meteorology, Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, Vol. 4, Springer-Verlag, Berlin, 1988, pp. 391–456.

Jaenicke, R., Tropospheric aerosols, in P. V. Hobbs (Ed.), *Aerosol-Cloud-Climate Interactions*, Academic, New York, 1993, pp. 1–31.

Koutsenogii, P. K., N. S. Bufetov, and V. I. Drosdova, Ion composition of atmospheric aerosol near Lake Baikal, *Atmos. Environ.*, *27*, 1629–1633, 1993.

Koutsenogii, P. K., and R. Jaenicke, Number concentration and size distribution of atmospheric aerosol in Siberia, *J. Aerosol Sci.*, *25*, 377–383, 1994.

Leaitch, W. R., and G. A. Isaac, Tropospheric aerosol size distributions from 1982 to 1988 over eastern North America, *Atmos. Environ.*, *25A*, 601–619, 1991.

Monahan, E. C., C. W. Fairall, K. L. Davidson, and P. Jones-Boyle, Observed interrelationships amongst 10m-elevation winds, oceanic whitecaps, and marine aerosols, *Q. J. R. Metereol. Soc.*, *109*, 379–392, 1983.

O'Dowd, C. D., and M. H. Smith, Physicochemical properties of aerosols over the northeast Atlantic: Evidence for wind speed related submicron sea-salt production, *J. Geophys. Res.*, *98*, 1137–1149, 1993.

Prospero, J., Mineral-aerosol transport to the North Atlantic and North Pacific: the impact of African and Asian sources, in A. H. Knap (Ed.), *The Long-Range Atmospheric Transport of*

Natural and Contaminant Substances, NATO ASI series, Kluwer Academic, New York, 1990, pp. 59–82.

Raes, F., R. Van Dingenen, E. Vignati, J. Wilson, J. P. Putaud, J. H. Seinfeld, and P. Adams, Formation and cycling of aerosols in the global troposphere, *Atmos. Environ.*, *34*, 4215–4240, 2000.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-sea-salt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, Wiley, New York, 1998.

Wexler, A. S., F. W. Lurmann, and J. H. Seinfeld, Modeling urban and regional aerosols. I. Model development, *Atmos. Environ.*, *28*, 531–546, 1994.

CHAPTER 13

PHOTOCHEMICAL SMOG: OZONE AND ITS PRECURSORS

SANFORD SILLMAN

1 INTRODUCTION

Photochemical smog refers to a number of species that are chemically produced in highly polluted environments in processes driven by sunlight. The most prominent of these species is ozone (O_3), which reaches levels that violate government health standards in urban areas throughout the world. Other components of photochemical smog include peroxyacetyl nitrate (PAN, $CH_3CO_3NO_2$) and nitric and sulfuric acid (HNO_3, H_2SO_4). The latter is associated with the formation of acid aerosols, which have serious impacts on both human health and visibility. Photochemical smog typically forms during conditions characterized by high sunlight (though often with haze and reduced visibility), light winds, and warm temperatures. This chapter focuses on ozone and its precursors.

Episodes with high ozone were first observed in Los Angeles in the 1950s (Haagen-Smit and Fox, 1954) and have generally been found in cities with high automobile traffic. This type of photochemical smog should be distinguished from the type of smog driven by primary emissions (primarily of coal-based SO_2, NO_2, CO, and soot), which characterized the city of London during the early 1900s (Brimblecombe, 1987) and Beijing today. In primary smog, high concentrations are associated with patterns of atmospheric circulation that "trap" the emitted pollutants in an atmospheric layer close to emission sources. The most severe events tend to occur in fall or winter, when atmospheric vertical mixing at ground level is minimal, and may coincide with fog (hence the origin of the word *smog* for the combination of smoke and fog). By contrast, photochemical smog can only occur in meteorological conditions that favor photochemical activity (i.e., high sunlight, warm temperatures)

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

and not necessarily during conditions with restricted meteorological dispersion. The most severe events have occurred in large urban areas with warm, dry climates (Los Angeles, Mexico City, Athens). However severe photochemical smog has been observed in virtually all major cities of North America and Europe and more recently in developing nations. Although the most severe episodes have occurred in locations with high automobile traffic, elevated O_3 has also been associated with coal-fired power plants (Miller et al., 1978; White et al., 1983; Gillani and Pleim, 1996; Ryerson et al., 2001). In the eastern United States and in Europe elevated O_3 occurs in regionwide events and is characterized by transport over distances of 500 km or more. Elevated O_3 has also been found in association with biomass burning in the tropics (see Chapter 14).

The relation between ozone formation and precursor emissions has been the subject of much uncertainty and controversy. Ozone is formed from two general classes of precursors: hydrocarbons (including oxygenated organic species)* and nitrogen oxides ($NO + NO_2$, or NO_x). The chemistry of ozone formation typically falls into one of two recognizable patterns: a NO_x-limited regime in which the rate of formation increases with NO_x and is largely independent of hydrocarbon concentrations and a hydrocarbon-limited (or light-limited) regime in which the rate of formation increases with hydrocarbons and decreases with increasing NO_x. An analogous split into NO_x-limited and light-limited regimes also occurs in the remote troposphere. The split into NO_x-limited and HC-limited regimes has generated a debate on policy, especially in the United States, concerning the best way to reduce urban ozone. Because this represents a major uncertainty associated with ozone formation, much of this chapter will address the complex relation between ozone, NO_x, and HC.

2 GENERAL FEATURES OF PHOTOCHEMICAL SMOG

Diurnal and Seasonal Cycle

Ozone and other secondary reaction products show a pronounced diurnal cycle with peak concentrations typically occurring in late afternoon. The diurnal cycle of O_3 shows a sharp contrast with the diurnal cycle of primary species, including NO_x, HC, and CO (see Fig. 1). The primary species typically have peak concentrations in early morning and much lower concentrations during the daytime as concentrations are diluted through the process convection-driven vertical mixing. Because production of O_3 requires sunlight, peak concentrations often occur at the time of maximum vertical mixing and often coincide with diurnal minima in precursor concentrations. The diurnal cycle of O_3 is also influenced by nighttime removal of O_3 near the ground (through surface deposition or through reaction with directly emitted NO). Especially in urban areas, O_3 concentrations near the surface are often very low at night. The characteristic increase in O_3 during the morning hours (6 to 10 A.M.) is usually driven by convective mixing that breaks up the nighttime inversion at the

*The family consisting of hydrocarbons and oxygenates such as formaldehyde, HCHO, is properly referred to by acronyms such as volatile organic compounds (VOC) or reactive organic gases (ROG). In this chapter they will be referred to collectively as hydrocarbons (HC).

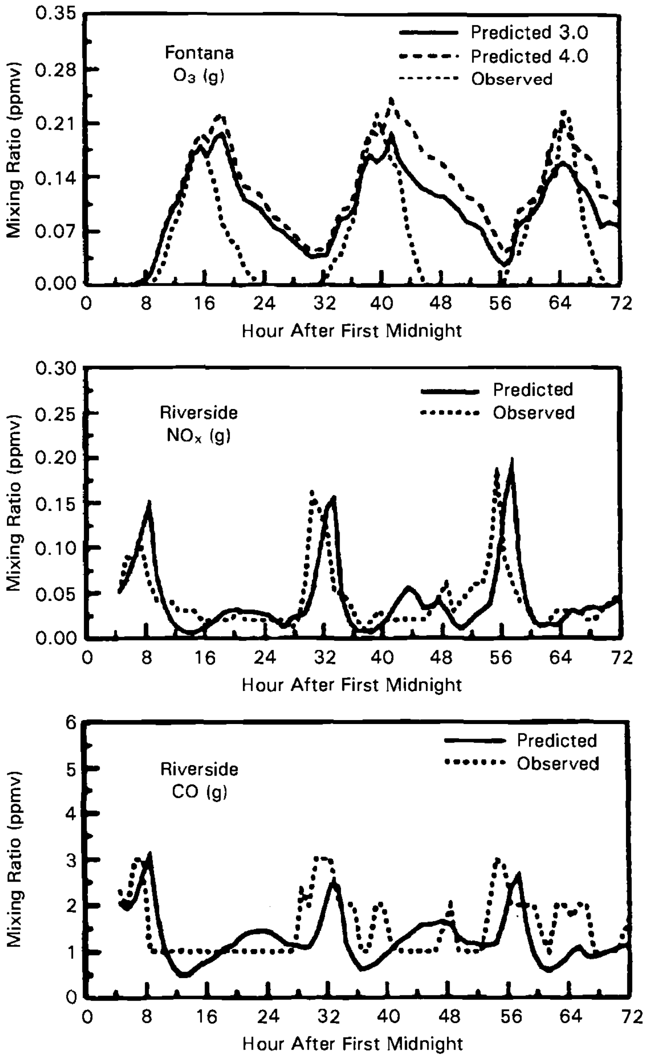


Figure 1 Time series for O_3, CO, NO, and NO_2 (in parts per million, ppmv) vs. hour at Riverside, CA, August 26–28, 1988. Dashed lines represent measurements, and solid lines represent model predictions. From Jacobson et al. (1996).

surface and mixes down air (from 100 to 1000 m above the surface) with higher O_3. The subsequent rise in O_3 after 10 A.M. is often associated with photochemical production.

The city of Los Angeles is especially susceptible to high-ozone events because it frequently sees a combination of high sunlight, warm temperatures, and a low-level thermal inversion (typically 500 to 1000 m above the surface) during the daytime. In most other cities the conditions that favor ozone formation (sunlight and warm temperatures) coincide with vigorous vertical mixing (up to 2000 m), which has a moderating effect on ozone concentrations. Thermal inversions, which trap pollutants near the ground, are more commonly associated with cold temperatures and often with fog. These conditions would produce high concentrations of primary pollutants but not ozone.

Unless stated otherwise, the discussion of ozone concentrations below refers to the diurnal peak or near-peak concentrations that occur during the afternoon.

Concentrations and Regional Transport

The global background concentration of O_3 near the surface is 20 to 40 ppb parts per billion (ppb), although these values probably represent an increase in comparison with preindustrial concentrations. There have also been episodes in which high concentrations of O_3 originating in the upper troposphere, ultimately of stratospheric origin, may have been transported to the surface.

During air pollution events in the United States and Europe, peak O_3 frequently exceeds 125 ppb, which is the current government health standard in the United States.* In Los Angeles during the 1970s and 1980s air quality violations (i.e., $O_3 < 125$ ppb) were reported on approximately 180 days per year. In the 1990s the frequency of violation has been lowered to 90 days per year. Most other major cities in the United States record violations on 5 to 10 days per year. In Europe, ozone exceeds 125 ppb on just a few days per year, while in Mexico City at present, ozone exceeds 125 ppb on 200 days per year. Concentrations above 200 ppb are found only during the most severe events, and concentrations as high as 490 ppb have been observed in Los Angeles and in Mexico City.

In addition, 80 to 100 ppb ozone is frequently observed in rural areas of the eastern United States and Europe during regional events. In these events, air with ozone concentrations above 80 ppb frequently extends over a 1000 × 1000 km region and extends vertically to 1000 to 2000 m above the surface (e.g., Clarke and Ching, 1983). These events are often associated with stagnant high pressure systems in which air may be trapped under a subsidence inversion at ~2000 m. An example in eastern North America is shown in Figure 2. In the example, ozone above 120 ppb

*As of 1997, a 1-h average concentration in excess of 125 ppb constitutes an air quality violation in the U.S. metropolitan areas that record violations of the 1-h standard on more than 3 days over a 3-year period are held in violation of clean air laws and are asked to submit a plan for pollution reduction. Since 1977, most U.S. cities have been continually in violation. It was recently proposed that the 1-h air quality standard be replaced with a standard based on 8-hour average concentrations, in which an 8-h average concentration in excess of 85 ppb is counted as an air quality violation.

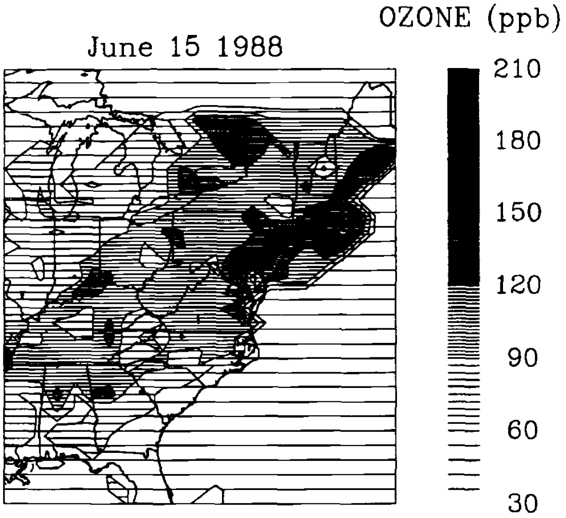


Figure 2 Peak ozone concentrations in the eastern United States during a severe air pollution event (June 15, 1988) based on surface observations at 350 EPA monitoring sites. The shadings represent values of 30 to 60 ppb (lightest shading) to 180 to 210 ppb (darkest shading) with 30-ppb intervals in between. Values reported for Canada and the Atlantic Ocean are inaccurate since no observations were available for these locations. First printed in Sillman (1993).

was found in many metropolitan areas, especially in the corridor extending from Washington to New York and Boston. However, ozone above 90 ppb covered a much larger area extending from Kentucky to Maine. Although ozone concentrations above 120 ppb were generally associated with plumes from specific urban areas (or from coal-fired power plants), concentrations of 90 to 100 ppb were found at rural sites throughout the region. In addition, unusually high ozone (> 200 ppb) was found in Acadia National Park in Maine. The high ozone in Maine is most likely due to transport from Boston (300 km distant) and the New York area (700 km distant).

Environmental and Health Impacts

The impact of ozone and acid aerosols on human health has been the subject of intense scrutiny. Ozone and aerosols have been associated with a variety of lung ailments. Short-term symptoms (including lung inflammation, asthmatic responses, and measured impairment of lung functions) have been found in experiments in response to ozone concentrations as low as 120 ppb. High-ozone events have been correlated with increased admissions to hospitals for respiratory diseases and with increased mortality rates. For a summary of findings, see Lippman (1993) and Bascomb et al. (1996).

In addition, ozone concentrations of 80 ppb have been found to cause damage both to forests and to agricultural crops. Crop damage from ozone in the United States has been estimated to cause monetary losses of \$1 to 2 billion per year. For a summary of findings, see U.S. Congress (1989) and National Research Council (NRC, 1991).

Dependence on Temperature

As stated above, ozone in polluted regions shows a strong dependence on temperature. This dependence on temperature is important as a basis for understanding variations in ozone concentrations from year to year or between cities. As shown in Figure 3, elevated ozone is always associated with temperatures in excess of 20°C and is often with temperatures above 30°C. In the eastern United States and Europe, year-to-year variations in ozone concentrations are often the result of variations in temperature and cloud cover, rather than in changes in emission of pollutants.

The reason for the dependence on temperature is due largely to the chemistry of ozone formation. Cardelino and Chameides, (1990) and Sillman and Samson (1995) found that the temperature dependence was associated with the temperature-dependent decomposition rate of PAN. PAN becomes longer lived at lower temperatures, and formation of PAN results in the removal of NO_x, hydrocarbons, and odd hydrogen radicals (described below), all of which suppress ozone formation. PAN, also a component of photochemical smog, tends to reach maximum values at intermediate temperatures (5 to 10°C). Jacob et al. (1993) proposed that ozone correlates with temperature partly because the meteorological conditions that favor ozone formation (high solar radiation and light winds) tend to be associated with warm temperatures. In addition, the emission rate of biogenic hydrocarbons (a major ozone precursor, discussed below) increase sharply with increasing temperature. Ozone is affected by temperature only in polluted regions. Temperature apparently has little impact on ozone production at the global scale (Sillman and Samson, 1995).

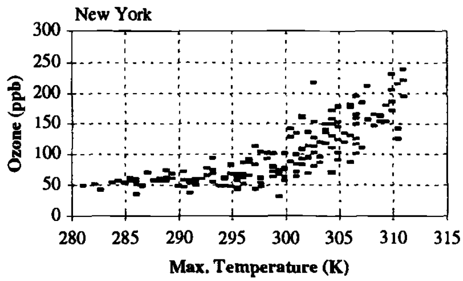


Figure 3 Diurnal peak O_3 (ppb) vs. maximum surface temperature observed in the New York–New Jersey–Connecticut metropolitan area for April 1 through September 30, 1988. From Sillman and Samson, 1995.

Role of Biogenic Hydrocarbons

The main precursors of photochemical smog, NO_x and hydrocarbons, are emitted into the atmosphere by a variety of human activities—transport (chiefly automobiles), coal-fired industry (especially electric power plants), and biomass burning. However, significant amounts of hydrocarbons occur naturally and are emitted by vegetation, primarily from trees. The most important of these biogenic hydrocarbons are isoprenes (C_5H_8), emitted by oaks and other deciduous trees, and α- and β-pinenes ($C_{10}H_{16}$), which are emitted from conifers. These species react chemically in the same way as anthropogenic hydrocarbons and can function as precursors to photochemical smog. In the United States it is estimated that emission of biogenic hydrocarbons equals or exceeds emission of anthropogenic hydrocarbons (Geron et al., 1994). Even in urban areas biogenic hydrocarbons can account for a significant fraction of total hydrocarbon emissions and can have a large impact on the formation of smog (Chameides et al., 1988). The impact of isoprene is especially large because it reacts rapidly, with a chemical lifetime of one hour or less. Consequently even small amounts of isoprene (0.5 ppb) can have a large impact on ozone.

It should be emphasized that naturally occurring hydrocarbons will not lead to the formation of photochemical smog in the absence of human activities because smog formation requires NO_x in addition to hydrocarbons. Although some NO_x is emitted naturally through biological activity, naturally occurring NO_x emissions are too small to allow significant formation of O_3 and other components of smog. Biogenic NO_x is estimated to be 7% of total NO_x emissions in the United States (Williams et al., 1992) and most of this is associated with agriculture (especially with the use of nitrate fertilizer).

Ozone Production Efficiency

The ozone production efficiency represents the rate of production of ozone divided by the loss rate for NO_x $[P(O_3)/L(NO_x)]$. Liu et al. (1987) first introduced the concept of ozone production efficiency and used it as a basis for estimating global production of ozone as a function of estimated NO_x emissions. A central feature of the ozone production efficiency is the tendency toward lower values in more polluted environments. Recent estimates suggest that ozone production efficiency is 10 to 30 in the remote troposphere but just 3 to 5 in urban areas.

3 RELATION BETWEEN OZONE, NO_x, AND HYDROCARBONS

The relation between ozone, NO_x and hydrocarbons can be illustrated by an isopleth plot (Fig. 4), which shows instantaneous rates of ozone production as a function of NO_x and hydrocarbon concentrations. It can be seen that ozone production is a highly nonlinear process, especially with regard to NO_x. Ozone production as a function of NO_x shows well-defined local maxima, usually at a specific HC/NO_x ratio. This region of maximum ozone (the "ridge line") can be thought of as a divide

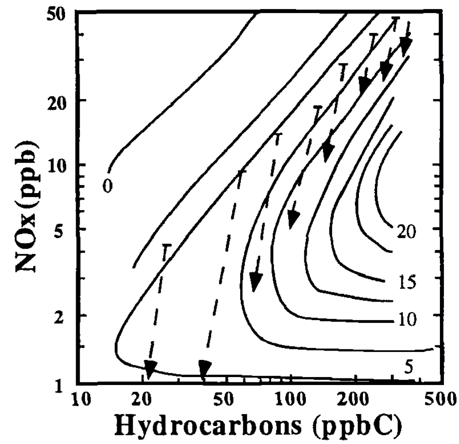


Figure 4 Isopleths giving net rate of ozone production (ppb per hour, daytime average, solid line) as a function of ROG (ppbC) and NO_x (ppb). The dashed lines and arrows show the calculated evolution of ROG and NO_x concentrations in a series of air parcels over an 8-h period (9 A.M. to 5 P.M.), each with initial $ROG/NO_x = 6$ and speciation typical of urban centers in the United States, based on calculations shown in Milford et al. (1994).

between two regimes with different photochemical behavior. Above the ridge line (with low HC/NO_x ratios), ozone production rates increase with increasing HC but decrease with increasing NO_x (hydrocarbon-limited regime). Below the ridge line (with high HC/NO_x ratios) ozone production rates increase with increasing NO_x and will be largely unaffected by changes in hydrocarbons (NO_x-limited regime). The existence of these two regimes has an enormous impact on public policy because it affects the choice of control strategies for reducing high ozone levels. If ozone production is dominated by NO_x-limited chemistry, then reductions in NO_x emissions would be necessary to reduce ozone concentrations. If production is dominated by HC-limited chemistry, then reductions in hydrocarbons would be needed. There is also a complex relation between NO_x, HC, and particulates, which also affects policy choices (Meng et al., 1997).

An important feature of HC–NO_x chemistry is the tendency for polluted air to evolve toward the NO_x-limited regime as the air mass ages and moves downwind. Air is most likely to show HC-limited chemistry when it is close to emission sources, especially in large cities. As the air mass ages, the HC/NO_x ratio increases and the chemistry shifts to the NO_x-limited regime. This shift from HC-limited to NO_x-limited chemistry as polluted air ages is illustrated by the air parcel trajectories in Figure 4.

In terms of photochemical mechanisms, the split into NO_x-limited and HC-limited regimes is closely related to the chemistry of odd hydrogen, defined as the sum of OH, HO_2, and RO_2 radicals (where R represents a hydrocarbon chain). The sequence of reactions that lead to photochemical smog (including production of both ozone and acid aerosols) is usually initiated by reactions that involve the OH radical, and availability of OH (which is produced from reactions involving sunlight, O_3 and H_2O) often controls the rate of ozone production. The split into NO_x- and HC-limited regimes is determined by the loss mechanism for odd hydrogen (Sillman, 1995; Sillman et al., 1990; Kleinman 1991, 1994; summarized in Sillman, 1999). The reaction sequences are shown in the postscript to this chapter.

Kleinman (1991, 1994) has shown that the split between NO_x- and HC-limited regimes can be explained simply in terms of the supply of NO_x relative to the source of odd hydrogen. NO_x-limited chemistry occurs when the supply of odd hydrogen exceeds the supply of NO_x, while HC-limited chemistry (also referred to as light-limited chemistry) occurs when the supply of NO_x exceeds that of odd hydrogen. This explanation is useful because it provides a conceptual basis for understanding how HC–NO_x chemistry varies from location to location and from event to event. For example, HC-limited chemistry is most likely to occur in large cities and during severe events, when the supply of NO_x is largest, and also during periods of low sunlight, which limits the source of odd hydrogen. NO_x-sensitive chemistry is more likely in smaller cities and during more moderate events (i.e., with lower NO_x) and in far downwind locations (Milford et al., 1994). These trends in HC–NO_x chemistry are all consistent with Kleinman's description.

The following is a summary of factors that affect the variation between HC-limited and NO_x-limited chemistry.

HC : NO_x Ratio

As illustrated in Figure 4, HC-sensitive chemistry is associated with low HC/NO_x ratios and NO_x-sensitive chemistry is associated with high HC/NO_x. The importance of HC/NO_x ratios was identified in the early research into the causes of photochemical smog in the 1950s (Haagen-Smit and Fox, 1954).

For many years, the U.S. Environmental Protection Agency (EPA) used a rule of thumb that HC-NO_x chemistry could be deduced based on the HC/NO_x ratio at 6 to 9 A.M., where a ratio of 10* or less was presumed to correspond with HC-sensitive chemistry and a ratio of 20 or more would correspond with NO_x-sensitive chemistry. This approach has been discredited (e.g., NRC, 1991) because it failed to take into account many of the other factors that affect HC–NO_x chemistry, described below. The chemical impact of hydrocarbons also depends on the reactivity of the hydrocarbon species, so that NO_x-sensitive chemistry is more likely when HC reactivity is higher. It is often useful to think of hydrocarbons and HC/NO_x ratios in terms of reactivity-weighted sums rather than just total concentration.

*Hydrocarbon concentrations are often expressed in parts per billion carbon (ppbC), which represents a sum of species concentrations in ppb weighted by the number of carbon atoms contained. HC/NO_x ratios are expressed in pppC/ppb.

Biogenic Hydrocarbons

The inclusion of biogenic hydrocarbons in analyses of photochemical smog often has a large impact on HC–NO_x chemistry. Biogenic hydrocarbons cause an increase in HC/NO_x ratios and therefore cause a shift toward NO_x-sensitive chemistry. The impact of biogenic hydrocarbons is often overlooked because (i) biogenic hydrocarbons are extremely reactive, and consequently have an impact on chemistry out of proportion to their ambient concentrations; and (ii) biogenic emissions are zero at night and low during the morning hours, and are therefore underrepresented in the traditional morning HC/NO_x ratio.

Historically, the role of biogenic hydrocarbons on urban ozone formation was not recognized until 1988. More recently, it has been shown that emission estimates used by the U.S. EPA underestimated biogenic emissions by factors of 3 or more (Geron et al., 1994). Unpublished results from model calculations suggest that use of the higher biogenic emission estimates would cause a shift from primarily HC-sensitive chemistry to primarily NO_x-sensitive chemistry in many cities of the eastern United States. The impact of biogenic hydrocarbons is smaller in Europe (Simpson, 1995).

Geographical Variation

The pattern of geographical variation in HC–NO_x chemistry is largely associated with the photochemical aging process. As stated above, fresh emissions are often in an HC-limited state but evolve toward NO_x-limited chemistry as the air mass ages. In addition, the total accumulation of ozone in a fully aged air mass appears to be controlled entirely by NO_x rather than by HC. In other words, a reduction in hydrocarbons in an HC-limited region has the effect of deferring ozone production until an air mass moves downwind and disperses, but may have little effect on the number of ozone molecules that is produced once the chemistry has run to completion.

The contrast between HC-limited chemistry in an urban center and NO_x-limited chemistry in downwind suburban regions has been dominated most extensively for Los Angeles (Milford et al., 1989, 1994). Extensive measurements and model calculations have supported the view that downtown Los Angeles has HC-limited chemistry. By contrast, ozone in rural locations in the eastern United States (representing photochemically aged air) is usually sensitive to NO_x rather than HC, although there are exceptions. The highest ozone concentrations typically occur in suburban locations approximately 6 h downwind of major urban centers. These locations represent an intermediate situation between the HC-limited chemistry of urban centers and NO_x-limited chemistry in far downwind locations. It is often uncertain whether peak ozone concentrations are associated with HC-limited or NO_x-limited chemistry, and predictions (derived from model calculations) are frequently dependent on model assumptions, e.g., about emission rates, winds, and vertical mixing.

It should be emphasized that predictions for HC–NO_x sensitivity for individual locations are all highly uncertain at this time. HC–NO_x predictions are often based on model calculations with little supporting evidence from ambient measurements.

The most detailed analyses, including both model calculations and analyses of ambient measurements, have been done for rural sites in the eastern United States (e.g., Buhr et al., 1995; Roselle and Schere, 1995; Jacob et al., 1995; Sillman, 1995; Trainer et al., 1993) and for Los Angeles (e.g., Milford et al., 1989; Jacobson et al., 1996) Other area evaluations have been done for Atlanta (Sillman et al., 1995), Mexico City (Sosa et al., 2000), and Europe (Simpson, 1995). For a more complete summary, see Sillman, (1999) and NRC (1991).

Evaluation of Ozone Production through Measurements

Two types of ambient measurements are especially important for evaluating the impact of photochemistry as a source of O_3. One is the correlation between O_3 and CO (e.g., Parrish et al., 1993). Because CO is primarily a product of human activities (either industry or biomass burning), a positive correlation between these species is interpreted as a signal for photochemical smog, especially in the remote troposphere. A second measurement is the correlation between O_3 and the sum of total reactive nitrogen (NO_y, including NO_x, PAN, HNO_3, and other organic nitrates) and between O_3 and the sum of NO_x reaction products (NO_y–NO_x, or NO_z). Because O_3 and NO_z are both produced by similar photochemical processes, there is a strong correlation between these species in polluted environments at times of photochemical activity.

The correlation between O_3 and NO_z is also interpreted as a measure of ozone production efficiency, defined as the ratio of net production of ozone to the loss rate for NO_x [$P(O_3)/L(NO_x)$]. The slope between O_3 and NO_z is determined partly by the ozone production efficiency but is also influenced by atmospheric removal processes, especially for HNO_3. The ozone production efficiency is often used as a basis for interpreting ozone chemistry (e.g., Liu et al., 1987). In addition, Sillman (1995, 1998, 1999) has proposed that the value of the ratio O_3/NO_z can be used as an "indicator" for NO_x-sensitive versus HC-sensitive ozone chemistry.

4 CHEMISTRY OF OZONE FORMATION

Ozone is produced by a reaction sequence that is initiated by reaction of hydrocarbons or CO with the OH radical. Although individual hydrocarbons follow complex reaction pathways, they often conform to the following pattern:

$$\mathrm{HC} + \mathrm{OH} \xrightarrow{[O_2]} \mathrm{RO_2} + \mathrm{H_2O} \qquad (1)$$

$$\mathrm{CO} + \mathrm{OH} \xrightarrow{[O_2]} \mathrm{HO_2} + \mathrm{CO_2} \qquad (2)$$

RO_2 represents a hydrocarbon chain with O_2 attached. The group of RO_2 radicals and HO_2 all react rapidly with NO:

$$RO_2 + NO \xrightarrow{[O_2]} R'CHO + HO_2 + NO_2 \qquad (3)$$

$$HO_2 + NO \xrightarrow{[O_2]} OH + NO_2 \qquad (4)$$

resulting in an intermediate hydrocarbon by-product (R'CHO) and NO_2. This is followed by photolysis of NO_2:

$$NO_2 + h\nu \rightarrow NO + O \qquad (5)$$

The resulting oxygen atom rapidly combines with O_2 to form ozone $(O + O_2 + M \rightarrow O_3 + M)$.

The rate of formation of ozone and other smog elements, including sulfate and nitrate aerosols, depends critically on the OH radical, which initiates the reaction sequence. The complex dependence of ozone on NO_x and hydrocarbons is closely linked to the chemistry of OH and associated radical species, including HO_2 and RO_2 radicals. Because the reaction sequence (1) through (4) operate on the radicals OH, HO_2, and RO_2 without changing the sum $OH + HO_2 + RO_2$, it is useful to regard the latter sum as a family of species (odd hydrogen). Much of the complexity of ozone chemistry can be understood by analyzing sources and sinks for this family. Odd hydrogen sources are almost all photolytic reactions and include the following:

$$O_3 + h\nu \xrightarrow{[H_2O]} 2OH \qquad (6)$$

$$HCHO + h\nu \xrightarrow{[O_2]} 2HO_2 + CO \qquad (7)$$

Odd hydrogen is removed by reactions that produce hydrogen peroxides and nitric acid:

$$HO_2 + HO_2 \rightarrow H_2O_2 + O_2 \qquad (8)$$

$$RO_2 + HO_2 \rightarrow ROOH + O_2 \qquad (9)$$

$$OH + NO_2 \rightarrow HNO_3 \qquad (10)$$

Formation of peroxyacetyl nitrate (PAN) is also a significant sink for odd hydrogen.

It is possible to derive an analytic solution for OH and for the rate of production of ozone as a function of NO_x and HC based on the above reactions (Sillman et al., 1990, 1995). The solution has the form of a fourth degree polynomial for OH, NO_x, and HC (or for ozone production, NO_x, and HC) and reproduces many of the qualitative features of OH and ozone production as a function of NO_x and HC (Fig. 4). HC-limited chemistry occurs when formation of nitric acid (10) represents the dominant loss mechanism for odd hydrogen. In this situation reactions (6), (7), and (10) form an approximate steady state that determines OH. Increased NO_x

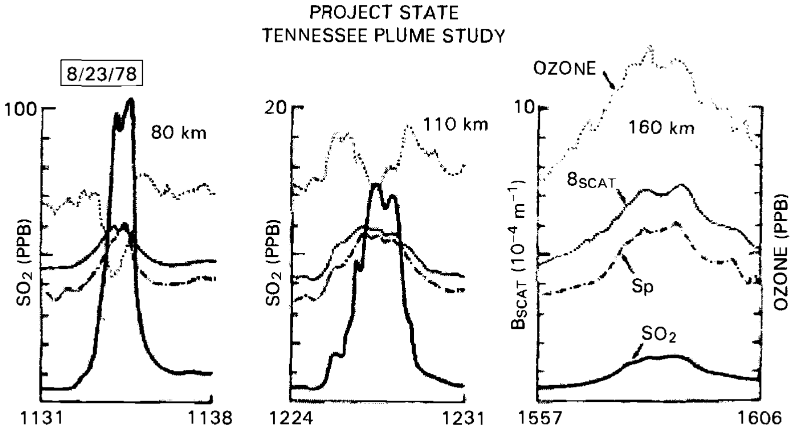


Figure 5 Stages in the chemical development of a power plant plume. The three sets of profiles show measurements of SO_2 (surrogate for NO_x, heavy solid line), ozone (dotted line), particulate sulfur (S_p, line-dot-line), all in ppb; and the light-scattering coefficient (B_{scat}, 10^{-4}/m, light solid line) made during crosswind aircraft traverses through the plume of the Cumberland power plant in NW Tennessee on August 23, 1978. The traverses at 80, 110, and 160 km downwind distances illustrate the "early," the "intermediate," and the "mature" stages of chemical development of the plume, respectively. From Gillani et al., 1996.

causes a decrease in OH and a consequent decrease in the rate of ozone production [approximately equal to the rate of (1)]. Increased HC causes a modest increase in OH [due to (7)] and a larger increase in the rate of (1), which leads to increased ozone production. NO_x-limited chemistry occurs when formation of peroxides [(8) and (9)] represents the dominant sink for odd hydrogen. In this situation the sum $HO_2 + RO_2$ is determined by the steady state between (6), (8), and (9) and is relatively insensitive to changes in NO_x or HC. The rate of ozone formation, approximately equal to the rate of reactions (3) and (4), increases with increasing NO_x and is largely unaffected by HC.

At nighttime O_3 is removed by reaction with NO, as follows:

$$NO + O_3 \rightarrow NO_2 + O_2 \qquad (11)$$

During the daytime reactions (5) and (11) both occur rapidly, but the combination has little effect on ozone concentrations. However, at nighttime, (5) does not occur and (12) results in removal of O_3. Reaction (11) also causes a decrease in ozone during the daytime in the vicinity of a large emission source of NO, e.g., coal-fired power plants. Power plant plumes typically show a decrease in O_3 immediately downwind of the plant, followed by recovery and subsequent increase in O_3 as the ozone-forming reactions (1) to (4) occur (see Fig. 5) (White et al., 1983; Gillani and Pleim, 1996). This pattern of reduced O_3 near the plume source followed by enhanced O_3 downwind is similar to the pattern of evolution of urban plumes with HC-limited chemistry near emission sources and NO_x-limited chemistry further downwind.

ACKNOWLEDGMENTS

Support for this work was provided by the U.S. National Science Foundation (grant #ATM-9713567).

REFERENCES

Bascomb, R., P. A. Bromberg, D. L. Costa, R. Devlin, D. W. Dockery, M. W. Frampton, W. Lambert, J. M. Samet, F. E. Speizer, and M. Utell. Health effects of outdoor air pollution. *Am. J. Resp. Crit. Care Med.*, *153*, 477–498, 1996.

Brimblecombe, P., *The Big Smoke: A History of Air Pollution in London since Medieval Times*, Methuen, London, 1987.

Buhr, M., D Parrish, J. Elliot, J. Holloway, J. Carpenter, P. Goldan, W. Kuster, M. Trainer, S. Montzka, S. McKeen, and F. C. Fehsenfeld, Evaluation of ozone precursor source types using principal component analysis of ambient air measurements in rural Alabama. *J. Geophys. Res.*, *100*, 22853–22860, 1995.

Cardelino, C. A., and W. L. Chameides, Natural hydrocarbons, urbanization, and urban ozone, *J. Geophys. Res.*, *95*, 13971–13979, 1990.

Chameides, W. L., R. W. Lindsay, J. Richardson, and C. S. Kiang, The role of biogenic hydrocarbons in urban photochemical smog: Atlanta as a case study, *Science*, 241, 1473–1474, 1988.

Clarke, J. F., and J. K. S. Ching, Aircraft observations of regional transport of ozone in the northeastern United States, *Atmos. Environ.*, *17*, 1703–1712, 1983.

Geron, C. D., A. B. Guenther, and T. E. Pierce, An improved model for estimating emissions of volatile organic compounds from forests in the eastern United States, *J. Geophys. Res.*, *99*, 12773–12791, 1994.

Gillani, N. V., and J. E. Pleim, Sub-grid-scale features of anthropogenic emissions of NO_x and VOC in the context of regional Eulerian models, *Atmos. Environ.*, *30*, 2043–2059, 1996.

Haagen-Smit, A. J., and M. M. Fox, Photochemical ozone formation with hydrocarbons and automobile exhaust, *J. Air Pollut. Control Assoc. 4*, 105–109, 1954.

Jacob, D. J., B. G. Heikes, R. R. Dickerson, R. S. Artz, and W. C. Keene, Evidence for a seasonal transition from NO_x- to hydrocarbon-limited ozone production at Shenandoah National Park, Virginia, *J. Geophys. Res.*, *100*, 9315–9324, 1995.

Jacob, D. J., J. A. Logan, G. M. Gardner, R. M. Yevich, C. M. Spivakowsky, S. C. Wofsy, S. Sillman, and M. J. Prather, Factors regulating ozone over the United States and its export to the global atmosphere, *J. Geophys. Res.*, *98*, 14817–14827, 1993.

Jacobson, M. Z., R. Lu, R. P. Turco, and O. P. Toon, Development and application of a new air pollution modeling system—Part I: Gas-phase simulations. *Atmos. Environ.*, *30*, 1939–1963, 1996.

Kleinman, L. I., Seasonal dependence of boundary layer peroxide concentration: The low and high NO_x regimes, *J. Geophys. Res.*, *96*, 20721–20734, 1991.

Kleinman, L. I., Low and high-NO_x tropospheric photochemistry, *J. Geophys. Res.*, *99*, 16831–16838, 1994.

Lippman, M., Health effects of tropospheric ozone: Review of recent research findings and their implications to ambient air quality standards, *J. Expos. Anal. Environ. Epidemiol.*, *3*, 103–128, 1993.

Liu, S. C., M. Trainer, F. C. Fehsenfeld, D. D. Parrish, E. J. Williams, D. W. Fahey, G. Hubler, and P. C. Murphy, Ozone production in the rural troposphere and the implications for regional and global ozone distributions, *J. Geophys. Res.*, *92*, 4191–4207, 1987.

Meng, Z., D. Dabdub, and J. H. Seinfeld, Chemical coupling between atmospheric ozone and particulate matter, *Science*, *277*, 116–119, 1997.

Milford, J., D. Gao, S. Sillman, P. Blossey, and A. G. Russell, Total reactive nitrogen (NO_y) as an indicator for the sensitivity of ozone to NO_x and hydrocarbons, *J. Geophys. Res.*, *99*, 3533–3542, 1994.

Milford, J., A. G. Russell, and G. J. McRae, A new approach to photochemical pollution control: Implications of spatial patterns in pollutant responses to reductions in nitrogen oxides and reactive organic gas emissions, *Environ. Sci. Technol.*, *23*, 1290–1301, 1989.

Miller, D. F., A. J. Alkezweeny, J. M. Hales, and R. N. Lee, Ozone formation related to power plant emissions, *Science*, *202*, 1186–1188, 1978.

National Research Council (NRC), Committee on Tropospheric Ozone Formation and Measurement, *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, National Academy Press, Washington, DC, 1991.

Parrish, D. D., J. S. Holloway, M. Trainer, P. C. Murphy, G. L. Forbes, and F. C. Fehsenfeld, Export of North American ozone pollution to the North Atlantic Ocean, *Science*, *259*, 1436–1439, 1993.

Roselle, S. J., and K. L. Schere. Modeled response of photochemical oxidants to systematic reductions in anthropogenic volatile organic compound and NO_x emissions, *J. Geophys. Res.*, *100*, 22929–22941, 1995.

Ryerson, T. B., M. Trainer, J. S. Holloway, D. D. Parrish, L. G. Huey, D. T. Sueper, G. J. Frost, S. G. Donnelly, S. Schauffler, E. L. Atlas, W. C. Kustler, P. D. Goldman, G. Hubler, J. F. Meagher, and F. C Frehsenfeld, Observations of ozone formation in power plant plumes and implications for ozone control stategies, *Science*, **292**, 719–723, 2001.

Sillman, S. Tropospheric ozone: The debate over control strategies, *Annu. Rev. Energy Environ.*, *18*, 31–56, 1993.

Sillman, S., The use of NO_y, H_2O_2 and HNO_3 as indicators for O_3-NO_x-ROG sensitivity in urban locations, *J. Geophys. Res.*, *100*, 14175–14188, 1995.

Sillman, S., The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Millenial review series, *Atmos. Environ.*, *33*(12), 1821–1845, 1999.

Sillman, S., K. Al-Wali, F. J. Marsik, P. Nowatski, P. J. Samson, M. O. Rodgers, L. J. Garland, J. E. Martinez, C. Stoneking, R. E. Imhoff, J-H. Lee, J. B. Weinstein-Lloyd, L. Newman, and V. Aneja, Photochemistry of ozone formation in Atlanta, GA: Models and measurements, *Atmos. Environ.*, *29*, 3055–3066, 1995.

Sillman, S., D. He, M. Pippin, P. Daum, L. Kleinman, J. H. Lee and J. Weinstein-Lloyd, Model correlations for ozone, reactive nitrogen and peroxides for Nashville in comparison with measurements: Implications for VOC-NO_x sensitivity, *J. Geophys. Res.*, *103*, 22629–22644, 1998.

Sillman, S., J. A. Logan, and S. C. Wofsy, The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes, *J. Geophys. Res.*, *95*, 1837–1851, 1990.

Sillman, S. and P. J. Samson, The impact of temperature on oxidant formation in urban, polluted rural and remote environments, *J. Geophys. Res.*, *100*, 11497–11508, 1995.

Simpson, D., Biogenic emissions in Europe, 2, Implications for ozone control strategies, *J. Geophys. Res.*, *100*, 22891–22906, 1995.

Sosa, G., J. West, F. San Martini, L. T. Molina and M. J. Molina, "Air Quality Modeling and Data Analysis for Ozone and Particulates in Mexico City." MIT Integrated Program on Urban, Regional and Global Air Pollution Report No. 15, 76 pages, October 2000, available from http://eaps.mit.edu/megacities/index.html.

Trainer, M., D. D. Parrish, M. P. Buhr, R. B. Norton, F. C. Fehsenfeld, K. G. Anlauf, J. W. Bottenheim, Y. Z. Tang, H. A. Wiebe, J. M. Roberts, R. L. Tanner, L. Newman, V. C. Bowersox, J. M. Maugher, K. J. Olszyna, M. O. Rodgers, T. Wang, H. Berresheim, and K. Demerjian, Correlation of ozone with NO_y in photochemically aged air, *J. Geophys. Res.*, *98*, 2917–2926, 1993.

U.S. Congress, Office of Technology Assessment, *Catching Our Breath: Next Steps for Reducing Urban Ozone*, OTA-O-412, U.S. Government Printing Office, Washington, DC, 1989.

White, W. H., D. E. Patterson, and W. E. Wilson, Jr., Urban exports to the nonurban troposphere: Results from project MISTT, *J. Geophys. Res.*, *88*, 10745–10752, 1983.

Williams, E. J., A. Guenther, and F. C. Fehsenfeld, An inventory of nitric oxide emissions from soils in the United States, *J. Geophys. Res.* *97*, 7511–7519, 1992.

CHAPTER 14

BIOMASS BURNING

ANNE M. THOMPSON

1 INTRODUCTION

Biomass fires are both natural and anthropogenic in origin. The natural trigger is lightning, which leads to mid- and high-latitude fires and episodes of smoke and pollution associated with them. Lightning is also prominent in tropical regions when the dry season gives way to the wet season and lightning in convective systems ignites dry vegetation.

Atmospheric consequences of biomass fires are complex. When considering the impacts of fires for a given ecosystem, inputs of fires must be compared to other processes that emit trace gases and particles into the atmosphere. Other processes include industrial activity, fires for household purposes, and biogenic sources, which may themselves interact with fires. That is, fires may promote or restrict biogenic processes (Fig. 1).

Several books have presented various aspects of fire interactions with atmospheric chemistry (Levine, 1991, 1996; Crutzen and Goldammer, 1993) and a cross-disciplinary review of a 1992 fire-oriented experiment appears in *SAFARI: The Role of Southern African Fires in Atmospheric and Ecological Environments* (van Wilgen et al., 1997). The IGAC/BIBEX core activity (see acronyms at end of chapter) has sponsored field campaigns that integrate multiple aspects of fires—ground-based measurements with an ecological perspective, atmospheric measurements with chemical and meteorological components, and remote sensing (Table 1).

This chapter presents two aspects of biomass fires and the environment. Namely, the relationship between biomass burning and ozone is described, starting with a brief description of the chemical reactions involved and illustrative measurements and interpretation. Second, because of the need to observe biomass burning and its consequences globally, a summary of remote-sensing approaches to the study of fires

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

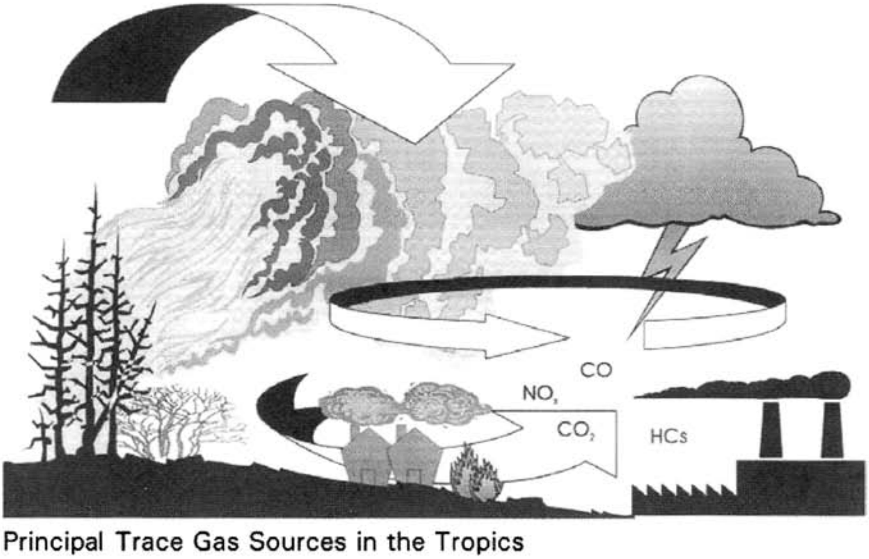


Figure 1 Schematic of processes in tropics with significant production of trace gases—CO, hydrocarbons, or NO—that contribute to tropospheric ozone formation. Biomass fires are major sources of CO, hydrocarbons, and NO, but lightning and soils contribute to NO in the upper troposphere and boundary layer, respectively. Soils release CO as well, under certain conditions and vegetative production is a large source of hydrocarbons. Isoprene from vegetation is oxidized to form CO and more highly reactive oxygenated hydrocarbon intermediates. Besides burning of biomass, burning of wood for fuel use and industrial combustion release the ozone precursors, CO, hydrocarbons, and NO.

and trace gases is given. Examples in this chapter are restricted to tropical burning for matters of brevity and because most burning activity globally is within this zone.

2 CHEMICAL REACTIONS: OZONE FORMATION AND EFFECTS OF FIRES ON ATMOSPHERIC OXIDIZING CAPACITY

Pyrogenic emissions of ozone precursors are abundant (Andreae, 1991), and ozone formation from biomass fires has been the subject of much study (Granier et al., 1996; Lelieveld et al., 1997). The steps in ozone formation are the same as smog reactions in urban environments, although non-gas-phase chemistry may also play a role because particulate emissions from fires are substantial. The release of reactive hydrocarbons (CH_4, but more importantly, nonmethane hydrocarbons), carbon monoxide and NO (nitric oxide) produces a mixture that enhances ozone formation. Table 2 shows the sequence of reactions with NO, CO, and nonmethane hydrocarbons (designated as RH).

TABLE 1 Campaigns with Significant Biomass Burning Observations

Date	Name (Acronym)	Location	Reference
1— 2—Jul–Aug., 1985 April–May, 1987 3—July–August, 1988, 1990	Atmospheric Boundary Layer Experiments (ABLE), 1, 2, 3	1—Tropical Atlantic Ocean 2—Brazilian Rain Forest 3—Alaskan northern wetlands	*JGR*[a] 93: (D2) Feb. 20 1988 *JGR* 95: (D10) Sep. 20 1990 *JGR* 97: (D15) Oct. 30 1992 *JGR* 99: (D1) Jan. 20 1994
I—1987 II—1991	TROPOZ	Europe to America to South Africa and return	I, Quad. Ozone—1988[b]
1988	DECAFE	Equatorial Africa	*JGR* 97: (D6), 6187–6193, 1992
1991	FOS		*J. Atmos. Chem.*, 22 (1), 1995
Aug.–Oct. 1992	TRAnsport and Chemistry near the Equatorial-Atlantic (TRACE-A)	South tropical Atlantic Ocean	*JGR* 101: (D19), 23515–24330, 1996
1992	Southern Africa Fire-Atmosphere Research Initiative (SAFARI-92)	Southern Africa	*JGR* 101: (D19) 23505–24330, 1996
May, 1994	Southern African Atmospheric Research Initiative (SA'ARI-94)	Southern Africa	*S. Afr. J. Sci.* 91: (7), 360–362, July 1995
May–June, 1996 Jan./Feb., 1997	EXPeriment for REgional Sources and Sinks of Oxidants (EXPRESSO)	Central African Republic and Republic of Congo	*JGR* 104: (D23) 30625–30657, Dec. 20 1999
Sep./Oct., 1997	SAFARI-97 Field Campaign in Kenya	Kenya	*JGR* 102: (D15) 18879–18888, Aug. 20 1997
Feb./March, 1998 October, 1999	Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) CLAIRE	Amazon, Brazil; Surinam	*Ann. Geophys-Atm.-Hydr.* 17: (8) 1095–1110, Aug. 1999

Note. Those since 1990 were ICAC/BIBEX sponsored.

[a] JGR = *Journal of Geophysical Research.*

[b] Quad. Ozone—1988 = *Ozone in the Atmosphere*, R. D. Bojkov and P. Fabian, A. Deepak, Eds., Pub., Hampton, VA, 1989.

TABLE 2 Photochemical Reactions Linking Methane, NMHC, CO, and NO with O_3, OH

OH Forms from Ozone Photolysis

$O_3 + h\nu \rightarrow O(^1D) + O_2 \quad O(^1D) + H_2O \rightarrow OH + OH$

Methane Oxidation

$OH + CH_4(+O_2) \rightarrow CH_3O_2 + H_2O$

$CH_3O_2 + NO \rightarrow NO_2 + CH_3O$

[Form formaldehyde: $CH_3O(+O_2) \rightarrow HCHO$]

$HCHO + h\nu \rightarrow H_2 + CO \leftarrow$ formation of CO

$HCHO + h\nu(+O_3) \rightarrow HO_2 + CHO$

NMHC Oxidation

$OH + NMHC(+O_2) \rightarrow RO_2 + H_2O$

$RO_2 + NO \rightarrow NO_2 + RO$

CO Oxidation by OH Produces HO_2

$OH + CO(+O_2) \rightarrow CO_2 + HO_2$

Conversion of NO to NO_2 by HO_2, RO_2, CH_3O_2

$CH_3O_2 + NO \rightarrow NO_2 + CH_3O$

$RO_2 + NO \rightarrow NO_2 + RO$

$HO_2 + NO \rightarrow NO_2 + OH$

Formation of O_3

$NO_2 + h\nu \rightarrow O + NO$

$O + O_2(+M) \rightarrow O_3 + M$

3 RESULTS OF TROPICAL FIELD CAMPAIGNS

Trace Gas Signatures and Ozone Photochemistry

In Brazil and Africa, experiments directed toward biomass burning (e.g., DECAFE, TRACE-A, SCAR-B, TROPOZ I and II, EXPRESSO) and biogenic emissions (ABLE 2, ABLE 3, LBA) have shown that both pyrogenic and biogenic emissions can lead to substantial ozone formation. Biogenic sources appear to be most important in the boundary layer, below canopy level, where soil NO emissions lead to ozone formation. This was seen during ABLE 2A (Jacob and Wofsy, 1988), where 1 ppbv NO built up near the surface producing > 15 ppbv O_3/day. In addition to NO, isoprene emissions were essential to ozone formation. During the SAFARI-92 experiment (September–October 1992), elevated NO levels over the savanna following precipitation (Harris et al., 1996; Zepp et al., 1996) signified biogenic emissions. Aircraft sampling showed that these higher NO signals extended well into the mixed layer and that they lasted 1 to 3 days. This source of NO could contribute to ozone formation, providing a significant fraction in tropical regions during the dry (burning) to wet (nonburning) transition (Swap et al., 1996).

Despite contributions from biogenic sources, persistently high ozone levels throughout the free tropical troposphere during the dry season usually originate from biomass burning and occasionally from urban areas. Figure 2 shows typical ozone and ozone precursor profiles in a region affected by biomass burning, during the October 6, 1992 TRACE-A flight over Zambia. Figure 3 summarizes mean

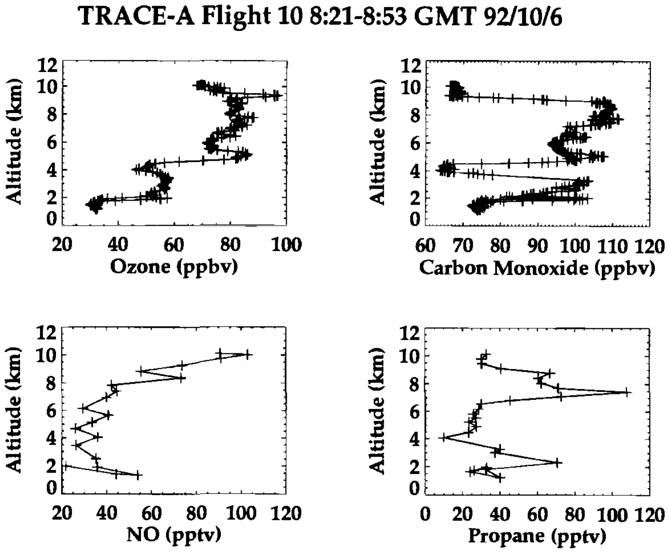


Figure 2 Profiles of ozone and ozone precursors CO, NO, and propane from sampling on board the NASA/DC-8 during the TRACE-A field experiment (on October 6, 1992, flight from Johannesburg to Zambia. Data available from: *http://www-gte.larc.nasa.gov.*)

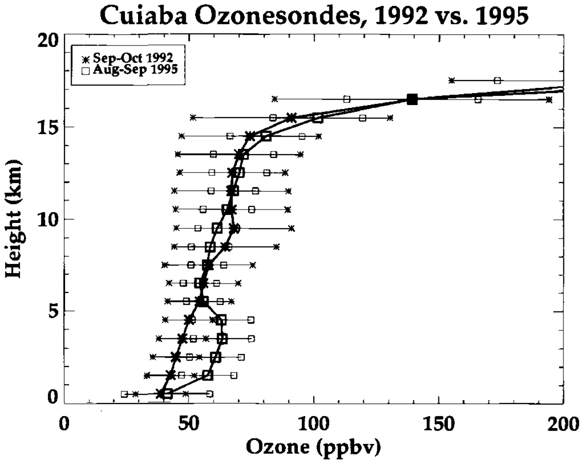


Figure 3 Ozone profiles from ozonesondes at Cuiabá, Brazil (16°S, 56°W) during biomass burning field experiments of 1992 (TRACE-A), which was a relatively low fire activity year and 1995 (SCAR-B), with greater burning activity. Mean 1-km profiles (to ±1-sigma) shown. Data from V. W. J. H. Kirchhoff.

profiles from ozonesondes launched at Cuiabá (Brazil, 16°S, 56°W) during two campaigns: TRACE-A and SCAR-B. Biomass burning (Artaxo et al., 1998) was much lower in 1992 (TRACE-A) than in 1995 (SCAR-B), and the boundary layer was more stable during the latter period; hence ozone levels in the lower troposphere were greater during SCAR-B.

Studies of ozone photochemical formation during sampling periods on TRACE-A have been made with photochemical steady-state ("point") models (Jacob et al., 1996; Thompson et al., 1996; Zenker et al., 1996; Mauzerall et al., 1998). The mixed layer, near the surface, usually has net ozone formation. For example, in the 4 km nearest the surface, relatively fresh emissions sampled over Brazil (September 27, 1992) and Zambia (October 6, 1992) during TRACE-A produced 10 to 15 ppbv O_3/day. During TROPOZ I, near the Ivory Coast in December 1987, air parcels with emissions less than 2 days old formed ozone at a 15 to 35 ppbv O_3/day rate (Jonquières et al., 1998). In PEM–Tropics A, near biomass burning in southeast Asia, near-surface ozone formation averaged >6 ppbv ozone/day (Schultz et al., 1999). Ozone formation increased with altitude because NO was supplied by peroxyacetylnitrate (PAN) transported into the region (Schultz et al., 1999). In contrast, during TRACE-A, above the mixed layer, ozone formation was in balance between production and loss or net negative during TRACE-A (Jacob et al., 1996; Thompson et al., 1996) because NO was depleted.

In the upper troposphere, photochemical formation often proceeds at modest rates (1 to 3 ppbv/day) and the ozone photochemical lifetime is 2 weeks to 2 months (Thompson et al. 1996; Jacob et al., 1996; Schultz et al., 1999). Usually, ozone formation is slightly positive because NO concentrations are sufficiently high. The requisite NO concentrations (>50 pptv) may come from lightning enhancement of NO, recycling of reactive nitrogen, or convective injection of NO_x. Venting of the boundary layer in convective cells produced by the intense heat of biomass fires is another mechanism whereby ozone precursors are injected into the free troposphere (Chatfield et al., 1996; 1998). This was seen during African aircraft sampling in TROPOZ (Jonquières et al., 1998) and TRACE-A (Thompson et al., 1996; Mauzerall et al., 1998).

Analyses of ozone formation show its evolution during transit away from the continents of biomass burning (Chatfield et al., 1996; Thompson et al., 1996; Jonquières and Marenco, 1998; Jonquières et al., 1998; Mauzerall et al., 1998; Schultz et al., 1999). Mauzerall et al. (1998) classified the age of air in terms of reactive nitrogen species (NO, HNO_3, PAN) and CO content, using ratios of tracers like CO_2, C_2H_2, C_2H_6, and CH_3COCH_3. In parcels sampled during TRACE-A, the limiting ozone-forming reactant was NO, which tended to be used up within a day or so. Older air parcels are refreshed with NO as PAN decomposes thermally, releasing NO_2, which is rapidly photolyzed to NO. Thus, downwind, ozone formed from NO that was supplied by PAN.

Schultz et al. (1999) examined photochemical characteristics of African plumes, observed thousands of kilometers from their sources, over the Pacific during the September PEM–Tropics A expedition. Figure 4 shows CO over the tropical Pacific (Blake et al., 1999), with many elevated CO segments due to pyrogenic sources from several continents (Olson et al., 1999). Schultz et al. (1999) also find the PAN mechanism for NO to be dominant, as in TRACE-A, but advection of ozone, not

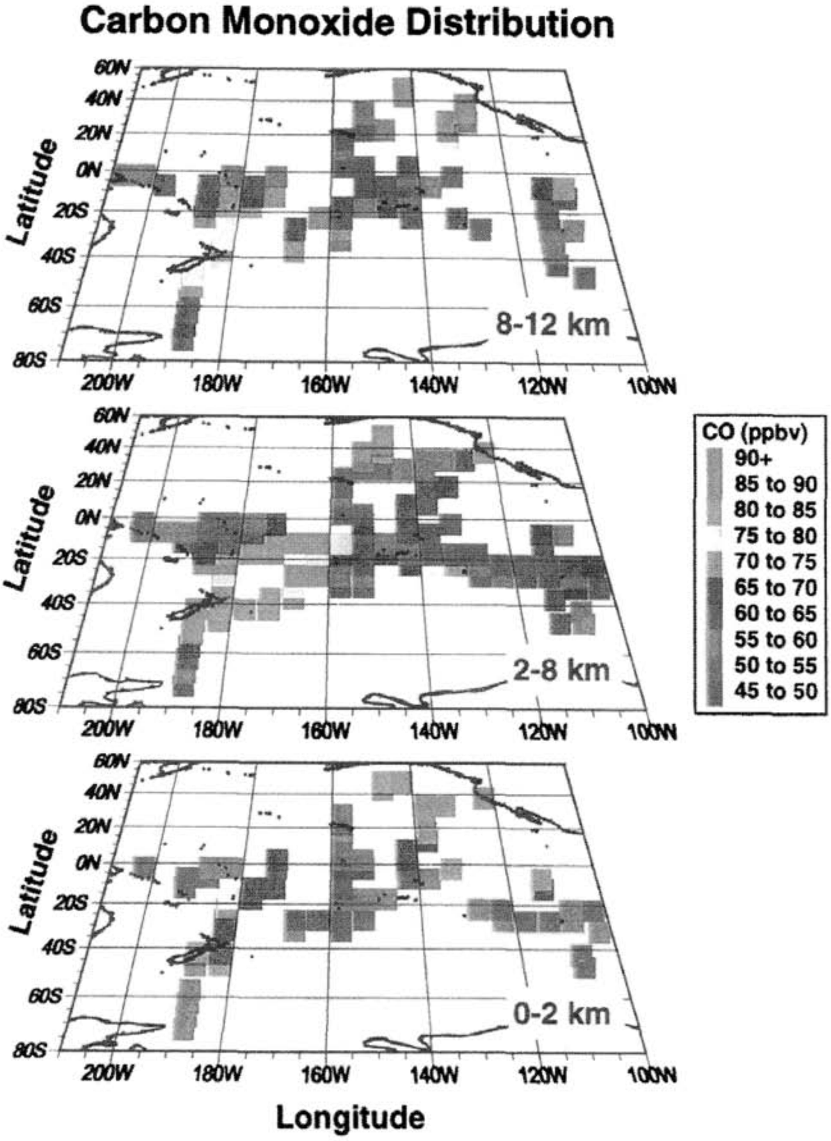


Figure 4 (see color insert) CO over tropical Pacific during September 1996 PEM–Tropics A sampling (from Blake et al., 1999). Measurements by G. W. Sachse with a lidar-based instrument. Analysis of possible fire sources is described by Olson et al. (1999). See ftp site for color image.

local photochemistry, is still a major tropical Pacific ozone source. A plume of ozone with African origins observed over the western Pacific appears in Figure 5.

Biomass burning is not the only large nonurban NO source that contributes to tropical ozone formation. Lightning is also a significant source. TRACE-A observa-

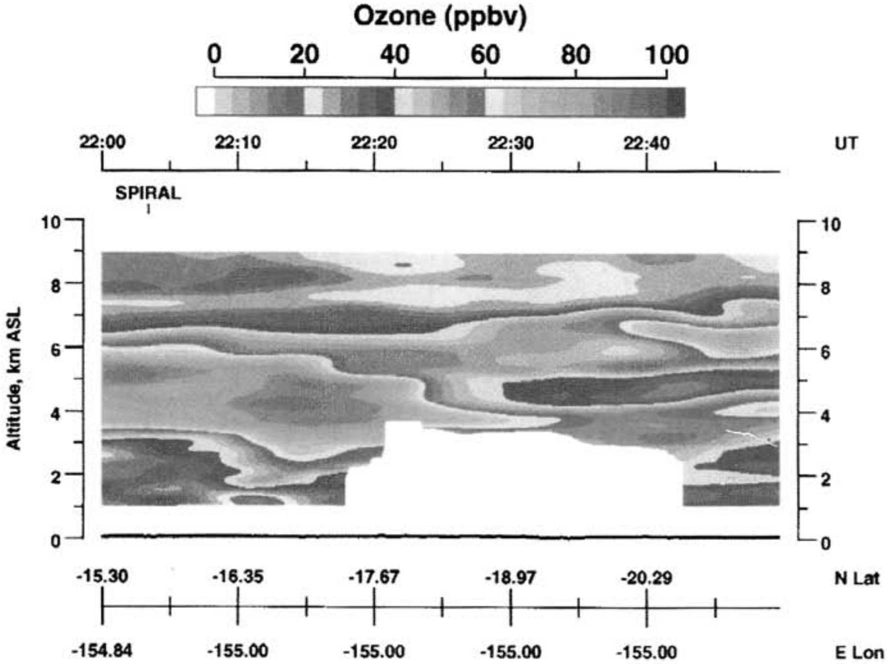


Figure 5 (see color insert) Ozone plume over the Pacific seen during the PEM–Tropics A aircraft mission in Sept.–Oct. 1996. (from Fenn et al., 1999). See ftp site for color image.

tions throughout the south tropical Atlantic, for example, showed elevated, relatively fresh NO in the upper tropospheric that did not always track other tracers of biomass burning (Smyth et al., 1996). A TRACE-A flight (September 27, 1992) over Brazil in which deep convection transported relatively fresh biomass burning emissions to the upper troposphere (Pickering et al., 1996) was also punctuated by lightning-produced NO. From comparison of NO enhancements to other biomass burning emittants (CO, hydrocarbons), it appeared that 35 to 40% of the NO from the September 27, 1992 flight on TRACE-A was due to lightning.

Transport of Trace Gas Emissions and Ozone from Biomass Burning

Aircraft, ground-based and sounder sampling, in combination with trajectory and regional dynamical modeling, elucidates the roles of convection and long-range transport in determining the distribution of smoke aerosol, tropical ozone, and ozone precursor distributions. Over large biomass burning regions of north equatorial Africa, the Harmattan winds cause large-scale transport of biomass burning products from the continent in a southwesterly flow to the Atlantic and toward South America (Jonquières and Marenco, 1998; Fig. 6*a*). From southern African savanna burning, convection on a large regional scale drives a Southern Hemisphere "Great Plume" (Chatfield et al., 1996, 1998) toward the tropical Atlantic and Indian

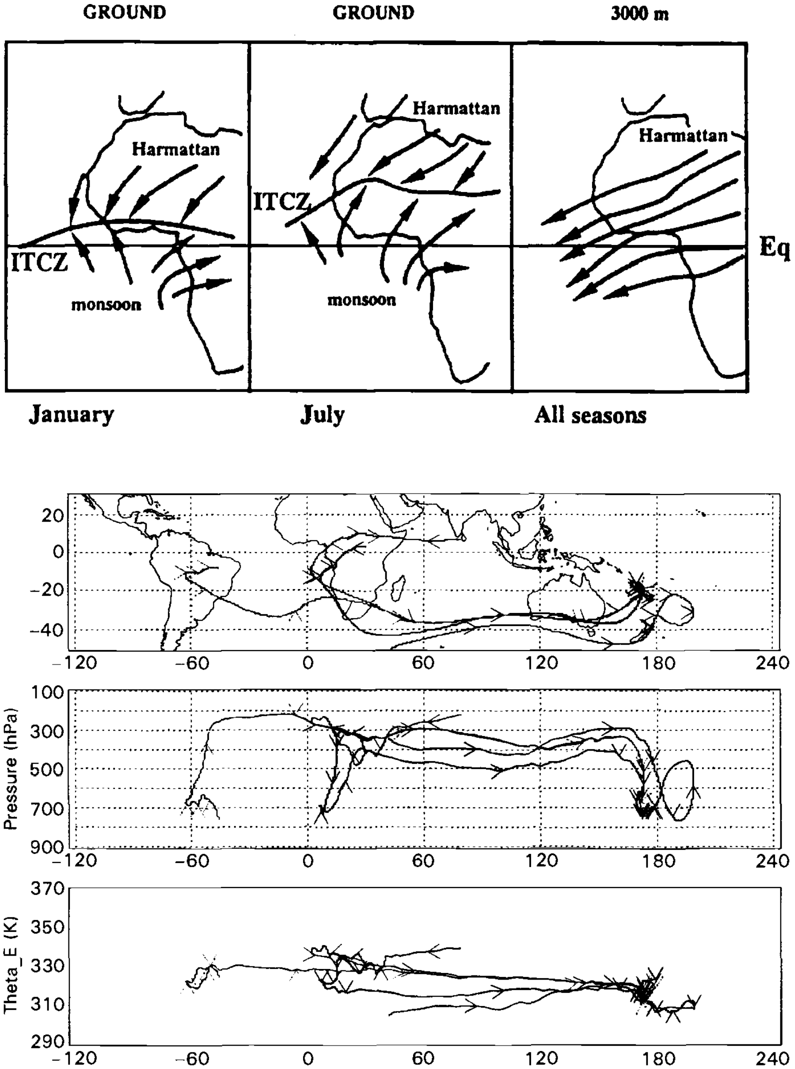


Figure 6 (*a*) Transport patterns from north equatorial Africa (Jonquières et al., 1998); (*b*) Aspects of the "Southern Global Plume" from Guo and Chatfield (1998). Back trajectories are initiated at several points along the NASA DC-8 flight of September 23–24, 1996 (PEM-Tropics A). Top panel shows traces back to origin areas in South America and Africa. Middle panel shows pressures of trajectories. Directional arrows are spaced every 2 days. Lowest panel shows potential temperature of trajectory, which is nearly conserved, demonstrating that trajectory does not cross into other air masses. Model is based on MM5 mesoscale model. See ftp site for color image.

Oceans. Figure 6*b* shows that fires from South America can also affect the Indian and Pacific Oceans.

Ozone from both southern Africa and South America appears seasonally over the south Atlantic in tropospheric ozone satellite retrievals (Chapter 5). South of 15°S, the predominant exit for biomass burning emissions and ozone, which can accumulate in stable layers (Garstang et al., 1996; Garstang and Tyson, 1997; Tyson et al., 1997) is toward the Indian Ocean. In both Atlantic and Indian Ocean exit routes from southern Africa, emissions from fires, vented by shallow or deep convection, inject most of the ozone precursors into the 4 to 8 km layer. These are readily detected in ozone profiles from balloon-borne sondes released over Réunion Island (21°S, 55°E; Baldy et al., 1996; Randriambelo et al., 1999). Examples of fire-affected layers at Réunion appear in Figure 7. In Guo and Chatfield (1998), tracers in the 5th version PennState/NCAR Mesoscale Model (MM5) simulate the flow of CO from industrial, biogenic, and biomass burning sources over southern Africa to the western Pacific Ocean. CO mixing ratios computed by the model agree with observations during PEM–Tropics A (Hoell et al., 1999).

The route of biomass burning emissions from Brazil has been studied on ABLE 2A, TRACE-A and SCAR-B. Figure 8, which is based on a composite of forward trajectories during the SCAR-B experiment (Longo et al., 1999), shows air parcels

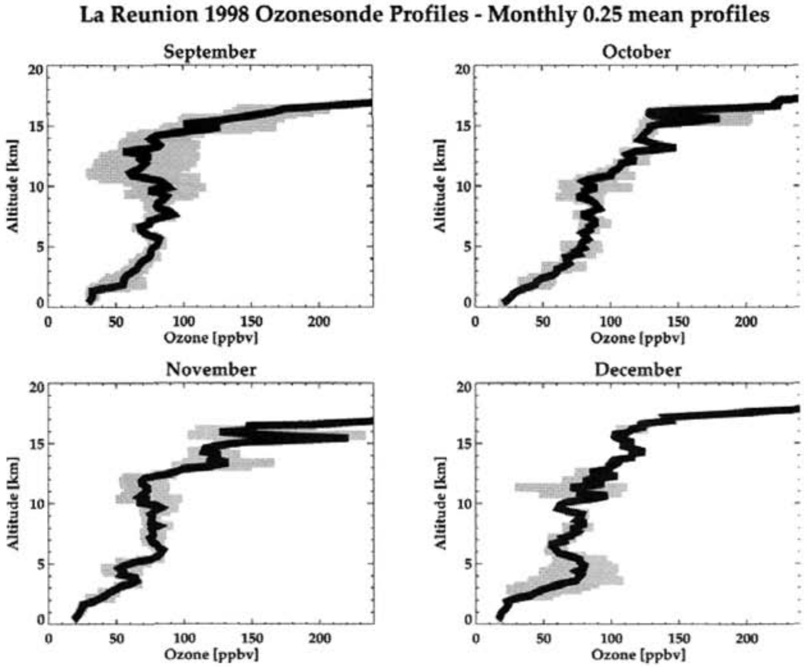


Figure 7 Ozone soundings over Réunion Island (21°S, 55°E) in the Indian Ocean, with layers of high ozone due to transport from African burning. Mean monthly profiles with l-sigma shading.

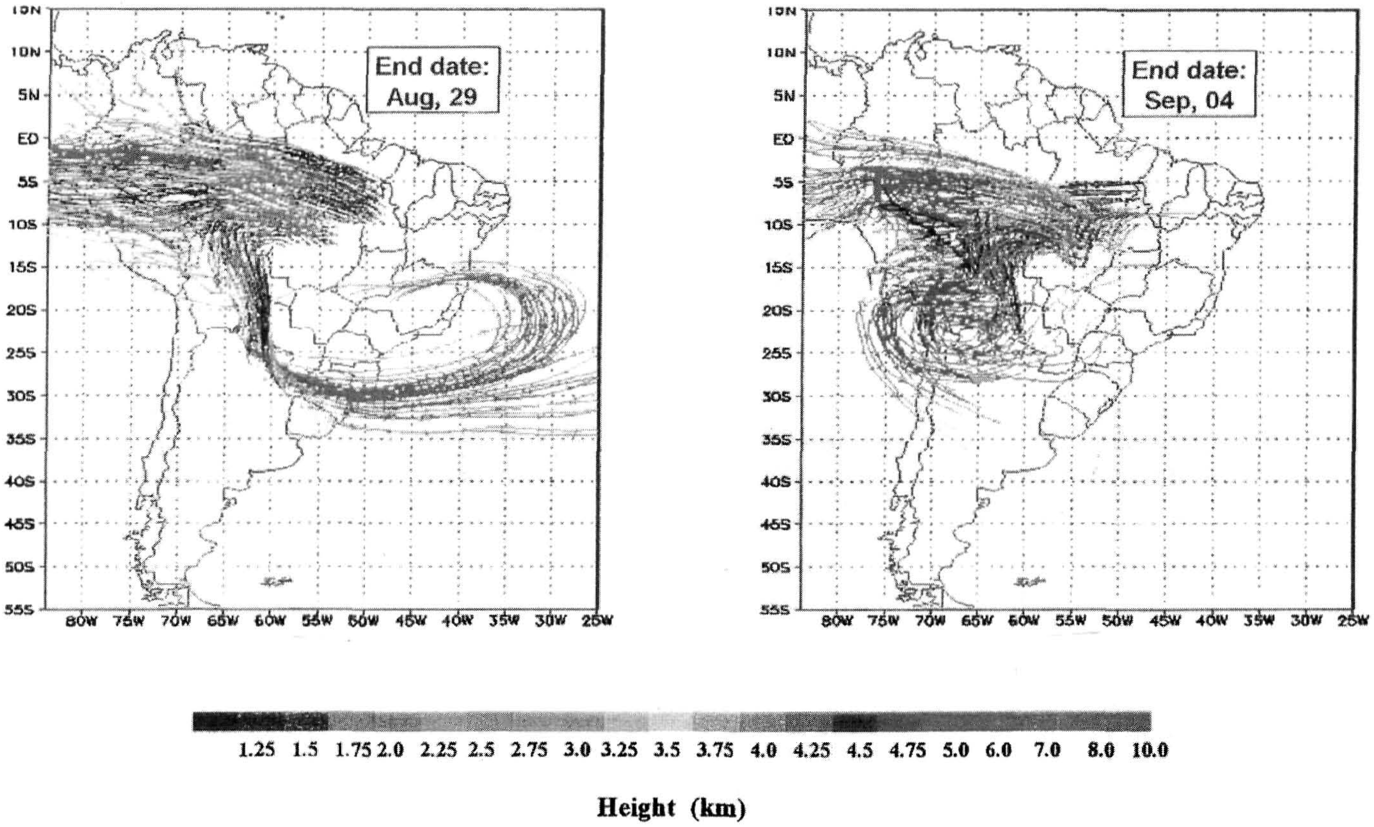


Figure 8 (see color insert) Composite of forward trajectories from Cuiabá during the 1995 SCAR-B field experiment. A Brazilian version of the Colorado State mesoscale RAMS model was used to provide winds for the University of São Paulo kinematic trajectory model (from Longo et al., 1999). See ftp site for color image.

from active burning regions sending ozone and ozone precursors over mountains toward the eastern Pacific Ocean. Unfortunately, there was no satellite remote sensor available in August 1995 to detect flows over the eastern Pacific during SCAR-B. However, satellite data from 1979 to 1992 (Kim and Newchurch, 1996; Thompson and Hudson, 1999) show a seasonal drift of ozone into this region. During TRACE-A, the predominant post-convective flow from Brazilian biomass burning areas at the onset of the wet season was in the westerlies toward the Atlantic. It was estimated that upper tropospheric ozone was largely supplied from the South American continent (Thompson et al., 1997), with additional ozone resulting from lightning-produced NO.

Ozonesondes over Africa and South America, near or downwind from sources, as well as ozonesondes at more remote locations—Réunion (21°S, 55°E), Ascension (8°S, 14°W), American Samoa (14°S, 170°W)—show impacts of biomass burning ozone (Cros et al., 1992; Fishman et al., 1992; Baldy et al., 1996; Oltmans et al., 1998). Examples of ozone profiles at Pretoria (25°S, 28°E) and Etosha Park (19°S, 15°E) during SAFARI-92 and TRACE-A appear in Figures 9*a* and 9*b*. Neither of these sites is in a burning region, but clusters of back trajectories initiated at the peaks with arrows show that they may be several days' transport time from African burning or a week from South American savanna burning. Back trajectories from the Etosha Park ozonesonde profile of October 11, 1992, indicated significant exposure to fires within 2 days (Fig. 9*c*). For the Pretoria ozonesonde sample on October 11, 1992, launched within 30 km of Johannesburg, the number of fires encountered in a 5-day back trajectory is less and travel time is greater than air parcel origins on October 11, 1992, at Etosha Park.

Airborne sampling and sounding profiles show that layers of enriched or depleted ozone are remarkably stable (Garstang et al., 1996; Garstang and Tyson, 1997; Newell et al., 1999). Using the SAFARI-92/TRACE-A soundings over Irene, Garstang et al. (1996) found very little vertical mixing and estimated that some of the stable layers observed had lifetimes greater than 50 days.

4 REMOTE SENSING

Remote sensing is an invaluable tool for looking more closely at biomass burning effects in the atmosphere. In terms of trace gases, ozone and CO instrumentation flown on aircraft and in space has seen the imprint of biomass fires on a widespread basis. Remote sensing of carbonaceous absorbing aerosols (soot, smoke) is made from airborne platforms and from a number of space-borne instruments. In addition, imagers are able to detect fires (Cahoon et al., 1992) and fire burn scars on the earth (Justice et al., 1996). Instrumentation is summarized in Table 3 and applications are described below.

Carbon Monoxide

The Shuttle-borne Measurement of Air Pollution by Space (MAPS; Reichle et al., 1990; *Journal of Geophysical Research*, Aug. 20, 1998) instrument is a gas correla-

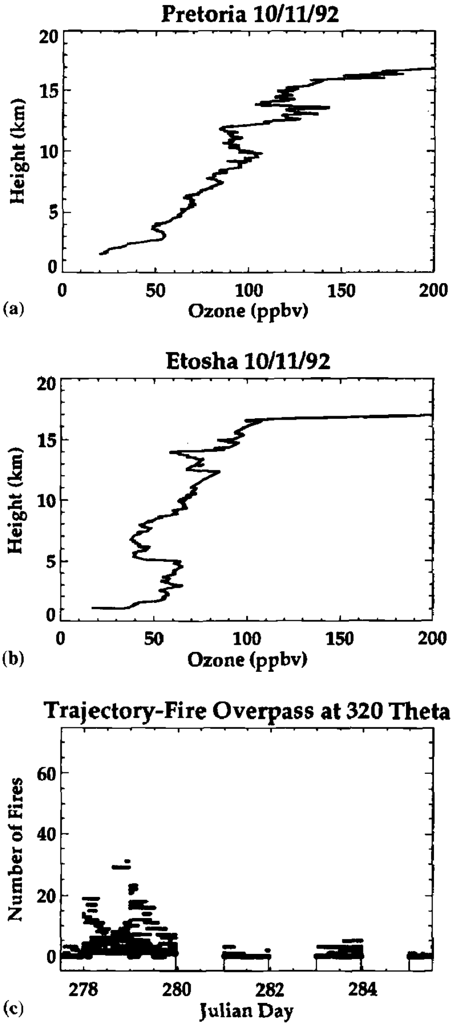


Figure 9 (*a*) Pretoria ozone sounding for October 11, 1992. (*b*) Etosha sounding for October 11, 1992. (*c*) Fires passed over by air parcels in a cluster of back trajectories initiated at ~5 km ($\theta = 320$ K) at the Etosha location on October 11, 1992. This suggests fire emissions contribution to ozone profile in (*b*). Satellite fire counts from Justice et al. (1996); gaps refer to days with missing fire data.

TABLE 3 Remote Sensing Instrumentation for Detection of Smoke, Fires, and Trace Gas Emissions

MAS: airborne surface imager
AVHRR, GOES: smoke detection from fires
AVHRR: surface imaging for active fires and burn scars
Cloud and smoke lidar: NASA/ER-2 instrument
DMSP: active fires
GOME: ozone, NO_2, HCHO, BrO
MAPS, MOPITT: CO
TOMS: ozone, smoke, and dust aerosol (also SO_2, sulfate aerosols)

tion radiometer that senses CO by differencing two cells. For the region of the atmosphere of greatest sensitivity, between 5 and 10 km, MAPS gives an accurate measurement of carbon monoxide. Operating on the Space Shuttle in 1981, twice in 1984 and in 1994, with data covering 55°N to 55°S, MAPS observed CO from urban pollution as well as from biomass burning. Biomass burning signatures in the tropics

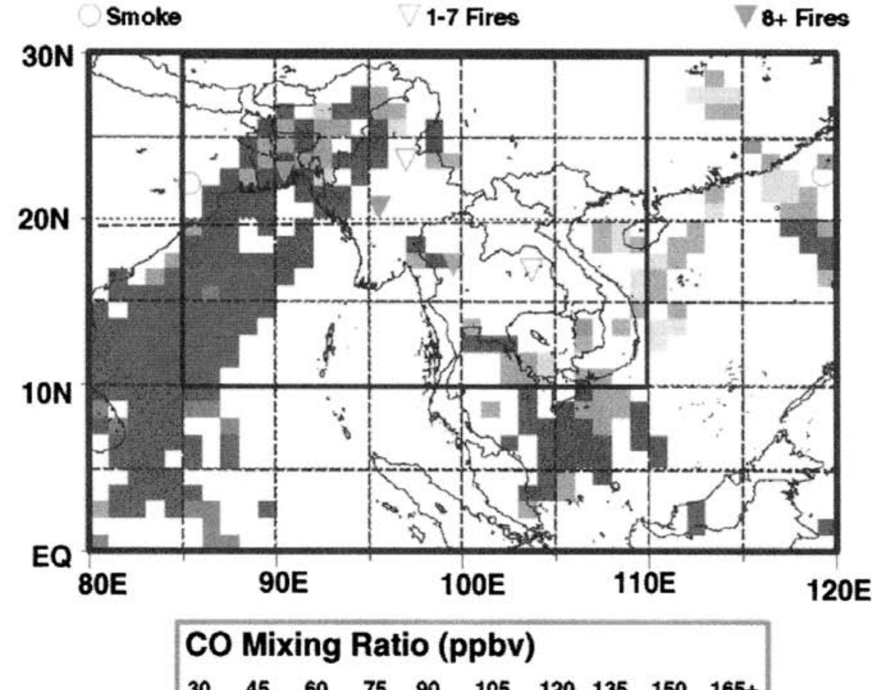


Figure 10 (see color insert) (*a*) MAPS CO, April 1994 (from Christopher et al., 1998). See ftp site for color image.

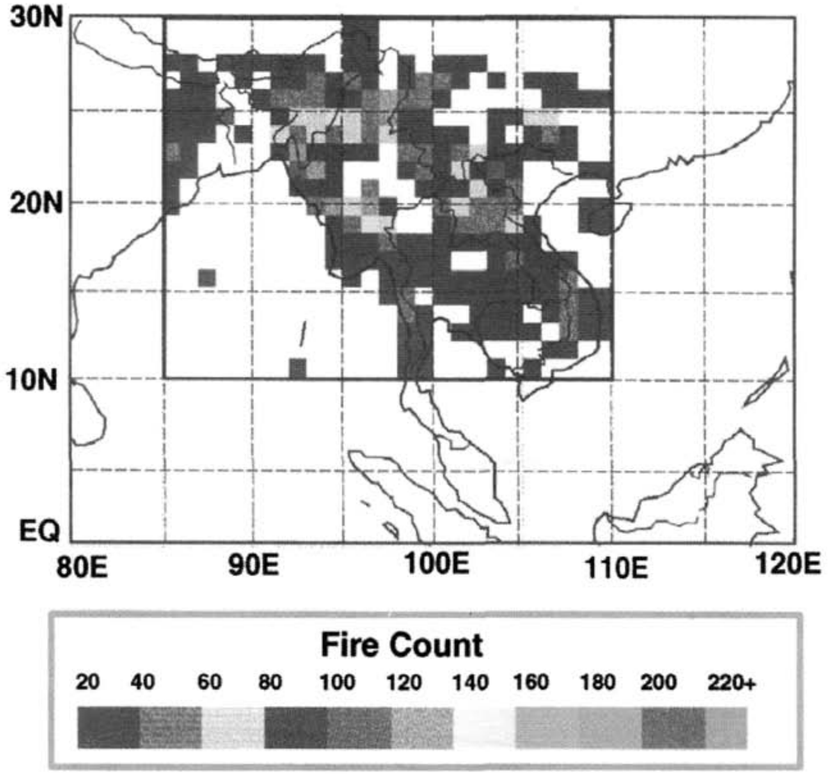


Figure 10 (see color insert) (*b*) coincident fires during April 1994 Space Shuttle flight (from Christopher et al., 1998). See ftp site for color image.

are evident as elevated CO concentrations, usually >60 ppbv. This was confirmed in an airborne campaign of validation measurements conducted during the 1994 MAPS operations. Because MAPS detects midtropospheric CO, it essentially detects areas of burning and convective transport in which CO from the boundary layer is transported to midtroposphere. Urban CO that escapes the boundary layer can also be detected. Figure 10 shows MAPS CO and remotely sensed fires contributing to the CO over southern Asia (from Christopher et al., 1998).

MOPITT, the new CO and methane sensor aboard the *Terra* spacecraft, was launched into orbit in December 1999. As of this writing, MOPITT observations had not yet begun.

Tropospheric Ozone

The application of ozone remote sensing to the troposphere, in a series of studies by Fishman and co-workers (Fishman et al., 1986, 1990; Fishman and Brackett, 1997),

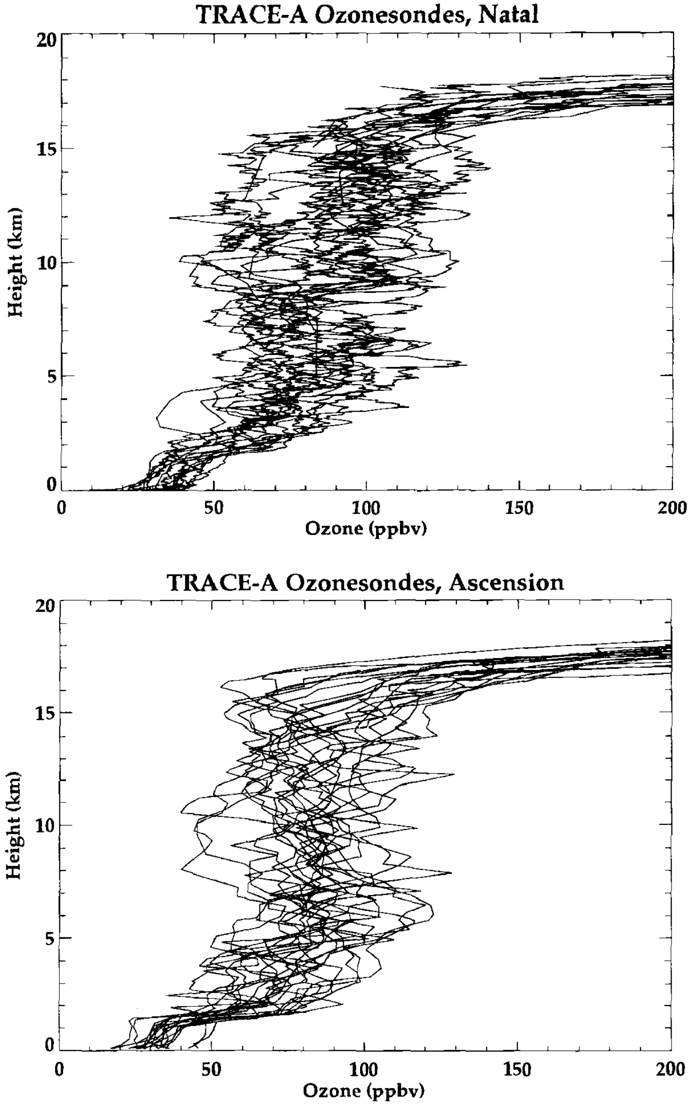


Figure 11 Ozonesonde profiles during the TRACE-A field experiment (September 9 to October 22, 1992) over (*a*) Natal, Brazil (6°S, 35°W) and (*b*) Ascension Island (8°S, 15°W).

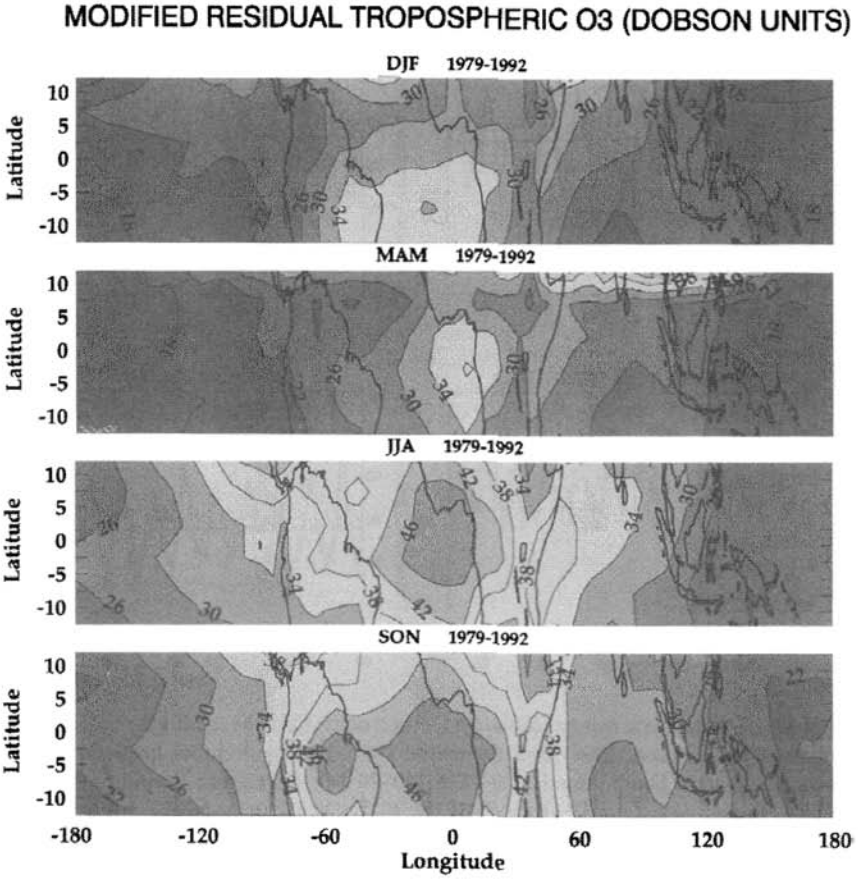


Figure 12 (see color insert) Wave-one pattern in tropospheric ozone apparent in TOMS satellite data, averaged from 2 maps/month during the 1979–1992 *Nimbus 7* observing period. Wave appears to be present throughout year. Scale is DU (Dobson units). Cf. Figure A1 in Thompson and Hudson (1999). See ftp site for color image.

gave the first insight into the extent of biomass burning effects on tropospheric ozone. The entire south Atlantic basin shows a tropospheric ozone maximum in the latter part of the Southern Hemisphere biomass burning season. Because the TOMS satellite instrument measures column ozone, and has limited sensing capacity below 500 mbar, the vertical characteristics of enhanced ozone seen from space had to be confirmed by ozonesondes (Fishman et al., 1992). Layering of ozone from the boundary layer to 15 km is evident in sondes from Natal (coastal Brazil at 6°S) and Ascension Island (8°S, 15°W; Fig. 11). These profiles were taken during the 1992 SAFARI/TRACE-A experiments.

The intensity of the ozone maximum feature varies from year-to-year, and the chemical consequences of biomass burning appear to overlie a persistent wave-one

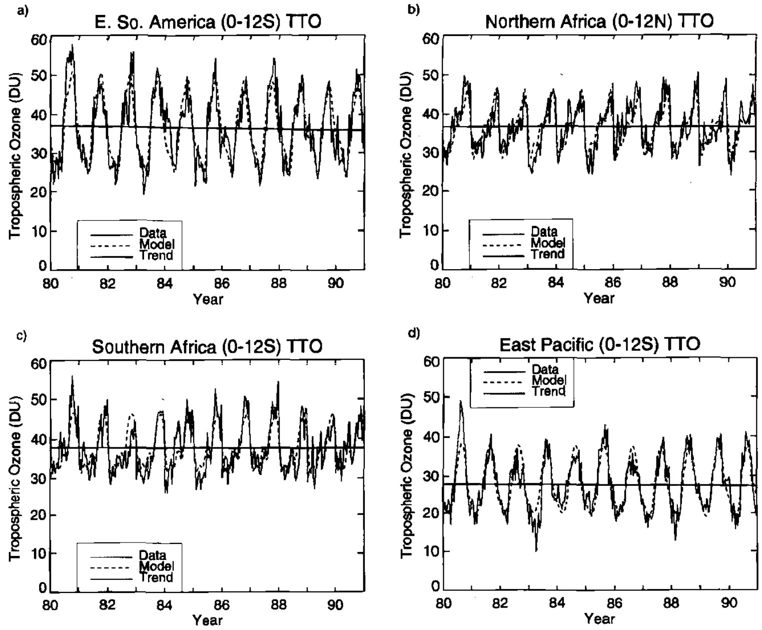


Figure 13 Tropospheric column ozone (in Dobson units) from the modified-residual method (Thompson and Hudson, 1999) over the period 1980–1990 within four tropical regions as follows: (*a*) eastern South America (0–12°S, 40–70°W); (*b*) eastern Pacific (0–12°S, 80–110°E); (*c*) northern equatorial Africa (0–12°N, 20°W–20°E); (*d*) southern equatorial Africa (0–12°S, 0–30°E). Data available at *metosrv2.umd.edu/~tropo/*

pattern that maintains nonpollution Atlantic tropical tropospheric ozone at an always greater column depth than nonpollution Pacific ozone (Hudson and Thompson, 1998; Thompson and Hudson, 1999; Ziemke et al., 1996). An example of wave-one patterns in tropospheric ozone, taken from tropical tropospheric ozone maps, appears in Figure 12. During the Southern Hemisphere dry season, tracers of savanna fires and photochemical analysis show that elevated ozone over the south Atlantic basin is dominated by biomass burning sources; see references for TROPOZ, DECAFE, SAFARI, TRACE-A, and SCAR-B (Table 1). This ozone amounts to 20 to 30 DU (Dobson unit) more over the Atlantic region than over the Pacific at the same season and 20 to 30 DU more than is over the Atlantic during its seasonal minimum (March–May). These features are apparent in a time series of TOMS-based tropospheric ozone data (Thompson and Hudson, 1999; Ziemke et al., 1998). Mean annual tropospheric ozone column in eastern South America and Africa (deseasonalized value given by straight lines in Figs. 13*a* to 13*c*) is 38 DU compared to 28 DU over the eastern Pacific (Fig. 13*d*).

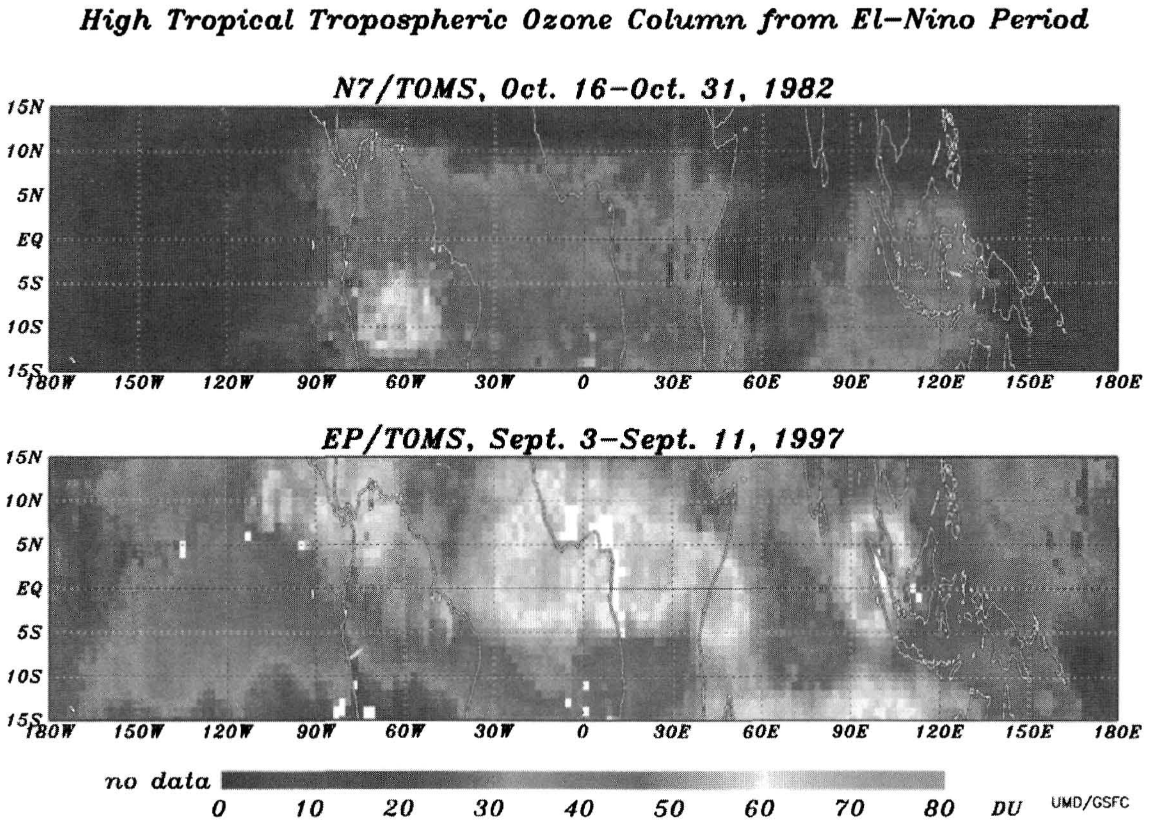

High Tropical Tropospheric Ozone Column from El-Nino Period
N7/TOMS, Oct. 16–Oct. 31, 1982
EP/TOMS, Sept. 3–Sept. 11, 1997
15N
10N
5N
EQ
5S
10S
15S
180W
150W
120W
90W
60W
30W
0
30E
60E
90E
120E
150E
180E
no data
0
10
20
30
40
50
60
70
80
DU
UMD/GSFC

Figure 14 (see color insert) Tropospheric column ozone (in DU, from modified-residual method; Thompson and Hudson, 1999) during El Niño-Southern Oscillation (ENSO) of late 1982 (upper panel) as seen in tropical tropospheric ozone map and for September 1997 (lower panel). See ftp site for color image.

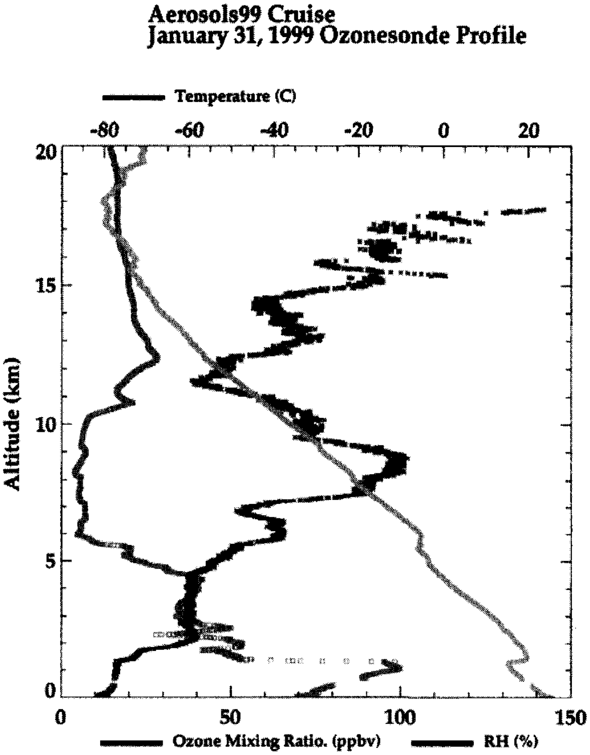


Figure 15 (see color insert) (*a*) Profiles of ozone, temperature and water vapor (as percent relative humidity) from 0 to 20 km on January 31, 1999 during Aerosols99 cruise of R/V *Ronald H. Brown.* Anti-correlation of high ozone between 7 and 10 km suggestive of aged stratospheric air. (*b*) Comparison of integrated tropospheric column ozone from sondes launched along Atlantic transect of R/V *Ronald H. Brown* (Thompson et al., 2000) in January–February 1999 and from sondes launched along January–February 1993 Atlantic transect of R/V *Polarstern* (Weller et al., 1996). See ftp site for color image.

The TOMS-based maps are used to characterize interannual variability and seasonality of tropical ozone. In Figure 13, two features stand out. One is that over the 11-year period illustrated, there is no significant trend in tropospheric ozone (Thompson and Hudson, 1999; Chandra et al., 1999), despite an apparent increase in smoke aerosols in some of these areas (Hsu et al., 1999). The second noteworthy feature is the presence of extremes in tropospheric ozone during the strong El Niño episode of late 1982 and early 1983. For example, over South America there was elevated ozone due to higher-than-usual biomass burning activity (Fig. 14, upper panel), whereas over the eastern Pacific, with enhanced convective activity, upward transport of ozone diluted (and reduced) column ozone. Very high ozone and biomass burning aerosol signals were observed by TOMS over the

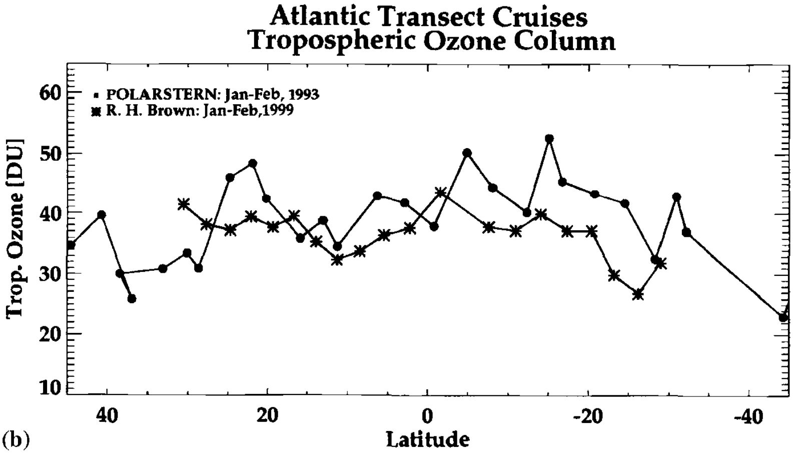


Figure 15b

Indonesian region (Fig. 14, lower panel) during the 1997–1998 El Niño event. Ozonesondes in Java, downwind of the most intensely burning regions of Indonesia (Liew et al., 1999) also registered high tropospheric column ozone (Fujiwara et al., 1999).

Krishnamurti et al. (1996) showed that ozone accumulates over the south Atlantic due to dynamical forces that tends to produce an Atlantic–Pacific ozone gradient (more ozone over the Atlantic) irrespective of chemical sources. Evidence of dynamical effects are easiest to isolate in ozonesondes recorded over the tropical Atlantic during the Southern Hemisphere wet season. For example, Atlantic oceanographic transects with ozonesonde launches (Smit et al., 1989; Weller et al., 1996; Thompson et al., 2000) have shown free tropospheric ozone in the Southern Hemisphere dominated by high ozone layers that may originate from cross-hemispheric transport (Jonquières and Marenco, 1998) or aged air parcels from the stratosphere (the latter inferred from very low water vapor, Fig. 15*a*). The result is a paradox with respect to biomass burning in that there is more tropospheric ozone in the Southern Hemisphere wet season than there is north of the Intertropical Convergence Zone (Fig. 15*b*), where there is active biomass burning over northern equatorial Africa.

ACRONYMS

ABLE	Amazon Boundary Layer Experiment (A = 1985; B = 1987)
AVHRR	Advanced Very High Resolution Radiometer
BIBEX	Biomass Burning Experiment
CLAIRE	Coordinated LBA (Large Basin Amazonia) Atmospheric Experiment

DECAFE	Dynamique Et Chimie Atmosphérique en Forêt Equatoriale [1988; FOS (Fires of Savannas)/DECAFE = 1991]
DMSP	Defense Mapping Satellite Project
EXPRESSO	Experiment for Regional Sources and Sinks of Oxidants (1996)
GOME	Global Ozone Monitoring Experiment (operating 1995–)
IGAC	International Global Atmospheric Chemistry Project
MAPS	Measurements of Air Pollution from Shuttle (1981, 1984, 1994)
MAS	MODIS (Moderate Resolution Imaging Spectrometer) Airborne Simulator
MOPITT	Measurements of Pollution in the Troposphere
PEM–Tropics A	Pacific Exploratory Mission (1996)
PEM–Tropics B	Pacific Exploratory Mission (1999)
SAFARI	Southern African Fire Atmospheric Research Initiative (1992)
SCAR-B	Smoke, Clouds and Radiation—Brazil (1995)
SEAFIRE	Southeast Asia Fire Experiment (1997)
TOMS	Total Ozone Mapping Spectrometer (*Nimbus 7*, 1978–1993; *Meteor*, 1991–1994; *ADEOS*, 1996–1997; *Earth-Probe*, 1996–)
TRACE-A	Transport and Atmsopheric Chemistry near the Equator—Atlantic (1992)
TROPOZ	Tropospheric Ozone Campaigns (I = 1987; II = 1991)

ACKNOWLEDGMENTS

Thanks to Robert Chatfield and Volker Kirchhoff for comments, discussion, and prepublication results. Graphical and manuscript assistance were provided by J. C. Witte, T. L. Kucsera (SSAI at NASA/Goddard), A. V. Cresce (University of Maryland), and J. R. Ziemke (SCA at NASA/Goddard).

REFERENCES

Andreae, M. O., Biomass burning: Its history, use and distribution and its impact on environmental quality and global climate, in J. S. Levine (Ed.), *Global Biomass Burning: Atmospheric, Climatic and Biospheric Implications*, MIT Press, Massachusetts, 1991, pp. 3–21.

Artaxo, P., E. T. Fernandes, J. V. Martins, M. A. Yamasoe, P. V. Hobbs, W. Maenhaut, K. M. Longo, and A. Castanho, Large-scale aerosol source apportionment in Amazonia, *J. Geophys. Res.*, *103*, 31837–31848, 1998.

Baldy S., G. Ancellet, M. Bessafi, A. Badr, and D. Lan Sun Luk, Field observations of tropospheric vertical distribution of tropical ozone at a remote marine site in the southern hemisphere, *J. Geophys. Res.*, *101*, 23835–23849, 1996.

Blake, N. J., D. R. Blake, O. W. Wingenter, B. C. Sive, L. M. McKenzie, J. P. Lopez, I. J. Simpson, H. E. Fuelberg, G. W. Sachse, B. E. Anderson, G. L. Gregory, M. A. Carroll, G. M. Albercook, and F. S. Rowland, Influence of southern hemispheric biomass burning on mid-tropospheric distributions of nonmethane hydrocarbons and selected halocarbons on the remote South Pacific, *J. Geophys. Res.*, *104*, 16213–16232, 1999.

Cahoon, Jr., D. R., B. J. Stocks, J. S. Levine, W. R. Cofer III, and K. P. O'Neill, Seasonal distribution of African savanna fires, *Nature*, *359*, 812–815, 1992.

Chandra, S., J. R. Ziemke, and R. W. Stewart, An 11-year solar cycle in tropospheric ozone from TOMS measurements, *Geophys. Res. Lett.*, *26*, 185–188, 1999.

Chatfield, R. B., J. A. Vastano, H. B. Singh, and G. W. Sachse, A general model of how fire emissions and chemistry produce African/Oceanic plumes (O_3, CO, PAN, smoke) seen in TRACE-A, *J. Geophys. Res.*, *101*, 24279–24306, 1996.

Chatfield, R. B., J. A. Vastano, L. Li, G. W. Sachse, and V. S. Connors, The Great African plume from biomass burning: Generalizations from a three-dimensional study of TRACE A carbon monoxide, *J. Geophys. Res.*, *103*, 28059–28077, 1998.

Christopher, S. A., C. Joyce, and R. M. Welsh, Satellite investigations of fire, smoke, and carbon monoxide during April 1994 MAPS mission: Case studies over tropical Asia, *J. Geophys. Res.*, *103*, 19327–19336, 1998.

Cros, B., D. Nganga, A. Minga, J. Fishman, and V. Brackett, Distribution of tropospheric ozone at Brazzaville, Congo, determined from ozonesonde measurements, *J. Geophys. Res.*, *97*, 12869–12875, 1992.

Crutzen, P. J., and J. G. Goldammer, *Fire in the Environment: The Ecological, Atmospheric, and Climatic Importance of Vegetation Fires: Report of the Dahlem Workshop*, Wiley, New York, 1993.

Fenn M. A., E. V. Browell, C. F. Butler, W. B. Grant, S. A. Kooi, M. B. Clayton, G. L. Gregory, R. E. Newell, Y. Zhu, J. E. Dibb, H. E. Fuelberg, B. E. Anderson, A. R. Bandy, D. R. Blake, J. D. Bradshaw, B. G. Heikes, G. W. Sachse, S. T. Sandholm, H. B. Singh, and R. W. Thornton, Ozone and aerosol distributions and air mass characteristics over the South Pacific during the burning season, *J. Geophys. Res.*, *104*, 16197–16212, 1999.

Fishman, J., V. G. Brackett, and K. Fakhruzzaman, Distribution of tropospheric ozone in the tropics from satellite and ozonesonde measurements, *J. Atmos. Terr. Phys.*, *54*, 589–597, 1992.

Fishman, J., and V. G. Brackett, The climatological distribution of tropospheric ozone derived from satellite measurements using version 7 Total Ozone Mapping Spectrometer and Stratospheric Aerosol and Gas Experiment data set, *J. Geophys. Res.*, *102*, 19275–19278, 1997.

Fishman, J., P. Minnis, and H. G. Reichle, Use of satellite data to study tropospheric ozone in the tropics, *J. Geophys. Res.*, *91*, 14451–14465, 1986.

Fishman, J., C. E. Watson, J. C. Larsen, and J. A. Logan, The distribution of tropospheric ozone determined from satellite data, *J. Geophys. Res.*, *95*, 3599–3617, 1990.

Fujiwara, M., K. Kita, S. Kawakami, T. Ogawa, N. Komala, S. Saraspriya, and A. Suripto, Tropospheric ozone enhancements during the Indonesian forest fire events in 1994 and in 1997 as revealed by ground–based operations, *Geophys. Res. Lett.*, *26*, 2147–2420, 1999.

Garstang, M. and P. D. Tyson, Atmospheric circulation, vertical structure and transport, in B. van Wilgen, M. Andreae, J. Goldammer, and J. Lindesay (Eds.), *Fire Southern African*

Savanna: Ecological and Atmospheric Perspectives, University of Witwatersrand Press, Johannesburg, 1997, Chapter 6.

Garstang M., P. D. Tyson, R. J. Swap, M. Edwards, P. Kållberg, and J. A. Lindesay, Horizontal and vertical transport of air over southern Africa, *J. Geophys. Res.*, *101*, 23721–23736, 1996.

Granier, C., W-M. Hao, G. Brasseur, and J-F. Müller, Land-use practices and biomass burning: Impact on the chemical composition of the atmosphere, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Massachusetts, 1996, pp. 140–148.

Guo, Z., and R. B. Chatfield, Meteorology of the Southern Global Plume: African and South American fires pollute the south Pacific, paper presented at the Sixth International Conference on Atmospheric Sciences and Application to Air Quality, Beijing, November 3–5, 1998.

Harris G. W., F. G. Wienhold, and T. Zenker, Airborne observations of strong biogenic NO_x emissions from the Namibian savanna at the end of the dry season, *J. Geophys. Res.*, *101*, 23707–23711, 1996.

Hoell J. M., D. D. Davis, D. J. Jacob, M. O. Rodgers, R. E. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM-Tropics A, August–September 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hsu, C. N., J. R. Herman, O. Torres, B. N. Holben, D. Tanre, T. F. Eck, A. Smirnov, B. Chatenet, and F. Lavenu, Comparisons of the TOMS aerosol index with Sun-photometer aerosol optical thicknesss: Results and applications, *J. Geophys. Res.*, *104*, 6269–6279, 1999.

Hudson, R. D., and A. M. Thompson, Tropical tropospheric ozone (TTO) from TOMS by a modified-residual method, *J. Geophys. Res.*, *103*, 22129–22145, 1998.

Jacob, D. J., and S. C. Wofsy, Photochemistry of biogenic emissions over the Amazon forest, *J. Geophys. Res.*, *93*, 1477–1486, 1988.

Jacob, D. J., B. G. Heikes, S. M. Fan, J. A. Logan, D. L. Mauzerall, J. D. Bradshaw, H. B. Singh, G. L. Gregory, R. W. Talbot, D. R. Blake, and G. W. Sachse, Origin of ozone and NO_x in the tropical troposphere: A photochemical analysis of aircraft observations over the South Atlantic Basin, *J. Geophys. Res.*, *101*, 24235–24250, 1996.

Jonquières, I., and A. Marenco, Redistribution by deep convection and long-range transport of CO and CH_4 emissions from the Amazon basin, as observed by the airborne campaign TROPOZ II during the wet season, *J. Geophys. Res.*, *103*, 19075–19091, 1998.

Jonquières, I., A. Marenco, A. Maalej, and F. Rohrer, Study of ozone formation and transatlantic transport from biomass burning emissions over West Africa during the airborne Tropospheric Ozone Campaigns TROPOZ I and TROPOZ II, *J. Geophys. Res.*, *103*, 19059–19073, 1998.

Justice, C. O., J. D. Kendall, P. R. Dowty, and R. J. Scholes, Satellite remote sensing of fires during the SAFARI campaign using NOAA-advanced very high resolution radiometer data, *J. Geophys. Res.*, *101*, 23851–23863, 1996.

Kim, J.-H., and M. J. Newchurch, Climatology and trends of tropospheric ozone over the Eastern Pacific ocean, *Geophys. Res. Lett.*, *23*, 3,723–3,726, 1996.

Krishnamurti, T. N., M. C. Sinha, M. Kanamitsu, D. Oosterhof, H. Fuelberg, R. Chatfield, D. J. Jacob, and J. Logan, Passive tracer transport relevant to the TRACE-A experiment, *J. Geophys. Res.*, *101*, 23889–23907, 1996.

Lelieveld, J., P. J. Crutzen, D. Jacob, and A. M. Thompson, Modeling of biomass burning influences on tropospheric ozone, in B. W. van Wilgen (Ed.), *Fire in the Southern Africa*

Savannas: Ecological and Atmospheric Perspectives, University of Witwatersrand Press, Johannesburg, 1997, Chapter 10.

Levine, J. S., *Biomass Burning: Atmospheric. Climatic and Biospheric Implications*, MIT Press, Cambridge, MA, 1991.

Levine, J. S., *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996.

Liew, S. C., L. K. Kwo, K. Padmanabhan, O. K. Lim, and H. Lim, Delineating land/forest fire burnt scars with ERS interferometric synthetic aperture radar, *Geophys. Res. Lett.*, *26*, 2409–2412, 1999.

Longo, K. M., A. M. Thompson, V. W. J. H. Kirchhoff, L. A. Remer, S. R. de Freitas, M. A. F. S. Dias, P. Artaxo, W. Hart, J. D. Spinhirne, and M. A. Yamasoe, Correlation between smoke and tropospheric ozone concentration in Cuiabã during Smoke, Clouds, and Radiation-Brazil (SCAR-B), *J. Geophys. Res.*, *104*, 12113–12129, 1999.

Mauzerall, D. L., J. A. Logan, D. J. Jacob, B. E. Anderson, D. R. Blake, J. D. Bradshaw, B. Heikes, G. W. Sachse, H. Singh, and B. Talbot, Photochemistry in biomass burning plumes and implications for tropospheric ozone over the tropical South Atlantic, *J. Geophys. Res.*, *103*, 8401–8423, 1998.

Newell, R. E., V. Thouret, J. Y. N. Cho, P. Stoller, A. Marenco, and H. G. Smit, Ubiquity of quasi-horizontal layers in the troposphere, *Nature*, *398*, 316–319, 1999.

Olson, J. R., B. A. Baum, D. R. Cahoon, and J. H. Crawford, Frequency and distribution of forest, savanna and crop fires over tropical region during PEM-Tropics A, *J. Geophys. Res.*, *104*, 5865–5876, 1999.

Oltmans, S. J., A. S. Lefohn, H. E. Scheel, J. M. Harris, H. Levy, I. E. Galbally, E. G. Brunke, C. P. Meyer, J. A. Lathrop, B. J. Johnson, D. S. Shadwick, E. Cuevas, F. J. Schmidlin, D. W. Tarasick, H. Claude, J. B. Kerr, and O. Uchino, Trends of ozone in the troposphere, *Geophys. Res. Lett.*, *25*, 139–142, 1998.

Pickering, K. E., A. M. Thompson, Y. Wang, W-K Tao, D. P. McNamara, V. W. J. H. Kirchhoff, B. G. Heikes, G. W. Sachse, J. D. Bradshaw, G. L. Gregory, and D. R. Blake, Convective transport of biomass burning emissions over Brazil during TRACE-A, *J. Geophys. Res.*, *101*, 23993–24012, 1996.

Randriambelo, T., J. L. Baray, S. Baldy, P. Bremaud, and S. Cautenet, A case study of extreme tropospheric ozone contamination in the tropics using in-situ, satellite, and meteorological data, *Geophys. Res. Lett.*, *26*, 1287–1290, 1999.

Reichle, H. G., V. S. Connors, J. A. Holland, R. T. Sherrill, H. A. Wallio, J. C. Casas, E. P. Condon, B. B. Gormsen, and W. Seiler, The distribution of middle tropospheric carbon-monoxide during early October 1984, *J. Geophys. Res.*, *95*, 9845–9856, 1990.

Schultz, M. G., D. J. Jacob, Y. H. Wang, J. A. Logan, E. L. Atlas, D. R. Blake, N. J. Blake, J. D. Bradshaw, E. V. Browell, M. A. Fenn, F. Flocke, G. L. Gregory, B. G. Heikes, G. W. Sachse, S. T. Sandholm, R. E. Shetter, H. B. Singh, and R. W. Talbot, On the origins of tropospheric ozone and NO_x, over the tropical South Pacific, *J. Geophys. Res.*, *104*, 5829–5844, 1999.

Smit, H., D. Kley, S. McKeen, A. Volz, and S. Gilge, The latitudinal and vertical distribution of tropospheric ozone over the Atlantic Ocean in the southern and northern hemispheres, in R. D. Bojkov and P. Fabian (Eds.), *Ozone in the Atmosphere*, 1989, pp. 419–422.

Smyth, S. B., S. T. Sandholm, J. D. Bradshaw, R. W. Talbot, D. R. Blake, N. J. Blake, F. S. Rowland, H. B. Singh, G. L. Gregory, B. E. Anderson, G. W. Sachse, J. E. Collins, and A. S. Bachmeier, Factors influencing the upper free tropospheric distribution of reactive nitrogen over the South Atlantic during the TRACE A experiment, *J. Geophys. Res.*, *101*, 24165–24186, 1996.

Swap, R. J., M. Garstang, S. A. Macko, P. D. Tyson, and P. Kållberg, Comparison of biomass burning emissions and biogenic emissions to the tropical south Atlantic, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996, pp. 396–402.

Thompson, A. M., B. G. Doddridge, J. C. Witte, R. D. Hudson, W. T. Luke, J. E. Johnson, B. J. Johnson, and S. J. Oltmans, Shipboard and satellite views of elevated tropospheric ozone over the tropical Atlantic in January–February 1999, *Geophys. Res. Lett.*, *22*, 3317–3320, 2000.

Thompson, A. M., and R. D. Hudson, Tropical tropospheric ozone (TTO) maps from Nimbus 7 and Earth-Probe TOMS by the modified-residual method: Evaluation, El Niño signals and trends based on Atlantic regional time series, *J. Geophys. Res.*, 26961–26975, 1999.

Thompson, A. M., K. E. Pickering, D. P. McNamara, M. R. Schoeberl, R. D. Hudson, J. H. Kim, E. V. Browell, V. W. J. H. Kirchhoff, and D. Nganga, Where did tropospheric ozone over southern Africa and the tropical Atlantic come from in October 1992? Insights from TOMS, GTE/TRACE-A and SAFARI-92, *J. Geophys. Res.*, *101*, 24251–24278, 1996.

Thompson, A. M., W-K. Tao, K. E. Pickering, J. R. Scala, and J. Simpson, Tropical deep convection and ozone formation, *Bull. Am. Meteorol. Soc.*, *78*, 1043–1054, 1997.

Tyson, P. D., M. Garstang, A. M. Thompson, P. D'Abreton, R. D. Diab, and E. V. Browell, Atmospheric transport and photochemistry of ozone over central Southern Africa during the Southern Africa Fire-Atmosphere Research Initiative, *J. Geophys. Res.*, *102*, 10623–10635, 1997.

van Wilgen, B. W., M. O. Andreae, J. G. Goldammer, and J. A. Lindesay, *Fire in the Southern Africa Savannas: Ecological and Atmospheric Perspectives*, Witwatersand University Press, Johannesburg, South Africa, 1997.

Weller, R., R. Lilischkis, O. Schrems, R. Neuber, and S. Wessel, Vertical ozone distribution in the marine atmosphere over the central Atlantic Ocean (56°S–50°N), *J. Geophys. Res.*, *101*, 1387–1399, 1996.

Zenker, T., A. M. Thompson, D. P. McNamara, T. L. Kucsera, F. G. Wienhold, G. W. Harris, P. LeCanut, M. O. Andreae, and R. Koppmann, Regional trace gas distribution and airmass characteristics in the haze layer over southern Africa during the biomass burning season (Sep./Oct. 1992): Observations and modeling from the STARE/SAFARI-92/DC-3, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996, pp. 296–308.

Zepp, R. G., W. L. Miller, R. A. Burke, D. A. B. Parsons, and M. C. Scholes, Effects of moisture and burning on soil-atmosphere exchange of trace carbon gases in a southern African savanna, *J. Geophys. Res.*, *101*, 23699–23706, 1996.

Ziemke, J. R., S. Chandra, and P. K. Bhartia, Two new methods for deriving tropospheric column ozone from TOMS measurements: The assimilated UARS MLS/HALOE and convective-cloud differential techniques, *J. Geophys. Res.*, *103*, 22115–22128, 1998.

Ziemke, J. R., S. Chandra, A. M. Thompson, and D. P. McNamara, Zonal asymmetries in southern hemisphere column ozone: Implication of biomass burning, *J. Geophys. Res.*, *101*, 14421–14427, 1996.

CHAPTER 15

ACID RAIN AND DEPOSITION

WILLIAM B. GRANT

1 INTRODUCTION

Early Concern

The first mention of acid rain in print was by Robert Boyle in which he referred to "nitrous or salino-sulphureous spirits" in air in his 1692 book *A General History of the Air*. The Scottish chemist Robert Angus Smith began to study acid rain in Manchester, England, in 1852 and extended the work in England, Scotland, and Germany for 20 years. His 1872 book, *Air and Rain: The Beginnings of a Chemical Climatology*, pointed out the link between sulfur pollution and "acid rain." He warned that acid rain was damaging plants and materials downwind of industrial regions, but his warning went largely unheeded.

While some research was conducted on acid deposition in the ensuing years, it was not until the 1950s and 1960s that E. Gorham, conducting research in England and Canada, built the major foundations for our present understanding of the causes of acid precipitation and its impact on aquatic ecosystems. However, it took the work of a Swedish scientist, S. Oden, in the 1960s, to arouse the scientific community and general public to engage in the debate about acid deposition. One newspaper account described his ideas about an insidious "chemical war" among the nations of Europe. Thus, by the 1970s, it was finally realized that Eastern Europe, Germany, Scandinavia, Canada, and the United States were experiencing widespread damage to forests and lakes as well as damage to stone and metal buildings and other structures from acid rain. In Germany, the term *Waldsterben* (forest death) was coined. Forests in parts of the Czech Republic, Slovakia, and Russia were practically devastated due to acid rain and heavy-metal ion deposition from uncontrolled industrial and power plant emissions. China and India are also experiencing significant effects of acid

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

rain, with the Taj Mahal losing much of its stonework surface material to acid deposition.

Acid Rain Chemistry

Acid rain is actually precipitation of various ions, both anions and cations, through precipitation, such as rain, snow, fog, as well as dry particles or aerosols. A typical ion balance is

$$\begin{aligned}&[H^+] + [Na^+] + [Na_4{}^+] + 2[Ca^{2+}]\\ &\quad = 2[SO_4{}^{2-}] + 2[SO_3{}^{2-}] + [NO_3^-] + [Cl^-] + [OH^-] + [HCO_3{}^-] + 2[CO_3{}^{2-}] \quad (1)\end{aligned}$$

The primary naturally occurring trace gas that affects the pH of precipitation is carbon dioxide (CO_2), which forms carboxylic acid in water. The aqueous reactions of carbon dioxide are as follows:

$$CO_2 \text{ gas} + H_2O \rightarrow H_2CO_3 \quad (2)$$

$$H_2CO_3 \rightarrow HCO_3{}^- + H^+ \quad (3)$$

$$HCO_3{}^- \rightarrow CO_3{}^{2-} + H^+ \quad (4)$$

Since pK_a of (4) is as high as 10.3, reaction (2) has the greatest influence on the acidity of natural atmospheric systems. For a partial pressure of CO_2 of 350 ppmv, Henry's law constant (K_H) is as follows:

$$K_H = [H_2CO_3]/[CO_2 \text{ gas}] = 3.97 \times 10^{-2} \text{ mol/L atm} \quad (5)$$

and the equilibrium constant (K_3) of reaction (3) is given by:

$$K_3 = [H^+][HCO_3{}^-]/[H_2CO_3] = 4.5 \times 10^{-7} \text{ mol/L} \quad (6)$$

By combining and rearranging these two expressions, one arrives at the following equation:

$$[HCO_3{}^-] = ([CO_2 \text{ gas}] \times K_H \times K_3)/[H^+] \quad (7)$$

If the concentration of bicarbonate in water is equal to the hydrogen ion concentration, then by substitution, one arrives at the following:

$$[H^+]^2 = ([CO_2 \text{ gas}] \times K_H \times K_3)/[H^+] \quad (8)$$

$$= 5.97 \times 10^{-12} \text{ mol}^2/\text{L}^{-2} \quad (9)$$

Therefore,

$$[\mathrm{H}^+] = 2.44 \times 10^{-6}\ \mathrm{mol/L} \tag{10}$$

and, hence

$$\mathrm{pH} = -\log[\mathrm{H}^+] = 5.6 \tag{11}$$

The bracketed quantities denote molar concentrations, with the cations on the left, the anions on the right.

Note that Henry's law can be expressed in terms of a pseudo-Henry's law constant to account for the increased uptake of gas in the liquid due to reactions in the liquid. For example,

$$K_{\mathrm{CO}_2^*} = [\mathrm{CO}_2 \times \mathrm{H}_2\mathrm{O} + \mathrm{HCO}_3{}^{-}\mathrm{CO}_3{}^{2-}]/P_{\mathrm{CO}_2} \tag{12}$$

However, 5.6 is not necessarily the natural pH of rain since other naturally occurring species also play a role. Nitrogen oxides are formed naturally during lightning discharges, and sulfur species are released into the atmosphere over the oceans from biological activity as dimethyl sulfide (DMS). Hydrocarbon acids such as carboxylic acids, HCOO_t, and methylcarboxylic acids, $\mathrm{CH_3COO}_t$, also contribute to the acidity, especially in remote, forested regions. On the other hand, base cations from soil dust, such as Ca, Mg, K, and P, etc., are alkaline and increase the pH. Thus, the natural acidity of precipitation can vary considerably depending on the upwind sources and, as will be discussed, meteorological conditions.

In addition, pollutants such as the nitrogen and sulfur oxides also contribute to acidity and are the focus of most of the concern regarding acid deposition. However, ammonia, often associated with agricultural operations, is alkaline. This chapter will examine the sources, the chemical transformations involved in the production of acid deposition, transport, deposition amounts and trends, and the effects on soils, plants, animals, and materials.

2 SOURCES

Natural

To put pollution contributions into perspective, it is worthwhile first to understand the role that naturally occurring materials play in acid deposition. Natural sources of sulfur account for 25 to 30% of the total, unless there are large volcanic eruptions, such as El Chichon in 1982 or Mount Pinatubo in 1991. Mount Pinatubo was estimated to emit 9 Tg of S into the stratosphere (total sources are 94 to 123 Tg S/yr), where the e^{-1} residence time for sulfuric acid aerosols is approximately one year.

Oxides of nitrogen ($NO_x = NO + NO_2$) are also produced naturally. As discussed in Chapter 4, natural sources such as soil emissions, lightning, stratospheric–tropospheric exchange, and a portion of biomass burning account for approximately one third of total NO_x.

Hydrocarbons are also involved in acid deposition. Carboxylic acids, $HCOO_t$, and methylcarboxylic acids, CH_3COO_t, are important hydrocarbon acids derived from direct terrestrial emissions as well as oxidation of emissions by marine or terrestrial biota.

Base cations are generally derived from soils through lofting of aeolian dust by wind. Deserts, such as the Sahara Desert in Africa and the Gobi Desert in China, generate large dust clouds each year that are transported thousands of kilometers. The dust from the Sahara Desert often reaches both North and South America and may provide significant base cations for vegetation in the rain forests. The dust from the Gobi Desert is often seen over Japan and the Korean Peninsula. The base cation deposition in Europe was studied for 1989. Using a $10 \times 20\ \mathrm{km}^2$ grid, maps were produced showing that base cations neutralize $SO_4{}^{2-} + NO_3{}^-$ by much more in southern regions than in northern regions. South of 45° to 50°N, more than 50% was neutralized, with more than 75% in some locations; in Norway and Sweden, the amount neutralized generally fell to less than 10%. The variations can be explained in terms of the amounts of acid ions and base cations in the air. Soil-derived dust in the United States used to provide the base cations to help neutralize the effects of sulfur and nitrogen. However, the amount of base cations in precipitation has been declining in the past 2 to 3 decades in the United States probably because of changes in farming and construction practices that leave fewer disturbed regions from which wind can raise dust.

Anthropogenic

Anthropogenic sources of sulfur accounted for 77% of global sulfur emissions in 1980. Combustion of fossil fuel for electric power production is responsible for most of the anthropogenic contributions to acid deposition, accounting for 67% of the anthropogenic SO_2 emissions in the United States in 1996. Industrial fuel combustion accounts for 17% of the U.S. SO_2 emissions, with various other sources accounting for the rest. Fuel used for transportation generates 7% of the SO_2 emissions, which has been linked to regional haze patterns in such places as the Los Angeles Basin and portions of the eastern United States.

NO is a by-product of combustion of all hydrocarbon fuels, both fossil fuel and fresh biomass, due to the high temperatures involved. In the United States in 1996, 30% of the NO emissions came from on-road vehicles, 28% from electric utilities, 19% from nonroad engines and vehicles, 13% from industrial fuel combustion, and 10% from other sources. On a global basis, anthropogenic NO emissions are highest where industry, fossil fuel power plants, and surface transportation are most densely sited, i.e., the northern midlatitudes.

Anthropogenic ammonia emissions are associated with fertilizers and livestock feedlots. Organic acids also contribute to the anthropogenic burden of acid deposition. The major organic acids found in the gas phase are formic acid (HCOOH) and acetic acid (CH_3COOH), with other organic acids found in minor amounts. Sources include automobile exhaust, biomass burning, and some food processing plants.

3 TRANSFORMATION

The emitted SO_2, NO, and NH_3 are transformed to aerosols and components of precipitation through both gas-phase and aqueous-phase chemical reactions.

Sulfur dioxide is transformed in the gas phase primarily by:

$$SO_2 + OH \rightarrow HOSO_2 \tag{13}$$

Ozone can also lead to the oxidization of SO_2. Such reactions would be especially important at night when OH radical concentrations are very small due to the absence of solar radiation. One way this can happen is for ozone to react with an alkene, such as ethene or propene by adding to the carbon double bond, creating a primary ozonide. Since ozonides are not stable, this can rapidly split into what is called a Criegee intermediate, named after the German chemist who proposed the mechanism. A Criegee intermediate can react with SO_2 in a series of steps that also result in the oxidation of SO_2, which can also be oxidized directly by ozone, but the reaction rate is slow. Note that the rate of oxidation of SO_2 has a seasonal cycle in middle latitudes, being as much as an order of magnitude lower in winter than in summer.

In the aqueous phase, other reactions can occur. For example:

$$SO_2 + H_2O \rightarrow SO_2 \times H_2O \tag{14}$$

$$SO_2 \times H_2O \rightarrow HSO_3^- + H^+ \tag{15}$$

$$HSO_3^- \rightarrow SO_3^{2-} + H^+ \tag{16}$$

These reactions establish equilibria of the various sulfur species, with mole fractions dependent on the pH of the solution and both Henry's law constant [for (14)] and equilibrium constants [for (15) and (16)]. Dissolved SO_2 (13) is favored at pH below 2, the bisulfite ion (15) for $2 < pH < 7$, and the sulfite ion (16) for $pH > 7$.

Aqueous-phase reactions with H_20_2 in cloud, fog, and raindrops are considered to be the dominant mechanisms for the oxidation of SO_2 to H_2SO_4. Thus, H_20_2 could be rate limiting. Field and modeling studies show that to explain the seasonal concentrations of H_20_2 (higher in summer than in winter) the initial rate of aqueous phase H_20_2 photoformation has to be linearly dependent on solar actinic flux, i.e., radiation that induces photochemical reactions. Organic chromophores are suggested to be responsible for the H_20_2 photoformation. One implication of this study is that the seasonal variability in the nonlinearity between SO_2 emissions and regional sulfate deposition may be largely explained. Other peroxides can also oxidize SO_2, but exist in lower concentrations than does H_20_2.

Other reactions leading to the oxidation of S(IV) include ozone, O_2 catalyzed by transition metal ions such as Fe^{3+} and Mn^{2+}, and carbonaceous particles. While the reactions with H_2O_2 are generally most important (2 to 20% per hour, independent of pH), the others are much weaker in general and have very strong pH dependences. Above a pH of 5, the reaction with ozone is comparable to that for H_2O_2, with the other reactions somewhat weaker.

Note that other sulfur species, such as hydrogen sulfide (H_2S) and carbonyl sulfide (OCS), emitted by biological sources, can also be oxidized, as well as dimethyl sulfide (DMS), emitted from marine sources. While OH is the primary source of DMS oxidation, NO_3 also reacts rapidly with DMS, and halogens, such as bromine, chlorine, and iodine are also potential reactants with DMS in the marine boundary layer.

Nitric oxide (NO) is rapidly oxidized to NO_2, especially by reacting with ozone:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{17}$$

From there, it is transformed to nitric acid by interaction with the hydroxyl radical:

$$NO_2 + OH \rightarrow HNO_3 \tag{18}$$

This reaction is about 10 times more rapid than that of (13).

Nitric acid can also be formed by the reaction with various organics, such as the alkanes and aldehydes. In this case, hydrogen is abstracted from the organic molecule. This reaction may account for 15% of the nitric acid formation, occurring primarily at night.

Both sulfate and nitrate aerosols are very hygroscopic and increase in diameter rapidly with increases in relative humidity above 50% to 70%. In the absence of cloud formation, they form the bulk of regional aerosols downwind of heavily industrialized/urbanized regions, such as the eastern United States. As acid haze becomes thicker and stays near the surface, it can become acid fog, such as has been observed in California and in eastern U.S. mountains. An aerosol/fog cycle can be set up in which aerosol particles grow by water condensation on existing nuclei, dilute and dissolve in fog droplets, where they undergo chemical conversions. The process can go the other way as solute concentrations increase due to evaporation of the water, leading back to aerosols. Thus, as the temperature cycles during the day, the fog–aerosol–fog cycle can be made.

An excellent overview of the chemistry of acid precipitation can be found in Finlayson-Pitts and Pitts (2000).

4 TRANSPORT

In addition to source regions and transformation mechanisms and rates, winds and other meteorological conditions also play important roles in determining where acid precipitation will occur. The pollution plumes will be transported at the rate of the prevailing winds. The source gases will be transformed at various rates depending on such factors as amount of solar radiation, concentrations of OH and water vapor, temperature, and the extent of clouds.

Sulfate can be transported up to 1100 km in normal downwind directions and up to 400 km in the normal upwind directions (i.e., during the reduced opportunities for transport in that direction), while nitrogen oxides are transported as nitrates as far as 200 to 800 km. It is found that turnover times for anthropogenic sulfate are 4.7 ± 1.1 days in the eastern United States.

The transport of ammonia and ammonium depends on the emissions of SO_2 and NO_x along the trajectory of the air mass containing them. The transport distance for ammonia and ammonium in northern Europe depends on the amount of SO_2 and NO_x present. When they are present, transport is reduced significantly because ammonium aerosols are formed rapidly. NH_x is most likely to be deposited in the country of origin in Europe, given the sizes of the countries, while for SO_x only 25 to 30% would be deposited, and for NO_x only 10%.

5 DEPOSITION

Acid deposition occurs in two primary forms—wet and dry. Wet deposition comprises rain, snow, and fog. Dry deposition involves turbulent transport of aerosol and gases to the surface layer. The relative amounts of wet and dry deposition depend on a number of factors, such as the amount of precipitation, whether the

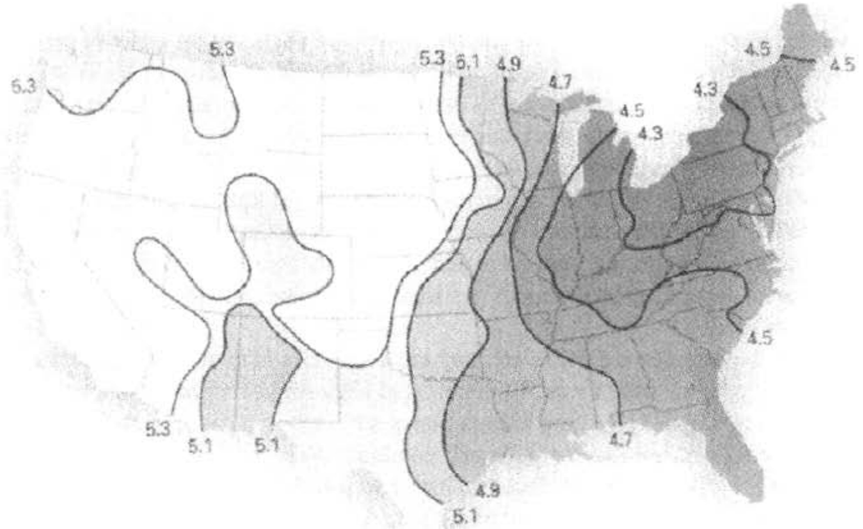


Figure 1 Annual pH of rain for the United States in 1990. The black lines indicate contours of equal pH. See ftp site for color image.

elevation is above the cloud line, how far the site is from the primary sources of the acid ions, etc. At U.S. Environmental Protection Agency (EPA) National Dry Deposition Network stations in the eastern United States in 1991, dry sulfate deposition accounted for approximately 10 to 60% (mean approximately 40%) of total sulfate deposition, with wet deposition accounting for the rest. For nitrates, the dry deposition fraction varied from 20 to 65% (mean approximately 45%). Due to the seasonal cycle in the rate of oxidation of SO_2, deposition rates for SO_2 tend to be higher than for SO_4 in winter, with the reverse occurring in summer.

The acidity of deposition depends on the difference between anions and cations in the precipitate. Thus, the nitrate and sulfate ions reduce the pH, while ammonium and soil-derived dust increase the pH. Figure 1 shows a map of the pH of rain for the United States in 1990, indicating that the pH is lowest just southeast of the Great Lakes, a consequence of the high amount of fossil fuel combustion in and to the west of the region.

6 MEASUREMENT

Instruments

Various instruments are used in the study of acid deposition. Since emission rates are generally estimated based on factors associated with fuel consumption, not many measurements are made at the source regions. Standard meteorological instruments and networks are used for the meteorological data input. The collectors generally use polypropylene funnels and bottles. The bottles may be refrigerated to 4°C to reduce evaporation and/or heated to melt snow. When wet and dry deposition collectors are used together, a lid is placed over the dry deposition bucket at the onset of precipitation, then back over the wet deposition bucket at the end of precipitation. However, it should be noted that measurement of dry deposition is notoriously difficult, and that buckets do not adequately represent the manner in which the local surfaces collect dry deposition. The three conceptual ways in which dry deposition is measured are: (1) direct collection on surrogate or natural surfaces, (2) flux measurements by eddy correlation or profile techniques, and (3) indirect estimation using atmospheric concentration monitoring and estimated deposition velocities. Which approach is used varies depending on the funds available and the accuracy to which the information is desired.

Once the samples are collected, they are taken to a laboratory for analysis. The analytical methods used by the National Acid Deposition Program/National Trends Network (NADP/NTN) in the United States are likely typical of such programs. A glass electrode is used to measure pH; conductivity is measured using a platinum electrode; chloride, nitrate, orthophosphate and sulfate are measured with ion chromatrography with a detection limit of 0.03 mg/L for all but orthophosphate, which is measured with a detection limit of 0.02 mg/L; ammonium is measured using automated phenate colorimetry with a detection limit of 0.02 mg/L; calcium, magnesium, potassium, and sodium are measured with flame atomic absorption spectrophotometer with a detection limit of 0.003 except for calcium, for which the detection limit is 0.09 mg/L. Sodium and/or magnesium can be used to estimate the fraction of material derived from sea salt. This is useful in determining how to apportion the sulfate values between terrestrial and oceanic sources.

The Acid Precipitation in Ontario Study (APIOS) deposition monitoring program has similar instrumentation with slightly different detection limits. The NADP/NTN detection limits were improved by instrument changes in 1985, while the APIOS instrumentation was established in 1980 and not updated as of 1990.

Surface networks

Collection instruments are often set out in networks. The NADP/NTN is an example of such a network. It is part of a cooperative program that includes federal, state, and private research organizations. The objectives of the program are:

1. To measure and characterize the supply of beneficial and injurious chemical substances in atmospheric deposition on a broad regional scale
2. To determine the spatial patterns and temporal trends in the distribution of chemical elements deposited on natural and managed ecosystems
3. To provide information needed to gain a better understanding of the sources, transport, and transformation of materials contributing to or associated with acidic atmospheric deposition in the United States

The NADP/NTN was made operational in July 1978 and continues to the present time. The sites were selected to represent major physiographic, agricultural, aquatic, and forested areas throughout the United States. In general, sites are located in rural areas away from sources that could affect the measurements. The program grew from 22 sites in late 1978 to about 200 sites in 1985, which were still in operation in 1990. The containers are heated to 4°C to melt snow but are not refrigerated. Samples are collected weekly and sent to the Central Analytical Laboratory in Champaign, Illinois.

7 INTENSIVE STUDY PROGRAMS

In the 1980s, a major study, the National Acid Precipitation Assessment Program (NAPAP) was funded by Congress to investigate the situation in the United States. The total cost was $500 million. Areas of investigation included acid deposition and effects on aquatic and terrestrial ecosystems. Both nitrate and sulfate depositions were found to be highest in the northeast United States near the eastern Great Lakes, centered on eastern Michigan, western New York and Pennsylvania, and northern West Virginia, and extending into southern Ontario, albeit with somewhat different geographical distributions. Ammonium deposition peaked in Michigan and southern

Ontario. As a consequence, annual pH of rain is lowest in New York, Pennsylvania, and West Virginia as shown in Figure 1 for 1990.

Similar programs have been carried out in a number of European countries, especially in terms of acid rain effects on forests, with a number of them reported in the Springer *Ecological Studies* series.

8 GLOBAL TRENDS IN EMISSIONS AND DEPOSITION

With accelerating economic development in Southeast Asia, anthropogenic NO_x emissions are expected to increase dramatically in the near future. It has been estimated that global NO_x emissions will increase from an estimated 19 Tg NO_2 in 1990 to 86 Tg NO_2 in 2020. The largest increases are expected in the power and transport sectors.

Trends of acid deposition should generally follow the regional trends for fossil fuel consumption, with coal and oil providing most of the sulfur, and all components contributing to the nitrogen oxides and organic acids. In the United States, wood was the primary source of fuel until 1880, being used to generate about 3×10^{15} Btu/yr at the peak in 1870. Coal started to be used in increasing amounts around 1850, rising to 15×10^{15} Btu/yr by 1916, staying in the range 10 to 17×10^{15} Btu/yr after that. Oil started to become an important fuel source after 1900, rising to 35×10^{15} Btu/yr by 1977 before leveling off. Natural gas also became important after 1900, rising to 24×10^{15} Btu/yr by 1970 before dropping slightly. Thus, in the United States, acid deposition should have risen steadily from 1900 to at least the 1980s. In the eastern and midwestern United States there has been an estimated 19% decrease in SO_2 emissions and a 16% decrease in NO_x emissions between 1975 and 1987. Since the U.S. Clean Air Act Amendments of 1990 mandated further decreases in sulfur emissions, they have continued to decrease. Between 1989 and 1995, sulfur dioxide decreased 35% and sulfate 26% in rural eastern United States. Nitrogen emissions have not been recognized as being very important until recently for a variety of scientific and political reasons, and it is more difficult to remove NO_x than SO_2 from the flue gases, so the regulations on nitrogen emissions are not as strong as for sulfur. Between 1989 and 1995 nitrogen concentrations in rural eastern United States had fallen only 8%.

Data for historical anthropogenic emissions of SO_2 are also available for Europe. A gradual increase is seen from 1880 (0.45 Tg/yr) to 1940 (1.4 Tg/yr), a dip in 1945, then a rapid increase to >36 Tg/yr in 1980, followed by a gradual decline thereafter. Ammonia emissions peaked in the mid-1980s.

Continued population growth and development are expected to lead to an increase of 25% in the deposition of nitrogen in the more-developed-country regions by the year 2020. Earth's population is projected to increase from 6 billion in 1999 to 8.5 billion in 2020, and per-capita energy consumption is expected to double compared to 1980. Much of the increase will be felt in Asia. The increases in nitrogen oxides may lead to larger ozone concentrations, thereby increasing the oxidizing capacity of the atmosphere and its ability to absorb thermal infrared radiation.

9 SOIL CHANGES

Bernhard Ulrich is credited with determining how acid deposition affects soil during the acidification process. His 1966 study set the stage for his later work. His review summarizes the effects of acid deposition on soil cation-anion budgets and lists a number of his key works. As soil acidity increases due to acid deposition (or plant biomass harvesting for that matter), the base cations (e.g., Ca, Mg, K, P) try to neutralize the acidity and are leached from the upper soil horizons in the process. As the process continues, the transition metal and aluminum oxides are dissolved, with these cations becoming more prevalent in the soil solution. Nitric acid is a stronger acid than sulfuric, so it has a greater ability to lower the soil pH. An interesting recent finding is that as the process continues, Al^{3+} seems to accelerate the base cation leaching process, making Al^{3+} more readily available. As acid deposition continues over a long period, the acid neutralizing capacity (ANC) or alkalinity decreases.

Additional influences on ANC arise from biogeochemical processes. Trees, for example, enhance the collection of dry deposition as well as remove base cations from the soil. Soil organic matter storage is followed by decay, which releases the trace minerals, nitrogen, and organic acids. In addition, forest defoliation by the gypsy moth has exacerbated the effects of acidic deposition. Changes in stream water composition following severe defoliation of forested mountain watersheds in western Virginia has included increased concentrations of nitrate and acidity, as well as accelerated export of base cations, and pH and ANC reached lower levels than previously observed, especially during storm flow conditions. To date, several years following the defoliation, stream water composition has not returned to pre-defoliation values.

Finally, there are interactions between the various processes. Changing acid–base status changes vegetation amounts and types. Reductions in vegetation cover can lead to reduction in enhanced collection of dry deposition as well as higher surface temperatures, thereby increasing microbial activities.

10 EFFECTS ON FORESTS, AQUATIC ECOSYSTEMS, AND MATERIALS

Forests

Paradoxically, one of the first effects of acid deposition on trees and forests is that of stimulating growth, rather than hindering it. Nitrogen in both ammonium (NH_4) and nitrate (NO_3) forms can be utilized by trees in building amino acids required for growth. Thus, nitrogen deposition first has the impact of fertilizing plants. This process eventually ceases in temperate ecosystems when the soil is nitrogen saturated. The impact of nitrogen deposition on carbon uptake by terrestrial ecosystems has been modeled using several different three-dimensional models. Both NO_y and NH_x deposition were considered. The bulk of the NO_y deposition was found to be in the eastern United States, Europe, and, to a lesser extent, in eastern Asia and Japan. All five models predict that most of the carbon will be sequestered in the forests of eastern United States and Europe. Without N saturation, C sequesterization was found to range from 6 to 13×10^{15} g C/yr, while with N saturation, the range was 5 to 10×10^{15} g C/yr. This implies that N saturation reduces the growth rate of forests, in line with what has been observed in forests in the northeastern United States.

Acid deposition also causes the soil solution pH to be lowered, in part through the increased biomass growth rate, since the tree has to give up hydronium ions in exchange for base cations. It should be noted that the impact of acid deposition on forests is mediated through the soils, with some better able to buffer the acid than others. Calcium carbonate or limestone, for example, has a high buffering capacity, and would take a long time to show serious effects from acid deposition. One response of trees is for the tree roots to try to grow away from the acid soil, which may take the form of growing more in the upper organic layer, rather than in the lower mineral horizons. This makes trees susceptible to other stresses, such as winds and drought. Another effect is that since trees obtain less calcium after long-term acid deposition, the strength of the boles (trunks) and branches is reduced, since plants rely upon calcium for cell wall structure, they are much more susceptible to falling during ice, snow, and wind storms, as was the case in the northeastern United States and southeastern Canada in early 1998.

Starting around the 1970s, researchers in the United States and Europe began to notice that trees were beginning to show evidence of decline for nonhistorical reasons. Acid deposition was identified as a likely suspect in the early 1970s, although the effects of acid deposition had been observed in the sixteenth century in Europe and discussed again in the midnineteenth century.

The effects of acid precipitation on European forests in the 1980s have been well documented, especially to the Norway spruce [*Picea abies* (*L.*) *Karst*]. A study investigating the spruce decline determined that a long history of acid deposition, mostly sulfate prior to the early part of the century, with nitrate added around 1915, led to the observed effects. The soils were somewhat deficient in calcium and magnesium, and by about 1980, there was a strong nutritional imbalance due to years of ammonium nitrate depositions, nitrate leaching from the soils, and soil acidification. The yellowing of the leaves was attributed to deficiencies in magnesium. While *Waldsterben* in Europe was less pronounced in the early-to-mid-1990s than in the mid-1980s, probably due to reductions in sulfur emissions, declines in forest health are still quite prevalent, especially in central Europe. Annual forest condition surveys in conjunction with the modeling studies of nitrogen deposition show increased soil acidity in the regions with highest forest decline symptoms. There, the mean plot defoliation was in the range of 20 to 40% in 1997, with evergreens affected more than deciduous trees.

Acid deposition has had an adverse impact on forests in the eastern United States. The decline of the red spruce forests in the northeastern United States has been attributed to acid deposition, as has the decline of red spruce forests in North Carolina. Acid deposition has also adversely affected the sugar maples (*Acer saccharum Marsh.*) in Pennsylvania and Quebec as well as red oaks (*Quercus rubra*) and white oaks (*Quercus alba*) in the eastern United States. Evidence linking acid deposition to U.S. forest condition is found using the U.S. Department of Agriculture Forest Service Forest Inventory and Analysis data in conjunction with acid ion deposition doses using the acid deposition data from NAPAP. Increased mortality rates for white oaks (*Quercus alba*) in the northeastern United States can be related statistically to increased acid ion doses.

Further evidence for the role of acid deposition affecting oaks is found in oak tree ring studies in North Carolina and Missouri in the United States. The growth spurts in the 1950s for oaks in decline compared with lower growth rates of healthier nearby oaks are consistent with the N fertilization effect; the gradual growth decline subsequently is consistent with impaired tree vitality due to both acid deposition and ozone exposure; and the rapid decline after major droughts in the 1980s is consistent with shallower root depth, leading to greater water stress in drought periods.

Aquatic Ecosystems

Aquatic ecosystems have borne much of the brunt of acid deposition, resulting in significant loss of invertebrate populations and fish production among other things.

There are several processes influencing acid–base chemistry of surface waters. Wet and dry deposition is one. The other important factor is the ANC (alkalinity) of the water body. In turn, the ANC is strongly affected by the soils and bedrock under and near the body. The difference between the sums of base cations and acid anions derived from the soils and bedrock is equal to the ANC. Location of a body of water in a region where the soils and rocks are more likely to contribute base cations than acid anions to the water are less likely to be acidified by acid deposition. The base cations involved at the higher ANC levels are generally, in approximate order of importance, calcium, magnesium, sodium, and potassium. The acid anions are, likewise, carbonate, organic acids, sulfate, chloride, and nitrate. Of course, local conditions affect the amounts and relative orders.

Both aquatic animals and plants are adversely affected by acidification. The processes affected by acidification include change rates and amounts of primary production, nutrient cycling, and decomposition. Aluminum plays an important role in acidified systems since it is detrimental or toxic to both animal and plant life. Normally, aluminum is tightly bound to oxygen or the hydroxyl radical, OH. As the pH is lowered below 6, the concentration of monomeric aluminum rises rapidly. Aluminum in acidified streams has been found to coat the gills of fish, leading to premature mortality.

It has recently been recognized that atmospheric deposition of nitrogen is playing a significant role in the eutrophication in estuaries and coastal waters, such as the Chesapeake Bay in the mid-Atlantic eastern United States. Until a landmark study was published in 1991, it was thought that most of the nitrogen reaching such bodies of water came from agricultural operations. More recent work has determined that approximately 20% of the nitrogen reaching the Chesapeake Bay as wet precipitation is in the form of dissolved organic nitrogen. In addition, a significant fraction comes from ammonium.

Another consequence of lake acidification is increased transparency. Most likely this is due to reduction in dissolved organic carbon or from a change in the chemical nature and light absorption capacity of dissolved organics in the water. This can lead to changes in primary productivity and thermal structure at lower depths. An additional consequence of increased transparency is increased transmission of ultraviolet B (UV-B) (280 to 320 nm) radiation. This leads to reductions in abundances of phytoplankton and zooplankton sensitive to UV-B.

The geographic overview of the regional case study areas is instructive. The key factors distinguishing among the regions are geology, soils, climate, hydrology, deposition chemistry, land use, vegetation, and landforms. All play important roles in determining the degree of acidification of aquatic ecosystems. Regions with bedrock highly resistant to chemical weathering are more likely to have low ANC lakes. Calcareous bedrock leads to high ANC waters. However, if the overlying till has different properties, it can counter the influences of the bedrock. In the northeast United States, glaciers brought in till from the calcareous Canadian Shield, leading to high ANC lakes. Among soils, the younger soils, more often found in the northern United States, lead to lower ANC water bodies, while the older soils, more often found in the southeastern United States, lead to higher ANC water bodies due to the accumulated organic matter that can lead to organic acidity.

Materials

Acid deposition also affects materials such as rocks and metals used in monuments and building construction through corrosion. Calcareous rock materials such as marble and sandstone are particularly vulnerable since the base cations are leached by the acids just as in soils. Mortar from limestone is also very susceptible to damage, but bricks are largely immune to the effects. Even ancient monuments are affected in a variety of ways including removal of material; development of rusty yellow patinas rich in Fe and Cu; firmly attached black crusts in contact with percolating water, where recrystallized calcite shields amorphous deposits rich in S, Si, Fe, and carbonaceous particles; and black loose deposits of gypsum and fly ash particles. Also, metals that react with hydrogen, nitrate, or sulfate, such as copper and iron, will be slowly eroded. Modern building practices have to consider effects of acid deposition and corrosion in the design phase.

11 POLICIES

Since acid rain has adverse impacts on animals, plants, and structures, there is concern that levels be reduced from current levels in many places and not increase rapidly in developing regions where fossil fuel combustion is increasing.

After completion of the NAPAP study, but not because of it, the 1990 amendments to the Clean Air Act mandated reductions in sulfur dioxide emissions from power plants in an effort to reduce the impacts of acid deposition on the environment. The key study in this regard was one published in *Science* showing essentially that what goes up must come down, i.e., that regions within a few hundred miles downwind of SO_2 (and NO_x) emission sources would be impacted by the emissions.

Given the fact that anthropogenic emissions of acid precursors are expected to rise, and that acid deposition has major adverse impacts on both aquatic and land ecosystems, it seems to be worthwhile to set local, national, and international policies that would tend to reduce the projected increases in emissions. The four main routes to cutting pollution emissions are: (1) using low-pollutant fuels, (2) preventing the formation of pollutants such as NO during combustion, (3) screening pollutants from exhaust and flue gases, and (4) energy conservation. Some of these routes would also help reduce the emissions of greenhouse gases. Choosing between these routes or some combination thereof involves consideration of the trade-offs including economic and political issues, e.g., the sources of the various fuels and whether the costs of emissions reductions outweigh the benefits, with the added complication that the groups incurring the costs are not necessarily the ones reaping the benefits.

A variety of policies has been identified that could be adopted to reduce the contribution of transport sector NO_x emissions at the local level in the Netherlands. The most important national policies identified relate to vehicles and fuels, pricing policy, public transport policies, and national guidelines for policies on parking and land use, while the most important local policies identified are those for parking, land use, cycling, and restrictions for motorized vehicles.

Regulations that would lead to further reductions in nitric oxide and sulfur emissions were proposed in the United States in late 1999. Oil refiners are being asked to remove 90% of the sulfur from gasoline. The manufacturers of sport utility vehicles (SUVs) and light-duty trucks are being asked to comply with the emission standards for passenger vehicles. Older electric power generating plants, which tried to escape emissions controls under the "grandfather" clause, are being asked to cut their nitrogen emissions. The proposed action affects 392 generating units at both electric generating (EGU) and non-electric-generating (non-EGU) facilities in 12 states. Affected EGUs will be required to reduce NO_x emissions to $0.15\,\mathrm{lb}\,(\mathrm{mm\,Btu})^{-1}$, while large non-EGUs will be required to reduce NO_x emissions by approximately 60% from baseline levels.

If changed regulations are not sufficient, Congress may consider additional legislation to reduce emissions. Of course, there would be a phase-in period, so it might take a decade or two for the changes to have an impact on the environment.

REFERENCES

Adriano, D. C., and A. H. Johnson (Eds.), *Acidic Precipitation*, Vol. 2: *Biological and Ecological Effects*, Springer-Verlag, Berlin, 1989.

Boyle, Robert, *The General History of the Air*, Awnsham and John Churchill, London, 1692.

Charles, D. F., and S. Christie (Eds.), *Acidic Deposition and Aquatic Ecosystems*, Springer-Verlag, Berlin, 1991.

Cowling, E. B., Acid precipitation in historical perspective, *Environ. Sci. Technol.*, *16*, 110A–123A, 1982.

Erisman, J. W., and G. P. J. Draaijers, *Atmospheric Deposition in Relation to Acidification and Eutrophication*, Elsevier, New York, 1995.

Finlayson-Pitts, B. J., and J. N. Pitts, Jr., Acid deposition: formation and fates of inorganic and organic acids in the troposphere, Ch. 8 in *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments and Applications*. Academic, New York, 2000, pp. 294–348.

Fisher, D. C., and M. Oppenheimer, Atmospheric nitrogen deposition to the Chesapeake Bay Estuary, *Ambio*, *23*, 102–108, 1991.

Graedel, T. E., and R. McGill, Degradation of materials in the atmosphere, *Environ. Sci. Technol.*, *20*, 1093–1100, 1986.

Hedin, L. O., and G. E. Likens, Atmospheric dust and acid rain, *Sci. Am.*, *275*(6), 88–92, 1996.

Johnson, D. W., and S. E. Lindberg (Eds.), *Atmospheric Deposition and Forest Nutrient Cycling*, Springer-Verlag, Berlin, 1992.

Likens, G. E., and F. H. Bormann, Acid rain: A serious regional environmental problem, *Science*, *184*, 1176–1179, 1974.

Lindberg, S. E., A. L. Page, and S. A. Norton (Eds.), *Acidic Precipitation*, Vol. 3: *Sources, Deposition and Canopy Interactions*, Springer-Verlag, Berlin, 1990.

Radojevic, M., and R. M. Harrison (Eds.), *Atmospheric Acidity, Sources, Consequences and Abatement*, Elsevier Applied Science, New York, 1992.

Schulze, E.-D., O. L. Lange, and R. Oren (Eds.), *Forest Decline and Air Pollution, Ecological Studies 77*, Springer-Verlag, Berlin, 1989.

Schütt, P., and E. B. Cowling, Waldsterben, a general decline of forests in Central Europe: Symptoms, development, and possible causes, *Plant Disease*, *69*, 548–558, 1985.

Schwartz, S. E., Acid deposition: Unraveling a regional phenomenon, *Science*, *243*, 753–763, 1989.

Sisterson, D. L., V. C. Bowersox, T. P. Meyers, A. R. Olsen, and R. J. Vong, *Deposition Monitoring: Methods and Results*, NAPAP Report 6, Argonne National Laboratory, Argonne, III, 1990.

Smith, Robert A., *Air and rain [microform]: The Beginning of a Chemical Climatology*, Longmans, London, 1872.

Sverdrup, H., and P. Warfvinge, Past and future changes in soil acidity and implications for forest growth under deposition scenarios, *Ecol. Bull.*, *44*, 335–351, 1995.

Ulrich, B., Nutrient and acid–base budget of Central European forest ecosystems, in *Effects of Acid Rain on Forest Processes*, Wiley-Liss, New York, 1994, pp. 1–50.

Ulrich, B. 1983(a). A concept of forest ecosystem stability and of acid deposition as a driving force for destabilization. In: Ulrich, B. and Pankrath, J (Eds), Effectsm of Accumulation of Air Polutants in Forest Ecosystems. D Reidel Publishing Company, 1–29.

Ulrich, B. 1983(b). Soil acidity and its relations to acid deposition. In: Ulrich, B and Pankrath, J (Eds), Effects of Accumulation of Air Pollutants in Forest Ecosystems. D Reidel Publishing Company, 127–146.

CHAPTER 16

FUNDAMENTALS OF VISIBILITY

WILLIAM C. MALM

1 INTRODUCTION

A definition of visibility, as it relates to management of the many visual resources found in national parks, wilderness areas, and urban centers, is a complex and difficult concept to address. Should visibility be defined in strictly technical terms that concern themselves with exact measurements of illumination, threshold contrast, and precisely measured distances? Or is visibility more closely allied with value judgments of an observer viewing a scenic vista?

Historically, *visibility* has been defined as the greatest distance at which an observer can just see a black object viewed against the horizon sky. An object is usually referred to as at threshold contrast when the difference between the brightness of the sky and the brightness of the object is reduced to such a degree that an observer can just barely see the object. Much effort has been expended in establishing the threshold contrast for various targets under a variety of illumination and atmospheric conditions. An important result of this work is that threshold contrast for the eye, adapted to daylight, changes very little with background brightness, but it is strongly dependent upon the size of the target and the time spent looking for the target.

However, visibility is really more than being able to see a black object at a distance for which the contrast reaches a threshold value. Coming upon a mountain such as one of those shown in Figures 1*a* and 1*b*, an observer does not ask, "How far do I have to back away before the vista disappears?" Rather, the observer will comment on the color of the mountain, on whether geological features can be seen and appreciated, or on the amount of snow cover resulting from a recent storm system. Approaching landscape features such as those shown in Figures 1*c*

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

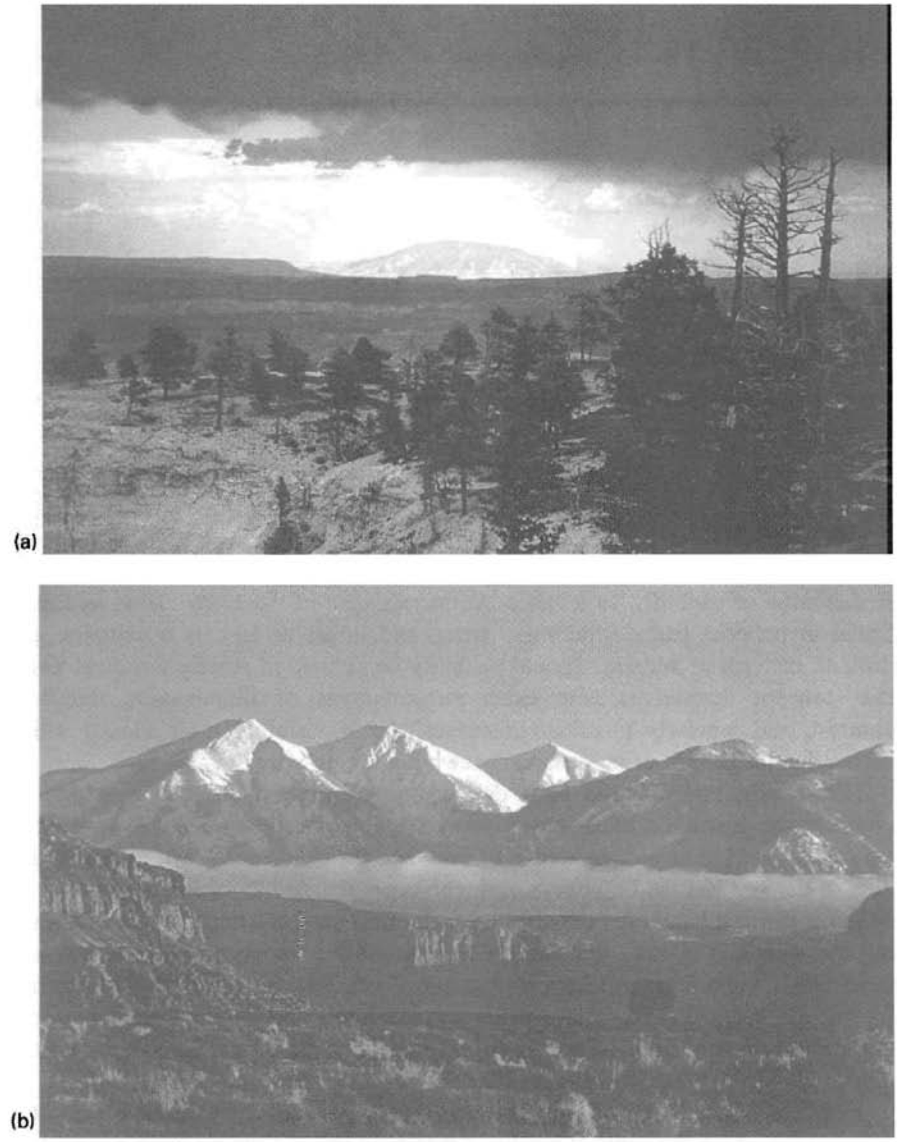


Figure 1 Photographs (*a*) through (*d*) show that, from a visual resource point of view, visibility is not how far a person can see but rather the ability of an observer to clearly see and appreciate the many and varied scenic elements in each vista. (*a*) The farthest scenic feature is the 130-km distant Navajo Mountain, as seen from Bryce Canyon National Park. (*b*) The La Sal Mountains, as seen from the Colorado River, are a dominant view from the distant horizon. (*c*) This view in Canyonlands National Park shows the highly textured foreground canyon walls against the backdrop of the La Sal Mountains. The La Sals are 50 km from the observation point. (*d*) Bryce Canyon as seen from Sunset Point. Notice the highly textured and brightly colored foreground features. See ftp site for color image.

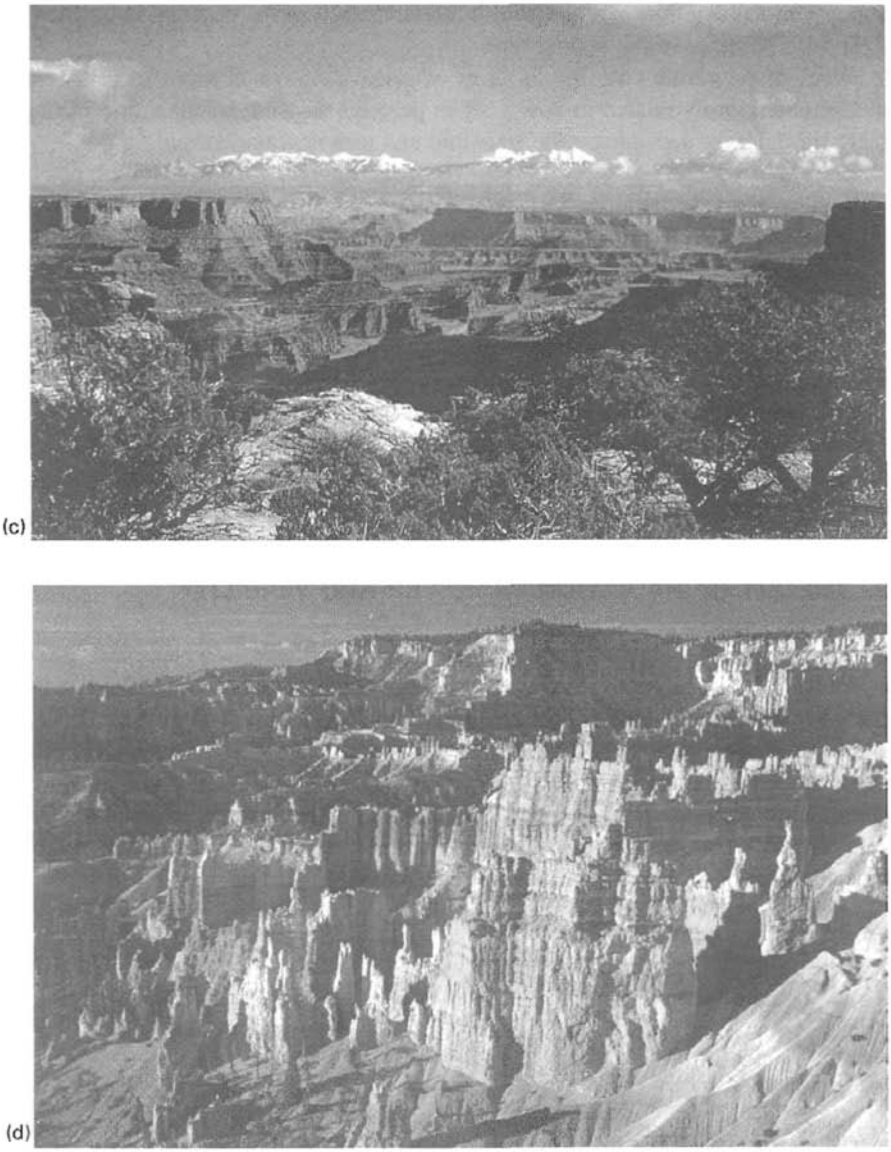


Figure 1 Continued

and 1*d*, the observer may comment on the contrast detail of nearby geological structures or on shadows cast by overhead clouds.

Visibility, in the context of viewing scenic vistas, is more closely associated with conditions that allow appreciation of the inherent beauty of landscape features. It is important to be able to see and appreciate the form, contrast detail, and color of near and distant features. Therefore, visibility includes psychophysical processes and concurrent value judgments of visual impacts, as well as the physical interaction of light with particles in the atmosphere.

Whether we define visibility in terms of visual range or in terms of some parameter more closely related to how visitors perceive a visual resource, the management of visibility depends on the scientific and technical understanding of:

- How aerosols are dispersed across land masses and into local canyons and valleys
- How they transform from a gas into particles that impair visibility
- How they interact with light
- The psychophysical processes involved in viewing scenic landscape features

Scientific understanding of some of these issues is more complete than others. The focus of this discussion is on developing a basic understanding of the interaction of light with aerosols and the psychophysical properties of the eye–brain system as they relate to visibility.

2 THEORY OF RADIATION TRANSFER AND VISIBILITY

The response of the human eye to radiant energy of different wavelengths is shown in Figure 2. The maximum response to a unit of energy is at 0.55 μm. When radiant energy is discussed in terms of the response of the human eye, photometric concepts and units are conventionally used. Conversely, when the entire radiation field of the sky is modeled or measured, radiometric units are employed. Usually, but not always, photometric parameters are derived from the more fundamental radiometric variables. Table 1 lists the various radiometric and corresponding photometric variables typically employed in radiation transfer calculations.

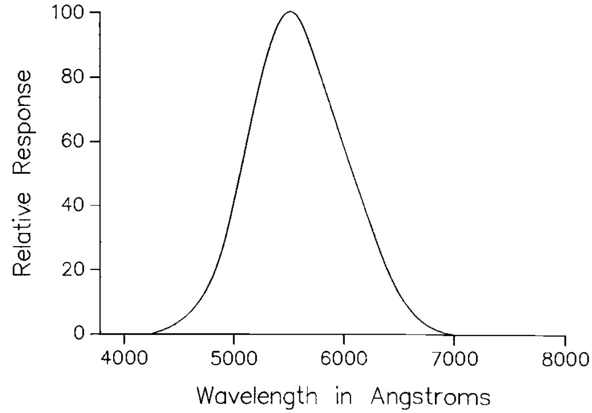


Figure 2 Spectral response of the human eye.

TABLE 1 Radiometric and Photometric Concepts and Units

Radiometric	Symbol	Units	Photometric	Symbol	Units
Radiant energy	U	joule	Luminous energy	Q	Talbot
Radiant flux	P	watt	Luminous flux	F	lumen
Radiant intensity	J	watt/steradian	Luminous intensity	I	lumen/steradian
Radiance	N	watt/m^2 steradian	Luminance	B	lumen/m^2 steradian
Irradiance	H	watt/m^2	Illuminance	E	lumen/m^2

Atmospheric Scattering and Extinction

The alteration of radiant energy as it passes through the atmosphere is due to scattering and absorption by gases and particles. The sum of scattering and absorption is referred to as the extinction coefficient. The effect of the atmosphere on the visual properties of distant objects theoretically can be determined if the concentration and characteristics of air molecules, particles, and absorbing gases are known throughout the atmosphere and most importantly along the line of sight between the observer and object. The extinction coefficient is made up of particle and gas scattering and absorption:

$$b_{\text{ext}} = b_{sg} + b_{ag} + b_{sp} + b_{ap} \tag{1}$$

where s, a, g, and p refer to scattering, absorption, gases, and particles, respectively.

Light scattering by gases is described by the Rayleigh scattering theory (vandeHulst, 1981). Important characteristics of Rayleigh scattering are:

- Its proportionality to molecular number density ($b_{sg} = 12\ \text{Mm}^{-1}$ at sea level and at 0.55 μm).
- The amount of scattered light varies as $1/\lambda^4$ where λ is the wavelength of light.
- Equal amounts of light are scattered in forward and backward directions.
- Light scattered at 90° is nearly completely polarized.

The only gas that is normally found in the atmosphere which absorbs light is nitrogen dioxide, NO_2. Absorption by NO_2 at 550 nm is $b_{ag} = 330[NO_2]$, where the units of b_{ag} are Mm^{-1} and the units of $[NO_2]$ are ppm (Nixon, 1940; Hodkinson, 1966). Furthermore, NO_2 absorbs more in the blue portion of the spectra than in the red portion. Therefore, NO_2 appears brown or yellowish if viewed against a background sky.

In most instances, particle scattering and absorption are primarily responsible for visibility reduction. Single-particle scattering and absorption properties can, with a number of limiting assumptions, be calculated using Mie theory (vandeHulst, 1981; Mie, 1908). However, before such calculations are carried out, appropriate boundary conditions must be specified. Typically aerosol models assume:

External Mixtures Particles exist in the atmosphere as pure chemical species that are mixed without interaction.

Multicomponent Aerosols Single particles are made up of two or more species.

Transfer of Radiant Energy

Visibility involves more than specifying how light is absorbed and scattered by the atmosphere. Important factors involved in seeing an object are outlined in Figure 3 and summarized below:

- Illumination of the overall scene by the sun, which includes illumination resulting from sunlight scattered by clouds and atmosphere as well as reflections by ground and vegetation
- Scene characteristics that include color, texture, form, and brightness
- Optical characteristics of intervening atmosphere:

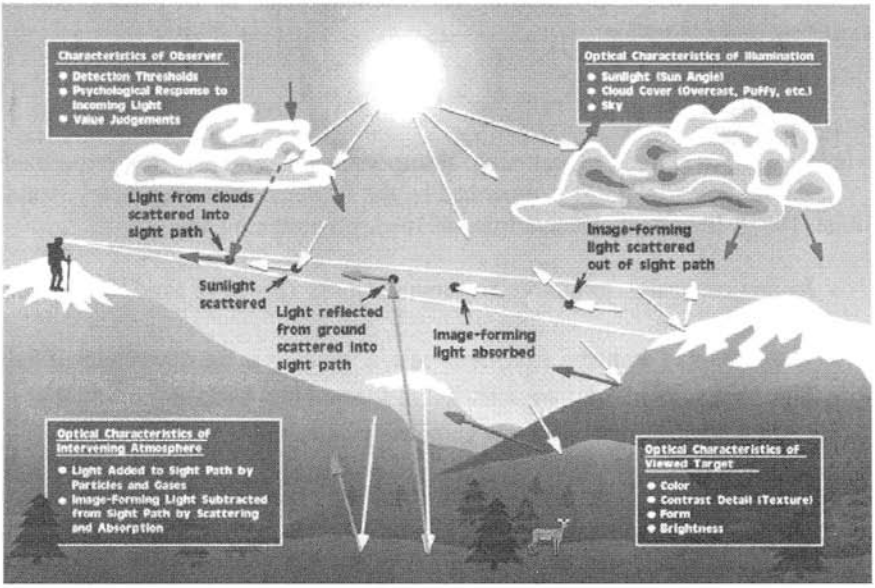

Characteristics of Observer
• Detection Thresholds
• Psychological Response to Incoming Light
• Value Judgements
Optical Characteristics of Illumination
• Sunlight (Sun Angle)
• Cloud Cover (Overcast, Puffy, etc.)
• Sky
Light from clouds scattered into sight path
Image-forming light scattered out of sight path
Sunlight scattered
Light reflected from ground scattered into sight path
Image-forming light absorbed
Optical Characteristics of Intervening Atmosphere
• Light Added to Sight Path by Particles and Gases
• Image-Forming Light Subtracted from Sight Path by Scattering and Absorption
Optical Characteristics of Viewed Target
• Color
• Contrast Detail (Texture)
• Form
• Brightness

Figure 3 Important factors involved in seeing a scenic vista are outlined. Image-forming information from an object is reduced (scattered and absorbed) as it passes through the atmosphere to the human observer. Air light is also added to the sight path by scattering processes. Sunlight, light from clouds, and ground-reflected light all impinge on and scatter from particulates located in the sight path. Some of this scattered light remains in the sight path, and at times it can become so bright that the image essentially disappears. A final important factor in seeing and appreciating a scenic vista are the characteristics of the human observer. See ftp site for color image.

  - Image-forming information (radiation) originating from landscape features is scattered and absorbed (attenuated) as it passes through the atmosphere toward the observer.
  - Sunlight, ground-reflected light, and light reflected by other objects are scattered by the intervening atmosphere into the sight path.
- Psychophysical response of the eye–brain system to incoming radiation

Image-forming information is lost by the scattering of imaging radiant energy out of the sight path and absorption within the sight path, while ambient light scattered into the sight path adds radiant energy to the observed radiation field. This process is described by:

$$\underset{\text{(loss)}}{\frac{dN_r(\theta, \varphi, \mathbf{r})}{dr}} = -b_{\text{ext}}N_r(\theta, \varphi, \mathbf{r}) \underset{\text{(gain)}}{+N_*(\theta, \varphi, \mathbf{r})} \tag{2}$$

where $N_r(\theta, \varphi, \mathbf{r})$ is the apparent radiance at some vector distance, $\mathbf{r}$, from a landscape feature, $N_*(\theta, \varphi, \mathbf{r})$ (referred to as the path function) is the radiant energy gain within an incremental path segment, and $b_{\text{ext}}N_r(\theta, \varphi, \mathbf{r})$ is radiant energy lost within that same path segment. The atmospheric extinction coefficient (b_{ext}) is the sum of both atmospheric scattering (b_s) and absorption (b_a). Although not explicitly stated, it is assumed that each variable in, and each variable derived from, Eq. (2) is wavelength dependent. The parenthetical variables ($\theta, \varphi, \mathbf{r}$) indicate that N_r and N_* are dependent both on the direction of image transmission and on the position within the path segment. For the sake of brevity, the parenthetical variables will be dropped in following equations. When the postscript r is appended to any symbol, it denotes that the quantity pertains to a path of length r. The subscript 0 always refers to the hypothetical concept of any instrument located at zero distance from the object—as, for example, in denoting the inherent radiance of a surface. Prescripts identify the objects; the prescript b referring to background and l to landscape feature.

When N_r has some special value, N_q, such that $b_{\text{ext}}N_q = N_*$, then $dN_q/dr = 0$; N_q is independent of r and is commonly referred to as the equilibrium radiance. Therefore, for every path segment

$$\frac{dN_r}{dr} = -b_{\text{ext}}(N_r - N_q) \tag{3}$$

If N_q is constant, Eq. (3) can be integrated to yield

$$\frac{N_r - N_q}{N_0 - N_q} = T_r \tag{4}$$

where T_r is the transmittance over path length r and is given by

$$T_r = \exp - \int_0^r b_{\text{ext}} r' dr' \tag{5}$$

Rearranging Eq. (4) yields

$$N_r = N_0 T_r + N_q(1 - T_r) \tag{6}$$

where the first term on the right of Eq. (6) is the residual image-forming radiance, while the second term is the path radiance (airlight), N_r^*, which results from scattering processes throughout the sight path. The parameter N_∞^* is the sky radiance:

$$N_\infty^* = N_q(1 - T_\infty) \tag{7}$$

If T_∞ is approximately zero, then $N_q = N_\infty^* = N_s$ and

$$N_r^* = N_s(1 - T_r) \tag{8}$$

where N_s is sky radiance. Equation (8) allows for a simple approximation of N_r^* when N_s is known.

The explicit dependence of N_r^* on illumination and directional scattering properties of the atmosphere are best examined by considering

$$N_r^* = \int_0^r N_* T_r \, dr \tag{9}$$

where

$$N_* = h_s \sigma + \int_{4\pi} N\sigma \, d\Omega \tag{10}$$

The second term on the right-hand side is the contribution to N_* from sky, cloud, and earth radiance and $d\Omega$ is an element of solid angle. The parameter h_s is sun irradiance, and σ is the volume scattering function defined in such a way that

$$b_s = \int_{4\pi} \sigma \, d\Omega \tag{11}$$

Therefore, σ describes the amount of radiant energy (light) scattered in some direction, while the sum of radiant energy scattered in all directions is proportional to the scattering coefficient b_s. The amount of energy scattered out of and into a sight path over some incremental distance, Δr, is proportional to b_s. It is a fundamental optical property of the atmosphere. Its measurement and characterization have been the focus of a number of studies.

Contrast Transmittance in Real Space

Any landscape feature can be thought of as consisting of many small pieces, or elements, with a variety of physical characteristics. For instance, the reflectivity of an element as a function of wavelength, along with characteristics of the incident radiation, determines its color and brightness. The brightness of an element at some observing distance and at one wavelength is referred to as monochromatic apparent spectral radiance. The monochromatic apparent spectral radiance of any element is given according to Eq. (6) by

$${}_{l}N_r = T_{rl}N_0 + N_r^* \tag{12}$$

where N_r^* is substituted explicitly for $N_q(1 - T_r)$. The subscript l indicates that the radiance is associated with a specific uniform landscape feature. In the early literature the subscript t (for target) was used instead of l because of the applicability of Eq. (12) and contrast to the seeing of military targets.

A scenic element is always seen against some background, such as the sky or another landscape feature. The apparent and inherent background radiance are related by an expression similar to Eq. (12)

$${}_{b}N_r = T_{r\,b}N_0 + N_r^* \tag{13}$$

Subtracting Eq. (13) from Eq. (12) yields the relation

$$({}_{l}N_r - {}_{b}N_r) = T_r({}_{l}N_0 - {}_{b}N_0) \tag{14}$$

Thus, radiance differences are transmitted along any path with the same attenuation as that experienced by each image-forming ray.

The image-transmitting properties of the atmosphere can be separated from the optical properties of the object by the introduction of the contrast concept. The inherent spectral contrast, C_0, of a scenic element is, by definition,

$$C_0 = ({}_{l}N_0 - {}_{b}N_0)/{}_{b}N_0 \tag{15}$$

The corresponding definition for apparent spectral contrast at some distance r is

$$C_r = ({}_{l}N_r - {}_{b}N_r)/{}_{b}N_r \tag{16}$$

If Eq. (14) is divided by the apparent radiance of the background ${}_{b}N_r$ and combined with Eqs. (15) and (16), the result can be written as

$$C_r = C_0 \frac{{}_{b}N_0}{{}_{b}N_r} T_r \tag{17}$$

Substituting Eq. (13) for ${}_bN_r$ and rearranging yields

$$\tau_r \equiv \frac{C_r}{C_0} = \frac{1}{1 + \dfrac{N_r^*}{{}_bN_0 T_r}} \tag{18}$$

The right-hand member of Eq. (18) is an expression for the contrast transmittance, τ_r, of the path of sight. Equation (18) is the law of contrast reduction by the atmosphere expressed in the most general form. It should be emphasized that Eq. (18) is completely general and applies rigorously to any path of sight regardless of the extent to which the scattering and absorbing properties of the atmosphere or the distribution of lighting exhibit nonuniformities from point to point.

Visual Range Concept

Substituting Eq. (5) into Eq. (17) yields

$$C_r = C_0 \frac{{}_bN_0}{{}_bN_r} e^{-b_{\text{ext}} r}. \tag{19}$$

If an object is viewed against a background sky under uniform illumination conditions and through a uniform haze, ${}_bN_0/{}_bN_r = 1$ and Eq. (19) becomes

$$b_{\text{ext}} = -\frac{1}{r} \ln \frac{C_r}{C_0} \tag{20}$$

Equation (20) forms the basis for using teleradiometer contrast measurements for approximating the extinction coefficient. If C_0 and distance r are known, b_{ext} can be calculated.

The distance at which C_r approaches a threshold contrast of between -0.02 or -0.05 defines the visual range, V_r. If $|C_0| = 1$ (black object) and -0.02 is taken to be a threshold contrast, then Eq. (20) becomes

$$V_r = \frac{3.192}{b_{\text{ext}}} \tag{21}$$

Equation (21) allows visual range data to be interpreted in terms of extinction and vice versa, extinction measurements to be interpreted in terms of visual range. There is some debate as to what threshold contrast to use.

Equivalent Contrast

The above mathematical formalism is limited in that it does not account for human visual system response to edge sharpness between adjacent scenic features or to changes in contiguous contrast for features with varying size. More modern psychophysical perception threshold formalisms can be constructed to incorporate the eye–brain system response to variations in edge sharpness between landscape features as well as variation in spatial frequency of landscape scenic elements (Carlson and Cohen, 1978; Campbell and Robson, 1964; Campbell et al., 1968; Campbell and Kulikowski, 1986; Henry, 1977; Malm, 1985; Malm et al., 1987). Any approach that incorporates the human response to spatial frequencies (size and shape effects) is most easily handled using linear system theory. A first step is to develop a quantitative descriptor of the scene itself.

A scene can be decomposed into light and dark bars of various spatial frequencies and intensities whose brightness change is proportional to a sine wave function. Equivalent contrast, C_{eq}, is just the average contrast of those sine waves within specified frequencies. Therefore, equivalent contrast can be calculated either for all spatial frequencies or only for those frequencies to which the human visual system responds. Then C_{eq} can be used in human visual system models to estimate the probability that a human observer will notice a change in the appearance of a landscape feature as aerosols are added or removed from the atmosphere.

Contrast Transmittance in Spatial Frequency Space (Modulation Transfer Function)

In a derivation similar to the contrast transmittance derivation, it can be shown that the transmittance of equivalent contrast through the atmosphere in the presence of aerosols is given by

$$C_{\mathrm{eq},r} = C_{\mathrm{eq},0} M_{\mathrm{tf},a} \tag{22}$$

where

$$M_{\mathrm{tf},a} = \frac{1}{1 + \dfrac{N_r^*}{a_\infty T_r}} \tag{23}$$

and $C_{\mathrm{eq},r}$ and $C_{\mathrm{eq},0}$ are the equivalent contrast at distance r and 0, respectively, while $M_{\mathrm{tf},a}$ is the atmospheric modulation transfer function. The parameter a_∞, the average scene radiance, is the zero-order term in a two-dimensional Fourier decomposition of the scene radiance field.

Comparison of Eqs. (18) and (23) shows that if ${}_bN_0 = a_\infty$, then contrast transmittance in real and spatial frequency space is identical. In most cases, the feature within the image of interest is small compared with its surroundings, and average radiance, a_∞, is very nearly the same as background radiance ${}_bN_0$. This is a very satisfying result. Whether one is interested in using modern psychophysical spatial frequency models to examine how much aerosol can be introduced into the atmosphere before it is noticed or how image contrast is changed as a function of aerosol load, the calculation is reduced to understanding the dependence of the atmospheric modulation transfer function, or contrast transmittance, on aerosol chemical and physical properties.

Dependence of Contrast Transmittance (τ_r) on Atmospheric Optical Variables

Because the contrast transmittance is the one variable that contains all the information required to describe how various physical descriptors of scenic landscape features are modified as a function of aerosol loading, illumination, and observer-vista geometry, it is of interest to examine how sensitive τ_r is to changes in atmospheric aerosol loading as a function of aerosol mass and average scene radiance. The average scene radiance, $\bar{N}$, was identified as a_∞ in Eq. (23).

Malm and Henry (1987) examined how the τ_r changes with changing image reflectivity, image distance, aerosol size distribution, and aerosol mass loading. For a sulfate aerosol, b_{ext} is almost entirely due to scattering and, as such, b_{ext} is proportional to aerosol mass. Therefore, the variation of τ_r with respect to b_{ext} is proportional to its variation with respect to aerosol mass. Figures 4*a* and 4*b* show $s \equiv |\Delta\tau_r \Delta b_{\text{ext}}|$ as a function of b_{ext}.

Figure 4*a* corresponds to a typical sulfate aerosol mass size distribution, scattering angle $\theta_s = 15°$, and $N_0 = 0.13N_s$, where N_s is the Rayleigh sky radiance. Figure 4*b* is also for a sulfate aerosol but with $\theta_s - 125°$ and $N_0 = 0.5N_s$. An immediately evident trend shown in Figures 4*a* and 4*b* is that there is a distance where S is maximum; S decreases to zero as $R \to 0$ and as $R \to \infty$. Secondly, the distance at which S is maximum increases as N_0 increases (brighter landscapes). In a forward scattering situation where landscapes are in a shadow ($C_0 \approx -0.90$), S is maximum in the 5- to 10-km range. Although not explicitly shown in Figure 4*a*, in a backscatter geometry ($\theta_s = 125°$), the most sensitive distance is still around 5 to 10 km if the landscape is dark. However, the maximum sensitivity drops by about a factor of 2 and is not nearly as sensitive to distance. On the other hand, Figure 4*b* shows that when the landscape is highly reflective and illuminated ($C_0 \approx 0.50$ and $\theta_s = 125°$) the distance of maximum sensitivity increases, is quite sensitive to background b_{ext}, and remains sensitive to changes in b_{ext} long after dark targets have lost their sensitivity (dark targets will have disappeared, while bright targets can still be seen).

Figure 5 examines in more detail the relative contribution of N_r^* and T to S. Figure 5 shows contributions of $N_r^*$and T to S for the case shown in Figure 4*a* at $T = 10$ km (forward scattering, sulfate aerosol, and dark target). Changes in N_r^* are primarily responsible for changes in $M_{\text{tf},a}$ as aerosol is added or subtracted from a clean atmosphere. As background aerosol loading is increased (larger b_{ext}), the relative importance of T to S increases to a point where T dominates the effect on S. However, it should be emphasized that this only occurs after the $M_{\text{tf},a}$ has increased to a point where landscape features would be barely visible. Figure 5*b* shows N_r^* and T contributions to S for the Figure 4*b* case at $R = 70$ km (backscatter, sulfate aerosol, and bright target). With this geometry, attenuation of image-forming information, T, is responsible for much of the change in $M_{\text{tf},a}$. In fact, N_r^* can

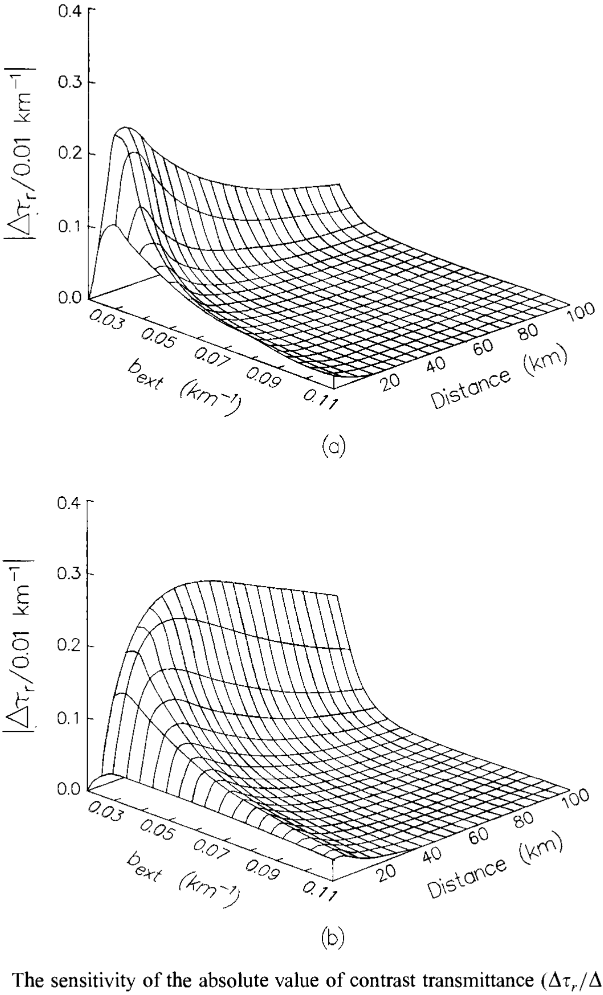


Figure 4 The sensitivity of the absolute value of contrast transmittance ($\Delta\tau_r/\Delta b_{\mathrm{ext}}$) plotted as a function of extinction coefficient and distance to landscape feature. (*a*) Scattering angle $\theta_s = 15°$, shadowed vista $\bar{N}_0 = 0.13\ N_s$, and sulfate aerosol, and (*b*) scattering angle $\theta_s = 125°$, illuminated vista $\bar{N}_0 = 0.5\ N_s$, and sulfate aerosol.

decrease as b_{ext} increases and compensate slightly (contribute to cause $M_{\mathrm{tf},a}$ to increase) for decreases in T.

The foregoing discussion shows that the effect of increasing b_{ext} (aerosol concentration) for a scattering aerosol in almost all situations causes $M_{\mathrm{tf},a}$ to decrease. However, under forward scattering situations where targets tend to be dark, N_r^*

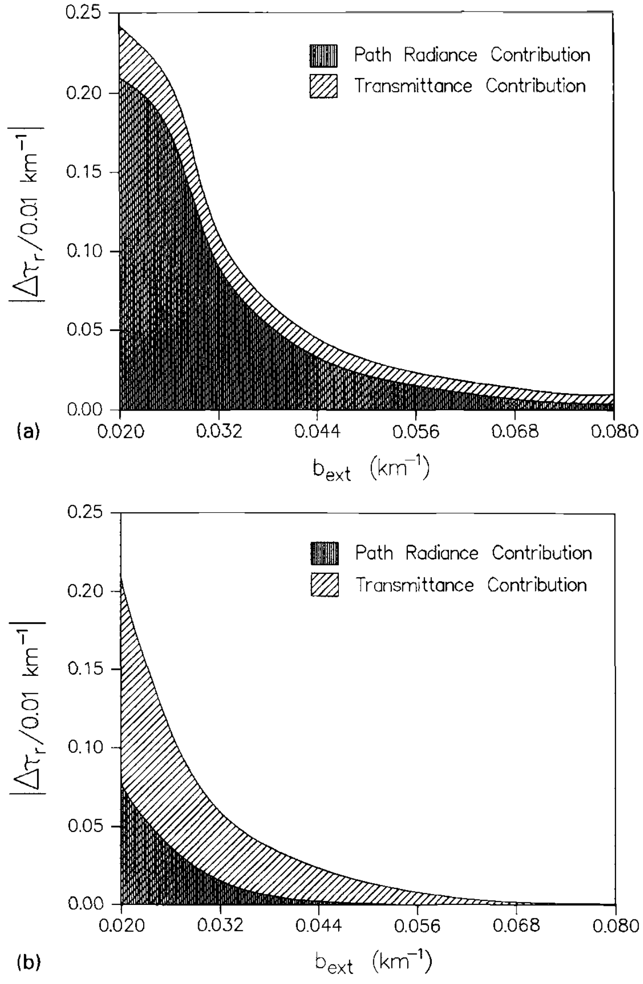
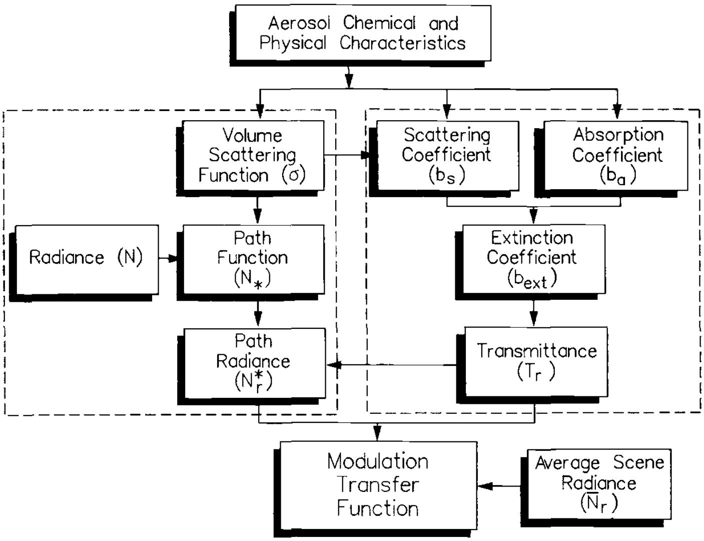


Figure 5 Sensitivity (S) expressed as changes in the modulation transfer function increase of $b_{\text{ext}} = 0.01\ \text{km}^{-1}$ plotted against b_{ext} for a sulfate aerosol. In (a) $\theta_s = 15°$, $R = 70$ km, and $\bar{N}_0 = 0.13\ N_s$, while in (b) $\theta_s = 125°$, $R = 70$ km, and $\bar{N}_0 = 0.5\ N_s$. Parts (a) and (b) show the relative contributions of path radiance and atmospheric transmittance to changes in $M_{\text{tf},a}$ as a function of b_{ext}.

dominates changes in $M_{\text{tf},a}$. On the other hand, when looking at brightly colored landscape features with the sun behind the observer's back (backscatter), the relative importance of N_r^* to visibility becomes smaller and changes in N_r as a result of increased b_{ext} are more dependent on image-forming radiance being attenuated over the sight path. However, for a specific scene under static illumination conditions, contributions of N_r^* and T to change in $M_{\mathrm{tf},a}$ as a function of aerosol concentration tend to track each other.

Because most research to date has focused on apportionment of b_{ext}, and therefore T, to aerosol species, it is fortunate that for scattering aerosols, such as sulfates, an understanding of this relationship yields significant insight into how aerosols affect visibility under a wide range of viewing conditions. However, under not uncommon circumstances, the major cause of visibility degradation can be associated with path radiance, and path radiance explicitly requires knowledge of the volume scattering function in addition to b_{ext}. Almost no effort has been expended on examining how path radiance is affected as a function of aerosol characteristics or on apportioning path radiance to aerosol species. Aerosols that absorb light contribute to path radiance differently than aerosols that only scatter light (such as sulfates), so the impact of scatterers and absorbers on path radiance is not additive. Conversely, the effect of scattering and absorbers in b_{ext} is additive. Therefore, when appreciable concentrations of light-absorbing particles or gases are present, knowledge of just b_{ext} (transmittance) may not be adequate to describe changes in visibility.

The concepts discussed above are summarized in Figure 6. Those variables enclosed in the box on the left side of Figure 6 are dependent on illumination observer geometry, while those on the right are not. Path radiance, a geometry-



Aerosol Chemical and Physical Characteristics
Volume Scattering Function (σ)
Scattering Coefficient (b_s)
Absorption Coefficient (b_a)
Radiance (N)
Path Function (N*)
Extinction Coefficient (b_ext)
Path Radiance (N*r)
Transmittance (Tr)
Modulation Transfer Function
Average Scene Radiance (N̄r)

Figure 6 Flow diagram showing how aerosol physical-chemical characteristics relate to the optical variables required to completely specify the atmospheric modulation transfer function.

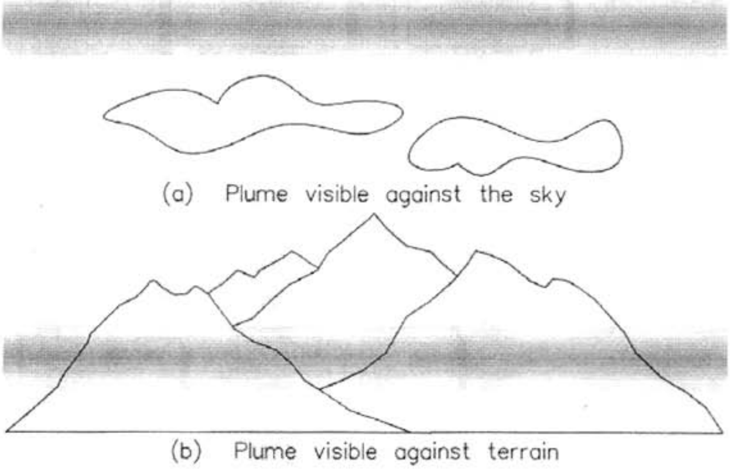
dependent variable, is combined with atmospheric transmittance, a geometry-independent parameter, and average scene luminance to yield contrast transmittance or modulation transfer function.

3 VISIBILITY IMPAIRMENT

Aerosols introduced into the atmosphere can result in visibility impairment that is manifested in two distinct ways: first, as a general alteration in the appearance of landscape features such as color, contiguous contrast between adjacent geologic features, etc., and secondly the aerosol haze may become visible in and of itself. Haze may be visible by the contrast or color difference between itself and its background, or (at great enough optical depths) uniform haze manifests itself as a semitransparent curtain that can be seen or perceived as a separate hazy entity disassociated from landscape features. Henry (1987) has referred to the phenomenon as atmospheric transparency, which is psychophysical in nature, and different from atmospheric transmittance.

Perceptibility Parameters for Quantification of Layered Haze (Plume Blight)

Figure 7 illustrates two situations in which a layered haze is visible: (a) when viewed against the sky and (b) when viewed against terrain features. In both cases, the layered haze will be visible as a distinct, horizontal layer if it is sufficiently brighter or darker than the viewing background.


(a) Plume visible against the sky

(b) Plume visible against terrain

Figure 7 Two viewing situations in which plumes may be visible.

The simplest way to characterize the relative brightness (or darkness) of plumes is through the use of plume contrast:

$$C = \frac{{}_pN_r - {}_bN_r}{{}_bN_r} \tag{24}$$

where ${}_pN_r$ and ${}_bN_r$ are the spectral radiances of the plume and its background, at some distance, r, and at wavelengths in the visible spectrum ($0.4 < \lambda < 0.7$ μm). A plume is visually perceptible only if it creates a nonzero contrast at different wavelengths in the visible spectrum greater than an observer's perceptibility threshold (generally in the range of ±0.01 to ±0.05).

An object can be perceived because it has a brightness different from that of the background or because it has a different color. Gases and particles in the atmosphere can give rise to coloration by their light-scattering properties (blue sky or white clouds) or by altering the color of objects seen through them (brown coloration due to NO_2). Several schemes have been used to quantify color. The Commission Internationale de l'Eclairage (CIE) has set colorimeter standards that form the basis of the CIE system of color specification. The most popular CIE index is the so-called ΔE parameter that not only quantifies differences in color but also differences in brightness. However, the CIE method, while accurate and acceptable for a laboratory situation, may not adequately represent color differences in a natural setting. In any case, a ΔE of 1 is a just noticeable difference in color and/or brightness in a laboratory setting and ΔE of 4 can be easily seen by the casual observer.

Layered Haze Thresholds. Psychophysical research (Cornsweet, 1970; Faugeras, 1979; Hall and Hall, 1977; Henry, 1986; Howell and Hess, 1978; Malm et al., 1987; Ross et al., 1987) has documented the fact that the human eye–brain system is most sensitive to spatial frequencies of approximately three cycles/degree (cpd). Spatial frequency is defined as the reciprocal of the distance between sine-wave crests (or troughs) measured in degrees of angular subtense of a sine-wave grating. Thus, spatial frequency has units of cycles/degree. Any pattern of light intensities, whether it is a sine wave, square wave, step function, or any other pattern, can be resolved by Fourier analysis into a sum of sine-wave curves of different magnitude and frequency. For instance, a rough estimate of the primary spatial frequency of a Gaussian plume can be made as follows. If it were assumed that a Gaussian distribution is nearly identical to a sine-wave pattern, a 2° width of the plume would correspond to the period of the sine wave. The spatial frequency would be the inverse of this, or 0.5 cpd. Figure 8 illustrates several estimates of the sensitivity of the human visual system to sine/square-wave gratings and single Gaussian and square-wave stimuli with various spatial frequencies.

The sensitivity of the human eye–brain system drops off significantly at high spatial frequency (due to visual acuity) and also to a lesser extent at low spatial frequency (i.e., broad, diffuse objects). The human visual system is more sensitive to images with sharp, distinct edges (e.g., square waves) than to images with diffuse, indistinct edges (e.g., sine waves or Gaussian plumes).

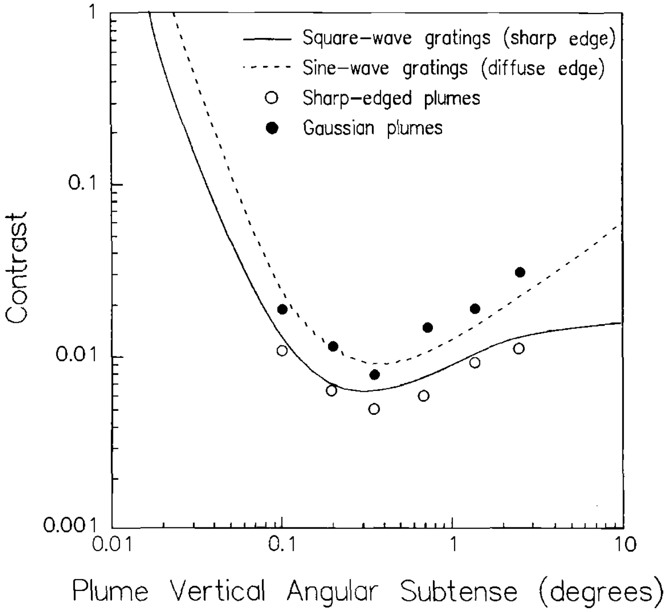


Figure 8 Sensitivity curves as reported by Howell and Hess (1978) for sine- and square-wave ratings and for sharp-edged (Malm et al., 1987) and Gaussian plumes (Ross et al., 1990).

Ross et al. (1997), based on an extensive literature review, designed a laboratory study to develop the information necessary to predict the probability of detection of plumes with a known size, shape, and contrast. The strategy taken was to develop probability of detection curves for computer-generated plume stimuli that encompasses the various plume geometries that could be encountered in the "real" world and interpolate between these measured thresholds to develop estimates for plumes with other shapes and geometries. In each case, the protocol for observer detection was the same for all experiments, the surround was kept at the same brightness, edge effects were dealt with uniformly, and stimuli representative of Gaussian plume brightness profiles were used.

Sixteen subjects were used for a full-length plume experiment. The stimuli used consisted of plumes with vertical angular sizes of 0.09°, 0.18°, 0.36°, 0.72°, 1.44°, and 2.88° and a horizontal angular extent of 16°. Contrast values of 0.050, 0.040, 0.030, 0.020, 0.017, 0.015, 0.013, 0.011, and 0.005 were used for all sizes. Figure 9 shows the predicted probability of detection curves. As plume contrast increases, the probability of detecting the plume increases. If a plume has a modulation contrast of greater than about 0.01 it will be detected nearly 100% of the time for all size plumes. Furthermore, these curves show that the size of the plume is quite impor-

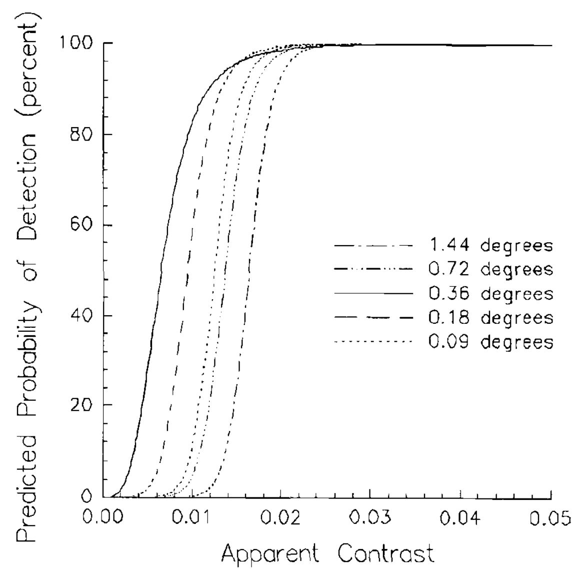


Figure 9 Predicted probability of detection curves for one subject used in the full-length plume study.

tant! Plumes that subtend an angle of about 3° can be detected more easily than plumes that are larger or smaller. Results for circular and oval-type plumes with Gaussian edges were similar but required higher contrasts to be detected.

To more clearly see how the three shapes compared, the modulation contrast corresponding to 50% probability of detection for each shape is plotted against plume size in Figure 10. Notice that the general trend for all stimuli is the same, with plumes subtending about a 3° width being the easiest to detect. However, observers are most sensitive to full-length plumes and least sensitive to circular stimuli with the oval plumes being intermediate. The full length, oval, and circular plume contrast threshold data have been incorporated into a linear interpolation algorithm that allows plumes of any size to be estimated.

Other studies have been carried out for brighter than background-layered hazes. They identified a 70% detection threshold contrast of 0.02 using photographs of a natural scene with light-colored layered hazes, which varied in size. The evidence for ΔE thresholds is not as clear-cut. The data of Jaeckel (1973) and Malm et al. (1980) support 70% detection thresholds for ΔE of three, while the estimates of Latimer et al. (1978) and the more recent data of Malm et al. (1987) and Henry and Matamala (1990) suggest a ΔE threshold of less than 1. This work is summarized in Table 2.

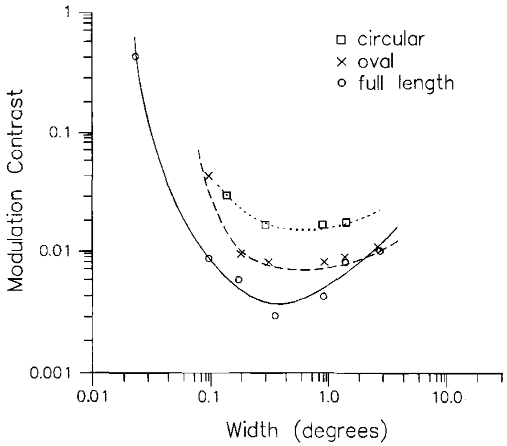


Figure 10 Threshold modulation contrast is plotted as a function of plume width in degrees for full-length, oval, and circular plumes. The human observer is most sensitive to all plumes if they have a width that is about 3°. Plumes larger or smaller than about 3° require increased contrast to be seen.

Perceptibility Parameters for Quantification of Uniform Haze Impairment

Whereas work discussed in the previous sections has emphasized detection thresholds of layered hazes, specifically plumes, other researchers have concentrated their efforts in establishing the change in image appearance required to just notice a difference in image sharpness.

Early work focused on establishing the just noticeable difference between a scene where an object viewed against the same background could just be seen and one where that object could not be identified. This threshold work was carried out in the context of establishing the "threshold" contrast for visual range determination.

More recent work has been directed toward incorporating results of basic psychophysical measurements into models that will predict the change in display modulation transfer function (MTF) required to evoke a just noticeable difference (JND) in display image sharpness. Displays of interest were television-type video displays. One model, the quadratic detection model (QDM), relies on the calculation of the image mean square luminance fluctuation, termed the image modulation depth. Henry (1979) and Henry et al. (1981) have suggested that modulation depth may be appropriate visibility indices because they incorporate all of the information content contained in a scenic vista.

Malm and Pitchford (1989) have suggested using the concept of a just noticeable change (JNC) in the appearance of a landscape feature as a psychophysical variable that relates directly to human perception. A JNC corresponds to the amount of absorbing gas or atmospheric particulate matter required to evoke a noticeable change in the appearance of a particular landscape. The effect of a change in aerosol concentration can then be expressed as the number of JNCs between landscape appearance under current conditions versus the appearance after a change in emissions. Malm and Pitchford (1989) have suggested using the QDM to predict a JNC; however, any psychophysical model relating changes in aerosol concentration to human eye–brain visual thresholds could be used for this purpose. It is emphasized that none of the currently used psychophysical models have been field validated.

More recently, Pitchford and Malm (1994) have proposed the deciview scale, which is based on the fact that all detection threshold models and experiments show that above contrasts of about 0.02, a just noticeable change in contrast, is directly proportional to the initial contrast, $\Delta C = LC$, where L is a proportionality constant. By assuming the availability of sensitive scenic targets at every distance, it can then be demonstrated that any specific fractional change in extinction coefficient is equally perceptible regardless of baseline visibility conditions. The index is defined so that its scale, which is expressed in deciview (dv), is linear with respect to fractional changes in extinction and is given by, $\text{dv} = 10 \ln(b_{\text{ex}}/0.01\ \text{km}^{-1})$, where extinction is expressed in inverse kilometers. A 1-dv change is about a 10% change in extinction.

Application of the Quadratic Detection Model. Typical scenes are made up of features that are quite varied with respect to size, shape, and luminance level. However, some attempts have been made to classify scenic structure into broad categories such as form, line, and texture. Form refers to large shapes seen either against sky or other uniform background, while line is usually associated with appearance of rivers or similar geological features. Texture refers to the periodic contrast associated with sparsely populated trees seen against a uniform background, varied geologic features, or other similar higher frequency scenic structures.

Studies investigating eye fixation and eye motion as observers look at pictures show that pictorial areas with little modulation receive very little attention, while higher modulated scenic features receive more (Boswell, 1975). Since high-contrast edges are most sensitive to changes in atmospheric modulation transfer function and since the discrimination of an atmospheric modulation change in a frequency-specific channel is a minimum when the contrast in that channel is largest, it can be concluded that high-contrast edges are good patterns for predicting the relationship between just noticeable changes in scenic appearance and increases in atmospheric aerosol load.

For many typical scenes, a JNC is equivalent to a change in atmospheric modulation of approximately 0.06. Figure 11 shows a typical JNC surface for an 80% reduction in atmospheric extinction as a function of observer distance and atmospheric extinction, assuming a change in MTF of 0.06 is perceptible. The scattering angle is 15°, $a_{\infty} = N_s$ where N_s is sky brightness, and the initial contrast, C_0, is equal to -1.0. A typical aerosol mass size distribution with typical chemical properties was assumed.

There are some general features that show up in all JNC surfaces. For any given distance there is a background extinction that is most sensitive to an incremental

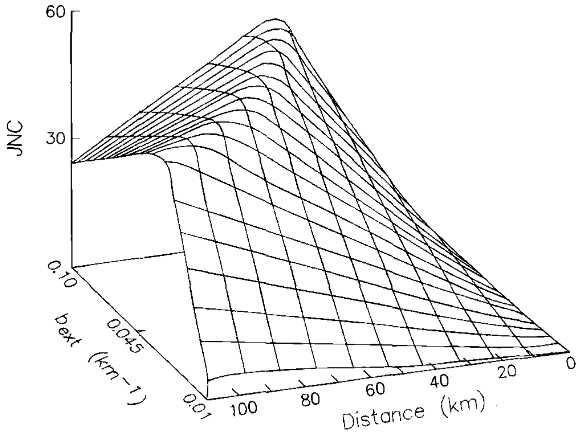


Figure 11 Just noticeable change surface plotted as a function of observer distance and atmospheric background extinction. The surface corresponds to a reduction of background extinction of 80%.

change in extinction, and for any given extinction there is observer distance that is most sensitive to extinction change. Secondly, for any given observer distance, the sensitivity of a scene to incremental reductions in atmospheric extinction drastically reduces as background extinction increases; and finally, the distance where the scene is most sensitive to a change in extinction decreases as background extinction increases.

Human Judgments of Visual Air Quality

The previous section discussed methodologies for establishing the change in atmospheric particulate loading required to be noticeable either as a layered haze or as a change in scenic quality. It should be emphasized that calculations of detection thresholds and JNCs are statements about changes in information content in an image. JNC changes in the appearance of an image are not necessarily good indicators of judged image quality. For instance, a change in 10 JNCs in a scene with low overall contrast may not be judged to have the same change in image quality as 10 JNCs in a high-contrast scene.

Studies by Malm et al. (1980, 1981), Latimer et al. (1980, 1983), Middleton et al. (1984), Stewart et al. (1983), and Hill (1990) have established relationships between judgments of image quality of natural scenes and various atmospheric and vista parameters such as mountain/sky contrast, solar angle, extinction coefficient, sky color, and percent cloud cover. Latimer et al. (1980) had observers judge scenic beauty (SBE) and visual air quality (VAQ) for a number of eastern and western national park vistas as they appeared under a variety of illumination and meteorological conditions. The results of their study were mixed and in some cases contradictory. In Latimer et al. (1980, p. 113), they conclude that "to different extents for different vistas, ratings of VAQ and SBE both increase with increasing visual range." In Latimer et al. (1983, pp. 49–50), they conclude that "ratings of SBE of a given vista were independent of visual range unless there was a dominant distant landscape feature in the landscape scenery." Since the visual range calculation "normalizes" out specific unique characteristics of vistas, these results are not surprising. The Latimer studies did conclude that changes in illumination did have a considerable effect on SBE ratings. Middleton et al. (1984) also concluded that illumination was important to VAQ judgments and were able to show at one site that there is a good correlation between VAQ and $\ln(b_{\text{scat}})$, where b_{scat} is the atmospheric scattering coefficient. Additionally, Hill (1990) emphasizes that color is extremely important to judgments of scenic beauty.

Malm et al. (1980) examined the relationship between VAQ and vista contrast. They showed that, under fixed illumination and meteorological conditions, apparent vista contrast of the most distant vista element was a good prediction of VAQ judgments. The study also showed that changes in foreground color (due to change in illumination), addition of clouds, or snow cover caused the VAQ ratings to be higher but did not cause the sensitivity of VAQ to change in vista contrast change. Malm et al. (1980) also presented a model of human perception of VAQ. The model is based on the observation that ratings of VAQ are proportional to the sum of the fraction of each scenic element subtended by various landscape features multiplied by the atmospheric transmittance between that landscape feature and observer. It was shown that when a single landscape feature, void of color and textural detail, dominates the perceived change in visual air quality, the model predicts a linear relationship between VAQ and the apparent contrast of that landscape feature (contrast of form).

Several researchers have found that judgments of photographs can be used as surrogates for judgments made in the field provided the experiments have been properly designed. This is an important finding since one way to reduce the per-observation cost of obtaining judgment-based measurements of visual air quality is to use judgments of photographs rather than field observations. For example, Stewart et al. (1984) found that although visual air quality tends to be judged slightly worse in photographs than in the field, the relative differences among scenes are approximately the same whether visual air quality is judged from photographs or in the field.

The implication of the visual air quality perception research described in the preceding paragraphs is that there are a number of variables such as sun angle, cloud cover, and scene composition that are firmly integrated into judgments of aesthetic value of a scenic resource. Therefore, studies designed to assess social, psychological, or economical value associated with a given change in atmospheric particulate concentration must be designed in such a way that these confounding variables do not affect the outcome of the experiment. For instance, a number of experiments have been carried out using photographs of landscape features under a variety of air quality conditions as the stimulus. To avoid extraneous variables such as sun angle from affecting the study, it is essential that the study be carried out using photographs taken at the same time of day and under similar lighting conditions.

4 EXAMPLES OF VISIBILITY IMPAIRMENT

The camera can be an effective tool in capturing the visual impact that pollutants have on a visual resource. Figures 12*a* to 12*d* show the effect that various levels of uniform haze have on a Glacier National Park vista. These photographs were taken of the Garden Wall from across Lake McDonald. Figures 13 and 14 show similar hazes of vistas at Mesa Verde and Bryce Canyon National Parks. The Chuska Mountains in Figure 13 are 95 km away. Navajo Mountain in Figure 14 is 130 km distant. This photograph should be compared with Figure 1*a*, a photograph of Navajo Mountain taken on a day in which the particulate concentration in the atmosphere was near zero.

Under stagnant air mass conditions aerosols can be "trapped" and produce a visibility condition usually referred to as layered haze. Figure 15 shows Navajo Mountain viewed from Bryce Canyon National Park with a bright layer of haze that extends from the ground to about halfway up the mountain. Figure 16 is a similar example of layered haze but with the top portion of the mountain obscured.

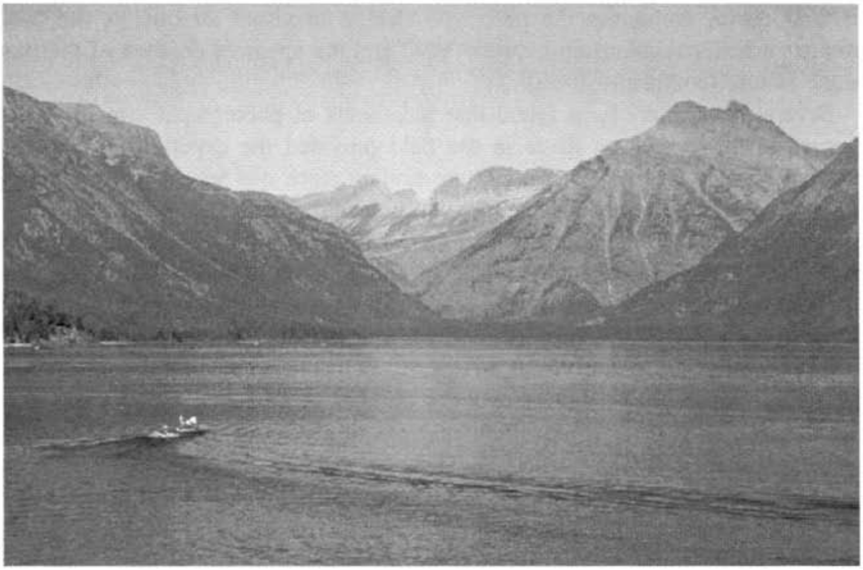


Figure 12 Effect of regional or uniform haze on a Glacier National Park vista. The view is of the Garden Wall from across Lake McDonald. Atmospheric particulate concentrations associated with photographs (*a*), (*b*), (*c*), and (*d*) correspond to 7.6, 12.0, 21.7, and 65.3 $\mu g/m^3$. See ftp site for color image.

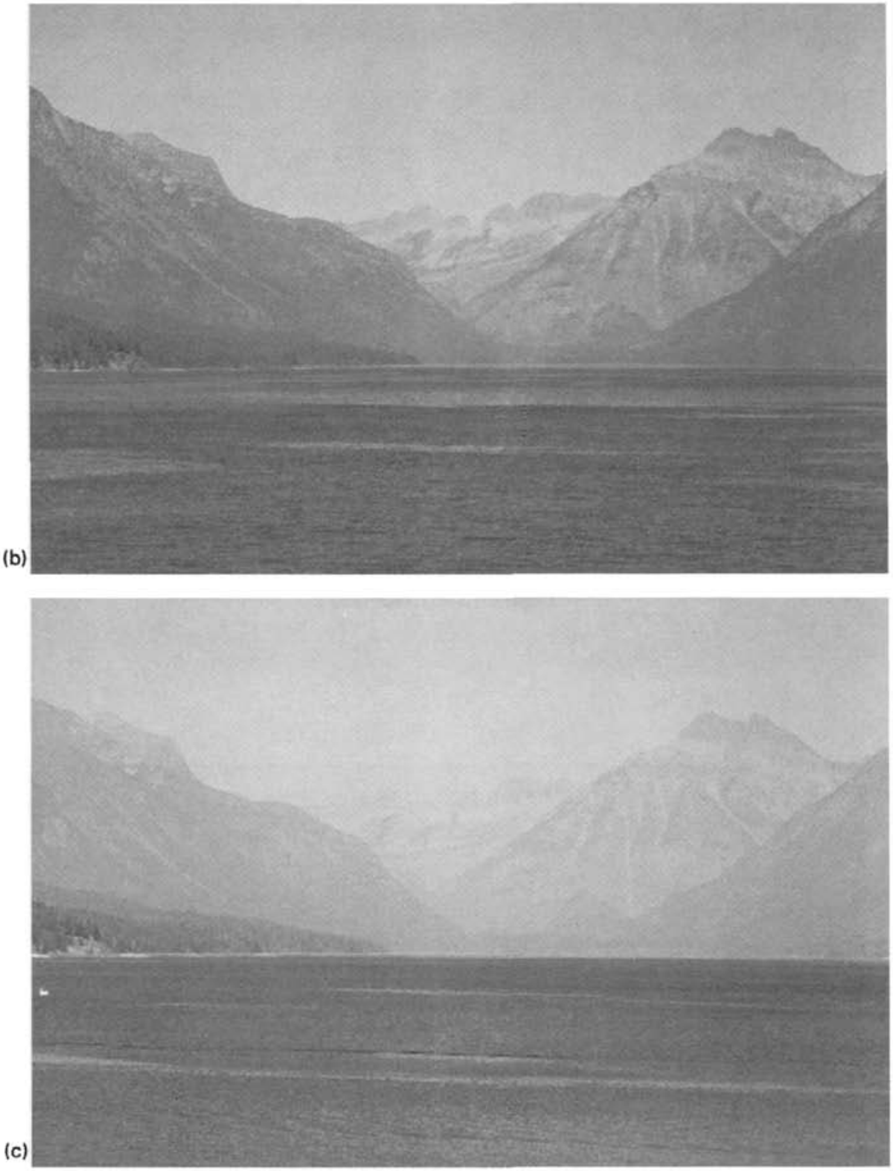


Figure 12 (*Continued*)

Figure 17 is a classic example of plume blight. In plume blight instances, specific sources such as those shown in Figure 18 emit pollutants into a stable atmosphere. The pollutants are then transported in some direction with little or no vertical mixing.

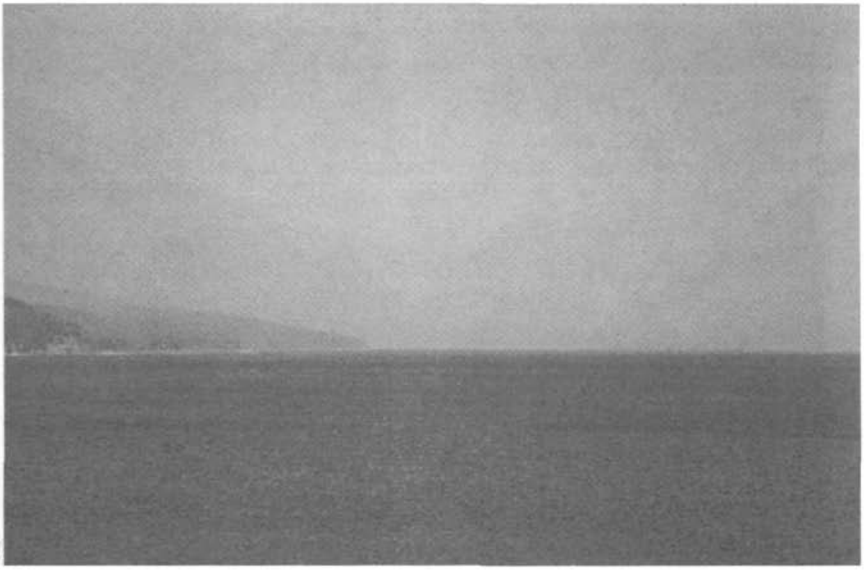


Figure 12 (*Continued*)

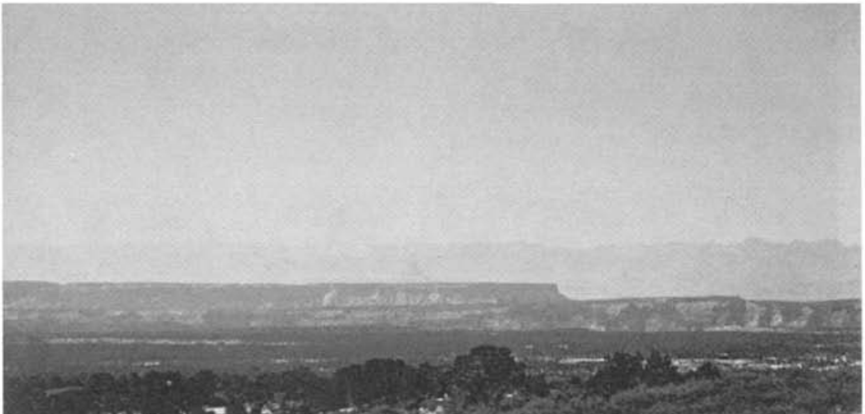


Figure 13 Effects of uniform haze on the Chuska Mountains as seen from Mesa Verde National Park. The atmospheric particulate concentration on the day this photograph was taken corresponded to 1 $\mu g/m^3$. See ftp site for color image.

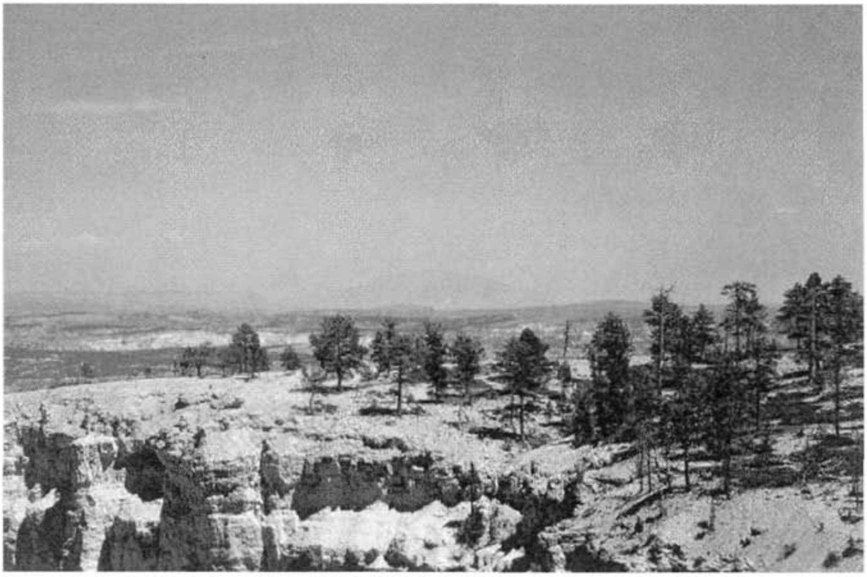


Figure 14 Uniform haze degrades visual air quality at Bryce Canyon National Park. The 130-km distant landscape feature is Navajo Mountain. Atmospheric particulate concentration on the day this photograph was taken is 3 $\mu g/m^3$. See ftp site for color image.

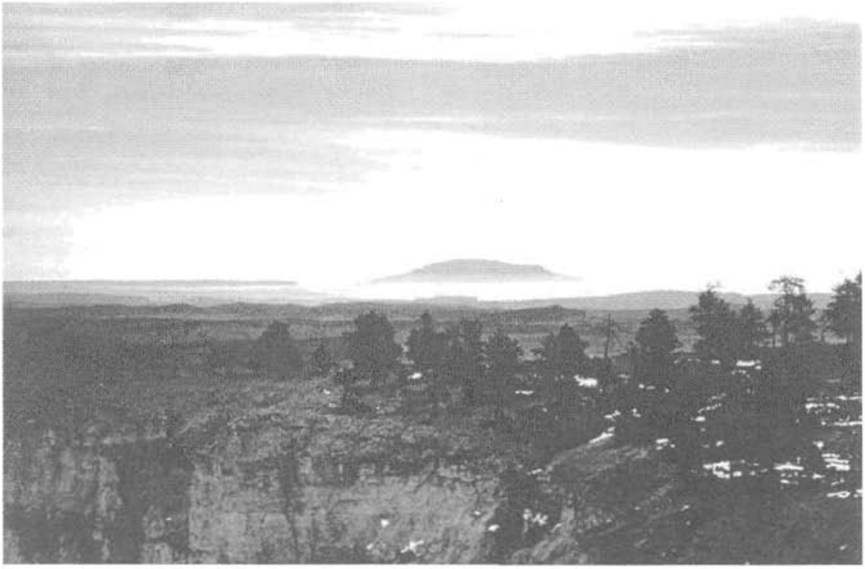


Figure 15 Navajo Mountain as seen from Bryce Canyon, showing the appearance of layered haze. The pollutants are trapped in a stable air mass that extends from the ground to about half-way up the mountain side. See ftp site for color image.

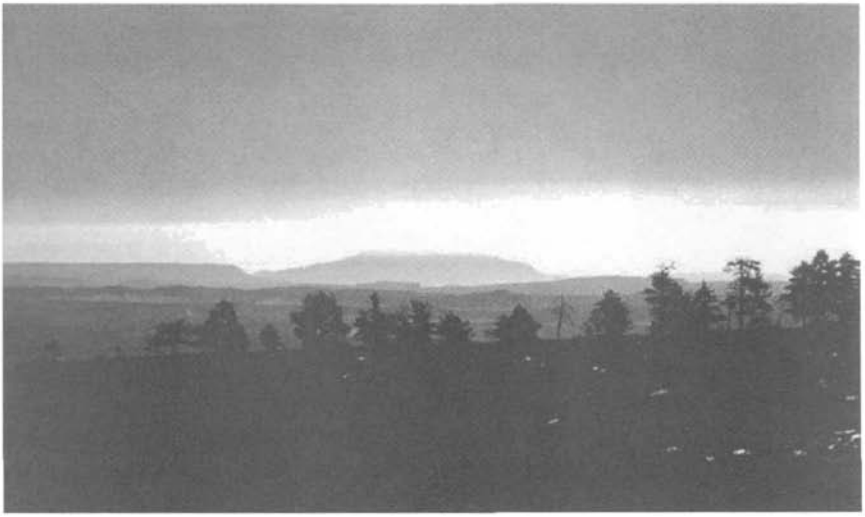


Figure 16 Photograph of Navajo Mountain similar to Figure 15 but with a suspended haze layer that obscures the top portion of the mountain. See ftp site for color image.

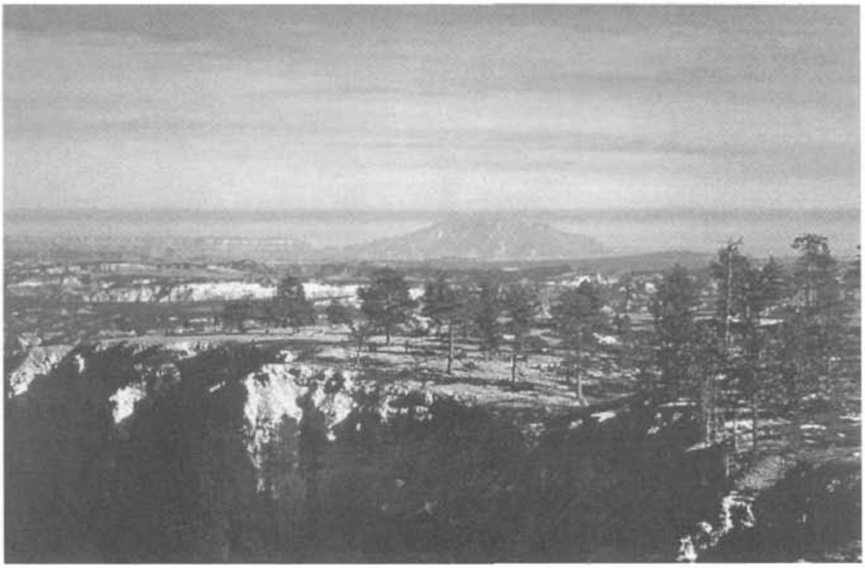


Figure 17 Classic example of "plume blight." The thin, dark plume on Navajo Mountain results from a point source emitting particulate matter into a stable atmosphere. See ftp site for color image.

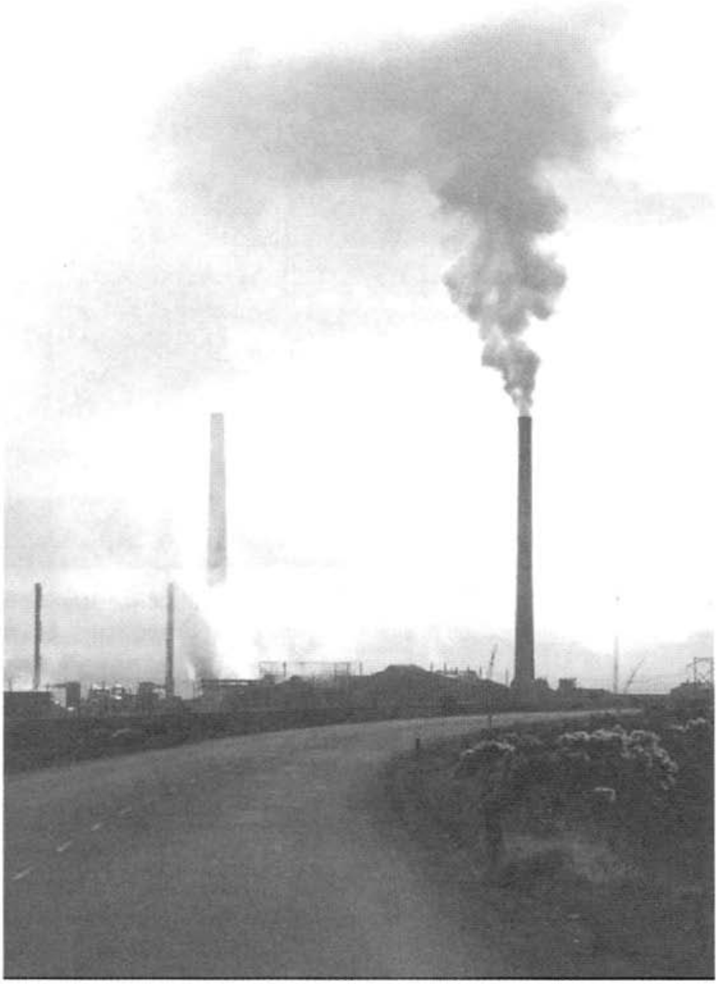


Figure 18 Example of one kind of point source that emits pollutants into the atmosphere. See ftp site for color image.

Figures 19, 20, 21, and 22 show other layered haze conditions that frequently occur at Grand Canyon and Mesa Verde National Parks. At Mesa Verde (Figure 22), much of the pollution comes from urban areas and the Four Corners and San Juan Power Plants, while at the Grand Canyon layered hazes are associated with smoke and nearby coal-fired power plants.

Figures 23 and 24 show the appearance of plumes containing carbon. In both of these cases the pollutants are being emitted from forest fires. However, Figure 23 shows the appearance of a specific forest fire plume, while Figure 24 shows the effect of viewing a vista through a concentration of particles containing carbon. In this instance, the vista is the north wall of the Grand Canyon as seen from the top of

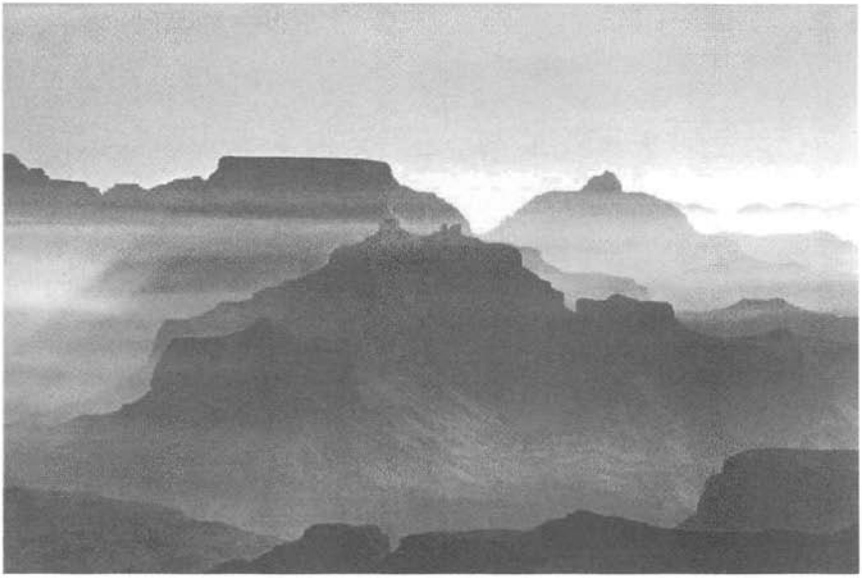


Figure 19 Smoke trapped by an inversion layer in the Grand Canyon. During the winter months, inversions are quite common in almost all parts of the United States. See ftp site for color image.

San Francisco Peaks in northern Arizona. Notice the overall "graying" and reduction of contrast of the distant scenic features. Remember that carbon absorbs all wavelengths of light and scatters very little. Thus the scene will always tend to be darkened.

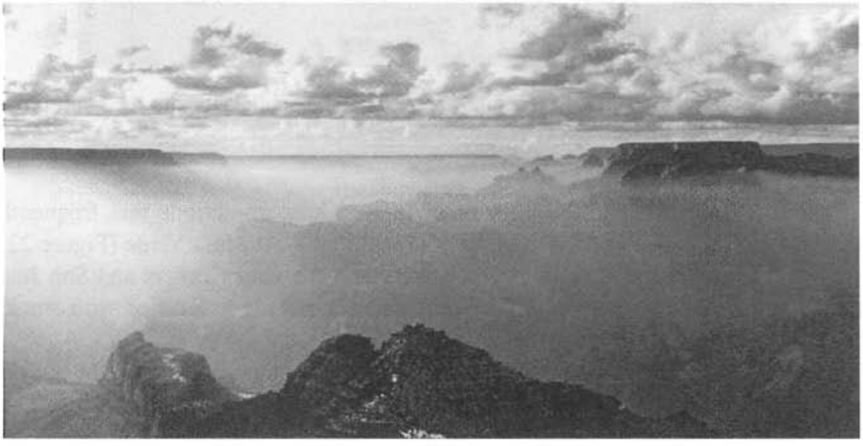


Figure 20 Example of power plant emissions trapped in an air inversion layer in the Grand Canyon. See ftp site for color image.

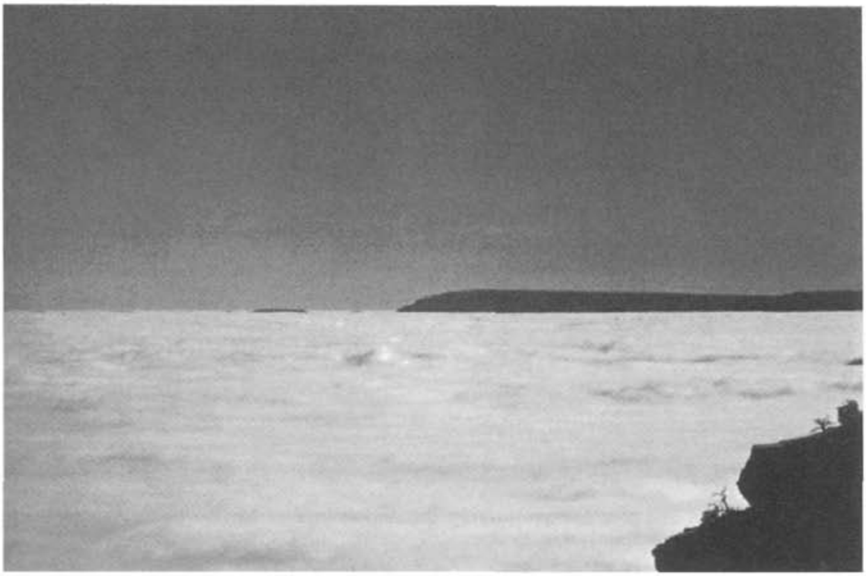


Figure 21 Effects of inversion layer in Grand Canyon. In this case, a cloud has formed within the canyon walls. See ftp site for color image.

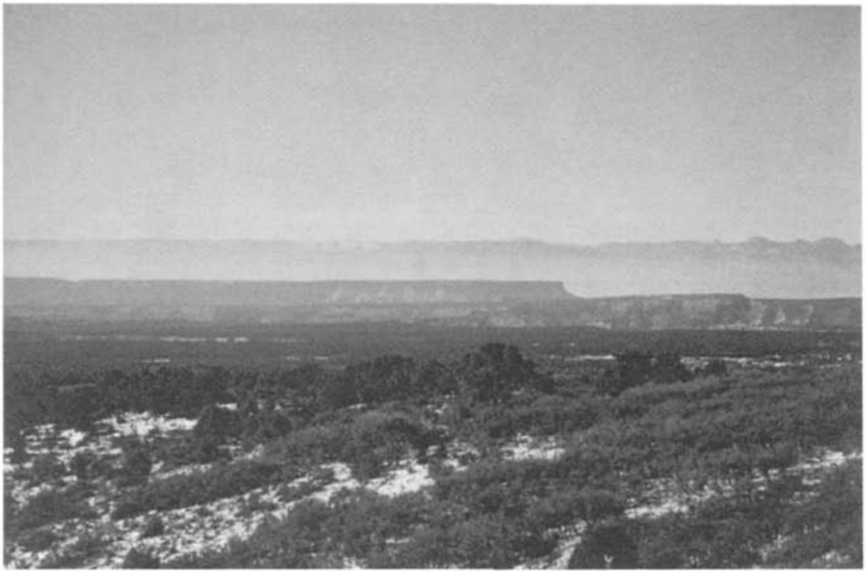


Figure 22 Effects of layered haze trapped in front of the Chuska Mountains as viewed from Mesa Verde National Park. This condition occurs 30 to 40% of the time during winter months. See ftp site for color image.

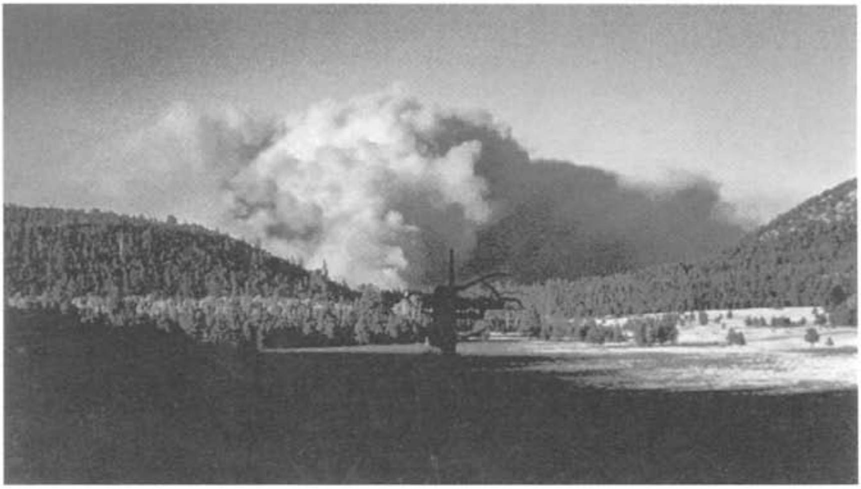


Figure 23 Forest fire plume exemplifying the appearance of carbon particles and demonstrating the effect of lighting. Where the plume is illuminated, it appears gray, but identical particles in the shadow of the plume appear dark or almost black. See ftp site for color image.

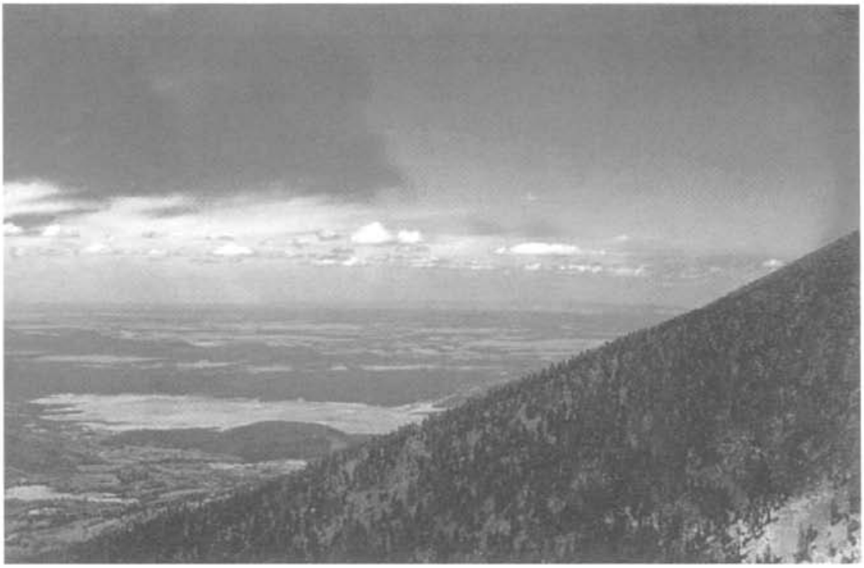


Figure 24 Example of how light-absorbing particles (in this case carbon) affect the ability to see a vista. Carbon absorbs all wavelengths of light and generally causes a "graying" of the overall scene. Shown here is the north wall of the Grand Canyon as seen from the top of the San Francisco Peaks in northern Arizona. See ftp site for color image.

Figure 25 shows the effects of illumination on the appearance of power plant plumes. The two plumes on the left are particulate plumes, while the two plumes on the right consist of water droplets. The plume on the far right, which is illuminated by direct sunlight, appears to be white. The second, identical water droplet plume, which is shaded, appears dark. The amount of illumination can have a significant effect on how particulate concentrations appear.

Figure 26 demonstrates how the effect of nitrogen dioxide gas (NO_2), in combination with varied background illumination, can combine to yield a very brown atmospheric discoloration. If a volume of atmosphere containing NO_2 is shaded and if light passes through this shaded portion of the atmosphere, the light reaching the eye will be deficient in photons in the blue part of the spectrum. As a consequence, the light will appear brown or reddish in color. However, if light is allowed to shine on, but not through, that same portion of the atmosphere, scattered light reaches the observer's eye and the light can appear to be gray in nature. Both of these conditions are shown in Figure 26. On the right side of the photo clouds shade the mixture of NO_2 and particulates. The same atmosphere, illuminated because the cloud cover is not present, appears almost gray in the middle portion of the photograph.

Effects of illumination are further illustrated in Figures 27*a*, and 27*b*. Figure 27 is an easterly view of the La Sal Mountains in southeastern Utah as seen from an elevated point that is some 100 km distant. The photograph in Figure 27*a* was taken at 9:00 A.M., while the photograph shown in Figure 27*b* was taken later in the day.

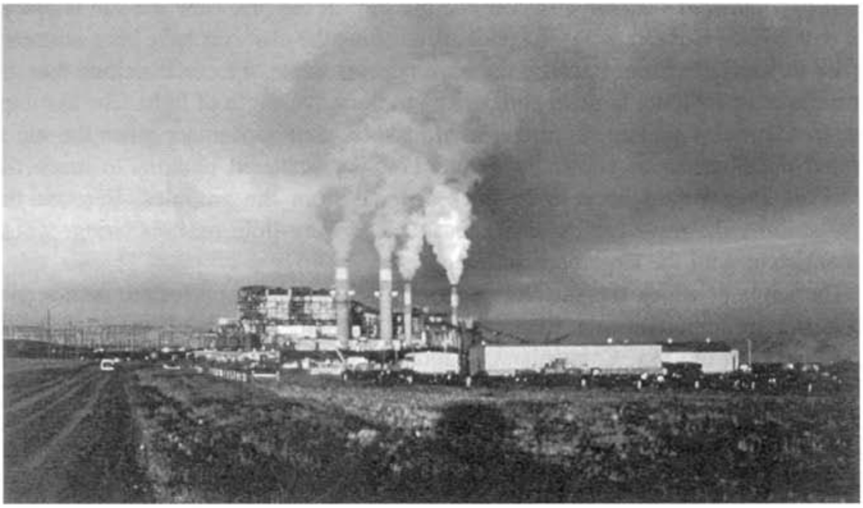


Figure 25 The effect of illumination on the appearance of plumes. The two plumes on the right are identical in terms of their chemical makeup, in that they are primarily water droplets. However, the far right plume is directly illuminated by the sun and the plume second from the right is shaded. The first plume appears white and the second appears almost black. The two plumes on the left are fly-ash plumes. See ftp site for color image.

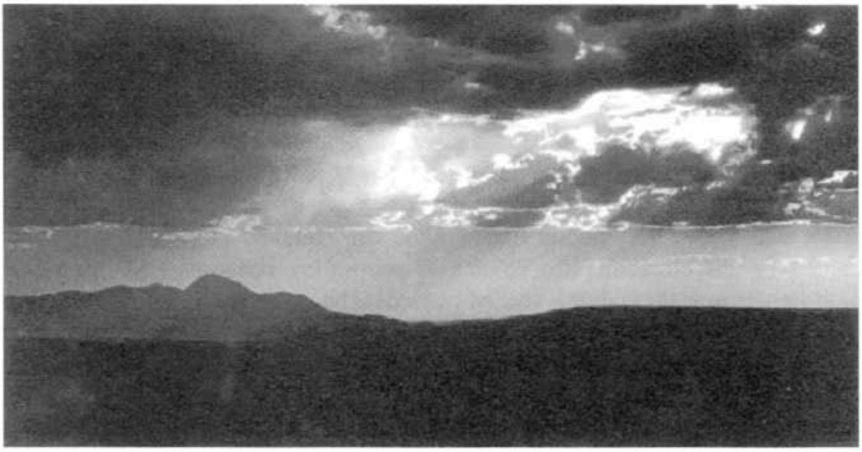


Figure 26 (see color insert) The brown discoloration resulting from an atmosphere containing nitrogen dioxide (NO_2) being shaded by clouds but viewed against a clear blue sky. Light scattered by particulate matter in the atmosphere can cominate light absorbed by NO_2, causing a gray or blue appearing haze (left side of photograph). See ftp site for color image.

These photographs show how these views, or vistas, appear when obscured by a layer of haze. In the first view the haze layer appears white, but the same air mass viewed later in the day has a dark gray appearance. This effect is entirely due to the geometry involved with the observer and the sun. In the first view the sun is low in the eastern sky. Consequently, the photons reaching the observer have been scattered in the forward direction. Because the haze appears white, we can conclude that the particles must be quite large in comparison to the wavelength of light. The assumption that particles are large is further reinforced by their appearance when the sun is behind the observer as shown in Figure 27*b*. For scattered photons to reach the observer, they would have to be backscattered from the particles. Because the haze appears dark, we can conclude that there is very little backscattering, which is consistent with the large particle hypothesis.

The angle at which the sun illuminates a vista or landscape feature (sun angle) plays another important role. Figures 28*a* to 28*d* exemplify this effect. The view is from Island in the Sky, Canyonlands National Park, looking out over Canyonlands with its many colorful features toward the 50 km distant La Sal Mountains. Figure 28*a* shows how the canyon appears when it is in total shadow (6:00 A.M.). Figures 28*a* to 28*c* show a progressively higher sun angle until in Figure 28*d* the scene is entirely illuminated. In each case, the air quality is the same. The only change is in the angle with which the sun illuminated the vista. There are primarily two reasons for the apparent change in visual air quality. First, at higher sun angles, there is less scattering of light by the intervening atmosphere in the direction of the observer. Second, the vista reflects more light; consequently, more image-forming information (reflected photons from the vista) reaches the eye. The contrast detail and scene are enhanced.

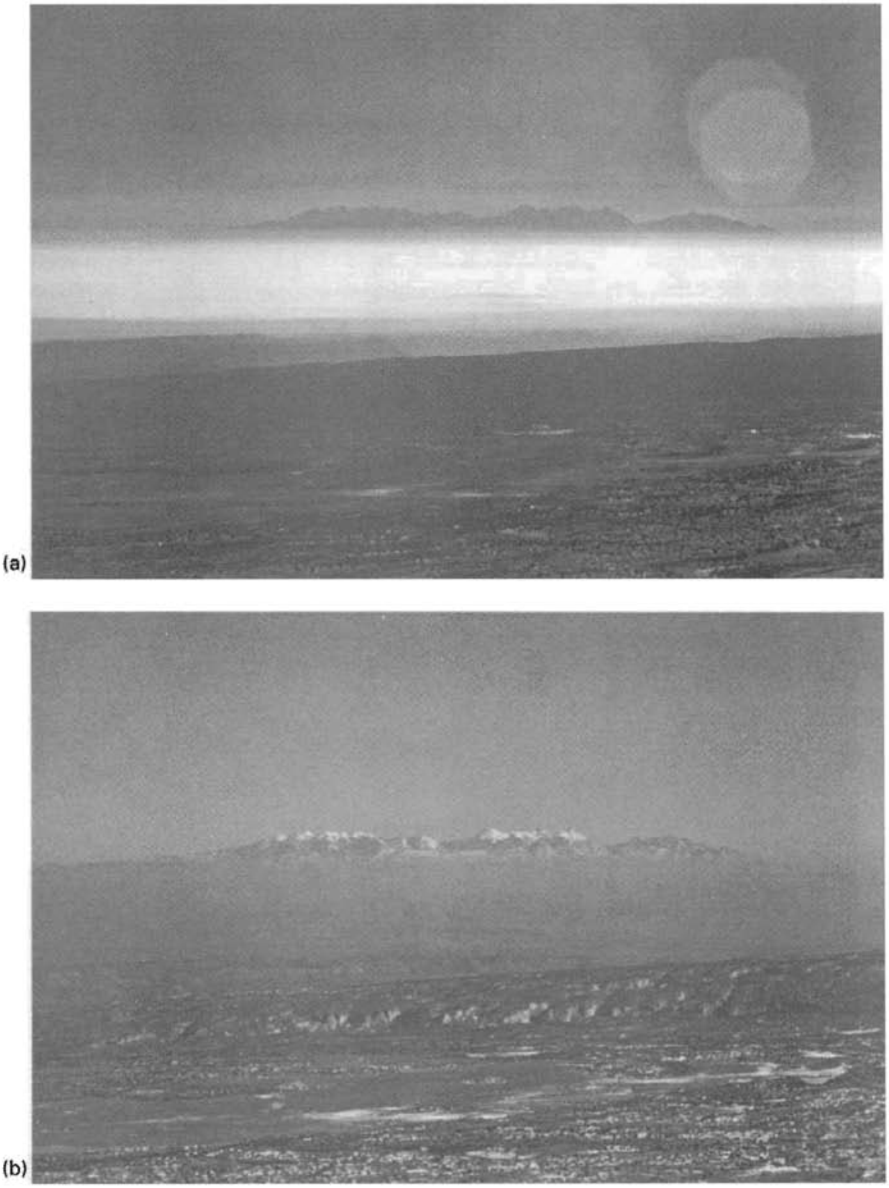


Figure 27 Photographs show how the same haze trapped in an inversion layer looks under forward and backscatter conditions. In (*a*) under forward scattering conditions (morning), the haze appears white; in (*b*) the identical haze, viewed in the afternoon during backscatter conditions, is dark or gray. Because most of the light energy is scattered in the forward direction (white haze), it can be concluded that the particles must be quite large in comparison to the wavelength of light. See ftp site for color image.

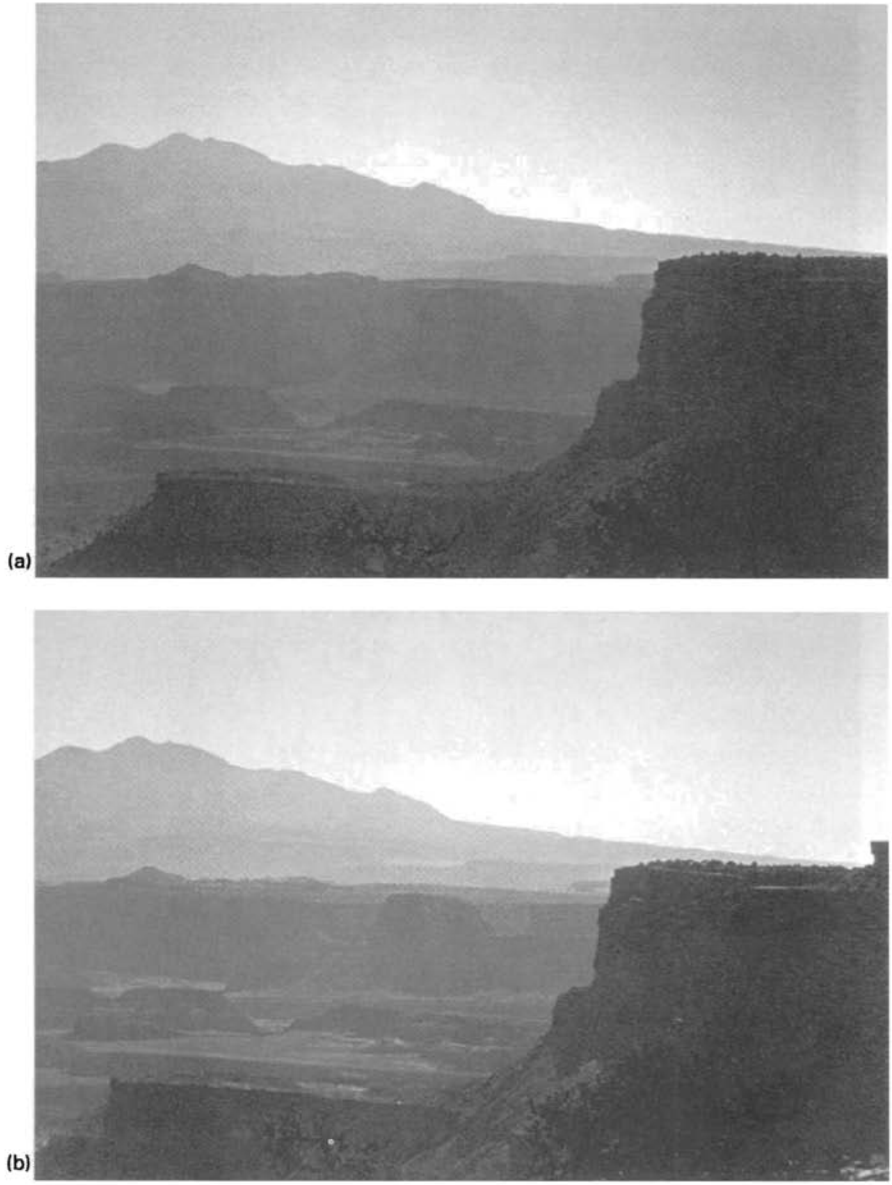


Figure 28 Four photographs showing the effect of shifting sun angle on the appearance of a vista as seen from Island in the Sky, Canyonlands National Park. In each photograph, the air quality is the same. In (*a*) (6:00 A.M.) the sun angle–observer–vista geometry results in a large amount of scattered air light (forward scattering) added to the sight path, but minimal amount of imaging light reflected from the vista. Figures 28*b* and 28*c* show a progressively higher sun angle until in Figure 28*d*, the scene is entirely illuminated. Scattered light is minimized and reflected; imaging light is at a maximum. See ftp site for color image.

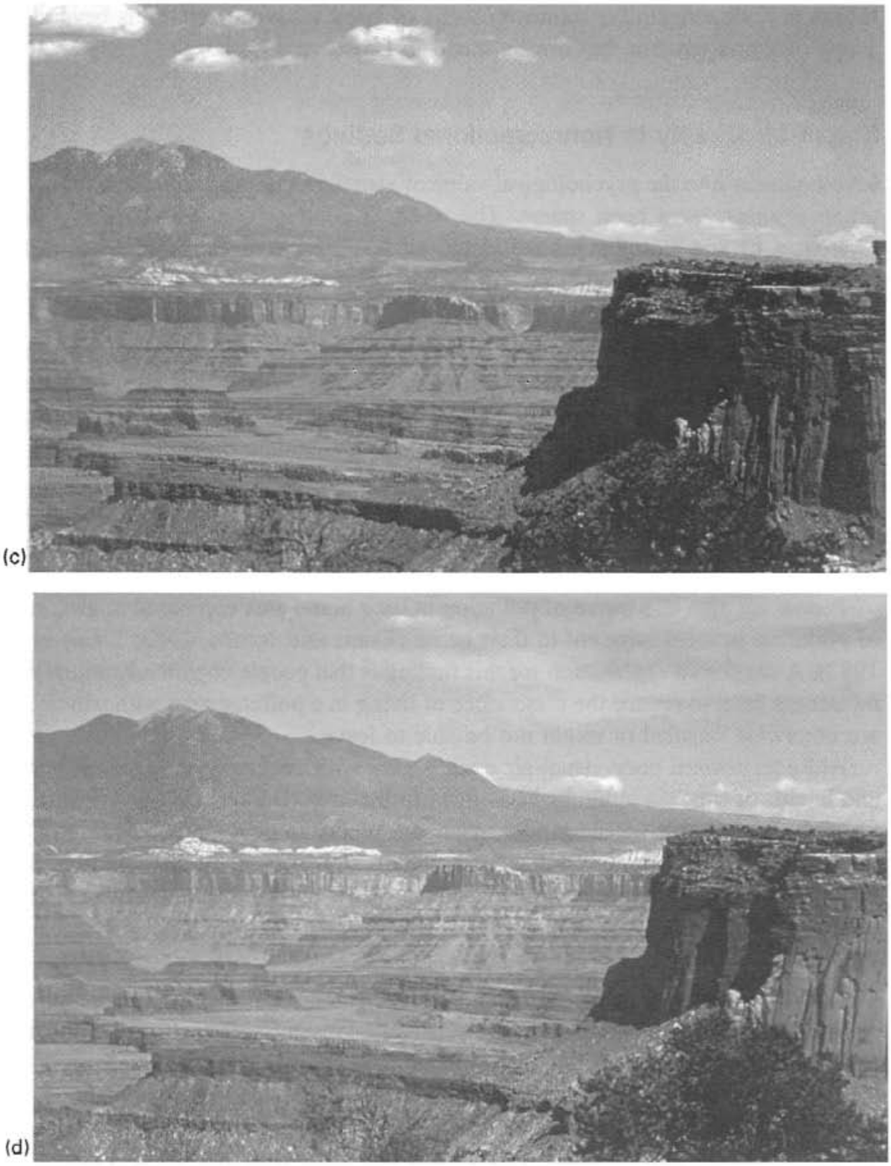


Figure 28 Continued

5 VALUE OF GOOD VISUAL AIR QUALITY

Efforts to define and quantify the value of good visual air quality have generally followed two courses. One emphasis has been on monetary costs to resource degradation and human health. The other emphasis has been on the psychological value of visual air quality in the context of recreational and nonrecreational settings. A thorough review regarding monetary value of good visual air quality is beyond the scope of this discussion but can be found in Brown and Callaway (1990).

Visual Air Quality in Nonrecreational Settings

Investigations into the psychological value of visual air quality in nonrecreational, or urban settings, have been sparse. The research conducted in this area examined awareness of and attitudes toward visual air quality and investigated relationships between visual air quality, stress, and human behavior.

Public Perception of Visual Air Quality. Survey research of public awareness of visual air quality using direct questioning typically reveals that 80% or more of the respondents are aware of poor visual air quality, and that poor visibility and media publicity are the primary factors that precipitate the awareness (Cohen et al., 1986). These surveys have also shown that awareness is not uniform across the general population of a given area. Persons with higher income and educational levels tend to be more aware of poor visual air quality than those with lower income and educational levels.

People are also less aware of pollution in their home area compared to awareness of pollution in areas adjacent to their home (Evans and Jacobs, 1982; Evans et al., 1982). A suggested explanation for this finding is that people cognitively adjust their awareness level to reduce the dissonance of living in a polluted area with which they are otherwise satisfied or might not be able to leave.

Attitudes toward poor visual air quality vary with socioeconomic status, health, and length of time an individual has lived in the area (Barker, 1976). Affluent and well-educated people consider poor visual air quality to be a more serious problem than others. People who are not economically tied to sources of air pollution, have respiratory ailments, or are new to an area also show the strongest negative reactions to reduced VAQ.

Visual Air Quality and Stress. Reduced visual air quality is an ambient environmental stressor because it is a relatively constant and unchanging situation over which one has little direct control (Campbell, 1983). The associated stress and lack of control is chronic, not salient, and may be manifested in heightened levels of anxiety, tension, anger, fatigue, depression, and feelings of helplessness (Evans et al., 1987; Ziedner and Shechter, 1988). How one deals with this stress is dependent on coping behavior and ability to adapt. The relationship between stress due to poor visual air quality and mental health is poorly understood. However, results from a study conducted by Rotton and Frey (1982) showed that as visual air quality decreased, emergency calls for psychiatric disturbances increased.

Visual Air Quality and Behavior. Evans et al. (1982) found that persons who recently moved to Los Angeles from areas with good visual air quality consistently reduced outdoor activities during periods of reduced visual air quality compared with longer-term residents. Studies have also reported reduced altruism and increased hostility and aggression during periods of poor air quality (Cunningham, 1979; Jones and Bogat, 1978; Rotton et al., 1979). The relationship between aggression, hostility, and visual air quality is curvilinear with feelings of aggression and hostility increasing to a certain point and then dropping off and yielding to a desire to withdraw and escape from the situation. Evans and Cohen (1987) suggest that individuals adjust to poor visual air quality through adaptation and coping behaviors by altering their judgment of air quality based on current and previous exposure.

Visual Air Quality in Recreational Settings

During the past decade, an experience-based demand model has been developed to assess demand for recreational opportunities. The model incorporates visitor demand for activities, for social/physical/managerial site attributes, and for the realization of specific psychological satisfactions.

The model was used to investigate the psychological value of good visual air quality at Grand Canyon, Mesa Verde, Great Smoky Mountains, Mount Rainier, and Everglades National Parks using on-site interviews and mail-back surveys. The purpose was to evaluate the importance of visual air quality relative to other park attributes, to determine if visitors were accurately aware of changes in visibility, and to ascertain whether relationships existed between visual air quality and visitor satisfaction (Ross et al., 1985, 1987).

Importance of Good Visual Air Quality. The importance of good visual air quality to park visitors was evaluated by having visitors rate how important specific park attributes were to their recreational experience. Cluster analysis was used to statistically identify similar types of attributes based on response patterns. Grand Canyon National Park's attributes, their corresponding mean importance scores, and cluster formation are shown in Figure 29. The "clean, clear air" attribute ranked third in importance and combined with attributes that are descriptive of a clean, natural setting, which, as a group, were slightly more important than the cluster of view-related attributes. This indicates that visitors interpret "clean, clear air" as being an integral part of the cleanliness of the park and as such, an important part of the overall recreational experience sought at Grand Canyon.

The importance of a natural, clean environment with clean air was not unique to Grand Canyon visitors. Figure 30 shows that similar findings resulted from the other studies regardless of park location or overall theme. The cleanliness attribute cluster, which included "clean, clear air," was the most important cluster at all five parks.

Visitor Awareness of Visual Air Quality. In a random sample, nearly 1800 visitors at Grand Canyon National Park were asked during an interview if they were aware of any haze and, if so, how hazy they thought it was. Results from correlation analysis between awareness of haze and standard visual range measures showed that visitors' awareness of haze increased as visibility decreased. Correlation coefficients

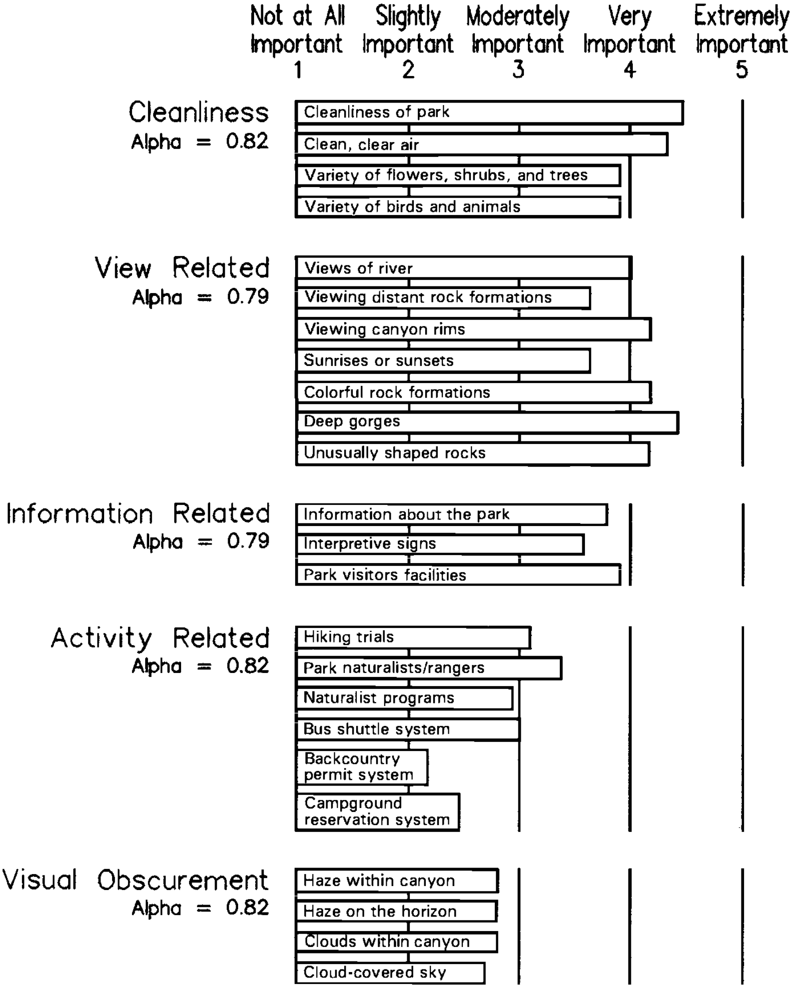


Figure 29 Attributes, attribute mean scores, attribute clusters, and attribute cluster mean scores for the Grand Canyon National Park visitor survey. See ftp site for color image.

were also calculated between visitor awareness of haze and ratings on enjoyment of the view, impact of haze on overall park enjoyment, and satisfaction with the "clean, clear air" attribute. Results showed that as awareness of reduced visibility increased, enjoyment with the view, overall park enjoyment, and satisfaction with the "clean, clear air" attribute decreased.

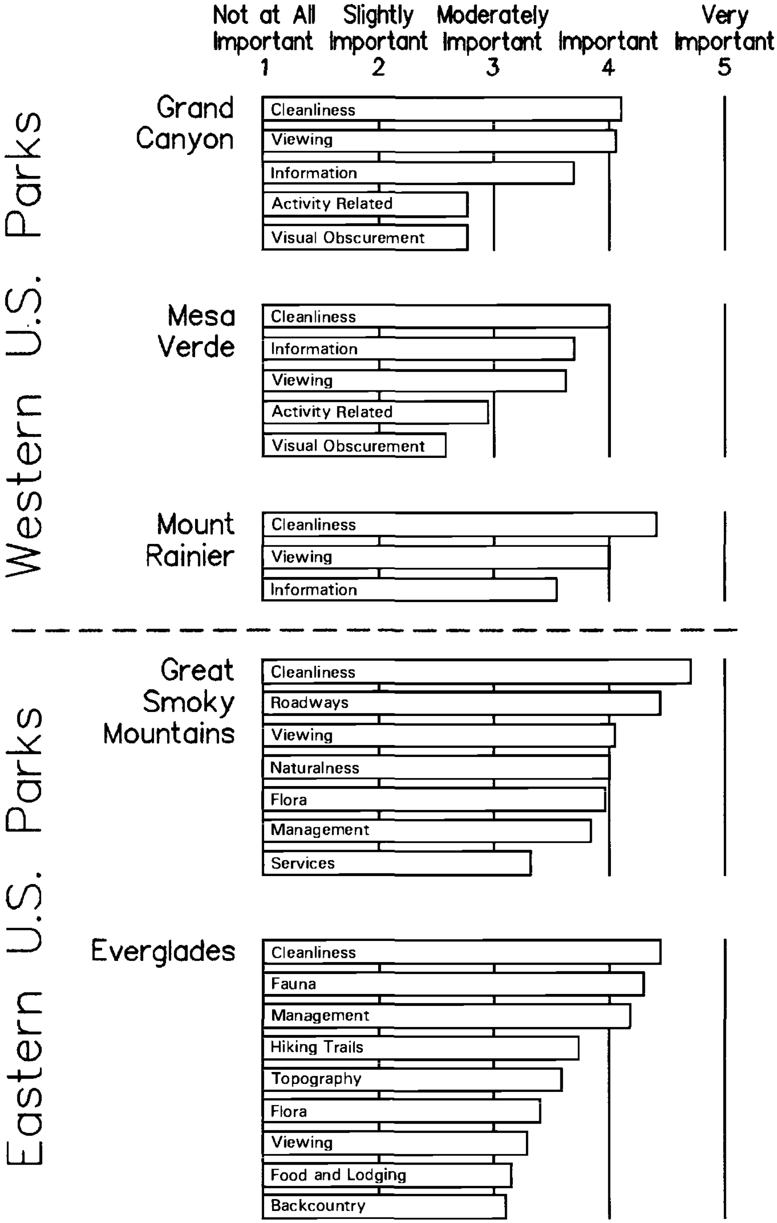


Figure 30 Relative importance of attribute clusters at five national parks. See ftp site for color image.

TABLE 2 Summary of Contrast and Color Change Threshold Data

Contrast	ΔE	Percent Detection	Edge	Reference
0.003[a]	—	50	Sharp	Blackwell (1946)
0.014	—	?	Sharp	Lowry (1931, 1951)
0.007[b]	—	?	Sharp	Howell and Hess (1978)
0.009[b]	—	?	Diffuse	
0.016[c]	—	?	Sharp	
—	1	30	Sharp	Jaeckel (1973)
—	2	50	Sharp	
—	3	70	Sharp	
—	4	90	Sharp	
0.006	1	10	Diffuse	Malm et al. (1980)
0.009	1.5	25	Diffuse	
0.014	2.3	50	Diffuse	
0.02	3.3	75	Diffuse	
0.025	4.2	90	Diffuse	
0.01	—	90	Sharp	Loomis et al. (1985)
0.005[d]	—	70	Sharp	Malm et al. (1985)
0.010[e]	—	70	Sharp	
0.020[f]	—	70	Diffuse	Ross et al. (1988)
0.007[d]	—	70	Diffuse	Ross et al. (1990)
0.025[e]	—	70	Diffuse	

[a] The most sensitive contrast reported for largest size of stimulus and largest luminance and longest response time evaluated (probably the minimum possible threshold).

[b] The most sensitive contrast reported at a spatial frequency of 3 cycles/degree.

[c] Threshold contrast for sharp objects at low spatial frequencies.

[d] Minimum threshold for 0.36° wide plumes.

[e] Maximum threshold for all size plumes tested.

[f] Threshold contrast reported for light-colored, diffuse edge hazes of varying size.

Visual Air Quality and Recreational Behavior. A laboratory study conducted by Malm et al. (1984) at Grand Canyon National Park examined how visual air quality might affect visitor behavior. Participants examined sets of photographs with different levels of visual air quality and indicated how they would be willing to spend a given amount of time either driving to a lookout point or touring an archaeological site. The study concluded that subjects place a high value on visual air quality and would be willing to significantly alter behavior for increased visual air quality. For example, subjects would be willing to spend an additional 2.5 h driving time to view a dominant distant landscape for a 0.01 km^{-1} reduction in atmospheric extinction. The study also showed that vistas that lacked color and texture were insensitive to increases in atmospheric extinction.

Disclaimer. The assumptions, findings, conclusions, judgments, and views presented herein are those of the author and should not be interpreted as necessarily representing official National Park Service policies.

REFERENCES

Barker, M. Planning for environmental indices: Observer appraisals of air quality, in K. Craig and E. Zube (Eds.), *Perceiving Environmental Quality: Research and Applications*, 175–203, Plenum, New York, 1976.

Blackwell, H. R., Contrast thresholds of the human eye, *J. Opt. Soc. Am.*, *36*, 624, 1946.

Boswell, G. T., *How People Look at Pictures*, University of Chicago Press, Chicago, IL, 1975.

Brown, G. M., and J. M. Callaway, Methods for valuing acidic deposition and air pollution efforts, National Acid Precipitation Assessment Program (NAPAP): State of Science, Report No. 27, Washington, DC, 1990.

Campbell, F. W., Ambient stressors, *Environ. Behavior*, *15*, 355–380, 1983.

Campbell, F. W., and J. J. Kulikowski, Orientational selectivity of the visual cell of the cat, *J. Physiol. (London)*, *187*, 437, 1986.

Campbell, F. W., B. Cleveland, G. F. Cooper, and C. Enroth-Cogell, The spatial selectivity of the visual cells of the cat, *J. Physiol. (London)*, *198*, 237, 1968.

Campbell, F. W., and J. G. Robson, Application of Fourier analysis to the modulation response of the eye, *J. Opt. Soc. Am.*, *54*, 581A, 1964.

Carlson, C. R., and R. W. Cohen, *Image Descriptors for Displays: Visibility of Displayed Information*, RCA Laboratories, Princeton, NJ, 1978.

Cohen, S., G. W. Evans, D. Stokols, and D. S. Krantz, *Behavior, Health, and Environmental Stress*, Academic, New York, 1986.

Cornsweet, T., *Visual Perception*, Academic, New York, 1970.

Cunningham, M., Weather, mood, and helping behavior: Quasi-experiments with the sunshine Samaritan, *J. Per. Soc. Psych.*, *37*, 1947–1956, 1979.

Evans, G. W., and S. Cohen, Environmental stress, in D. Stokols and I. Altman (Eds.), *Handbook of Environmental Psychology*, 571–602, Wiley, New York, 1987.

Evans, G. W., and S. V. Jacobs, Air pollution and human behavior, in G. W. Evans (Ed.), *Environmental Stress*, 105–132, Cambridge University Press, New York, 1982.

Evans, G. W., S. V. Jacobs, and N. B. Frager, Behavioral responses to air pollution, in A. Baum and J. Singer (Eds.), *Advances in Environmental Psychology*, Vol. 4, 237–270, Erlbaum, New York, 1982.

Evans, G. W., S. V. Jacobs, D. Dooley, and R. Catalano, The interaction of stressful life events and chronic strains on community mental health, *Am. J. Comm. Psych.*, *15*, 23–24, 1987.

Faugeras, O. D., Digital color image processing within the framework of a human visual model, *IEEE Trans. Acoust. Speech Sig. Process*, *ASSP-27*, 380–393, 1979.

Hall, C. F. and E. L. Hall, A nonlinear model for the spatial characteristics of the human visual system, *IEEE Trans. Syst. Man. Cybernet. SMC-7*, 161–170, 1977.

Henry, R. C., The application of the linear system theory of visual acuity to visibility reduction by aerosols, *Atmos. Environ.*, *11*, 697, 1977.

Henry, R. C., The Human Observer and Visibility—Modern psychophysics applied to visibility degradation, in *View on Visibility: Regulatory and Scientific*, 27–35, Air Pollution Control Association, Pittsburgh, PA, 1979.

Henry, R. C., Improved predictions of plume perception with a human visual system model, *J. Pollut. Control Assoc.*, *36*, 1353–1356, 1986.

Henry, R. C., Psychophysics, visibility, and perceived atmospheric transparency, *Atmos. Environ.*, *21*, 159–164, 1987.

Henry, R. C., and L. V. Matamala, Prediction of color matches and color differences in the outdoor environment, in C. V. Mathai (Ed.), *Transactions of Visibility and Fine Particles*, 554–561, Air and Waste Management Association, Pittsburgh, PA, 1990.

Henry, R. C., J. F. Collins, and D. Hadley, Potential for quantitative analysis of uncontrolled routine photographic slides, *Atmos. Environ.*, *15*, 1959, 1981.

Hill, A. C., Measuring How Landscape Color Affect Aesthetic Value, in *Transactions of Visibility and Fine Particles*, Air and Waste Management Association, Pittsburgh, PA, 570–581, 1990.

Hodkinson, J. R., Calculations of color and visibility in urban atmospheres polluted by gaseous NO_2, *J. Air Water Pollut. Int.*, *10*, 137–144, 1966.

Howell, E. R., and R. F. Hess, The functional area for summation to threshold for sinusoidal gratings, *Vision Res.*, *18*, 369–374, 1978.

Jaeckel, S. M., Utility of color-difference formulas for match acceptability decisions, *Appl. Opt.*, *12*, 1299–1316, 1973.

Jones, J. W., and G. A. Bogat, Air pollution and human aggression, *Psych. Rpts.*, *43*, 721–722, 1978.

Latimer, J. A., R. W. Bergstrom, S. R. Hayes, M. K. Liu, J. H. Seinfeld, G. Z. Whitten, M. A. Wojcik, and M. J. Hillyer, *The Development of Mathematical Models for the Prediction of Anthropogenic Visibility Impairment*, EPA Report No. 4503-78-110a,b,c, Environmental Protection Agency, Research Triangle Park, NC, 1978.

Latimer, D. A., T. C. Daniel, and H. Hogo, *Relationship between Air Quality and Human Perception of Scenic Areas*, Publication No. 4323, American Petroleum Institute, Washington, DC, 1980.

Latimer, D. A., H. Hogo, D. H. Hern, and T. C. Daniel, Effects of visual range on the beauty of national parks and wilderness area vistas, in R. D. Rowe and L. G. Chestnut (Eds.), *Managing Air Quality and Scenic Resources at National Parks and Wilderness Areas*, Westview Press, Boulder, Co., 1983.

Loomis, R. J., M. J. Kiphart, D. B. Garnaard, W. C. Malm, and J. V. Molenar, Human perception of visibility impairment, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Detroit, MI, 1985.

Lowry, E. M., The photometric sensibility of the eye and the precision of photometric observations, *J. Opt. Soc. Am.*, *21*, 32, 1931.

Lowry, E. M., The luminance discrimination of the human eye, *J. Soc. Motion Pictures Television Eng.*, 1951.

Malm, W. C., An examination of the ability of various physical indicators to predict judgment of visual air quality, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Pittsburgh, PA, 1985.

Malm, W. C., and R. C. Henry, Regulatory perspective of visibility research needs, paper presented at the Eightieth Annual Meeting of the Air Pollution Association, New York, NY, 1987.

Malm, W. C., and M. Pitchford, The use of an atmospheric quadratic detection model to assess change in aerosol concentrations to visibility, paper presented at the Eighty-Second Annual Meeting of the Air Pollution Control Association, Anaheim, CA, 1989.

Malm, W. C., M. Kleine, and K. Kelley, Human perception of visual air quality (layered haze), paper presented at the Conference on Visibility at the Grand Canyon, AZ, 1980.

Malm, W. C., K. Kelley, J. Molenar, and T. Daniel, Human perception of visual air quality (uniform haze), *Atmos. Environ.*, *15*, 1875, 1981.

Malm, W. C., P. Bell, and G. E. McGlothin, Field testing: A methodology for assessing the importance of good visual air quality, in *Proceedings of the Seventy-Seventh Annual Meeting of the Air Pollution Control Association*, Pittsburgh, PA, 1984.

Malm, W. C., D. M. Ross, R. Loomis, J. Molenar, and H. Iyer, An examination of the ability of various physical indicators to predict perception thresholds of plumes as a function of their size and intensity, in P. J. Bhardwaja (Ed.), *Visibility Protection Research and Policy Aspects*, Air Pollution Control Association. Pittsburgh, PA, 1987.

Middleton, P., T. R. Stewart, D. Ely, and C. W. Lewis, Physical and chemical indicators of urban visual air quality, *Atmos. Environ.*, *18*, 861–870, 1984.

Mie, G. *Ann. Phy. Bd. 2*, *25*, *IV*, Fogle, Netherlands, 1908.

Nixon, J. K., Absorption coefficient on NO_2 in the visible spectrum, *J. Chem. Phys.*, *8*, 157, 1940.

Pitchford, M. L., and W. C. Malm, Development and applications of a standard visual index, *Atmos. Environ.*, *28*, 1049–1054.

Ross, D. M., W. C. Malm, and R. J. Loomis, The psychological variation of good visual air quality by national park visitors, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Detroit, MI, 1985.

Ross, D. M., W. C. Malm, and R. J. Loomis, An examination of the relative importance of park attributes at several national parks, in P. S. Bhardwaja (Ed.), *Transactions of Visibility Protection: Research and Policy Aspects*, Air Pollution Control Association, Pittsburgh, PA, 1987.

Ross, D. M., W. C. Malm, H. K. Iyer, and R. J. Loomis, Human detection of layered haze using natural scene slides with a signal detection paradigm, paper presented at the Eighty-First Annual Meeting of the Air Pollution Control Association, Dallas, TX, Malm, 1988.

Ross, D. M., W. C. Halm, H. K. Iyer, and R. J. Loomis, Human visual sensitivity to layered haze using computer generated images, in C. V. Mathai (Ed.), *Transactions of Visibility and Fine Particles, Air and Waste Management Association*, 582–595, Pittsburgh, PA, 1990.

Ross, D. M., W. C., Malm, and H. K. Iyer, Human visual sensitivity to plumes with a Gaussian luminance distribution: Experiments to develop an empirical probability of detection model, *J. Air Waste Mgmt. Assoc.*, *47*, 370–382, 1997.

Rotton, J., and J. Frey, Atmospheric conditions, seasonal trends, and psychiatric emergencies, in *Replications and Extensions*, American Psychological Association, Washington, DC, 1982.

Rotton, J., T. Barry, M. Milligan, and M. Fitzpatrick, The air pollution experience and interpersonal aggression, *J. Appl. Psych. 9*, 397–412, 1979.

Stewart, T. R., P. Middleton, and D. W. Ely, Urban visual air quality judgements: Reliability and validity, *J. Environ. Psych.*, *3*, 129–145, 1983.

Stewart, T. R., P. Middleton, M. Downton, and D. Ely, Judgements of photographs versus field observations in studies of perception and judgment of the visual environment, *J. Environ. Psych.*, *4*, 283–302, 1984.

vandeHulst, H. C., *Light Scattering by Small Particles*, Dover, New York, 1981.

Ziedner, M., and M. Shechter, Psychological responses to air pollution: Some personality and demographic correlates, *J. Environ. Psych. 8*, 191–208, 1988.

CHAPTER 17

CLOUD CHEMISTRY

STEPHEN E. SCHWARTZ

1 INTRODUCTION

The term *cloud chemistry* is considered here to comprise both cloud composition and reactions that take place in clouds. Clouds are a very special subset of the atmosphere because they present substantial amounts of condensed-phase water (liquid or solid) that can dissolve gases that would otherwise be present in the gas phase, and, as a consequence of condensed-phase reactions, permit reactions to occur that would not otherwise occur or would be much slower. In this sense clouds may be considered to serve as catalysts of atmospheric reactions.

The uptake and reaction of material in clouds, especially sulfur and nitrogen oxides and acids, has received particular attention in the context of gaining improved understanding of the processes responsible for acid deposition. Consequently, the examples developed here focus on these chemical systems. However, much of the resulting understanding of these phenomena is applicable more generally to other systems.

2 CLOUD PHYSICAL PROPERTIES PERTINENT TO CLOUD CHEMISTRY

Clouds consist of a suspension of liquid or solid (ice) particles in air. Thus, formally, a cloud is an aerosol, a suspension of particles in air. However, it is useful to distinguish clouds from clear-air (noncloud) aerosols. The cloud environment is slightly supersaturated with respect to liquid water or ice, respectively. The typical amount of condensed-phase water is 0.1 to 1 g/m^3 (roughly equivalent to 0.1 to 1/kg

This chapter has been co-authored under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

of air). The amount of condensed-phase water is substantially lower in cirrus clouds and in polar stratospheric clouds. For condensed-phase amounts substantially exceeding 1 g/m^3, coagulation occurs and precipitation rapidly develops, removing condensed-phase water from the cloud.

A liquid water content of 1 g/m^3 corresponds (within the approximation that the density of water is 1 kg/m^3) to a liquid water volume fraction $L = 1 \times 10^{-6}$, or one part per million by volume. On dimensional grounds the separation between cloud droplets is $\sim L^{-3}$ times the diameter of the droplets; for $L = 1 \times 10^{-6}$, the average interdrop separation is ~100 times the drop diameter. Thus clouds must be considered a sparse suspension of condensed-phase water. Clouds are mostly air. Thus any consideration of cloud chemistry must deal with both the gas phase and the condensed phase.

Despite this sparseness, clouds still contain much more condensed-phase material than cloud-free air. Consider a clear-air aerosol of mass loading of 100 μg/m^{-3}; within the approximation of density equal to 1 kg/m^3, the corresponding condensed-phase volume fraction is 1×10^{-10}. The much greater mass loading of a cloud leads among other things to its greater light scattering, the most distinguishing feature of clouds.

Clouds form when air, containing water vapor, is cooled to a temperature below its dew point. Typically this occurs when air is lifted, for example, buoyant rise of a convective parcel, or larger scale gentle upward motion of warm air over denser cooler air. Cooling by conduction can also be important, for example, in ground fogs, as can radiative cooling. The condensation process defines the number concentration of cloud droplets by activating a certain fraction of preexisting aerosol particles into cloud droplets (see Chapter 19). The number concentration is typically 100 to 1000/cm^3 or 10^8 to 10^9/m^3. Thus within the cloud the condensed-phase water is finely suspended. For droplet concentration of 1×10^9/m^3 and liquid water volume fraction of 1×10^{-6} m^3/m^3, the corresponding volume of an individual droplet is 1×10^{-15} m^3 and the corresponding diameter $\sim 1 \times 10^9$ m or 10 μm.

Invariably there is a dispersion in the diameter of drops; that is, there is a spectrum of cloud droplet sizes. This influences mass transport processes, which are faster for smaller droplets, affecting uptake and reaction of gases in clouds. Typically cloud droplet distributions are rather sharply peaked. This is a consequence of the fact that mass transport of condensing water is faster for smaller droplets thereby allowing the smaller droplets, to "catch up" with the larger ones early in the cloud formation process.

Clouds persist in the atmosphere for a few tens of minutes (short-lived cumulus) to a few tens of hours (persistent stratus). Most clouds evaporate, rather than precipitate, thereby returning dissolved nonvolatile material to the clear air as aerosol particles.

3 SOURCES OF CLOUDWATER COMPOSITION

Cloudwater composition is very much a function of location, being dominated by availability of soluble ionic species. Principal ionic species present in cloudwater include sodium and chloride, from seawater, sulfate and nitrate anions, and ammonium and hydrogen ion as cations. In regions influenced by industrial emissions of sulfur and nitrogen oxides, cloudwater concentrations of H^+ are commonly 10^{-4} mol/L (molar, M) and not uncommonly 10^{-3} M or higher (Daum et al., 1984).

The fact that cloud droplets form on existing aerosol particles has immediate implications for cloudwater composition. Consider an ammonium sulfate aerosol particle of dry diameter 0.1 μm that serves as a nucleus of a cloud droplet of 10 μm diameter. The volume of the particle is $\sim 10^{-21}$ m^3. For density ~ 1000 kg/m^{-3} and molecular weight 100 g/mol (~ 0.1 kg/mol), the amount of ammonium sulfate contained in the particle is 10^{-17} mol. For this material dissolved in a 10-μm droplet ($\sim 10^{-15}$ m^3) the solution concentration is $\sim 10^{-2}$ mol/m^{-3} or $\sim 10^{-5}$ M. This concentration is at the low end of the range of concentrations of sulfate in cloudwater (and also in precipitation) in regions influenced by industrial emissions (see Chapter 15). It should be stressed that this figure varies as the third power of the particle diameter, that is, an order of magnitude for a factor of 2 in particle diameter. Thus for the particle diameter 0.2 μm, the concentration is 10^{-4} M.

Consider the correspondence between aqueous-phase concentration and the equivalent mixing ratio of the material in air. For one thousand 0.1-μm-diameter, unit-density particles per cm^3, the corresponding mass loading is 1 $\mu g/m^{-3}$, a loading that is rather low in the context of industrialized regions (See Chapter 16), albeit still substantially greater than that characteristic of regions remote from industrial sources. For molecular weight 100, this corresponds to a molar mixing ratio relative to air, $x \approx 0.3\,\text{nmol/mol(air)}$ (ppb). For a substance S that dissolves entirely in cloudwater, the relation between the mixing ratio of the substance x_S in air and concentration in cloudwater is

$$[S] = x_S p_{\text{atm}}/LR_g T$$

where [S] is aqueous concentration, p_{atm} is the atmospheric pressure, R_g is the universal gas constant, and T is the absolute temperature. In SI units p_{atm} is in units of pascal and $R_g = 8.3$ J/mol/K. The resulting concentration [S] is in units moles per cubic meters. In practical units (concentration in mol/L and pressure in bar; 1 bar = 10^5 Pa)

$$[S](\text{mol/L}) = 10^2 x_S p_{\text{atm}}(\text{bar})/LR_g T$$

In general the fractional uptake of soluble (ionic) aerosol species into cloudwater is fairly high, approaching unity at low aerosol loading and/or high updraft velocities leading to fairly high maximum supersaturation governing activation of aerosol particles (Leaitch et al., 1996). However, in the case of gases the uptake varies substantially depending on the solubility and/or reactivity of the gas in question.

4 UPTAKE OF GASES INTO CLOUDWATER

In general, a gaseous substance does not dissolve entirely in cloudwater in view of the rather limited solubility of most atmospheric gases in water—if the gas were highly soluble in cloudwater, it would be rapidly rained out and no longer in the atmosphere. The equilibrium concentration of a gaseous substance, S, physically dissolved in a liquid, is given by Henry's law (See Chapter 19).

$$[\mathrm{S(aq)}] = H_\mathrm{S} p_\mathrm{S} = H_\mathrm{S} x_\mathrm{S} p_\mathrm{atm}$$

where H_S is the Henry's law solubility coefficient of the gas. (In practical units, p_S in bar and [S(aq)] in mol/L, i.e., M, H_S has units M/bar.) Abundance of a gas-phase species is expressed in terms of the molar mixing ratio in air x, which is applicable equivalently to substances in gas, aerosol, or solution phases (Schwartz and Warneck, 1995). Characterization of the Henry's law solubility is the first step to understanding the uptake and reaction of a gas in cloudwater. Henry's law solubility coefficients of many gases of atmospheric importance are given in Figure 1.

The ratio of the amount of material in solution to gas phase (distribution ratio), under assumption of Henry's law equilibrium, is given by

$$D_{\mathrm{aq}/g} \equiv \frac{\text{moles in aqueous phase}}{\text{moles in gas phase}} = 10^{-2} LH(\mathrm{M/bar}) R_g T$$

If this is written as $D_{\mathrm{aq}/g} = H/H_{1/2}$ where $H_{1/2} = 10^{-2}\, LR_g T$, then for any specified value of L, the value of Henry's law solubility coefficient for which the gas is equally distributed between the gas phase and cloudwater is given by $H_{1/2}$. Consider a cloud of rather high liquid volume fraction $L = 10^{-6}$ (i.e., ~1 g/m^{-3} liquid water content); the corresponding value of $H_{1/2}$ is $\sim 4 \times 10^4$ M/bar; $H_{1/2}$ would be correspondingly higher for lower values of L. Comparison with the values of Henry's law solubility coefficients given in Figure 1 shows that virtually all such coefficients are orders of magnitude less than this value, supporting the assertion that reaction of the dissolved gas is required for substantial uptake into cloudwater.

In the case of gases that undergo rapid reversible reaction with water, for example, hydration or acid dissociation, it is necessary to consider the overall solubility equilibrium, not just the Henry's law equilibrium. Consider the solubility equilibrium for the dissolution of an acidic gas, for example, formic acid, HCOOH. The overall equilibrium for this dissolution may be thought to consist of the following steps:

$$\mathrm{HCOOH(g) = HCOOH(aq)}$$
$$\mathrm{HCOOH(aq) = H^+(aq) + COOH^-(aq)}$$

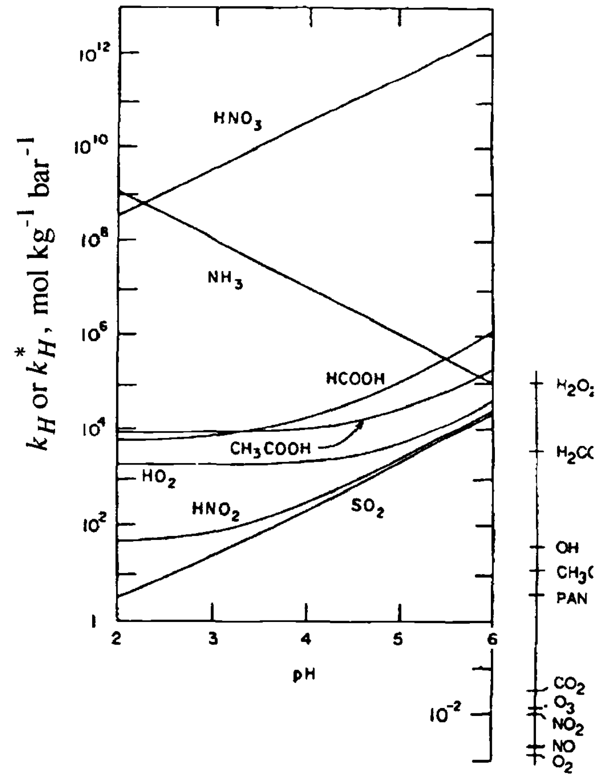


Figure 1 pH dependence of the effective Henry's law coefficient for gases that undergo rapid acid–base dissociation reactions in dilute aqueous solution, as a function of solution pH. Buffer capacity of solution is assumed to greatly exceed incremental concentration from uptake of indicated gas. Also indicated at the right of the figure are Henry's law coefficients for nondissociative gases. $T \sim 300$ K. Modified from Schwartz (1986a).

These reactions sum to give the overall reaction

$$\mathrm{HCOOH(g)} = \mathrm{H^+(aq)} + \mathrm{COOH^-(aq)}$$

The corresponding equilibrium expressions are

$$H_{\mathrm{HCOOH}} = \frac{[\mathrm{HCOOH(aq)}]}{x_{\mathrm{HCOOH}}p_{\mathrm{atm}}} \qquad K_a = \frac{[\mathrm{H^+}][\mathrm{COOH^-}]}{[\mathrm{HCOOH(aq)}]} \qquad K_{\mathrm{eq}} = \frac{[\mathrm{H+}][\mathrm{COOH^-}]}{x_{\mathrm{HCOOH}}p_{\mathrm{atm}}}$$

where K_a is the acid dissociation constant of aqueous formic acid. Depending on the situation, it may be more useful to deal with the overall solubility or with the individual equilibria.

The total concentration of the dissolved gas can be written (here staying with the example of formic acid) as

$$[\text{Formic acid}] \equiv [\text{HCOOH}] + [\text{COOH}^-] = H_{\text{HCOOH}} x_{\text{HCOOH}} p_{\text{atm}} \left(1 + \frac{K_{\text{eq}}}{[\text{H}^+]}\right)$$

It is often a good assumption that the cloudwater is well buffered against change in acid concentration $[\text{H}^+]$ resulting from the incremental uptake of gases present at low partial pressures characteristic of the ambient atmosphere. Under this assumption, $[\text{H}^+]$ is a constant and hence the aqueous concentration is linear in gas-phase partial pressure with an effective Henry's law solubility coefficient defined as:

$$H^*_{\text{HCOOH}} \equiv H_{\text{HCOOH}} \left(1 + \frac{K_{\text{eq}}}{[H^+]}\right)$$

so that one obtains a Henry's law-like expression for the overall solubility,

$$[\text{Formic acid}] = H^*_{\text{HCOOH}} x_{\text{HCOOH}} p_{\text{atm}}$$

In the case of SO_2 there are two acid dissociation equilibria. The effective Henry's law solubility coefficient for S(IV) (the Roman numeral IV denotes the oxidation state) is

$$H^*_{\text{S(IV)}} \equiv H_{\text{SO}_2} \left(1 + \frac{K_{\text{a1}}}{[H^+]} + \frac{K_{\text{a1}} K_{\text{a1}}}{[\text{H}^+]^2}\right)$$

where K_{a1} and K_{a2} denote the first and second dissociation constants, respectively.

Values of effective Henry's law solubility coefficients are shown in Figure 1 as a function of solution pH for the range of pH values typical of cloudwater. The effective solubility coefficient can greatly exceed the Henry's law coefficient for physical dissolution, especially for strong acids, such as nitric acid, and also for ammonia, which is highly soluble in the form of ammonium ion $NH_4{}^+$. These effective Henry's law solubility coefficients can also substantially exceed $H_{1/2}$, indicating that at equilibrium, such highly soluble gases as HNO_3 are essentially entirely taken up by cloudwater.

Because the chemical kinetics of acid dissociation reactions are generally quite rapid, the uptake of acidic gases such as HNO_3 is itself quite rapid, under control of mass transport processes rather than chemical kinetics. The mass transport processes governing this uptake are essentially identical to those governing the transfer of water vapor itself to and from cloud droplets, and the solubility of a gas such as HNO_3 is such that the uptake of soluble gases occurs on the time scale of cloud droplet activation and growth, that is taking place on a time scale of a few seconds to a few tens of seconds. This can result in such soluble gases being preferentially concentrated in the initially formed drops rather than being distributed uniformly throughout the cloud droplet spectrum (Wurzler et al., 1995); this can influence subsequent uptake and reaction of less soluble gases such as SO_2. A gas such as HNO_3 that dissolves in a growing cloud droplet contributes soluble material to the droplet, thereby adding to the Raoult effect of the solute already serving as the cloud condensation nucleus and increasing its cloud nucleating potential. This can have a further influence on cloud droplet composition and can also lead to situations of free cloud droplet growth at relative humidity slightly below 100% (Kulmala et al., 1997).

5 REACTIVE UPTAKE OF GASES BY CLOUDWATER

Without further reaction the fractional uptake of SO_2 into cloudwater is low, even at fairly high pH. The same is true *a fortiori* for NO_2, which does not undergo acid dissociation reaction in aqueous solution. However, there is a strong thermodynamic driving force in clouds for the reactive uptake of these gases to form sulfuric and nitric acids, respectively, the principal species contributing to acid deposition. This situation has stimulated substantial research interest in the processes whereby these gases are transformed into the acids and incorporated into cloudwater. The present understanding of these reactive uptake processes is that in the case of SO_2, the process consisting of uptake of SO_2 followed by aqueous-phase oxidation contributes substantially to the uptake of sulfuric acid by cloudwater and to the deposition of this material in precipitation. In contrast, the uptake process for NO_2 to form nitric acid appears to be dominated by gas-phase oxidation followed by uptake of the oxidized species. This section presents the formalism by which the rate of aqueous-phase reaction in cloudwater may be evaluated treating these two gases as examples.

Consider the rate of aqueous-phase reaction of dissolved sulfur-IV to be given by

$$\frac{-d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(1)}[\mathrm{S(IV)}]$$

where $k^{(1)}$ denotes an effective first-order rate coefficient, which in general may be equal to a second-order rate coefficient times the concentration of a second reagent. The reaction need not be first-order in the reagent sulfur-IV; additional power(s) of [S(IV)] could be incorporated within the effective first-order rate coefficient $k^{(1)}$. It is useful to refer the reaction rate to the total S(IV) concentration because of equilibration of individual sulfur-IV species within solution, $SO_2(aq)$, HSO_3^- or bisulfite, and $SO_3{}^{2-}$ or sulfite, that is rapid relative to depletion by reaction. The aqueous-phase reaction rate can be related to the gas-phase mixing ratio of SO_2 by solubility equilibria between aqueous-phase concentration of S(IV) and gas-phase partial pressure of SO_2, under assumption that these equilibria apply. This phase equilibrium is expected to hold if mass transport rates coupling the two phases are sufficiently fast to replenish the aqueous-phase material that is depleted by reaction, a situation that is normally expected to obtain, as discussed below. Hence

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(1)} H^{*}_{\mathrm{S(IV)}} x_{\mathrm{SO_2}} p_{\mathrm{atm}}$$

Under assumption that the aqueous-phase rate is uniform within a given region of a cloud, then the rate of reaction, expressed as a rate of decrease in the mixing ratio of SO_2, is

$$\frac{dx_{\mathrm{SO_2}}}{dt} = -\{10^{-2} L R_g T k^{(1)} H^{*}_{\mathrm{S(IV)}}\} x_{\mathrm{SO_2}} = -k^{(1)}_{\mathrm{eff}} x_{\mathrm{SO_2}}$$

The quantity in braces is an effective first-order rate coefficient of aqueous-phase reaction, referred to the gas-phase mixing ratio; this quantity may be directly employed in evaluating rates of reactions or in comparison to rate coefficients for loss by gas-phase reactions. Note that $k^{(1)}_{\mathrm{eff}}$ scales linearly with liquid water content and with Henry's law solubility coefficient. The more water present to serve as volume of reactor, the faster the reaction. Likewise, the more soluble the reagent gas, the faster the reaction. Evaluation of the rate of a specific reaction requires knowledge of the effective first-order rate coefficient of aqueous-phase reaction, $k^{(1)}$. For this, one must identify the mechanism and rate of aqueous-phase reaction.

There is a strong thermodynamic driving force for oxidation of dissolved SO_2 by molecular oxygen, which, because of its abundance, might be thought to be the key oxidant of SO_2 in cloudwater. However, this reaction is quite slow unless catalyzed, for example, by transition metal ions. Although catalyzed oxidation of dissolved sulfur-IV by dissolved molecular oxygen may be of some importance in some circumstances, the species that have been identified as of principal importance in oxidation of sulfur-IV in cloudwater are the strong oxidants ozone (O_3) and hydrogen peroxide (H_2O_2). Ozone is commonly present in the atmosphere at a mixing ratio of 30 to 50 nmol/mol. Hydrogen peroxide is present at much lower abundance, ~1 nmol/mol. These mixing ratios compare with those for SO_2 of order 10 nmol/mol in regions influenced by industrial emissions, to much lower at locations well removed from sources.

Consider first the ozone reaction. The rate of aqueous-phase reaction is given as

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(2)}[\mathrm{S(IV)}][\mathrm{O_3}], \qquad \text{i.e., } k^{(1)} = k^{(2)}[\mathrm{O_3}]$$

where $k^{(2)}$ is a second-order rate constant that must be determined by laboratory measurement and has been found to exhibit a strong pH dependence, increasing with increasing pH (Fig. 2a). The concentration of dissolved ozone is related to the gas-

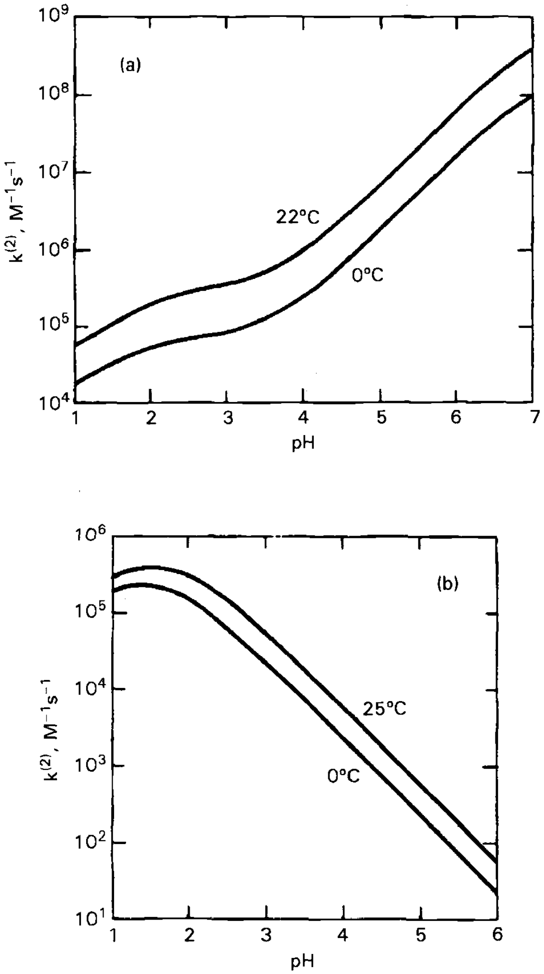


Figure 2 Effective second-order rate coefficients for aqueous-phase reaction of S(IV) with O_3 (a) and with H_2O_2 (b) as a function of pH. Modified from Schwartz (1988).

phase mixing ratio of this species again under assumption of solubility equilibrium, as

$$[O_3] = H_{O_3} x_{O_3} p_{\text{atm}}$$

so that

$$k^{(1)}_{\text{eff}} = 10^{-2} L R_g T k^{(2)} H_{O_3} x_{O_3} H^{*}_{S(IV))} p_{\text{atm}}$$

Combining the kinetic and solubility data permits the rate of reaction to be evaluated for known or assumed conditions of cloud liquid water content and partial pressures of reagent gases (Fig. 3). Here the left-hand ordinate gives the rate of aqueous-phase reaction. The right-hand ordinate gives the effective first-order rate coefficient of aqueous-phase reaction, referred to the gas-phase mixing ratio; $k^{(1)}_{\text{eff}}$ for indicated conditions, expressed in units of percent per hour. Note the strong pH dependence, rate increasing with pH, resulting from the pH dependences of sulfur-IV solubility (Fig. 1) and kinetic rate constant (Fig. 2). The ozone reaction is quite rapid at high pH. However, because of production of sulfuric acid as the reaction

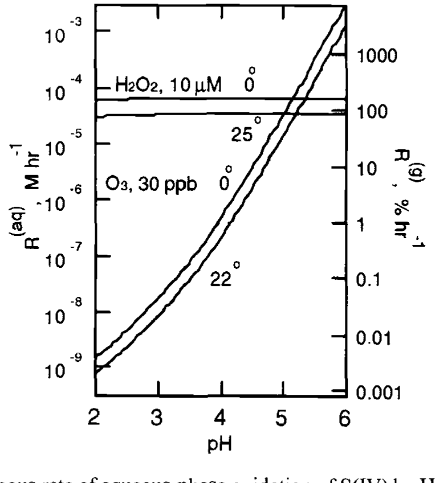


Figure 3 Instantaneous rate of aqueous-phase oxidation of S(IV) by H_2O_2 and O_3, evaluated as a function of pH for representative nonurban reagent concentrations. The rates scale approximately linearly with reagent concentrations. The right-hand ordinate gives the oxidation rate of SO_2 referred to the gas-phase partial pressure and expressed as percent per hour for a liquid water content $L = 1 \times 10^{-6}$ (1 cm^3/m^{-3}), the rate scales approximately linearly with L. For the H_2O_2 reaction the indicated aqueous-phase concentration of H_2O_2 corresponds to total mixing ratio of this species (gas plus aqueous phase; the two are comparable) of ~0.6 × 10^{-9}. See ftp site for color image.

proceeds, the pH rapidly becomes lower, decreasing the rate. Although a strong acid concentration of 10 μM (pH 5) is quickly reached, in perhaps 10 min, a much greater time, ~10 h, is required to reach an acid concentration of 50 μM. For this reason the ozone reaction is unlikely to account for cloudwater acidities of 10^{-4} to 10^{-3} M commonly observed in regions influenced by industrial emissions of SO_2.

Now consider the hydrogen peroxide reaction, whose aqueous-phase rate is given by:

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(2)}[\mathrm{S(IV)}][\mathrm{H_2O_2}]$$

This reaction is acid catalyzed; that is, the second-order aqueous-phase rate constant increases with decreasing pH (Fig. 2*b*). The pH dependences of solubility and reaction kinetics now cancel to yield a reaction rate that is roughly independent of pH throughout the pH range pertinent to cloudwater acidification (Fig. 3). For the conditions given in Figure 3 the rates of the O_3 and H_2O_2 reactions are equal roughly at pH 5. An effective first-order reaction rate of 100%/h corresponds to a $1/e$ lifetime of SO_2 of 1 h; the actual lifetime would depend on actual conditions. The H_2O_2 reaction is the only identified atmospheric reaction capable of maintaining the SO_2 oxidation rate sufficiently rapid to produce observed cloudwater H^+ and ${SO_4}^{2-}$ concentrations on time scales pertinent to cloud acidification.

In the case of the ozone reaction, ambient mixing ratios of O_3 are generally sufficiently in excess of those of SO_2 that depletion of O_3 need not be considered. However, ambient concentrations of H_2O_2, which are typically below 3 nmol/mol are often much less than ambient SO_2 mixing ratios. This leads to a situation where the reaction proceeds rapidly to completion by exhausting the H_2O_2 reagent. If on the other hand SO_2 mixing ratios are the lesser, then the reaction can rapidly and completely exhaust ambient SO_2. The time scale of this process, a few tens of minutes for representative mixing ratios in the nmol/mol region, leads to a situation where the extent of reaction is controlled by the limiting reagent. Such appears to be the case as indicated by field measurements simultaneously examining H_2O_2 and SO_2 mixing ratios in clouds. A survey of nonprecipitating stratiform clouds indicated that although either species is frequently present at nmol/mol mixing ratios, appreciable mixing ratios of the two species are virtually never simultaneously present, (Daum, 1990). The contribution of this reaction to cloudwater acidification has been directly confirmed by field measurements under well-defined flow conditions, including experiments with artificially introduced SO_2 and inert tracers, showing concomitant decreases in SO_2, and H_2O_2 and increases in H^+ and ${SO_4}^{2-}$ consistent with this reaction. This reaction is now thought to be the major contributor to atmospheric oxidation of SO_2, contributing both to acid precipitation and, in the likely event of cloud evaporation to sulfate aerosol, a principal component of atmospheric aerosols.

Based on laboratory and industrial experience nitrogen dioxide (NO_2) is known to be highly reactive with liquid water forming nitric and/or nitrous acids, the initial reaction being

$$2NO_2(N_2O_4) + H_2O \rightarrow 2H^+ + NO_3^- + NO_2^-$$

for which there is a strong thermochemical driving force; the N_2O_4 in parentheses indicates the possible participation of the NO_2 dimer, dinitrogen tetroxide. (This reaction is the basis for industrial manufacture of nitric acid.) It was therefore assumed by many atmospheric chemists that NO_2 would be rapidly taken up in cloudwater in the ambient atmosphere. Consideration of the mechanism of this reaction gives the aqueous-phase rate expression,

$$-\frac{d[NO_2]}{dt} = 2k^{(2)}H^2_{NO_2}x^2_{NO_2}p^2_{atm}$$

Determination of the Henry's law coefficient for NO_2 and the second-order reaction rate constant permitted evaluation of this rate for atmospheric conditions. Such evaluations have indicated that this rate is much too slow to contribute appreciably to NO_2 uptake by cloudwater at ambient concentrations. The reason for this, and for the great difference with experience at high NO_2 concentrations, is that the reaction is second-order in the concentration of a very weakly soluble gas. Comparisons of NO_2 mixing ratios in clouds with those in clear air in the vicinity of clouds indicates that the fractional uptake of NO_2 into cloudwater is quite small, lending confirmation to the above picture. Alternative possible mechanisms for NO_2 uptake include reaction with reducing species dissolved in cloudwater, as NO_2 is a fairly strong oxidant.

An alternative mechanism that may be important, especially at night, is initiated by the gas-phase reactions

$$NO_2 + O_3 \rightarrow NO_3 + O_2$$

followed by

$$NO_3 + NO_2 \rightarrow N_2O_5$$

with N_2O_5 being taken up by cloudwater by reaction to form nitric acid:

$$N_2O_5 + H_2O(l) \rightarrow 2H^+ + 2NO_3^-$$

The rate of reaction is controlled by the rate of the initiation reaction of NO_2 with O_3, which is several percent per hour at typical ozone mixing ratios of a few tens of nmol/mol. The reason that this appears to be important at night but not during the day is that photolysis of NO_3 by visible radiation

$$NO_3 + h\nu \rightarrow NO + O_2$$

is the major sink of NO_3 during the day, thereby cutting off the overall reaction.

In addition to these acidification reactions several other in-cloud reactions have been identified as of importance or potential importance in atmospheric chemistry. The hydroperoxy radical, HO_2, which plays an important role in gas-phase photochemistry as part of the chain of reactions leading to ozone formation by oxidation of NO and hydrocarbons, is thought to be rather soluble in water because of its weak-acid dissociation:

$$HO_2 \rightarrow H^+ + O_2^-$$

The dissolved material undergoes rapid self-reaction to form hydrogen peroxide. It has been suggested that the occurrence of this process can substantially influence the ozone budget in the remote troposphere. However, the process remains somewhat speculative in view of the lack of firm information on the solubility of the HO_2 radical.

Several studies have demonstrated substantial aqueous-phase formation of H_2O_2 by photochemical reactions in collected cloudwater. The exact processes are not yet elucidated but evidently involve trace organic species, which are difficult to characterize. Such reactions may contribute substantially to SO_2 oxidation in situations where this oxidation is limited by the amount of H_2O_2 initially present. More generally, it may be noted that photochemical reactions, in both gas and solution phases, may be enhanced in the tops of clouds because of enhanced photolysis fluxes, by a factor of 5 or more, that result from multiple scattering of solar radiation within clouds.

6 COUPLED MASS TRANSPORT AND CHEMICAL REACTION

As indicated above, quantitative evaluation of the rates of aqueous-phase reactions in clouds are predicated on the assumption that the rate of mass transport processes coupling the gas-phase reservoir of reagent gas to the solution phase within individual cloud droplets is sufficiently fast to maintain the Henry's law equilibrium in competition with the sink of dissolved material by aqueous-phase reaction. The pertinent mass transfer processes are gas-phase diffusion, from the bulk of the gas phase to the gas–liquid interface; transfer across the interface, as governed by the gas-kinetic collision rate and the mass accommodation coefficient, the fraction of collisions resulting in transfer of material across the interface, a property characteristic of individual gases and solutions; and aqueous-phase diffusion of the dissolved gas occurring concomitantly with aqueous-phase reaction. In general, if the reaction is sufficiently slow, mass transport is sufficiently rapid to maintain the solubility equilibria, but departure from equilibrium occurs for sufficiently rapid reaction rates. Criteria for the onset of this "mass transport limitation" of the rate of aqueous-phase reactions in clouds have been developed in terms of drop radius, Henry's law coefficient, effective first-order reaction rate coefficient, diffusion coefficients, and mass accommodation coefficients. For the most part, the rate of reaction of SO_2 in cloudwater appears only minimally limited by mass transport rates, the exception being the ozone reaction at high pH, under which condition both the solubility and effective first-order rate coefficient are quite large.

7 SUMMARY

Clouds present substantial concentrations of liquid-phase water, which can potentially serve as a medium for dissolution and reaction of atmospheric gases. The important precursors of acid deposition, SO_2, and nitrogen oxides NO and NO_2 are only sparingly soluble in clouds without further oxidation to sulfuric and nitric acids. In the case of SO_2, aqueous-phase reaction with hydrogen peroxide and to lesser extent ozone are identified as important processes leading to this oxidation, and methods have been described by which to evaluate the rates of these reactions. The limited solubility of the nitrogen oxides precludes significant aqueous-phase reaction of these species, but gas-phase reactions in clouds can be important especially at night.

REFERENCES

Daum, P. H., Observations of H_2O_2 and S(IV) in air, cloudwater and precipitation and their implications for the reactive scavenging of SO_2, *Atmos. Res.*, *25*, 89–102, 1990.

Daum, P. H., T. J. Kelly, S. E. Schwartz, and L. Newman, Measurements of the chemical composition of stratiform clouds, *Atmos. Environ.*, *18*, 2671–2684, 1984.

Kulmala, M., A. Laaksonen, R. J. Charlson, and P. Korhonen, Clouds without supersaturation, *Nature*, *388*, 336–337, 1997.

Leaitch. W. R., C. M. Banic, G. A. Isaac, M. D. Couture, P. S. K. Liu, I. Gultepe, S.-M. Li, K. I. Kleinman, P. H. Daum, and J. I. MacPherson, Physical and chemical observations in marine stratus during the 1993 North Atlantic Regional Experiment: Factors controlling cloud droplet number concentrations, *J. Geophys. Res.*, *101*, 29123–29135, 1996.

Schwartz, S. E, Chemical conversions in clouds, in S. D. Lee, T. Schneider, L. D. Grant, and P. J. Verkerk (Eds.), Lewis Publishers, Chelsea, MI, 1986a, pp. 349–375.

Schwartz, S. E., Mass-transport considerations pertinent to aqueous-phase reactions of gases in liquid-water clouds, in, W. Jaeschke (Ed.), *Chemistry of Multiphase Atmospheric Systems*, Springer, Heidelberg, 1986b, pp. 415–471.

Schwartz, S. E., Mass-transport limitation to the rate of in-cloud oxidation of SO_2: Reexamination in the light of new data, *Atmos. Environ.*, *22*, 2491–2499, 1988.

Schwartz, S. E., and P. Warneck, Units for use in atmospheric chemistry, *Pure Appl. Chem. 67*, 1377–1406, 1995.

Wurzler S., A. I. Flossmann, H. R. Pruppacher, and S. E. Schwartz, The scavenging of nitrate by clouds and precipitation. I. A theoretical study of the uptake and redistribution of $NaNO_3$ particles and HNO_3 gas by growing cloud drops using an entraining air parcel model, *J. Atmos. Chem.*, *20*, 259–280, 1995.

CHAPTER 18

DRY DEPOSITION

M. L. WESELY

1 INTRODUCTION

Dry deposition refers to the removal of trace substances from the atmosphere without the aid of precipitation. Dry deposition cleanses the air and delivers substances to the Earth's surface. In air, the vertical transfer is accomplished primarily by turbulent mixing; for particles whose diameters are greater than 1 to 2 μm, vertical transfer is aided by gravitational settling. As air comes into contact with surface elements, gas molecules can react with surface materials or dissolve in them. Particles can be captured by interception or impaction with the surface elements. In comparison to wet deposition, in which substances are carried in precipitation to the surface, dry deposition is a slower, more continuous process and is profoundly affected by the physical, chemical, and biological properties of the surface.

The downward vertical mass flux density (deposition rate per unit area), divided by concentration C at a specified height, is conventionally known as the deposition velocity V_d. The value of this velocity can vary greatly depending on the properties of the substance of interest and local atmospheric and surface conditions. Nevertheless, the concept of a deposition velocity is highly useful in many applications because it produces the deposition rate when multiplied by a measured or modeled concentration. As a general guideline, a deposition velocity of 0.1 cm/s is small, 1.0 cm/s is moderately large, and several centimeters per second is near the limit that is physically possible on the basis of turbulent mixing alone. The corresponding residence times of the constituents in the lower atmosphere can be weeks, days, or hours, respectively, when they are controlled only by dry deposition. Relatively inert gases with very small deposition velocities can have lifetimes of several years in the atmosphere if other sources of removal, such as chemical transformations, are weak.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

Because dry deposition is a surface process, it can be treated as a boundary condition in models that compute atmospheric chemical budgets. In numerical models, the deposition rate is usually estimated on the basis of deposition velocity fairly close to the surface, typically at a height less than 50 m, with the assumption that the vertical flux below such small heights does not change substantially with height. This assumption is valid if the lower atmosphere has concentrations and mixing properties that are horizontally uniform and steady with time over the period of a few hours. Another requirement for nearly constant fluxes with height is that the substance of interest does not undergo chemical and physical changes that are rapid in comparison to the time scales of local vertical mixing. When these requirements are met, the vertical flux can be estimated as

$$F = -V_d C \tag{1}$$

where we have adopted the convention that a flux directed downward is negative. This type of formulation is intended only for the case of a flux being directed downward, when V_d is positive; it has little merit for substances that are emitted from the surface because ambient concentration often has little effect on emission rates.

This chapter provides a brief, somewhat introductory, overview of dry deposition. The reader can find considerable additional information in the scientific literature. For example, a review of the state of the science after considerable research on acidic deposition during the 1980s in the United States was provided by Hicks et al. (1989), and measurement techniques were reviewed by Businger (1986). Some European perspectives on acidic deposition were provided by Erisman and Draaijers (1995), and reviews of the status of dry deposition knowledge were conducted recently by Lovett (1994), Seinfeld and Pandis (1998), and Wesely and Hicks (2000).

2 FORMULATION OF DEPOSITION VELOCITY

A common simple method of evaluating Eq. (1) is by analogy to Ohm's Law, where F corresponds to current, V_d corresponds to the inverse of the total resistance, and C corresponds to the voltage, referenced to electrical ground corresponding to a concentration of zero that is assumed to occur somewhere in the surface. From this analogy, the deposition velocity can be expressed in terms of three resistances in series.

$$V_d = (R_a + R_b + R_c)^{-1} \tag{2}$$

Here, R_a represents the aerodynamic resistance to transfer associated with turbulent mixing above the surface, R_b is the resistance of the quasilaminar sublayer of air in contact with surface elements, and R_c is the bulk resistance of the surface. The values of R_a and R_b can be estimated with readily available micrometeorological formulations. Various schemes exist in the literature for depicting and evaluating these resistances (e.g., Hicks et al., 1987; Wesely, 1989), and Figure 1 shows a relatively complex version that includes some of the many resistances in series and parallel that can be constructed for R_c.

The existence and the relative importance of each flux pathway shown in Figure 1 tend to be unique for each type of surface and each substance. Each resistance term itself normally must be parameterized in terms of surface properties. In field experiments, R_c is often found as the residual, unmeasured quantity in Eqs. (1) and (2). Some components of R_c are not measured directly but are typically inferred over a

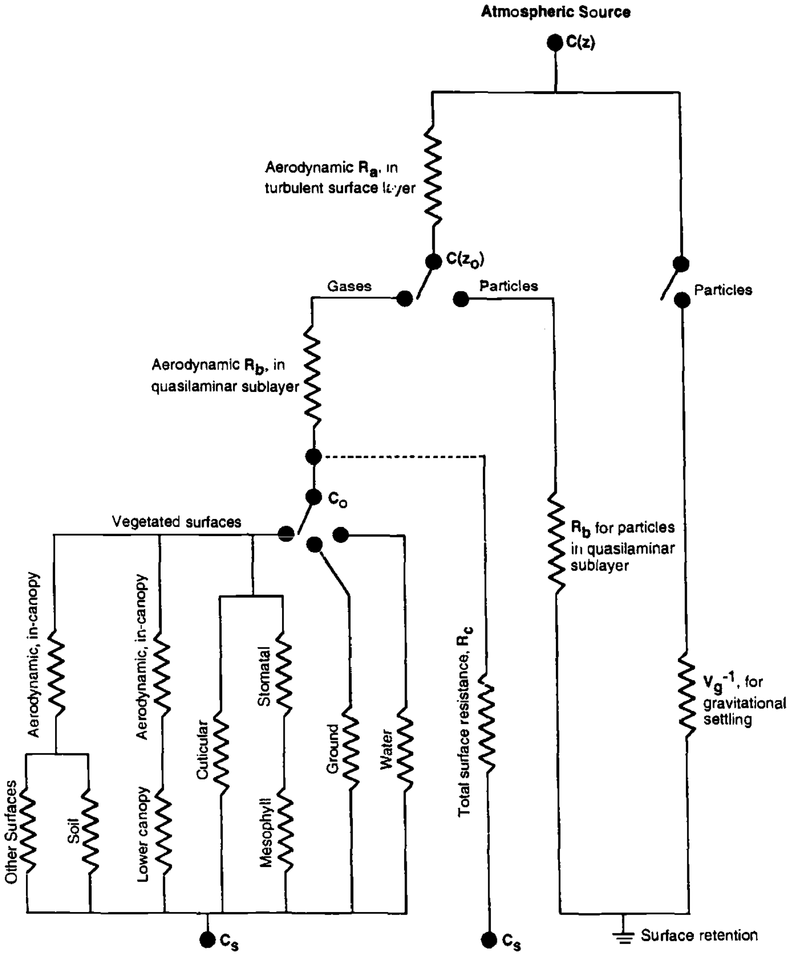


Figure 1 Resistance scheme for dry deposition.

particular surface as environmental conditions change. When leaf stomatal openings close at night, for example, the bulk resistance to deposition on the outer surfaces of leaves and the ground below vegetative canopies can be inferred. The resistance of the waxy cuticle is sometimes measured in the laboratory, and the resistance of the ground surface beneath the canopy is occasionally evaluated with flux measurements there. In the parameterizations that are generated, important variables include environmental factors (such as solar radiation, temperature, air humidity, wetness of the surface caused by dew and rainfall, and soil moisture content) and details of the surface (such as height of vegetative canopy, amount of leaf area, species of vegetation, and soil pH). For bodies of water, the structure and size of waves can be important.

For particle deposition, R_c is not commonly considered explicitly and the deposition velocity is expressed in terms of R_a, R_b, and gravitation settling velocity V_g. However, R_b embodies several somewhat complex processes involving transport through the quasilaminar sublayer, interception of particles by fine elements of the surface, and inertial impaction of particles on the surface. Theoretical formulations for both R_b and V_g usually include a strong dependency on particle size.

Although R_a, R_b, and aerodynamic resistances in canopies are considered separately, they are all strongly affected by turbulence parameters. The turbulent mixing induced by buoyancy forces associated with surface heating by solar radiation can directly alter R_a and in-canopy resistances. The roughness of the surface, a primary factor in evaluating R_a, is linked indirectly to the vegetative properties that affect the resistances of elements of a vegetative canopy. Somewhat more confusing is the fact that in-canopy resistances are implied in Figure 1 to be controlled purely by turbulence, but the distribution of "sinks" in canopies can alter the value of the in-canopy aerodynamic resistance because the latter is actually a composite of vertically distributed air-phase resistances to many surface elements. For this and other reasons noted below, the approach that uses Eq. (2) and Figure 1 is considered by many researchers to be oversimplified.

Several other difficulties exist with the resistance analogy for dry deposition. Some experiments have shown, for example, that weak mixing at night in tall canopies might lead to storage of chemicals in air in the sheltered areas. Deep snowpack might store some relatively insoluble substances. Gusty winds can resuspend particles. Relatively inert gases might temporarily dissolve in surface materials and later be reemitted. Some trace gases are emitted from natural surfaces, sometimes at greater rates when the ambient concentrations are small. To overcome such difficulties, experimental observations of deposition velocities sometimes provide the primary information used to evaluate Eq. (1), or more sophisticated models of air–surface exchange are developed.

3 DEPOSITION VELOCITY ESTIMATES

The great diversity of airborne trace chemical properties, surface conditions, and environmental conditions prevents the generation of universally applicable dry deposition parameterization schemes. Studies have tended to focus on substances important in atmospheric chemistry or likely to be harmful to human health, biota, and man-made materials. The types of surface most often chosen for investigation are ones that are fairly common in a given region, because the total amount of substances removed from the atmosphere depends on the amount of areal exposure to the surfaces. For example, the deposition of O_3 to agricultural areas and forests in the eastern half of the United States has been studied fairly frequently. Figure 2 shows the results of some studies conducted by Argonne National Laboratory in the 1990s. As can be seen, the deposition velocity to soybean fields tends to be larger than for maize, a pattern related to the structures and physiological properties of the plants. Cloudy conditions and the resulting reduction in solar radiation for the central day in Figure 2 caused leaf stomata to be partially closed and the value of R_c to be fairly large. At night, the stomata were closed for all three canopies, leaving open only the deposition pathway to the outer surfaces of the vegetation and the soil surface beneath. Deposition to the tallgrass prairie tended to be suppressed in general because of the effects of fairly dry soil and the tendency of the grass to increase stomatal resistance in such conditions to reduce transpiration of moisture.

Ozone is taken up through plant stomata because destruction of O_3 occurs fairly rapidly in the substomatal cavities. The oxidization by O_3 of some organic substances in solution in the film of water that envelops the mesophyll cells is the primary reason for the very small mesophyll resistance. Ozone also reacts strongly with many surface materials, although the waxy outer coating of leaves is an effective barrier. Because O_3 is poorly soluble in water, flux pathways are insignificant to water that is free from significant amounts of substances with which O_3 reacts. In

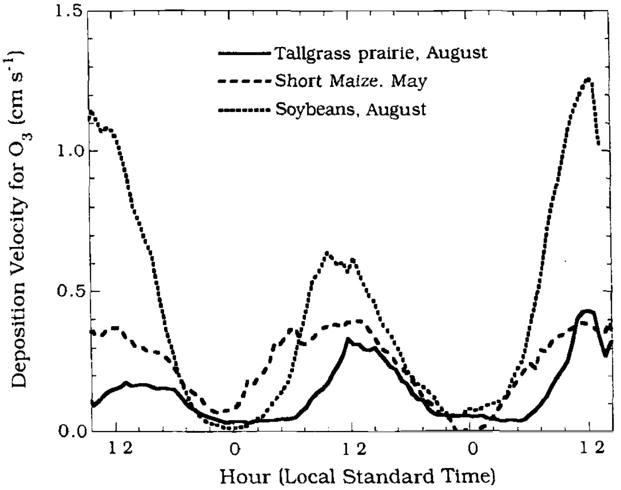


Figure 2 Observations of the dry deposition velocity for ozone over three types of surfaces.

general, measures of the oxidizing capacity of various substances provide a means of evaluating their ability to be destroyed at surfaces of various types.

Many studies have shown that SO_2 is also taken up by vegetation with practically no mesophyll resistance. The primary factor that affects SO_2 removal is its fairly large effective solubility in water. Usually the amount of exposure to water is a major factor in assessing the deposition velocity of substances that dissolve and dissociate rapidly in water. Table 1 shows deposition velocity values for SO_2, O_3, and NO_2 for various surfaces, based on experimental observations and resistance models. As can be seen, the deposition velocities for NO_2 tend to be smaller than for SO_2 and O_3, mainly because the water solubility of NO_2 is small and its ability to oxidize surface materials is weak compared to that of O_3. In general, measures of the effective solubility and oxidizing capacity of gases provide a means of estimating their deposition velocities relative to those seen experimentally for SO_2 and O_3.

TABLE 1 Typical Deposition Velocities (cm/s^1) for SO_2, O_3, and NO_2 at a Height of 10 m[a]

	Soybeans		Grassland, Maize		Deciduous Forest		Coniferous Forest	
Substance	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
Midsummer with Lush Vegetation								
SO_2	1.4	0.4	0.8	0.3	0.9	0.1	0.6	0.1
O_3	1.0	0.2	0.7	0.2	0.8	0.1	0.5	0.1
NO_2	0.8	0.1	0.4	0.05	0.7	0.03	0.4	0.03
Autumn with Unharvested Cropland								
SO_2	0.4	0.2	0.4	0.2	0.2	0.1	0.3	0.1
O_3	0.4	0.2	0.4	0.2	0.2	0.1	0.3	0.1
NO_2	0.1	0.1	0.1	0.05	0.05	0.03	0.2	0.03
Late Autumn after Frost, No Snow								
SO_2	0.5	0.2	0.2	0.1	0.1	0.1	0.1	0.1
O_3	0.5	0.2	0.4	0.2	0.2	0.1	0.2	0.1
NO_2	0.1	0.1	0.1	0.05	0.1	0.04	0.1	0.03
Winter, Snow on Ground and Near Freezing								
SO_2	0.5	0.2	0.5	0.2	0.1	0.1	0.3	0.1
O_3	0.1	0.03	0.1	0.03	0.2	0.03	0.1	0.04
NO_2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Transitional Spring with Partially Green Short Annuals								
SO_2	1.0	0.3	0.7	0.2	0.5	0.1	0.3	0.1
O_3	0.7	0.2	0.5	0.2	0.5	0.1	0.3	0.1
NO_2	0.4	0.1	0.2	0.05	0.3	0.04	0.2	0.03

[a] Cases 1 and 2 for each surface type correspond to solar irradiances of 500 and 0 W/m^2, respectively. Dry surfaces and moderate wind speeds are assumed.

The deposition of nonpolar, nonreactive gases such as some organic compounds is usually assumed to be small, although solubility in lipids in vegetation might slightly enhance deposition. Studies have shown that this pathway is measurable but very small. Deposition velocities of less than 0.1 cm/s are likely.

Particle deposition velocities can be strongly dependent on particle size. For particles smaller than 0.1 to 0.2 μm in diameter, deposition by transport through the quasilaminar sublayer can be fairly strong; the extremely fine particles diffuse through air similarly to molecules of gas. Particles larger than 1 to 2 μm are deposited mainly by gravitational settling, for which the associated deposition velocities can be several centimeters per second. For the so-called accumulation size mode, in which particle diameters are larger than 0.1 to 0.2 μm and smaller than 1 to 2 μm, mechanisms of deposition are often thought to be ineffective. Some field studies have shown, however, that processes of interception and impaction in gusty wind conditions can enhance deposition velocities substantially, to values exceeding 0.5 cm/s during daytime conditions over typical terrestrial surfaces. Such deposition velocities have been seen over grass for sulfate, which usually exists primarily in the accumulation size mode; values exceeding 1.0 cm/s for sulfate and nitrate have been seen over partially wetted coniferous forests.

4 MODELS OF DEPOSITION VELOCITY

Models have become significantly more sophisticated during the past two decades and are becoming more effective tools for the environmental worker who must make estimates of deposition rates of trace chemicals. "Big-leaf models" that use Eq. (2) with little breakdown of R_c into component resistances have been supplanted to some extent by multilayer canopy models for vegetated surfaces at specific sites where local conditions are observed directly (e.g., Meyers and Baldocchi, 1988; Meyers et al., 1998). Variations of big-leaf models in which R_c is represented by several possible flux pathways have been used extensively in dry deposition modules intended for regional- and large-scale numerical models of atmospheric chemistry (e.g., Pleim et al., 1984; Wesely, 1989; Padro and Edwards, 1991; Benkovitz et al., 1994; Ganzeveld and Lelieveld, 1995).

The potential is high for advancing the accuracy of dry deposition estimates by using advanced atmospheric models with notably improved descriptions of the surface conditions that affect dry deposition. Third-generation models are expected to have capabilities that will reduce the dependency on empirically derived resistance values and provide a means of coupling deposition and emission more closely (Peters et al., 1995). Third-generation models are also likely to incorporate better simulations of the structure of the planetary boundary layer, to provide estimates of soil moisture content and evapotranspiration that can be valuable inputs to dry deposition modules, and to allow the use of parameterizations of vegetative processes that are based on physiological processes, such as processes that control photosynthesis and uptake of carbon dioxide.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under contract W-31-109-Eng-38, as part of the Atmospheric Chemistry Program of the Office of Science, Office of Biological and Environmental Research, Environmental Sciences Division.

REFERENCES

Benkovitz, C. M., C. M. Berkowitz, R. C. Easter, S. Nemesure, R. Wagener, and S. E. Schwartz, Sulfate over the North Atlantic and adjacent continental regions: Evaluation for October and November 1986 using a three-dimensional model driven by observation-derived meteorology, *J. Geophys. Res.*, *99*, 20725–20756, 1994.

Businger, J. A., Evaluation of the accuracy with which dry deposition can be measured with current micrometeorological techniques, *J. Climate Appl. Meteorol.*, *25*, 1100–1124, 1986.

Erisman, J. W., and G. P. J. Draaijers, *Atmospheric Deposition in Relation to Acidification and Eutrophication*. Elsevier, New York, 1995.

Ganzeveld, L., and J. Lelieveld, Dry deposition parameterization in a chemistry general circulation model and its influence on the distribution of reactive trace gases, *J. Geophys. Res.*, *100*, 20999–21012, 1995.

Hicks, B. B., D. D. Baldocchi, T. P. Meyers, R. P. Hosker, Jr., and D. R. Matt, A preliminary multiple resistance routine for deriving dry deposition velocities from measured quantities, *Water Air Soil Pollut.*, *36*, 311–330, 1987.

Hicks, B. B., R. R. Draxler, D. L. Albritton, F. C. Fehsenfeld, J. M. Hales, T. P. Meyers, R. L. Vong, M. Dodge, S. E. Schwartz, R. L. Tanner, C. I. Davidson, S. E. Lindberg, and M. L. Wesely, *Atmospheric Processes Research and Process Model Development*, State of the Science/Technology, Report No. 2, National Acid Precipitation Assessment Program, Superintendent of Documents, Government Printing Office, Washington, D. C., 1989.

Lovett, G. M., Atmospheric deposition of nutrients and pollutants in North America: An ecological perspective, *Ecol. Appl.*, *4*, 629–650, 1994.

Meyers, T. P., and D. D. Baldocchi, A comparison of models for deriving dry deposition fluxes of O_3 and SO_2 to a forest canopy, *Tellus*, *40B*, 270–284, 1988.

Meyers, T. P., P. Finkelstein, J. Clarke, T. G. Ellestad, and P. F. Sims, A multilayer model for inferring dry deposition using standard meteorological measurements, *J. Geophys. Res.*, *103*, 22645–22661, 1998.

Padro, J., and G. C. Edwards, Sensitivity of ADOM dry deposition velocities to input parameters: A comparison with measurements for SO_2 and NO_2 over three land-use types, *Atmos.-Ocean.*, *29*, 667–685, 1991.

Peters, L. K., C. M. Berkowitz, G. R. Carmichael, R. C. Easter, G. Fairweather, S. J. Ghan, J. M. Hales, L. R. Leung, W. R. Pennell, F. A. Potra, R. D. Saylor, and T. T. Tsang, The current status and future direction of Eulerian models in simulating the tropospheric chemistry and transport of trace species: A review, *Atmos. Environ.*, *29*, 189–222, 1995.

Pleim, J. E., A. Venkatram, and R. Yamartino, *ADOM/TADAP Model Development Program*, Vol. 4: *The Dry Deposition Module*, Ontario Ministry of the Environment, Rexdale, Canada, 1984.

Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics* Wiley, New York, 1998.

Wesely, M. L., Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models, *Atmos. Environ.*, *23*, 1293–1304, 1989.

Wesely, M. L., and B. B. Hicks, A review of the current status of knowledge on dry deposition, *Atmos. Environ.*, *34*, 2261–2282, 2000.

CHAPTER 19

FATE OF ATMOSPHERIC TRACE GASES: WET DEPOSITION

CHRIS WALCEK

1 INTRODUCTION

Some trace gases and pollutants are readily removed from the atmosphere by becoming incorporated into cloudwater and then falling to Earth's surface in precipitation. In cloud-free air, small amounts of condensable species such as sulfates, ammonia, and nitrates coagulate with water vapor to form or nucleate extremely small atmospheric aerosol particles. Through gaseous diffusion and aerosol coagulation, smaller aerosols generally grow in size with time, while continuously maintaining an approximate equilibrium with water vapor and other surrounding condensable trace gases. Some aerosols eventually become large enough to develop an appreciable fall speed that overcomes frictional air drag, and thus slowly settle toward Earth's surface. However, well before particles grow to sizes where gravitational settling becomes important, larger aerosol particles are readily incorporated into clouds when aerosol-laden air cools during lifting, mixing with colder air, or other radiative cooling processes. Cloud drops initially form directly on aerosol particles, immediately incorporating a significant fraction of aerosol-borne trace constituents into the liquid phase. Once clouds form, soluble gases rapidly diffuse toward and dissolve into cloud droplets, contributing to trace chemical concentrations in cloudwater.

Most of the time, cloudwater evaporates, releasing the aerosols and gases that were absorbed during condensation back into the atmosphere. Under some conditions, cloud drops and ice particles grow by vapor diffusion and coalesce with other cloud drops and become large enough to develop appreciable fall speeds, at which point precipitation-sized cloud particles are formed. Within a cloud, falling liquid

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

and ice precipitation rapidly scavenges and accretes smaller cloud drops, and precipitation particles grow large enough to leave the cloud and fall to the surface. Outside the cloud, falling precipitation evaporates before reaching the surface, which releases some dissolved constituents back to the atmosphere, and concentrates the remaining dissolved constituents in precipitation before reaching the surface.

Thus there are many pathways through which trace constituents are transferred from the gas phase into precipitation at Earth's surface. Within a cloud there is nucleation scavenging, which takes place as ice and liquid water condenses on condensation and ice nuclei. Impaction scavenging of aerosols and small cloud particles by other cloud drops and other classes of hydrometeors (cloud ice, graupel, snow etc.) involves collisions between hydrometeors and interstitial cloud particles. Gases are also readily absorbed by all categories of hydrometeors through direct gaseous diffusion. Below cloud base, falling precipitation absorbs gases through diffusion, and aerosols are incorporated into falling precipitation by impaction scavenging.

Rudimentary calculations of the physics of these scavenging processes, as well as more sophisticated simulations reveal that the largest fraction of trace constituents in precipitation usually originate from the direct nucleation scavenging of soluble aerosols by cloud drops when they initially condense during cloud formation. Probably the next most important source of trace constituents in precipitation arises from the dissolution of soluble trace gases into cloudwater. Other scavenging mechanisms, such as impaction scavenging of aerosols by cloud or precipitation drops, and diffusion or impaction of gases or aerosols by larger precipitation particles are typically much less efficient. In the following sections, these important scavenging pathways are further discussed and quantified.

2 NUCLEATION SCAVENGING

It is well established that in the atmosphere the phase transition from water vapor to liquid water depends on the presence of cloud condensation nuclei (CCN). CCN are composed of water-soluble substances that bind with water molecules and significantly lower the equilibrium partial pressure of water vapor, allowing water to condense or change phase into these "salty" solutions when water vapor concentrations are well below saturation with respect to pure water.

Raoult's Law. The equilibrium partial pressure of water vapor over a solution containing a dissolved salt is very close to the mole fraction of water molecules in the solution times the saturated partial pressure of water vapor over pure water. This reduction of the equilibrium vapor pressure of water over a "salty" solution is known as "Raoult's law." Thus a droplet composed of 50% water molecules and 50% sodium chloride ions (25% Na^+ and 25% Cl^-) will be at equilibrium with an environment where the relative humidity is 50%. Therefore, the mole fraction of water molecules in "wetted" aerosols is very close to the ambient relative humidity, and therefore at humidities close to 100%, CCN and the wetted aerosols become nearly "pure" water solutions. Even in the cleanest environments, there are several tens to hundreds of CCN per cubic centimeter of air. Therefore, there are always small amounts of wetted surfaces present in the atmosphere, and the time it takes H_2O vapor to diffuse toward these wetted aerosols is sufficiently small so that condensation or evaporation occurs rapidly, maintaining the environment near a saturated equilibrium with respect to these solution droplets at all times.

Clouds form when relative humidities (RH) exceed 100%, and, since the equilibrium partial pressure of water vapor is only a function of temperature, cloud formation is usually induced by the cooling of air. If the total amount of water in an air parcel remains the same, as an air parcel cools, the relative humidity rises, and a small amount of water condenses onto the wetted aerosols, and they swell in size. When air is cooled at humidities below 100%, the wetted aerosol absorbs water vapor, increasing the liquid-phase mole fraction, thus changing the equilibrium pressure of water vapor over the wetted aerosols following Raoult's law. However, when the RH exceeds 100%, the aerosol solutions become essentially "pure water," and additional water added to the droplets does not increase the equilibrium partial pressure of water vapor. Therefore, all vapor in excess of saturation rapidly condenses. Models and measurements in clouds show that the RH in clouds rarely exceeds about 101%, and typical supersaturations in a cloud are on the order of a few tenths of a percent (RH = 100.1 to 100.5%).

Kelvin Effect. For extremely small spherical drops, there is insufficient surface tension to "hold" condensed water in a liquid phase, and thus the equilibrium partial pressure of water vapor over a spherical droplet is higher than the equilibrium partial pressure over a flat liquid surface. For example, a spherical liquid drop with a radius of 0.01 μm requires a relative humidity of about 110% to maintain its size without evaporating. A pure water drop with a radius of 1 μm requires RH = 100.1% to maintain its size without growth or evaporation.

Köhler Curves. The *increase* of the partial pressure over a spherical surface due to the Kelvin effect counters the *reduction* in vapor pressure due to Raoult's effect. Together, the Raoult and Kelvin effects produce the classical "Köhler curve" (Köhler, 1926) describing the equilibrium partial pressure (e_{CCN}) over a spherical droplet of radius r containing a specified amount of dissolved ions relative to pure water:

$$\frac{e_{\mathrm{CCN}}}{e_s} \approx 1 + \frac{a}{r} - \frac{b}{r^3} \tag{1}$$

where e_s is the equilibrium partial pressure of water vapor over a flat pure water surface, and the term involving a/r is a curvature (Kelvin) term, and the b/r^3 term is the solution (Raoult) term. Numerically, $a \approx 3.3 \times 10^{-5}/\mathrm{T}$ (K) (cm), and $b \approx 4.3\, im_s/M_s$ (cm^3) where i is approximately the number of dissolved molecules produced when the soluble CCN dissolves [e. g., sodium chloride (NaCl) produces 2 ions; sulfuric acid (H_2SO_4) produces 3], M_s is the molecular weight and m_s is the dry mass of the soluble component of the CCN $(=\frac{4}{3}\pi r_s^3\rho_s$, r_s = dry salt particle radius, ρ_s = salt density).

The equilibrium vapor pressure over a solution droplet may be larger or smaller than the vapor pressure over a plane surface of pure water, depending on whether the solute term is smaller or larger than the curvature term. Figure 1 shows examples of the saturation vapor pressure around drops that have condensed on two sizes and typical compositions of soluble aerosols. These curves show that there is a local maximum that occurs at a *critical* radius $[r_c = (3b/a)^{1/2}]$ and critical supersaturation $[s_c = 100(e/e_s - 1) = (4a^3b/27)^{1/2}]$. If a cloud drop is smaller than the critical size, and the local water vapor pressure is less than the critical supersaturation, drops grow only until they reach equilibrium with respect to the environmental water vapor pressure. If the water vapor concentration is greater than the critical supersaturation, a CCN will grow indefinitely, and a cloud drop is considered "activated" or "nucleated." Notice that the critical supersaturation is a strong function of the dry aerosol radius, with smaller particles requiring a higher supersaturation before they are activated relative to larger dry aerosols.

Within a cloud updraft, the saturation vapor pressure is continuously decreasing as air adiabatically expands and cools during lifting. When the RH initially exceeds

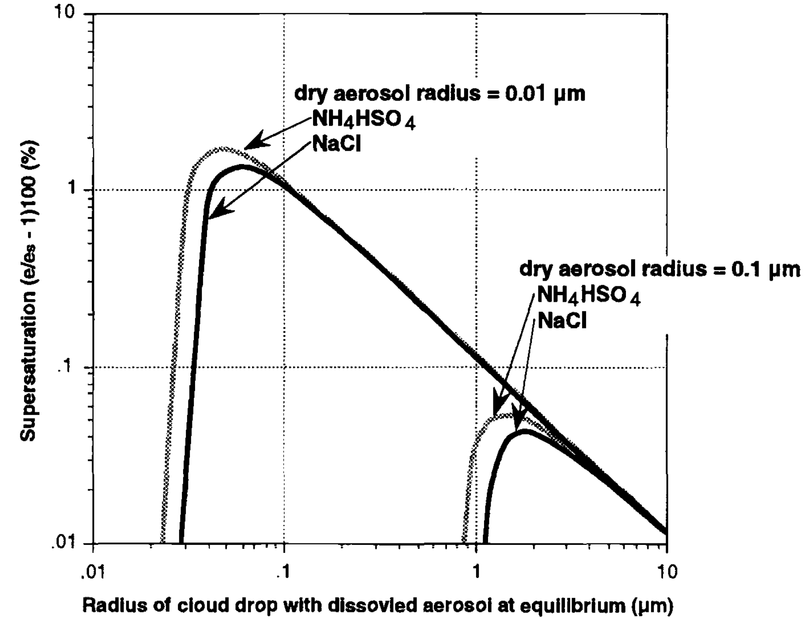


Figure 1 Equilibrium partial pressure of water vapor (expressed as supersaturation percent) as a function of the radius of a droplet containing a specified amount of dissolved salt (specified in terms of dissolved dry aerosol radius).

100%, water vapor primarily diffuses toward the few CCN containing the largest amounts of dissolved salts, which usually are the largest dry aerosols with the lowest critical supersaturations. Once a few cloud drops are activated, "excess" water vapor can either condense on existing drops or water vapor can diffuse toward and activate new CCN, depending on how rapidly the parcel is cooling. If not enough droplets are activated, the supersaturation increases, and more CCN are activated, incorporating smaller dry soluble aerosols into the cloud water. If sufficient numbers of nucleated drops present, water vapor diffuses toward existing drops, the supersaturation does not increase, and no additional drops are "activated."

The total number of cloud droplets nucleated during cloud formation depends in a very complicated way on the rate of cooling (i.e., lifting rate, or updraft velocity), and the size spectra and composition of soluble aerosols present in the cloud updraft.

Figure 2 shows one example of an explicit simulation of water vapor diffusion in an aerosol-laden cloud updraft. Supersaturations in a cloud updraft increase during the initial few meters above cloud base, and the greatest supersaturations are reached within 20 m or less above cloud base, usually within a few seconds above cloud base. Above the level of highest supersaturation, no additional CCN are activated, and condensing water vapor readily diffuses toward the already activated and growing cloud drops, which provide adequate surface area for condensation.

Numerous measurements show that condensation nuclei concentrations are typically a factor of 3 to 10 higher in continental areas relative to maritime areas, depending on the supersaturation at which CCN concentrations are measured. Figure 2 shows that in maritime environments containing relatively few CCN, supersaturations in a cloud updraft reach considerably higher values since there are so few

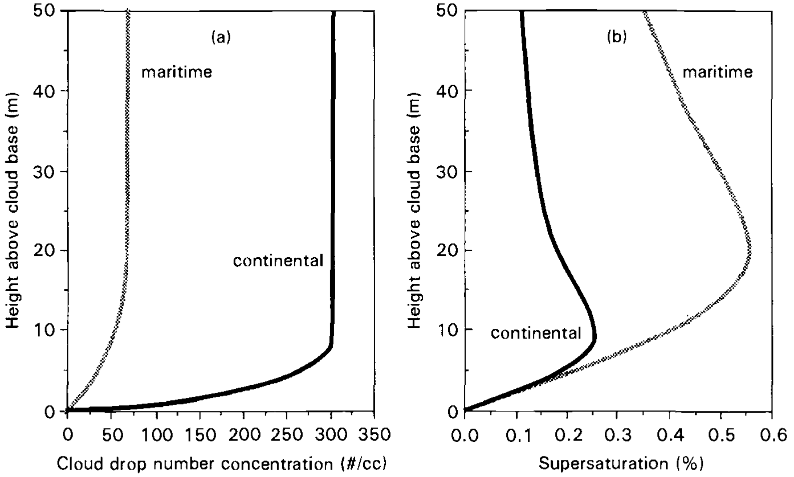


Figure 2 Initial development of cloud properties in air ascending at 1 m/s (*a*) number of cloud drops and (*b*) supersaturation. Typical maritime and continental CCN spectra assumed.

condensation nuclei present during cloud formation. Therefore, fewer drops are ultimately nucleated. In contrast, in continental areas where significantly greater numbers of CCN are present, more cloud drops form, and since vapor more efficiently diffuses toward these drops, supersaturations are appreciably lower than maritime clouds.

Figure 3 shows typical droplet number concentrations, maximum supersaturations, and the minimum dry radii of aerosols nucleated in a cloud updraft according to the Köhler theory following the approach of Twomey (1959). The influence of different aerosol size distributions on cloud properties is crudely accounted for here by assigning typical "continental" and "maritime" CCN distributions. Usually maritime conditions have lower numbers of CCN, and the differences between the maritime and continental distributions shown in Figures 2 and 3 qualitatively show the range of the natural variations that occurs in various cloud environments. This figure shows that greater numbers of cloud drops are nucleated at higher updraft velocities, due to increased cooling rate, and therefore condensation rate within the rising cloud parcel. Under continental conditions containing greater numbers of CCN and aerosols, more drops are nucleated and peak supersaturations are lower.

Figure 4 shows the fraction of soluble aerosol mass that is activated or incorporated into cloud drops during cloud formation for several updraft velocities under typical "maritime" or "continental" CCN and aerosol distributions, estimated two ways. One method is to add up the mass of the largest aerosols observed in a typical

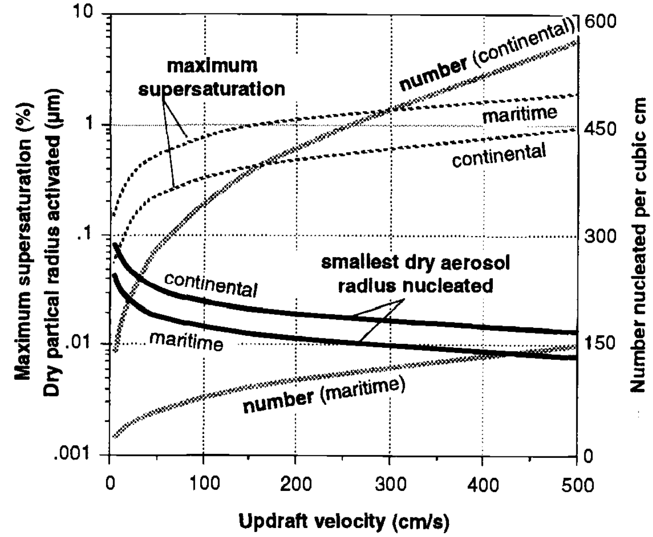


Figure 3 Maximum supersaturation, dry radius of the smallest aerosol activated, and number of cloud drops formed during condensation within a cloud updraft. Ammonium bisulfate aerosol, and typical supersaturation activation for maritime or continental air masses assumed.

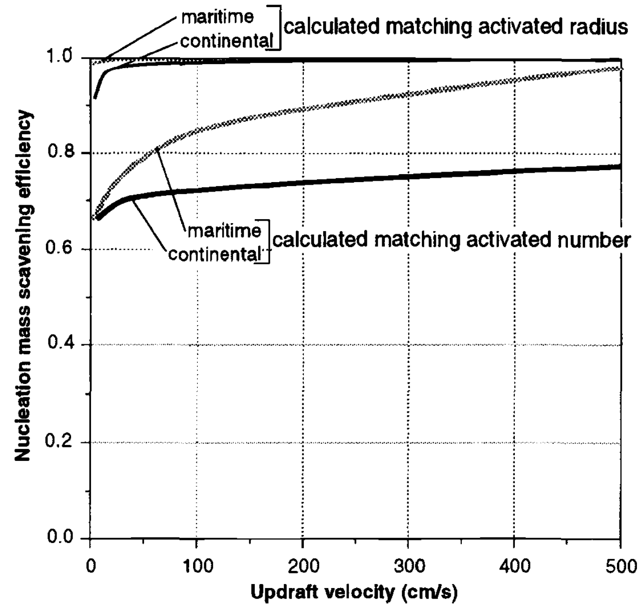


Figure 4 Fraction of aerosol mass scavenged during condensation within a cloud updraft. Activation spectra and dry aerosol size distributions typical of continental or maritime for ammonium bisulfate aerosol assumed.

dry aerosol size distribution that were nucleated during condensation. Thus if 300 cloud drops are nucleated, one can calculate the mass of the 300 largest dry aerosols in a measured aerosol size distribution, and compare this mass to the total aerosol mass in all sizes. Another method for estimating nucleation scavenging involves using the Köhler equation. Knowing the maximum supersaturation in a cloud updraft, one can calculate the size of the smallest soluble aerosol particle nucleated. Knowing this size, one can calculate the mass fraction contained in aerosol particles greater than this size from a measured aerosol size distribution. These two methods yield slightly different mass fractions and suggest that there are some inconsistencies and uncertainties in our scientific understanding of the nucleation processes and how it relates to the size distribution of dry aerosols entering a cloud. Despite these uncertainties, Figure 4 shows that a large fraction of the mass of soluble aerosols is activated and incorporated into cloudwater when a cloud forms. Irrespective of these minor uncertainties, the fraction of aerosol mass nucleated is proportional to the updraft velocity and inversely related to the number concentration of aerosol in air. Thus clouds forming under continental conditions typically scavenge a slightly smaller fraction of the aerosol mass.

The concentration of aerosol-laden trace constituents in cloud water can be given by:

$$C_l = \frac{\varepsilon_{\text{aer}} 10^6 C_T}{L} \tag{2}$$

where C_l is the liquid-phase concentration (moles per liter$_{\text{water}}$), ε_{aer} is the mass scavenging fraction of the aerosol-borne trace constituent, shown in Figure 4, C_T is the total concentration (moles per liter$_{\text{air}}$) of trace constituent in air from which the cloud forms, and L is the condensed water content of the cloud (grams water per cubic meter of air). As shown in Figure 4, the mass-scavenging fraction is usually large, and nearly all aerosol-borne constituents are incorporated into the aqueous phase within clouds during condensation. Low scavenging efficiencies for soluble aerosols occur under highly polluted, high particle number concentration conditions, or within clouds that form slowly at low cooling rates, such as fogs.

Trace Gas Scavenging

After clouds form, soluble gases rapidly diffuse toward and dissolve into the liquid phase. In the presence of liquid water, gases partition themselves between gas and aqueous phases, and the liquid-phase concentration (moles per liter) divided by the partial pressure (atm) of the dissolved constituent over the liquid at equilibrium is defined as the Henry's law coefficient (K_h moles/liter/atm), a standard measure of trace gas solubility.

For typical clouds, interstitial gases diffuse toward and establish an equilibrium with a condensed phase within a few seconds or less. Therefore, soluble gases are very close to Henry's law equilibrium with cloud drops. Under equilibrium conditions, the liquid-phase concentration of a soluble gas in cloudy air can be written in terms similar to the expression for the concentration of soluble aerosol [Eq. (2)]:

$$C_l = \frac{\varepsilon_{\text{gas}} 10^6 C_T}{L} \tag{3}$$

where C_l is the liquid-phase concentration (moles per liter$_{\text{water}}$), and as previously defined for aerosols, C_T is the total concentration (moles per liter$_{\text{air}}$) of trace gas in the cloudy air. The "scavenging efficiency" (ε_{gas}) of soluble gases can be calculated from mass conservation and equilibrium constraints as

$$\varepsilon_{\text{gas}} = \frac{1}{10^6/(K_h LRT) + 1} \tag{4}$$

Here R is the universal gas-law constant (0.082 atm liter/mol K) and T is the temperature (K). For highly soluble gases that partition predominantly into the liquid phase, the term involving K_h in (4) $\ll 1$, ε_{gas} is close to unity, and $C_1 = 10^6 C_T/L$. At the other extreme, for low-solubility gases that remain predominantly in the gas phase, the K_h term in (4) $\gg 1$, ε_{gas} is small, and therefore $C_l = C_T K_h RT$, independent of the cloud liquid water content.

Many gases rapidly dissociate into several chemical forms when they dissolve in cloud water. For example, SO_2 is a weak acid dissociating into three chemical forms: a nonionic hydrated complex (SO_2H_2O), bisulfite (HSO_3^-), and sulfite ($SO_3^=$) ions. Equilibrium expressions are defined to quantify this dissociation as

$SO_2(g) \leftrightarrow SO_2H_2O\ (aq)$	$K_h = [SO_2H_2O(aq)]/P_{SO_2}(g)$ mol/liter atm
$SO_2H_2O(aq) \leftrightarrow HSO_3^- + H^+$	$K_1 = [HSO_3^-][H^+]/[SO_2H_2O\ (aq)]$ mol/liter
$HSO_3^- \leftrightarrow SO_3^= + H^+$	$K_2 = [SO_3^=][H^+]/[HSO_3^-]$ mol/liter

In addition to Henry's law constant for SO_2, laboratory-measured equilibrium coefficients K_1 and K_2 are used to quantify the first and second dissociation of SO_2 in solution. Other gases such as organic acids, CO_2, NH_3, HNO_3, and HCl dissociate in a similar manner, and one can define an "effective" Henry's law solubility, which accounts for the concentrations of *all* dissolved chemical forms of the trace gas, which can generally be expressed as:

$$K_{\mathrm{he}} = K_h\left[1 + \frac{K_1}{[H^+]}\left(1 + \frac{K_2}{[H^+]}\right)\right] \tag{5}$$

This effective solubility must be used in (4) for estimating liquid-phase concentrations in a cloud. This effective Henry's law solubility is therefore usually a function of the concentration of hydrogen ion $[H^+]$, or the acidity of the cloudwater, which is proportional to the concentrations of the acids and bases in cloudy air, and strongly influenced by the cloud liquid water content.

Figure 5 shows the mass-scavenging fraction for gases as a function of their effective Henry's law solubility. Also shown on this figure is the approximate solubility and scavenging fraction of several trace gases of interest in atmospheric chemistry. For gases with Henry's law constants less than about 100 mol/liter atm, only an extremely small fraction enters into the liquid phase in a typical cloud. This includes most organic gases, NO, NO_2, and CO, on the order of a few percent of formaldehyde and SO_2 dissolve in a cloud. In contrast, 50 to 80% of hydrogen peroxide (H_2O_2) dissolves in most clouds, and nearly all ammonia (NH_3), nitric acid (HNO_3), and sulfuric acid (H_2SO_4) are scavenged within cloudy air.

Wet Deposition Fluxes

The rate at which trace chemicals are removed from the atmosphere via wet deposition is intimately related to the life cycle of liquid water in the atmosphere. The flux of dissolved constituents to the surface in precipitation is the product of the precipitation rate P_r (mm/h = $\mathrm{kg_{water}/m^{-2}}$ h) and the liquid-phase concentration of trace constituents in precipitation (C_l moles per liter of solution)

$$\text{Flux (mol/m}^{-2}\text{/h)} = C_1 P_r \tag{6}$$

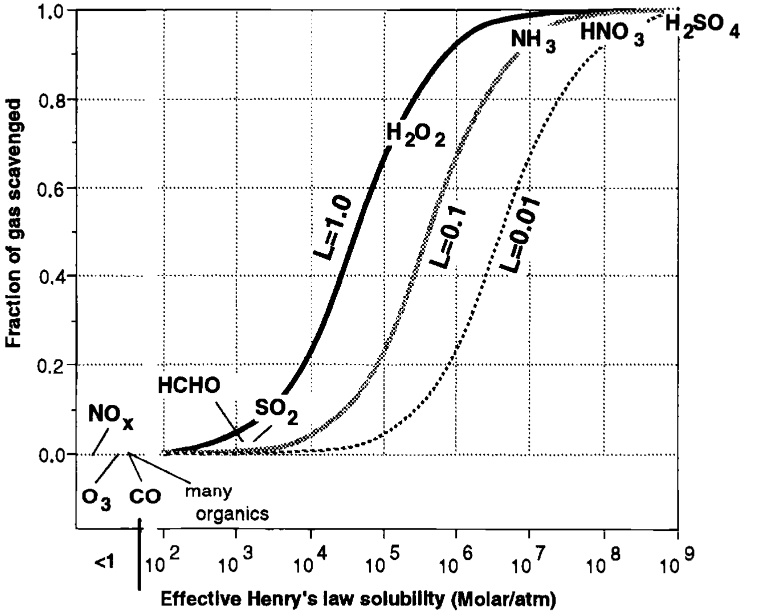


Figure 5 Fraction of gases scavenged in the presence of cloud liquid water as a function of the effective Henry's law solubility. Scavenging fraction shown for several cloud liquid water contents (L, grams liquid water per cubic meter). Approximate solubility of numerous gases shown.

Since precipitation forms by collecting and scavenging small cloud droplets from throughout a cloudy layer, the concentration of dissolved constituents in precipitation to a first approximation is proportional to the average concentration of constituents in the cloudwater from which the precipitation forms. Evaporation of precipitation below cloud base increases concentrations in precipitation, especially during the initial stages of a precipitation event. As noted above, the concentrations of aerosols and highly soluble gases in cloudwater are inversely proportional to the condensed water content in a cloud, which is in turn proportional to the amount and speed of cooling during cloud formation. The amount of cooling is proportional to the amount of lifting in convective or orographic clouds, and other radiative factors determine the cooling rate in fogs and some stratiform clouds. Thus the water content of a cloud is often proportional to the depth or vertical displacement of air within a cloud, although entrainment and mixing of dry air from outside a cloud often evaporates and dilutes condensed water in clouds.

Liquid water contents in clouds typically increases with altitude above cloud base, and therefore liquid-phase concentrations generally decrease with altitude above cloud base as the dissolved gases and aerosols are diluted by the increasing amounts of liquid water.

Figure 6 shows vertical profiles of "adiabatic" and "typical" water contents within a convective cloud updraft. Here, adiabatic water contents are calculated assuming that the updraft remains saturated as it cools, and *all* water vapor in excess of saturation remains in the updraft as liquid water. Adiabatic water contents are rarely observed in clouds and represents an upper limit to the amount of condensed water within a cloud. Measurements (Warner, 1970) typically show that water contents in convective clouds are ~20 to 40% of adiabatic values.

Since liquid-phase concentrations are proportional to total concentrations of trace substances in air [Eqs. (2)–(4)], one can derive a simple expression for the rate of change of a trace chemical concentration in a precipitating environment due to precipitation scavenging:

$$\begin{aligned}\frac{dC_T}{dt} &= -\frac{\varepsilon 10^3 P_r}{\bar{L}\,\Delta z}C_T \\ &= -\frac{C_T}{\tau_s} = -\left(\frac{\varepsilon}{\tau_{\text{cw}}}\right)C_T\end{aligned} \tag{7}$$

where Δz is the depth of the cloudy layer experiencing precipitation, and L is the water content averaged over the cloud depth (g/m^3); τ_s is a time constant for soluble species to be removed from the cloudy environment due to precipitation scavenging,

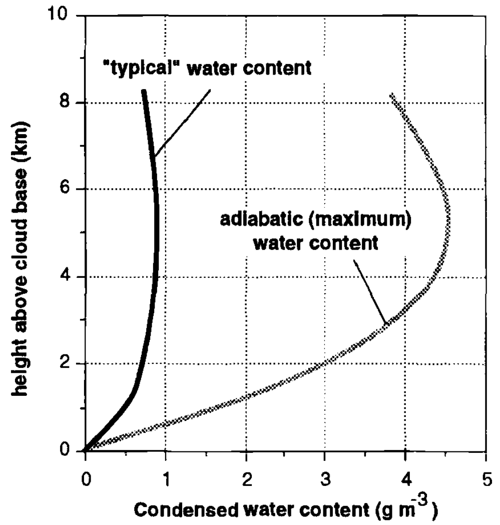


Figure 6 Condensed water content within a convective updraft vs. height above cloud base. Upper limit (adiabatic) and typical horizontal average through a cloud shown. Cloud base 10°C at 900 mbar. Typical water contents are adiabatic water contents scaled by Warner's (1970) compilation of observed reduction factors.

and is proportional to the time constant for the removal of condensed water τ_{cw} in a cloud; ε is the scavenging efficiency for either soluble aerosols (Fig. 4) or trace gases [Eq. (4), Fig. 5]. The time constant for the removal of condensed water in a cloud is the condensed water path in a cloud (mm $= L\ \Delta z/10^3$) divided by the precipitation rate (mm/h) from the cloud. For highly soluble species, the scavenging efficiency shown in Figures 4 and 5 are very high ($\varepsilon \sim 1$), and therefore the wet deposition time scale for liquid water is identical to the time constant for removal of soluble species from the cloudy environment.

Using reasonable estimates for the parameters in Eq. (7), one can quantify the wet deposition time scales for the removal of trace constituents under precipitating conditions. For a convective cloud, Figure 7 shows precipitation rate and the time scale for washout of condensed water as a function of cloud depth at various storm efficiencies. Updraft at cloud base is 1 m/s, and precipitation rates (and washout times) scale linearly with this assumed velocity. Storm efficiency is defined here as the surface precipitation rate divided by the condensation rate in the updraft. Storm efficiencies range from 10%, in relatively dry, high-wind-shear environments to 100% in saturated, low-shear environments (Weisman and Klemp, 1982; Lipps and Hemler, 1986).

Average condensed water contents for calculating the washout lifetime are taken from "typical" water contents shown in Figure 6, or closer to adiabatic (shaded gray region on Fig. 7*b*) conditions. The main point of Figure 7 is to show that time scales for removal of condensed water substance are on the order of an hour or less, and therefore soluble constituents that are completely absorbed into cloudwater are removed rapidly from the atmosphere when precipitation is occurring. For less soluble constituents, such as SO_2, which partitions only a few percent into the aqueous phase in a cloud, the time constant for wet removal is longer by a factor of 10 to 100, depending on the cloud depth and microphysical storm efficiencies.

Therefore we conclude from these semiqualitative estimates that in the immediate vicinity of precipitation systems, soluble gases and a large fraction of the mass of soluble aerosols are efficiently removed from the atmosphere on a time scale of an hour or less. Chemical species that are rapidly removed by precipitation include gaseous HNO_3, NH_3, and H_2O_2. Soluble constituents of CCN including aerosol sulfates, nitrates, and sea salts are also rapidly scavenged from the atmosphere under most precipitating conditions.

From a global perspective, the rate-limiting factor determining how quickly trace constituents are removed from the atmosphere is determined by how often and where precipitation occurs. At any given time, clouds typically cover approximately half of Earth's surface, but only a small fraction of clouds are precipitating. Pruppacher and Klett (1978) suggest that from a global perspective the average residence time of condensed water in the atmosphere is on the order of 7 h, and the residence time of all water substance (vapor + condensed water) is on the order of 9 days. Therefore on a global perspective we expect similar removal time scales for soluble trace chemicals, scaled by the relative partitioning of trace gases in condensed and vapor phases.

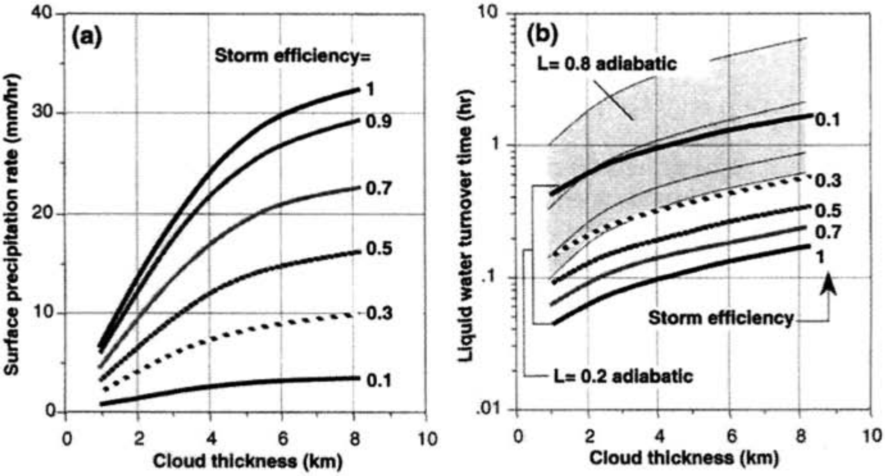


Figure 7 (*a*) Precipitation rate and (*b*) condensed water washout time scale within convective clouds as a function of cloud depth at various storm efficiencies. Updraft at cloud base is 1 m/s, and precipitation rates (and washout times) scale linearly with this assumed velocity. Storm efficiency defined as the surface precipitation rate divided by the condensation rate in the updraft. Average condensed water contents for calculating lifetime taken from typical water contents shown in Figure 6 or closer to adiabatic (shaded gray region).

Acid Deposition

An "acid" is essentially any substance that releases hydrogen ions (H^+) when dissolved in water. Several atmospheric trace constituents dissociate into positive and negative ions when dissolved in water, and some are acidic to varying degrees. The strongest acids in atmospheric waters are dissolved sulfuric acid (H_2SO_4) and nitric acid (HNO_3). Numerous other acidic substances have been identified in the atmosphere, such as sulfur dioxide (SO_2), organic acids, hydrochloric acid (HCl), carbon dioxide (CO_2), and even water itself, but these substances are either relatively weak acids or are present at relatively small concentrations and thus do not usually contribute appreciably to measured acidity. For any solution, there is an equal concentration of dissolved positive and negative ion charge, and a typical ion balance in cloudwater and precipitation is

$$\begin{aligned}[H^+] + [Na^+] + [NH_4{}^+] + [\text{soil ions}] &= (\text{positive ions}) \\ &= 2[SO_4{}^=] + [NO_3{}^-] + [Cl^-] + [HCO_3{}^-]\ (\text{negative ions})\end{aligned}$$

Na^+ and Cl^- ions arise from dissolved sea salt aerosols and are usually present in approximately equal concentrations. $NH_4{}^+$ is dissolved ammonia, "soil ions" refers to calcium and magnesium cations that are typically associated with carbonates

(HCO_3^-) in soil dust, and $SO_4^=$ and HNO_3^- are sulfuric and nitric acid. Since an ion balance is always maintained in atmospheric waters, the concentration of H^+ is

$$[H^+] = 2[SO_4^=] + [NO_3^-] + [HCO_3^-] - [NH_4^+] - [\text{soil ions}]$$

Thus the concentration of the hydrogen ion is proportional to the concentrations of sulfates, nitrates, bicarbonate (dissolved CO_2), ammonia, and carbonate-laden soil dust dissolved in cloudwater and precipitation. Measured concentrations of H^+ vary over several orders of magnitude, and therefore a logarithmic pH scale is used to quantify acidity levels in water

$$\text{pH} = -\log_{10}[H^+]$$

Using the pH scale, a decrease in one pH unit corresponds to a 10-fold increase in acidity or H^+ concentration. Also, as the pH decreases, the concentration of H^+ and acidity increases. Pure water in equilibrium with atmospheric CO_2 has a pH near 5.6, but the concentrations of sulfate, nitrates, ammonia, or soil cations in cloud and rainwater usually greatly exceed the concentrations of dissolved CO_2, even in remote areas. Typical "clean" atmospheric waters have a pH of 4.5 to 5.5. In more polluted areas, pHs in precipitation range from 3 to 4, and in some low liquid water content clouds, pHs as low as 2 to 3 have been measured.

Formation of Acids. Sulfuric and nitric acids are produced by reactions between atmospheric oxidants and emitted sulfur and nitrogen oxides (SO_2 and NO_x), which are by-products of fossil fuel combustion and other industrial activities. The dominant reactions converting SO_2 to sulfuric acid include reactions with hydrogen peroxide (H_2O_2) in clouds and hydroxyl radical (HO) in air. Nitric acid is produced by the oxidation of NO_2 by the HO radical, and also at night by a heterogeneous reaction involving ozone, NO_2, and NO_3 radicals. Atmospheric oxidants responsible for acid formation are produced via a complex sequence of photochemical reactions, and some acid-generating chemical reactions occur among dissolved gaseous constituents in atmospheric clouds or aerosols. Typically, emitted NO_x is converted to nitric acid within a day or less, and SO_2 is converted to sulfuric acid within several days following emission. The concentrations of the oxidants, and the time scale for chemical reactions vary strongly with season, latitude, time of day, sunlight intensity, background concentrations of NO_x and organic compounds, and many other chemical and meteorological factors.

Strong acids have an affinity for water, and therefore hygroscopically grow or combine with water vapor to form "haze" aerosols containing sulfuric acid, nitric acid, and varying degrees of neutralizing ammonia (NH_3), especially when atmospheric relative humidities are above 60 to 70%. Typically, ammonia and nitric acid are present as both gases and aerosols in the atmosphere, while sulfate partitions predominantly into condensed aerosols. These sulfate, nitrate, and ammonium-containing aerosol particles constitute a significant fraction of cloud condensation nuclei (CCN), and thus acid-containing aerosols are readily incorporated into clouds.

Precipitation forming within clouds therefore contains dissolved CCN together with other soluble gases such as HNO_3 and NH_3.

Undesirable Effects. At high concentrations or exposures, acidic solutions induce numerous undesirable reactions with surfaces. In conjunction with other pollutants, acid deposition contributes to potentially deleterious effects on aquatic, agricultural, and forest ecosystems. Chemical changes attributed to the deposition of acidity from the atmosphere have been measured in forest ecosystems and surface waters.

Concentrations of acids in lakes have been correlated with the concentrations and deposition rates of atmospheric acids, and high concentrations of acids in lakes and streams can adversely affect fish populations. Health effects associated with exposure to acid-containing particulates in humans remains an area of uncertainty since current studies of these effects are too limited to unambiguously discern dose–response relationships in humans. Acid deposition from the atmosphere has been shown to accelerate the deterioration rate of exposed metals, painted finishes, and concrete or stone surfaces.

In industrialized areas, concentrations of sulfuric and nitric acids in cloud water and precipitation are up to 50 to 100 times greater than values measured in areas that are not influenced by upwind emissions of anthropogenic pollutants. The relative concentrations of deposited sulfur and nitrogen acids are correlated with the relative emission rates of sulfur and nitrogen pollutants over larger areas.

The amount of acidity within precipitation is strongly influenced by numerous meteorological and chemical factors, as well as the emission rates of precursor sulfur and nitrogen pollutants. Therefore a thorough understanding of larger-scale meteorology, cloud dynamics and microphysics, and atmospheric chemistry is required to fully quantify and study atmospheric acidity

REFERENCES

Köhler H., Zur thermodynamic der kondensation an hygroskopichen kernen und bemerkungen über das zusammenfliessen der tropfen, *Medd. Met. Hydr. Anst. Stockholm*, 3(8), 1926.

Lipps, F. B., and R. S. Hemler, Numerical simulation of deep tropical convection associated with large-scale convergence, *J. Atmos. Sci.*, *43*, 1796–1816, 1986.

Pruppacher, H. R., and J. D. Klett, Microphysics of Clouds and Precipitation, D. Reidel Publishing, 714 pp., 1978.

Twomey, S., The nuclei of natural cloud formation: The supersaturation in natural clouds and the variation of cloud droplet concentration, *Geophys. Pura Appl.*, *43*, 243–249, 1959.

Warner, J., *J. Atmos. Sci.*, *27*, 1035, 1970.

Weisman, M. L., and J. B. Klemp, The dependence of numerically simulated convective storms on vertical wind shear and buoyancy, *Monthly Weather Rev.*, *110*, 504–520, 1982.

CHAPTER 20

LARGE-SCALE CIRCULATION OF THE STRATOSPHERE

WILLIAM L. GROSE

1 GOVERNING EQUATIONS

In their most basic form, the governing equations that determine the circulation and thermal structure of the atmosphere are referred to as the primitive equations. The following version of those equations do include some approximations, and the reader is referred to a standard text such as Holton (1992) for a fuller discussion of the details and the advantages of using pressure, p, as a vertical coordinate. In vector form the horizontal momentum equation can then be written in an (x, y, p) coordinate system as

$$\frac{DV}{Dt} = -f\mathbf{k}xV_h - \nabla_p\Phi \tag{1}$$

with the total derivative given as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + \omega\frac{\partial}{\partial p}$$

and ∇_p is the horizontal gradient operator (with partial derivatives taken holding p constant). Here V_h is the horizontal velocity, Φ is the geopotential, f is the Coriolis parameter, and $\mathbf{k}$ is a unit vector in the vertical direction. Also, x is the west-to-east coordinate, y is the south-to-north coordinate, and u and v are the zonal and meri-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

dional components, respectively, of the horizontal velocity, V_h. The variable, ω is defined as

$$\omega = \frac{Dp}{Dt}$$

and becomes the surrogate for the vertical velocity, w, in the (x, y, p) coordinate system.

For large-scale motions, the vertical component of velocity is typically several orders of magnitude smaller than for the horizontal velocity, and vertical accelerations can be neglected to a good approximation. The vertical component of the momentum equation then reduces to a diagnostic equation and can be expressed as

$$\frac{\partial \Phi}{\partial p} = \frac{RT}{p} \tag{2}$$

with R the gas constant for air and T the temperature.

The mass continuity equation in this coordinate system becomes

$$\nabla_p \cdot V_h + \frac{\partial \omega}{\partial p} = 0 \tag{3}$$

Finally, the thermodynamic energy equation becomes

$$\frac{DT}{Dt} = \frac{\omega RT}{C_p p} + \frac{Q}{C_p} \tag{4}$$

with C_p the specific heat at constant pressure for air. The variable Q is the diabatic heating rate. For the stratosphere, Q is typically the radiative heating rate and can be calculated as a function of the other dependent variables, knowing the distribution of radiatvely active species such as ozone, water vapor, and carbon dioxide (Goody, 1995).

Equations (1) to (4) represent the primitive equations in the (x, y, p) coordinate frame and form a determinate system of equations for the variables u, v, ω, Φ and T. These equations are inherently nonlinear and must be integrated in time with suitable initial and boundary conditions. Direct solution of the primitive equations requires the use of sophisticated numerical techniques implemented on digital computers.

A relationship of fundamental importance can be derived from the primitive equations [see Pedlosky (1979) for the derivation] in terms of Ertel's potential vorticity, Π, namely

$$\frac{D\Pi}{Dt} = 0 \tag{5}$$

for frictionless, adiabatic conditions. Here, Π, is given by

$$\Pi = \frac{1}{\rho}(\zeta + f)\nabla\Theta \tag{6}$$

Note that in Eq. (6) the total derivative is now expressed in the Cartesian coordinate system (x, y, z) as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}$$

Also, ∇ in Eq. (6) is now the three-dimensional gradient operator in the (x, y, z) coordinate system with ρ the density, ζ the relative vorticity (curl of the velocity V), and Θ the potential temperature. For frictionless, adiabatic flow, this relationship requires that Π be conserved following the motion. Physically, potential vorticity is representative of the ratio of the rotation of a fluid vortex column to the depth of the column. Conservation of potential vorticity for a compressible fluid can be thought of as analogous to conservation of angular momentum for a solid body. This conservation property for Π represents a very powerful constraint on the motions, particularly in the lower stratosphere where Eq. (5) is a reasonable approximation for time scales up to about 10 days. Distributions of Π on an isentropic surface (i.e., a constant potential temperature surface) thus represent a conserved dynamical tracer and, as we shall later see, provide much useful insight into the nature of the transport.

2 VERTICAL TEMPERATURE STRUCTURE

A traditional means of separating Earth's atmosphere into regions with quite distinct characteristics originated from considerations of the vertical temperature structure. A reference temperature profile for middle latitudes is shown in Figure 1 as a function of the geometric height using data from the compilation, *U. S. Standard Atmosphere* (1976). The difference between the stratosphere and troposphere in terms of the vertical temperature profile is readily apparent. Note that in the troposphere, the temperature T, decreases with increasing height z, up to the level of the tropopause. In contrast, the temperature in the lower stratosphere is nearly constant and then increases with increasing height through the middle and upper stratosphere until the level of the stratopause is reached. It is this difference in the temperature lapse rate, Γ, where

$$\Gamma = -dT/dz \tag{7}$$

that results in a difference in the stability characteristics between these two regions of the atmosphere.

The atmosphere is said to be statically stable, if after an air parcel is adiabatically displaced in the vertical dimension, the net forces acting on the parcel tend to restore

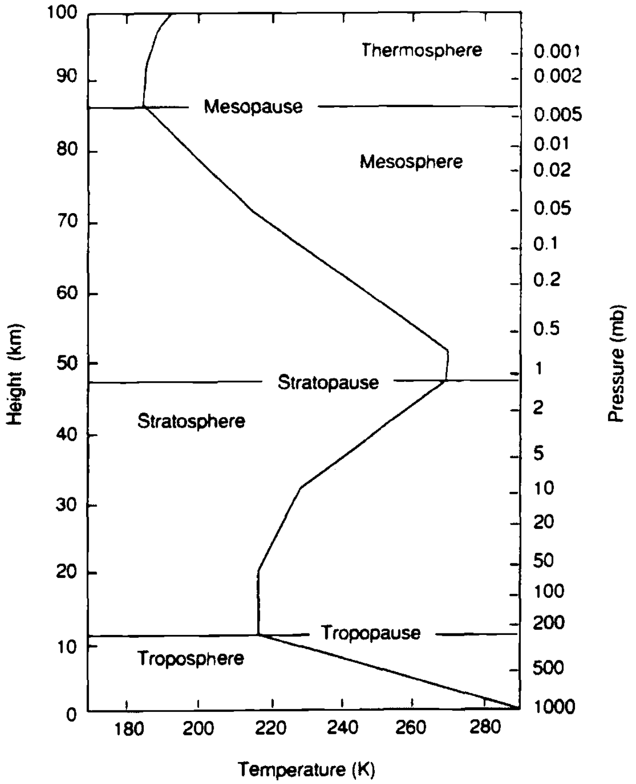


Figure 1 Mean temperature profile at midlatitudes based upon the *U. S. Standard Atmosphere* (1976) (from Holton, 1992).

it to its unperturbed position. The usual diagnostic for examining the static stability of the atmosphere is the buoyancy frequency or Brunt–Vaisala frequency (Holton, 1992), N, where

$$N^2 = [g(\Gamma_d - \Gamma)/T] \tag{8}$$

with g being the acceleration due to gravity and Γ_d being the dry adiabatic lapse rate (9.8°/km). The buoyancy frequency is the frequency of oscillation of an air parcel that has been adiabatically displaced from its equilibrium position in an atmosphere at rest. A situation with $N^2 > 0$ ($\Gamma < \Gamma_d$) corresponds to a stably stratified atmosphere. An inspection of Figure 1 reveals that N^2 would be greater than zero (Γ is zero or negative) in the stratosphere, and hence, this region should be stable. Indeed, in practice it has been found that vertical mixing in the stratosphere is much slower than is typical for the troposphere, where air parcels can be transported vertically between the surface and the tropopause on time scales of several days or even as little as a few minutes in the case of very strong convective activity associated with the largest thunderstorms (Wallace and Hobbs, 1977).

This stable stratification with relatively slow vertical mixing is an important influence on the stratospheric circulation and gives rise to the long residence times in the stratosphere that have been deduced for long-lived constituents such as the chlorofluorocarbons (CFC) [appproximately 50 years for CFC-11 and 100 years for CFC-12, the two most abundant CFC compounds (WMO, 1994)]. Here, long-lived is used in the sense that the characteristic time scale for chemical change is very much greater than that associated with the transport of a constituent.

3 ZONAL-MEAN CLIMATOLOGY OF TEMPERATURES AND ZONAL WINDS

Climatologies of zonal-mean (spatial averages at constant latitude) values of temperatures and zonal (east–west component) winds are displayed as a function of height and latitude in Figure 2*a* and 2*b*, respectively. An inspection of the temperature cross section shown in Figure 2*a* reveals marked variations in temperatures between the winter and summer hemispheres in the stratosphere. Note that the tropopause above the equator occurs at a much higher altitude than above the polar regions and that distinct discontinuities or "breaks" occur in the tropopause at middle latitudes. In the lower, summer stratosphere the temperature increases from the equator toward the pole. In contrast, there is a midlatitude warm belt (Ramanathan and Grose, 1978) in the lower, winter stratosphere. Coldest temperatures occur in the lower stratosphere over the south polar region during winter. The northern polar regions exhibit more dynamical variability during winter and are typically not quite as cold. In the upper stratosphere the temperature increases monotonically from winter to summer pole.

The corresponding cross section of zonal winds is shown in Figure 2*b*. Note the presence of easterly winds (from east toward the west) in the summer hemisphere and westerly winds (from west toward the east) in the stratosphere. The seasonal reversal of summer easterlies to winter westerlies in each hemisphere is observed to be a prominent feature of the stratospheric circulation. The axis of the wintertime, westerly jetstream tilts poleward with decreasing height in the stratosphere resulting in a high latitude jetstream in the lower stratosphere, the so-called polar night jet.

4 ZONAL-MEAN MERIDIONAL CIRCULATION

The zonally averaged, meridional (north–south component) and vertical winds are, on average, at least an order of magnitude smaller than the zonal winds described in the previous section. A useful depiction of the circulation in the meridional plane (latitude vs. height), was originally derived by Murgatroyd and Singleton (1961).

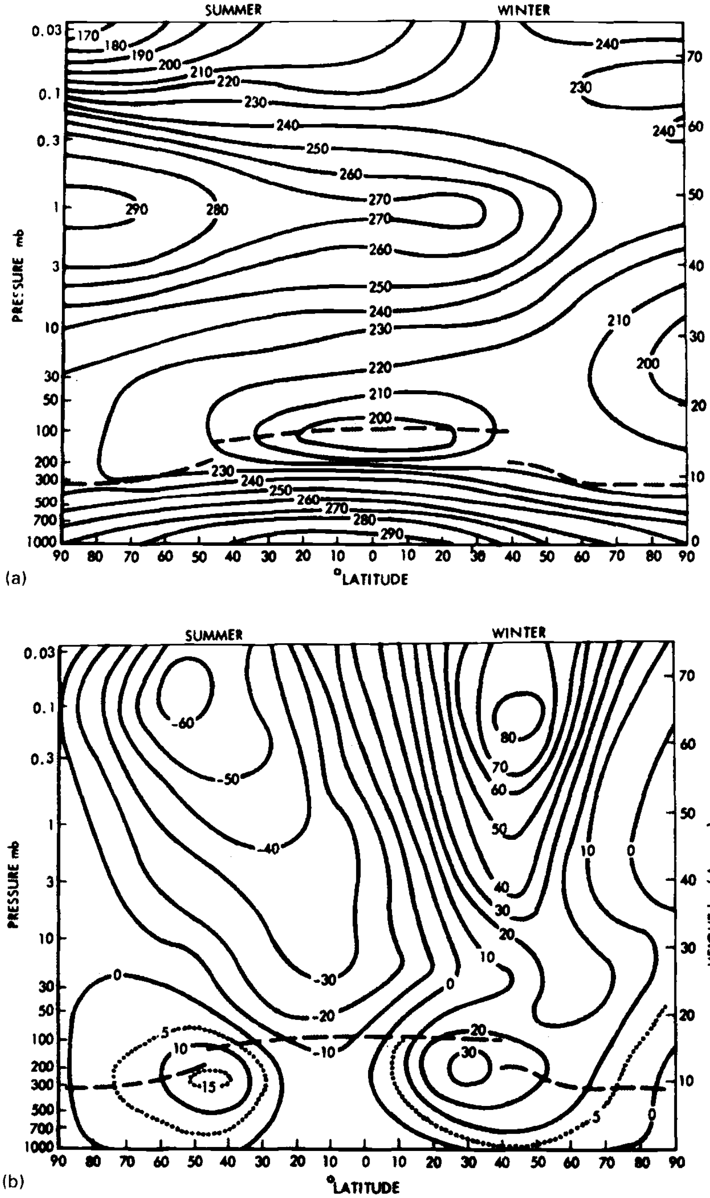


Figure 2 Latitude–altitude cross section at solstice conditions of zonally averaged (*a*) temperature (K) and (*b*) zonal wind component (m/s) (from Murgatroyd, 1969).

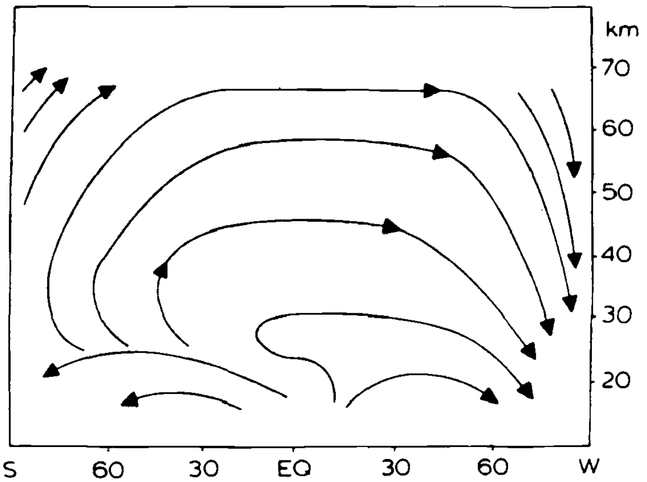


Figure 3 Schematic streamlines of the diabatic circulation for solstice conditions. S, the summer pole and W the winter pole (from Dunkerton, 1978).

This circulation is now generally referred to as the diabatic circulation. It is conceptually useful in that it illustrates the sense of the actual mass motions in the meridional plane. Using net radiative heating rates originally compiled by Murgatroyd and Singleton for solstice conditions, Dunkerton (1978) derived the vertical velocities necessary to balance these heating rates. From considerations of mass continuity, the corresponding meridional velocities were then calculated. Figure 3 shows the resultant streamlines inferred by Dunkerton (1978) from the calculated velocity fields. The large-scale stratospheric circulation in the meridional plane exhibits rising motion in the summer hemisphere with a slow (seasonal or longer time scale) meridional drift and subsidence over the winter pole (Andrews et al. 1987). A Coriolis torque associated with the meridional drift influences the production of the summertime (wintertime) zonal-mean easterlies (westerlies) in the stratosphere seen in Figure 2*b*.

5 WAVE MOTIONS

The depiction of the large-scale stratospheric circulation discussed in the preceding two sections is a zonally averaged perspective as noted. However, wavelike disturbances (commonly referred to as "waves"), which propagate vertically from the troposphere producing departures from zonal symmetry, are known to be important in determining the circulation and the transport of constituents in the stratosphere.

Important departures from the climatological zonal mean state shown in Figure 2*a* and 2*b* occur as a result of sudden stratospheric warmings, the quasibiennial oscillation (QBO) and the semiannual oscillation (SAO). Occurrence of these latter two phenomena is presently believed to result at least in part from vertically propagating Kelvin, Rossby-gravity, and/or gravity waves. The QBO manifests itself as an oscillation of the zonal winds with alternating westerlies and easterlies in the equatorial lower stratosphere. The period of the oscillation varies, but averages about 27 months. The SAO is also an oscillation of equatorial zonal winds with alternating westerlies and easterlies, but this phenomenon occurs in the upper stratosphere (and lower mesosphere) on a semiannual basis as the name implies. The mechanisms responsible for the QBO and SAO are quite different. The interested reader is referred to Andrews et al. (1987) for further details.

The sudden stratospheric warming phenomenon occurs during some years in association with anomalous enhancement of the amplitude of vertically propagating, planetary-scale disturbances into the wintertime stratosphere. Major warming events are characterized by very rapid increases in polar temperatures (50 to 70 K in a week or less) and severe disruptions of the wintertime westerly polar vortex with zonal easterlies replacing zonal westerlies at high latitudes. These warming events can occur in either hemisphere, although generally Southern Hemisphere warming events tend to be somewhat less spectacular than their counterpart in the Northern Hemisphere.

The various different types of waves [e. g., gravity waves, Rossby waves, Kelvin waves, and Rossby-gravity waves; see Andrews et al. (1987), for further description of these and other types of waves] are often distinguished by the restoring mechanism that is responsible for producing the wavelike motions. In particular, both gravity waves and Rossby waves play a most significant role with respect to the large-scale motions and the concomitant transport of constituents observed in the stratosphere.

The restoring force for the gravity wave is the buoyancy force, which is proportional to N^2 and results from stable density stratification in the atmosphere. One of the most important consequences of gravity waves propagating upward into the stratosphere is their role in determining the structure of the stratospheric jets. Gravity waves propagate upward and "break" (Lindzen, 1981) at some level in the stratosphere or in the mesosphere. Breaking occurs as the gravity wave grows in amplitude, eventually producing an unstable lapse rate with resultant turbulence and mixing. Momentum deposition occurs as a result of the breaking process and effectively creates a net drag on the zonal momentum budget and decelerates the zonal-mean flow. Gravity wave breaking in the lower stratosphere is believed to contribute to the separation of the tropospheric and stratospheric jets as seen in Figure 2*b*. In a similar fashion, gravity wave breaking in the mesosphere contributes to deceleration and closing off above the stratospheric jets in the mesosphere.

The Rossby wave restoring force ultimately results from the latitudinal gradient in the Coriolis parameter, f,

$$f = 2\Omega \sin \phi \qquad (9)$$

where Ω is the angular velocity of rotation of Earth and ϕ is the latitude. The Rossby wave is responsible for much of the irreversible quasi-horizontal transport that occurs in the extratropical wintertime stratosphere by a process termed Rossby wave breaking (McIntyre and Palmer, 1984), which is fundamentally different from the gravity wave breaking process just discussed. Contours of constant values of Ertel's potential vorticity, Π, on an isentropic surface (a constant Θ surface) represent material lines when Eq. (5) is valid. Rossby wave breaking occurs as the wave amplifies, and material lines buckle and deform irreversibly. A graphic illustration of the process can be seen in Figures 4*a* and 4*b*. This sequence of figures depicts conditions in the midstratosphere (about 30 km) during a stratospheric warming in January 1979. The shaded area in Figure 4*a* correlates very well with the polar vortex, which was nearly centered on the pole on January 17, 1979. Ten days later (Fig. 4*b*), a large-amplitude wave resulted in elongation of the polar vortex with a tongue of low potential vorticity air from the vortex being drawn around the Aleutian anticyclone (which is centered near 65°N, 150°W at this time), and discrete parts of this tongue are seen to be scattered around the anticyclone. McIntyre and Palmer (1983) constructed these isentropic distributions of Π from meteorological analysis data and refer to them as a "coarse-grained" view of the potential vorticity distribution in the stratosphere.

Concurrent observations of ozone (Leovy et al., 1985) show a very high degree of correlation with the potential vorticity distribution and testify to the essential correctness of the Rossby wave-breaking paradigm of McIntyre and Palmer (1983) for transport and mixing in the extra-tropical stratosphere.

6 SUMMARY

Earth's atmosphere is a thin layer of gas rotating with the planet and externally forced by differential radiative heating by the Sun. The somewhat simplified picture of the large-scale circulation in the stratosphere that has been presented here has been separated into two component parts for convenience, a zonally averaged representation and a contribution from wavelike disturbances. The zonally averaged circulation described here is one that is thermally driven with air parcels heated and ascending at low latitudes, drifting slowly poleward, and finally cooling and descending at higher latitudes. Concurrently, wavelike disturbances propagating upward produce departures from the zonally averaged state. These disturbances transport and mix heat, momentum, and constituents. The reader should be reminded that chemical transformations are also taking place that not only alter the composition of the stratosphere, but also affect the radiative heating as the composition changes (principally changes in ozone, water vapor, and carbon dioxide). It should also be noted that the process by which air enters and leaves the stratosphere (the stratospheric–tropospheric exchange process) is considerably more complicated than described here. A recent comprehensive review of stratospheric–tropospheric exchange can be found in Holton et al. (1995). The stratosphere should be viewed as a very complex system in which radiative, chemical, and dynamical processes mutually interact to determine the structure and composition.

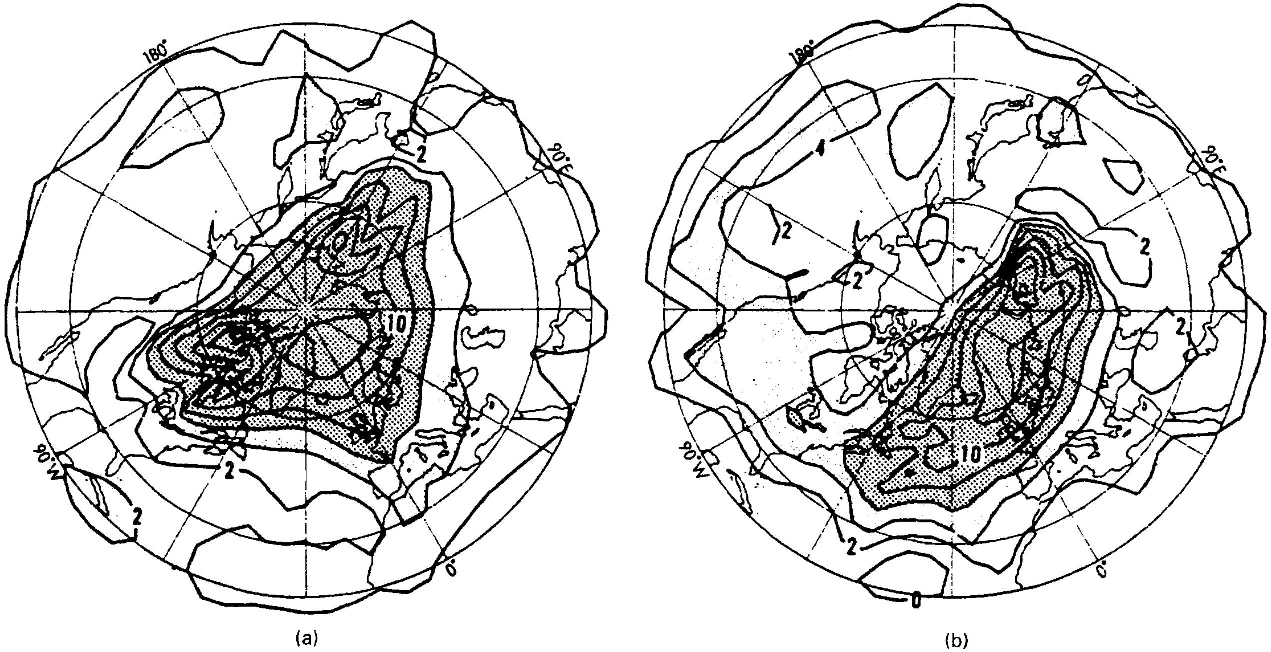


Figure 4 Potential vorticity distribution on the 850 K isentropic surface for (*a*) January 17, 1979 and (*b*) January 27, 1979. Polar stereographic projection with outermost latitude circle 20°N (from McIntyre and Palmer, 1983).

REFERENCES

Andrews, D. G., J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, Academic, London, 1987.

Dunkerton, T. J., On the mean meridional mass motions of the stratosphere and mesosphere, *J. Atmos. Sci.*, *35*, 600–614, 1978.

Goody, R., *Principles of Atmospheric Physics and Chemistry.* Oxford University Press, New York, 1995.

Holton, J. R., *An Introduction to Dynamic Meteorology*, Academic, San Diego, CA, 1992.

Holton, J. R., P. H. Haynes, M. E. McIntyre, A. R. Douglass, R. B. Rood, and L. Pfister, Stratosphere-troposphere exchange, *Rev. Geophys.*, *33*, 403–439, 1995.

Lindzen, R. S. Turbulence and stress owing to gravity wave and tidal breakdown, *J. Geophys. Res.*, Vol. *86*, 9707–9714, 1981.

McIntyre, M. E., and T. N. Palmer, Breaking planetary waves in the stratosphere, *Nature*, *305*, (5935), 593–600, 1983.

Murgatroyd, R., The structure and dynamics of the stratosphere, in G. A. Corby (Ed.), *The Global Circulation of the Atmosphere*, Royal Meteorological Society, London, 1969.

Murgatroyd, R., and F. Singleton, Possible meridional circulations in the stratosphere and mesosphere, *Q. J. R. Meteorol. Soc.*, *87*, 125–135, 1961.

Pedlosky, J., *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1979.

Ramanathan, V., and W. L. Grose, A numerical simulation of seasonal stratospheric climate: Part I. Zonal temperatures and winds, *J. Atmos. Sci.*, *35*, 600–614, 1978.

U. S. Standard Atmosphere. U.S. Government Printing Office, Washington, DC, 1976.

Wallace, J. M., and P. V. Hobbs, *Atmospheric Science: An Introductory Survey*, Academic, San Diego, 1977.

WMO Scientific Assessment of Ozone Depletion, WMO Global Ozone Research and Monitoring Project, Report No. 37, World Meteorological Organization, Geneva, Switzerland, 1994.

CHAPTER 21

STRATOSPHERIC OZONE OBSERVATIONS

JACK A. KAYE AND JACK FISHMAN

1 INTRODUCTION

The accurate knowledge of the distributions of ozone (O_3) in the global atmosphere is important for several reasons. First, the amount of ozone in the atmosphere plays a significant role in determining the amount of biologically damaging ultraviolet (UV) radiation that can reach Earth's surface. Second, ozone both absorbs and emits radiation in the atmosphere; this must be accounted for in atmospheric circulation models if they are to correctly represent the temperature and wind distributions in the atmosphere, especially in the upper troposphere and lower stratosphere. Finally, ozone together with the hydroxyl (OH) radical formed in the atmosphere in ozone photochemistry are key atmospheric oxidants. Hydroxyl plays a particularly important role by initiating much of the chemistry associated with air pollution and by being the primary destruction mechanism for several long-lived chemical compounds that contribute to global warming.

Unlike any other atmospheric phenomenon, the U.S. Congress has mandated that the National Aeronautics and Space Administration (NASA) prepare reports describing the status of our current understanding of the upper atmosphere (Public Law 101-549). In accordance with this mandate, several documents have been issued since 1985 in the form of World Meteorological Organization reports; these summaries contain a plethora of information about stratospheric ozone data as well as supporting measurements of other trace gases critical to the destruction of the stratospheric ozone layer. Much of the information in this section can be found in the last few of these reports (Albritton and Watson, 1991; Albritton et al., 1994, 1998), and the reader is referred to these studies for additional in-depth information.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

For the purposes of understanding surface UV radiation, the main quantity for which knowledge is needed is the total column amount of ozone (the integrated ozone amount in a single "column" of the atmosphere). For atmospheric circulation models and both air pollution and chemical oxidation studies, the distribution of ozone as a function of altitude (the vertical profile) must be known. Accurate, long-term knowledge of ozone distributions is important if changes in surface UV flux, upper atmosphere temperature distributions, and concentrations of both long- and short-lived pollutants are to be quantitatively understood.

For ozone distributions to be used in quantitative studies, they must be measured with high accuracy and precision. A particularly important form of precision is long-term measurement stability, as there is strong evidence for long-term changes in ozone distribution (both total column and vertical profile) over much of Earth's surface. Without excellent (and well-characterized) long-term measurement stability, it is difficult to differentiate gradual long-term changes in actual ozone distributions from inaccuracies or drift in the measurement systems.

The measurement of ozone distributions in the atmosphere presents both a challenge and an opportunity. The challenge comes mainly from its limited amount—the amount of ozone in a given volume of air is always quite small, nearly never exceeding a mixing ratio (mole fraction) of 10 parts per million by volume (ppmv) in the stratosphere and on the order of 100 parts per billion by volume (ppbv) in the polluted troposphere. Amounts in the unpolluted troposphere can be significantly smaller than the latter figure (see Chapter 3.) The mixing ratio of ozone can vary significantly with altitude, including a significant increase with altitude on going from the troposphere into the stratosphere, as well as the existence of thin layers (laminae) of air with concentrations of ozone that differ from those of the surrounding air. In regions of severe ozone depletion (such as the lower stratosphere over the Antarctic in the Austral spring), ozone may be nearly completely absent.

The total column amount of ozone typically varies in the range of 3×10^{18} molecules/cm^2 to $\sim 1.5 \times 10^{19}$ molecules/cm^2. This column amount is expressed in Dobson units (DU), which correspond to the thickness of the layer of ozone (in thousandths of a centimeter) that would be formed if all the ozone in a column of air were brought to the surface. The conversion factor is $1\ \mathrm{DU} = 2.69 \times 10^{16}$ molecules/cm^2. The amounts given above correspond to the Antarctic in Austral springtime (~100 DU) and the Arctic region in late winter (~500 DU).

Other challenges to ozone measurement can stem from potential interferences. Measurement techniques based on chemical reactivity may be affected by the presence of other oxidizing species, such as sulfur dioxide (SO_2), while spectroscopic measurements may be affected by the presence of species with absorption features in the same wavelength regions as those of ozone. Spectroscopic measurements may also be significantly affected by the presence of aerosol particles, and most such techniques (microwave and far infrared measurements being the exception) cannot penetrate clouds. Finally, it is worth noting that ozone is present in several isotopic forms. The dominant one (~99%) consists of three atoms of the dominant isotopic form of oxygen atoms (^{16}O), but there are also forms involving ^{18}O and ^{17}O, whose chemical abundance is ~0.3 and ~0.04% of that of ^{16}O.

Although this chapter focuses on the measurement of ozone, it is important to remember that ozone measurements cannot be understood (and especially, the causes of observed changes be understood) if measurements of related parameters, such as temperature, aerosol distributions, and other chemical constituents, are not also made. Indeed, many measurement networks, aircraft-based research platforms, and satellites make several of the measurements together to allow for maximum utility of the information obtained.

2 PROPERTIES OF OZONE AFFECTING ITS MEASUREMENT

The ozone molecule contains three oxygen atoms and is shaped in the form of an isosceles triangle, with a bond angle of 117° and the length of the bond is ~0.13 nm. The electronic spectroscopy of ozone is very rich owing to the multiplicity of electronic states that arise from the combination of a triply degenerate oxygen molecule $[O_2(^3\Sigma_g{}^-)]$ with a ninefold degenerate oxygen atom $[O(^3P)]$, as well as the presence of relatively low-lying states of both atomic oxygen $[O(^1D)]$ and molecular oxygen $[O_2(^1\Delta), O_2(^1\Sigma)]$. The resulting electronic spectra consist of several important band systems covering the range from the ultraviolet to the near infrared. Figure 1 shows the absorption properties of the ozone molecule in the ultraviolet and visible parts of the electromagnetic spectrum. These strong spectral features—the Hartley and Huggins bands in the ultraviolet and the Chappuis band in the visible—demonstrate the potential for use of ozone spectra in its measurement. In particular, the sharp variation in ozone absorption near 320 nm shows that ultraviolet measurements shortward of this wavelength should be very useful for ozone measurements. The much weaker Chappuis band (near 600 nm) may be useful where long path lengths are available or where ozone amounts are sufficiently high that near saturation could occur with shorter wavelengths.

3 TOTAL COLUMN MEASUREMENTS

Ground-Based Methods

The measurement of the total column amount of ozone in the atmosphere goes back more than 70 years with the development of an ultraviolet technique by Dobson. This technique, still used today, has formed the backbone of all global measurement programs for ozone columns. The basic physics of this technique is relatively straightforward. UV radiation from the sun will be absorbed by ozone in the atmosphere, so ground-based measurements of surface UV flux will contain information about the integrated ozone amount in the atmosphere. As noted in the previous section, other processes, such as Rayleigh and Mie scattering involving atmospheric aerosol particles will also affect ozone measurements. The Dobson technique involves the use of pairs of UV wavelengths corresponding to features with different strengths in ozone's UV spectrum. Since the wavelength sensitivity over a relatively

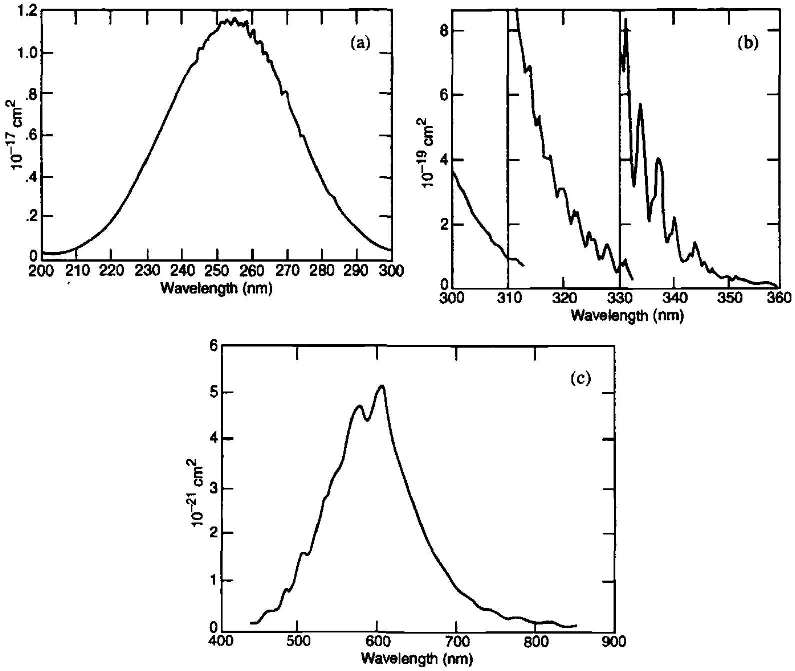


Figure 1 Electronic absorption spectrum of ozone: (*a*) Hartley band and (*b*) Huggins bands in the ultraviolet, and (*c*) the weaker Chappuis band, centered near 600 nm, in the visible.

short spectral region of the scattering by aerosols is much less than that of ozone's UV absorption spectrum, a much improved estimate of total ozone amounts can be retrieved (assuming knowledge of the ultraviolet flux from the sun is available at the two wavelengths used). Different pairs of wavelengths may be used for different amounts of ozone; typically used pairs include 312/331 nm and 318/340 nm.

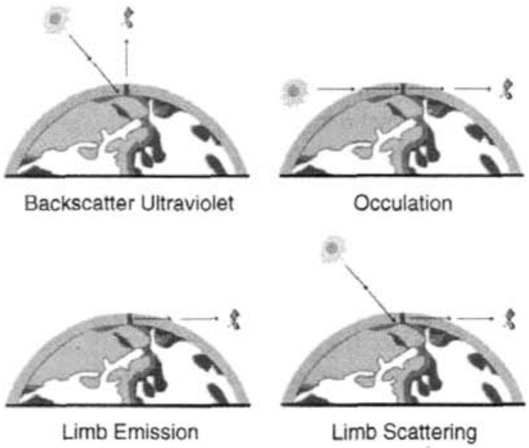
The distribution of Dobson instruments increased dramatically in the 1950s, and now there is excellent coverage over much of the world with Dobson-type instruments (which included not only those operating on the above principle using a limited number of fixed wavelengths, but also other instruments such as the Brewer spectrophotometer and filter photometers used primarily in the former Soviet Union). Like all surface-based instruments, the Dobson network lacks coverage over much of the ocean-dominated Southern Hemisphere and has fewer stations in developing countries than in industrialized nations. As noted above, such surface-based measurements will not provide data in the presence of clouds, which can be a significant limitation in the tropics, where cloudiness associated with the upward part of the Brewer–Dobson circulation is a common occurrence.

Although the Dobson technique is the most common surface-based one for measurements of the total ozone column, other approaches have been used in the past. These include those using both infrared and visible/UV wavelengths. In the latter, a variant of the Dobson technique, known as differential optical absorption spectroscopy (DOAS) is used. These other techniques have the advantage of not requiring direct sunlight to make measurements, which may be of particular importance in attempts to measure ozone in polar night (when moonlight can be sufficient for ozone measurements).

Space-Based Measurements

The primary space-based measurement technique used for measurements of total column ozone is the backscatter ultraviolet (BUV) technique. This is really a spaceborne analog of the Dobson technique, except when used from space one must account for the fact that the UV radiation passes through the atmosphere at least twice—once on the way from the sun to the surface (or scattering/absorbing layer) and once on its return to the measuring spacecraft. Account must also be taken for the UV reflectivity of the underlying ground or cloud surface. A schematic diagram of how satellite measurements are taken using four different methods is given in Figure 2. The backscatter ultraviolet (as its name implies) utilizes the ozone absorption characteristics in the ultraviolet portion of the spectrum. Occultation techniques use the properties of ozone absorption in the visible and ultraviolet wavelengths whereas the limb emission and limb scattering techniques use a knowledge of ozone absorption in the infrared and microwave portions of the electromagnetic spectrum.

In the BUV technique, the solar flux can be measured directly, although to reduce the flux to manageable levels for the observing instrument, a "diffuser plate" is typically deployed when the instrument looks at the sun (it is retracted for Earth


Backscatter Ultraviolet

Occulation

Limb Emission

Limb Scattering

Figure 2 Schematic diagram showing the spacecraft and atmosphere geometry for the four common methods of acquiring trace gas measurements from satellite instruments. See ftp site for color image.

viewing). In applying the BUV technique, it is also helpful to have the measurement time close to local noon, so that the optical path lengths through the atmosphere are close to their potential minimum for the corresponding surface location. In its simplest form, the instrument looks straight down (nadir viewing) to make measurements below the satellite track (Fig. 2*a*). Maps of ozone column can be created by either scanning the instrument's field of view across the orbital track or using some sort of imaging detector so that observations are made corresponding to different ground locations.

The first use of this technique was on the BUV instrument aboard the *Nimbus 4* satellite launched in 1970. This instrument, which was a purely nadir-viewing one, obtained data for several years. The data gave an excellent picture of both the latitudinal and seasonal nature of global column ozone distributions. These data are still of scientific interest; recently they were reexamined to help characterize Antarctic ozone amounts in the early 1970s and show that there was no evidence for significant depletion of Antarctic ozone in the springtime then.

The most significant application of the BUV techniques has been in the total ozone mapping spectrometer (TOMS) and solar backscatter ultraviolet (SBUV) series of instruments. The TOMS instruments use measurements at six wavelengths to measure ozone. For the first two TOMS instruments (one on the *Nimbus 7* satellite that obtained data from October 1978 to May 1993 and one aboard a Russian *Meteor-3* satellite that obtained data from September 1991 to December 1994), the wavelengths used were 312.5, 317.5, 331.2, 339.8, 360, and 380 nm. The latter two are essentially unaffected by the presence of ozone and were used to provide information on surface UV reflectivity. The TOMS instruments were also shown to have information about concentrations of sulfur dioxide, especially during times of enhancement following large volcanic eruptions, aerosols (both tropospheric aerosols including those from biomass burning and volcanic dust, among other types and stratospheric aerosols following large volcanic eruptions), and surface UV radiative flux.

The ground resolution of these instruments is approximately 50 × 50 km at nadir (resolution is degraded as the instrument field of view scans sideways). The orbit of the *Nimbus 7* satellite (sun synchronous, polar orbiting) was excellent for TOMS measurements, while that of the *Meteor-3* satellite was less so since it was not sun synchronous. Roughly half the time the *Meteor-3* orbit led to TOMS observations at local times sufficiently far away from noon that results must be used at great care if at all.

The newer TOMS instruments, which operated aboard the Japanese ADEOS spacecraft (Aug. 1996–May 1997) and NASA's *Earth Probe* (EP) satellite, use a slightly different wavelength set from the previous TOMS instruments. For these new instruments, there is an additional channel to help in the measurement of ozone at high solar zenith angles, as well as a channel to monitor the behavior of the TOMS instrument. The *Earth Probe* TOMS instrument was originally launched into a relatively low (~500 km) orbit to provide for better ground resolution (~26 × 26 km at nadir) than that of the *ADEOS* TOMS instrument (42 × 42 km), which flew aboard the higher orbiting *ADEOS* spacecraft. At the lower altitude, TOMS could not obtain full daily maps over the entire sunlit Earth, however, as there were interorbit gaps equatorward of approximately 60° latitude. Following the *ADEOS* failure, the *EP* satellite was boosted into a higher orbit (~750 km) to allow for near global spatial coverage.

The SBUV series of instruments includes the original SBUV instrument, which flew aboard the *Nimbus 7* satellite, and updated instruments (SBUV/2) that flew aboard several of the operational meteorological satellites of the American National Oceanic and Atmospheric Administration (NOAA) on the *NOAA-9*, *NOAA-11*, and *NOAA-14* satellites, to date. The SBUV instruments, which also have a capability to determine ozone vertical profile (see Section 4 of this chapter) do not have any cross-track scanning capability and thus do not obtain contiguous daily maps as do the TOMS instruments; they simply obtain data along the daytime subsatellite tracks (the nature of the BUV technique, which requires the presence of sunlight, precludes nighttime data). Long-term calibration information for the SBUV instruments was provided by the Shuttle-borne SBUV (SSBUV) instrument, which flew eight times on the Space Shuttle from 1989 to 1996.

The TOMS and SBUV series of instrument have provided an invaluable database on the total ozone distribution of Earth's atmosphere and its many variations. In Figure 3, a two-dimensional representation (latitude–time) of total ozone distributions derived from the TOMS satellite is shown. Key elements of the total ozone distribution are evident—low total ozone with little seasonal variation in the tropics, highest total ozone values in late winter at high northern latitudes, and lowest total ozone values associated with Antarctic ozone depletion at high southern latitudes during the Austral spring. The long-term changes in the global amount of total ozone determined from *Nimbus-7* TOMS is shown in Figure 4*a*. These data have gone through extensive calibration procedures and comparisons with ground-based Dobson stations to ensure the greatest possible accuracy. The greatest changes have occurred over the Antarctic continent and is seasonal in extent. On the other hand, no statistically significant ozone depletion has been noted in the tropics (see

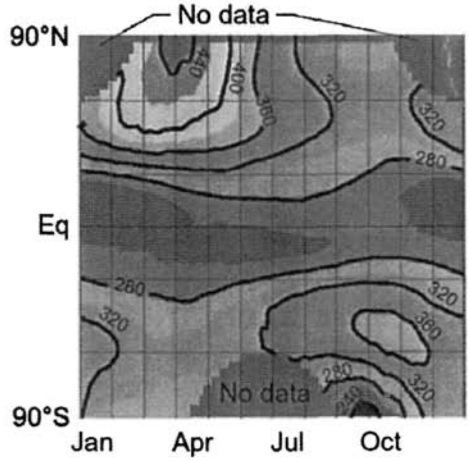


Figure 3 (see color insert) Two-dimensional (latitude/season) representation of total column ozone as measured by TOMS for the period 1978 to 1993. See ftp site for color image.

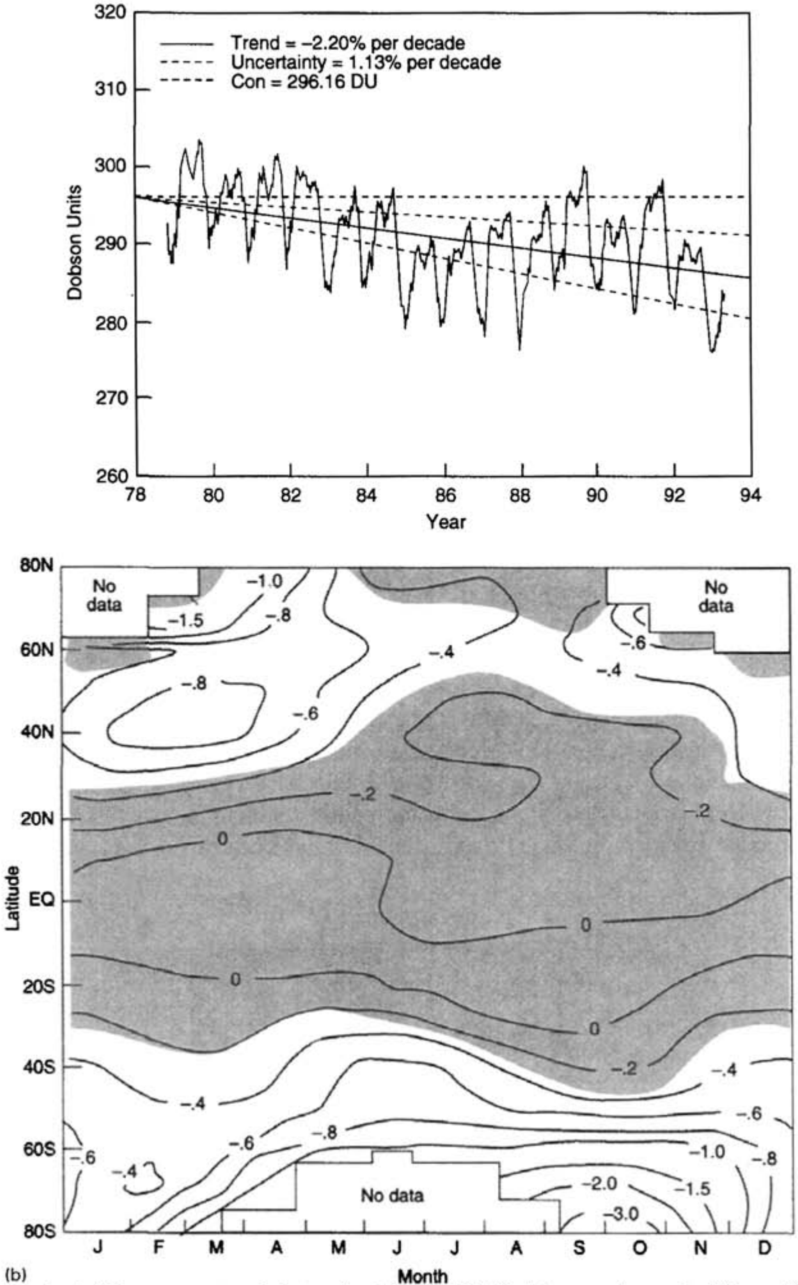


Figure 4 (*a*) Long-term trend determined from TOMS data over the period from 1978 to 1994 after correcting for seasonal cycles, the 11-year solar cycle, and the quasi-biennial oscillation; (*b*) month vs. latitude analysis of ozone trend plotted in contours of percentage loss per year. Shaded areas indicate areas where trend is not definitive. See ftp site for color image.

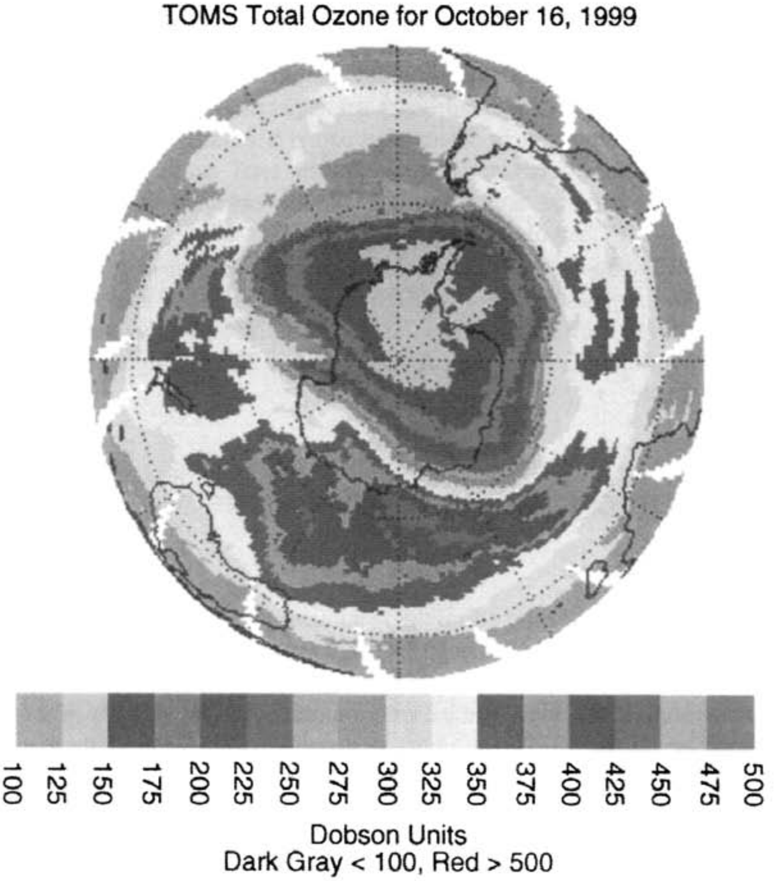

TOMS Total Ozone for October 16, 1999
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
475
500
Dobson Units
Dark Gray < 100, Red > 500

Figure 5 (see color insert) Map of total column ozone over the Antarctic as determined from TOMS October 16, 1999. See ftp site for color image.

Fig. 4*b*). An example of the daily mapping capability of the TOMS instruments is shown in Figure 5, in which a contour map of total ozone distributions during the height of the Antarctic ozone depletion season is presented. The regions of ozone depletion, surrounded by the "collar" region of higher ozone amounts, are clear.

Several limitations of TOMS and SBUV data are worth noting. In particular the UV wavelengths used do not penetrate clouds and are less sensitive to ozone in the lowest few kilometers of the troposphere. Thus, variations in cloudiness or the nature of the tropospheric ozone profile can affect the retrieved ozone column amounts. Improved understanding of these limitations has been an important goal of much recent research.

Another measurement technique used to obtain measurement of a "total ozone-like quantity" is infrared emission. The TIROS operational vertical sounder (TOVS) instruments measure infrared radiation at 9.6 μm (corresponding to one of the fundamental vibrational modes of ozone). TOVS is mainly sensitive to lower stratospheric ozone and as such does not provide a true total column measurement. However, the correlations between total ozone and lower stratospheric ozone are well established (since it is the lower stratosphere where most ozone is found), so the TOVS product has many of the same characteristics as TOMS total ozone. Since the TOVS measurement uses infrared emission, data can be obtained in regions without sunlight, such as high latitudes during polar night.

A space-based version of the DOAS technique has been implemented aboard the European Space Agency's *ERS-2* satellite using the Global Ozone Monitoring Experiment (GOME). *ERS-2* was launched in April 1995. The GOME instrument uses a broader wavelength range than do the TOMS or SBUV instruments, including longer wavelengths.

4 OZONE VERTICAL PROFILE MEASUREMENTS

Ground-Based Ozone

The first ground-based ozone profiling technique to be used was the Umkehr method, in which the solar zenith angle dependence of Dobson-type measurements is used to determine the vertical profile. This technique obtains data at ~5 km vertical resolution. It provides little information on the lowest ~20 km of the atmosphere, however. In the middle and upper stratosphere, the Umkehr data record has provided an important source of information, especially on long-term ozone trends.

Another ground-based technique for obtaining information on the ozone vertical profile is the use of microwave emission. Since ozone molecules occupy a broad range of rotational states at atmospheric temperatures, there are numerous transitions that will take place for which emission-based remote sensing may be used. Information on the vertical profile comes from the shape of the observed emission lines because of the pressure-broadened nature of the emission lines—emission from ozone in the upper stratosphere will take place near the center of the spectral band, while that from lower down will occur in the wings of the line. Although the vertical resolution of this technique is somewhat limited, it can provide valuable information, especially when measured together with distributions of ozone-destroying free radicals such as ClO or HO_2.

Another technique for ground-based measurement of ozone is lidar. In the lidar technique, a pulse of laser light is sent up from the ground, and the scattered signal that returns to the ground provides information on the composition of the air mass being observed, while the time delay between the laser shot and the return signal is used to provide altitude information. Since the air mass being sampled will interact with laser light by other processes besides ozone absorption (such as molecular Rayleigh scattering, as well as aerosol scattering), lidar systems typically employ two laser wavelengths, one of which is more strongly absorbed by ozone than the other. The wavelength dependence of the aerosol and Rayleigh scattering is typically much less than that of ozone (and relatively well understood) allowing for retrieval of ozone amounts. The wavelengths used will depend to some extent on the altitude range at which ozone measurements are desired. For stratospheric measurements, where larger ozone abundances are typically observed than in the troposphere, wavelengths with a smaller absorption cross section are needed than for tropospheric measurements. Typical wavelength pairs used are 308 and 355 nm for stratospheric ozone lidar and 288 and 299 nm for tropospheric lidar. Lidar can also be implemented from aircraft, in both upward and downward looking configurations.

In Situ Measurement Techniques

The primary in situ measurement technique used for determination of the ozone vertical profile is that used on ozonesondes. In one standard implementation, an iodine/iodide redox concentration cell is used. An electric current is generated when air containing ozone is pumped into the cell, with the amount of current being related to the partial pressure of ozone in the air mass being sampled. This technique is capable of providing excellent vertical resolution, and is unparalleled at determining the existence of "tongues" or "laminae" of air masses with ozone contents that differ from those of their surroundings (see examples in Chapter 1). Because of the limitations of the balloon on which they are flown, ozonesondes rarely rise above $\sim$30 km. Ozonesonde measurements are usually only made from a fairly limited set of observing stations, and except during certain intensive field campaigns, are typically made at most weekly. Ozonesondes can be flown at various locations and do not require the presence of sunlight. Ozonesondes from the South Pole have provided an important part of our knowledge of the vertical distribution of ozone over Antarctica, for instance, especially on its seasonal variation in springtime. In Figure 6, a plot of ozone vertical profiles over Antarctic measured before and during the presence of the ozone hole are shown. The ozonesondes provide clear evidence for the near total absence of ozone in the 12 to 22 km altitude range during the period of ozone depletion.

The measurement technique used by ozonesondes requires very careful emphasis on calibration and intercomparison. In some cases due to uncertainties about operations, ozonesonde profiles are "normalized to Dobson" so that the observed profile is modified based on a scaling of the calculated integrated ozone column to that observed at a co-located or nearby Dobson station. There are also several different types of ozonesondes, whose operational characteristics differ slightly. In spite of these uncertainties, the ozonesonde record has been critical in the assessment of ozone trends in the lower stratosphere, a region ($\sim$15 to 20 km) that is very difficult to observe at high accuracy using space-based instruments.

Other in situ techniques for ozone measurement also exist. One used extensively aboard research aircraft is a spectroscopic technique in which the absorption of air at UV wavelengths is accurately determined. This is a very accurate technique, as the spectral information is well known and there is little opportunity for interference because of the significantly smaller abundance of most potential contaminants. This technique has been used aboard NASA's *ER-2* aircraft in its flights in the lower

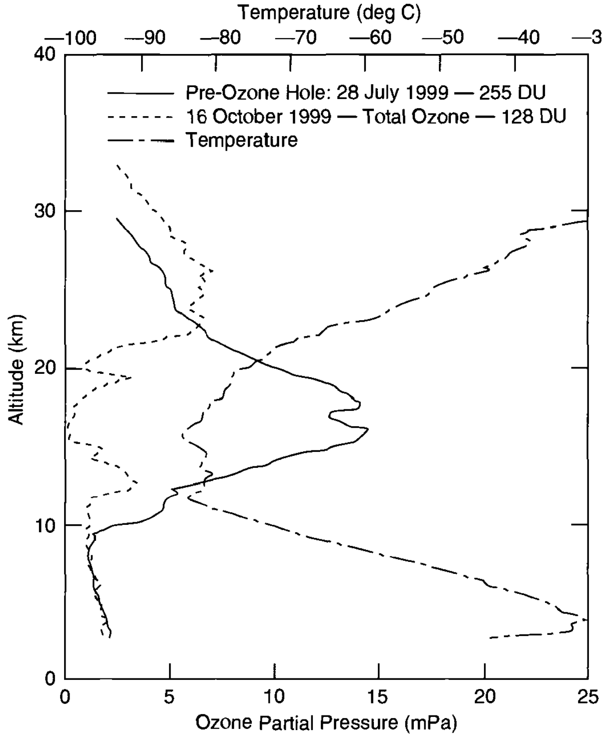


Figure 6 (see color insert) Plot of vertical profile of ozone (blue and red lines) over the South Pole as measured from ozonesondes during austral winter (July 28) and spring (October 16), 1999; temperature profile for October 16 is also shown (green line). See ftp site for color image.

stratosphere and upper troposphere, for instance. A chemiluminescent system has also been used.

Space-Based Remote Sensing

The ozone profiling technique with the greatest heritage uses the BUV technique. By using several wavelengths shorter than those used for total column measurements, information on the ozone distribution in the middle and upper stratosphere can be obtained. This technique makes use of the fact that with decreasing wavelengths light is absorbed at higher altitudes in the atmosphere because of the corresponding increased absorption cross section (see Fig. 1). The vertical resolution of this technique is quite broad, however, some 7 km in the middle and upper stratosphere and close to 10 km below the peak in the ozone layer (typically 20 to 25 km depending on latitude). Through its use on the SBUV series of instruments, this technique has provided extensive information on the latitude and seasonal dependence of ozone's vertical structure in the middle and upper stratosphere. One strong conclusion to come from this is the clear demonstration of statistically significant ozone losses near 40 km, especially at high latitudes.

The other space-based technique with the longest heritage uses the absorption of radiation at occultation (the rising and setting of the sun with each orbit). The occultation technique (see Fig. 2*b*) has several notable advantages. First, because it involves an along-path length against a rising or setting sun, signals are strong. Second, the technique is inherently "self-calibrating" in that for each measurement an observation is made at the top of the atmosphere and in darkness, so any changes in the performance of the instrument can be determined, at least to first order. Third, the technique has the capability for relatively high (~1 km) vertical resolution to be implemented fairly easily, although in the lower stratosphere, where ozone mixing ratios vary rapidly with altitude and one must view through the peak in the ozone layer, vertical resolution may be degraded.

The main limitation of the solar occultation technique is in spatial coverage. Since the sun rises and sets only once per orbit, at most two latitudes of data per orbit are obtained. These latitudes will change fairly rapidly with time. Depending on the orbital inclination and the orientation of the spacecraft orbit, a complex pattern of observations versus time is obtained; this may complicate the determination of seasonal distributions. If the spacecraft is in a polar sun-synchronous orbit, sunrises and sunsets are only obtained at high latitudes. Another disadvantage of this technique is high sensitivity to aerosol loading. When stratospheric aerosol loading is high, such as following a major volcanic eruption (e.g., Mt. Pinatubo in 1991), the high aerosol abundance provides a great deal of extinction that can complicate the retrieval of ozone amounts. This technique cannot penetrate clouds, and therefore can provide little information on the tropics below the tropopause, as high-level clouds are typically present near the tropical tropopause. An additional limitation of this method is that it actually observes number density as a function of altitude; most atmospheric scientists work with mixing ratios as a function of pressure. Unless temperature is measured together with ozone amounts, the conversion from the observed to desired quantities requires externally supplied (and noncollocated) meteorological information.

The occultation technique has been implemented using both UV/visible/near-infrared and purely infrared wavelengths. The Stratospheric Aerosol and Gas Experiment (SAGE) series of instruments has been the longest term implementation of this technique. SAGE I, which flew aboard the *AEM-2* satellite, obtained data from 1979 to 1981, while the SAGE II instrument, which flies aboard the *Earth Radiation Budget Satellite* (ERBS), has obtained data since its launch in late 1984. Both instruments flew in a 57° inclination orbit; the latitude and time dependence of the solar occultations for SAGE II is shown in Figure 7. It is seen that it typically takes ~3 weeks for the occultations to scan the full range of latitudes. The nonuniform nature of the coverage is quite obvious—in some months, some latitudes are not sampled at all. High latitudes are only sampled occasionally. The SAGE instruments, which make measurements at a total of 4 (SAGE I) and 7 (SAGE II) chan-

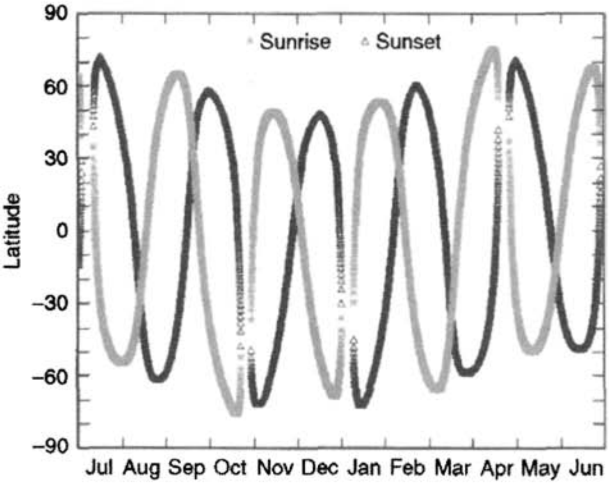


Figure 7 Spatial coverage of solar occultations from the SAGE II instrument as a function of season. Sunrise and sunset occultations are indicated with different symbols. There is excellent year-to-year repeatability of the times and locations of the occultations. See ftp site for color image.

nels, also measured both nitrogen dioxide (NO_2) and stratospheric aerosols; the SAGE II instrument also measures water vapor (H_2O). The main ozone measurement channel uses the Chappuis bands at 600 nm, but the measurement must account for the presence of aerosols, which also contribute to the 600-nm extinction.

Along with ozonesondes, the SAGE II instruments provide a critical data set for the long-term variation of stratospheric ozone. In Figure 8, a zonally and seasonally averaged representation of the long-term trend in stratospheric ozone as a function of altitude is shown. Clear evidence for both upper stratospheric loss in ozone amounts (largest at high latitudes) and lower stratospheric amounts is observed. An accurate characterization of this change as a function of altitude remains an important research area.

The occultation technique using very similar wavelengths was also implemented using the polar ozone aerosol monitor (POAM-2) instrument, which flew aboard the French *SPOT-3* satellite and obtained data for 3 years from late 1993 to late 1996 (Bevilacqua et al., 1997). Since the SPOT-3 satellite was in a polar sun-synchronous orbit, all occultations were at middle and high latitudes, and the POAM data provided an important picture of how high-latitude ozone profiles vary over the course of the year. An example of this is in Figure 9 in which the ozone number density at 20 km is shown inside the Antarctic polar vortex from May through December for the 3 years of POAM-2 observation. The slow decline of ozone in the winter, the rapid falloff in September, and the slow recovery in October and November are all clearly evident. Interannual variability in these 3 years is quite small.

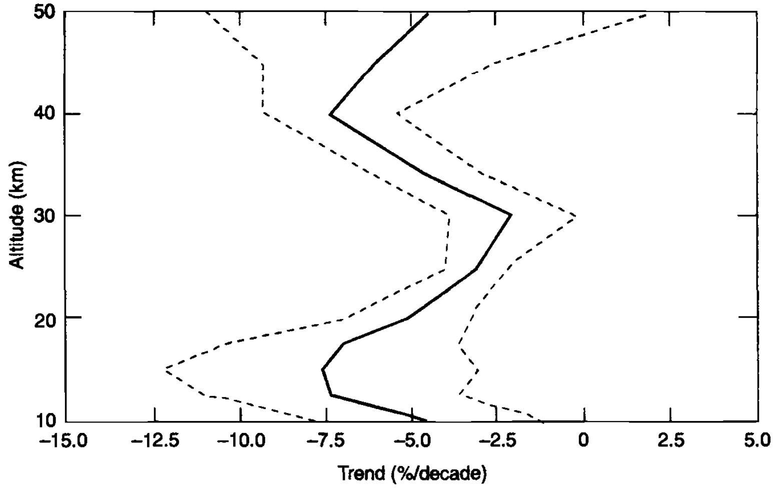


Figure 8 Two-dimensional (latitude/altitude) representations of the long-term trends in stratospheric ozone distributions determined from the SAGE I and SAGE II instruments.

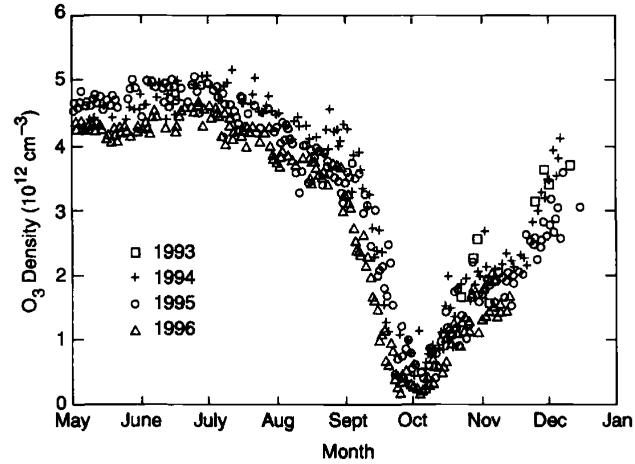


Figure 9 Plot showing the evolution of ozone amounts in the lower stratosphere inside the polar vortex during the fall–winter of the 3 years for which POAM-2 obtained data. Different symbols are used for each year.

The occultation technique has also been applied using infrared wavelengths. The atmospheric trace molecule spectroscopy (ATMOS) instrument, a Fourier transform spectrometer, flew four times aboard the Space Shuttle (1985, 1992, 1993, 1994). ATMOS is notable for its measurements of a very broad range of trace constituents based on its very high spectral resolution ($\sim$0.01 cm^{-1}). Its measurement of ozone (and many other constituents) helped serve as an important validation tool for measurements made from NASA's upper atmosphere research satellite (UARS), launched in September 1991. The Halogen Occultation Experiment (HALOE), a combination broadband radiometer and gas cell correlation radiometer has measured ozone (and several other trace gases) during the more than 6 years of UARS operations. Since UARS is in a 57° inclination orbit, the spatial coverage of HALOE is very similar in character to that of SAGE. Most recently, the improved limb atmospheric spectrometer (ILAS) instrument flew aboard Japan's *ADEOS* satellite and obtained data for nearly a year from August 1996 to June, 1997. Since *ADEOS* was in a polar sun-synchronous orbit ILAS data were restricted to high latitudes.

Emission technology has also been used for measurement of atmospheric ozone. These involve looking at the limb of the atmospheres and measuring the thermal emission from ozone (or some other species). Such measurements do not require the presence of a light source and can thus be made over a complete orbit (both day and night). When made from a polar sun-synchronous orbit, they are made at roughly the same time every day (typically once in the daytime and once in the nighttime), which facilitates studies of diurnal variation. On an inclined orbiter, such as UARS, the measurement time will vary and therefore intermix diurnal and seasonal dependence. As typically implemented, limb observations have vertical resolution of $\sim$3 km, although the exact amount can vary higher or lower depending on the measuring instrument. Since limb emission is a thermally driven process, simultaneous measurement of temperatures to high accuracy is required. Infrared emission observations of ozone involve the measurement of the emission from the relatively small population of vibrationally excited molecules that exist in thermal equilibrium with the large majority of ground-state molecules, while in microwave emission, it is emission from rotationally excited molecules that is measured. In some cases (especially daytime in the mesosphere), nonthermal process may populate the excited vibrational states of ozone, and these must be accounted for in determining ozone concentrations from infrared emission measurements. Such nonthermal distribution of population states typically does not occur in the microwave, where the smaller energy quantum allows for thermal equilibrium to be maintained.

Implementations of emission techniques for ozone include the limb infrared monitor of the stratosphere (LIMS) instrument, which flew aboard the *Nimbus 7* satellite and obtained data from October 1978 to May 1979. Later implementations include two instruments aboard UARS—the improved stratospheric and mesospheric sounder (ISAMS) and the cryogenic limb array etalon spectrometer (CLAES) instruments. The CLAES instrument, which used a solid cryogen cooler, worked for approximately 20 months until the depletion of the cryogen. The ISAMS instrument on UARS provided 7 months of data.

These instruments have provided significant information on the behavior of ozone at a given pressure level, especially the relationship between ozone amounts and the meteorology of the stratosphere. An example is shown in Figure 10, in which the variation of ozone in the lower stratosphere (~30 mbar) during a major stratospheric warming in the Northern Hemisphere in the winter of 1979 is shown using LIMS data. In this figure, the polar vortex region of high ozone usually found over the pole (e.g., February 6, 1979) is split into two, as shown in the later analyses for February 16 and February 23. The March 1 panel shows that the ozone has filled in the low region subsequent to a stratospheric warming that had occurred during this time.

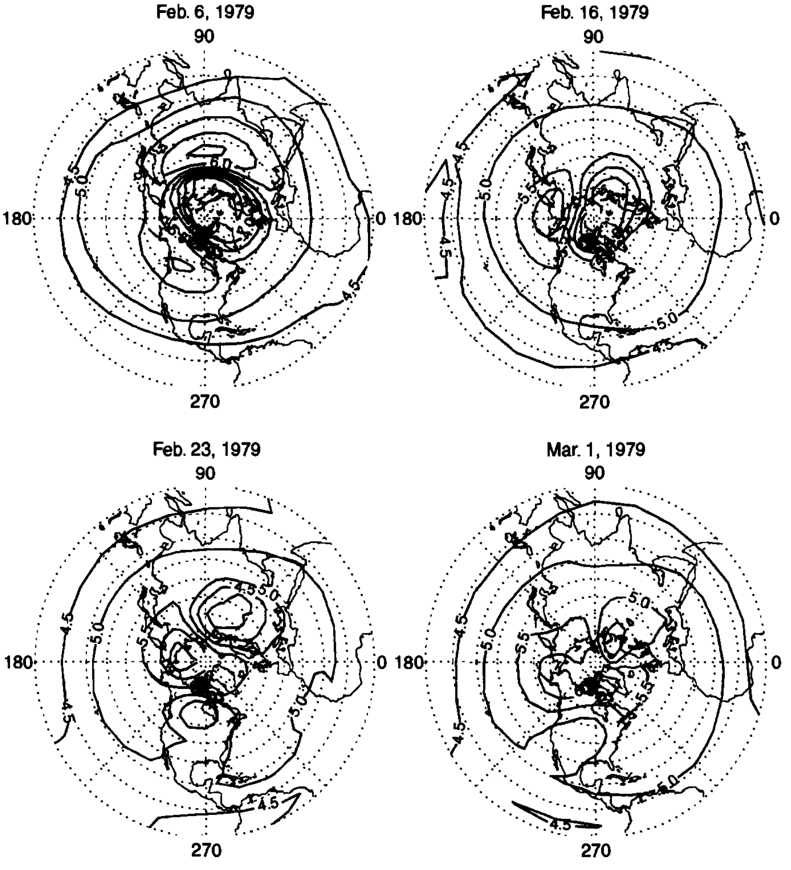


Figure 10 Contour maps showing the evolution of lower stratospheric ozone obtained from LIMS during the vicinity of a major stratospheric warming.

As typically implemented, limb emission observations are made only at longitude as the satellite moves along its orbital track; no cross-track scanning (as is done in TOMS) or imaging is carried out to "fill in" the interorbit gaps. One instrument that is an exception to this is the cryogenic infrared spectrometers and telescopes for the atmosphere (CRISTA) instrument, which was deployed from the Space Shuttle in 1994 and 1997. The CRISTA instrument has three telescopes and infrared detectors viewing 18° apart from each other, thus obtaining higher horizontal resolution than any other atmospheric chemistry profiling instrument.

Several other space-based techniques have been used for measurement of atmospheric ozone. In one, a variant of the infrared emission technique, emission is measured not from vibrationally excited ozone (as was done on LIMS, CLAES, and ISAMS), but from electronically excited molecular oxygen [$O_2(^1\Delta)$] produced following ultraviolet photolysis of ozone. This technique, which was used on the *Solar Mesosphere Explorer* (*SME*) satellite in the early 1980s, is applicable mainly in the thermosphere and mesosphere; at lower attitudes, the $O_2(^1\Delta)$ is quenched so rapidly that there is insufficient signal for detection. UV limb scattering was also implemented on *SME* but was limited to observation in the mesosphere and topmost part of the stratosphere. Since this is a scattering technique, it provides the possibility for good spatial coverage and also can potentially allow for good vertical resolution. More recently, this technique was tested with a pair of instruments (the Shuttle Ozone Limb Sounder Experiment and the Limb Ozone Retrieval Experiment) that flew aboard the Space Shuttle in 1997. Finally, the technique of stellar occultation (in which stars are the source of the radiation) has been tested using the ultraviolet imaging and spectral imagers (UVISI) instrument that flies aboard the U.S. Department of Defense's *Midcourse Space Experiment* (*MSX*) spacecraft. The stellar occultation technique has the potential to overcome the spatial limitation of the solar occultation technique because of the existence of many stars that can serve as a source in a given orbit. The technique also has potential for high vertical resolution; the main complication is the need to overcome the much reduced photon flux for a star as opposed to that of the sun.

5 FUTURE MEASUREMENTS

There will be significant activity in the next few years in ozone measurements, especially in the implementation of several space-based measurement systems. These include additional copies of existing instruments, next-generation instruments based on current techniques and significantly new instruments. The first decade of the twenty-first century will also see to at least two space platforms devoted to studying the composition of the atmosphere and providing new insight into both the distribution of ozone and the chemical processes responsible for its measured abundance.

TOMS/SBUV

An updated version of the TOMS instrument, which will probably have a much wider spectral range than TOMS, as well as a large number of channels, is an instrument to be provided by the Dutch government aboard the *Aura* spacecraft of NASA's Earth Observing System, scheduled for launch in 2004. This instrument, currently known as OMI (ozone monitoring instrument), will also make use of spatial imaging through the use of a detector array, eliminating the need for cross-track scanning and the use of a photomultiplier tube for detection. In the longer-term, a total ozone mapping instrument is scheduled for inclusion in the National Polar Orbiting Environmental Satellite System (NPOESS) being developed by the United States. The exact nature of this instrument has not yet been determined, although it will likely have many of the observing goals of TOMS. The first NPOESS spacecraft will fly no earlier than 2010. Additional SBUV instruments are planned for several NOAA operational meteorological spacecraft through 2004.

SAGE III

An improved version of the SAGE instrument was launched aboard a Russian *Meteor-3M* satellite in late 2001. The SAGE III instrument uses the occultation technique pioneered with the earlier SAGE instruments but has significant advances, including a wider spectral range (from 290 to 1500 nm), and the use of relatively high resolution spectral information within several channels to help provide much improved detection of ozone and other trace species, as well as separation of ozone and aerosol extinction. SAGE III also makes collocated measurements of temperature and pressure to facilitate conversion of profiles from number density versus altitude to mixing ratio versus pressure. SAGE III also has a lunar occultation capability that will remove some of the spatial limitations associated with solar occultations. This is especially important for the *Meteor-3M* SAGE, which will be in a polar sun-synchronous orbit in which solar occultations are restricted to high latitude. The lunar occultations cover a much broader range of latitudes.

Platforms Dedicated to Atmospheric Chemistry

ENVISAT. ENVISAT, a program of the European Space Agency, was launched in 2002 and has three significant instruments for measurements of atmospheric ozone and related trace constituents. These include the scanning imaging absorption spectrometer for atmospheric chemistry (SCIAMACHY), the Michelsen interferometer for passive atmospheric sounding (MIPAS), and the global ozone monitoring through stellar occultations (GOMOS) instruments. SCIAMACHY is an enhanced version of the GOME instrument flying aboard *ERS-2*; it will utilize the DOAS technique as did GOME but will also have limb and occultation modes and will have infrared wavelengths that GOME did not have. MIPAS is a high-resolution infrared emission instrument, and GOMOS will use the stellar occultation technique to determine ozone profiles (Burrows, 1999).

EOS Aura. The EOS *Aura* spacecraft planned for launch in 2004 will have three ozone-measuring instruments in addition to the OMI. These are the high-resolution dynamics limb sounder (HIRDLS), the microwave limb sounder (MLS), and the troposphere emission spectrometer (TES). HIRDLS will use the technique of infrared emission to determine vertical profiles of ozone, temperature, aerosols, and a host of trace constituents. Unlike most previous infrared emission measurements, however, HIRDLS will have high vertical (~1 km) and horizontal (~5°) resolution; the latter comes from its making several scans away from the nadir as the spacecraft moves in its orbit. MLS is a significant improvement over the UARS MLS, with special attention being given to improving measurements in the upper troposphere and lower stratosphere, as well as the measurement of many trace constituents, such as OH, BrO, and N_2O not measured with the UARS MLS. TES is designed to measure ozone and its precursors in the troposphere using high-resolution infrared spectral measurements. It will use both nadir and limb viewing geometries to do this.

REFERENCES

Albritton, D. L., and R. T. Watson (Eds.), *Scientific Assessment of Ozone Depletion: 1991*, Global Ozone Research and Monitoring Project, Report No. 25, World Meteorological Organization, Geneva, 1991.

Albritton, D. L., R. T. Watson, and J. J. Aucamp (Eds.), *Scientific Assessment of Ozone Depletion: 1994*, Global Ozone Research and Monitoring Project, Report No. 37, World Meteorological Organization, Geneva, 1994.

Albritton, D. L., J. J. Aucamp, G. Megie and R. T. Watson (Eds.), *Scientific Assessment of Ozone Depletion: 1998*, Global Ozone Research and Monitoring Project, Report No. 44, World Meteorological Organization, Geneva, 1998.

Bevilacqua, R. M., et al., Use of POAM II data in the investigation of the Antarctic ozone hole, *J. Geophys. Res.*, *102*, 23643–23657, 1997.

Burrows, J. P., Current and future passive remote sensing techniques used to determine atmospheric constituents, in A. F. Bouwman (Ed.), *Approaches to Scaling of Trace Gas Fluxes in Ecosystems*, Elsevier, Amsterdam, 1999, pp. 317–347

CHAPTER 22

AEROSOL PROCESSES IN THE STRATOSPHERE

MARIO J. MOLINA

1 INTRODUCTION

It is now well established that chemical reactions involving aerosol particles play a key role in stratospheric ozone depletion.[1–3] Some of these reactions take place on the surface of solid particles, while others occur inside liquid particles; both are commonly referred to as heterogeneous processes because they involve both the gas and the condensed phase.

The aerosol layer is located in the lower stratosphere and consists predominantly of aqueous sulfuric acid droplets, commonly labeled SSAs (sulfate stratospheric aerosols). At mid and low latitudes their concentration is 70 to 80% by weight H_2SO_4, corresponding to mole fractions between ~0.3 and 0.5; at high latitudes and in the winter and spring months the SSAs may grow significantly in size, becoming polar stratospheric clouds (PSCs). As they cool, they absorb water vapor and also nitric acid vapor but remain in liquid form, becoming type Ia PSCs. If they freeze, they are labeled type Ib PSCs, and at sufficiently low temperatures they become ice crystals (type II PSCs). The mechanism of conversion between the various stratospheric aerosol types is not well established and is discussed in Section 5 below. Typical particle sizes are ~0.1, 1, and 10 μm diameter for SSAs and PSCs type I and type II, respectively; their abundance is ~1 to 10 particles/cm^3.

2 CHEMICAL REACTIONS ON STRATOSPHERIC AEROSOLS

Stratospheric trace species include sources, free radicals (species with an unpaired electron), and temporary reservoirs. Source species are produced at Earth's surface

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

and are stable enough to eventually reach the stratosphere. Temporary reservoirs are generated in the stratosphere, but they are ultimately transported downwards into the troposphere, where they are rapidly removed by rainout or washout. Both sets of species decompose in the stratosphere producing free radicals—either by photolysis or by reaction with another radical. Free radicals can participate in ozone destruction cycles but can also react with each other to produce stable reservoirs. Thus photochemistry is a source of radicals, while gas-phase reactions interconvert radicals into different forms or else destroy free radicals by producing stable reservoirs. Practically all gas-phase chemical reactions involve free radicals; reactions in the gas phase between nonradical (saturated) species are usually too slow to matter at atmospheric temperatures. However, aerosols provide a pathway for such reactions to take place.

The two most important sets of heterogeneous reactions in the stratosphere are chlorine activation and nitrogen deactivation. Chlorine activation reactions transform temporary chlorine reservoirs to a form that is photolytically active; the most important ones are the following:

$$HCl + ClONO_2 \rightarrow HNO_3 + Cl_2 \tag{1}$$

$$HCl + HOCl \rightarrow H_2O + Cl_2 \tag{2}$$

The species Cl_2 is photolytically very active; it readily absorbs near-ultraviolet (UV) and visible light to produce free Cl atoms, which in turn react rapidly with ozone:

$$Cl_2 + h\nu \rightarrow 2Cl \tag{3}$$

$$Cl + O_3 \rightarrow ClO + O_2 \tag{4}$$

The most important nitrogen deactivation reaction is

$$N_2O_5 + H_2O \rightarrow 2HNO_3 \tag{5}$$

The N_2O_5 species is produced in turn from nitrogen oxides:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \tag{6}$$

$$NO_2 + NO_3 \rightarrow N_2O_5 \tag{7}$$

The net effect of reactions (5), (6), and (7) is to convert active forms of nitrogen (NO_x) to the relatively stable temporary reservoir HNO_3, a species which, however, may regenerate radicals by solar photolysis:

$$HNO_3 + h\nu \rightarrow OH + NO_2 \tag{8}$$

On the other hand, a significant fraction of the gas-phase HNO_3 is incorporated at low temperatures into PSCs, where it is further stabilized and protected against solar photolysis. Yet another effect of this HNO_3 scavenging process is denitrification: if some fraction of the PSC particles grow sufficiently large, they may settle to lower altitudes, thus removing the NO_x source more permanently.

The combined effect of chlorine activation and nitrogen deactivation is accelerated ozone depletion: Chlorine free radicals destroy ozone more rapidly in the absence of nitrogen radicals because the two sets of radicals tend to react with each other to produce temporary reservoirs, slowing ozone depletion. The most important example of this process is the formation of $ClONO_2$ (chlorine nitrate) through the following radical recombination reaction:

$$ClO + NO_2 + M \rightarrow ClONO_2 + M \tag{9}$$

Reaction (9) is termolecular, and M is the "third" body (mostly N_2 or O_2) required to stabilize the newly formed bond. In contrast to other radical recombination products such as ClNO (nitrosyl chloride), CONO (chlorine nitrite), and $ClNO_2$ (nitryl chloride), the species $ClONO_2$ is relatively stable toward solar photolysis.[4] However, in the presence of HCl it rapidly regenerates the chlorine radicals by the chlorine activation reaction discussed above [reaction (1)].

3 HETEROGENEOUS REACTION RATES AND MECHANISMS

Rate constants for heterogeneous reaction rates are commonly expressed in terms of a reaction probability γ, which is the probability per collision of a gas-phase reactant molecule with the aerosol surface that chemical reaction will occur. For reaction on a solid aerosol surface of a species with a mean gas-phase concentration [C] molecule/cm^3), the overall rate may be approximated by the following expression, which is based on the "resistance" model[3]:

$$-d[\mathrm{C}]/dt = k_t S[\mathrm{C}] \tag{10}$$

$$1/k_t = 1/k_{\mathrm{diff}} + 1/k_{\mathrm{coll}} = R_p/D_g + 4/[\langle v\rangle\gamma/(1-\gamma/2)] \tag{11}$$

where t is time (s), k_t is the effective overall first-order rate constant (cm/s) for surface reaction, S is the aerosol surface area per unit volume (cm^2/cm^3), $1/k_{\mathrm{diff}}$ is the resistance associated with gas-phase diffusion, $1/k_{\mathrm{coll}}$ is the resistance associated with molecular collisions with the particle surface, D_g is the gas diffusion coefficient (cm^2/s), R_p is the average particle radius (cm), and $\langle v\rangle$ is the mean molecular speed (cm/s) of the gas-phase reactant. For values of γ of less than ~0.2, the expression $\gamma/(1-\gamma/2)$ may be approximated by γ. For certain conditions Eq. (11) can be further simplified, e.g., at low pressures and for small particles (large D_g and small R_p), the effect of gas-phase diffusion may be neglected ($1/k_{\mathrm{diff}} \approx 0$); etc. On the other hand, for small particlcs with sizes approaching the gas-phase mean free path, additional correction factors are needed.[5] For reaction on liquid particles, liquid-phase diffusion also needs to be taken into account, leading to additional resistance terms in Eq. (11).[3]

The rate of reaction (5) on sulfuric acid solutions is nearly independent of the acid concentration,[3,4] hence the reaction occurs readily at low and midlatitudes. In contrast, the rates of reactions (1) and (2) are negligible at those latitudes: The reaction mechanism involves as a first step incorporation of HCl vapor into the condensed phase, and HC1 is practically insoluble in concentrated H_2SO_4 solutions. On the hand, as the sulfuric acid particles cool and become more dilute at higher latitudes, the solubility of HCl increases sharply, and the reaction probability increases accordingly, reaching values larger than ~0.1 for temperatures below 200 K.

As the particles freeze at high latitudes, the reaction probabilities may be strongly affected: hydrolysis of N_2O_5 [reaction (5)] becomes very slow, while reactions (1) and (2) occur very efficiently on ice surfaces, requiring only a few collisions of the reactant $ClONO_2$ or HOCl with the particles exposed to HCl vapor.[4] Thus nitrogen deactivation [reaction (5)] occurs predominantly at mid and low latitudes and also at high latitudes as long as the aerosol particles remain liquid. In contrast, chlorine activation [reactions (1) and (2)] occurs efficiently on both liquid and solid particles, but only at high latitudes where the temperature drops below a threshold value of about 195 K.

The mechanism of reactions (1) and (2) is ionic in nature: HCl solvates in aqueous phase forming hydrochloric acid, and hence chloride anions. The chlorine atom in HOCl and $ClONO_2$ is slightly electropositive; both of these species react very fast with the chloride anions to produce molecular chlorine, which is rapidly desorbed from the ice surface.

The first step in the mechanism of reactions (1) and (2) on ice particles involves incorporation of HCl vapor into the surface layers. The high affinity of HCl for the ice surface is a consequence of ion pair formation, which takes place because the surface layers of ice are not as ordered as the ice crystal itself; they form a "liquid-like" aqueous layer in the presence of trace amounts of HCl (this species can depress the freezing point of water down to ~195 K). The amount of energy associated with physical adsorption involving only a hydrogen bond is too small to explain the experimental observations of HCl uptake by ice[6]; hence, reaction mechanisms involving weak physical adsorption and resorting to conventional Langmuir-type adsorption isotherms are not suitable.

The second step in the reaction mechanism involves incorporation into the surface layers of HOCl for reaction (2), or $ClONO_2$ for reaction (1). All the reactants have a high mobility on the surface: Once incorporated into the condensed phase, the HOCl molecules almost always find chloride ions before returning to the gas phase. Similarly, the $ClONO_2$ molecules in the surface layers also find chloride ions before reacting with water. This explains the experimental observation of a lack of dependence of the reaction rate on the concentration of HCl vapor: as long as HCl is in excess, the overall reaction is nearly zero order in HCl and first order in HOCl (or $ClONO_2$), and is only very weakly dcpcndcnt on temperature.

Reactions (1) and (2) also occur rapidly on nitric acid trihydrate (NAT) surfaces, with a mechanism similar to that on ice surfaces. However, there is an additional parameter that should be taken into account, namely the relative humidity. When

NAT is in equilibrium with ice, its H_2O vapor pressure is the same as that of ice, and the reaction probability γ for reactions (1) and (2) is practically the same as that on ice. As the relative humidity (and hence the H_2O vapor pressure of NAT) decreases, the reaction probability γ initially remains high, but it decreases for relative humidity values below ~50% (with respect to ice), and eventually it reaches values more than two orders of magnitude smaller. This behavior can be explained with a reaction mechanism involving the availability of water at the NAT surface to induce solvation and uptake of HCl vapor: at very low relative humidities, there is excess HNO_3 on the surface and solvation is hindered.

4 THERMODYNAMIC PROPERTIES OF STRATOSPHERIC AEROSOLS

To investigate the nature and chemical identity of stratospheric aerosols, it is useful to consider first the thermodynamic properties of the aerosols, and subsequently the rates of transformation between the various phases for the different chemical systems of interest. The primary thermodynamic properties of interest are the mole fractions of the various chemical components of the particles in the condensed phase and their vapor pressures, the partial pressures or concentrations of these components in the gas phase, and the temperature. The vapor pressures for low-volatility components (e.g., NaCl, and sometimes H_2SO_4) need not be considered explicitly; furthermore, the effect of total pressure on thermodynamic properties is negligible for atmospheric conditions.

The thermodynamic properties that determine the stability and equilibrium composition of the various phases can be represented conveniently by phase diagrams. A comparison of the atmospheric partial pressures with the vapor pressures displayed in the phase diagrams provides a useful guideline to establish the identity of the various condensed phases that can exist under atmospheric conditions. A specific atmospheric condition or state can be represented by a point in a phase diagram, while an atmospheric process or trajectory is represented by a line.

To illustrate the use of phase diagrams, consider the H_2SO_4/H_2O system. Figure 1 shows the equilibrium freezing temperatures for this system as a function of composition, and Figure 2 is a logarithmic plot of the water vapor pressure versus inverse temperature. The dashed lines in Figure 2 represent the H_2O vapor pressure of solutions with constant composition; it follows from the Clausius–Clapeyron equation that these lines are nearly straight, their slopes being equal to the partial molar enthalpy of evaporation of H_2O. The solid lines in Figures 1 and 2 represent conditions of coexistence of two condensed phases; the lines separate regions of stability for the different phases. Note also that the stability region for a particular hydrate is represented in Figure 2 by a surface, whereas in Figure 1 it is a vertical line, as the composition of the hydrate is fixed.

Phase diagrams represent properties at thermodynamic equilibrium. However, it often happens that a new phase does not form because of the presence of kinetic barriers to nucleation. This is the case for sulfuric acid solution droplets: They remain liquid throughout most of the stratosphere, even though their temperature

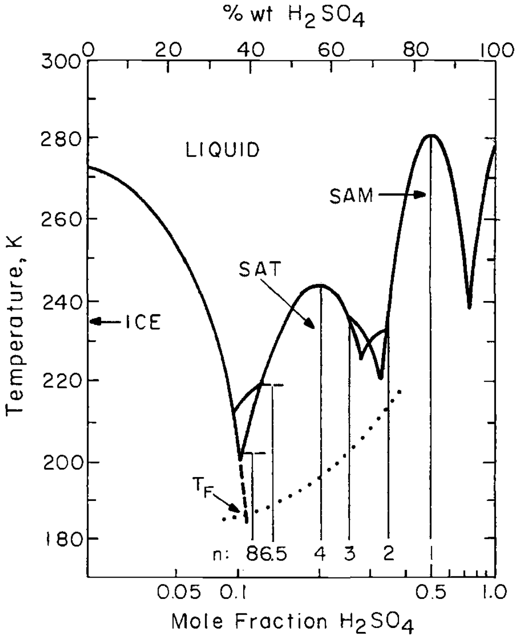


Figure 1 Temperature vs. composition phase diagram for the H_sSO_4/H_2O system. The solid line represents freezing temperatures (solid–liquid coexistence conditions). The thin vertical lines give the composition of the solids. The dotted line represents the equilibrium composition of metastable liquid particles in the stratosphere as a function of temperature, in an air parcel containing 3 ppmv of water vapor at ~16 km altitude; T_F indicates the ice frost point for this air parcel.

is below the freezing point; that is, they supercool very readily. Phase diagrams still provide useful representations of such metastable phases; for example, in Figure 2 the vapor pressures of supercooled solutions are given by extensions of the dashed lines into the solid stability regions. Consider, for example, a stratospheric air parcel around 16 km altitude containing 3 parts per million (ppm) of water vapor, and cooling between 220 and 190 K; the properties of liquid sulfuric acid aerosols in such a parcel are represented by the dotted lines in Figures 1 and 2. The droplets swell and become less concentrated as the temperature drops.

There are similar phase diagrams for the HNO_3/H_2O system; however, because of the relatively high volatility of HNO_3 compared to H_2SO_4, it is useful to consider an additional phase diagram consisting of a logarithmic plot of the HNO_3 vapor pressure versus inverse temperature in order to elucidate the nature of the phases that are stable in the stratosphere for this system. Yet another version of a phase diagram is shown in Figure 3: It is a logarithmic plot with the vapor pressure of one component (HNO_3) in one axis and the vapor pressure of the other component (H_2O) in the

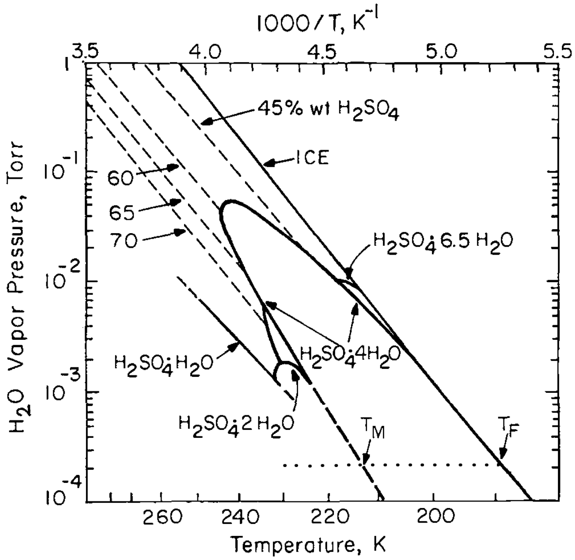


Figure 2 Water vapor pressure vs. temperature phase diagram for the H_2SO_4/H_2O system. The solid lines represent solid–liquid coexistence conditions; the thin dashed lines represent vapor pressures of liquids of constant composition given as wt % H_2SO_4, and the thick dashed line represents equilibrium coexistence conditions for sulfuric acid tetrahydrate (SAT) with the liquid (supercooled with respect to the dihydrate and monohydrate). The dotted line corresponds to that in Figure 1.

other axis. Temperature is a parameter in the figure: Isotherms are straight lines in the solid stability regions, and it follows from the Gibbs–Duhem equation that the slope of these isotherms is related to the composition of the condensed phase: The value of the slope is three for NAT, one for nitric acid monohydrate, and the isotherms are vertical for water ice. The isotherms are curved for the liquid stability region since the composition of the liquid can vary continuously.

Figure 3 shows that the most stable solid phase for the HNO_3/H_2O system in the polar stratosphere is NAT; however, there are indications from laboratory studies that nitric acid dihydrate (NAD) (which is not represented in Fig. 3), may nucleate first,[3] even though for most conditions NAD is metastable with respect to NAT. Note also that Figure 3 shows that the vapor pressure of H_2O over NAT at a particular temperature can have values ranging from the vapor pressure of pure water ice to that of the monohydrate; as discussed above, the reaction probability for reactions (1) and (2) remains large along the isotherm as long as the H_2O vapor pressure does not drop to a value below a half to a third of the value in the ice stability region.[6]

The freezing points and vapor pressure values required to generate phase diagrams such as those shown in Figures 1 to 3 are usually determined directly

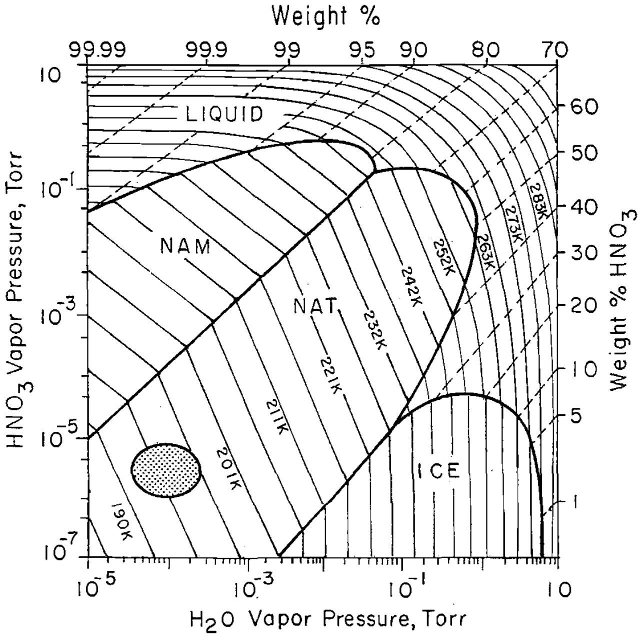


Figure 3 Nitric acid vs. water vapor pressure phase diagram for the HNO_3/H_2O system. The thick solid lines represent coexistence conditions for two condensed phases; the thin lines are isotherms (labeled with T in Kelvin). The dashed lines represent vapor pressures of liquids with constant composition (labeled as wt % HNO_3 in the upper and right axis). The dashed region in the lower left corner represents typical conditions in the lower stratosphere over the poles.

from laboratory experiments. However, the vapor pressures can also be determined indirectly by other means, e.g., from voltage measurements in electrochemical cells—accurate phase diagrams can be constructed from measurements of such voltages together with calorimetric measurements of the enthalpies and temperatures of the various phase transitions of interest. For ternary systems such as $H_2SO_4/HNO_3/H_2O$ the phase diagrams are more complicated but are nevertheless just extensions of the binary diagrams to one more dimension. On the other hand, the vapor pressures for such multicomponent systems can be reliably estimated using semiempirical thermodynamic models.[7]

5 MECHANISM OF FORMATION OF STRATOSPHERIC AEROSOLS

The source of sulfuric acid in the stratosphere is carbonyl sulfide (COS), which is of biological origin. Although emitted from the ground, it is sufficiently stable to reach the stratosphere, where it oxidizes to form sulfur dioxide, SO_2. This species is further oxidized to produce H_2SO_4 through the following mechanism:

$$SO_2 + OH \rightarrow HOSO_2 \tag{12}$$

$$HOSO_2 + O_2 \rightarrow HO_2 + SO_3 \tag{13}$$

$$SO_3 + 2H_2O \rightarrow H_2SO_4 + H_2O \tag{14}$$

The rate-determining step is reaction (12). Reaction (14) is second order in water vapor[2]; it is fast throughout the atmosphere, except in the upper stratosphere, where the water vapor concentration is relatively small. There is no net consumption of radicals with this mechanism in the atmospheric oxidation of SO_2; the overall effect is merely the conversion of OH into HO_2.

A second important sulfur source consists of volcanic eruptions, a few of which inject SO_2 directly into the stratosphere, such as El Chichón in Mexico, in 1982, and Mount Pinatubo in the Philippines, in 1991. Mount Pinatubo introduced enough SO_2 to increase the stratospheric H_2SO_4 burden by a factor of ~30,[8] inducing noticeable global cooling. Satellite observations indicate that the SO_2 oxidation process takes several weeks, and that the excess particles remain in the stratosphere a couple of years; the sulfuric acid haze formed is the origin of bright red sunsets.

As mentioned above, at high latitudes and in the winter months, the sulfuric acid/water droplets cool and grow to become PSCs, absorbing water and nitric acid vapor. Observations in the lower stratosphere of a rapid growth in the volume of these aerosol particles around 195 K were originally interpreted as resulting from the formation of nitric acid trihydrate (NAT); however, a more recent analysis of the field observations indicates that the particles often remain liquid, reaching compositions such as 30 wt % H_2SO_4 and 30 wt % HNO_3[7], with the particle growth being a consequence of rapid H_2O and HNO_3 uptake below a threshold temperature, which happens to approximately coincide with the temperature below which NAT becomes stable.

There is a large nucleation barrier for these supercooled liquid particles to freeze, and laboratory observations show that freezing does not occur until the temperature has dropped several degrees below the ice frost point, which is around 185 K in the lower stratosphere. Under such conditions water ice crystallizes first, leading to the formation of type II PSCs. Some atmospheric observations indicate the presence of solid particles at temperatures above the frost point; it is likely, however, that such particles had reached lower temperatures earlier and that water ice induced the formation of the acid hydrates. As the particles warm up, ice evaporates first and eventually the hydrates melt at the equilibrium phase transition temperatures expected from the phase diagrams (i.e., temperature T_M in Fig. 1), since there is essentially no nucleation barrier for the melting process.

Many questions remain, however, regarding the nature and the rates of liquid–solid phase transformations in PSCs. For example, in the Arctic stratosphere temperatures fall below the frost point much less frequently than over Antarctica, and yet solid PSCs do form, perhaps as a consequence of mesoscale temperature fluctuations.[9] There are also questions regarding denitrification, the PSC sedimentation process referred to above leads to the removal of nitric acid, and hence, of nitrogen oxides. The process is not sufficiently well understood to permit reliable predictions, for example, of H_2O, HNO_3, and NO_x levels at high latitudes for scenarios involving emissions from proposed future supersonic transports that would fly in the lower stratosphere.

REFERENCES

1. Molina, M. J., L. T. Molina, and D. M. Golden, Environmental chemistry (gas and gas-solid interactions): The role of physical chemistry, *J. Phys. Chem.*, *100*, 12888, 1996.
2. Molina, M. J., L. T. Molina, and C. E. Kolb, Gas phase and heterogeneous chemical kinetics of the troposphere, *Ann. Rev. Phys. Chem.*, *47*, 327, 1996.
3. Kolb, C. E., D. R. Worsnop, M. S. Zahniser, P. Davidovits, C. F. Keyser, M. T. Leu, M. J. Molina, D. R. Hanson, A. R. Ravishankara, L. R. Williams, and M. A. Tolbert, Laboratory studies of atmospheric heterogeneous chemistry, in J. R. Barker (Ed.), *Advanced Series in Physical Chemistry: Progress and Problems in Atmospheric Chemistry*, World Scientific Publishing, Singapore, 1995, pp. 771–875.
4. DeMore, W. B., S. P. Sander, D. M. Golden, R. F. Hampson, M. J. Kurylo, C. J. Howard, A. R. Ravishankara, C. E. Kolb, and M. J. Molina, *Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling, Evaluation No. 12*, JPL Publication 97-4, Jet Propulsion Laboratory, Pasadena, CA, 1997.
5. Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics from Air Pollution to Climate Change*, Wiley, New York, 1997.
6. Molina, M. J., The probable role of stratospheric "ice" clouds: Heterogeneous chemistry of the "ozone hole," in J. G. Calvert (Ed.), *The Chemistry of the Atmosphere: Its Impact on Global Change*, Blackwell, Oxford, 1994, pp. 27–38.
7. Peter, T., Microphysics and heterogeneous chemistry of polar stratospheric clouds. *Ann. Rev. Phys. Chem.*, *48*, 785, 1997.
8. McCormick, M. P., L. W. Thomason, and C. R. Trepte, Atmospheric effects of the Mt Pinatubo eruption, *Nature*, *373*, 399, 1995.
9. Carslaw, K. S., M. Wirth, A. Tsias, B. P. Luo, A. Dörnbrack, M. Leutecher, H. Volkert, W. Renger, J. T. Bacmeister, and T. Peter, Particle microphysics and chemistry in remotely observed mountain polar stratospheric clouds, *J. Geophys. Res.*, *103*, 5785–5796, 1998.

SECTION 2

HYDROLOGY

CHAPTER 23

HYDROLOGY OVERVIEW

SOROOSH SOROOSHIAN AND MARTHA P. L. WHITAKER

1 INTRODUCTION

This chapter provides a brief overview of the hydrologic cycle and discusses the role of hydrology, not only in the global contexts of weather and climate but also in the local and regional contexts of weather as it affects water resources management. This chapter contains a description of the hydrologic cycle and the identification of its specific reservoirs and fluxes. In each description, their relevance to various scales of the hydrologic cycle is discussed. The concept of the water balance is subsequently introduced as the basic tool with which one can understand the effects of perturbations on the hydrologic cycle, regardless of the scale of interest. Provided with this knowledge, stakeholders with concerns ranging from global climate change to flood forecasting will be better informed to responsibly manage water resources.

This section of the handbook also contains in-depth discussions on each flux of the hydrologic cycle, including precipitation of rain and snow (Chapters 24 and 25, respectively), evaporation and transpiration (Chapter 26), infiltration and soil moisture (Chapter 27), groundwater flow (Chapter 28), and runoff generation (Chapter 29). The final four chapters describe various types of mathematical tools with which one can analyze hydrologic phenomena: Chapters 30 to 32 specifically describe tools to better understand hydrologic events such as high river flows, runoff, and floods, and Chapter 33 explores the uses of remote sensing and geographic information systems to both visualize and quantify large-scale hydrologic phenomena.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

2 THE HYDROLOGIC CYCLE

Figure 1 represents a conceptual model of the hydrologic cycle and shows Earth's water movement between the ocean, land, and atmosphere. As with all cycles, it is ongoing and continuous, and there is no specific start or end point; however, because the main focus of this handbook is meteorology, precipitation is an appropriate place to begin an evaluation. Precipitation is water released from the atmosphere in the form of rain, snow, sleet or hail. During precipitation, some of the moisture is evaporated back into the atmosphere before ever reaching the ground. Some precipitation is intercepted by plants, a portion infiltrates the ground, and the remainder flows off the land into lakes, rivers, or oceans. An important difference between the roles of snow and rain is that runoff occurs relatively quickly following the rain event, whereas snow usually melts much more slowly over days, weeks, or months. The subsequent surge of snowmelt runoff can provide seasonal recharge to groundwater resources but can also trigger flood conditions if the snowmelt occurs too rapidly and in excessive amounts. In addition, the solid snow or ice may change directly into a gas, skipping the liquid state, in the process called *sublimation*.

When precipitation is intercepted by plants, it is eventually evaporated back to the atmosphere. When it infiltrates the ground, it can be taken up by roots and transpired by plants, it can be evaporated from the soil, or it may recharge an aquifer. The water in an aquifer is called *groundwater*, and its rate of flow in the subsurface is such that water can reside in aquifers for days to centuries before discharging to a surface body of water (e.g., river, lake, ocean). Once groundwater has discharged into a river, lake, or ocean, the surface of the water body is exposed for evaporation, causing moisture to collect and concentrate in the atmosphere, eventually returning to the earth as precipitation as the cycle begins again. In addition to natural discharge, groundwater can more rapidly discharge when an aquifer is pumped. With the advent of motorized pumps, the rapid removal of groundwater from aquifers is a relatively recent phenomenon that has greatly affected the depletion of the aquifers and the water balance of many catchments.

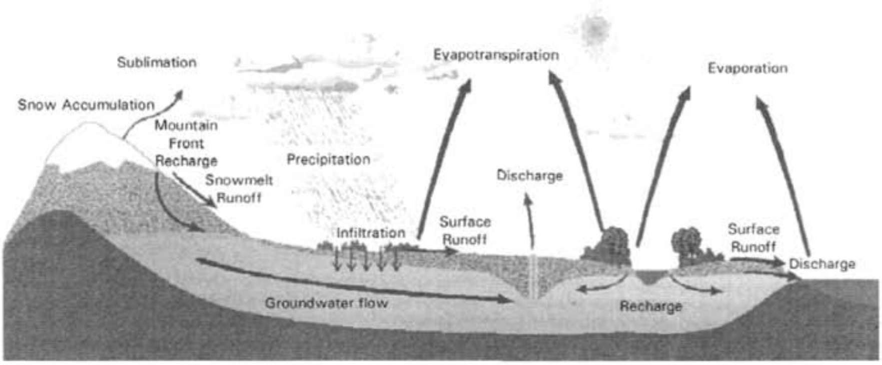


Figure 1 Schematic representation of the hydrologic cycle. (Courtesy B. Imam.)

While the hydrological cycle is a continuous process, it is by no means uniform throughout the globe: the residence time of water varies—often dramatically—among different portions of the cycle. For example, water is continuously evaporated from the surfaces of water bodies (such as oceans, lakes, and rivers). Similarly, precipitation that is intercepted by plants and other surfaces is often evaporated within a matter of hours. Once evaporated, it takes an average of 10 days for a water molecule to cycle through the atmosphere, but if it infiltrates to the water table, or if the precipitation occurs in a polar region, it may reside for hundreds of years before transferring to another step in the hydrologic cycle. In addition to variable residence times, the processes associated with the hydrologic cycle are not evenly distributed over the globe; they vary by climatic region. For example, evapotranspiration occurs readily in semiarid and arid regions, but subsequent precipitation may not occur within the same basin or region. The dramatic differences in how the cycle operates are especially evident when one evaluates the hydrologic cycle at the catchment scale.

Additional, variably detailed discussions of the hydrologic cycle may be found in Horden (1998), Maidment (1993), Driscoll (1986), and Freeze and Cherry (1979). Chahine (1992) offers a particularly thorough discussion of the hydrologic cycle in the context of climate studies and hydrologic modeling.

3 RESERVOIRS

On a global scale, the important reservoirs in the hydrologic cycle are the ocean, atmosphere, polar ice, groundwater, and moisture from land surfaces. At the global scale, water is transferred between reservoirs via four fluxes: precipitation, evapotranspiration, sublimation, and runoff. On a catchment scale, the availability of fresh water is the focus. Critical reservoirs on this scale are the atmosphere, lakes, rivers, and groundwater. Oceans and polar ice are typically irrelevant at the catchment scale, although seasonal snowmelt can contribute significantly (or destructively, in the case of floods) to a basin's water resources. Fluxes within a catchment are more strongly weighted toward the recharge and withdrawal of potable groundwater, as well as the occurrence of surface water flows.

Fresh water comprises only 2.5% of the world's total water supply. Of this scant freshwater supply, 69.6% is immobilized in ice and snow, primarily in the polar regions; nonsaline groundwater accounts for 30.1%; and the remainder of fresh water (0.3%) is distributed among lakes, rivers, wetlands, atmospheric water, and biological water found in plants and animals. While groundwater is Earth's second largest source of fresh water, on the average it accounts for less than 1% of the earth's total water supply; however, freshwater availability varies greatly on a regional basis. Figure 2 shows the distribution of terrestrial water in terms of Earth's total water supply and Earth's freshwater supply.

The major reservoirs of the hydrologic cycle are described below, and their role in the global- and/or catchment-scale hydrologic cycle is discussed briefly.

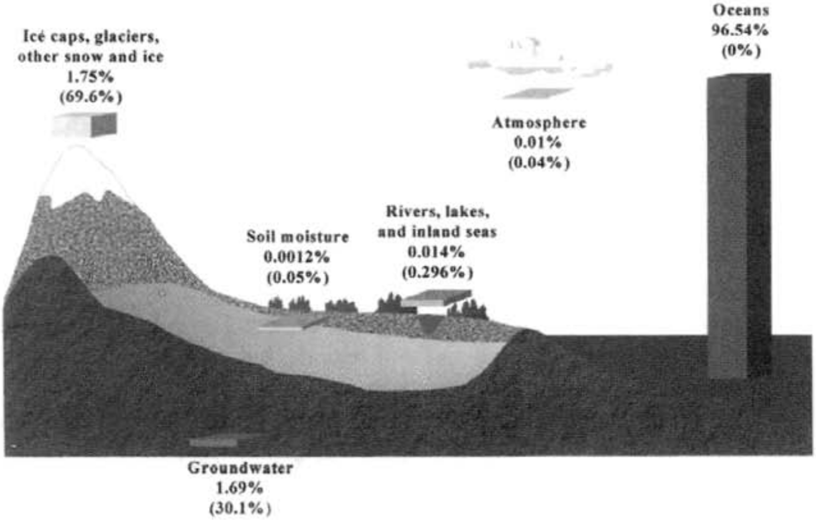


Figure 2 Distribution of terrestrial water partitioned in reference to Earth's total water supply (top percentage) and Earth's freshwater supply (bottom percentage, in parentheses).

Oceans

Roughly 70% of Earth's surface area and 96.5% of Earth's water volume is ocean water. In other words, the oceans comprise 96.5% of the water in the global hydrologic cycle, and with the considerable volume and energy circulated in this vast reservoir, they have a tremendous effect on climate. In particular, ocean surface temperatures (e.g., El Niño Southern Oscillation, or ENSO) can powerfully impact atmospheric circulation patterns.

Polar Ice

While polar ice represents only 1.7% of Earth's total water supply, it comprises almost 67% of Earth's freshwater reserves. Snow and ice-covered surfaces significantly impact Earth's climate because they have a very high ability to reflect solar (short-wave) radiation, but they contribute only marginally to Earth's hydrologic cycle.

Seasonal Snow and Ice

Although seasonal snow and ice represent only 1% of the world's fresh water, the annual melt cycles can play a significant role in water resources management at the catchment scale (Chapter 25). For example, snowmelt can be a welcome source of replenishment for lakes, rivers, reservoirs, and groundwater; however, rapid melting of large volumes of snow can cause flooding and subsequent contamination (e.g., sewer backups) of water resources.

Land-Based Surface Water

Land-based surface water includes rivers, lakes (both fresh and saline), surface soil moisture, and wetlands. On the global scale, the volume of land-based surface water is a small part (0.0153%) of the hydrologic cycle, but the rate of flux through these components, and hence the availability of fresh water, is critically important to human activities within individual watersheds. For thousands of years, humans have interacted with land-based surface water by building aqueducts, digging irrigation canals, and more recently, by diverting or damming rivers, and by pumping, which captures groundwater without allowing it to discharge naturally to a surface water body. With the occurrence of global climate change and the possibility of increased frequency of floods or droughts, responsible hydrologic management will be key to achieving sustainable water resources within catchments. This is possible only through scientists' continuous progress in understanding the various components of the hydrologic cycle in each catchment, improved modeling of all components of the hydrologic system, and narrowing the uncertainty bounds on hydrologic predictions.

Biological water is the primary constituent of living tissue in all plants and animals. It is another form of land-based surface water; yet it is only a minuscule percentage (0.0001%) of the total water on Earth. Regardless of its small percentage, the critical role of plants in the vertical transfer of water from soil and subsurface reservoirs to the atmosphere—particularly in semiarid regions—cannot be ignored (see Chapter 26).

Groundwater

Groundwater generally refers to the water that exists in saturated layers of porous geologic materials, called *aquifers*. On the global scale, groundwater is a slow-moving reservoir that comprises approximately 0.5% of the world's total water, yet it accounts for 30% of Earth's freshwater reserves. Accordingly, the existence and replenishment of these reserves is critical for maintaining a water supply for many human communities.

The global hydrological cycle generally depicts groundwater as slowly discharging to oceans, lakes, and rivers, but groundwater discharge at the catchment scale is rapidly accelerated by the electric turbine pump. A major change in groundwater discharge rates in the United States came about with the widespread agricultural use of electric turbine pumps with the electrification of rural America following World War II. The effect of widespread use of the electric turbine pump for irrigated agriculture has been a decrease in water tables (up to 400 ft within 50 years in fast-growing metropolitan areas such as Tucson and Phoenix, Arizona) and, in some cases, groundwater capture of surface water resources. Only within the past century have scientists and water managers begun to understand the depleting effects of groundwater pumping on surface water flows (e.g., Bouwer and Maddock, 1997; Maddock and Vionnet, 1998; Glennon and Maddock, 1997).

Regional precipitation patterns often determine whether groundwater supplies are a sustainable resource for a catchment's population. In semiarid and arid regions where precipitation is light and water demands increase with growing metropolitan populations, natural recharge is insufficient to maintain a long-term, dependable water supply, and can often result in land subsidence leading to considerable property damage. For some arid regions, the mining of groundwater as a nonrenewable resource is the only viable alternative (e.g., El Geriani et al., 1998; Gijsbers and Loucks, 1999), but in other areas, water conservation and artificial recharge efforts can prolong a basin's water supplies.

Atmosphere

Water resides in the atmosphere for approximately 8 to 10 days before falling back to Earth. The presence of water vapor in the atmosphere affects weather and climate in several ways. First, it is the source for all forms of precipitation (such as rain, snow, sleet, and hail). In addition, it serves to regulate Earth's surface temperature by absorbing and reflecting incoming short-wave solar radiation and absorbing Earth's emission of long-wave radiation. Finally, atmospheric water is a source of latent heat that can mobilize large air masses. Although the averagc volume of atmospheric water is only 0.001% of Earth's total water reserves, its role in climate and weather is substantial for both global and catchment considerations.

4 FLUXES

The major fluxes within the hydrologic cycle are described below, and their role in the global- and/or catchment-scale hydrologic cycle is addressed.

Precipitation

Precipitation is the process by which liquid and solid-phase aqueous particles, such as rain, snow, sleet, and hail, fall from the atmosphere to Earth's surface. The occurrence of precipitation over land is typically cited as the driving force of the hydrologic cycle, since it triggers the commencement of other fluxes (evapotranspiration, runoff, infiltration) by providing a new source of moisture to the system. The intensity and frequency of precipitation vary considerably both spatially and temporally, and the effects of precipitation can be both welcome (e.g., during droughts) or undesirable if it occurs in excess and causes subsequent flooding. In some regions, where dry air dominates the weather conditions, precipitation may fall from the clouds but evaporate before ever reaching the ground; this is a phenomenon known as *virga*. Measurements and estimates of precipitation (volume and intensity) are critical to any study or modeling effort involving the hydrologic cycle. Rain gages have been the primary mechanisms for observation, but their sparse distributions and other limitations do not provide the spatial and temporal resolution needed for various modeling and research efforts. Recent advances are rapidly improving the situation by merging satellite and radar with gage information. Examples of such work are discussed in Smith (Chapter 24), Bales and Cline (Chapter 25), Sorooshian et al. (2000), Adler et al. (1993), Arkin and Xie (1994), and Xie and Arkin (1995, 1996, 1997).

Evapotranspiration

Evaporation is defined as "the rate of liquid water transformation to vapor from open water, bare soil or vegetation with soil beneath" (Shuttleworth, 1993), and *transpiration* is the rate of water added to the atmosphere as it moves from soil through the stomata of vegetation. Evapotranspiration (ET) is thus a compound term that describes the collective effect of evaporation of water and transpiration of plants. It is the primary process that moves moisture from Earth's surface to the atmosphere. The only other natural means by which water is transferred from the earth to the atmosphere is the process of *sublimation*, where solid phases of water (e.g., snow and ice) transition directly to atmospheric vapor in the absence of melting. Sublimation typically occurs in regions of cool temperatures and low relative humidity. Evapotranspiration is often an elusive variable to quantify, as it varies diurnally, seasonally, and with changes in precipitation events. A more thorough discussion of evaporation, including a description of various evaporation measurement techniques, may be found in Chapter 26.

Runoff

Runoff is generally thought of as the movement of excess rainfall across the land surface into rivers, lakes, or the ocean. It occurs when the rate of precipitation exceeds the rate of infiltration at the soil surface, or when soil is saturated. Runoff is a particularly important process at the catchment scale, since it can recharge reservoirs and replenish rivers that may subsequently recharge the groundwater; runoff can also cause soil erosion, and excess runoff can lead to flooding. In Chapter 29, Beven offers a broader historical description of the definition of runoff and also describes various hydrological components that contribute to its generation.

Groundwater

Natural groundwater fluxes are typically slow; water may reside in an aquifer for as little as a few hours (as in the case of river bank storage) or for hundreds of years. Accordingly, groundwater itself is often perceived, on the average, as a relatively slow-moving reservoir in the global hydrologic cycle. At the catchment scale, however, where stream–aquifer interactions are relatively rapid and substantial, the average groundwater fluxes are relatively fast moving. They comprise: (1) the natural flow of water between watersheds, (2) the water pumped from an aquifer, (3) mountain-front recharge (seasonal infiltration of snowmelt at the base of mountain ranges), (4) event-based infiltration (infiltration from precipitation and subsequent rises in surface water levels, especially rivers), and (5) artificial recharge via anthropogenic conservation projects. To better comprehend such complex hydrologic flow scenarios, it is critical to first understand the basic principles of groundwater flow. In Chapter 28, Yeh not only describes Darcy's law, the fundamental flow equation for fluid in porous media, but also reviews flow equations for various aquifer conditions (e.g., confined, leaky, unconfined) and describes the use of groundwater flow models used for water resources management.

The Water Balance: Global to Catchment Scale

The *water balance* simply refers to the volumes of water that flow through various components of the hydrologic cycle. More specifically, it is another useful conceptual model in which the components of the hydrologic cycle are evaluated as storage units that are affected by various inputs and outputs. If the various components of the cycle can be quantified or at least estimated, it is possible to gain an understanding of how alteration of a component might affect the balance of the hydrologic cycle. The most simplistic formulation of a water balance is denoted by the elementary continuity equation that conveys the notion that "input to a hydrologic system equals the output from the system, plus or minus any changes in storage":

$$I = O \pm \Delta S$$

where, for a given domain, I is the total inflow, comprised of surface runoff (into the domain), groundwater inflow and precipitation; O is the outflow of evapotranspiration, surface runoff (out of the domain), and groundwater; and ΔS is the change in storage, whose variables are determined by the scale of the domain.

The concept of a water balance is useful at both global and regional scales. At the basin or watershed scale, where groundwater–surface water interactions might encompass the primary focus, precipitation and groundwater inflow would be a model's input, while overland flow, groundwater outflow, and evapotranspiration

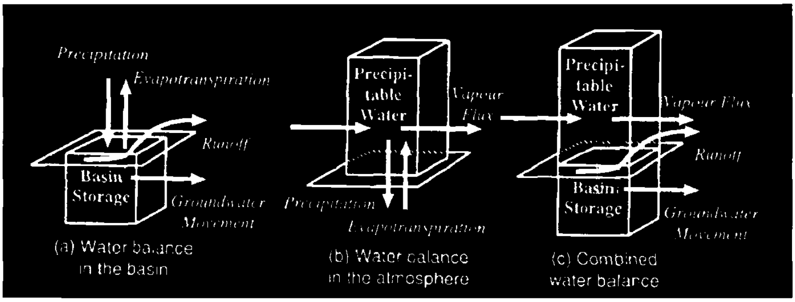


Figure 3 (*a*) Water balance at the catchment scale, (*b*) atmospheric water balance, (*c*) combined land surface atmosphere water balance (after Oki, 1995, 1999).

would be its outputs. Figure 3*a* shows a conceptual model for the water balance at a watershed scale. Note that the figure does not represent changes in storage caused by anthropogenic activities such as pumping, artificial recharge, or surface water diversions from or to other basins. The consideration of such anthropogenic effects in determining the water balance may be critical, depending on the spatial and temporal scales under consideration.

For meteorologists, the most relevant transfers of water in the hydrological cycle are the vapor flux and moisture exchanges between the atmosphere and Earth. With Earth's surface as the focal point of the cycle, we can evaluate precipitation as the major input to the system, while evaporation and transpiration output moisture to the atmosphere. The change in storage could include, for example, the infiltrated water that is not reevaporated into the atmosphere, or water that becomes frozen in polar ice caps. Such water is temporarily and relatively static in the land surface–atmosphere system. That is, given that atmospheric scientists tend to evaluate the hydrologic cycle over a time frame of about 8 to 10 days (i.e., the average amount of time that water cycles through the evaporation–condensation–precipitation cycle), water that resides as ice or becomes slow-moving groundwater is seen as a very slow change in storage of the land surface–atmosphere system. Figures 3*b* and 3*c* show conceptual models of the water balance in the atmosphere and the combined basin–atmosphere water balance, respectively.

In addition to catchment-scale analyses, the evaluation of the water balance of the hydrologic cycle at increasingly larger scales is important because the issues and

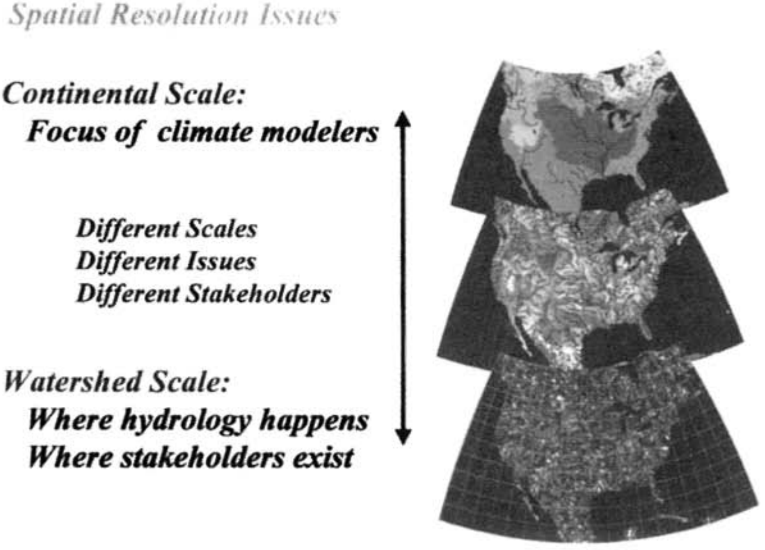


Figure 4 Identification of the spatial scales at which hydrologic phenomena are measured. Different scales delineate different stakeholders, and also determine the various levels of water management issues.

stakeholders are different for each scale. Figure 4 illustrates the different stakeholder affiliations at various scales of hydrologic investigation. The uppermost illustration of North America shows outlines of continental-scale basins, and the adjacent text indicates that the corresponding stakeholders are climate modelers. The results of climate modelers' research potentially affect international, global policies and may influence industrial emissions standards for greenhouse gases. In the middle illustration of North America, sub-basins are delineated, and in the bottom illustration, copious individual watersheds are outlined. The sub-basin to watershed scales are where hydrology happens on scales at which most people can observe more immediate and obvious impacts to their local water supply. Water resources management issues at the watershed scale are thus clearly different than those of the sub-basin and continental basin; however, our understanding of the water cycle at all scales—both spatial *and* temporal—is critically important to addressing the needs of various stakeholders.

Figure 5 shows how specific water resource issues vary in space and time. The spatial scale varies from 10 to 10^6 km^2, and the temporal scale varies from days to centuries. Across the top of the diagram, the types of prediction are identified that can be made for a corresponding time scale: the shortest predictions are weather forecasts (ranging from less than a day to several days), while the longest predictions are climate change (on the order of centuries). The center of the diagram identifies the types of water resources management issues corresponding with different permutations of spatial and temporal ordinates.

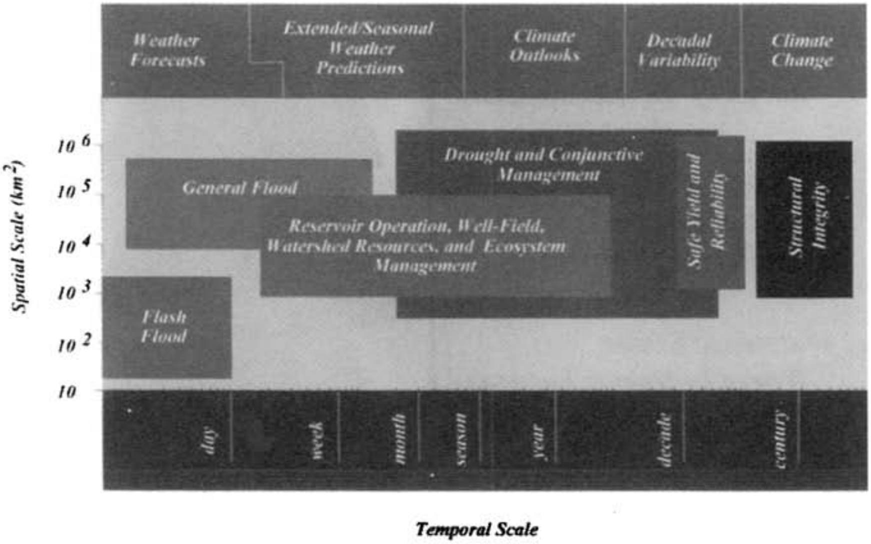


Figure 5 Schematic illustration of how water resources issues vary across spatial and temporal scales (after National Research Council, 1998).

5 MODELING AND REMOTE SENSING OF THE GLOBAL HYDROLOGIC CYCLE: MODELING GLOBALLY, BENEFITING LOCALLY

Modeling efforts can help us understand the potential impact of human activities on the hydrologic cycle and climate in particular. Unfortunately, our current efforts are hampered by a lack of quantitative data on the distribution and flux of water in its various states, and by our uncertainty of the interactive functioning within the hydroclimatological system (Chahine, 1992). To address these uncertainties, the Global Energy and Water Cycle Experiment (GEWEX) was initiated by the World Climate Research Programme (WCRP) in 1988, to observe and model the hydrological cycle and energy fluxes in the atmosphere, at the land surface, and in the upper oceans. A complementary, parallel program, known as the Biospheric Aspects of the Hydrologic Cycle (BAHC), was initialized by the International Geospheric-Biosphere Programme (IGBP) to complement the GEWEX program by placing the emphasis on the biological aspects of the hydrologic cycle—particularly the role of plants in the vertical transfer of water and carbon between the land and atmosphere. For its part, GEWEX has served as a coordinating body of scientists who initiate and facilitate communication among numerous international research teams investigating various aspects of hydrometeorological processes. The hydrological cycle between the land surface and upper atmosphere has subsequently received considerable attention (Chapter 27). Scientists have begun to suggest that we should also consider how land–atmosphere interactions at the basin-scale affect or are affected by climate.

The National Research Council (1998) stated that most water resources management problems are addressed at the sub-basin and watershed scales. Five GEWEX continental-scale experiments (CSEs) have been making promising contributions to improving our understanding of the water balance at scales small enough to be useful for water resources management purposes. For example, the first CSE to be established was GCIP, the GEWEX Continental-Scale International Project, a large-scale study of the Mississippi Basin. During its early phases, GCIP developed data sets, models, and a research framework to better understand and predict land–atmosphere interactions on climatic time scales (seasonal and annual) in the Mississippi River Basin. In fact, GCIP succeeded in meeting most of its objectives, and the project has since transformed to encompass the entire continental United States, as well as part of northern Mexico. This follow-on research project is called the GEWEX America Prediction Project (GAPP). Additional CSEs were selected to represent different climatic conditions than in the Mississippi River Basin. Evaluated together, and separately, the resulting coupling of land surface models with atmosphere and ocean models is a primary step toward improved climate prediction (Chahine, 1992). Such improvements of operational hydrologic and water resources management tools are critical in helping to bring global and GCIP/GEWEX-scale climate predictions down to a scale important for addressing local and regional water resources issues (National Research Council, 1998).

With the ever-increasing popularity of geographic information systems (GIS) and remote sensing (RS), we are witnessing many new advances in hydrologic modeling, particularly distributed models, which more accurately represent spatial features. Engman and Mittikalli (Chapter 35) provide a brief summary of GIS and RS issues.

6 STOCHASTIC MODELS OF HYDROLOGIC PROCESSES

A *stochastic* process is described by a randomly determined set of observations, each of which is a sample of one element from a probability distribution. Virtually all hydrologic processes can be characterized as stochastic. It is therefore not surprising that the development and application of statistical and stochastic methods in hydrology date back several decades (e.g., Fiering (1967, 1976); Haan (1997); Chow et al. (1998), among many others). The application of flood frequency analysis in hydrologic design and operation of water resources systems is a good example of how influential and powerful these methods have become. Valdés et al. (Chapter 34) address the methods used for stochastic *forecasting*, while Salas et al. (Chapter 33) discuss stochastic *simulations* in the context of precipitation and streamflow. Forecasts are generally applied to operational and management scenarios, while simulations are used in the context of design and planning. More recently, it has become increasingly popular to apply stochastic simulation tools to more thoroughly address the uncertainties of hydroclimatic processes. Salas and Pielke (Chapter 32) provide an excellent review of the current state of the literature in this area.

7 CONCLUSION

The discussion provided above is a brief overview of the various elements of the hydrologic cycle (fluxes and processes) and also offers a summary of ongoing related research activities. It is expected that research and development activities in hydrology and water resources will continue to follow two general paths: theoretical and applied. Theoretical research most related to this handbook will be driven by the need to more accurately close the water budget and quantify the energy cycle at various spatial and temporal scales. As discussed in the chapters that follow, we expect to see future advances in observational tools and applications (e.g., remote sensing, GIS, etc.). We may also expect more advanced modeling of hydrologic processes both at the catchment scale as well as scales that are intended to provide coupling with other components of Earth's systems (i.e., atmosphere, ocean, and biogeochemical processes). On the more applied side, the future requirements for adequate water supplies (quantity and quality) will demand further development of both deterministic and stochastic tools that take advantage of more advanced forms of observational, GIS, and computational techniques. These tools will provide prediction and simulation capabilities for assessing the ramifications of hydroclimatic scenarios (e.g., droughts and floods, regional groundwater depletions, existence and movement of contaminants in both surface water and groundwater, etc.).

Acknowledgments

We gratefully acknowledge the support provided by SAHRA (Sustainability of Semi-Arid Hydrology and Riparian Areas), an NSF Science and Technology Center at the University of Arizona, as well as the Global Energy and Water Cycle Experiment (GEWEX). Our sincere gratitude is also extended to Terri Hogue, Thomas Pagano, and Corrie Thies, for their thoughtful comments and editing of this document in its various stages of completion.

REFERENCES

Adler, R. F., A. J. Negri, P. R. Keehn, and I. M. Hakkarinen, Estimation of monthly rainfall over Japan and surrounding waters from a combination of low-orbit microwave and geosynchronous IR data, *J. Appl. Meteor.*, *32*, 335–356, 1993.

Arkin, P.A., and P. Xie, The global precipitation and climatology project: First algorithm intercomparison project, *Bull. Am. Meteor. Soc.*, *75*, 401–419, 1994.

Bouwer, H., and T. Maddock III, Making sense of the interactions between groundwater and streamflow: Lessons for water masters and adjudicators. *Rivers*, *6*(1), 19–31, 1997.

Chahine, M. T., The hydrological cycle and its influence on climate, *Nature*, *359* (Oct. 1), 373–380, 1992.

Chow, V. T., D. R. Maidment, and L. W. Mays, *Applied Hydrology*, McGraw-Hill Higher Education, New York, 1988.

Driscoll, F. G. (Ed.), *Groundwater and Wells*, Johnson Division, St. Paul, MN, 1986.

El Geriani, A. M., O. Essamin, P. J. A. Gijsbers, and D. P. Loucks, Cost-effectiveness analysis of Libya's water supply system, *J. Water Resourc. Plan. Manag.*, *124*(6), 320–329, 1998.

Fiering, M. B., *Streamflow Synthesis*, Harvard University Press, Cambridge, MA, 1967.

Fiering, M. B., Reservoir planning and operation, in H. W. Shen (Ed.), *Stochastic Approaches to Water Management*, Vol. 2, Ft. Collins, CO, 1976, pp. 17:1–17:21.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Gijsbers, P. J. A., and D. P. Loucks, Libya's choices: Desalinization or the great man-made river project. *Phys. Chem. Earth (B)*, *24*(4), 385–389, 1999.

Glennon, R. J., and T. Maddock III, The concept of capture: The hydrology and law of stream/aquifer interactions, in *Proceedings of the Forty-Third Annual Rocky Mountain Mineral Law Institute*, Denver, CO, 1997, Chapter 22.

Haan, C. T., *Statistical Methods in Hydrology*, Iowa State University Press, Ames, Iowa, 1977.

Horden, R. H., The hydrologic cycle, in R. W. Herschy and R. W. Fairbridge (Eds.), *Encyclopedia of Hydrology and Water Resources*, Kluwer Academic Publishers, Dordrecht, pp. 400–404, 803 p. The Netherlands, 1998.

Maddock III, T., and L. B. Vionnet, Groundwater capture processes under a seasonal variation in natural recharge and discharge, *Hydrogeol. J.*, *6*, 24–32, 1998.

Maidment, D. R., Hydrology, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 1.

National Research Council, *GCIP Global Energy and Water Cycle Experiment (GEWEX) Continental-Scale International Project: A review of progress and opportunities*, National Academy Press, Washington, DC, 1998.

Oki, T., K. Musiake, H. Matsuyama, and K. Masuda, Global atmospheric water balance and runoff from large river basins, *Hydrol. Proc.*, *9*, 655–678, 1995.

Oki, T., The global water cycle, in K. A. Browning and R. J. Gurney (Eds.), *Global Energy and Water Cycles*, Cambridge University Press, Cambridge, 1999.

Shuttleworth, W. J., Evaporation, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 4.

Sorooshian, S., K.-L. Hsu, X. Gao, H. Gupta, B. Imam, and D. Braithwai'e, Evaluation of PERSIANN system satellite-based estimates of tropical rainfall, *Bull. Am. Meteorol. Soc.*, *81*(9), 2035–2046, 2000.

Xie, P., and P. A. Arkin, A comparison of gauge observations and satellite estimates of monthly precipitation, *J. Appl. Meteor.*, *34*, 1143–1160, 1995.

Xie, P., and P. A. Arkin, Analysis of global monthly precipitation using gauge observations, satellite estimates and numerical model predictions, *J. Climate*, *9*, 840–858, 1996.

Xie, P., and P. A. Arkin, Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates and numerical model outputs, *Bull. Appl. Meteor.*, *34*, 1143–1160, 1997.

CHAPTER 24

RAINFALL

JAMES A. SMITH

The capability to measure rainfall advanced dramatically in the last quarter of the twentieth century. The advances have been paced by remote-sensing technologies including both ground-based weather radar and satellite-borne instruments. The most dramatic developments have centered around the capability to monitor precipitation globally from satellite sensors. This measurement capability provides a variety of avenues for hydroclimatological analysis and forecasting. Advances in ground-based radar technologies and deployment of dense networks of rain gages has enhanced the ability to measure rainfall at short time scales (less than 1 h) and small spatial scales (less than 1 km). These time and space scales are often most relevant for water management applications. A brief summary of rainfall measurement and analysis capabilities is presented in the following three sections and organized by the three principal measurement technologies: rain gage, radar, and satellite.

1 RAIN GAGES

Networks of rain gages play a key role in hydrologic applications ranging from flood forecasting to design of high-hazard structures and water supply management. A wide variety of recording and nonrecording rain gages are used for hydrologic applications. Review and discussion of rain gage technologies are presented in the work by Sumner (1988).

There exist several inherent sources of error that affect all types of rain gages. All rain gages suffer from errors due to modification of the wind field by the gage [see Robinson and Rodda (1969) for detailed discussions]. The magnitude of errors depends on wind speed, siting characteristics, and type of precipitation (Groisman

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

and Legates, 1994; Sevruk, 1982, 1989; Nystuen et al., 1996; Steiner et al., 1999; McCollum and Krajewski, 1998; Larson and Peck, 1974). Rain gage measurement of rainfall is especially difficult in a variety of settings, including mountain ridges, forests, and water bodies. Measurement errors for snow are typically much larger than for rain and are generally in the form of catch deficiencies (Groisman and Legates, 1994).

Rain gage networks serve as the basis for climatological assessments of precipitation that are used for a wide range of applications (see, e.g., Frei and Schaer, 1998). Three of the principal types of climatological analyses that are used for water management applications are illustrated in Figures 1 to 3. Assessments of average rainfall conditions, in a variety of forms, are central to activities involving industrial, municipal, and agricultural water use. Mean annual precipitation is shown in Figure 1 [see also Groisman and Legates (1994) for a discussion of biases in rain gage analyses of mean precipitation]. Global assessments of continental precipitation have been developed from rain gage observations by Legates (1987) [see also Legates and Wilmott (1990)]. Precipitation frequency analysis plays a central role in engineering design problems, especially in urban areas (Urbonas and Roesner, 1993). The 15-min, 100-year rainfall magnitude for the United States (Frederick et al., 1977) is illustrated in Figure 2. The network of gages that have the temporal resolution to provide short-term precipitation frequency analyses, such as those in Figure 2, is far less dense than the rain gage network used to produce mean annual precipitation maps. Consequently, it is difficult to assess the true geographic variability of extreme rainfall rates. It is likely that geographic features, such as mountains and land–water boundaries, exert a pronounced influence on the frequency of extreme rainfall rates. The density of the network, however, is not adequate to resolve these geographic variations. Design of high-hazard structures, such as spillways on major dams, is determined through probable maximum precipitation (PMP) analyses (Hansen, 1987; WMO, 1986). Rain gage data sets, in the form of storm catalogs, play a central role in PMP analyses. Storm catalogs for PMP analyses consist of gage observations from specific events. Consequently, the density of gage observations in regions experiencing catastrophic rainfall is critical for PMP analyses. The 6-h, 200 mi^2 PMP for the eastern United States is shown in Figure 3. The greatest uncertainties in PMP analyses are for small areas (less than 200 mi^2), short time periods (6 h and less), and for regions of complex terrain (National Research Council, 1994).

2 RADAR

Implementation of the NEXRAD (next-generation weather radar) system of WSR-88D (weather surveillance radar—1988 Doppler) radars has resulted in dramatic advances in rainfall measurement capabilities for the United States (Klazura and Imy, 1993). Operational National Weather Service (NWS) rainfall products derived from WSR-88D observations provide rainfall analyses for the United States at 1-h time resolution and spatial resolution of approximately 4 km (Hudlow et al., 1991).

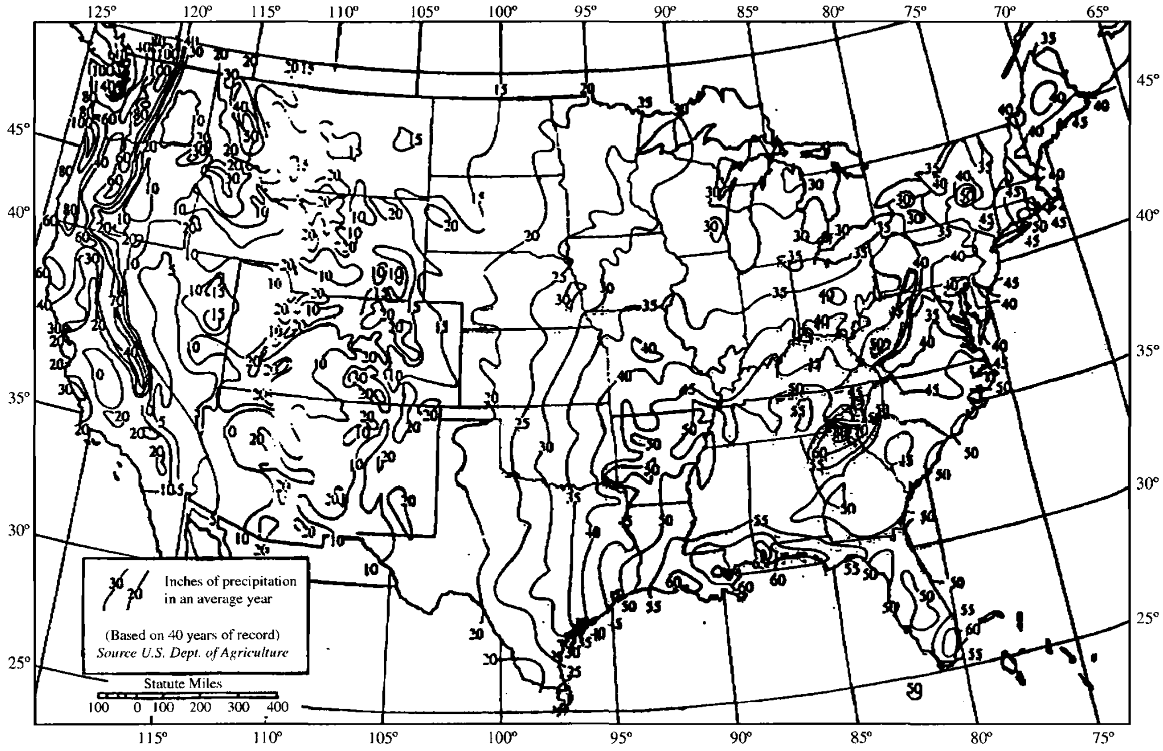


Figure 1 Mean annual precipitation (inches) for the United States from rain gage observations.

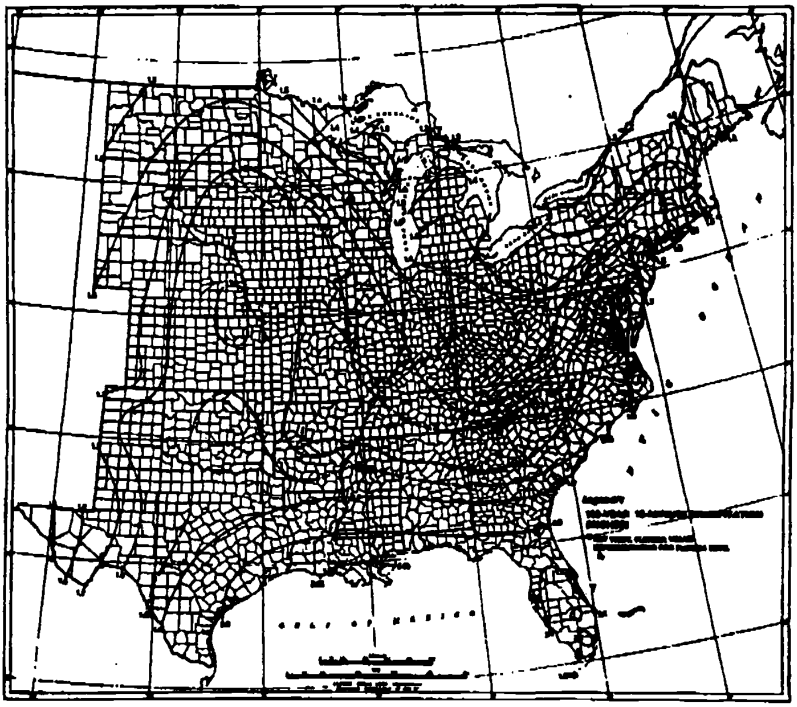


Figure 2 The 100-year, 15-min rainfall magnitudes (inches) for the United States east of the Rocky Mountains.

The hourly digital product (HDP) rainfall estimates are created by the WSR-88D radar product generator on a 131 × 131, 4-km grid centered at each radar site. The range over which rainfall products are constructed for each site is approximately 230 km. The algorithm used to construct this product (Fulton et al., 1998) consists of the following steps: (1) quality control, including identification and elimination of anomalous propagation returns, (2) conversion of radar reflectivity factor to rainfall rate through a Z-R relationship, (3) correction for range effects, (4) aggregation of rainfall estimates to hourly, 4-km grid scale, and (5) bias correction using rain gage observations. The HDP product is the base rainfall product from the NEXRAD system. Detailed assessments of HDP algorithm performance are presented in Smith et al. (1996b) and Baeck and Smith (1998) [see also Joss and Waldvogel (1989), Wilson and Brandes (1979), and Anagnostou and Krajewski (1998)].

In a second stage of WSR-88D rainfall processing, multisensor precipitation analyses employ rain gage observations and the 4-km HDP rainfall fields in an optimal estimation framework developed by Krajewski (1987) and Seo (1998a, 1998b). These rainfall fields are subsequently composited into a regional mosaic.

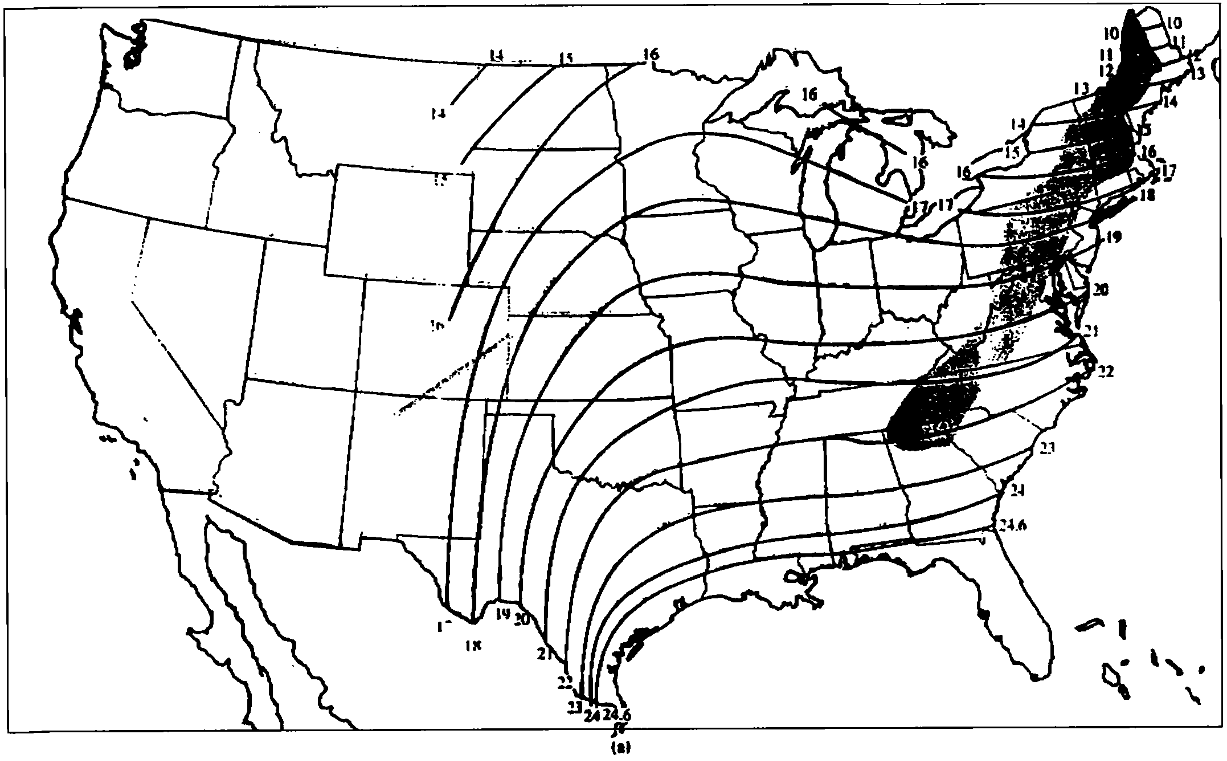


Figure 3 The 6-h, 200 mi^2 PMP magnitudes (inches) for the eastern United States.

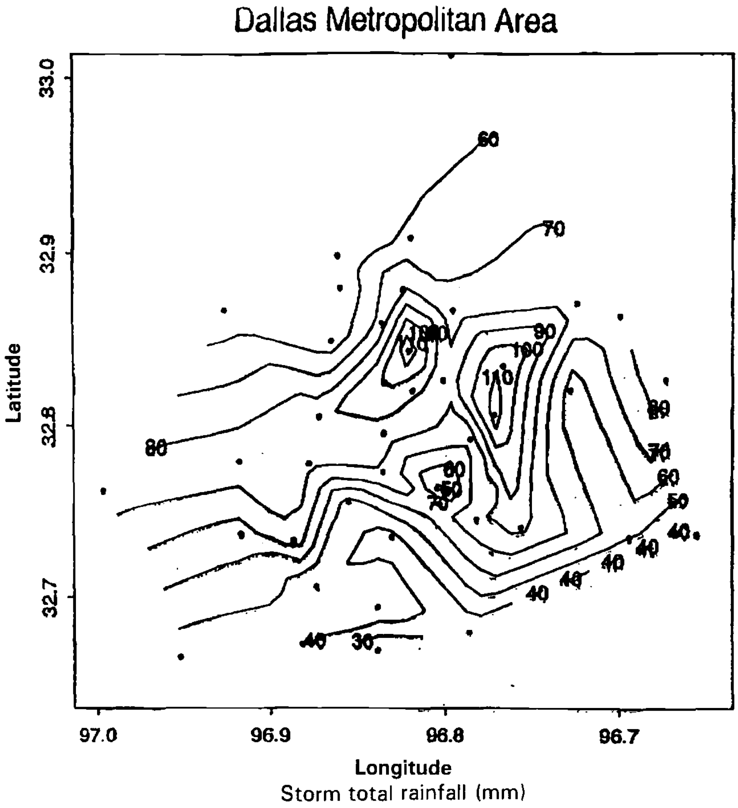


Figure 4 Storm total rainfall (mm) from the Dallas Metropolitan Area mesonet for the Dallas hailstorm of May 5, 1995. The dimensions of the surrounding box are approximately 30 × 30 km.

The regions that comprise the individual mosaics correspond to the watershed boundaries that delimit the NWS River Forecast Center areas of coverage. Algorithms used for mosaicking of multiple, overlapping radar coverages are described in Seo et al. (1998). A national, hourly precipitation analysis is produced at the National Centers for Environmental Prediction (NCEP).

The national 4-km, hourly rainfall mosaic produced by NWS from rain gage and WSR-88D rainfall products will provide an important source of rainfall information for climatological analyses, especially as the observing period increases. Radar observations have not generally served as the basis for climatological analyses of rainfall [see, however, Baeck and Smith (1995) for an exception]. Issues of bias in radar rainfall estimation must be addressed for radar-based rainfall databases to be most useful for climatological studies (Smith et al., 1996b).

Radar polarimetric measurements (Zrnic, 1996; Zrnic and Ryzhkov, 1996; Ryzhkov and Zrnic, 1996; Aydin et al., 1995), which utilize the capability of radar to transmit and receive electromagnetic radiation at alternating polarization, hold promise for providing significant improvements in rainfall estimates. Polarization measurements have been shown to be quite useful for quality control algorithms, including detection of bright band, hail, and AP [anomalous propagation (of radar waves, due to sharp gradients of water and air density)], as well as for algorithms for estimating rainfall rate (Peterson et al., 1999; Zrnic, 1996). The NEXRAD network was designed for eventual implementation of polarization measurements by the WSR-88D.

Radar has provided a significant component of the observational basis for studying storms that produce extreme rainfall. Chappell (1989) and Doswell et al. (1996) summarize key elements of heavy rainfall producing storms with particular emphasis on storms that produce large point rainfall accumulations through small net storm motion [see also Maddox et al. (1979)]. These storms have been termed quasi-stationary convective systems (Chappell, 1989). Houze et al. (1990) provide a detailed summary of radar-derived storm structure for severe thunderstorms in the central United States [see also Perica and Foufoula-Georgiou (1996) and Steiner et al. (1995)].

WSR-88D observations, and the rainfall products derived from these observations, have provided a new playing field for hydrologic application and science. Many hydrologic problems that were previously not possible to address due to an absence of information concerning rainfall, have been attacked from an observational perspective. Numerous examples can be drawn from flood hydrology. Figure 5 illustrates a storm total rainfall analysis constructed for the rapidan storm of June 27, 1995 (Smith et al., 1996a). More than 600 mm of rain fell on the east slope of the Virginia Blue Ridge during a 12-h period resulting in record unit discharge for the United States east of the Mississippi River and catastrophic landslides and debris flows. Fluvial and geomorphic impacts of the rapidan storm rival those described in the classic study by Hack and Goodlett (1960) for the June 1949 storm in the Shenandoah Mountains. The chief difference between studies of the 1949 and 1995 storms is rainfall measurement at the 1-km horizontal scale and 6-min time scale for the 1995 storm that allows direct assessment of hydrologic processes.

3 SATELLITE

Satellite-borne instruments have proven useful for monitoring precipitating cloud system since the 1960s. Steady progress has been made in developing algorithms for retrieving rainfall accumulations from passive satellite observations in the microwave (Negri et al., 1994; Adler et al., 1994) and infrared (Vicente and Scofield, 1997; Huffman et al., 1995; Adler and Negri, 1988) portions of the electromagnetic spectrum. This progress is reflected in rapidly advancing capabilities for hydroclimatological analysis (Kummerow et al., 2000; Adler et al., 2000; Huffman et

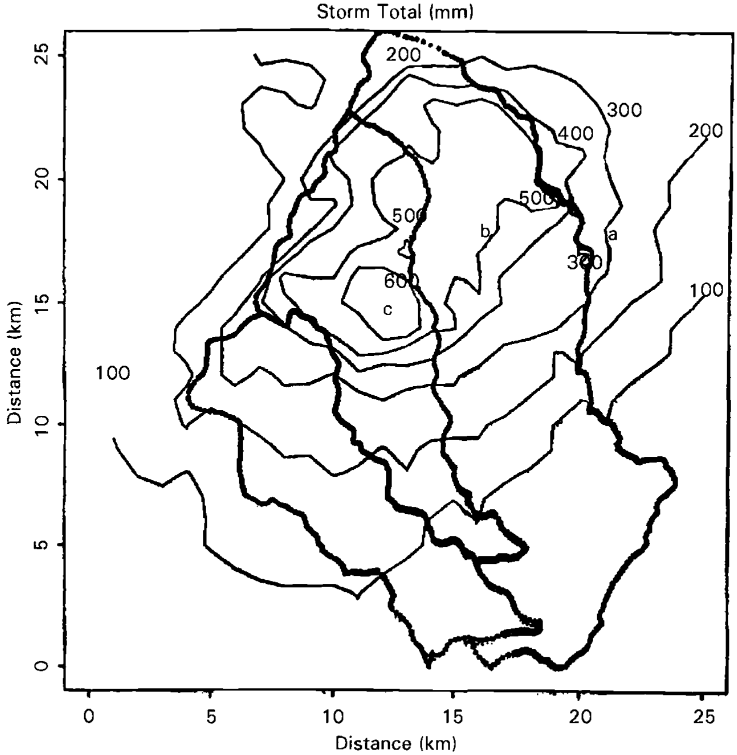


Figure 5 Storm total rainfall (mm) from the Sterling, Virginia, WSR-88D for the rapidan storm of June 27, 1995. The basin boundary for the 295-km^2 watershed and boundaries for 3 subwatersheds are shown in solid lines.

al., 1997, 2001; Krajewski et al., 2000). Techniques for quantitative precipitation estimation from satellite sensors are reviewed below.

A geostationary, infrared-based satellite algorithm (Vicente and Scofield, 1997) has been developed and implemented for heavy rainfall measurement. This algorithm is specifically designed for deep, moist convective systems. Estimated precipitation rates are based on the cloud top temperature obtained from the 10.7-μm infrared channel. The empirical equations used to relate cloud top temperature and rainfall rate were calibrated from radar data sets consisting of observations from thunderstorm systems. A moisture correction factor obtained from the precipitable water and mean relative humidity is used to adjust the estimates for different moist environments. The technique of relating rain rate and cloud top temperature tends to overestimate the rain area in some cases and rain rate in others. The technique is also subject to underestimation of rain rates in warm cloud top environments and overestimation of cold top storms in strong wind shear environments.

The *Tropical Rainfall Measuring Mission* (*TRMM*) satellite (Simpson et al., 1988) is designed to measure tropical precipitation and its variation. With the inclusion of a precipitation radar, TRMM provides the first opportunity to estimate the vertical profile of the latent heat that is released through condensation. The TRMM rainfall data will be particularly important for studies of the global hydrological cycle and for testing the ability of climate models to simulate climate accurately on the seasonal time scale.

The TRMM instruments for rainfall observation consist of a precipitation radar, a multifrequency microwave radiometer, and a visible and infrared (VIS/IR) radiometer. The precipitation radar provides measurements of the three-dimensional rainfall distribution over both land and ocean. The precipitation radar will permit the measurement of rain over land where passive microwave channels have difficulty. The horizontal resolution is approximately 4 km, the range resolution is 250 m, and the scanning swath width is 220 km. The multichannel microwave radiometer provides information on vertically integrated precipitation, its areal distribution, and its intensity. Rainfall analyses using the microwave radiometer are best suited for open ocean conditions. The visible infrared (IR) scanner provides high-resolution information on cloud coverage, type, and cloud top temperatures and serves as the link between these data and the long and virtually continuous coverage by the geosynchronous meteorological satellites. The instrument, with a swath width of 720 km, will provide cloud distributions by type and height and rain estimates from brightness temperatures at a horizontal resolution of approximately 2 km.

Satellite IR observations from geostationary satellites have been used extensively for assessing the climatology of extreme rainfall producing storms. An extensive climatology has been developed for mesoscale convective complexes (Maddox, 1980) based on IR-based assessments of cloud properties. Numerous studies have examined the links between mesoscale convective complexes (MCCs), and the more general category of mesoscale convective systems, and heavy rainfall [see Houze (1993)].

REFERENCES

Adler, R. F., and A. J. Negri, A satellite infrared technique to estimate tropical convective and stratiform rainfall, *J. Appl. Meteor.*, *27*, 30–51, 1988.

Adler, R. F., G. J. Huffman, and P. R. Keehn, Global tropical rain estimates from microwave-adjusted geosynchronous IR data, *Remote Sensing Rev.*, *11*, 125–152, 1994.

Adler, R. F., G. J. Huffman, D. T. Bolvin, S. Curtis, and E. J. Nelkin, Tropical rainfall distributions determined using TRMM combined with other satellite and rain gauge information, *J. Appl. Meteor.*, *39*(12), 2007–2023, 2000.

Anagnostou, E., and W. F. Krajewski, Calibration of the WSR-88D precipitation processing

Aydin, K., V. N. Bringi, and L. Liu, Rain rate estimation in the presence of hail using S-band specific differential phase and other radar parameters, *J. Appl. Meteor.*, *34*, 404–410, 1995.

Baeck, M. L., and J. A. Smith, Climatological analysis of manually digitized radar data for the United States, *Water Resour. Res.*, *31*(12), 3033–3049, 1995.

Baeck, M. L., and J. A. Smith, Rainfall estimation by the WSR-88D for heavy rainfall events, *Weather Forecast.*, *13*, 413–436, 1998.

Barros, A. P., and D. P. Lettenmaier, Dynamic modeling of orographically induced precipitation, *Rev. Geophys.*, *32*(3), 265–284, 1994.

Chappell, C., Quasistationary convective events, in P. Ray (Ed.), *Mesoscale Meteorology*, American Meteorological Society, Boston, 1989.

Doswell III, C. A., H. E. Brooks, and R. A. Maddox, Flash flood forecasting: An ingredients-based methodology, *Weather Forecast.*, *11*(4), 560–581, 1996.

Frederick, R. H., V. A. Myers, and E. P. Auciello, Five- to 60-minute precipitation frequency for the eastern and central United States, NOAA Technical Memo, NWS Hydro-35, June 1977.

Frei, C., and C. Schaer, A precipitation climatology of the Alps from high-resolution raingauge observations, *Int. J. Climatol.*, *18*(8), 873–900, 1998.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, and D. A. Miller, The WSR-88D rainfall algorithm, *Weather Forecast.*, *13*(2), 377–395, 1998.

Groisman, P. Y., and D. R. Legates, The accuracy of United States precipitation data, *Bull. Am. Meteor. Soc.*, *75*(2), 215–227, 1994.

Hack, J., and J. Goodlett, Geomorphology and forest ecology of a mountain region in the central Appalachians, *U.S. Geological Survey Professional Paper 347*, 1960.

Hansen, M., Probable maximum precipitation for design floods in the United States, *J. Hydrol.*, *96*, 267–278, 1987.

Houze, R. A., *Cloud Dynamics*, Academic, New York, 1993.

Houze, R. A., B. F. Smull, and P. Dodge, Mesoscale organization of springtime rainstorms in Oklahoma, *Monthly Weather Rev.*, *118*, 613–654, 1990.

Hudlow, M. D., J. A. Smith, M. L. Walton, and R. C. Shedd, NEXRAD—New era in hydrometeorology, in I. Cluckie and C. Collier (Eds.), *Hydrological Applications of Weather Radar*, Ellis-Norwood, New York, 1991, pp. 602–612.

Huffman, G. J., R. F. Adler, B. Rudolf, U. Schneider, and P. R. Keehn, A technique for combining satellite-based estimates raingauge analysis, and NWP model precipitation information into global precipitation estimates, *J. Climate*, *8*(5), 1284–1295, 1995.

Huffman, G. J., R. F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider, The Global Precipitation Climatology Project (GPCP) combined precipitation dataset, *Bull. Am. Meteor. Soc.*, *78*(1), 5–20, 1997.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, Global precipitation at one-degree daily resolution from multisatellite observations, *J. Hydrometeor.*, *2*(1), 36–50, 2001.

Joss, J., and A. Waldvogel, Precipitation measurement and hydrology: A review, in *Battan Memorial and Radar Conference*, David Atlas, editor, American Meteorological Society, Boston, 1989.

Klazura, G. E., and D. A. Imy, A description of the initial set of analysis products available from the NEXRAD WSR-88D System, *Bull. Am. Meteor. Soc.*, *74*, 1293–1311, 1993.

Krajewski, W. F., Co-kriging of radar and rain gage data, *J. Geophys. Res.*, *92*(D8), 9571–9580, 1987.

Krajewski, W. F., G. J. Ciach, and J. R. McCollum, Initial validation of the global precipitation climatology project monthly rainfall over the United States, *J. Appl. Meteorol.*, *39*(7), 1071–1086, 2000.

Kummerow, C., J. Simpson, O. Thiele, W. Barnes, A. T. C. Chang, E. Stocker, R. F. Adler, A. Hou, R. Kakar, F. Wentz, P. Ashcroft, T. Kozu, Y. Hong, K. Okamoto, T. Iguchi, H. Kuroiwa, E. Im, Z. Haddad, G. Huffman, B. Ferrier, W. S. Olson, E. Zipser, E. A. Smith, T. T. Wilheit, and G. North, The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit, *J. Appl. Meteor.*, *39*(12), 1965–1982, 2000.

Larson, L., and E. Peck, Accuracy of precipitation measurements for hydrologic modeling, *Water Resour. Res.*, *10*(4), 857–863, 1974.

Legates, D. R., A climatology of global precipitation, *Climatology*, *40*(1), 1987.

Legates, D. R., and C. J. Willmott, Mean Seasonal and Spatial Variability in Gauge-Corrected, Global Precipitation, *Int. J. Climatol.*, *10*(2), 111–127, 1990.

Maddox, R. A., Mesoscale convective complexes, *Bull. Am. Meteor. Soc.*, *61*(11), 1374–1387, 1980.

Maddox, R., C. Chappell, and L. Hoxit, Synoptic and meso-α scale aspects of flash flood events, *Bull. Am. Meteor. Soc.*, *60*(2), 115–123, 1979.

McCollum, J. R., and W. F. Krajewski, Uncertainty of monthly rainfall estimates from rain gauges in the Global Precipitation Climatology Project, *Water Resour. Res.*, *34*(10), 2647–2654, 1998.

National Research Council, *Estimating Bounds on Extreme Precipitation Events*, National Academy Press, Washington DC, 1994.

Negri, A. J., R. F. Adler, E. J. Nelkin, and G. J. Huffman, Regional rainfall climatologies derived from special sensor microwave imager (SSM/I) data, *Bull. Am. Meteor. Soc.*, *75*, 1165–1182, 1994.

Nystuen, J. A., J. R. Proni, P. G. Black, and J. C. Wilkerson, A comparison of automatic rain gauges, *J. Atmos. Oceanic Technol.*, *13*, 62–73, 1996.

Perica, S., and E. Foufoula-Georgiou, Linkage of scaling and thermodynamic parameters of rainfall: Results from midlatitude mesoscale convective systems, *J. Geophys. Res.*, *101*(D3), 7431–7448, 1996.

Petersen, W. A., L. D. Carey, S. A. Rutledge, J. C. Knievel, N. J. Doesken, R. H. Johnson, T. B. McKee, T. Vonder Haar, and J. F. Weaver, Mesoscale and radar observations of the Fort Collins flash flood of 28 July 1997, *Bull. Am. Meteor. Soc.*, *80*(2), 191–216, 1999.

Robinson, A. C., and J. C. Rodda, Rain, wind and the aerodynamic characteristics of raingauges, *Met. Mat.*, *98*, 113–120, 1969.

Ryzhkov, A. V., and D. S. Zrnic, Assessment of rainfall measurement that uses specific differential phase, *J. Appl. Meteorol.*, *35*(11), 2080–2090, 1996.

Seo, D.-J., Optimal estimation of rainfall fields using radar and rain gage data, *J. Hydrol.*, *208*(1,2), pp. 25–36, 1998.

Seo, D.-J., Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions, *J. Hydrol.*, *208*(1,2), pp. 37–52, 1998.

Seo, D. J., J. Breidenbach, and E. Johnson, Real-time estimation of mean field bias in radar rainfall data, *J. Hydrol.*, *223*(3,4), pp. 131–147, 1999.

Seo, D. J., R. Fulton, and J. Breidenbach, Rainfall estimation in the WSR-88D era for operational hydrologic forecasting in the National Weather Service, Symposium on Hydrology, Phoenix, American Meteorological Society, 1998, pp. J60–J62.

Sevruk, B., Methods of correction for systematic error in point precipitation measurement for operational use, WMO Publication No. 589, OHR No. 21, 1982.

Sevruk, B., Wind-induced measurement error for high intensity rains, Proceedings, WMO/IAHS International Workshop on Precipitation Measurements, St. Moritz, Switzerland, 1989, pp. 199–204.

Simpson, J., R. F. Adler, and G. North, A proposed tropical rainfall measuring mission (TRMM), *Bull. Am. Meteor. Soc.*, *69*, 278–295, 1988.

Smith, J. A., M. L. Baeck, M. Steiner, and A. J. Miller, Catastrophic rainfall from an upslope thunderstorm in the Central Appalachians: the Rapidan Storm of June 27, 1995, *Water Resour. Res.*, *32*(10), 3099–3113, 1996a.

Smith, J. A., D.-J. Seo, M. L. Baeck, and M. D. Hudlow, An intercomparison study of NEXRAD precipitation estimates, *Water Resour. Res.*, *32*(7), 2035–2045, 1996b.

Steiner, M., R. A. Houze, Jr., and S. E. Yuter, Climatological characterization of three-dimensional storm structure from operational radar and raingage data, *J. Appl. Meteor.*, *34*, 1978–2007, 1995.

Steiner, M., J. A. Smith, S. J. Burges, C. V. Alonso, and R. W. Darden, Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation, *Water Resour. Res.*, *35*(8), 2487–2503, 1999.

Sumner, G., Precipitation measurement and observation, in *Precipitation: Process and Analysis*, Wiley, Chichester, 1988, Chapter 7.

Vicente, G. A., and R. A. Scofield, Real time rainfall rate estimates derived from the GOES-8/9 satellites for flash flood forecasting, numerical modeling and operational hydrology, Preprints, 13th Conference on Hydrology, Long Beach, CA, American Meteorological Society, 1997, pp. J115–J118.

Urbonas, B. R., and L. A. Roesner, Hydrologic design for urban drainage and flood control, in D. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1992.

World Meteorological Organization (WMO), *Manual for estimation of Probable Maximum Precipitation*, 2nd ed., WMO-No. 332, WMO, Geneva, Switzerland, 1986.

Wilson, J. W., and E. A. Brandes, Radar measurement of rainfall—a summary, *Bull. Am. Meteor. Soc.*, *60*, 1048–1058, 1979.

Zrnic, D. S., Weather radar polarimetry—trends toward operational applications, *Bull. Am. Meteorol. Soc.*, *77*(7), 1529–1534, 1996.

Zrnic, D. S., and A. V. Ryzhkov, Advantages of rain measurements using specific differential phase, *J. Atmos. Ocean. Tech.*, *13*(2), 454–464, 1996.

CHAPTER 25

SNOW HYDROLOGY AND WATER RESOURCES (WESTERN UNITED STATES)

ROGER C. BALES AND DON CLINE

1 INTRODUCTION

Seasonally snow-covered areas of Earth offer special challenges for water resources management, challenges that arise from both hydrologic and social factors. Seasonal snowpacks account for the major source of the runoff for streamflow and groundwater recharge over wide areas of the midlatitudes. For example, in the western United States over 85% of the annual runoff from the Colorado River basin originates as snowmelt. Most of this is from a few small source areas in four western states, mostly above 2700 m, which comprise only 12% of the basin area. Globally, snowmelt runoff from Earth's mountains fills the rivers and recharges the aquifers that over a billion people depend on for their water resources. Future climate variability and change are expected to result in major changes in the partitioning of snow and rainfall and the timing of snowmelt, which will have important implications for water use and resource management in these regions. It is therefore important to understand the processes controlling snowmelt runoff for both water resources as well as other resource management purposes.

2 CURRENT HYDROCLIMATIC CONDITIONS IN THE WESTERN UNITED STATES

Much of the variability in snow cover found in Earth's seasonally snow-covered regions can be found in the relatively well-studied western United States, which

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

has regions ranging from the high-precipitation Pacific Northwest coast to the semi-arid Southwest. The area from the Rocky Mountains to the Pacific Coast can be conveniently divided into seven regions with different hydroclimatic regimes (Fig. 1). Throughout the region, much of the annual streamflow is directly attributable to springtime melting of snow accumulation from the previous winter; however, there are also lower-elevation areas within the region that experience snowmelt throughout the winter and spring. Extended winter snowpack water storage in alpine areas, with often gradual melt rates, results in annual hydrographs having rising limbs of characteristically low slope, usually superimposed with small diel fluctuations reflecting daily melt cycles (Fig. 2). Beyond this basic similarity, however, wide differences in the source, delivery, and amount of moisture each region receives, the amount of water typically stored in their snowpacks, and the rate of release of that water result in different streamflow regimes between regions.

The climate of the western United States is dominated by large-scale atmospheric circulation originating over the north Pacific. In winter, the Pacific/North American (PNA) anomaly pattern forms a series of pressure centers of alternating sign stretching across the Pacific into southern Canada and down toward the Gulf of Mexico (Lin et al., 1990). The PNA typically forms a series of dry longwave ridges and wet troughs across North America with cold polar air masses to the north of the frontal boundary and warm subtropical air masses to the south. The position of the ridges and troughs influences the seasonal moisture cycle. The mountain ranges of the western United States, like all major mountain ranges, strongly influence global atmospheric circulation; in doing so, they affect the seasonal moisture cycle.

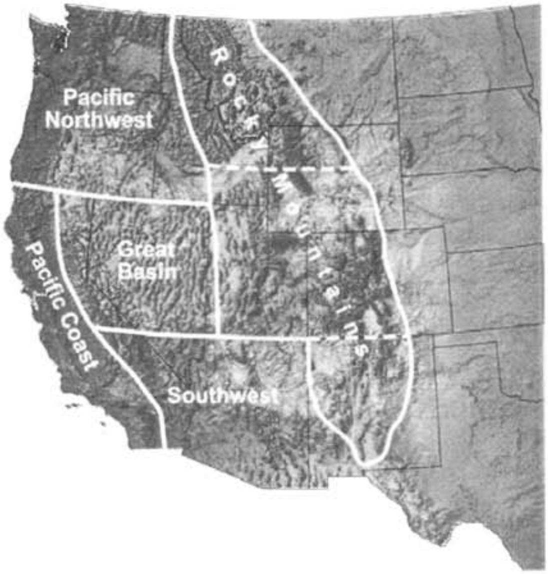


Figure 1 Seven hydroclimatic regimes of the western United States [after Paulson et al. (1991)].

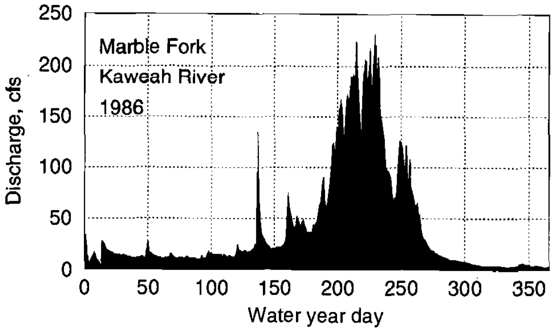


Figure 2 Discharge hydrograph for 19 km^2 Marble Fork of Kaweah River, southern Sierra.

The major source of moisture for all of the western United States is the Pacific Ocean; during fall and winter, orographic lifting and cooling of Pacific air masses laden with moisture results in precipitation either as rain or snow. The coast ranges, Cascades and Sierra Nevada, form a major orographic barrier for the Pacific moisture, causing much of the winter precipitation to fall as rain on the western side of the mountains. Winter precipitation on the eastern side of the Cascades and Sierra Nevada, although less, generally falls as snow in higher elevations. Relatively warm winter temperatures usually result in warm, wet snowpacks that often are nearly isothermal, and susceptible to rapid melting from warm temperatures and rainfall.

In the spring and summer, moisture from the Gulf of Mexico and subtropical Atlantic Ocean becomes important in most of the western states, with the exception of the coastal states. Early spring Gulf and subtropical Atlantic moisture often precipitates as snow, especially at higher elevations. Additional summertime moisture is provided in the southwestern states by subtropical Pacific air masses. With the exception of "land-recycled" moisture from land surface evapotranspiration (Paulson, 1991), these are the three sources of moisture that provide the western United States with precipitation and runoff. Peak snow accumulation and snowpack water storage in most of the region is found at higher elevations and generally occurs in March or April with snowmelt runoff occurring through May to July, depending on elevation and latitude.

Frontal activity associated with low-pressure systems is responsible for much of the winter precipitation in the northern Rocky Mountains, and upslope transport of moisture from east to west is important at lower elevations on the eastern side of the mountains. Summer precipitation, much of which ends up as evapotranspiration in the semiarid western United States, is mostly influenced by convective activity. However, snowpack storage serves as the major water supply for the summer months. The continentality of the northern Rocky Mountain region leads to cold, dry snowpacks. Significant energy is required to raise the temperature of the snowpack to the isothermal and melting stage; as a result the snowpack tends to remain well into spring. Rainfall generally does not contribute sufficient energy to drive snowmelt, until perhaps very late in the season.

High elevations in the central Rocky Mountains receive most of this region's annual precipitation as winter snowfall. Pacific frontal systems bringing most of the winter moisture to this region can arrive from the west, northwest, or southwest, and this influences the distribution of precipitation. Westerly tracks are orographically lifted to some extent by the Wasatch Plateau in Utah and are lifted further by the ranges along the Continental Divide in central Colorado, resulting in the heaviest precipitation west of the Continental Divide. Northwesterly tracks are lifted by the Wasatch Range and the Uinta Mountains in Utah and by the ranges along the Divide in north central Colorado, resulting in heavier precipitation at these locations. Storm tracks arriving from the southwest do not encounter major orographic effects until they reach the San Juan Mountains in southwestern Colorado, resulting in typically heavy winter precipitation in this part of the region from these storm tracks. In general, precipitation declines markedly throughout areas east of the Continental Divide. However, low-pressure systems east of the Divide can bring significant moisture in from the Gulf of Mexico during spring, resulting in sometimes heavy snowfall in the foothills at lower elevations on the eastern side of the Divide. Lower elevation areas of the central Rockies receive considerably less precipitation; most of the region's snowpack storage is concentrated in the higher mountains.

3 MEASUREMENT AND ESTIMATION OF SNOW PROPERTIES

Historical Background

Undoubtedly, the main recurring question in snow hydrology in the western United States is: How much snow is out there? Water resources managers forecast the amount of seasonal runoff, based in part on estimates of the amount of snow accumulation, or snow water equivalent (SWE), across a watershed or region and in part on forecasts of future precipitation. Estimates of SWE and snow-covered area (SCA) are used for a variety of purposes that are vital to the economy of a region, including: reservoir management, snow load maps, annual precipitation maps (for planning), drought monitoring, fish and game management, recreation (e.g., skiing, river trips), acid precipitation monitoring, and avalanche forecasting. (See Section 5 by Doesken.)

Historically, the Natural Resource Conservation Service (NRCS) has been charged with coordinating snow surveys, or point measurements of SWE. It also prepares seasonal water supply outlooks in the western United States. Predictions of water availability in the western United States are made by inventorying snowpacks in winter and early spring using measurements at over 2000 snow courses, including about 1000 snowpack telemetry (SNOTEL) sites that provide continuous data. The remaining sites are manual and are visited monthly. Empirical relationships between these observations and measured streamflow are used to forecast streamflow at over 500 points. In California, the California Cooperative Snow Survey (CCSS) coordinates measurements; it depends on 40 cooperating agencies for data collection. CCSS makes seasonal water supply forecasts, as do many program cooperators; weekly updates are made for major streams (Hart and Gehrke, 1990).

Estimation of the spatial distribution of SWE is challenging because of the many factors that affect its distribution and the small correlation length of the SWE spatial distribution. Topographic heterogeneity and variability in precipitation patterns also present problems in accurately determining the time of maximum accumulation. The simplicity of regression models makes them an attractive means of estimating SWE because of the large amount of work required to directly measure SWE on the catchment scale.

Estimates of the spatial distribution of SWE for the western United States come primarily from two sources. Operationally, the National Weather Service's (NWS's) National Operational Hydrologic Remote Sensing Center (NOHRSC) assimilates in situ and airborne snow survey data with satellite observations of snow cover to estimate SWE distributions throughout the winter and spring (Figs. 3*a* and 3*b*). Second, atmospheric models estimate the distribution of SWE based on modeled snowfall and surface energy balance parameters. The NWS SWE estimates are based on an interpolation procedure called snow estimation and updating system (SEUS), which interpolates between observed points to produce gridded SWE estimates. Point SWE estimates come from the NRCS sites and remote sensing. The NOHRSC conducts airborne SWE surveys along 1800 flight lines throughout the United States, many of which are located in the west (Fig. 4). Water contained in the snowpack attenuates terrestrial gamma radiation. The attenuation is measured relative to snow-free conditions to estimate SWE.

Remote Sensing

Remote sensing provides important spatial information about snow that can be used to improve the accuracy and timeliness of hydrologic forecasts for seasonally snow-covered areas, with commensurate gains in water resources management. At present, the only remotely sensed snow information used in operational hydrologic forecasting is the areal extent of snow cover (Fig. 3*a*). Over the past decade there has been an expanded development of remote sensing as a tool for determining other snow properties, which can be used to assist in estimating snow distributions and snowmelt runoff. There has also been a move toward development of physically based snowmelt models to use with this emerging data, particularly for alpine areas. The coupling of remote sensing and physically based approaches will enable making not only more accurate basin-scale forecasts but will also provide spatially distributed estimates of snowmelt.

The possibilities for detecting snowpack properties are largely determined by the wavelength being recorded by the remote-sensing instrument. Visible and near-infrared wavelengths, because they do not penetrate far into the snowpack, mainly provide information about the surface of the snowpack (e.g., snow-covered area, grain size, and albedo). However, microwave wavelengths can penetrate the snow-

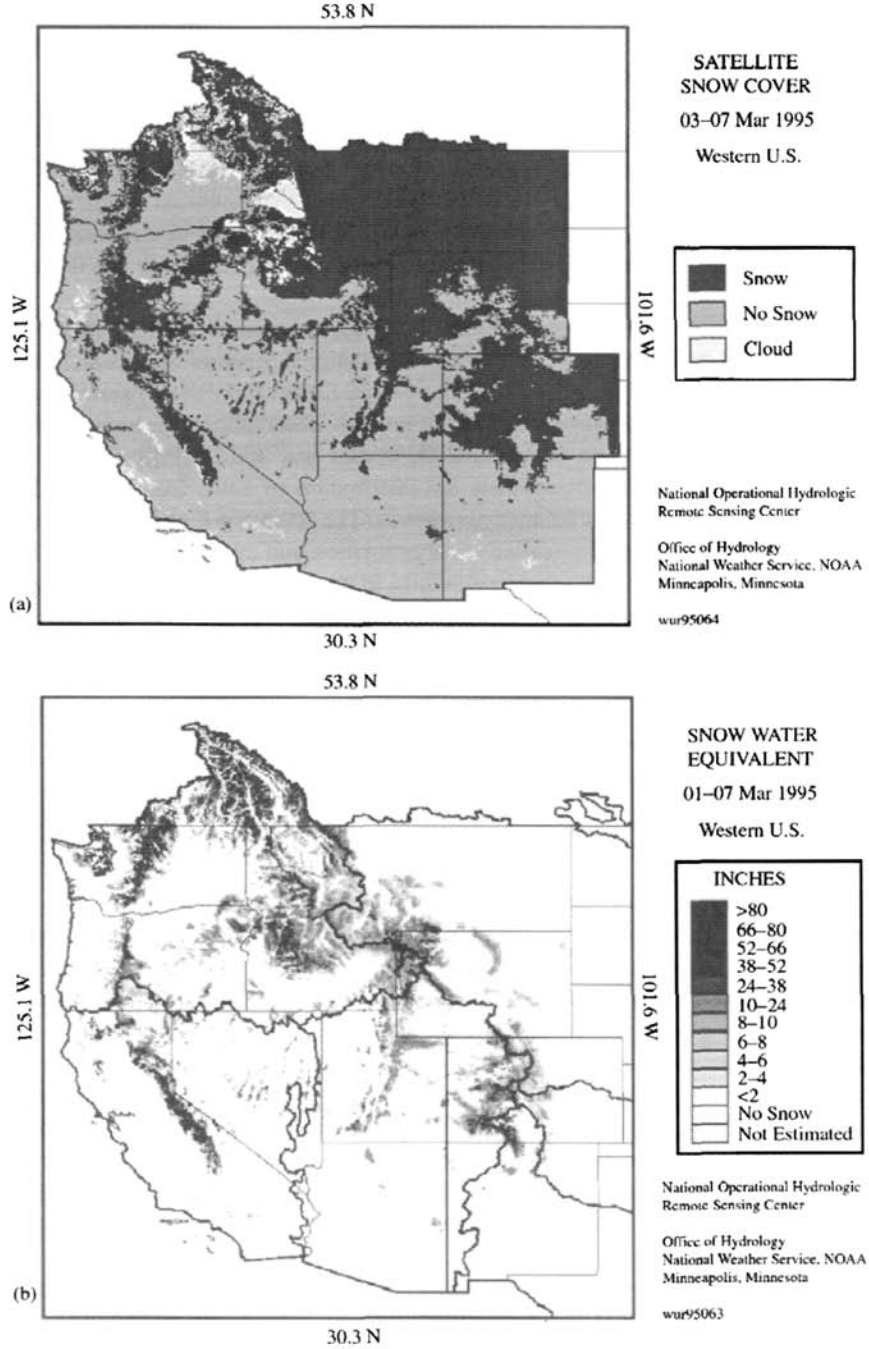


Figure 3 Operational snow products from the National Weather Service: (*a*) Satellite snow cover for March 1995 and (*b*) snow water equivalence for March 1995.

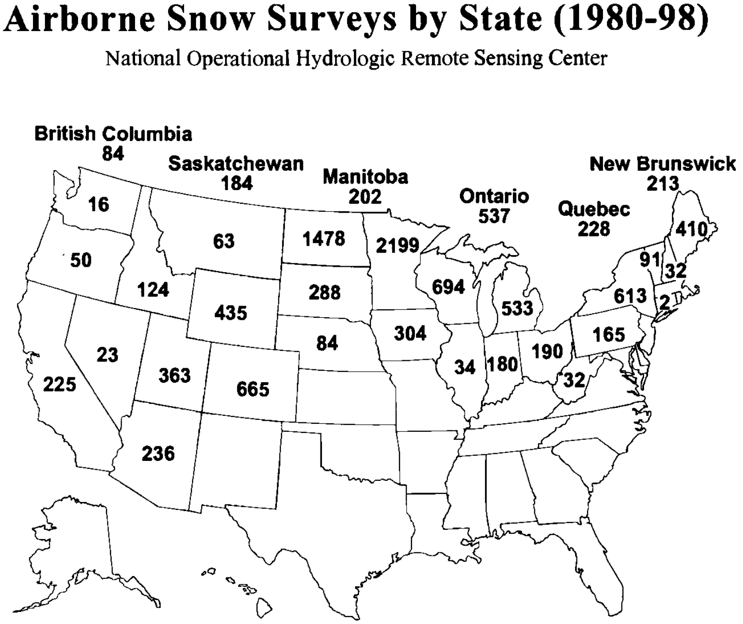


Figure 4 Number of airborne snow survey flight lines by state (1980–1998). National Operational Hydrologic Remote Sensing Center (NOHRSC).

pack, thereby providing an opportunity to collect volume integrated data (e.g., SWE).

Because of the difficulty of making field measurements in snow-covered mountainous regions, remote sensing has been pursued as a means of measuring snowcover properties. The National Oceanic and Atmospheric Administration's (NOAA's) advanced very-high resolution radiometer (AVHRR) data have been routinely used for classification of snow-covered versus snow-free area (Matson et al., 1986; Matson, 1991; Xu et al., 1993). Differences between the spatial, temporal, spectral, and radiometric resolutions of different remote-sensing instruments result in tradeoffs between instruments for hydrologic applications. Optimization of one type of resolution generally involves some sacrifice in other types of resolution. For example, the *Landsat* thematic mapper (TM) has a much better spatial resolution than the AVHRR (30 m versus 1 km pixel size, respectively); however, the AVHRR can provide daily coverage of a given point, whereas the TM can only provide biweekly coverage. Development of accurate snow-cover information for areas with steep, variable topography characteristic of the western United States requires higher resolution data than are currently available from operational remote-sensing instruments or improved processing of the current data.

Passive microwave sensors are used to monitor snow, but three problems have limited its application. First, uncertainty in snow texture results in a significant noise in the calibration of a brightness temperature index with measured snow properties. Second, passive microwave imagery has a large pixel size that results in significant mixed pixels over areas with forest, mountains, or lakes. It appears unlikely that snow properties could be "unmixed" in these situations. This restricts operational use to large flat areas such as prairies and tundra. Finally, the signature from snow is indistinguishable from bare ground when the snow is wet, which requires special processing of time series and inference.

For mapping of snow properties of greatest hydrological importance, a synthetic aperture radar (SAR) with some special characteristics is necessary. A SAR is sensitive to many snow parameters such as snow density, depth grain size, free-liquid water content, and snow-pack structures that hydrologists use. It can image day or night in all weather, and it has a fine spatial resolution compatible with topographic variation affecting snow distribution. Experiments as part of the SIR-C/X SAR missions have significantly advanced and demonstrated the capabilities of new multifrequency and multipolarization SAR—to map both wet and dry snow covers (Shi and Dozier, 1997), to infer snow wetness (Shi and Dozier, 1995), and to estimate snow density and depth and thereby snow water equivalent (Shi and Dozier, 1996).

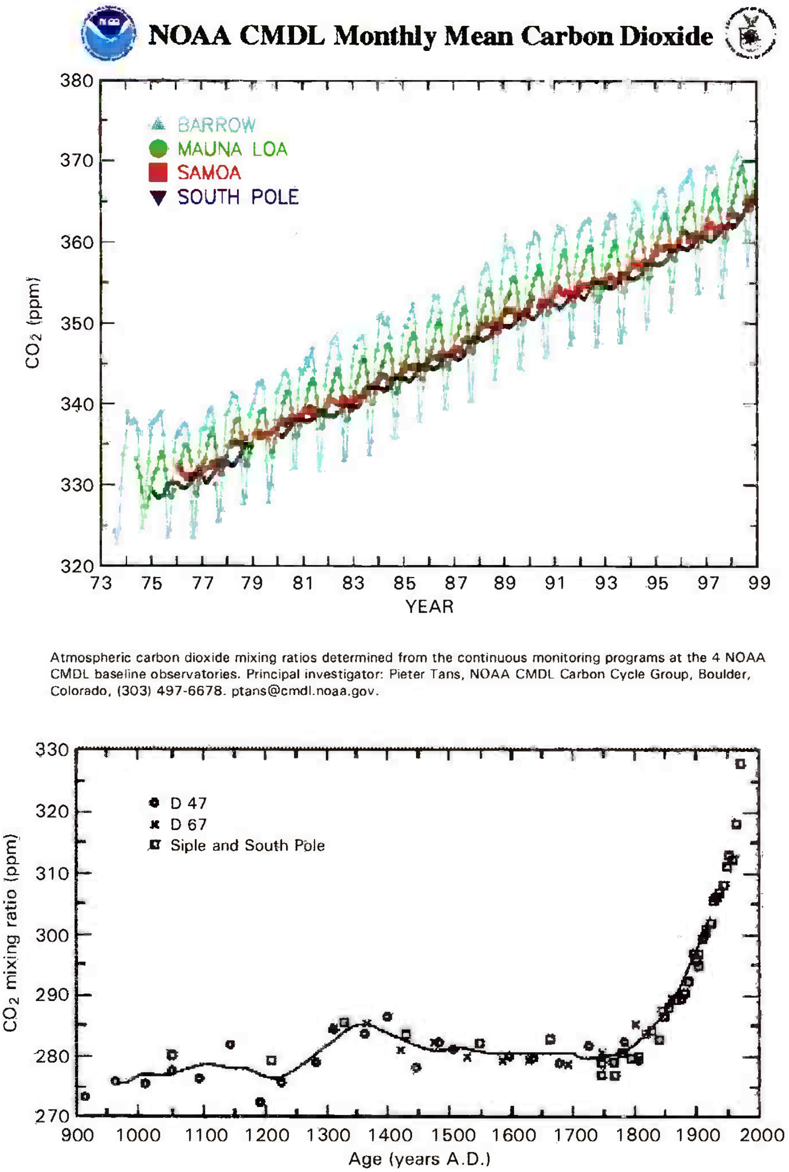
However, operational satellites do not provide the necessary data. To achieve sufficiently high spatial resolution to measure the variability of SWE in mountainous areas, a multiple-frequency, multiple-polarization SAR is required. At present, the measurement of the spatial distribution of SWE, and total snow volume within a mountainous basin, must be performed by intensive field sampling to attempt to represent the large spatial variability of alpine snowpacks. Logistics and safety limitations generally restrict the number of field samples that may be so obtained (Elder et al., 1991a). Thus the problem of determining the volume and distribution of snowpack water storage within mountain basins remains acute.

Climate Models

While still not an operational tool, the use of high-resolution regional climate models for simulating seasonal and interannual changes in snowpack in areas such as the western United States is quite promising. At a 60-km resolution such a model can reproduce the overall patterns of measured precipitation and snow cover and to some extent year-to-year variations (Figs. 5*a* and 5*b*) (Seth et al., 1999). Errors in simulating the actual magnitude of seasonal snow accumulation arise in large part from lack of realism in model topography, with inadequacies in model parameterization also a factor. However, climate modeling offers great promise as a tool that can be integrated with ground-based and satellite observations.

Emerging Technologies

The National Aeronautic and Space Administration's (NASA's) moderate resolution imaging spectrometer (MODIS) will provide near-daily global coverage (comparable



Atmospheric carbon dioxide mixing ratios determined from the continuous monitoring programs at the 4 NOAA CMDL baseline observatories. Principal investigator: Pieter Tans, NOAA CMDL Carbon Cycle Group, Boulder, Colorado, (303) 497-6678. ptans@cmdl.noaa.gov.

Figure 5 (Chapter 1) (*a*) Monthly concentrations of CO_2 measured from gas samples at four monitoring sites operated by NOAA's Climate Monitoring and Diagnostics Laboratory from the early 1970s; (*b*) CO_2 concentrations determined from ice core samples estimated to go back ~1000 years.

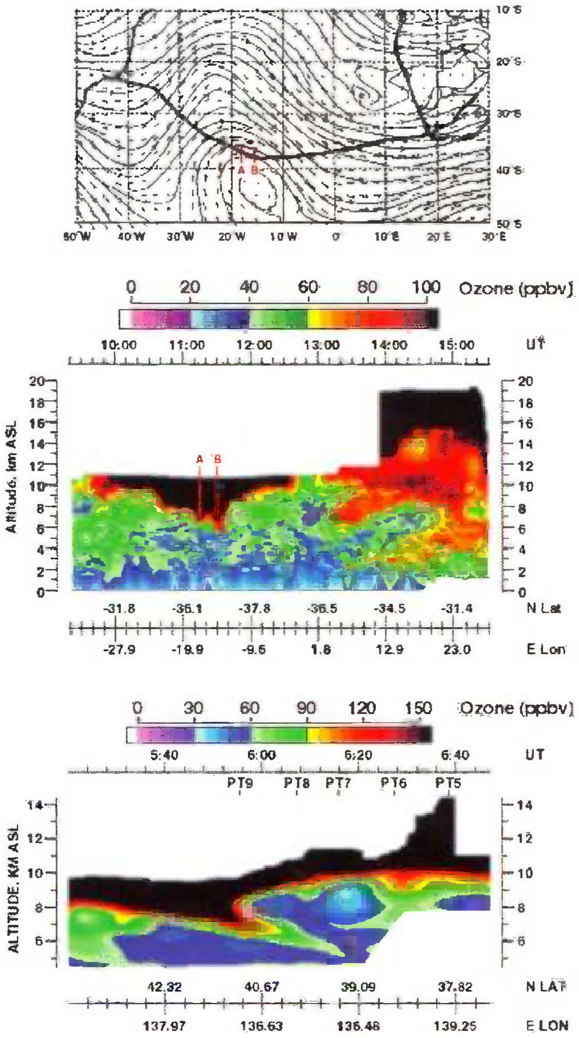
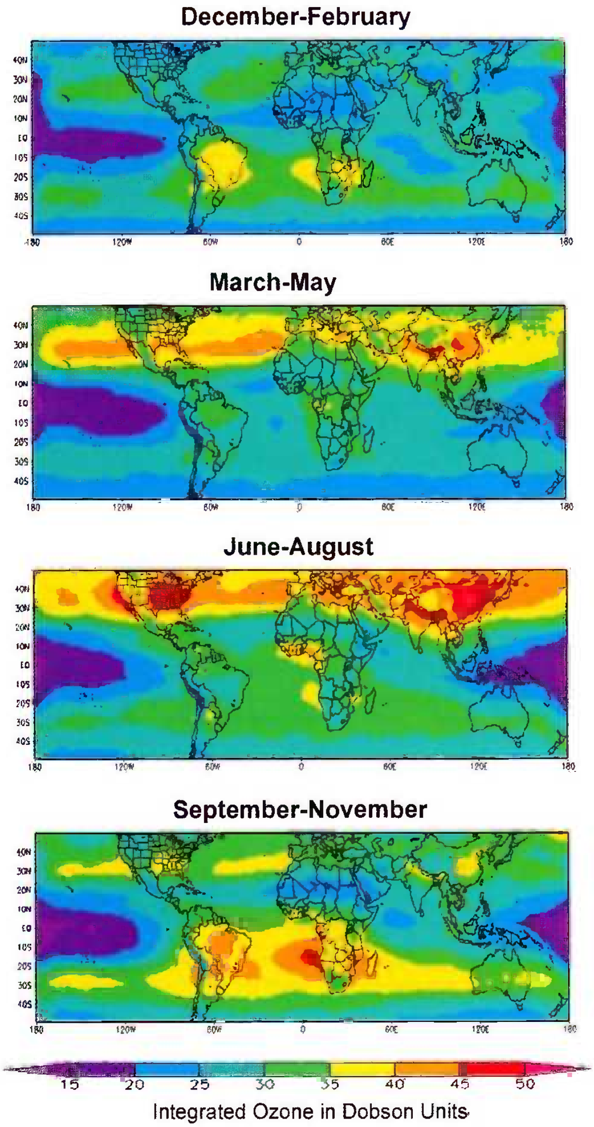


Figure 10 (Chapter 1) Three-panel figure showing evidence of ozone input from the stratosphere into the troposphere in both hemispheres. The top panel shows the flight path (heavy line) of a DC-8 airplane on October 3, 1992, from South America to Africa that intersected a trough protruding from higher latitudes. Points A and B on that flight path show high concentrations of ozone being transported to altitudes below 6 km in the middle panel; the data depicted in this panel were obtained from a differential absorption lidar system that measured ozone below the 11-km flight level of the DC-8. The lowest panel shows a similar feature for a flight on March 11, 1994, in the Northern Hemisphere. As the airplane flies from north to south in this panel, note the higher tropopause height south of the fold.



December-February

March-May

June-August

September-November

Figure 1 (Chapter 3) Climatological distribution of tropospheric ozone derived from satellite measurements between 1979 and 2000 (from Fishman et al., 2002). Units of contours and Dobson Units (DU). Regions greater than 40 DU have been shaded.

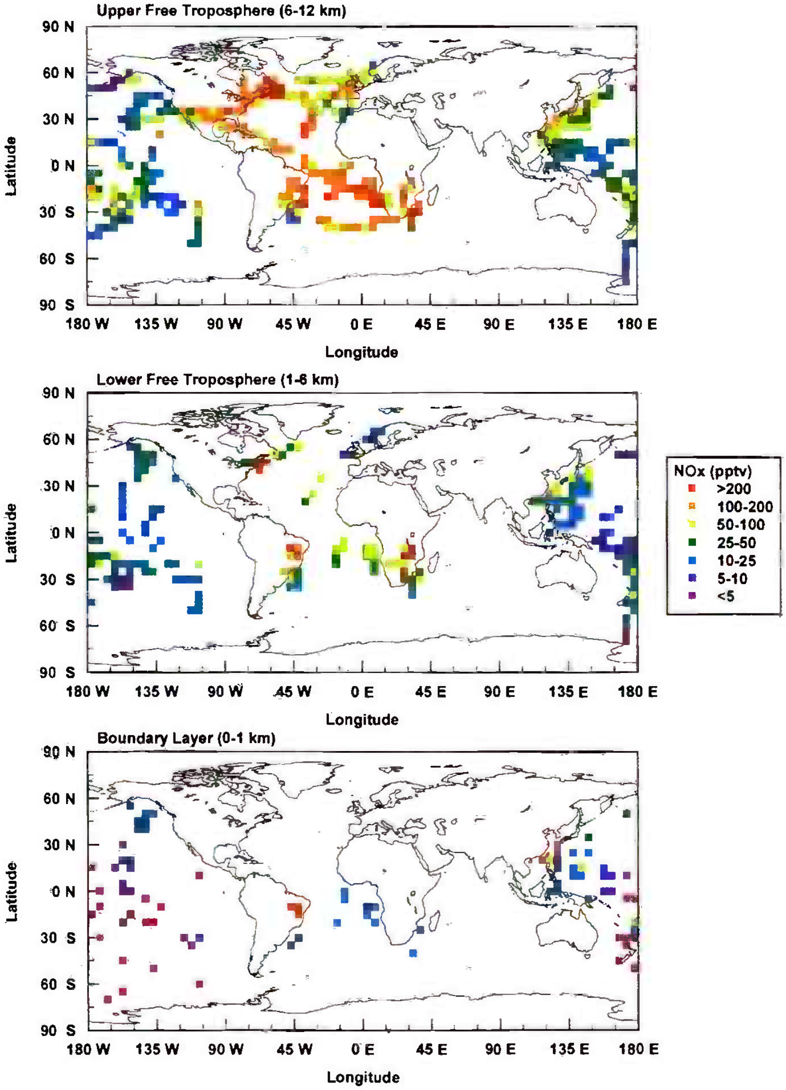


Figure 3 (Chapter 4) Distribution of NO$_x$ based on measurements taken from NASA's DC-8 aircraft during fall (see text for details). Data are averaged on a 5° × 5° latitude–longitude grid for three altitude ranges.

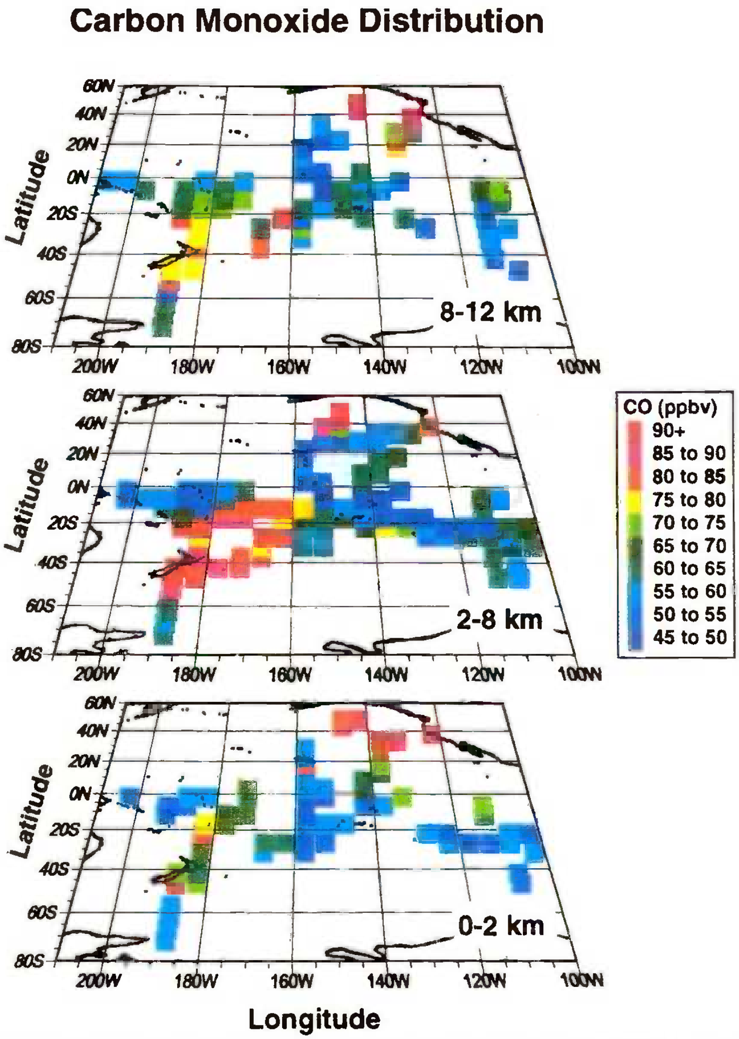


Figure 4 (Chapter 14) CO over tropical Pacific during September 1996 PEM–Tropics A sampling (from Blake et al., 1999). Measurements by G. W. Sachse with a lidar-based instrument. Analysis of possible fire sources is described by Olson et al. (1999).

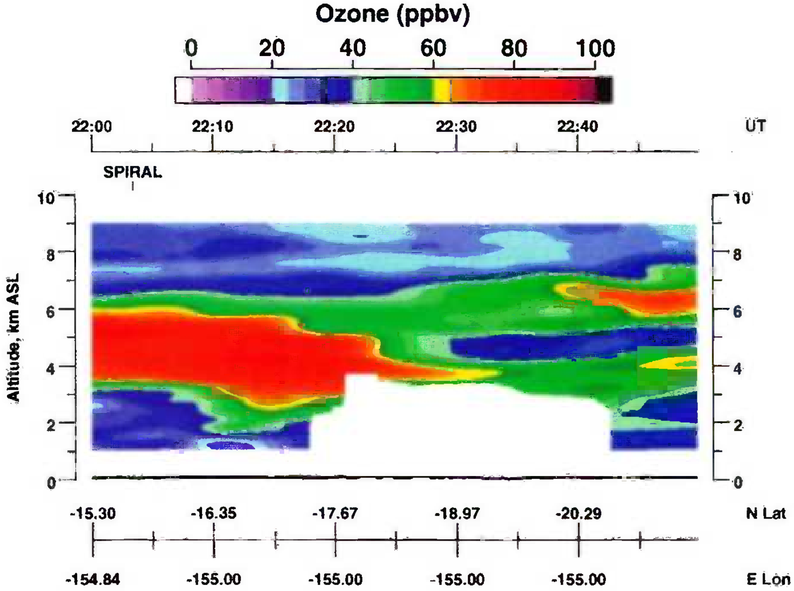


Figure 5 (Chapter 14) Ozone plume over the Pacific seen during the PEM–Tropics A aircraft mission in Sept.–Oct. 1996. (from Fenn et al., 1999).

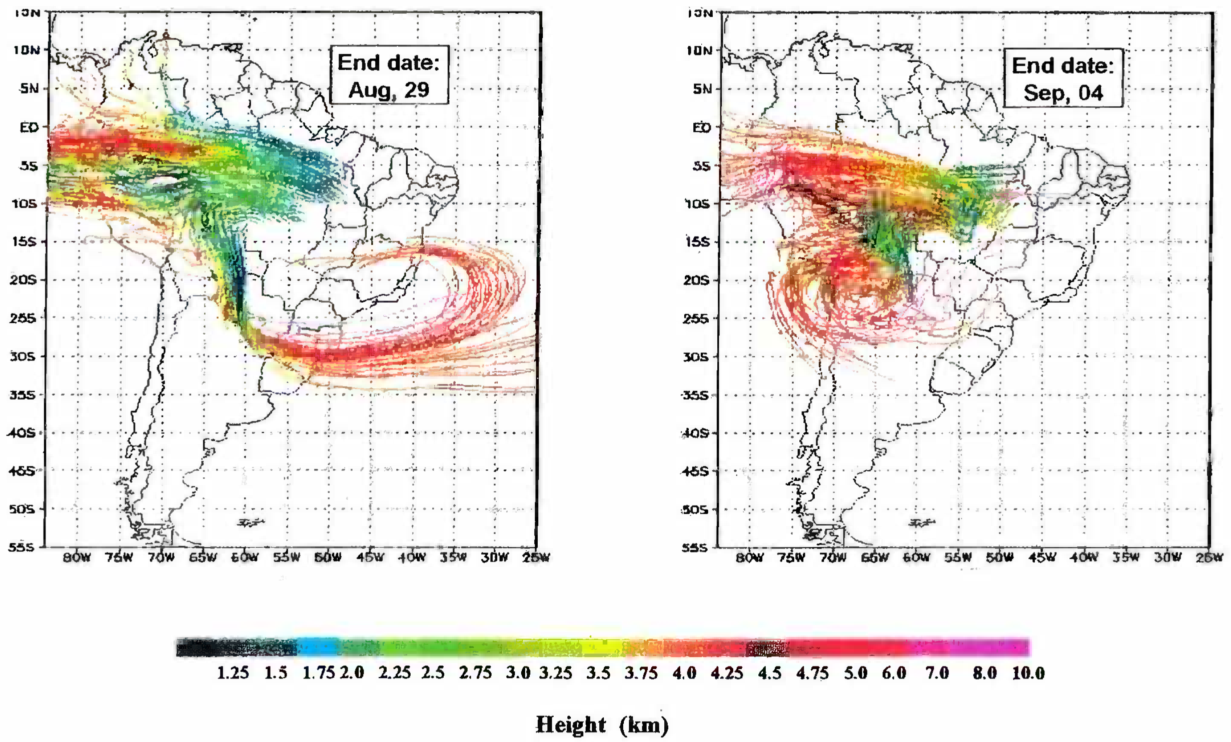


Figure 8 (Chapter 14) Composite of forward trajectories from Cuiabá during the 1995 SCAR-B field experiment. A Brazilian version of the Colorado State mesoscale RAMS model was used to provide winds for the University of São Paulo kinematic trajectory model (from Longo et al., 1999).

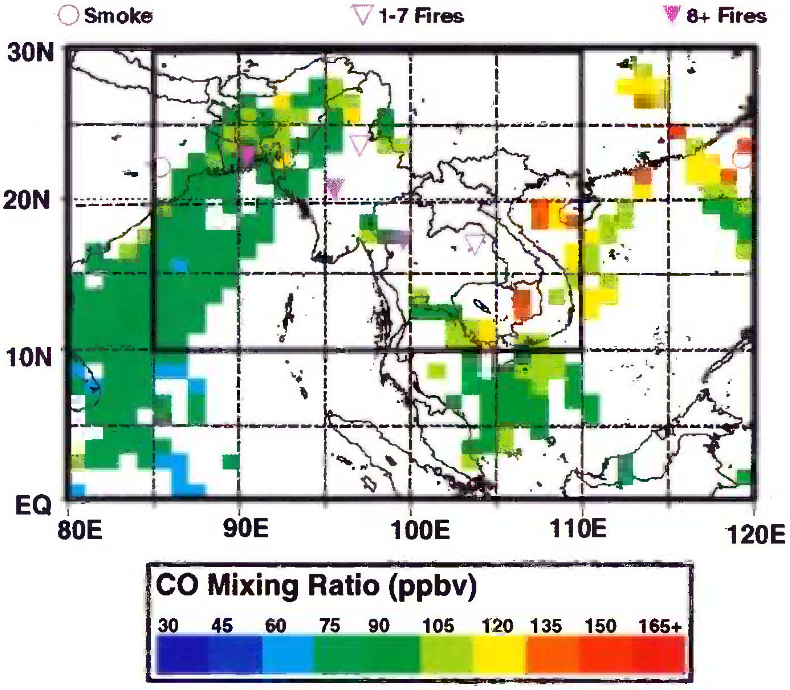


Figure 10 (Chapter 14) (*a*) MAPS CO, April 1994 (from Christopher et al., 1998).

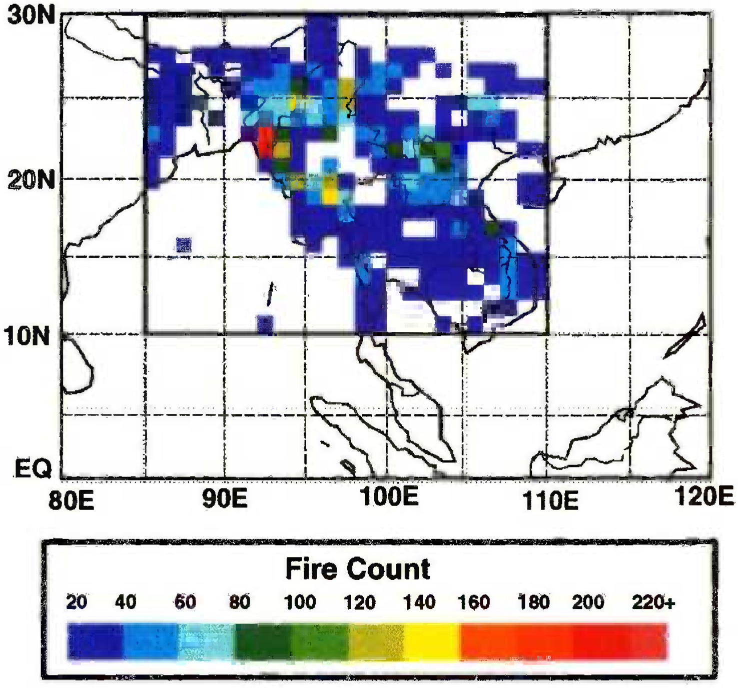


Figure 10 (Chapter 14) (*b*) coincident fires during April 1994 Space Shuttle flight (from Christopher et al., 1998).

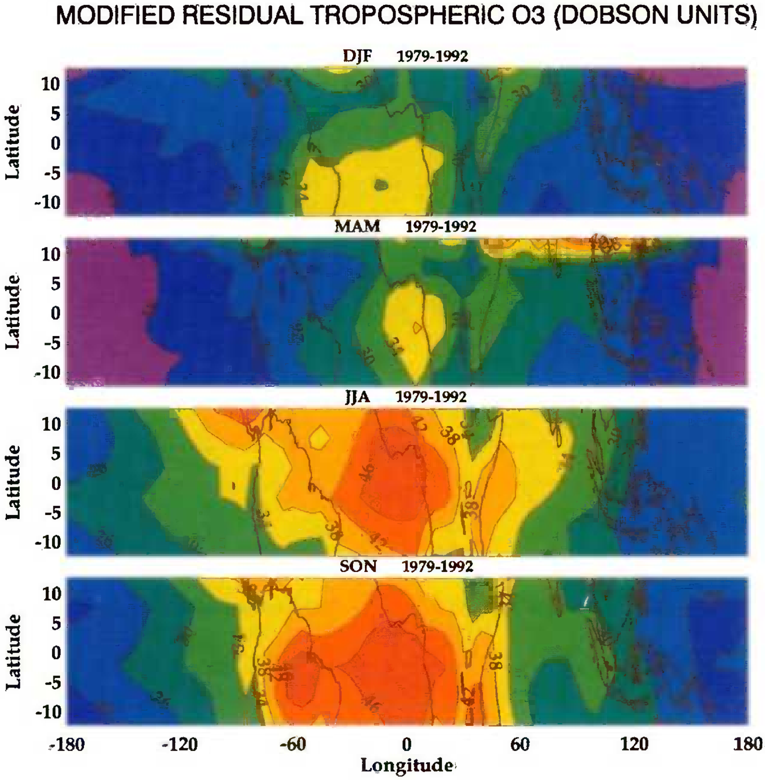


Figure 12 (Chapter 14) Wave-one pattern in tropospheric ozone apparent in TOMS satellite data, averaged from 2 maps/month during the 1979–1992 *Nimbus* 7 observing period. Wave appears to be present throughout year. Scale is DU (Dobson units). Cf. Figure A1 in Thompson and Hudson (1999).

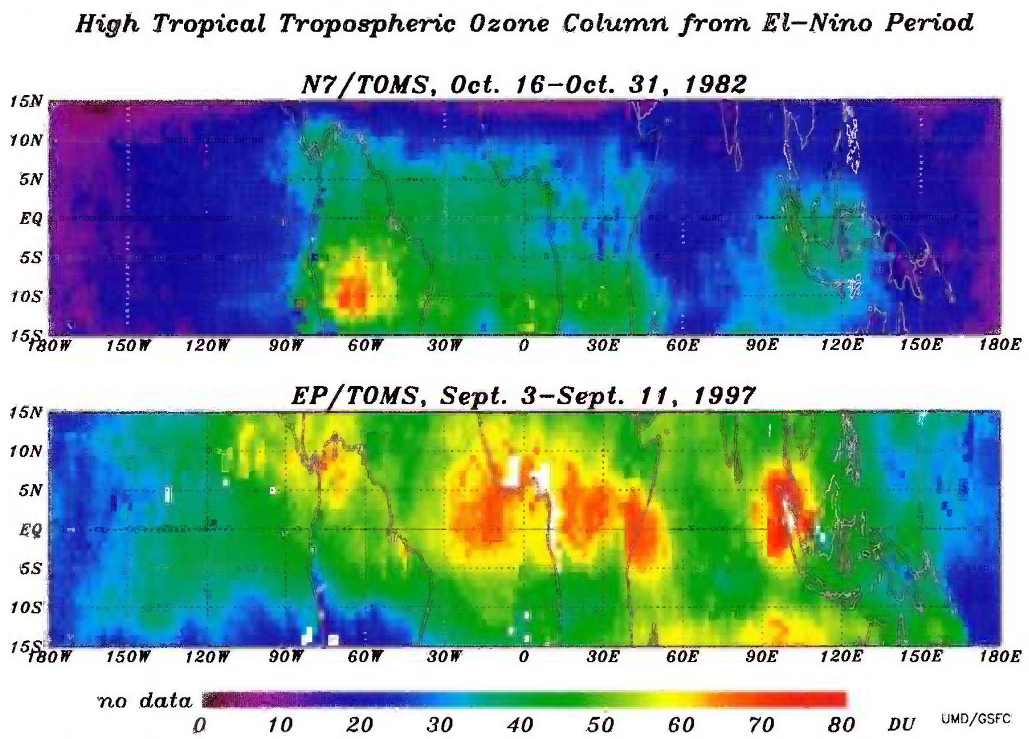


Figure 14 (Chapter 14) Tropospheric column ozone (in DU, from modified-residual method; Thompson and Hudson, 1999) during El Niño-Southern Oscillation (ENSO) of late 1982 (upper panel) as seen in tropical tropospheric ozone map and for September 1997 (lower panel).

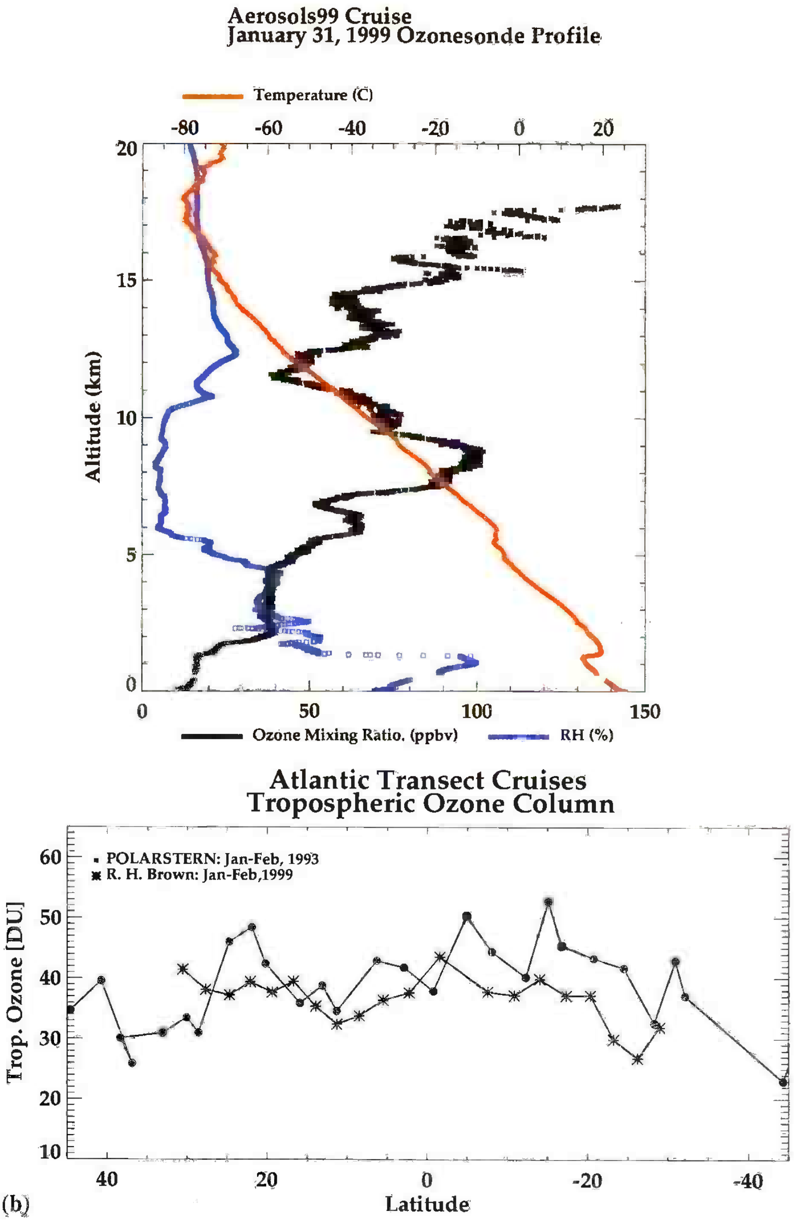


Figure 15 (Chapter 14) (*a*) Profiles of ozone, temperature and water vapor (as percent relative humidity) from 0 to 20 km on January 31, 1999 during Aerosols99 cruise of R/V *Ronald H. Brown*. Anti-correlation of high ozone between 7 and 10 km suggestive of aged stratospheric air. (*b*) Comparison of integrated tropospheric column ozone from sondes launched along Atlantic transect of R/V *Ronald H. Brown* (Thompson et al., 2000) in January–February 1999 and from sondes launched along January–February 1993 Atlantic transect of R/V *Polarstern* (Weller et al., 1996).

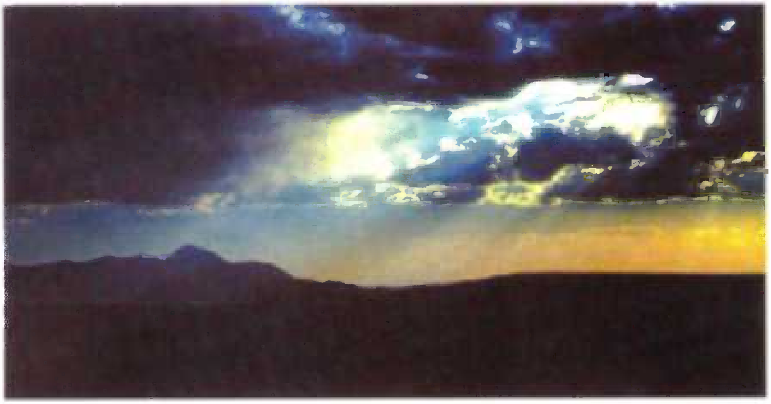


Figure 26 (Chapter 16) The brown discoloration resulting from an atmosphere containing nitrogen dioxide (NO_2) being shaded by clouds but viewed against a clear blue sky. Light scattered by particulate matter in the atmosphere can cominate light absorbed by NO_2, causing a gray or blue appearing haze (left side of photograph).

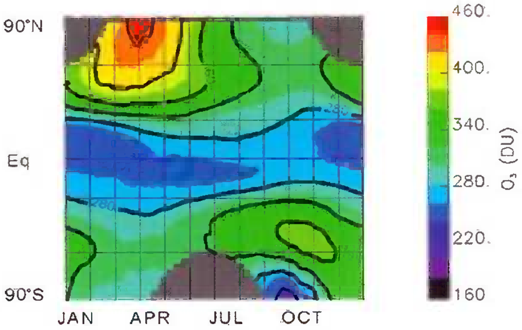


Figure 3 (Chapter 21) Two-dimensional (latitude/season) representation of total column ozone as measured by TOMS for the period 1978 to 1993.

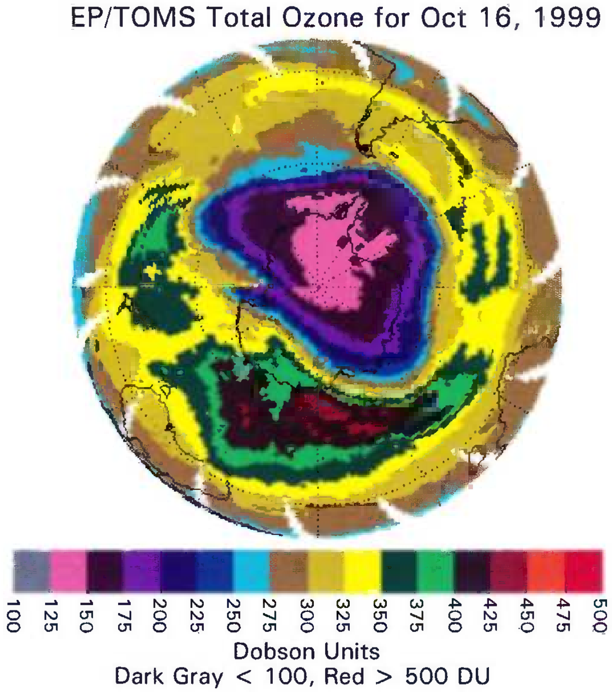

EP/TOMS Total Ozone for Oct 16, 1999
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
475
500
Dobson Units
Dark Gray < 100, Red > 500 DU

Figure 5 (Chapter 21) Map of total column ozone over the Antarctic as determined from TOMS October 16, 1999.

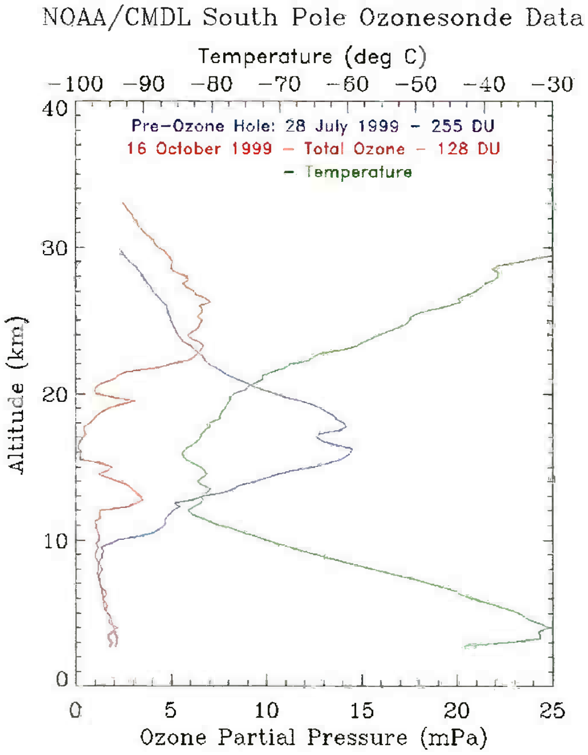


Figure 6 (Chapter 21) Plot of vertical profile of ozone (blue and red lines) over the South Pole as measured from ozonesondes during austral winter (July 28) and spring (October 16), 1999; temperature profile for October 16 is also shown (green line).

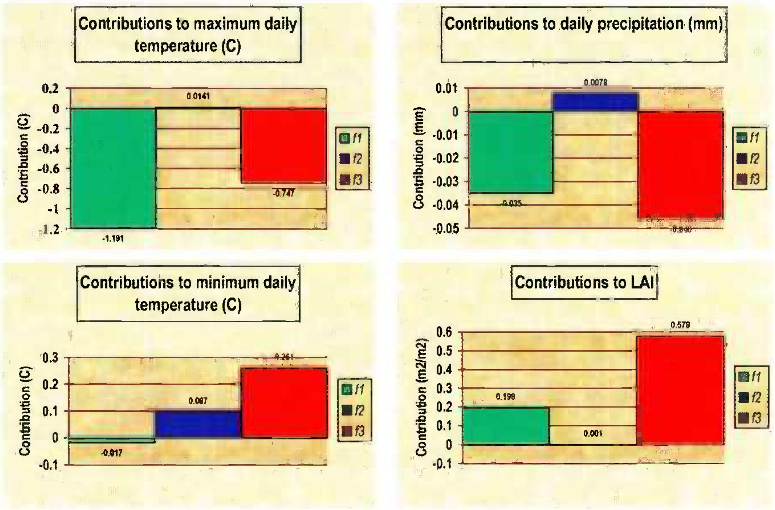


Figure 2 (Chapter 32) RAMS/GEMTM coupled model results—the seasonal domain-averaged (central Great Plains) for 210 days during the growing season, contributions to maximum daily temperature, minimum daily temperature, precipitation, and leaf area index due to f1 = natural vegetation, f2 = 2XC02 radiation, and f3 = 2xC02 biology. (*Adapted from Eastman et al., 2001*).

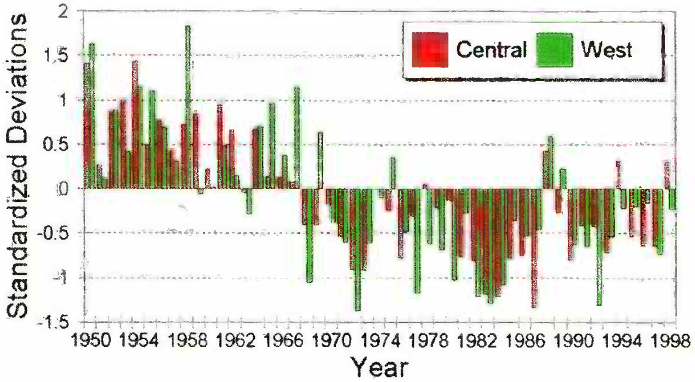
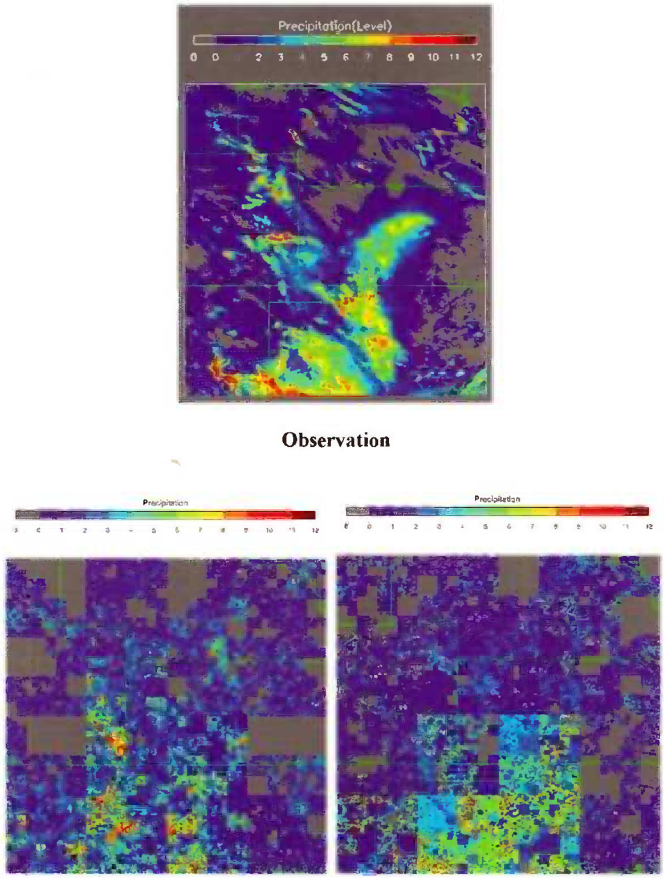


Figure 3 (Chapter 32) Swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998. (*After Landsea et al., 1999.*)



Observation

β-lognormal model

Nonparametric hierarchical model

Figure 6 (Chapter 33) Comparison of observed and downscaled rainfall fields (July 6, 1997) (from Kang and Ramirez[46]).

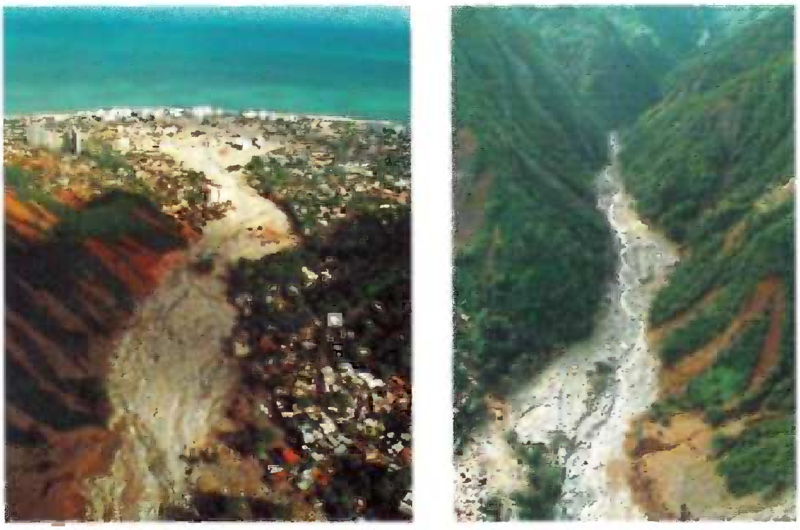


Figure 1 (Chapter 36) Quebrada San Julián upstream of Caraballeda showing evidence of recent debris flows and flash floods. Note the high slope angles, large numbers of debris flow scars, and abundance of new alluvium and colluvium in the channel bed and fan surface.

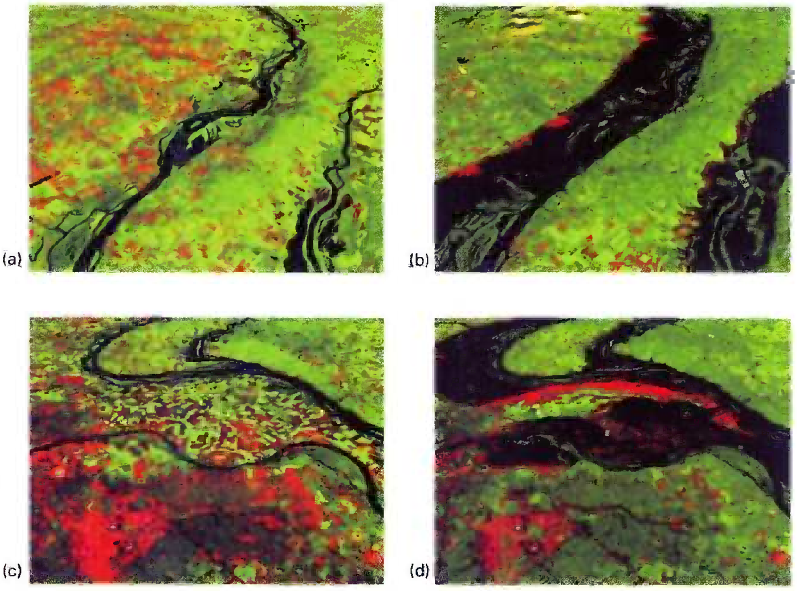


Figure 2 (Chapter 36) (4 panels) These scenes show various sections of the Mississippi River near St. Louis before and just after the 1993 floods, which peaked in late July/early August. The images show the area as seen by the LandSat Thematic Mapper (TM) instrument. The short-wave infrared (TM band 5), infrared (TM band 4), and visible green (TM band 2) channels are displayed in the images as red, green, and blue, respectively. In this combination, barren and/or recently cultivated land appears red to pink, vegetation appears green, water is dark blue, and artificial structures of concrete and asphalt appear dark gray or black. Reddish areas in the scenes during the flood show where water had started to recede, leaving barren land.

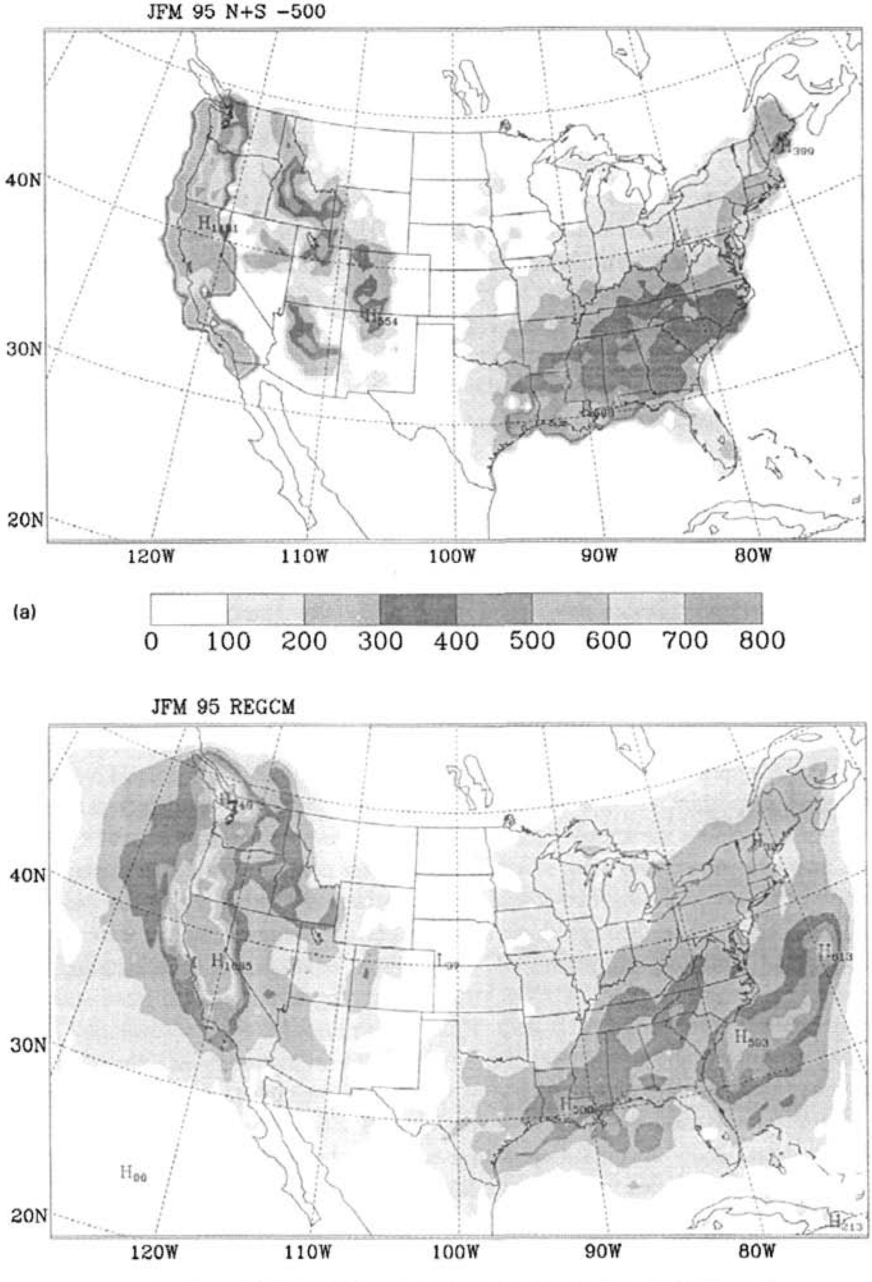


Figure 5 January, February, March 1995 total precipitation (mm) from (*a*) NCDC+SNOW-SNOWTEL; and (*b*) simulated by RegCM.

to AVHRR), but at spatial resolutions ranging from 250 m to 1 km. The sensor has two channels in the visible and near-infrared spectral bands at 250-m resolution, five channels in the visible, near-infrared, and short-wave infrared at 500-m resolution, and the remaining 29 MODIS channels have a spatial resolution of 1 km. The MODIS sensor has on-board visible/near-infrared calibrators, while the AVHRR does not; thus one will be able to derive radiances over snow using some of the MODIS sensors. At least one of the visible MODIS sensors will not saturate over snow. This will be an advancement over the AVHRR and TM sensors that experience significant detector saturation over snow and ice targets in the visible channels. The standard MODIS snow and ice products will consist of 500-m or 1-km resolution binary maps of snow and ice cover, respectively, produced on a global, daily basis in most months.

An approach to modeling spatially distributed snowmelt in steep, alpine basins was proposed using net potential radiation, distributed across the basin using a digital elevation model, as the main factor determining relative snowmelt (Elder et al., 1991b). However, to date this approach has only been applied to small, headwater catchments. It is also possible to infer SWE after the fact from measurements of snow-cover depletion. With a time series of snow cover, e.g., from TM, AVHRR, or MODIS imagery, one can tell when the snow cover disappears, i.e., when snow water equivalence goes to zero. Then using a spatially distributed snowmelt model, one can back calculate from the time snow cover disappears at a point, and then infer the starting value of snow water equivalence. This method has been implemented using TM scenes for a small watershed in the Sierra Nevada, California (Cline et al., 1998).

Research shows that remote sensing also allows estimation of several hydrologic variables important for snowmelt modeling. From airborne hyperspectral sensors [currently NASA's advanced visible/infrared imaging spectrometer (AVIRIS)] one can estimate snow grain size, albedo, liquid water content in the surface layer, and subpixel coverage. Using two-frequency, co-polarized synthetic aperture radar, one can map both snow through thick cloud cover and estimate liquid water content accurately. Work on estimation of snow water equivalence is continuing, with promising results from SIR-C/X-SAR, from photogrammetry and from snowmelt modeling with time-series SCA data. These capabilities have been developed largely using experimental sensors that are not included in currently scheduled satellite launches. A fully automated method of subpixel snow cover mapping uses *Landsat* TM data to map snow cover in the Sierra Nevada and make quantitative estimates of the fractional snow-covered area within each pixel (Rosenthal and Dozier, 1996). Snow fraction estimates from the satellite data can be as accurate as those attainable with high-resolution aerial photography, but they are obtained faster, at much lower cost, and over a vastly larger area.

An important emerging technology is the use of spatially distributed, energy balance snow models to describe the accumulation and ablation of snowpacks. These models organize a wide variety of hydrometeorological and terrain information and permit an improved understanding of snowpack evolution throughout an area of interest. These models have been implemented primarily in well-instrumented research basins, where high-quality meteorological measurements are available to drive the models. However, recent applications have extended their use to larger regions, where mesoscale numerical weather prediction (NWP) model analyses are used to drive the models (e.g., Cline and Carroll, 1999). The NWS NOHRSC is currently developing a four-dimensional data assimilation system for snow estimation that will use a spatially distributed, physically based snow energy and mass balance model, mesoscale NWP analyses, and an updating scheme to provide operational SWE estimates for the United States.

Gaps in Measurement and Understanding

The most significant constraint on the development of snow hydrology is the general lack of measurements of SWE. Measurements are very sparse throughout the United States, especially in the eastern United States. Most operational in situ SWE measurements are collected to support empirical models that are used to estimate snowmelt runoff. For this purpose, the location of the measurement does not necessarily have to be representative of the surrounding area—it is the relationship between the SWE at particular sites and runoff that is important. This means that most operational SWE measurements (e.g., snow courses and snow pillows) are not reliable indicators of the distribution of SWE in a given area. Rather they are index sites that have snow cover for much of the season, to better support empirical modeling. Furthermore, there are simply too few measurements available to adequately characterize the spatial variability of SWE. The general lack of SWE measurements imposes a significant constraint on the development of improved remote-sensing procedures and distributed snow models, which requires "ground truth" for validation and parameter estimation.

The spatial variability of SWE and snowmelt processes in models needs to be better understood. This is a key area for future development of snow hydrology. Most physical snow process work has been carried out at the point scale, or at very local scales. However, most hydrologic and atmospheric modeling scales are run at much coarser spatial scales that involve significant variability of important processes. These models must parameterize the subgrid variability in some manner, but little effort has been devoted to this problem in snow hydrology. Improved understanding of the variability of SWE and snowmelt processes is also needed to design improved sampling strategies for field measurements. High spatial resolution remote-sensing observations of SWE and other snowpack characteristics using SAR would significantly improve our understanding of the spatial variability of snow properties.

4 ESTIMATION OF SNOWMELT RUNOFF

Historical Approach for Operational Forecasts

Both conceptual and physical approaches have been employed in snowmelt runoff modeling. Conceptual models propose a mathematical relationship between snowmelt and measured quantities; thus melt can be calculated without treating in detail all of the physical processes and parameters that affect snowmelt. Conceptual models have the benefit of requiring less informational input but suffer from the uncertainty that the conditions hypothesized under different model scenarios are modeled sufficiently.

Operationally, the NWS is tasked with forecasting streamflow, floods, and seasonal water supplies in the United States. Various operational data are used by the algorithms making up the NWS River Forecast System (NWSRFS) to produce streamflow and water supply projections that extend several hours to months into the future. Snow accumulation and ablation is modeled in NWSRFS using an empirical approach that is driven using inputs of air temperature and precipitation (Anderson, 1973; Day, 1990). During periods of precipitation, a simplified energy balance approach is used to estimate snowpack state conditions. When no precipitation is occurring, a simple temperature index method is used. The models are spatially lumped, that is, they model the "average" snowpack state conditions over entire basins or subbasins.

Short-term streamflow forecasts are made using NWSRFS and meteorological forecasts of temperature and precipitation. Because meteorological forecasts become less reliable the further out they extend, this method of streamflow prediction is limited to periods of a few hours to a few days. Beyond that time, the short-term meteorological forecasts are blended into the climatologically average conditions. The long-range forecasting component of NWSRFS, called the extended streamflow prediction (ESP) technique, uses present-day streamflow, soil moisture, and snowpack conditions along with a historical time series of precipitation and temperature to estimate streamflow weeks or months into the future.

Two major factors contribute to high uncertainty in estimates of snowpack conditions, particularly in mountainous regions. First, the NWSRFS snow models contain several empirical parameters that must be calibrated in conjunction with the rest of the NWSRFS algorithms. The empirical calibrations are useful, as they help overcome certain problems, such as poorly representative temperature or precipitation measurement sites. The snow models perform best when conditions are near the average conditions used in the calibration. During extreme or unusual conditions, the models often produce spurious results. Second, accurate estimation of precipitation is critical for these models, but this is especially challenging in mountainous regions. Like the distribution of SWE, the distribution of precipitation is typically highly variable in mountain regions, and there are too few measurement sites to adequately represent this variability. The problem is compounded by the fact that precipitation gages do not measure snowfall well. Consequently, large uncertainties in precipitation inputs to the snow models propagate to the estimates of snow state.

Since current snow state conditions in the model serve as initial conditions for streamflow forecasts, the large uncertainties in modeled snowpack state conditions must be reduced where possible prior to extending the model forward in time. This is accomplished by updating the SWE in the snow models with observed SWE.

In the western United States, two data assimilation schemes are currently used to provide SWE updates for the snow models. The first is a relatively simple approach that uses satellite observations of snow cover to indicate areas without snow, plus an interpolation of all available surface- and airborne-based SWE observations. The second data assimilation scheme, SEUS (McManamon et al., 1993), assimilates surface, airborne, and satellite snow information in an optimal interpolation framework. Grid points in a basin are classified and the amount of SWE and snowmelt for each class for each year in a historical record is estimated. The historical mean SWE fields are then used to estimate the actual SWE values from current gridded data.

Spatially Distributed Modeling of Melt and Runoff

While empirical snowmelt runoff models have traditionally been useful for operational runoff volume forecasts, they provide little information on the timing, rate, or magnitude of discharge, and they are inappropriate in situations outside the boundary conditions governing the development of the relevant empirical parameters. Thus they may fail to adequately predict water yield in extreme or unusual years, and they cannot be reliably used in investigations examining snowmelt responses to climate variability and change. These problems, and the increasing importance of understanding intrabasin snowmelt for environmental analysis of such factors as basin ecology (Baron et al., 1993), water chemistry (Wolford et al., 1996), and hillslope erosion (Tarboton et al., 1991) have motivated the development of physically based, spatially distributed snowmelt models in recent years. Such models require information on the spatial distribution of snowpack water storage. But mountain snowpacks are spatially heterogeneous, reflecting the influences of rugged topography on precipitation, wind redistribution of snow, and surface energy fluxes during the accumulation season (Elder et al., 1991a), and no widely suitable method yet exists to directly map SWE or simulate distributed snowmelt in rugged mountain regions. One of the main obstacles to physically based modeling is the compilation of the necessary meteorological and snow-cover data to run, calibrate, and validate such models. For example, basin discharge has frequently been used as the sole physical criterion of model calibration and performance assessment for conceptual snowmelt models. But as it is an integrated response to melt and runoff, basin discharge is not sufficient to discriminate between the effects of the multiplicity of data inputs driving physical models and that distributed snow-cover data are required to assess model performance (Bloschl et al., 1991a, 1991b).

A distributed snowmelt model consists of two general components: a model for calculating melt at a single point (given a set of prescribed snowpack and meteorological conditions), and a method of developing the requisite snowpack and meteorological data for all points within the basin. Snowmelt modeling efforts require several steps necessary to couple basinwide energy balance snowmelt models with remote sensing and flow routing. The model topographic distribution of solar radiation (TOPORAD) (Dozier and Frew, 1990) uses information on watershed topography (i.e., a digital elevation model) to spatially distribute radiation. Simpler models are used to spatially distribute other energy balance components. These radiation maps are then used as inputs to models to estimate the distribution of snow around the basin prior to snowmelt (i.e., initial conditions for melt), and as inputs to the point snowmelt model.

Climate Change and Variability Issues

Expected changes in global climate are cause for concern for future water resources in the western United States, where limited water supplies are already in great demand, and a constrained water regulatory system struggles to satisfy water users. The main question for the regions is: What does potential climate change mean to snowmelt-dominated western water resources? Certainly the assumption of stationarity of the present snowmelt and streamflow regime must be questioned (Dracup and Kendall, 1990); however, it is not clear what exactly should take its place. One criterion with which many global climate models are run is the anticipated doubling of greenhouse forcing from atmospheric greenhouse gases; this doubling and the resulting modeled climate changes are expected to occur by the mid-twenty-first century, well within the "design life" of many existing water control structures and regulations, and indeed within the careers of the next generation of water planners. For this reason, the problem cannot be easily ignored; however, rapid advances in understanding of the potential effects of climate change on snowmelt-dominated hydrologic systems are necessary before an informed response by water resource planners can be made.

A variety of approaches have been used to estimate the effects of or sensitivity to climate change in snowmelt-dominated basins in the western United States. The methods range from purely statistical techniques (Duell, 1992), regression models, to complex deterministic spatial modeling strategies that explicitly account for every process in as detailed a manner as possible (Leavesley et al., 1992). Potential effects of climate change on snowmelt and streamflow in the west include reduced annual streamflow, earlier peak flows in spring, and less winter precipitation occurring as snow. However, due to the great uncertainty in the input data, and the general inapplicability of the models to forecasts beyond the range of their prior calibration, magnitudes remain largely unknown. There are also issues with the variety of models in use, and few cases where a common approach has been employed in more than one hydroclimatic region.

Two points illustrate the tentative and qualitative nature of work to date and highlight the need for a more rigorous approach. First, examination of the major topographic effects on temperature and precipitation in the western United States shows that only regional-scale climate models are appropriate for use in this part of North America. The results of any modeled climate change scenario for the western United States that has not incorporated reasonable representations of topography must be viewed as spurious at best (Seth et al., 1999). In addition to, and because of, the lack of adequate topographic representation in general circulation models (GCMs), several of the important consequences of topographically controlled precipitation and temperature, such as albedo differences due to the presence of snow at high elevations, and forced uplift and cooling of large-scale airflow over individual mountain ranges are also missing from GCMs. This suggests that it may not be feasible to treat GCM results as even a general trend or rough estimate of realistic climate changes. This is not so much a criticism of GCMs in general, but rather a criticism of the use of their output for modeling the hydrology of basins with topography of which the GCMs are not even aware. Thus accurate quantitative climate change scenarios are conspicuously absent from exercises attempting to predict climate change effects on streamflow in the west.

The second point of the argument is that hydrologic simulations are severely limited by the excessive need to calibrate snowmelt runoff models to achieve a good fit to a basin hydrograph. Under the assumption of stationarity, such calibration is acceptable if runoff is the only variable of interest. Unfortunately, there is absolutely no reason to assume that if climate changes the hydrologic characteristics of a basin will remain stationary. If internal basin hydrologic characteristics are also of interest, then most existing snowmelt runoff models cannot be used since they do not provide any information on internal basin processes. In fact, most of the models depend on sufficient aggregation of internal processes so that accurate outflow predictions may be made. Distributed parameter models appear to address the internal basin processes to a larger degree, but often require more extensive data.

In most studies to date of the potential effects of climate change on snowmelt runoff, the "effect" receiving the greatest attention is the basin hydrograph. Internal basin characteristics should be given greater attention for two reasons. First, basins themselves are important, and we should be concerned about how they may be affected by changing climate. Second, we need to begin incorporating the notion of nonstationarity of basin properties into hydrologic models. For example, climate-change-induced changes in the extent and characteristics of forest cover within a basin might have a greater effect on both the timing and magnitude of snowmelt runoff than changes in temperature or precipitation. Changes in the proportion of rainfall/snowfall could lead to important changes in sediment transport, mass wasting, and other geomorphic characteristics of a basin, which would affect appropriate selection and use of hydrologic models. The hydrochemistry of streams and rivers is largely dependent on the interaction of water with various basin component; thus intrabasin biological and geomorphic changes due to climate change could have important implications for stream water chemistry. Improvements in distributed-parameter hydrologic models and particularly in methods of collecting sufficient input data for these models will certainly be necessary before internal basin processes can be adequately examined.

REFERENCES

Anderson, E. A., *National Weather Service River Forecast System: Snow Accumulation and Ablation Model*, NOAA Technical Memorandum, NWS, Hydro-17, U.S. Department of Commerce, Washington, DC, 1973.

Baron, J. S., L. E. Band, S. W. Running, and D. Cline, The effects of snow distribution on the hydrologic simulation of a high elevation Rocky Mountain watershed using Regional HydroEcological Simulation Systems, RHESSys, *Eos, Trans. Am. Geophys. Union*, *74*(43), 237, 1993.

Bloschl, G., D. Gutknecht, and R. Kirnbauer, Distributed snowmelt simulations in an Alpine catchment 1. Model evaluation on the basis of snow cover patterns, *Water Resour. Res.*, *27*(12), 3171–3179, 1991a.

Bloschl, G., D. Gutknecht, and R. Kirnbauer, Distributed snowmelt simulations in an Alpine catchment 2. Parameter study and model predictions, *Water Resour. Res.*, *27*(12), 3181–3188, 1991b.

Cline, D., K. Elder, and R. Bales, Scale effects in a distributed snow water equivalence and snowmelt model for mountain basin, *Hydrol. Process.*, *12*(10–11), 1527–1536, 1998.

Cline, D., and T. Carroll, Inference of snow cover beneath obscuring clouds using optical remote sensing and a distributed snow energy and mass balance model, *J. Geophys. Res. Atmos.*, *104*(D16), 19631–19644, 1999.

Day, G. N., *A Methodology for Updating a Conceptual Snow Model with Snow Measurements*, NOAA Technical Report, NWS 43, U.S. Department of Commerce, Washington, DC, 1990.

Dozier, J., and J. E. Frew, Rapid calculation of terrain parameters for radiation modeling from digital elevation data, *IEEE Trans. Geosci. Remote Sensing*, 28, 963–969, 1990.

Dracup, J. A., and D. R. Kendall, Floods and droughts, in P. E. Waggoner (Ed.), *Climate Change and U.S. Water Resources*, Wiley, New York, pp. 243–267, 1990.

Duell, L. F. W., Jr., Use of regression models to estimate effects of climate change on seasonal streamflow in the American and Carson River Basins, California-Nevada, in Hermann, Raymond, ed., Managing water resources during global change: 28th Annual Conference, American Water Resources Association, Reno, Nev., November 1992, Proceedings, p. 731–740.

Elder, K., R. E. Davis, and R. C. Bales, Terrain classification of snow-covered watersheds, *Proc. Eastern Snow Conf.*, 48, 39–49, 1991a.

Elder, K., J. Dozier, and J. Michaelsen, Snow accumulation and distribution in an alpine watershed, *Water Resour. Res.*, 27, 1541–1552, 1991b.

Hart, D., and F. Gehrke, Status of the California Cooperative Snow Survey Program, in *58th Western Snow Conference Proceedings*, Sacramento, CA, 1990, pp. 9–14.

Leavesley, G. H., M. D. Branson, and L. E. Hay, Using coupled atmospheric and hydrologic models to investigate the effects of climate change in mountainous regions, in Hermann, Raymond, ed., Managing water resources during global change: 28th Annual Conference, Americal Water Resources Association, Reno, Nev., November 1992, Proceedings, p. 691–700.

Lin, H. F., F. K. Hare, and K. P. Singh, Influence of the atmosphere, in M. G. Wolman and H. C. Riggs (Eds.), *Surface Water Hydrology: Boulder, Colorado*, Geological Society of America, The Geology of North America, 1990, pp. 11–53.

McManamon, A., T. L. Szliga, R. K. Hartman, G. N. Day, and T. R. Carroll, Gridded snow water equivalent using ground-based and airborne snow data, in *Proceedings of the 50th Eastern Snow Conference*, Quebec City, 1993, pp. 75–81.

Matson, M., C.F. Roeplewski, and M.S. Varnadore, 1986: An Atlas of Satellite-Derived Northern Hemisphere Snow Cover Frequency, National Weather Service, Washington D.C., 75 p.

Matson, M., 1991: NOAA satellite snow cover data, Palegeography and Paleoecology, 90: 213–280

Paulson, R.W., E.B. Chase, R.S. Roberts, and D.W. Moody, compilers, *National Water Summary 1988–89: Hydrologic Events and Floods and Droughts*, U.S. Geological

Survey Water Supply Paper 2375, U.S. Government Printing Office, Washington, DC, 1991, p. 591.

Rosenthal, W., and J. Dozier, Automated mapping of montane snow cover at subpixel resolution from the Landsat Thematic Mapper, *Water Resour. Res.*, *32*(1), 115–130, 1996.

Seth, A., R. C. Bales, and R. E. Dickinson, A framework for the study of seasonal snow hydrology and its interannual variability in the alpine regions of the Southwest, *J. Geophys. Res.*, *104*(D18), 22117–22135, 1999.

Shi, J. and J. Dozier, Inferring snow wetness using C-band data from DIR-C's polarimetric synthetic aperture radar, *IEEE Trans. Geosci. Remote Sensing*, *33*(4), 905–914, 1995.

Shi, J., and J. Dozier, Estimation of snow water equivalence using SISR=C/X-SAR, in *Proceedings IGARRS 96* , IEEE No. 96Ch35875, 1996, pp. 2002–2004.

Shi, J., and J. Dozier, Mapping seasonal snow with SIR-C/X-SAR in mountainous areas, *Remote Sensing Environ.*, *59*, 294–307, 1997.

Shi, J. and J. Dozier, Estimation of Snow Water Equivalence Using SIR-C/X-SAR Image Data, Proceedings Progress in Electromafnetic Research Symposium, p. 1041, 1995.

Tarboton, D. G., M. J. Al-Adhami, and D. S. Bowles, Preliminary comparisons of snowmelt models for erosion prediction, *Proc. Western Snow Conf.*, *59*, 79–90, 1991.

Wolford, R. A., R. C. Bales, and S. Sorooshian, Development of a hydrochemical model for seasonally snow-covered alpine watersheds: Application to Emerald Lake Watershed, Sierra Nevada, California, *Water Resour. Res.*, *32*(4), 1061–1074, 1996.

Xu, H., J. O. Baily, E. C. Barrett, and R. E. J. Kelly, Monitoring snow area and depth with integration of remote-sensing and GIS, *Int. J. Remote Sensing*, *14*(17), 3259–3268, 1993.

CHAPTER 26

EVALUATING THE SPATIAL DISTRIBUTION OF EVAPORATION

WILLIAM P. KUSTAS, M. SUSAN MORAN, AND JOHN M. NORMAN

1 INTRODUCTION

Evaporation of water from soil and plant surfaces forms the connecting link between the energy balance and the water balance at Earth's surface. This phenomenon influences the large-scale circulation of the planetary atmosphere, affects soil moisture content that in turn affects hydrologic response, and regulates the microscale carbon dioxide uptake of stomata in individual plant leaves. The vast range of scales encompassed by the process of evaporation makes it of vital environmental interest.

Over the past century, theoretical, modeling, and experimental efforts have greatly expanded our ability to evaluate water loss due to evaporation at local scales using conventional instrumentation. In recent decades, a concerted effort has been made to develop techniques for evaluating the spatial distribution of evaporation at regional and global scales. This effort has been largely focused on the use of remotely sensed information available from sensors aboard orbiting satellite platforms. The result has been a variety of methods that vary in complexity from statistical approaches to physically based analytical approaches and ultimately to numerical process models that simulate the flow of heat and water through the soil, vegetation, and atmosphere.

This chapter will present a brief discussion of the physics of evaporation, highlight conventional methods for estimating evaporation rates, and then will focus on the use of remote sensing for evaluation of the spatial distribution of evaporation at the local, regional, and global scales. Emphasis will be placed on methods for estimating evaporation at an hourly to daily time frame, which is most appropriate for atmospheric, hydrological, and agricultural applications. This work will conclude

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

with a synthesis of the most important research and development issues related to the implementation of such approaches on an operational basis. Although much of the material in Sections 4 and 5 is from the work of Kustas and Norman (1996), new information and results from more recent studies are included.

2 SHORT HISTORY

Although the evaporation process has intrigued humankind for centuries, progress in understanding the physics of evaporation remained slow until the twentieth century when Bowen (1926) showed how the partitioning of available energy between the fluxes of sensible and latent heat could be determined from gradients of temperature and humidity:

$$\lambda E = -(R_n + G)/(1 + \beta) \tag{1}$$

where λE^* is the latent heat flux (W/m^2), R_n is the net radiation flux at the surface (W/m^2), G is the sensible heat flux conducted to the soil (W/m^2), and β is the Bowen ratio (Table 1). The ratio of sensible heat (H) to latent heat flux density is

$$\beta = H/\lambda E \tag{2}$$

In Eq. (1), fluxes away from the surface are negative and those toward the surface are positive. The Bowen ratio can be derived from temperature and humidity measurements:

$$\beta = \gamma(K_h/K_v)(\Delta T/\Delta e) \tag{3}$$

where γ is referred to as the psychrometric constant (2.453 MJ/kg at 20°C), K_h and K_v are the eddy transfer coefficients for sensible and latent heat, respectively, and ΔT and Δe are the differences in temperature in degrees centigrade and vapor pressure in kilopascals over the same elevation difference, Δz.

Following the work of Bowen (1926), Penman (1948) combined the thermal energy balance with certain aerodynamic aspects of evaporation and developed an

*Evaporation (E) is often represented in units of mm/day or mm/h but can also be expressed in energy units, where E is the evaporation rate (kg/s m^2), λ is the heat of evaporation (J/kg), and λE is the latent heat flux density (W/m^2). Though expressed in different units, the terms E and λE are interchangeable. To avoid confusion herein, the term E^* will represent evaporation rate in units of depth (mm/h or mm/d), E will represent mass flux density (kg/s m^2 or kg/d m^2), and λE will represent latent heat flux density (in units of W/m^2 or MJ^{-2} d^{-1}). For further clarification on evaluation of Eqs. (1) to (9), readers are encouraged to review Table 1, and consult the treatise by Monteith (1981) and the books by Brutsaert (1982) and Jensen et al. (1989).

TABLE 1 Summary of Scientific and Technical Notation

$\acute{\alpha}$	Surface shortwave albedo
α	Priestley–Taylor coefficient, $\alpha = 1.26$ for regions with no or low advective conditions
β	Bowen ratio, where $\beta = H/\lambda E$
C_p	Specific heat at constant pressure (kJ/kg°C)
d_0	Displacement height (m)
γ	Psychrometric constant (in units of MJ/kg or kPa/°C)
γ*	$\gamma(1 + r_c/r_a)$ (kPa/°C)
ΔT	Difference in temperature (°C) over the elevation Δz
Δe	Difference in vapor pressure (kPa) over the elevation Δz
Δz	Elevation difference (m)
Δ	Slope of the saturation vapor pressure–temperature curve (kPa/°C)
$e_z{}^o$	Saturation vapor pressure at the z level above the surface (kPa)
e_z	Actual vapor pressure at the z level above the surface (kPa)
$e_z{}^o - e_z$	Vapor pressure deficit (kPa)
e'	Instantaneous deviation of the partial water vapor pressure from the mean at height z
E	Mass flux density (kg/s m^2 or kg/d m^2)
E*	Evaporation rate in units of depth (mm/h or mm/d)
EF	Evaporative fraction, where EF$= -\lambda E/(R_n + G)$
ε_s	Surface emissivity
f_g	Fraction of green or actively transpiring vegetation
f_{gr}	Fraction of green vegetation viewed by the radiometer
G	Soil heat flux density (W/m^2)
H	Sensible heat flux density to the air (W/m^2)
$H + \lambda E$	Turbulent fluxes (W/m^2)
H_c	Sensible heat flux density from the canopy (W/m^2)
H_s	Sensible heat flux density from the soil (W/m^2)
k	von Karman's constant ($\approx$0.4)
K_h, K_v	Eddy transfer coefficients for sensible and latent heat, respectively
λE	Latent heat flux density (W/m^2 or MJ^{-2} d^{-1})
λE_c	Latent heat flux density from the canopy (W/m^2)
λE_p	Potential latent heat flux density (W/m^2)
N	Day length (h)
ρ	Air density (kg/m^3)
$\rho_{\Delta\lambda}$	Surface reflectance factor for the spectral range $\Delta\lambda$
ρ_{NIR}, ρ_{Red}	Surface reflectance factors in the near-infrared (NIR) and red spectrum, respectively
P	Atmospheric pressure (kPa)
r_a	Aerodynamic resistance (s/m)
r_c	Canopy resistance to vapor transport (s/m)
r_s	Resistance to heat flow in the boundary layer immediately above the soil surface (s/m)
R_n	Net radiant flux density at the surface (W/m)
$R_n + G$	Available energy (W/m^2)
R_{nc}	Absorbed net radiant flux density by the plant canopy (W/m^2)

(*continued*)

TABLE 1 (*continued*)

R_s	Incoming shortwave solar radiant flux density (W/m^2)
R_{ld}	Incoming longwave radiant flux density (W/m^2)
R_{lu}	Upwelling longwave radiant flux density, represented by $\varepsilon_s \sigma T_{\mathrm{sh}}^4$
σ	Stefan–Boltzman constant (5.67 x 10^{-8} W/m^2 K^4)
t	Time starting at sunrise (h)
T_a	Air temperature (°C)
T_{aero}	Surface aerodynamic temperature (°C)
T_c	Canopy temperature (°C)
T_{rad}	Radiometric temperature measured by an infrared radiometer from a space-borne platform
T_s	Soil surface temperature (°C)
T_{sh}	Hemispherical radiometric temperature (C or K)
u	Horizontal wind speed (m/s)
u_s	Horizontal wind speed (m/s) about 5 cm above the soil surface
w	Mean vertical wind at height z (m/s)
w'	Instantaneous deviation of vertical wind speed from w (m/s)
W_f	Wind function [generally, $a + b(u)$, where u is the wind speed in m/s]
Φ_h, Φ_m	Stability corrections for heat and momentum, respectively
z	Height above the surface at which u is measured (m)
z_{om}, z_{oh}	Roughness lengths for momentum and heat (m), respectively
subscript i	Instantaneous values
subscript d	Daily values
subscript m	Midday values

equation for estimating evaporation that was soon adopted by hydrologists and irrigation specialists. The general form of the Penman combination equation is

$$\lambda E = -[(\Delta/(\Delta + \gamma))(R_n + G) + (\gamma/(\Delta + \gamma))6.43 W_f(e_z^o - e_z)] \tag{4}$$

where Δ is the slope of the saturation vapor pressure–temperature curve (kPa/°C), γ is the psychrometic constant (kPa/°C), W_f is a wind function [generally, $a + b(u)$, where u is the wind speed in m/s)], e_z^o and e_z are the saturation and actual vapor pressures at the z level above the surface (kPa), and $(e_z^o - e_z)$ is vapor pressure deficit (kPa).

The Penman formula was recast in terms of an aerodynamic resistance and a surface resistance for application to single leaves (Penman, 1953) and vegetation canopies (Rijtema, 1965; Monteith, 1965). This result, now referred to as the Penman–Monteith equation, is probably the most universally used equation for calculating evaporation:

$$\lambda E = -[\Delta(R_n + G) + \rho C_p(e_z^o - e_z)/r_a]/[\Delta + \gamma^*] \tag{5}$$

where ρ is air density (kg/m^3), C_p is specific heat at constant pressure (kJ/kg°C), and the aerodynamic resistance, r_a (s/m) is

$$r_a = \{[\ln((z - d_0)/z_{0m}) + \ln(z_{0m}/z_{0h}) - \Phi_h][\ln((z - d_0)/z_{0m}) - \Phi_m]\}/k^2u \tag{6}$$

and z is the height above the surface at which u is measured (m), d_0 is the displacement height (m), z_{0m} and z_{0h} are the roughness lengths for momentum and heat (m), respectively, Φ_h and Φ_m are the stability corrections for heat and momentum, respectively, and k is von Karman's constant ($\approx$0.4). The integral stability functions were summarized by Beljaars and Holtslag (1991) for the stable and unstable conditions. The value of γ^* (kPa/°C) in Eq. (5) is a function of r_a and the canopy resistance to vapor transport [r_c (s/m)], where

$$\gamma^* = \gamma(1 + r_c/r_a) \tag{7}$$

Priestley and Taylor (1972) proposed a simplified version of the Penman combination equation for computation of potential evaporation heat flux density (λE_p) for a surface that has minimal resistance to evaporation. Under these conditions, the aerodynamic component was ignored and the energy component was multiplied by a coefficient,

$$\lambda E_p = -\alpha(\Delta/(\Delta + \gamma))(R_n + G) \tag{8}$$

where $\alpha = 1.26$ for regions with no or low advective conditions.

Regional-scale estimates of evaporation have been made using properties of the atmospheric boundary layer (ABL). One approach applies similarity theory to humidity, temperature, and wind in the ABL (Brutsaert and Mawdsley, 1976). Another approach involves the development of simplified conservation equations for the ABL (McNaughton and Spriggs, 1986). This links the surface fluxes to temporal changes in temperature and humidity in the mixed layer. There are problems in employing either approach. The former has difficulties related to the specification of appropriate roughness parameters, especially in heterogeneous terrain, while the latter must develop parameterizations for advection and entrainment processes that commonly exist in the ABL.

3 CONVENTIONAL APPROACHES FOR MEASURING EVAPORATION

Theoretical developments such as those described in the previous section are generally dependent upon experimental data for verification. There are a variety of conventional approaches for measuring evaporation, ranging from simple to complex and having a range of accuracies and spatial scales.

Most simply, evaporation can be measured under field conditions by monitoring the change in soil water storage over a period of time. Though this can be accomplished fairly easily with a neutron soil water probe, this method does not account for the drainage from the zone sampled or the upward movement of water from a saturated zone into the zone sampled. Discussions of the problems encountered in determining evaporation by soil sampling were presented by Robins et al. (1954) and Jensen and Wright (1978).

Weighing lysimeters are open-top tanks filled with soil in which crops are grown under natural conditions. Evaporation from the contained soil and plants is generally determined either by weighing the entire unit with a mechanical scale or with a counterbalanced scale and load cell; the reduction in the unit's weight over time equals the rate of water transfer to the atmosphere by evaporation. For accurate results, the soil conditions within the lysimeter should be the same as those without, and the lysimeter must be surrounded by the same vegetation that is growing in the lysimeter for a desired radius of about 100 m. A detailed summary of the use of lysimeters for estimation of evaporation can be found in publications by van Bavel and Myers (1962) and Howell et al. (1985).

Commercial instrumentation is available for determining evaporation using an energy balance approach (Bowen ratio) and a mass transfer method (eddy correlation). The Bowen ratio method [based on Eqs. (1) to (3)] allows values of evaporation to be obtained hourly during daylight hours. The accuracy of the method decreases with decreasing flux of water vapor, or when there is low evaporative demand (e.g., at night). A description of the Bowen ratio equipment was provided by Spittlehouse and Black (1980) and Gay and Greenberg (1985).

The eddy correlation method was proposed by Swinback (1951) based on the theoretical description of the mean vertical flux of water vapor:

$$E = (0.622/P)\rho w'e' \tag{9}$$

where P is atmospheric pressure (kPa), w' is the instantaneous deviation of vertical wind speed from the mean vertical wind (w) at height z, and e' is the instantaneous deviation of the partial water vapor pressure from the mean at height z. Evaluation of Eq. (9) is accomplished using vertical anemometers and vapor pressure sensors with short sampling intervals (hundredths of seconds) to determine w' and e' in short, successive periods of time (tenths of seconds). This method is amenable to field use in routine measurements for extended periods, e.g., months or years (Kanemasu et al., 1979).

Other approaches that have been used to measure evaporation rates include the inflow–outflow method for monitoring evaporation from catchments (Holmes, 1984) and portable gas assimilation chambers (Reicosky, 1981). A limitation of all the techniques described in this section is that they yield essentially point values of evaporation and, therefore, are applicable only to a homogeneous area surrounding the equipment that is exposed to the same environmental factors. An evaluation of the spatial distribution of evaporation over large heterogeneous areas would be prohibitive using these conventional point measurement techniques. There are advantages and disadvantages of these conventional methods and the remote-sensing techniques discussed in the following sections. Conventional methods yield data at one location but operate continuously over time. Techniques that utilize remotely sensed inputs yield data for each resolution element of the sensor, thus spatially distributed values of evaporation, but at only an instant in time.

4 APPROACHES FOR ESTIMATING EVAPORATION USING REMOTE SENSING

An alternative means of estimating the spatial distribution of evaporation is through the use of remotely sensed images, obtained by either aircraft- or spacecraft-based sensors. Images obtained from existing satellite sensors can cover swaths ranging from 60 to 2050 km (at resolutions ranging from 10 m to 1 km) and include information about surface reflectance, temperature, and general backscatter properties (Table 2).

In this section, recent developments in the evaluation of evaporation using remotely sensed images are discussed, with emphasis on several problems that must be resolved before an operational satellite-based system for monitoring areal evaporation from land surfaces can be realized. These methods have been divided into two basic classes: (a) statistical and analytical approaches that calculate H and λE "directly" from the remote-sensing data and (b) modeling approaches that use remote-sensing data to "define" or serve as boundary conditions in the estimation of λE and H.

Determination of λE Directly from the Remote-Sensing Data

Many approaches for determination of λE directly from remote-sensing data use the surface energy balance equation as the primary boundary condition to be satisfied; that is,

$$R_n + G + H + \lambda E = 0 \tag{10}$$

where $R_n + G$ is often termed the available energy and $H + \lambda E$ are the turbulent fluxes. Evaluation of the available energy is relatively straightforward and will be addressed first, followed by the discussion of more complex evaluation of the turbulent fluxes H and λE.

Approaches for Determining Available Energy

A number of approaches using remote sensing have been developed for estimating the available energy components in Eq. (10). Generally, R_n is evaluated in terms of its four radiation components (Sellers et al., 1990), namely,

$$R_n = (1 - \acute{\alpha})R_s + \varepsilon_s R_{\mathrm{ld}} - \varepsilon_s \sigma T_{\mathrm{sh}}^4 \tag{11}$$

where R_s is the incoming shortwave solar radiation (W/m^2), R_{ld} is the incoming longwave radiation (W/m^2), $\acute{\alpha}$ is the surface shortwave albedo, ε_s is the surface

TABLE 2 Some Current Satellite-Based Sensors

		Spectral Region							
Satellite	Sensor	Reflective (μm)	Thermal (μm)	Microwave (GHz)	Pixel Resolution (PR)	Orbital Characteristics	Repeat Cycle	Time of Data Acquisition	Delivery time from acquisition to user (T_D)
GOES-8	Imager	0.52–0.72 3.8–4.0 6.5–7.0	10.2–11.2 11.5–12.5		1 km (visible) 4 km (all others)	Geostationary	Stationary	Every 30 min	Instantaneous at ground station
METEOSAT	VISSR	0.4–1.1 5.7–7.1	10.5–12.5		Acquired at 1 km Archived at 8 km	Geostationary	Stationary	Every 30 min	Instantaneous at ground station
NOAA-12,14	Advanced Very High-Resolution Radiometer (AVHRR-2)	0.58–0.68 0.725–1.1	3.55–3.93 10.5–11.5 11.5–12.5		1.1 km (local area coverage) 4 km (global area coverage)	Near-polar, sun-synchronous	12 h,every 9.2 days	19.30 (ascending) and 07.30 (descending)	Instantaneous at ground station
Landsat-5	Thematic Mapper (TM)	0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90 1.55–1.75 2.08–2.35	10.4–12.5		30 m (Vis-IR) 120 m (thermal IR)	Near-polar, sun-synchronous	16 days	Midmorning	72 hours at best, generally 2 weeks to 1 month
Landsat-7	Enhanced Thematic Mapper Plus (ETM+)	0.50–0.90 0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90 1.55–1.75 2.08–2.35	10.4–12.5		30 m (Vis-IR) 60 m (thermal, IR) 15 m (panchromatic)	Near-polar, sun-synchronous	16 days	Midmorning	48 h
SPOT-1 to *SPOT-3*	High Resolution Visible (HRV)	0.50–0.75 0.50–0.59 0.62–0.66 0.77–0.87			10 m (panchromatic) 20 m (multispectral)	Near-polar, sun-synchronous	26 days, and pointing capability provide shorter cycles	Late morning	48 hours at best, generally 2 weeks to 1 month

ERS-1 to *ERS-2*	Active Microwave (AM-I) Along-Track Scanning Radiometer (ATSR)	1.6 3.7	11 12	5.3 VV (C-band)	1 km (optical) 30 m (3 looks, SAR) 100 m (@ radiometric resolution of 1 dB)	Near-polar, sun-synchronous	3 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
RADARSAT	Synthetic Aperture Radar (SAR)			5.3 HH (C-band)	28 m (4 looks, standard product)	Near-polar, sun-synchronous	24 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
JERS-1	OPtical Sensor (OPS) Visible and Near IR (VNIR) Radiometer Short wavelength InfraRed (SWIR) Radiometer Synthetic Aperture Radar SAR)	0.52–0.60 0.63–0.69 0.76–0.86 1.60–1.71 2.01–2.12 2.13–2.25 2.27–2.40		1.275 HH (L-band)	20 m (OPSVNIR and SWIR) 18 m (3 looks, SAR)	Near-polar, sun-synchronous	44 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
Space Imaging	IKONOS	0.45–0.90 0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90			1 m (panchromatic) 4 m (multispectral)	Inclination 98.1 °, sun-synchronous	1–3 days	Late morning	24–48 h

(*continued*)

TABLE 2 (*continued*)

Satellite	Sensor	Spectral Region			Pixel Resolution (PR)	Orbital Characteristics	Repeat Cycle	Time of Data Acquisition	Delivery time from acquisition to user (T_D)
		Reflective (μm)	Thermal (μm)	Microwave (GHz)					
Terra	MOderate Resolution Imaging Spectrometer (MODIS-N)	MODIS 0.66–0.87 (2 bands) 0.47–2.13 (4 bands) 0.42–0.94 (12 bands)	3.8–14.2 (17 bands)		MODIS 0.25 km (Visible, NIR) 0.5 km (Vis, NIR, SWIR) 1 km (Vis, NIR, TIR)	Polar orbiting, sun-synchronous	MODIS 1–2 days	10:30	48 h
	Advanced Space-borne Thermal Emission and Reflectance Radiometer (ASTER)	ASTER 0.52–0.86 (3 bands) 1.60–2.43 (6 bands)	8.3–11.3 (5 bands)		ASTER 15 m (Visible, NIR) 30 m (SWIR) 90 m (Thermal)		ASTER VNIR 5 days SWIR&T 16 days MISR 9 days		
	Multiangle Imaging Spectro Radiometer (MISR)	MISR 0.40–0.88 (4 bands)			MISR 240 m, 1.92 km				

emissivity, σ is the Stefan–Boltzman constant ($5.67 \times 10^{-8}\,\mathrm{W/m^2\,K^4}$), T_{sh} is the hemispherical radiometric temperature (K) as defined by Norman and Becker (1995), so that the quantity $\varepsilon_s \sigma T_{\mathrm{sh}}^4$ represents the upwelling longwave radiation flux, R_{lu}. The radiometric temperature measured by an infrared radiometer from a space-borne platform, T_{rad}, is assumed to approximate T_{sh}.

Both R_s and $\acute{\alpha}$ have been estimated from Geosynchronous operational environmental satellites (GOES) using empirical/statistical and physically based models (Pinker et al., 1995). On a daily basis, the estimate of R_s from satellite data has an uncertainty of approximately 10%, but at shorter time scales, for example hourly, the uncertainty may be greater (probably on the order of 20 to 30%, on average), especially for partly cloudy conditions (Pinker et al., 1994). Validating R_s at hourly or shorter time scales under partly cloudy skies is especially difficult due to sampling problems associated with the limited network of ground-based measurements typically available from field experiments (Pinker et al., 1994).

Satellite estimates of the contribution of the net longwave flux at the surface have been developed using sounding data (Darnell et al., 1992). The Tiros Operational Vertical Sounder (TOVS) of the National Oceanic and Atmospheric Administration (NOAA) satellites contains infrared and microwave sensors that can be used for estimating both R_{ld} and T_{rad}. Other approaches have utilized meteorological data collected near ground level with semiempirical relationships for estimating R_{ld}, and then used T_{rad} for calculating the upwelling longwave component (Jackson et al., 1987). Sellers et al. (1990) raise the concern that estimating the four components of R_n could lead to error accumulation, especially in estimating the net longwave flux because both R_{ld} and R_{lu} are large components, so the difference would be small and prone to significant uncertainty. This has led some to estimate surface R_n from the top of the atmosphere (TOA) R_n (Pinker and Tarpley, 1988). While it has been shown that there is little correlation between surface and TOA net longwave flux (Harshvardhan et al., 1990), there is a strong correlation between R_s and R_n at the surface. This has lead to statistical approaches using slowly varying surface properties such as surface albedo and soil moisture with remotely sensed estimates of R_s for estimating R_n (Kustas et al., 1994b). Other techniques use narrow-band reflectance data and T_{rad} from aircraft and satellite-based platforms for estimating the upwelling components $\acute{\alpha} R_s$ and R_{lu} and use meteorological data for estimating the downwelling components R_s and R_{ld} (e.g., Moran et al., 1989; Daughtry et al., 1990). Comparisons with ground-based observations at meteorological time scales (i.e., half-hourly to hourly) indicate that the differences are within the uncertainty in the measurements, namely 5 to 10%.

The soil heat flux (G) can be solved as a function of the thermal conductivity of the soil and the vertical temperature gradient. This temperature gradient cannot be measured remotely, hence numerical models solve for G by having several soil layers (Campbell, 1985). This requires detailed information about soil properties. Models using routine weather data may provide satisfactory predictions of soil heat flux (e.g., Camillo, 1989). An alternative approach takes G/R_n as a constant under daytime conditions that varies as a function of the amount of vegetation cover or leaf area index (LAI), which can be estimated by use of remotely sensed vegetation indices (VI)* (Choudhury et al., 1994). Several studies have shown that the value of G/R_n typically ranges between 0.4 for bare soil and 0.05 for full vegetation cover (Choudhury et al., 1987). Observations (Clothier et al., 1986; Kustas et al., 1993a) indicate that a linear relationship between VI and G/R_n exists, although analytically it has been shown that the relationship should be nonlinear (Kustas et al., 1993a).

Statistical Approaches for Determination of λE

Statistical methods for estimating λE have mainly been developed to predict daily λE using instantaneous remote-sensing observations and assumptions about the relationship between midday H and λE and $R_n + G$. One of the most widely applied approaches, using a T_{rad} observation near midday, was pioneered by Jackson et al. (1977) whereby they observed that daily differences between λE and R_n could be approximated by this linear expression:

$$R_{n,d} + \lambda E_d = A + B(T_{\text{rad},i} - T_{a,i}) \tag{12}$$

where the subscript i and d represent instantaneous and daily values, respectively, A and B are statistical regression coefficients, and T_a is the air temperature (°C) at about 2 m above the surface. A more general form of this expression was proposed by Seguin and Itier (1983) based on theoretical and experimental observations; namely,

$$R_{n,d} + \lambda E_d = B'(T_{\text{rad},i} - T_{a,i})^n \tag{13}$$

where B' was dependent on surface roughness and the value of n depended on stability ($n = 1$ for stable and 1.5 for unstable conditions). A variant of Eq. (13) was introduced by Nieuwenhuis et al. (1985) where they replaced $T_{a,i}$ and $R_{n,d}$ with a reference canopy temperature ($T_{c,i}$) corresponding to conditions of potential λE ($\lambda E_{d,p}$). The linear form of Eq. (12) has been verified experimentally and theoretically (Carlson and Buffum, 1989; Lagouarde, 1991). Carlson et al. (1995) used a soil vegetation atmospheric transfer (SVAT) model to show that a systematic relationship exists between the B' and n parameters in Eq. (13) and fractional cover, which can be estimated with remotely sensed data. Theoretical and experimental work by Lagouarde and McAneney (1992) resulted in the derivation of an equation for estimating daily sensible heat flux (H_d) using T_{rad} measured around the time of the NOAA–AVHRR (advanced very high resolution radiometer) overpass (1400 local standard time) and maximum T_a. The equation is similar in form to Dalton's evaporation equation (see Brutsaert, 1982) and requires the determination of two empirical parameters relating instantaneous to daytime average values of wind speed and surface–air temperature differences. On a daily basis the above techniques appear to have an uncertainty of ±1 mm/day or 20 to 30%.

* Spectral vegetation indices (VI) are a ratio or linear combination of reflectances in the red and NIR wavebands that is particularly sensitive to vegetation amount (Jackson and Huete, 1991) or the amount of photosynthetically active plant tissue in the plant canopy (Wiegand et al., 1991).

The approaches described above attempt to extrapolate "instantaneous" remote-sensing observations of the derived fluxes to daily totals, which is required for many hydrological and agricultural applications. Interest in daily fluxes led Jackson et al. (1983) to develop a procedure using the assumption that the temporal trend in λE would follow the course of solar radiation during the daylight period. They showed that for a clear day the ratio of daily to midday R_s (R_{sm}) could be approximated by an analytical expression:

$$R_{sd}/R_{sm} = 2N/[\pi \sin(\pi t/N)] \tag{14}$$

where N is the daylength in hours, and t is the time starting at sunrise. Several studies have shown this technique can yield satisfactory estimates of λE using the assumed equivalence $\lambda E_d/\lambda E_m = R_{sd}/R_{sm}$ (Brutsaert and Sugita, 1992).

Experimental observations analyzed by Hall et al. (1992) suggest that the evaporative fraction [EF $= -\lambda E/(R_n + G)$] remains fairly constant over the daytime period. With this assumption, an instantaneous estimate of the fluxes and hence EF from a remote-sensing observation would have the potential to provide daily λE as long as one can estimate the daytime average available energy ($R_n + G$). Several studies have found this technique can give reasonable results with differences in daily E^* of less than 1 mm/d (Sugita and Brutsaert, 1991; Brutsaert and Sugita, 1992; Hall et al., 1992; Kustas et al., 1994a). The estimates of daily λE derived from either Eq. (14) or from assuming EF is constant, however, should be adjusted for the contribution of nighttime λE. Nighttime λE can be anywhere from 10 to 30% of the daily total (Owe and van de Griend, 1990). This percentage of the daily total will largely depend upon the climate and season. For temperate climates in the summer, 10 to 20% of the daily total is probably typical (Brutsaert and Sugita, 1992).

Recently, Zhang and Lemeur (1995) examined the underlying assumptions of both Eq. (14) and constant EF using the Penman–Monteith equation, and compared the results to measurements from a mixed agricultural and forested region during HAPEX–MOBILHY (Hydrological Atmospheric Pilot Experiment–Modelisation du Bilan Hydrique; see, e.g., André et al., 1986) under clear skies. They found that EF is fairly constant for short vegetation but may not be for forests. Furthermore, the midday values of EF tended to be smaller than the daytime average and the daytime total available energy is required to use this method. Therefore they felt the approach of Jackson et al. (1983) was more suitable since it required only one instantaneous estimate of λE and Eq. (14) to compute daily λE. However, Eq. (14) will only be suitable for clear-day conditions whereas Sugita and Brutsaert (1991) and Kustas et al. (1994a) found that EF was reasonably constant under a wider variety of conditions.

Analytical Approaches for Determination of H and λE

Price (1980) proposed a model for obtaining daily integrated fluxes directly by integrating Eq. (10) over a 24-h period with some simplifying assumptions. The result is an analytical expression for computing daily λE. It requires as primary input a 24-h max–min difference in T_{rad} and daily average climate data obtained by routine weather station observations (i.e., wind speed, air temperature, and vapor pressure). This model readily lends itself to the NOAA–AVHRR series of satellites, which provide day–night pairs of radiometric surface temperature. Further refinements to the technique were made by Price (1982) resulting in a prognostic model that appears to give appropriate λE values when compared to local estimates using standard meteorological and pan evaporation data. However, the amplitude of the max–min difference in T_{rad} is affected by more than surface soil moisture when vegetation is present and therefore it is less directly coupled to the relative magnitude of λE (Norman et al., 1995a).

Other methods generally compute λE by evaluating R_n, G and H and solving for λE by residual in Eq. (10). At least one radiometric surface temperature observation is required. Unfortunately, most of the approaches that are described below provide only an instantaneous estimate of the fluxes because these models require T_{rad}, which means that only one estimate of λE can be computed during the daytime except when using T_{rad} observations from satellites such as *GOES* or *METEOSAT.*

With R_n and G estimated by the remote-sensing methods described earlier, sensible heat flux is normally computed using the following expression:

$$H = -\rho C_p(T_{\mathrm{aero}} - T_a)/r_a \tag{15}$$

where T_{aero} is the surface aerodynamic temperature (°C) (Norman and Becker, 1995) and T_a is the air temperature (°C) either measured at screen height or the potential temperature in the mixed layer (Brutsaert and Sugita, 1991; Brutsaert et al., 1993). The resistance to heat transfer (r_a) is affected by windspeed, atmospheric stability, and surface roughness (Brutsaert, 1982).

Since T_{aero} cannot be measured by remote sensing, it is usually replaced by T_{rad}. For uniform canopy cover, the difference between T_{aero} and T_{rad} is typically less than 2°C (Choudhury et al., 1986; Huband and Monteith, 1986), but for partial vegetation cover the differences can reach 10°C (Kustas, 1990). This has forced many investigators to adjust r_a via empirical methods related to the scalar roughness for heat (Kustas et al., 1989; Sugita and Brutsaert, 1990; Kohsiek et al., 1993) or to use an additional resistance term (Stewart et al., 1994). However, these adjustments to Eq. (15) are not generally applicable because they have not been related to physical quantities causing differences between momentum and scalar transport (McNaughton and Van den Hurk, 1995). This is supported by Sun and Mahrt (1995) who analyzed T_{rad} observations collected over heterogeneous surfaces and found that existing scalar roughness parameterizations for predicting reliable H fluxes with Eq. (15) were not generally applicable. Efforts have been made to develop dual-source models (Norman et al., 1995b; Lhomme et al., 1994; Chehbouni et al., 1996) to account for differences between T_{aero} and T_{rad}, and thus avoid the need for empirical adjustments to r_a. As a result, dual-source models may have broader application for heterogeneous surfaces (Kustas et al., 1996).

In dual-source modeling approaches, the energy exchange is partitioned between the soil/substrate and the vegetation. An example of a dual-source model was presented by Norman et al. (1995b), based on the assumption that soil surface and vegetation canopy fluxes can be taken in parallel, where

$$H = H_c + H_s = -\rho C_p\{[(T_c - T_a)/r_a] + [(T_s - T_a)/(r_a + r_s)]\} \tag{16}$$

and H_c and H_s are the sensible heat fluxes from the canopy and soil, respectively, r_s is the resistance to heat flow in the boundary layer immediately above the soil surface, and T_c and T_s are the canopy and soil temperatures, respectively. Though a dual-source approach such as that presented in Eq. (16) has the advantage over single-source approaches [represented by Eq. (15)] of accounting for different sources and sinks of energy fluxes, difficulties arise in specifying the resistances to sensible and latent heat transport from the soil and vegetation. However, relatively simple parameterizations have been proposed. For example, Norman et al. (1995b) proposed that the value of r_s be computed from the equation developed by Sauer et al. (1995)

$$r_s = (a + bu_s)^{-1} \tag{17}$$

where u_s is the wind speed (m/s) about 5 cm above the soil surface, estimated with equations of Goudriaan (1977), and $a \approx 0.004\,\text{m/s}$ and $b \approx 0.012$. Further, they proposed that values of T_c and T_s be derived from T_{rad} using the expression

$$T_{\text{rad}} = [f_{\text{gr}}T_c^4 + (1 - f_{\text{gr}})T_s^4]^{1/4} \tag{18}$$

where f_{gr} is the fraction of green vegetation viewed by the radiometer; and that the absorbed net radiation by the plant canopy, R_{nc}, be partitioned between H_c and λE_c according to the Priestley–Taylor approximation (Priestley and Taylor, 1972), where

$$R_{nc} = -H_c/[1 - 1.3f_g\Delta/(\gamma + \Delta)] \tag{19}$$

where f_g is the fraction of green or actively transpiring vegetation.

A recent study by Zhan et al. (1996) compared several single- and dual-source models for computing H with T_{rad} over different land cover types. They showed that models containing the least empiricism to account for the differences between T_{rad} and T_{aero} gave the best results with differences less than 30%, on average. The dual-source model by Norman et al. (1995b) generally gave the smallest differences with measured H fluxes. The average difference was around 20%, which is considered the level of uncertainty in eddy correlation and Bowen ratio techniques for determining the surface fluxes in heterogeneous terrain (Nie et al., 1992).

Another approach to solve this problem relates to performing detailed simulations using microclimate and radiative transfer models that can predict the relationship between T_{rad} and T_{aero} as a function of surface conditions such as vegetation cover or LAI and surface soil moisture and solar zenith and azimuth angles (Prévot et al., 1994). Some preliminary results from the simulations indicate that LAI is a major factor in determining the order of magnitude of the scalar roughness needed in Eq. (15) if T_{aero} is replaced by T_{rad}. A similar result using a Lagrangian approach was obtained by McNaughton and Van de Hurk (1995) who represented the difference between momentum and scalar transport using an excess resistance term.

The analytical approaches outlined above require an estimate of T_a. Air temperature is not measured in many regions, and where it is measured it only represents local conditions near the site of the measurement and not at each satellite image pixel. With most current satellite observations of T_{rad} at the 0.10- 1-km pixel scale, significant variations in near surface meteorological conditions may exist depending on surface conditions. Methods using satellite data indicate at least ±3°C uncertainty in the estimate of T_a when compared to standard weather station observations (Goward et al., 1994). Zhan et al. (1996) showed that two-source models are generally more sensitive to errors in $T_{\text{rad}} - T_a$ than to most other model parameters; thus it is a major advantage for a model not to require a measurement of T_a. Kustas and Norman (1997) revised the Norman et al. (1995b) dual-source model for computing the turbulent fluxes without the need for T_a via the use of T_{rad} observations at two sensor viewing angles, ~0° and ~50° zenith angles. Such viewing angles from a satellite-based platform have been available from the along track scanning radiometer (ATSR) instrument aboard the *ERS-1* satellite (Prata et al., 1990; Prata, 1993). With the ATSR data, there would be no need to extrapolate T_a from a sparse network of meteorological observations to each satellite pixel, a very unreliable approach. Moreover, the model is essentially unaffected by the typical 1 to 2°C error in estimating T_{rad} from satellites. With these two attributes, the model is well suited for computing regional-scale surface fluxes with an ATSR type of sensor.

Other methods avoid the need for estimating T_a on a pixel-by-pixel basis by relying on air temperature in the ABL, which is much more uniform over a region (Brutsaert and Sugita, 1991; Brutsaert et al., 1993). However, the variability of evaporation is more difficult to quantify. Other approaches attempt to use remotely sensed data in the optical wavebands to define variation in meteorological conditions (Bastiaanssen et al., 1998; Gao et al., 1998). It remains to be seen how universal these relationships are for different climates.

Modeling Approaches That Use Remote-Sensing Data to Define Boundary Conditions

Numerical Models. Several numerical models have been developed over the past decade to simulate surface energy flux exchanges using remote sensing data (usually observations of T_{rad}) for updating the model parameters (Camillo et al., 1983; Carlson et al., 1981; Soer, 1980; Taconet et al., 1986). The advantage of these approaches is that the temporal trend of the fluxes can be simulated and periodically updated with the remote-sensing data. Taconet et al. (1986) show the feasibility of using this approach with AVHRR data and, more recently, included the geostationary satellite data (*METEOSAT*) to increase the stability of the model inversion and atmospheric correction of the satellite observations (Taconet and Vidal-Madjar, 1988).

Unfortunately, these models require many input parameters related to soil and vegetation properties not readily available at regional scales. This has prompted some to simplify numerical models in order that remote sensing could potentially be used to estimate most of them (Bougeault et al., 1991). An extreme example of this is given by Brunet et al. (1991) who use an atmospheric boundary layer (ABL) model to calculate regional-scale energy fluxes with a Penman–Montieth equation for parameterizing the energy transport across the soil–vegetation–atmosphere interface. The surface resistance is the main adjustable parameter and is adjusted in order for the model to match the early afternoon infrared surface temperature observation from the NOAA–AVHRR satellite. Preliminary tests using observations under different moisture and crop conditions and surface temperatures from ground-based stations indicate the model adequately simulates the temporal trace and magnitudes of both the energy fluxes and surface temperature.

Numerical models have several advantages over the statistical and analytical approaches. First, they typically better represent the physics of energy transport in the soil–vegetation–atmosphere system. Second, with initial and boundary conditions, they can simulate the energy fluxes continuously. Yet many numerical models still require continuous weather data such as wind speed, air temperature, and vapor pressure, or in the case of atmospheric models that can simulate the near-surface weather, they require radiation data. In practice, few of these models can be used at regional scales with remote-sensing data because of the large amount of vegetation and soils information required to evaluate necessary parameters. Some success in bridging this gap has been achieved by combining a physically based robust model simulating the energy fluxes with remote-sensing data, which provides necessary information for determining key surface parameters in an operational mode (Sellers et al., 1992; Crosson et al., 1993). Two such approaches that appear to have great potential for estimating λE operationally are discussed below in some detail.

Atmospheric Climate Models. An important conceptual step in improving the procedure for estimating soil moisture and the surface energy balance came with the idea of using the time rate of change of T_{rad} from a geostationary satellite such as *GOES* with an atmospheric boundary layer model (Wetzel et al., 1984). By using time rate of change of T_{rad}, one reduces the need for absolute accuracy in satellite sensing and atmospheric corrections, both major challenges. Diak (1990) improved this approach further with a method for partitioning the available energy ($R_n + G$) into H and λE by using the rate of rise of T_{rad} from the *GOES* satellite and ABL rise from the 12 Greenwich mean time (GMT) synoptic sounding to the 00 GMT sounding. The model is initialized with the 12 GMT sounding of temperature, humidity, and wind speed. Then the surface Bowen ratio (i.e., the ratio of the turbulent fluxes $H/\lambda E$) and the "effective" surface roughness are varied until the predicted 12-h rise in ABL height and T_{rad} match the observations. This effective surface roughness combines the effects of the surface aerodynamic roughness, viewing angle, and fractional vegetative cover. Estimates of surface albedo and emissivity are required by the model.

Diak and Whipple (1993) further refined the model by including a procedure to account for effects of horizontal and vertical temperature advection and vertical motions above the ABL. Sensitivity of the model to the determination of the surface energy balance and to the effective roughness was performed with a case study using data from the Midwest and Great Plains areas in the continental United States. They also verified their model estimates of the surface energy balance with in situ measurements from the FIFE (First ISLSCP Field Experiment; see Sellers et al., 1988) site for 2 days. The model-derived λE values were within 10% of the measurements, suggesting this technique may provide reliable λE estimates at regional scales. Additional comparisons of 12-h averages of sensible heat flux with FIFE observations support the utility of their model (see Fig. 2 from Diak et al., 1995). They also found that temperature advection usually does not significantly impact the surface energy balance estimates given by the model on a daily basis, although for areas that are routinely affected by advection the biasing could impact longer term averages of λE (i.e., at climate time scales).

In a related approach, Anderson et al. (1997) recently developed and tested a two-source surface energy balance model requiring measurements of the time rate of change of surface temperature and an early morning ABL sounding. With this model, many of the problems associated with the use of radiometric surface temperature were avoided. The model accommodated the first-order dependence of the radiometric surface temperature on view angle, avoided the need for atmospheric corrections and precise emissivity evaluation, and did not require in situ measurements of air temperature. The performance of the model was evaluated with experimental data from FIFE and from a semiarid rangeland experiment (Monsoon'90; see Kustas and Goodrich, 1994). The model yielded uncertainties in flux estimates comparable to models needing in situ air temperature observations and were comparable to the uncertainties in surface energy flux measurements.

Recognizing the fact that using T_{rad} requires detailed information on the characteristics of the surface and the structure of the overlying atmosphere, which is often incomplete for many regions, Diak et al. (1994) have proposed a method that employs the High Resolution Interferometer-Sounder (HIS) for estimating the turbulent heat fluxes, H and λE. The premise is that the temporal changes in the radiances observed by the HIS implicitly measure changes in the lower atmosphere, which are a measure of the absolute amount of energy added to the ABL. The HIS radiance changes were described by coefficients obtained by an eigenvalue decomposition procedure. These coefficients were in turn related to various components of the surface energy balance equation using multiple linear regression. Diak et al. (1994) provide convincing evidence that this method responds to temperature changes in the lower atmosphere as well as surface temperature changes. Consequently, this method is equivalent to the method of Diak (1990), but without requiring any ancillary data, just two remote radiance measurements. However, even when HIS becomes operational, co-located flux measurements will be required to establish a database to use the HIS technique. One possible solution is to identify sites that have sufficiently detailed surface information to permit some of the other techniques described above to be used to calibrate this procedure. In any event, the HIS technique offers tremendous potential since it can evaluate the surface energy balance relying only on remotely sensed data.

Alternative Approach: Exploiting the VI/T_{rad} Relation. Numerous studies have found a significant negative correlation between the normalized difference vegetation index (NDVI) and T_{rad} over a variety of surfaces (Goward et al., 1985; Hope and McDowell, 1992; Nemani and Running, 1989; Nemani et al., 1993), where

$$\text{NDVI} = (\rho_{\text{NIR}} - \rho_{\text{Red}})/(\rho_{\text{NIR}} + \rho_{\text{Red}}) \tag{20}$$

and ρ_{NIR} and ρ_{Red} are the measured reflectance factors of the surface in the near-infrared (NIR) and red spectrum, respectively. They suggest that this relationship is related to the amount of available energy partitioned into λE, which is driven by variation in transpiration or evaporative cooling. Hope et al. (1986) showed theoretically that with VI and T_{rad} one can extract canopy resistance. However, this assumes complete canopy cover, which does not usually exist in most natural land surfaces.

Nemani and Running (1989) used an ecological model for forested regions and observed a nonlinear relationship between the slope of the NDVI–T_{rad} curve and the canopy resistance. Goward and Hope (1989) also proposed that the slope was a measure of the surface resistance. These approaches will be difficult to apply to most landscapes with partial canopy cover since variability in fractional cover and surface soil moisture cause significant scatter in the VI/T_{rad} relationship. Furthermore, studies suggest that the relationship between surface resistance and the NDVI/T_{rad} slope will vary significantly with vegetation type. Nemani et al. (1993) showed that the NDVI/T_{rad} slope responded to changes in water status of forested areas, but not of the grasslands. The variability in slope for the grasslands appeared to be mainly caused by variation in fractional cover rather than in λE. Smith and Choudhury (1991) used a coupled dual-source soil–vegetation model to show that the NDVI/T_{rad} slope largely depended on whether the drying soil surface is the source of the decline in λE or whether it was the vegetation. They also observed that the linear relationship between NDVI and T_{rad} did not exist for forests but only for agricultural and native pastures.

Others have used an energy balance model for computing spatially distributed fluxes from the variability within the NDVI–T_{rad} plot from a single scene (Price, 1990). Price (1990) used NDVI to estimate the fraction of a pixel covered by vegetation. From the NDVI/T_{rad} plot Price (1990) showed how one could derive bare soil and vegetation temperatures and, with enough spatial variation in surface moisture, estimate daily λE for the limits of full cover vegetation, dry and wet bare soils.

Following Price (1990), Carlson et al. (1990, 1994) combined an ABL model with a SVAT for mapping surface soil moisture, vegetation cover, and surface fluxes. Model simulations were run for two conditions: 100% vegetative cover with the maximum NDVI being known a priori, and with bare soil conditions knowing the minimum NDVI. Using ancillary data (including a morning atmospheric sounding, vegetation and soil-type information) root-zone and surface soil moisture were varied, respectively, until the modeled and measured $T_{\rm rad}$ were closely matched for both cases, and fractional vegetated cover and surface soil moisture were derived. Further refinements to this technique have been developed by Gillies and Carlson (1995) for potential incorporation into climate models. Comparisons between modeled-derived fluxes and observations have been made recently by Gillies et al. (1997) using high-resolution aircraft-based remote-sensing measurements from a grassland ecosystem during FIFE and Monsoon'90. Approximately 90% of the variance in the fluxes was explained by the model.

In a related approach, Moran et al. (1994) defined theoretical boundaries in the SAVI/$(T_{\rm rad} - T_a)$ two-dimensional space using the Penman–Monteith equation, where SAVI is the soil-adjusted vegetation index proposed by Huete (1988). The boundaries define a trapezoid, which has at the upper two corners unstressed and stressed 100% vegetated cover and at the lower two corners wet and dry bare soil conditions. To calculate the vertices of the trapezoid, measurements of R_n, vapor pressure, T_a, and wind speed are required as well as vegetation-specific parameters; these include maximum and minimum SAVI for the full-cover and bare soil case, maximum leaf area index, and maximum and minimum stomatal resistance. Moran et al. (1994) analyzed and discussed several of the assumptions underlying the model, especially those concerning the linearity between variations in canopy–air temperature and soil–air temperatures and transpiration and evaporation. Information about λE rates is derived from the location of the SAVI/$(T_{\rm rad} - T_a)$ measurements within the date and time-specific trapezoid. This approach permits the technique to be used for both heterogeneous and uniform areas and thus does not require having a range of NDVI and surface temperature in the scene of interest as required by Carlson et al. (1990) and Price (1990). Moran et al. (1994) compared the method for estimating relative rates of λE with observations over agricultural fields and showed it could be used for irrigation scheduling purposes. More recently, Moran et al. (1996) showed that the technique had potential for computing λE over natural grassland ecosystems.

5 SYNTHESIS

In this chapter, numerous methods were reviewed for using remote sensing to estimate λE. Based on a similar review conducted by Kustas and Norman (1996), a series of issues were identified as important for remote sensing of λE from measurements, modeling studies, and theoretical considerations. A slightly revised list of these issues is included here:

1. $T_{\rm rad}$ is not equal to $T_{\rm aero}$.
2. Most models are sensitive to errors in $T_{\rm aero} - T_a$ and u, yet the measurement of T_a and u at the time and location of the $T_{\rm rad}$ observation is not typically available.

3. T_{rad} dependence on view angle cannot generally be neglected because differences in vegetation and soil temperatures can be significant depending on soil moisture conditions.
4. Thermal emissivity is only known approximately on the pixel scale.
5. Atmospheric corrections and satellite calibrations contribute significant errors in the measurements of $\rho_{\Delta\lambda}$ and T_{rad} that are not always adequately known.
6. Remote observations are instantaneous, while integrated fluxes are desired on hourly, daily, or longer time scales.
7. Satellites with larger pixel sizes (1 to 4 km) can provide sufficiently frequent observations in time (i.e., *GOES*), but may have uncertainties related to the averaging over heterogeneous subpixel areas.
8. Continuous (hourly or daily) surface flux estimates are most useful, and clouds cause remote observations to be intermittent.

Kustas and Norman (1996) provided a representative list of models using remote observations to estimate λE and attempted to characterize which of the above eight issues each of these models addressed. None of the models address *all* the important issues at the present time, but several of the models address some of the important issues (1, 3, 4, and 6). Fewer models addressed the most critical issues of spatially distributed meteorological data and atmospheric correction of satellite image data (2 and 5). Related to issue 2, meteorological data acquired at a time or location other than that of the T_{rad} or VI observation can cause substantial error in the estimate of λE. Moran and Jackson (1991) reported that errors in extrapolation of T_a greater than 1°C were unacceptable for estimation of λE using the energy balance approach. They also reported that measurements of T_a measured at 2 m height over adjacent fields of bare soil and lush vegetation differed by up to 3°C at midday. Similarly disturbing results have been reported for wind speed estimation. Rahman (1996) compared a wind speed map constructed by simple interpolation of u values from local weather stations with a map of wind speed derived from the Regional Atmospheric Modeling System (RAMS; Pielke et al., 1992) that accounted for topographic effects. The RAMS-derived map of u was a substantial improvement over the simple interpolation because it accounted for the relatively strong winds in the passes between mountain ranges and relatively light winds in the lee of the ranges.

Related to issue 5, accounting for the attenuation of the radiances received by satellite-based sensors is not a trivial matter (Kaufman, 1989; Price, 1989). In correcting thermal-infrared data, whether using radiative transfer models or split-window techniques, the uncertainty is 1 to 3°C over land surfaces (Becker and Li, 1990; Perry and Moran, 1994). Model sensitivity to such an uncertainty in T_{rad} can be significant, especially over large vegetation where errors can be $\sim$100 W/m^2 for hourly to daily time scales (Norman et al., 1995a). However, the 150 W/m^2 uncertainty in estimating sensible heat flux from radiometric surface temperature observations suggested by Sellers et al. (1995b) is in many cases two to three times larger than errors reported by other researchers (Choudhury, 1994). All the methods reviewed in this chapter are based on the assumption that accurate remotely sensed estimates of surface reflectance, temperature, and backscatter will be readily available. At this time, they are not. A primary challenge will be to improve the accuracy and consistency of remotely sensed information with an insight into the accuracy requirements of operational models and algorithms.

None of the models explicitly addressed the issue of subpixel averaging, often termed *aggregation* (issue 7). Aggregation refers to spatial averaging of some heterogeneous surface variable to obtain an effective value representative of an area. In an assessment of the state of the art in aggregation research, Michaud and Shuttleworth (1997) concluded that, over flat terrain, simple aggregation rules applied to surface properties could result in simulated values of λE within 10% of fluxes from models with full representation of heterogeneity. Furthermore, they concluded that aggregation rules for vegetation characteristics were relatively straightforward in the case of patch-scale heterogeneity (variability of 100 to 1000 m). However, mesoscale heterogeneity (10 to 100 km) in surface cover will need to be addressed through more complicated types of parameterization and, in mountainous terrain, the influence of topography on near-surface meteorology must be considered. In an aggregation study related to the use of remote-sensing data for energy balance evaluation, Moran et al. (1997a) found that aggregation of remotely sensed measurements in sparse canopies could be accomplished with little error (such as aggregation of T_{rad} from 1 m^2 to 1 km^2) but not others (such as aggregation of H to 1 km^2). Kustas and Humes (1996) applied the Norman et al. (1995b) dual-source model for computing basin-scale fluxes with T_{rad} at 120-, 1000-, and ~8000-m pixel resolution over a semiarid rangeland landscape. They found minor changes in the fluxes aggregated from the different resolutions. Sellers et al. (1995a) investigated the impact of spatial variation in topography, vegetative cover, and soil moisture on area-averaged fluxes simulated by a SVAT model over a 2×15 km domain. They found simple averages of these parameters introduced minor errors in the SVAT simulations of the area-averaged fluxes. Still, other studies (Crosson et al., 1993; Sellers et al., 1992) suggest that issue 7 may be a significant problem at the 1-km scale but may average out at the 10-km scale (Norman and Divakarla, 1995).

None of the current models address the issue of continuous surface fluxes even with clouds, but studies are in progress to combine the thermal infrared remote-sensing approaches discussed in this chapter with mesoscale models and with a simplified land–atmosphere exchange model (Anderson et al., 2000). If issues 1 to 7 are addressed adequately, issue 8 will not limit remote estimation of regional λE fluxes.

6 CONCLUDING REMARKS

All the methods and models reviewed in this chapter have potential for operational evaluation of the spatial distribution of evaporation for agricultural and hydrological applications. Toward that goal, relatively simple methods using one-time-of-day remote sensing observations for quantifying daily ET have been applied operationally (Seguin et al., 1989, 1991). However, for many regions of Earth's land surface, meteorological data (primarily wind speed and air temperature) essential for driving model computations are not available. Approaches using remotely sensed data for estimating the variation of these quantities are being developed and tested (Bastiaanssen et al., 1998; Gao et al., 1998). How reliable the algorithms are for different climatic regimes needs to be evaluated. For air temperature, another approach is in the utilization of radiometric temperature observations from significantly different view angles in a dual-source model (Kustas and Norman, 1997). SVAT models using remote-sensing observations and linked to operational climate and hydrologic models (Ottlé and Vidal-Madjar, 1994; Gillies and Carlson, 1995; Mecikalski et al., 1999; Nouvellon et al., 2001) probably have the greatest potential for operational, regional application. This is because both the surface boundary conditions and atmospheric variables are simulated over time. For heterogenous and mountainous landscapes, further work should be focused on the development of robust aggregation techniques (e.g., Shuttleworth, 1998).

One of the greatest obstacles to the assimilation of remotely sensed information in physical models has been the inherent limitations of currently available sensors. Satellite-based sensors have the advantages of good geometric and radiometric integrity; the disadvantages include fixed spectral bands that may be inappropriate for a given application, spatial resolutions too coarse or too fine for the application, long time periods between image acquisition and delivery to user, and inadequate repeat coverage due to sensor or weather limitations. With the exception of the limitations due to weather, many of the existing limitations may be resolved with the newly launched *Terra*, *Landsat-7*, and *Space Imaging* satellites (Table 2).

Regarding the effects of clouds on image acquisitions, more work should be directed toward utilizing microwave remote sensing, which has some critical advantages over the use of optical data, including little atmospheric attenuation, cloud penetration, high spatial resolution, and day/night acquisitions. Microwave data have been used to derive soil moisture and other vegetation properties (Jackson et al., 1995; Moran et al., 1997b). Microwave data have also been used for estimating the partitioning of available energy into H and λE, for estimating soil evaporation, and in determining soil surface temperatures (Kustas et al., 1993b; Chanzy and Kustas, 1995; Troufleau et al., 1994). More recently, the dual-source model of Norman et al. (1995b) was revised to use remotely sensed near-surface moisture from a passive microwave sensor for estimating the soil surface energy balance (Kustas et al., 1998). With remotely sensed images of near-surface soil moisture, land cover classification and LAI, the model was applied over a semiarid area in southern Arizona. Comparison of model-predicted fluxes simulated over the daytime period with ground observations showed good results, with 15% differences in evaporation estimates, on average. It is also shown that it may be possible to simulate the daytime fluxes with only a single microwave observation.

The development of methods for combining microwave and optical data with SVAT schemes will likely produce the greatest advancement in the quantification of spatially distributed evaporation. This requires collection of remote-sensing data in concert with ground observations as part of large-scale field projects conducted in different climatic regions. This is a critical part in the further development and validation of model algorithms. Thus the conventional approaches for estimating evaporation outlined in this chapter play a key role in this effort.

ACKNOWLEDGMENT

The authors acknowledge funding support from NASA under grant numbers NAGW-4138, NASA-S-41396-F, and NASA NAGW-2425 through the NASA Earth Science Enterprise. The authors would like to thank Drs. Bruce Goff and Martha Anderson for their comments on an early version of this chapter.

REFERENCES

Anderson, M. C., J. M. Norman, G. R. Diak, W. P. Kustas, and J. R. Mecikalski, A two-source time-integrated model for estimating surface fluxes from thermal infrared satellite observations, *Remote Sensing Environ.*, *60*, 195–216, 1997.

Anderson, M. C., J. M. Norman, T. P. Meyers and G. R. Diak, An analytical model for estimating canopy transpiration and carbon assimilation fluxes based on light-use efficiency, *Agric. For. Meteor.*, *100*, 265–289, 2000.

Andre, J. C., J. P. Goutorbe, and A. Perrier, HAPEX-MOBILHY: A hydrologic atmospheric experiment for the study of water budget and evaporation flux at the climatic scale, *Bull. Am. Meteor. Soc.*, *67*, 138–144, 1986.

Bastiaanssen, W. G. M., R. A. Feddes, and A. A. M. Holtslag, A remote sensing surface energy balance algorithm for land (SEBAL) Part 1: Formulation, *J. Hydrol.*, 212–213: 198–2121998.

Becker, F., and Z. L. Li, Towards a local split window method over land surfaces, *Int. J. Remote Sensing*, *11*, 369–393, 1990.

Beljaars, A. C. M., and A. A. M. Holtslag, Flux parameterization over land surfaces for atmospheric models, *J. Appl. Meteor.*, *30*, 327–341, 1991.

Bougeault, P., J. Noilhan, P. Lacarrer, and P. Mascart, An experiment with an advanced surface parameterization in a mesobeta-scale model. Part I: Implementation. *Monthly Weather Rev.*, *119*, 2358–237, 1991.

Bowen, I. S., The ratio of heat losses by conduction and by evaporation from any water surface, *Phys. Rev.*, *27*, 779–789, 1926.

Brunet, Y., M. Nunez, and J.-P. Lagouarde, A simple method for estimating regional evapotranspiration from infrared surface temperature data, *ISPRS J. Photogram. Remote Sensing*, *46*, 311–327, 1991.

Brutsaert, W., *Evaporation into the Atmosphere*, Reidel, Dordrecht, 1982.

Brutsaert, W., and J. A. Mawdsley, Applicability of planetary boundary layer theory to calculate regional evapotranspiration, *Water Resour. Res.*, *12*, 852–858, 1976.

Brutsaert, W., and M. Sugita, A bulk similarity approach in the atmospheric boundary layer using radiometric skin temperature to determine regional fluxes, *Boundary-Layer Meteor.*, *55*, 1–23, 1991.

Brutsaert, W., and M. Sugita, Application of self-preservation in the diurnal evolution of the surface energy budget to determine daily evaporation, *J. Geophys. Res.*, *97*(D17), 18377–18382, 1992.

Brutsaert, W., A. Y. Hsu, and T. J. Schmugge, Parameterization of surface heat fluxes above a forest with satellite thermal sensing and boundary layer soundings, *J. Appl. Meteor.*, *32*, 909–917, 1993.

Camillo, P., Estimating soil surface temperatures from profile temperature and flux measurements. *Soil Sci.*, *148*, 233–243, 1989.

Camillo, P. J., R. J. Gurney, and T. J. Schmugge, A soil and atmospheric boundary layer model for evapotranspiration and soil moisture studies, *Water Resour. Res.*, *19*, 371–380, 1983.

Campbell, G. S., *Soil Physics with Basic*, Elsevier, New York, 1985.

Carlson, T. N., and M. J. Buffum, On estimating total daily evapotranspiration from remote surface measurements, *Remote Sensing Environ.*, *29*, 197–207, 1989.

Carlson, T. N., J. K. Dodd, S. G. Benjamin, and J. N. Cooper, Satellite estimation of the surface energy balance, moisture availability and thermal inertial, *J. Appl. Meteor*, *20*, 67–87, 1981.

Carlson, T. N., E. M. Perry, and T. J. Schmugge, Remote estimation of soil moisture availability and fractional vegetation cover for agricultural fields, *Agric. For. Meteor.*, *52*, 45–69, 1990.

Carlson, T. N., R. R. Gillies, and E. M. Perry, A method to make use of thermal infrared temperature and NDVI measurements to infer soil water content and fractional vegetation cover, *Remote Sensing Rev.*, *52*, 45–59, 1994.

Carlson, T. N., W. J. Capehart, and R. R. Gillies, A new look at the simplified method for remote sensing of daily evapotranspiration, *Remote Sensing Environ.*, *54*, 161–167, 1995.

Chanzy, A., and W. P. Kustas, Evaporation monitoring over land surface using microwave radiometry, in B. J. Choudhury, Y. H. Kerr, E. G. Njoku, and P. Pampaloni (Eds.), *ESA/NASA International Workshop*, VSP, Utrecht, 1995, pp. 531–550.

Chehbouni, A., D. Lo Seen, E. G. Njoku, and B. M. Monteney, Examination of the difference between radiative and aerodynamic surface temperatures over sparsely vegetated surfaces, *Remote Sensing Environ.*, *58*, 177–186, 1996.

Choudhury, B. J., Synergism of multispectral satellite observations for estimating regional land surface evaporation, *Remote Sensing Environ.*, *49*, 264–274, 1994.

Choudhury, B. J., J. R. Reginato, and S. B. Idso, An analysis of infrared temperature observations over wheat and calculation of latent heat flux, *Agric. Forest Meteor.*, *37*, 75–88, 1986.

Choudhury, B. J., S. B. Idso, and J. R. Reginato, Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared-temperature based energy balance equation, *Agric. Forest Meteor.*, *39*, 283–297, 1987.

Choudhury, B. J., N. U. Ahmed, S. B. Idso, R. J. Reginato, and C. S. T. Daughtry, Relations between evaporation coefficients and vegetation indices studied by model simulations, *Remote Sensing Environ.*, *50*, 1–17, 1994.

Clothier, B. E., K. L. Clawson, P. J. Pinter, Jr., M. S. Moran, R. J. Reginato, and R. D. Jackson, Estimation of soil heat flux from net radiation during the growth of alfalfa, *Agric. Forest Meteor.*, *37*, 319–329, 1986.

Crosson, W. L., E. A. Smith, and H. J. Cooper, Estimation of surface heat and moisture fluxes over a prairie grassland. 4: Impact of satellite remote sensing of slow canopy variables on performance of a hybrid biosphere model, *J. Geophys. Res.*, *98*(D3), 4979–4999, 1993.

Darnell, W. L., W. F. Staylor, S. K. Gupta, N. A. Ritchey, and A. C. Wilber, Seasonal variation of surface radiation budget derived from International Satellite Cloud Climatology Project C1 data, *J. Geophys. Res.*, *97*, 15741–15760, 1992.

Daughtry, C. S. T., W. P. Kustas, M. S. Moran, P. J. Pinter, Jr., R. D. Jackson, P. W. Brown, W. D. Nichols, and L. W. Gay, Spectral estimates of net radiation and soil heat flux, *Remote Sensing Environ.*, *32*, 111–124, 1990.

Diak, G. R., Evaluation of heat flux, moisture flux and aerodynamic roughness at the land surface from knowledge of the PBL height and satellite-derived skin temperatures, *Agric. Forest Meteor.*, *52*, 181–198, 1990.

Diak, G. R., and M. A. Whipple, Improvements to models and methods for evaluating the land-surface energy balance and "effective" roughness using radiosonde reports and satellite-measured "skin" temperatures, *Agric. Forest Meteor.*, *63*, 189–218, 1993.

Diak, G. R., C. J. Scheuer, M. S. Whipple, and W. L. Smith, Remote sensing of land-surface energy balance using data from the high-resolution interferometer sounder (HIS): A simulation study, *Remote Sensing Environ.*, *48*, 106–118, 1994.

Diak, G. R., R. M. Rabin, K. P. Gallo, and C. M. U. Neale, Regional-scale comparisons of NDVI, soil moisture indices from surface and microwave data and surface energy budgets evaluated from satellite and in-situ data, *Remote Sensing Rev.*, *12*, 355–382, 1995.

Gao, W., R. L. Coultier, B. M. Lesht, J. Qui, and M. L. Wesely, Estimating clear-sky regional surface fluxes in the southern Great Plains atmospheric radiation measurement site with ground measurements and satellite observations, *J. Appl. Meteorol.*, *37*, 5–22, 1998.

Gay, L. W., and R. J. Greenberg, The AZET battery-powered Bowen ratio system, in *Proceedings of the 17th Conf. on Agric. and Forest. Meteorol.*, 21–23 May, 1985, Scottsdale, AZ. American Meteorological Society, Boston, MA, 1985, pp. 181–182.

Gillies, R. R., and T. N. Carlson, Thermal remote sensing of surface soil water content with partial vegetation cover for incorporation into climate models, *J. Appl. Meteor.*, *34*, 745–756, 1995.

Gillies, R. R., T. N. Carlson, J. Cui, W. P. Kustas, and K. S. Humes, Verification of the "triangle" method for obtaining surface soil water content and energy fluxes from remote measurements of Normalized Difference Vegetation Index (NDVI) and surface radiant temperature, *Int. J. Remote Sensing*, *18*, 3145–3166, 1997.

Goudriaan, J., *Crop Micrometeorology: A Simulation Study*, Center for Agric., 1977.

Goward, S. N., and A. S. Hope, Evapotranspiration from combined reflected solar and emitted terrestrial radiation: Preliminary FIFE results from AVHRR data, *Adv. Space Res.*, *9*, 239–249, 1989.

Goward, S., G. D. Cruickshanks, and A. Hope, Observed relation between thermal emission and reflected spectral radiance of a complex vegetated landscape, *Remote Sensing Environ.*, *18*, 137–146, 1985.

Goward, S. N., R. H. Waring, D. G. Dye, and J. Yang, Ecological remote sensing at OTTER: Satellite macroscale observations, *Ecol. Appl.*, *4*, 322–343, 1994.

Hall, F. G., K. F. Huemmrich, S. J. Geotz, P. J. Sellers, and J. E. Nickerson, Satellite remote sensing of surface energy balance: Success, failures and unresolved issues in FIFE, *J. Geophys. Res.*, *97*(D17), 19061–19090, 1992.

Harshvardhan, R. D. A., and D. A. Dazlich, Relationship between the longwave cloud radiative forcing at the surface and the top of the atmosphere, *J. Clim.*, *3*, 1435–1443, 1990.

Holmes, J. W., Measuring evapotranspiration by hydrological methods, *Agric. Water Mgmt.*, *8*, 29–40, 1984.

Hope, A. S., and T. P. McDowell, The relationship between surface temperature and a spectral vegetation index of a tallgrass prairie: Effects of burning and other landscape controls, *Int. J. Remote Sensing*, *13*, 2849–2863, 1992.

Hope, A. S., D. E. Petzold, S. N. Goward, and R. M. Ragan, Simulated relationships between spectral reflectance, thermal emissions, and evapotranspiration of a soybean canopy, *Water Resour. Bull.*, *22*, 1011–1019, 1986.

Howell, T. A., R. L. McCormick, and C. J. Phene, Design and installation of large weighing lysimeters, *Trans. Am. Soc. Agric. Eng.*, *28*, 106–112, 117, 1985.

Huband, N. D. S., and J. L. Monteith, Radiative surface temperature and energy balance of a wheat canopy. Part I: Comparison of radiative and aerodynamic canopy temperature, *Bound.-Layer Meteor.*, *36*, 1–17, 1986.

Huete, A. R., A soil-adjusted vegetation index (SAVI), *Remote Sensing Environ.*, *27*, 47–57, 1988.

Jackson, R. D., and A. R. Huete, Interpreting vegetation indices, *Prev. Vet. Med.* 11, 185–200, 1991.

Jackson, R. D., R. J. Reginato, and S. B. Idso, Wheat canopy temperature: A practical tool for evaluating water requirements, *Water Resour. Res.*, *13*, 651–656, 1977.

Jackson, R. D., J. L. Hatfield, R. J. Reginato, S. B. Idso, and P. J. Pinter, Jr., Estimates of daily evapotranspiration from one time of day measurements, *Agric. Water Mgmt.*, *7*, 351–362, 1983.

Jackson, R. D., M. S. Moran, L. W. Gay, and L. H. Raymond, Evaluating evaporation from field crops using airborne radiometry and ground-based meteorological data, *Irrig. Sci.*, *8*, 81–90, 1987.

Jackson, T. J., P. E. O'Neill, W. P. Kustas, E. Bennett, and C. T. Swift, Passive microwave observation of diurnal soil moisture at 1.4 and 2.65 GHz, in *Proceedings of the 1995 International Geoscience and Remote Sensing Symposium*, T. I. Stein (Ed.), Vol. I, Institute of Electrical and Electronics Engineers, New York, pp. 492–494, 1995.

Jensen, M. E., and J. L. Wright, The role of evapotranspiration models in irrigation scheduling, *Trans. Am. Soc. Agric. Eng.*, *21*, 82–87, 1978.

Jensen, M. E., R. D. Burman, and R. G. Allen (Eds.), *Evapotranspiration and Irrigation Water Requirements: A Manual*, No. 70, Am. Soc. Civil. Eng. (ASCE) New York, NY, 1989.

Kanemasu, E. T., M. L. Wesely, B. B. Hicks, and J. L. Heilman, Techniques for calculating energy and mass fluxes, in B. J. Barfield and J. F. Gerber (Eds.), *Modification of the Aerial Environment of Crops*, American Society of Agricultural Engineers, St. Joseph, MI, 1979, pp. 156–182.

Kaufman, Y. J. The atmospheric effect on remote sensing and its corrections, in G. Asrar (Ed.), *Theory and Applications of Optical Remote Sensing*, Wiley, New York, 1989, pp. 336–428.

Kohsiek, W., H. A. R. De Bruin, H. The, and B. van den Hurk, Estimation of the sensible heat flux of semi-arid area using surface radiative temperature measurements, *Boundary-Layer Meteor.*, *63*, 213–230, 1993.

Kustas, W. P., Estimates of evapotranspiration with a one- and two-layer model of heat transfer over partial canopy cover, *J. Appl. Meteor.*, *29*, 704–715, 1990.

Kustas, W. P., and D. C. Goodrich, Preface to Monsoon '90 Special Section, *Water Resour. Res.*, *30*, 1211–1225, 1994.

Kustas, W. P., and K. S. Humes, Variations in the surface energy balance for a semi-arid rangeland using remotely sensed data at different spatial resolutions, in J. B. Stewart, E. T. R. Engman, A. Feddes, and Y. Kerr (Eds.), *The Scaling Issue in Hydrology*, Wiley, New York, 1996, pp. 127–145.

Kustas, W. P., and J. M. Norman, Use of remote sensing for evapotranspiration monitoring over land surfaces, *Hydrol. Sci. J. Sci. Hydrol.*, *41*, 495–516, 1996.

Kustas, W. P., and J. M. Norman, A two-source approach for estimating turbulent fluxes using multiple angle thermal infrared observations, *Water Resour. Res.*, *33*, 1495–1508, 1997.

Kustas, W. P., B. J. Choudhury, M. S. Moran, R. J. Reginato, R. D. Jackson, L. W. Gay, an H. L. Weaver, Determination of sensible heat flux over sparse canopy using thermal infrared data, *Agric. Forest Meteor.*, *44*, 197–216, 1989.

Kustas, W. P., C. S. T. Daughtry, and P. J. van Oevelen, Analytical treatment of the relationships between soil heat flux/net radiation ratio and vegetation indices, *Remote Sensing Environ.*, *46*, 319–330, 1993a.

Kustas, W. P., T. J. Schmugge, K. S. Humes, T. J. Jackson, R. Parry, M. A. Weltz, and M. S. Moran, Relationships between evaporative fraction and remotely sensed vegetation index and microwave brightness temperature for semiarid rangelands, *J. Appl. Meteor.*, *32*, 1781–1790, 1993b.

Kustas, W. P., E. M. Perry, P. C. Doraiswamy, and M. S. Moran, Using satellite remote sensing to extrapolate evapotranspiration estimates in time and space over a semiarid rangeland basin, *Remote Sensing Environ.*, *49*, 275–286, 1994a.

Kustas, W. P., R. T. Pinker, T. J. Schmugge, and K. S. Humes, Daytime net radiation estimated for a semiarid rangeland basin from remotely sensed data, *Agric. Forest Meteor.*, *71*, 337–357, 1994b.

Kustas, W. P., K. S. Humes, J. M. Norman, and M. S. Moran, Single- and dual-source modeling of surface energy fluxes with radiometric surface temperature, *J. Appl. Meteor.*, *35*, 110–121, 1996.

Kustas, W. P., X. Zhan, and T. J. Schmugge, Combining optical and microwave remote sensing for mapping energy fluxes in a semiarid watershed, *Remote Sensing Environ.*, 64, 116–131, 1998.

Lagouarde, J.-P., Use of NOAA AVHRR data combined with an agrometeorological model for evaporation mapping, *Int. J. Remote Sensing*, *12*, 1853–1864, 1991.

Lagouarde, J.-P., and K. J. McAneney, Daily sensible heat flux estimation from a single measurement of surface temperature and maximum air temperature, *Boundary-Layer Meteor.*, *59*, 341–362, 1992.

Lhomme, J.-P., B. Monteny, and M. Amadou, Estimating sensible heat flux from radiometric temperature over sparse millet, *Agric. Forest Meteor.*, *68*, 77–91, 1994.

McNaughton, K. G., and T. W. Spriggs, A mixed-layer model for regional evaporation, *Boundary-Layer Meteor.*, *34*, 243–262, 1986.

McNaughton, K. G., and B. J. J. M. Van den Hurk, A "Lagrangian" revision of the resistors in the two-layer model for calculating the energy budget of a plant canopy, *Boundary-Layer Meteor.*, *74*, 262–288, 1995.

Mecikalski, J. R., G. R. Diak, M. C. Anderson and J. M. Norman, 1999. Estimating fluxes on continental scales using remotely-sensed data in an atmospheric-land exchange model. *J. Appl. Meterol.*, *38*, 1352–1369.

Michaud, J. D., and W. J. Shuttleworth, Executive summary of the Tucson Aggregation Workshop, *J. Hydrol.*, *190*, 176–181, 1997.

Monteith, J. L., Evaporation and environment, *Symp. Soc. Exp. Biol.*, *19*, 205–234, 1965.

Monteith, J. L., Evaporation and surface temperature, *Q. J. R. Meteor. Soc.*, *107*, 1–27, 1981.

Moran, M. S., and R. D. Jackson, Assessing the spatial distribution of evapotranspiration using remotely sensed inputs, *J. Environ. Q.*, *20*, 725–737, 1991.

Moran, M. S., R. D. Jackson, L. H. Raymond, L. W. Gay, and P. N. Slater, Mapping surface energy balance components by combining LANDSAT Thematic Mapper and ground-based meteorological data, *Remote Sensing Environ.*, *30*, 77–87, 1989.

Moran, M. S., T. R. Clarke, Y. Inoue, and A. Vidal, Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index, *Remote Sensing Environ.*, *49*, 246–263, 1994.

Moran, M. S., A. F. Rahman, J. C. Washburne, D. C. Goodrich, M. A. Weltz, and W. P. Kustas, Combining the Penman–Monteith equation with measurements of surface temperature and reflectance to estimate evaporation rates of semiarid grassland, *Agric. Forest Meteor.*, *80*, 87–109, 1996.

Moran, M. S., K. S. Humes, and P. J. Pinter, Jr., The scaling characteristics of remotely sensed variables for sparsely-vegetated heterogeneous landscapes, *J. Hydrol.*, *190*, 338–363, 1997a.

Moran, M. S., A. Vidal, D. Troufleau, J. Qi, T. R. Clarke, P. J. Pinter, Jr., T. Mitchell, Y. Inoue, and C. M. U. Neale, Combining multifrequency microwave and optical data for farm management, *Remote Sensing Environ.*, *61*, 96–109, 1997b.

Nemani, R. R., and S. W. Running, Estimation of regional surface resistance to evapotranspiration from NDVI and thermal-IR AVHRR data, *J. Appl. Meteor.*, *28*, 276–284, 1989.

Nemani, R., L. Pierce, S. Running, and S. Goward, Developing satellite derived estimates of surface moisture status, *J. Appl. Meteor.*, *32*, 548–557, 1993.

Nie, D., E. T. Kanemasu, L. J. Fritschen, H. L. Weaver, E. A. Smith, S. B. Verma, R. T. Field, W. P. Kustas, and J. B. Stewart, An intercomparison of surface energy flux measurement systems during FIFE 1987, *J. Geophys. Res.*, *97*(D17), 18715–18724, 1992.

Nieuwenhuis, G. J. A., E. A. Schmidt, and H. A. M. Tunnissen, Estimation of regional evapotranspiration of arable crops from thermal infrared images, *Int. J. Remote Sensing*, *6*, 1319–1334, 1985.

Norman, J. M., and F. Becker, Terminology in thermal infrared remote sensing of natural surfaces, *Remote Sensing Rev.*, *12*, 159–173, 1995.

Norman, J. M., and M. Divakarla, Scaling carbon, water and energy fluxes from 30 m to 15 km, in *Agronomy Abstracts*, American Society of Agronomy Madison, WI, 1995.

Norman, J. M., M. Divakarla, and N. S. Goel, Algorithms for extracting information from remote thermal-IR observations of the earth's surface, *Remote Sensing Environ.*, *51*, 157–168, 1995a.

Norman, J. M., W. P. Kustas, and K. S. Humes, A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature, *Agric. Forest Meteor.*, *77*, 263–293, 1995b.

Nouvellon, Y., M. S. Moran, D. Lo Seen, R. B. Bryant, W. Ni, A. Begue, A. G. Chehbouni, W. E. Emmerich, P. Heilman and J. Qi, Coupling a grassland ecosystem model with Landsat imagery for a 10-year simulation of carbon and water budgets, *Rem. Sens. Env. 78*:131–149, 2001.

Ottlé, C., and D. Vidal-Madjar, Assimilation of soil moisture inferred from infrared remote sensing in a hydrological model over the HAPEX-MOBILHY region, *J. Hydrol.*, *158*, 241–264, 1994.

Owe, M., and A. A. van de Griend, Daily surface moisture model for large area semiarid land application with limited climate data, *J. Hydrol.*, *121*, 119–132, 1990.

Penman, H. L., Natural evaporation from open water, bare soil and grass, *Proc. R. Soc. A.*, *193*, 120–145, 1948.

Penman, H. L., The physical bases of irrigation control, *Rep. 13th Int. Hort. Cong.*, *2*, 913–923, 1953.

Perry, E. M., and M. S. Moran, An evaluation of atmospheric corrections of radiometric surface temperatures for a semiarid rangeland watershed, *Water Resour. Res.*, *30*, 1261–1269, 1994.

Pielke, R. A., W. R. Cotton, R. L. Walko, C. J. Tremback, W. A. Lyons, L. D. Grasso, M. E. Nicholls, M. D. Moran, D. A. Wesley, T. J. Lee, and J. H. Copeland, A comprehensive meteorological modeling system: RAMS, *Meteor. Atmos. Phys.*, *49*, 69–91, 1992.

Pinker, R. T., and J. D. Tarpley, The relationship between the planetary and surface net radiation: An update, *J. Appl. Meteor.*, *27*, 957–964, 1988.

Pinker, R. T., W. P. Kustas, I. Laszlo, M. S. Moran, and A. R. Huete, Satellite surface radiation budgets on basin scale in semi-arid regions, *Water Resour. Res.*, *30*, 1375–1386, 1994.

Pinker, R. T., R. Frovin, and Z. Li, A review of satellite methods to derive surface shortwave irradiance, *Remote Sensing Environ.*, *51*, 108–124, 1995.

Prata, A. J., Land surface temperatures derived from the AVHRR and ATSR I: Theory, *J. Geophys. Res.*, *89*(D9): 16689–16702, 1993.

Prata, A. J., R. P. Cechet, I. J. Barton, and D. T. Llewellyn-Jones, The along track scanning radiometer for ERS-1-scan geometry and data simulation, *IEEE Trans. Geosci. Remote Sensing*, *28*, 3–13, 1990.

Prévot, L., K. T. Brunet, U. Paw, and B. Seguin, Canopy modelling for estimating sensible heat flux from thermal infrared measurements, in *Proceedings of the Workshop on Thermal Remote Sensing of the Energy and Water Balance over Vegetation in Conjunction with Other Sensors*, Cemagref-Engref, Montpellier, France, 1994, pp. 17–26.

Price, J. C., Thc potcntial of remotely sensed thermal infrared data to infer surface soil moisture and evaporation, *Water Resour. Res.*, *16*, 787–795, 1980.

Price, J. C., Estimation of regional scale evaporation through analysis of satellite thermal-infrared data, *IEEE Trans. Geosci. Remote Sensing*, *GE-20*, 286–292, 1982.

Price, J. C., Quantitative aspects of remote sensing in the thermal infrared, in G. Asrar (Ed.), *Theory and Applications of Optical Remote Sensing*, Wiley, New York, 1989, pp. 578–603.

Price, J. C., Using spatial context in satellite data to infer regional scale evapotranspiration, *IEEE Trans. Geosci. Remote Sensing*, *GE-28*, 940–948, 1990.

Priestley, C. H. B., and R. J. Taylor, On the assessment of surface heat flux and evaporation using large scale parameters, *Monthly Weather Rev.*, *100*, 81–102, 1972.

Rahman, A. F., Monitoring regional-scale surface hydrologic processes using satellite remote sensing, Ph.D. dissertation, University of Arizona, Department of Soil and Water Science, Tucson, AZ, 1996.

Reicosky, D. C., A research tool for evapotranspiration measurements for model validation and irrigation scheduling, in *Irrigation Scheduling for Water and Energy Conservation in the 80's, Proc. of the Am. Soc. of Agric. Engineers Irrig. Scheduling Conf.*, December 1981, ASAE, Chicago, IL, 1981, pp. 18–26.

Rijtema, P. R., An analysis of actual evapotranspiration, *Agric. Res. Rep.*, *659*, 1–107, 1965.

Robins, J. R., W. O. Pruitt, and W. H. Gardner, Unsaturated flow of water in field soils and its effect on soil moisture investigations, *Soil. Sci. Soc. Am. Proc.*, *18*, 344–347, 1954.

Sauer, T. J., J. M. Norman, C. B. Tanner, and T. B. Wilson, Measurement of heat and vapor transfer coefficients at the soil surface beneath a maize canopy using source plates, *Agric. Forest Meteor.*, *75*, 161–189, 1995.

Seguin, B., and B. Itier, Using midday surface temperature to estimate daily evaporation from satellite thermal IR data, *Int. J. Remote Sensing*, *4*, 371–383, 1983.

Seguin, B., E. Assad, J. P. Freaud, J. Imbernon, Y. H. Kerr, and J. P. Lagouarde, Use of meteorological satellites for rainfall and evaporation monitoring, *Int. J. Remote Sensing*, *10*, 847–854, 1989.

Seguin, B., J.-P. Lagouarde, and M. Saranc, The assessment of regional crop water conditions from meteorological satellite thermal infrared data, *Remote Sensing Environ.*, *35*, 141–148, 1991.

Sellers, P. J., F. G. Hall, G. Asrar, D. E. Strebel, and R. E. Murphy, The first ISLSCP field experiment (FIFE), *Bull. Am. Meteor. Soc.*, *69*, 22–27, 1988.

Sellers, P. J., S. I. Rasool, and H.-J. Bolle, A review of satellite data algorithms for studies of the land surface, *Bull. Am. Meteor. Soc.*, *71*, 1429–1447, 1990.

Sellers, P. J., M. D. Heiser, and F. G. Hall, Relations between surface conductance and spectral vegetation indices at intermediate ($100\,m^2$ to $15\,km^2$) length scales, *J. Geophys. Res.*, *97*(D17), 19033–19059, 1992.

Sellers, P. J., M. D. Heiser, F. G. Hall, S. J. Goetz, D. E. Strebel, S. B. Verma, R. L. Desjardins, P. M. Schuepp, and J. I. MacPherson, Effects of spatial variability in topography, vegetation cover and soil moisture on area-averaged surface fluxes: A case study using FIFE 1989 data, *J. Geophys. Res.*, *100*(D12), 25607–25629, 1995a.

Sellers, P. J., B. W. Meeson, F. G. Hall, G. Asrar, R. E. Murphy, R. A. Schiffer, F. P. Bretherton, R. E. Dickinson, R. G. Ellingson, C. B. Field, K. F. Huemmrich, C. O. Justice, J. M. Melack, N. T. Roulet, D. S. Schimel, and P. D. Try, Remote sensing of the land surface for studies of global change: Models–algorithms–experiments, *Remote Sensing Environ.*, *51*, 1–17, 1995b.

Shuttleworth, W. J., Aggregation algorithms, *Q. J. R. Meteor. Soc.*, 1998.

Smith, R. C. G., and B. J. Choudhury, Analysis of normalized difference and surface temperature observations over southeastern Australia, *Int. J. Remote Sensing*, *12*, 2021–2044, 1991.

Soer, G. J. R., Estimation of regional evapotranspiration and soil moisture conditions using remotely sensed crop surface temperatures, *Remote Sensing Environ.*, *9*, 27–45, 1980.

Spittlehouse, D. L., and T. A. Black, Evaluation of the Bowen ratio/energy balance method for determining forest evapotranspiration, *Atmos.-Ocean*, *18*, 98–116, 1980.

Stewart, J. B., W. P. Kustas, K. S. Humes, W. D. Nichols, M. S. Moran, and H. A. R. de Bruin, Sensible heat flux-radiometric surface temperature relationship for eight semiarid areas, *J. Appl. Meteor.*, *33*, 1110–1117, 1994.

Sugita, M., and W. Brutsaert, Regional surface fluxes from remotely sensed skin temperature and lower boundary layer measurements, *Water Resour. Res.*, *26*, 2937–2944, 1990.

Sugita, M., W. Brutsaert, Daily evaporation over a region from lower boundary layer profiles, *Water Resour. Res.*, *27*, 747–752, 1991.

Sun, J., and L. Mahrt, Determination of surface fluxes from the surface radiative temperature, *J. Atmos. Sci.*, *52*, 1096–1106, 1995.

Swinback, W. C., The measurement of vertical transfer of heat and water vapour by eddies in the lower atmosphere, *J. Meteor.*, *8*, 135–145, 1951.

Taconet, O., and D. Vidal-Madjar, Applications of a flux algorithm to a field-satellite campaign over vegetated area, *Remote Sensing Environ.*, *26*, 227–239, 1988.

Taconet, O., T. Carlson, R. Bernard, and D. Vidal-Madjar, Evaluation of a surface/vegetation parameterization using satellite measurements of surface temperature, *J. Clim. Appl. Meteor.*, *25*, 1752–1767, 1986.

Troufleau, D., A. Vidal, A. Beaudoin, M. S. Moran, M. A. Weltz, D. C. Goodrich, J. Washburne, and A. F. Rahman, Using optical-microwave synergy for estimating surface energy fluxes over semi-arid rangeland, in *Proceedings of Physical Measurements and Signatures in Remote Sensing*, 17–21 January 1994, Intl. Soc. of Photogrammetry and Remote Sensing (ISPRS), Val d'Isere France, 1994, pp. 1167–1174.

van Bavel, C. H. M., and L. E. Myers, An automatic weighing lysimeter, *Agric. Eng.*, *43*, 580–583, 586–588, 1962.

Wetzel, P. J., D. Atlas, and R. Woodward, Determining soil moisture from geosynchronous satellite infrared data: A feasibility study, *J. Clim. Appl. Meteor.*, *23*, 375–391, 1984.

Wiegand, C. L., A. J. Richardson, D. E. Escobar, and A. H. Gerbermann, Vegetation indices in crop assessments, *Remote Sensing Environ.*, *35*, 105–119, 1991.

Zhan, X., W. P. Kustas, and K. S. Humes, An intercomparison study on models of sensible heat flux over partial canopy surfaces with remotely sensed surface temperature, *Remote Sensing Environ.*, *58*, 242–256, 1996.

Zhang, L., and R. Lemeur, Evaluation of daily evapotranspiration estimates from instantaneous measurements, *Agric. Forest Meteor.*, *74*, 139–154, 1995.

CHAPTER 27

INFILTRATION AND SOIL MOISTURE PROCESSES

PAUL R. HOUSER

Infiltration is the process of water entry from surface sources such as rainfall, snowmelt, or irrigation into the soil. The infiltration process is a component in the overall unsaturated *redistribution* process (Fig. 1)[1] that results in *soil moisture* availability for use by vegetation transpiration, exfiltration (or evaporation) processes, chemical transport, and groundwater recharge. Soil moisture, in turn, controls the partitioning of subsequent precipitation into infiltration and runoff, and the partitioning of available energy between sensible and latent heat flux.

Because of the importance of soil moisture on multiple processes, its definition can be elusive[2]; however, it is most often described as moisture in the unsaturated surface layers (first 1 to 2 m) of soil that can interact with the atmosphere through evapotranspiration and precipitation.[3]

1 CONTROLS ON INFILTRATION AND SOIL MOISTURE

To characterize soil moisture and infiltration, the physical controls on these processes must be considered. The primary soil controls will be considered in this chapter; however, other factors such as soil chemistry, thickness, soil layering or horizons, and preferential flow paths, as well as vegetation cover, tillage, roughness, topography, temperature, and rainfall intensity also exert important controls.[4]

A soil's particle size distribution has a large impact on its hydraulic properties. Soil particles less than 2 mm in diameter are divided into three texture groups (sand, silt, and clay) that help to classify broad soil types and soil water responses (Fig. 2).[5] The type of clay and the coarse material over 2 mm in diameter can also have a

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

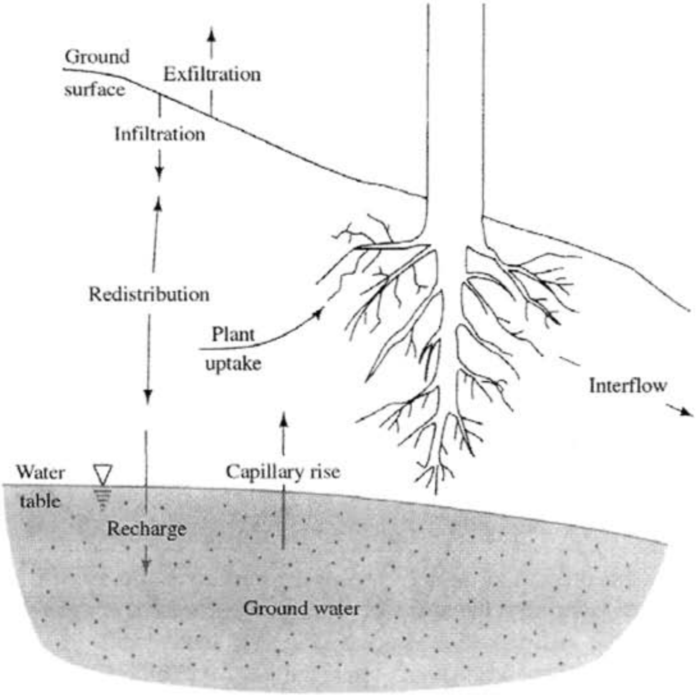


Figure 1 Unsaturated zone definition and active processes.[1]

significant impact on soil water properties. An overview of methods for determining particle size properties is given by Gee and Bauder.[6]

Bulk density, ρ_b $(\mathrm{M/L^3})$ is the ratio of the weight of dry solids to the bulk volume of the soil, and *porosity*, φ $(\mathrm{M^3/M^3})$, is the total volume occupied by pores per unit volume of soil:

$$\varphi = \frac{V_a + V_w}{V_s} = 1 - \frac{\rho_b}{\rho_m} \tag{1}$$

where V_s $(\mathrm{L^3})$ is the total volume of soil, V_a $(\mathrm{L^3})$ is the volume of air, V_w $(\mathrm{L^3})$ is the volume of water, and ρ_m $(\mathrm{M\,L^{-3}})$ is the particle density (normally about $2.65\ \mathrm{g/cm^3}$).

The volumetric water content, or soil moisture, θ $(\mathrm{L^3L^{-3}})$ is the ratio of water volume to soil volume:

$$\theta = \frac{V_w}{V_s} = \frac{W_w}{W_d}\frac{\rho_b}{\rho_w} \tag{2}$$

where W_w (M) is the weight of water, W_d (M) is the weight of dry soil, and ρ_w $(\mathrm{M/L^3})$ is the density of water. Soil moisture can vary in both time and space, with a

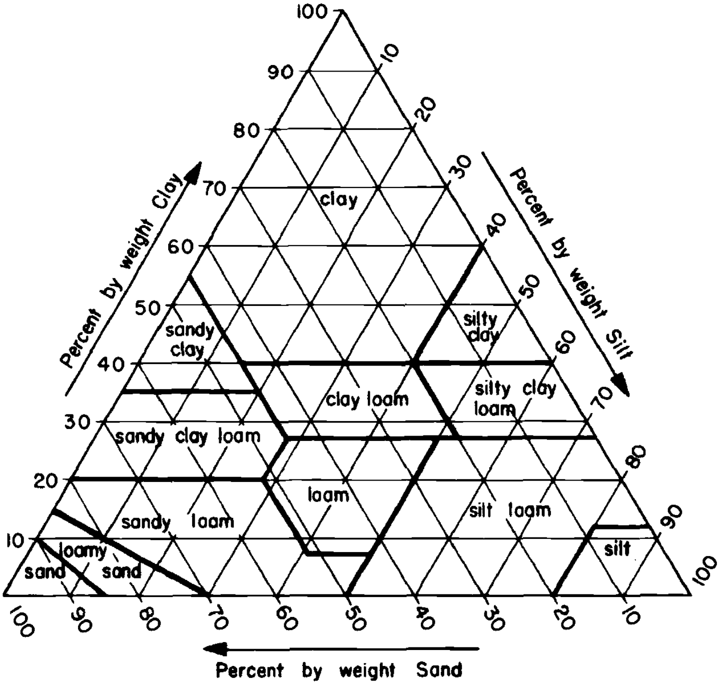


Figure 2 Soil textural triangle describing the relationship between texture and particle size distribution.[5]

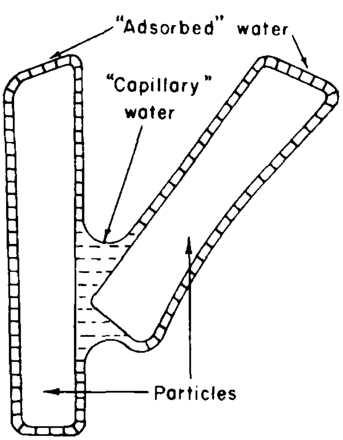


Figure 3 Capillarity and adsorption combine to produce suction.[7]

theoretical range from 0 to φ, but for natural soils the range is significantly reduced due to isolated pore space and tightly held or "adsorbed" water (Fig. 3).[7] If a soil is saturated, then allowed to drain until the remaining water held by surface tension is in equilibrium with gravitational forces, it is at *field capacity*, θ_f. Vegetation can remove water from the soil until the *permanent wilting point*, θ_w, is reached. Therefore, the *available water content* for plant use, $\theta_a = \theta_f - \theta_w$. Typical ranges of porosity, field capacity, and wilting point for different soils are given in Fig. 4.[8]

In unsaturated soils, water is held in the soil against gravity by surface tension (Fig. 3). This tension, suction, or *matric potential*, ψ (L), increases as the radii of curvature of the meniscus or water content decreases (Fig. 5).[9] Matric potential is expressed in reference to atmospheric pressure, so for saturated soil $\psi = 0$ and for unsaturated soil $\psi < 0$.

The *hydraulic conductivity*, *K* (L/T), is a measure of the ability of the soil to transmit water that varies nonlinearly over a large range depending on both soil

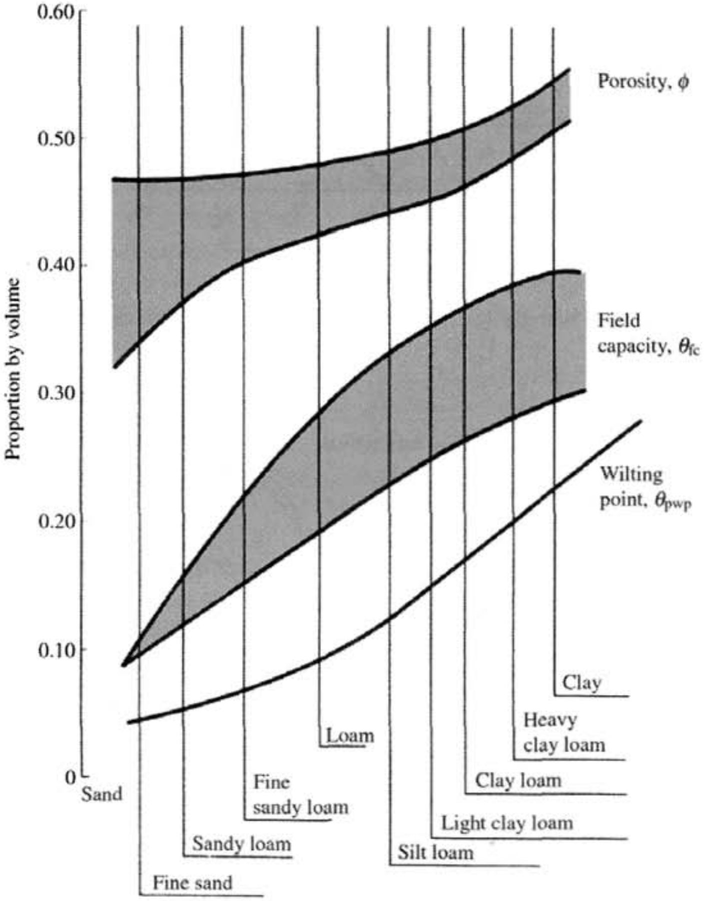


Figure 4 Water holding properties of various soils.[8]

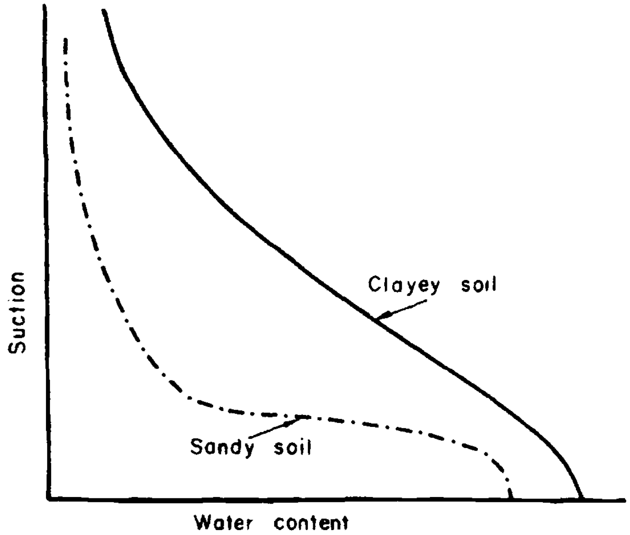


Figure 5 Effect of texture on water retention characteristics.[9]

properties and water content (Fig. 6).[10] Many laboratory and field hydraulic conductivity measurement methods exist for use with various soils; see Bouwer and Jackson[11] or Green et al.[12] for details.

Soil water content can significantly impact infiltration by (1) increasing the hydraulic conductivity, which increases infiltration, and (2) reducing the surface tension that draws moisture into the soil, which reduces infiltration. The net effect of these impacts depends on the water content itself, the water input rate, and duration and the distribution of hydraulic conductivity.

The *water retention characteristic* describes a soil's ability to store and release water and is defined by the relationship between soil moisture and the matric potential (Fig. 5). This is a power function relationship that has been described by Brooks and Corey[13] and Van Genuchten,[14] among others. The water tension characteristic is usually measured in air pressure chambers where the water content of a soil sample can be monitored over a wide pressure range.[15]

The water retention relationship may actually change between drying and wetting due to the entrapment of air in soil pores (Fig. 7).[16] For practical applications, this effect, called *hysteresis*, is usually neglected.[17]

2 PRINCIPLES OF SOIL WATER MOVEMENT

Through experiments on saturated water flow through sand beds, Darcy[18] found that the rate of flow, Q (L^3/T), through a cross-sectional area A (L^2), is directly propor-

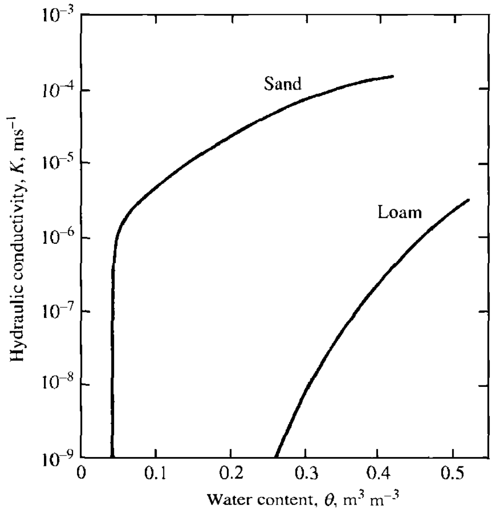


Figure 6 Effect of texture and soil moisture on hydraulic conductivity.[10]

tional to head loss (e.g., water elevation difference), ΔH (L), and inversely to the flow path length, Δl (L):

$$Q = KA\frac{\Delta H}{\Delta l} \tag{3}$$

Combining *Darcy's law* with the law of conservation of mass results in a description of unsaturated flow called *Richards equation*[19]:

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial z}\left(\frac{K}{C}\frac{\partial\theta}{\partial z}\right) - \frac{\partial K}{\partial z} \tag{4}$$

where $C = -\partial\theta/\partial\psi$ is the water content change in a unit soil volume per unit matric potential, ψ change. The Richards equation is the basis for most simulations of infiltration and redistribution of water in unsaturated soil. Using some approximations, analytical solutions of the Richards equation are available[20,21] that show good agreement with observations.[22] The Richards equation is based on saturated flow theory, and does not account for all of the processes active in natural systems, so it may not always perform well.[23]

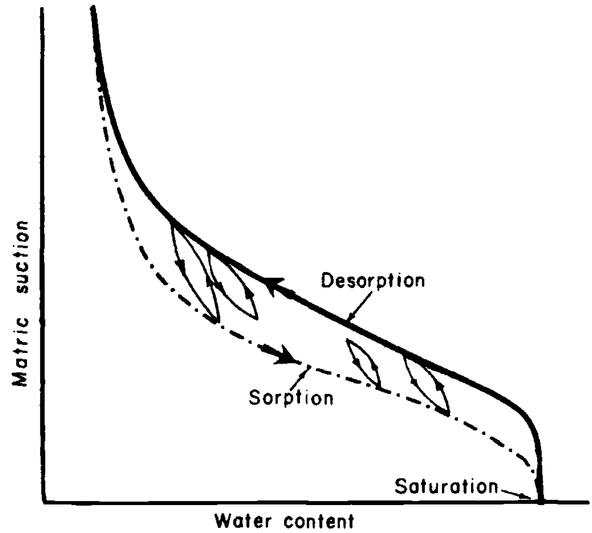


Figure 7 Changes in water retention characteristics between sorption and desorption.[16]

3 INFILTRATION ESTIMATION

Some basic principles that govern the movement of water into the soil can be used to predict infiltration. The *infiltration capacity, f* (L), is the maximum rate that a soil in a given condition can absorb water and generally decreases as soil moisture increases. If the *rainfall rate* is less than the infiltration capacity, then infiltration proceeds at the capacity rate. However, if the rainfall rate exceeds the infiltration capacity, then infiltration proceeds at the capacity rate, and the excess rainfall ponds on the surface or runs off. As the time from the onset of rainfall increases, infiltration rates decrease due to soil moisture increases, raindrop impact, and the clogging of soil pores, until a steady-state infiltration rate is reached (Fig. 8).[24] Existing infiltration models use empirical, approximate, or physical approaches to predict infiltration.[25]

Empirical. Empirical infiltration models generally utilize a mathematical function whose shape as a function of time, t, matches observations and then attempts a physical explanation of the process.

Kostiakov[26] proposed the simple infiltration rate, f (L/T) model:

$$f = \alpha\gamma t^{\alpha-1} \tag{5}$$

where α and γ are constants that have no particular meaning and must be evaluated by fitting the model to experimental data.

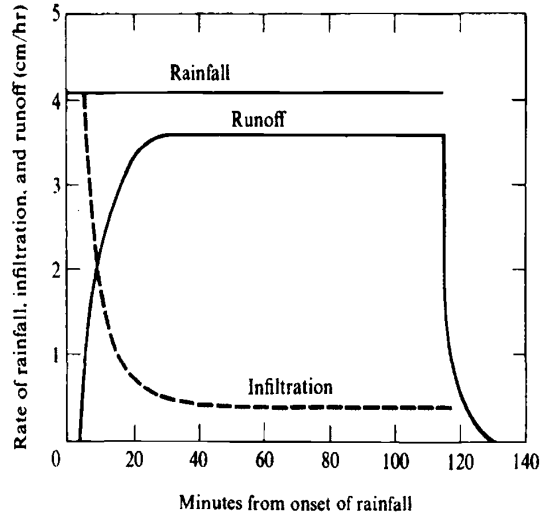


Figure 8 Idealized relationship between rainfall, infiltration, and runoff rates.[24]

Horton's[27] infiltration model has been widely used in hydrologic simulation. It relates infiltration capacity to initial infiltration rate, and f_0, the constant infiltration rate at large times, f_c:

$$f = f_c + (f_0 - f_c)e^{-\beta t} \tag{6}$$

where β is a soil parameter describing the rate of decrease of infiltration.

Approximate. Analysis approximations to the Richards equation are possible if several simplifying assumptions are made. Most approximate infiltration models treat the soil as a semi-infinite medium, with the soil saturating above a wetting front.

Green and Ampt[28] assumed in a soil with constant hydraulic properties, the matric potential at the moving wetting front is constant, leading to a discontinuous change in soil moisture at the wetting front:

$$f = K\left[1 + \frac{(\varphi - \theta_i)S_f}{F}\right] \tag{7}$$

where S_f (L) is the effective suction at the wetting front, θ_i is the initial water content, and F (L) is the accumulated infiltration.

Phillip[29] proposed that the first two terms in a series of powers of $t^{1/2}$ could be used to approximate infiltration:

$$f = \tfrac{1}{2}St^{1/2} + A \tag{8}$$

where S is a parameter called sorptivity, t is time from ponding, and A is a constant that depends on soil properties. In this model, the infiltration rate approaches a constant equal to the hydraulic conductivity at the surface water content, and the wetting front advances without changing its shape and approaches a constant velocity.

Physical. Recent advances in numerical methods and computing has facilitated the practical application of the Richards equation to realistic flow problems. Such packages can simulate water infiltration and redistribution using the Richards equation and including precipitation, runoff, drainage, evaporation, and transpiration processes.[30]

4 INFILTRATION MEASUREMENT

Infiltration rates can be measured at a point using a variety of methods described here, each appropriate for certain conditions. However, because of the large temporal and spatial variability of infiltration processes, catchment average infiltration rates may be desired, which can be obtained through the water balance analysis of rainfall–runoff observations.[31]

Ring Infiltrometer. This simple method is most appropriate for flood irrigation or pond seepage infiltration. A cylindrical metal ring is sealed at the surface and flooded. Intake measurements are recorded until steady-state conditions are reached.[32] If the effects of lateral flow are significant, then a double-ring infiltrometer can be used. Due to ponding conditions within the ring, observed infiltration rates are often higher than under natural conditions.[33]

Sprinkler Infiltrometer. This method is appropriate for quantifying infiltration from rainfall. Artificial rainfall simulators are used to deliver a specified rainfall rate to a well-defined plot. Runoff from the plot is measured, allowing computation of the infiltration rate.[34,35]

Tension Infiltrometer. The tension or disk infiltrometer employs a soil contact plate and a water column that is used to control the matric potential of the infiltrating water. By varying the tension, the effect of different size macropores can be determined.[36,37]

Furrow Infiltrometer. This method is useful if information on infiltration of flowing water in irrigation furrows is desired. Either the water added to a small section of blocked off furrow to maintain a constant depth or the inflow–outflow of a furrow segment can be monitored to determine the infiltration characteristics of the system.[38]

5 SOIL MOISTURE MEASUREMENT

Soil water content can be determined directly using gravimetric techniques or indirectly by inferring it from a property of the soil.[39,40]

Gravimetric. The oven-drying soil moisture measurement technique is the standard for calibration of all other methods but is time consuming and destructive. The method involves obtaining a wet soil sample weight, drying the sample at 105°C for 24 h, then obtaining the dry sample weight [see Eq. (2)].

Neutron Thermalization. High-energy neutrons are emitted by a radioactive source into the soil and are preferentially slowed by hydrogen atoms. The number of slow neutrons returning to the detector are a measure of soil moisture.

Gamma Attenuation. The attenuation in soil of gamma rays emitted from caesium-137 is directly related to soil density. If the soil's bulk density is assumed to be constant, then changes in attenuation reflect changes in soil moisture.[41]

Time-Domain Reflectometry (TDR). TDR measures the soil's dielectric constant, which is directly related to soil moisture, by measuring the transmit time of a voltage pulse applied to a soil probe.

Tensiometric Techniques. This method measures the capillary or moisture potential through a liquid-filled porous cup connected to a vacuum gage. Conversion to soil moisture requires knowledge of the water retention characteristic.

Resistance. The electrical resistance or conductivity of a porous block (nylon, fiberglass, or gypsum) imbedded in the soil depends primarily on the water content of the block. However, because of salinity and temperature sensitivity, measurements of these sensors are of limited accuracy.[42]

Heat Dissipation. Changes in the thermal conductivity of a porous block imbedded in the soil depend primarily on the water content of the block. The dissipation of a heat pulse applied to the block can be monitored using thermistors, then the soil water content can be determined from calibration information.

Remote Sensing. Soil moisture can be remotely sensed with just about any frequency where there is little atmospheric absorption.[43] But, it is generally accepted that long wavelength, passive microwave sensors have the best chance of obtaining soil moisture measurements that contain little error introduced by vegetation and roughness and offer great potential to remotely sense soil moisture content with depth due to differential microwave absorption with varying dielectric constant.[44]

6 SPATIAL AND TEMPORAL VARIABILITY

Natural soils exhibit considerable spatial heterogeneity in both the horizontal and vertical directions, and at all distance scales from the pore to the continent, to a degree that it is difficult to capture this variability in routine measurements.[45,46] This large variation in soil properties, infiltration, and soil moisture over relatively small areas makes it difficult to transfer the understanding of processes developed at a point to catchment scales. Many hydrological models assume that a single spatially representative average soil property can be used to characterize catchment (or even larger) scale processes. It is clear from the nonlinear character of soil water processes [Eq. 94)] that catchment average infiltration cannot be computed based on catchment average soil properties. It is also clear that the physical meaning of a soil property, say porosity, is relative to the volume over which it is averaged.[47] However, there is a need to understand and reduce this complexity for the purposes of prediction and management. Several approaches, including dividing the catchment into hydrologically similar subareas,[48] various statistical approaches,[49] and scaling and similarity theory[50,51] have made headway toward an understanding of infiltration and soil moisture spatial variability, but are not being widely used in practical applications.

One of the most important recent findings in this regard is the scale invariance of soil water behavior. If a heterogenous field is the union of homogenous spatial domains, each with associated characteristic length scales, then heterogeneity simplifies into the spatial variability of these length scales, while the functional relationships that describe soil water movement (i.e., the Richards equation) remain uniform across spatial scales.[52] This new understanding of the underlying symmetry of the Richards equation may help to facilitate a workable scale invariant analytical soil water dynamical model.

Finally, there is a continuing need for the observation of soil properties, soil moisture, and infiltration processes at multiple scales to facilitate understanding and prediction of these complex and socially significant processes. It is likely that remote sensing of soil moisture and other land surface factors will be instrumental in this respect.

REFERENCES

1. Dingman, S. L., *Physical Hydrology*, Prentice-Hall, Englewood Cliffs, NJ, 1994, p. 211.
2. Lawford, R. G., An overview of soil moisture and its role in the climate system, in F. J. Eley, R. Granger, and L. Martin (Eds.), *Soil Moisture; Modeling and Monitoring for Regional Planning*, National Hydrology Research Centre Symposium No. 9 Proceedings, 1992, pp. 1–12.
3. Schmugge, T., T. J. Jackson, and H. L. McKim, Survey of in-situ and remote sensing methods for soil moisture determination, in *Satellite Hydrology*, American Water Resources Association, 1979.
4. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and soil water movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.1–5.51.

5. Hillel, D., *Introduction to Soil Physics*, Academic, New York, 1982, p. 29.
6. Gee, G. W., and J. W. Bauder, Particle size analysis, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 383–411.
7. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 143.
8. Dunne, T., and L. Leopold, *Water in Environmental Planning*, W. H. Freeman, New York, 1978, p. 175.
9. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 150.
10. Marshall, T. J., and J. W. Holmes, *Soil Physics*, 2nd ed., Cambridge University Press, 1988, p. 87.
11. Bouwer, H., and R. D. Jackson, Determining soil properties, in J. van Schilfgaard (Ed.), *Drainage for Agriculture*, American Society of Agronomy, Madison, WI, 1974, pp. 611–672.
12. Green, R. E., L. R. Ahuja, and S. K. Chong, Hydraulic conductivity, diffusivity, and sorptivity of unsaturated soils—field methods, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 771–798.
13. Brooks, R. H., and A. T. Corey, *Hydraulic Properties of Porous Media*, Hydrology Paper 3, Colorado State University, Fort Collins, CO, 1964.
14. Van Genuchten, M. Th., A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, *44*, 892–898, 1980.
15. Cassel, D. K., and A. Klute, Water potential: Tensiometry, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 563–596.
16. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 153.
17. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and Soil Water Movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.5–5.6.
18. Darcy, H., *Les fontaines publiques de la ville de Dijon*, Dalmont, Paris, 1856.
19. Richards, L. A., Capillary conduction of liquids in porous mediums, *Physics*, *1*, 318–333, 1931.
20. Kuhnel, V., V. C. I. Dooge, G. C. Sander, and J. P. J. O'Kane, Duration of atmosphere-controlled and soil-controlled phases of infiltration for constant rainfall at a soil surface, *Ann. Geophys.*, *8*, 11–20, 1990.
21. Sposito, G., Recent advances associated with soil water in the unsaturated zone, in *Reviews of Geophysics*, Supplement, U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, 1995, pp. 1059–1065.
22. Whisler, F. D., and H. Bouwer, A comparison of methods for calculating vertical drainage and infiltration in soils, *J. Hydrol.*, *10*, 1–19, 1970.
23. Nielsen, D. R., J. W. Biggar, and J. M. Davidson, Experimental consideration of diffusion analysis in unsaturated flow problems, *Soil Soc. Am. Proc.*, *26*, 107–111, 1962.
24. Dunne, T., and L. Leopold, *Water in Environmental Planning*, W. H. Freeman, New York, 1978, p. 169.
25. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and soil water movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.21–5.23.

26. Kostiakov, A. N., On the dynamics of the coefficient of water-percolation in soils and on the necessity for studying it from a dynamic point of view for purposes of amelioration, *Trans. Sixth Comm. Intern. Soil. Sci. Soc. Russian*, part A, 1932, pp. 17–21.
27. Horton, R. E., An approach toward a physical interpretation of infiltration-capacity, *Soil Sci. Soc. Am. J.*, *5*, 399–417, 1940.
28. Green, W. H., and G. A. Ampt, Studies on soil physics: 1. Flow of air and water through soils, *J. Agric. Sci.*, *4*, 1–24, 1911.
29. Phillip, J. R., The theory of infiltration: 1. The infiltration equation and its solution, *Soil Sci.*, *83*, 345–357, 1957.
30. Ross, O. J., Efficient numerical methods for infiltration using Richards equation, *Water Resour. Res.*, *26*, 279–290, 1990.
31. Soil Conservation Service, Hydrology, in *SCS National Engineering Handbook*, U.S. Department of Agriculture, Washington, DC, 1972, Sec. 4.
32. Bouwer, H., Intake rate: Cylinder infiltrometer, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 825–844.
33. Lukens, R. P. (Ed.), *Annual Book of ASTM Standards*, Part 19: *Soil and Rock, Building Stones*, 1981, pp. 509–514.
34. Agassi, M., I. Shainberg, and J. Morin, Effects on seal properties of changes on drop energy and water salinity during a continuous rainstorm, *Aust. J. Soil Res.*, *26*, 1–10, 1988.
35. Peterson, A., and G. Bubenzer, Intake rate sprinkler infiltrometer, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 845–870.
36. Ankeny, M. D., T. C. Kaspar, and R. Horton, Design for an automated tension infiltrometer, *Soil Sci. Soc. Am. J.*, *52*, 893–896, 1988.
37. Perrouix, K. M., and I. White, Designs of disc permeameters, *Soil Sci. Soc. Am. J.*, *52*, 1205–1215, 1988.
38. Kincaid, D. C., Intake rate: Border and furrow, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 871–887.
39. Gardner, W. H., Water content, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 493–544.
40. Schmugge, T. J., T. J. Jackson, and H. L. McKim, Survey of methods for soil moisture determination, *Water Resour. Res.*, *16*(6), 961–979, 1980.
41. Jury, W. A., W. R. Gardner, and W. H. Gardner, *Soil Physics*, 5th ed., Wiley, 1991, pp. 45–47.
42. Hillel, D., *Introduction to Soil Physics*, Academic, New York, 1982, pp. 57–89.
43. Schmugge, T. J., Remote sensing of soil moisture, in *Hydrological Forecasting*, Wiley, New York, 1985.
44. Ulaby, F. T., G. A. Bradley, and M. C. Dobson, Potential application of satellite radar to monitor soil moisture, in *Satellite Hydrology*, American Water Resources Association, 1979, pp. 363–370.
45. Jury, W. A., W. R. Gardner, and W. H. Gardner, *Soil Physics*, 5th ed. Wiley, New York, 1991, pp. 268–293.

46. Dingman, S. L., *Physical Hydrology*, Prentice-Hall, Englewood Cliffs, NJ, 1994, pp. 247–250.

47. Bear, J., *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.

48. Springer, E. P., and T. W. Cundy, Field-scale evaluation of infiltration parameters from soil texture for hydrologic analysis, *Water Resour. Res.*, *23*, 325–334, 1987.

49. Berndtsson, R., and M. Larson, Spatial variability of infiltration in a semi-arid environment, *J. Hydrol.*, *90*, 117–133, 1987.

50. Sharma, M. L., G. A. Gander, and G. C. Hunt, Spatial variability of infiltration in a watershed, *J. Hydrol.*, *45*, 101–122, 1980.

51. Wood, E. F., M. Sivaplan, and K. Beven, Similarity and scale in catchment storm response, *Rev. Geophys.*, *28*, 1–18, 1990.

52. Sposito, G., Recent advances associated with soil water in the unsaturated zone, in *Reviews of Geophysics*, Supplement, U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, 1995, pp. 1059–1065.

CHAPTER 28

GROUNDWATER FLOW PROCESSES

WILLIAM W-G. YEH

1 INTRODUCTION

In the hydrologic cycle, groundwater occurs whenever surface water occupies and saturates the pores or interstices of the rocks and soils beneath Earth's surface. The geologic formations that store and transmit the subsurface water are known as aquifers. Aquifers, aquitards (semipermeable formations), or aquicludes (nonpermeable formations) may underlie a geographic area, watershed, or drainage basin, and all may hold water. But drawing water from aquitards and aquicludes is impractical and economically prohibitive, whereas groundwater stored in aquifers can be removed economically and is often a dependable source of water supply (Todd, 1980). Most aquifers can be considered as underground storage reservoirs that receive recharge from both natural and artificial sources.

Depending on local geological formation and boundary conditions, groundwater may flow out of the aquifer, contributing to surface runoff. In most cases, each aquifer formation has spatially varying properties, such as transmissivity and storativity, which affect the basin's response to pumping and artificial recharge. These formations are collectively referred to as a groundwater reservoir or groundwater system (Willis and Yeh, 1987). Groundwater aquifers can be classified as confined or unconfined, depending on the existence of a water table. A leaky confined aquifer represents a geological formation that leaks and allows water to flow through the confining layer.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

2 DARCY'S LAW

The fundamental law that governs groundwater flow in a laminar flow regime is Darcy's law. If we assume the porous medium is homogeneous and isotropic, Darcy's law states that the specific discharge is proportional to the gradient of hydraulic head:

$$\mathbf{q} = -K\nabla h \tag{1}$$

where $\mathbf{q}$ is the specific discharge vector (volume flow rate per unit cross-sectional area normal to the direction of flow), K is the hydraulic conductivity, h is the head, and ∇h is the gradient vector of the head,

$$\nabla h = \left(\frac{\partial h}{\partial x}i + \frac{\partial h}{\partial y}j + \frac{\partial h}{\partial z}k\right) \tag{2}$$

where i, j, and k are unit vectors in the x, y, and z coordinate directions, respectively. The hydraulic conductivity (K) is a function of both fluid and medium properties. As can be shown by dimensional analysis using the basic units of length (L), mass (M), and time (T), K can be expressed as (see, e.g., DeWiest, 1965):

$$K = \frac{Cd^2\gamma}{\mu} = k\frac{\gamma}{\mu} \tag{3}$$

where d (L) is some characteristic length of the medium, e.g., the average pore size or mean grain diameter of the granular material, μ M/(LT) is the dynamic viscosity, γ M/(L^{2}T^2) is the specific weight of the fluid (water), and C is a constant or shape factor, which accounts for the effects of stratification, packing, arrangement of grains, size distribution, and porosity. Parameter k is referred to as the intrinsic permeability and is solely dependent on the medium properties ($k = Cd^2$).

The porous medium is said to be homogeneous if the hydraulic conductivity is independent of the position (x, y, z) within the aquifer. If not, the aquifer is inhomogeneous, i.e., $K = K(x, y, z)$. The isotropy or anisotropy of the aquifer reflects the directional variability of the hydraulic conductivity. If the hydraulic conductivity varies with the direction of flow, the aquifer is anisotropic. On the other hand, if the hydraulic conductivity is independent of the direction of flow, the aquifer is isotropic. The conditions of inhomogeneity and anisotropy are common occurrences in the soils and geologic formations of aquifers.

Because the specific discharge may not be collinear with the gradient of the hydraulic head, nor have equal specific discharge components in the x, y, and z

directions, the hydraulic conductivity may be represented as a second-order tensor quantity (Eagleson, 1970). Darcy's law can then be generalized as:

$$\begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} = -\begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \begin{bmatrix} \dfrac{\partial h}{\partial x} \\ \dfrac{\partial h}{\partial y} \\ \dfrac{\partial h}{\partial z} \end{bmatrix} \tag{4}$$

If the coordinate system is aligned with the principal directions of the hydraulic conductivity tensor, Darcy's law in three dimensions can be written as:

$$q_x = -K_{xx}\frac{\partial h}{\partial x} \tag{5}$$

$$q_y = -K_{yy}\frac{\partial h}{\partial y} \tag{6}$$

$$q_z = -K_{zz}\frac{\partial h}{\partial z} \tag{7}$$

The governing equations for groundwater flow are generally derived by combining Darcy's law with the continuity equation (conservation of mass).

3 FLOW EQUATION FOR A CONFINED OR LEAKY AQUIFER

In a confined aquifer, the amount of water released from groundwater storage is dependent on the compressibility of the water and of the porous medium. Confined aquifers are bounded above and below by confining layers. In contrast, leaky or semiconfined aquifers have semipermeable confining layers that are capable of leakage and storage. A multilayered aquifer system is a system in which the aquifers are hydraulically interdependent as changes in head in one layer, caused by pumping, or recharge, can induce flow to and from adjacent layers. If we assume that the leakage or flow between layers and aquitards occurs only in the vertical direction and that any storage effects in the aquitards are negligible, the governing equation characterizing two-dimensional horizontal flow in a semiconfined or leaky aquifer can be expressed by

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right) = S\frac{\partial h}{\partial t} - L \pm Q \tag{8}$$

where T_{xx}, T_{yy} = components of transmissivity along the x and y coordinate axes, the product of the hydraulic conductivity and the aquifer's thickness (L^2/T)

S = storage coefficient (dimensionless)

h = head in the main aquifer (L)

t = time (L)

L = leakage from the overlying semiconfining stratum (L/T)

Q = sink/source term (L/T)

The leakage term can be calculated by Darcy's law:

$$L = K_z \frac{H - h}{b'} \tag{9}$$

where K_z = vertical hydraulic conductivity of the overlaying semiconfining stratum (L/T)

b' = thickness of the overlying semiconfining stratum (L)

H = external head in the overlying semiconfining stratum (L)

The effect of pumping and injection wells in the main aquifer can be simulated by representing the wells as point sources or point sinks under the assumption that the wells fully penetrate the thickness of the aquifer. If we let the index set Ω be the locations of all the pumping and injection wells, then the point sources/sinks can be expressed as:

$$\pm \sum_{w \in \Omega} Q_w \delta(x - x_w, y - y_w) \tag{10}$$

where $+Q_w$ is the pumping ($-Q_w$ recharge) from the wth pumping (injection) well located at (x_w, y_w) and $\delta(x - x_w, y - y_w)$ is the Dirac delta function, where

$$\delta(x - x_w, y - y_w) = \begin{cases} 1, & \text{if } x = x_w, y = y_w \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Equation (8) applies to a leaky aquifer system in which the main aquifer is overlain by a semiconfining stratum that leaks water into the main aquifer through the semiconfining stratum. If the term L goes to zero, Eq. (8) applies to a strictly confined system.

4 FLOW EQUATION FOR AN UNCONFINED AQUIFER

In contrast to confined aquifers, an unconfined aquifer has a free surface (water table) boundary, a boundary at atmospheric pressure. Water released from storage occurs due to gravity drainage as the water table in the aquifer responds to pumping, drainage, or natural or artificial recharge. The unconfined flow problem is commonly analyzed using the Dupuit assumptions: (1) uniform and horizontal flow within any vertical cross section, and (2) the velocity at the free surface may be expressed as $q_x = -K(\partial h/\partial x)$. The second assumption implies small slopes of the free surface.

Using the concept of vertical averaging, the governing equation characterizing two dimensional horizontal flow in an unconfined aquifer can be expressed as:

$$\frac{\partial}{\partial x}\left(K_{xx}h\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}h\frac{\partial h}{\partial y}\right)=[S_y+S_sh]\frac{\partial h}{\partial t}-R \tag{12}$$

where K_{xx}, K_{yy} = components of hydraulic conductivity along the x and y coordinate axes (L/T)
S_y = specific yield of the aquifer (dimensionless)
S_s = specific storage of the aquifer (1/L)
R = net recharge (L/T)

The specific storage effect is generally negligible when compared to the specific yield and can be dropped to give

$$\frac{\partial}{\partial x}\left(K_{xx}h\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}h\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{13}$$

which is the nonlinear Boussinesq equation. Pumping and injection wells may also be incorporated via Eq. (10) in the recharge term of the equation as point sources or sinks. There are several ways to linearize Eq. (13). The first method is based on the assumption that the depth of the flow varies slightly in the flow domain, e.g., mildly sloping aquifers. The head may then be expressed by

$$h=\bar{h}+\hat{h} \tag{14}$$

where $\bar{h}$ is the average depth of flow and $\hat{h}$ is the derivation of the head from $\bar{h}$. If we assume $\hat{h} \ll \bar{h}$, the Boussinesq equation becomes

$$\frac{\partial}{\partial x}\left(K_{xx}\bar{h}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\bar{h}\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{15}$$

or

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{16}$$

where $T_{xx}=K_{xx}\bar{h}$ and $T_{yy}=K_{yy}\bar{h}$. It can be seen that Eq. (16) is identical to the governing equation of the confined flow.

The second method of linearization is based on the temporal variation of the temporal derivative. Rewriting the temporal derivative as

$$S_y \frac{\partial h}{\partial t} = \frac{S_y}{2h} \frac{\partial h^2}{\partial t} \tag{17}$$

and assuming $\bar{S} = S_y/2h$ is approximately constant and equal to $S_y/2\bar{h}$, the Boussinesq equation is intrinsically linear in h^2,

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h^2}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h^2}{\partial y}\right) = 2\bar{S}\frac{\partial h^2}{\partial t} - 2R \tag{18}$$

If the initial and boundary conditions are also linear in h^2, Eq. (18) has been shown to be more accurate in predicting the water level.

The third method of linearization, as is used in MODFLOW (McDonald and Harbaugh, 1988), is to use the calculated head value from the last iteration h' to replace h, i.e., $T_{xx} = K_{xx}h'$ and $T_{yy} = K_{yy}h'$, and iteration continues until a convergence criterion is met.

5 INITIAL AND BOUNDARY CONDITIONS

The most common types of boundary conditions are Dirichlet, Neumann, and Cauchy.

1. Dirichlet conditions occur when a portion of the boundary is at a prescribed head level. For example, if an aquifer is adjacent to a stream or lake, then

$$h(x, y, t) = f_1(x, y, t) \qquad (x, y) \in \partial\Omega_1 \tag{19}$$

where f_1 is a known function and $\partial\Omega_1$ is the boundary.

2. Neumann conditions occur when a portion of the boundary has a specified flow transversing it normal to the boundary. For example, if a portion of the aquifer boundary is subject to recharge, then

$$T\frac{\partial h}{\partial n} = f_2(x, y, t) \qquad (x, y) \in \partial\Omega_2 \tag{20}$$

where f_2 is a known function and $\partial h/\partial n$ is the normal derivative to the boundary $\partial\Omega_2$.

3. The Cauchy, or mixed boundary, condition occurs when both the head and its normal derivative are specified on the boundary $\partial\Omega_3$, for example, the induced infiltration in a coupled aquifer and stream system (Prickett and Lonnquist, 1971).

Prior to solving the groundwater flow equation, the initial condition must be specified in the flow domain.

Typically, the initial condition can be expressed by

$$h(x, y, 0) = f_3(x, y) \qquad (x, y) \in \Omega \tag{21}$$

where Ω is the flow domain and f_3 is a known function.

6 DATA COLLECTION

Before a groundwater model can be used for prediction or management purposes, it must be calibrated using historical observations, historical operational records, and initial and boundary conditions. Calibration, in the general sense, concerns the estimation of model parameters (parameter identification) in a given conceptual model. Without doubt, the more data (observations) we have, the more reliable the calibrated model will be. Typical data available include the following:

1. *Topographical and geographical maps*. This information assists in defining the region of the groundwater basin and boundary types, i.e., impermeable, given head, given flux, and so on.
2. *Well logs*, which contain the vertical distribution of geological formations, including depth, color, character, size of material, and structure of the strata. This information helps to determine the layer structure and parameter value range in each layer.
3. *Groundwater level observations*. This information is most crucial to model calibration, as most of the groundwater models are calibrated against historical water level observations.
4. *Historical precipitation and streamflow data*. This information provides recharge and boundary conditions.
5. *Historical groundwater pumping/injection records*. This information provides the sink/source term.
6. *Land use data*, which contain acreages of human activities, such as residential, commercial, industrial, and agricultural uses. This information helps to estimate the groundwater consumption and return flow.

7 SELECTION OF NUMERICAL MODELS

Analytical solutions of groundwater flow have been reported for idealized groundwater systems; however, a more common approach to solving the distributed-parameter, time-dependent partial differential equations that govern the groundwater flow is through numerical techniques such as the finite-difference or finite-element methods. These techniques transform the partial differential equations into a system of algebraic equations. The solution of the system of algebraic equations determines the head values at a predetermined set of discrete nodal points within the aquifer system.

Finite-difference approximation is based on the Taylor series representation of the time and spatial derivatives. It is conceptually more straightforward than the finite-element approximation and easy to implement. Finite-element approximation is based on the method of weighted residuals. For many groundwater problems, the finite-element method may be more advantageous over the finite difference-method. Medium heterogeneity and irregular boundary conditions are handled easily by the finite element method. This contrasts with finite-difference approximation that requires complicated interpolation schemes to approximate complex boundary conditions. Moreover, in the finite-element method, the size of the elements can easily be varied to reflect rapidly changing state variables or parameter values. The piecewise continuous representation of the dependent variables and, possibly, the parameters of the groundwater system can also increase the accuracy of numerical approximation (Willis and Yeh, 1987).

Since these two types of numerical methods have been applied to many fields, references are abundant in the literature. Willis and Yeh (1987), Anderson and Woessner (1992), and Sun (1994b) have provided detailed analyses of how these two methods are to be applied to groundwater modeling. It is also worth noting that many established groundwater modeling softwares are available in the public domain. Bedient et al. (1994) provided a summary listing of existing numerical models of groundwater flow and solute transport.

8 PARAMETER ESTIMATION (PARAMETER IDENTIFICATION)

Thc accuracy of model prediction depends on the reliability of the estimated model parameters as well as on the accuracy of the prescribed initial and boundary conditions. In general, parameters used in deriving the governing equations are not directly measurable from the physical point of view. In practice, model parameters are required to be estimated from historical input–output observations using an inverse procedure of parameter estimation.

The inverse problem of parameter estimation in distributed-parameter systems has been studied extensively during the last three decades. The term distributed system implies that the response of the system is governed by a partial differential equation [(Eq. (8) or (13)] and parameters embedded in the equation (T_{xx}, T_{yy}, S) are spatially dependent. A review of the inverse problem of parameter identification in groundwater hydrology was presented by Yeh (1986), Carrera (1988), Sun (1994a), and McLaughlin and Townley (1996). In general, the inverse problem seeks to identify the model parameters by observing the output of the dependent variable (head) in the spatial and time domain. Frequently, point estimates of transmissivity and storage coefficient are also available and they can be used as prior information to regulate the inverse solution.

9 PARAMETERIZATION

The number of observations is finite and limited, whereas the spatial domain is continuous. For an inhomogeneous aquifer, the dimension of, for example, the transmissivity is theoretically infinite. In practice, the infinite parameter dimension must be reduced to a finite dimensional form. The reduction of the number of parameters from the infinite dimension to a finite dimensional form is called parameterization (Yeh and Yoon, 1976, 1981; Yeh, 1986; Sun, 1994a). Parameterization can be achieved by either a deterministic method or by a stochastic model. In general, parameterization can be achieved by the following methods.

Zonation Method

In this approach, the flow region of the aquifer is divided into a number of zones, and a constant parameter value is used to characterize the aquifer property in each zone. The unknown transmissivity function is then represented by a number of constants, which is equal to the number of zones. Here, we mention the work of Coats et al. (1970), Emsellem and de Marsily (1971), Yeh and Yoon (1976), and Cooley (1977, 1979).

In principle, the zonation pattern and its corresponding parameter values should be determined simultaneously (Sun and Yeh, 1985; Sun et al., 1998).

Interpolation Method

If, for example, finite elements are used as the interpolation method, the unknown parameter distribution in the flow region is discretized into a number of elements connected by a number of nodes. Each node is associated with a chosen local basis function. The unknown transmissivity distribution is then approximated by a linear combination of the basis functions, where the parameter dimension is equal to the number of unknown nodal transmissivity values (DiStefano and Rath, 1975; Yoon and Yeh, 1976; Yeh and Yoon, 1981):

$$T_e = \sum_j T_j \phi_j^e(x, y) \tag{22}$$

where the basis function ϕ_j^e is chosen in such a way that it equals 1 at the particular node j and 0 at all other nodes on the element (e). Other interpolation methods for approximating the transmissivity distribution include simple polynomial approximation, cubic spline, and kriging.

Stochastic Method

In this approach, the unknown parameter is treated as a random field, characterized by its first two moments, the mean (or drift) and the covariance function. A common approach is to assume that the logarithm of the hydraulic conductivity, $Y = \log K$, is normally distributed (Freeze, 1975; Hoeksema, 1985a). Also, the random field is represented by a constant mean and an isotropic, exponential covariance (Dagan, 1985; Hoeksema and Kitanidis, 1985b; Wagner and Gorelick, 1989):

$$E(Y) = \mu_Y \tag{23}$$

$$\mathrm{Cov}_{YY}(x_i, x_j) = \sigma_Y^2 \exp\left(-\frac{d_{ij}}{l_Y}\right) \tag{24}$$

where σ_Y^2 = log hydraulic conductivity variance
l_Y = log hydraulic conductivity correlation scale
d_{ij} = distance between points x_i and x_j

The hydraulic conductivity can thus be estimated by identifying the three statistical parameters μ_Y, σ_Y^2, and l_Y. In this approach, overparameterization is generally avoided, and the inverse solution obtained by the maximum-likelihood estimate and cokriging is highly stable.

In addition to the traditional approaches for parameterization mentioned above, Sun et al. (1995) suggested a geological parameterization method in which the unknown parameter (hydraulic conductivity) is directly related to the geological materials, and the geological structure of the aquifer is determined by the geostatistical method of kriging.

10 PARAMETER UNCERTAINTY, PARAMETER STRUCTURE, AND OPTIMUM PARAMETER DIMENSION

Parameter identification in a distributed-parameter system should, in principle, include the determination of both the parameter structure and its value. If zonation is used to parameterize the unknown parameters, the zonation pattern (parameter structure) is represented by the number and shape of zones. On the other hand, if the finite-element method is used for parameterization, parameter structure is represented by the number and location of nodal values of parameters.

Identifying parameter structure is much more difficult than identifying parameter values for a given parameter structure. In the past three decades, only a few studies have contributed to this topic. The question of how to determine an appropriate zonation pattern was first considered by Emsellem and de Marsily (1971), who suggested that the number of zones be gradually increased until model fit no longer improved. This approach ignores the reliability of the estimated parameters. Yeh and Yoon (1976) were the first to consider both the error in model fitting and the error associated with parameter uncertainty in determining zonation pattern; to determine if a particular zone should be subdivided into smaller zones, they used the variance of the estimation error. Sun and Yeh (1985) proposed a systematic approach that can automatically identify the optimal pattern of parameter structure and its corresponding parameter values by solving a combinatorial optimization problem. They clearly pointed out that the identified parameter values vary with the parameter structure. As a consequence, if the parameter structure is incorrect, the identified parameter values will also be incorrect. In Carrera and Neuman (1986), the dimension of parameterization is determined by Akaike's information criteria (Akaike, 1972); these criteria can also be used to compare different zonation patterns. Bellout (1992) considered the stability of pattern identification from a mathematical analysis. Recently, Zheng and Wang (1996) used the tabu search (TS) method to find the optimal zonation structure for one-dimensional problems. Eppstein and Doupherty (1996) presented an extended Kalman filter for simultaneously estimating transmissivity values and zonation pattern. A general formulation of the inverse problem that incorporates the identification of parameter structure and its parameter values is given in Sun et al. (1998). To estimate the parameter structure, some authors have attempted to incorporate directly into the solution of the inverse problems the geological structure information obtained from well-logs and seismic measurements (Rubin et al., 1992, Sun et al., 1995; Hyndman and Gorelick, 1996; Koltermann and Gorelick, 1996).

Shah et al. (1978) showed the relationship between the optimal dimension of parameterization and observations in considerable depth. The necessity to limit the dimension of parameterization has been further studied by Yeh and Yoon (1981), Yeh, et al. (1983), and Kitanidis and Vomvoris (1983). The dimension of parameterization is directly related to the quantity and quality of data (observations). In practice, the number of observations is limited and observations are corrupted with noise. Without controlling parameter dimension, instability in the inverse solution often results (Yakowitz and Duckstein, 1980). If instability occurs, parameter values will become unreasonably small (sometimes negative, which is physically impossible) and/or large, if parameter values are not properly constrained. In the constrained minimization, instability is characterized by the fact that during the inverse solution process parameter values are bouncing back and forth between the upper and lower bounds. Reduction of parameter dimension can make the inverse solution stable. As the number of zones (in the zonation case) is increased, the modeling error (least squares) decreases while the parameter uncertainty error at some point will start to increase (Shah et al., 1978; Yeh and Yoon, 1981). A trade-off of these two types of errors can then be made, from which an optimum parameter dimension can be determined. A standard procedure is to gradually increase the parameter dimension, starting from the lowest dimension, i.e., the homogeneous case, and calculate the two types of errors for each parameterization. The error in parameter uncertainty can be represented by a norm of the covariance matrix of the estimated parameters (Yeh and Yoon, 1976; Shah et al., 1978).

An approximation of the covariance matrix of the estimated parameters in nonlinear regression can be represented by the following form (Bard, 1974; Yeh and Yoon, 1976, 1981; Shah et al., 1978; Yeh, 1986):

$$\mathrm{Cov}(\hat{\mathbf{T}}) = \frac{J(\hat{\mathbf{T}})}{M-L}[\mathbf{A}(\mathbf{T})]^{-1} \qquad (25)$$

where $\mathbf{J}(\hat{\mathbf{T}})$ = least-squares error

M = number of observations

L = parameter dimension

$\mathbf{A} = [\mathbf{J}_D^T \mathbf{J}_D]$

$\mathbf{J}_D$ = Jacobian matrix of $\mathbf{h}$ with respect to $\mathbf{T}$

A norm of the covariance matrix has been used to represent the error in parameter uncertainty. Norms, such as trace, spectral radius (maximum eigenvalue), and determinant have been used in the literature. Equation (25) also assumes homoscedasticity and uncorrelated errors. This assumption is generally not satisfied and the actual covariance may be much higher than given by Eq. (25).

The covariance matrix of the estimated parameters also provides information regarding the reliability of each of the estimated parameters. A well-estimtated parameter is generally characterized by a small variance as compared to an insensitive parameter that is associated with a large variance. By definition, the correlation matrix of the estimated parameter is

$$\mathbf{R} = \begin{bmatrix} \dfrac{c_{11}}{(c_{11}c_{11})^{1/2}} & \cdots & \dfrac{c_{1L}}{(c_{11}c_{LL})^{1/2}} \\ \vdots & & \vdots \\ \dfrac{c_{L1}}{(c_{LL}c_{11})^{1/2}} & \cdots & \dfrac{c_{LL}}{(c_{LL}c_{LL})^{1/2}} \end{bmatrix} \tag{26}$$

where c_{ij}'s are elements of the covariance matrix of the estimated parameters. The more sensitive the parameter, the closer and quicker the parameter will converge. A correlation analysis of the estimated parameters would indicate the degree of interdependence among the parameters with respect to the objective function. Correlation of parameters is called the *collinearity* problem. Such problems can cause slow rate of convergence in minimization and in most cases result in nonoptimal parameter estimates. A more rigorous treatment of the collinearity problem is to use the more sophisticated statistical techniques, such as ridge regression (Cooley, 1977) and the method of principal components.

11 MODEL STRUCTURE ERROR (PARAMETER STRUCTURE ERROR)

Sun et al. (1998) presented a procedure whereby the model structure error of using one model structure to replace another model structure is defined by a max–min problem that is based on the distance between the two models measured in the parameter, observation, and prediction/management space. Parameter structure error resulting from a different level of parameterization is a special case of model structure error. Without losing generality, we will use parameter structure error to represent model structure error.

The parameter structure error, SE(G_A, G_B), of using parameter structure G_B to replace parameter structure G_A can be defined by the following max–min problem (Sun, 1994a, 1996):

$$\mathrm{SE}(G_A, G_B) = \max_{p_A} \min_{p_B} d(G_A, \mathbf{p}_A; G_B, \mathbf{p}_B) \tag{27}$$

where d is the distance (to be defined later) between the two models, $M_A(G_A, \mathbf{p}_A)$ and $M_B(G_B, \mathbf{p}_B)$; and parameters $\mathbf{p}_A$ and $\mathbf{p}_B$ must be in their admissible regions P_A and P_B. In general, SE(G_A, G_B) $\neq$ SE(G_B, G_A). When G_A is a simplification of G_B, we have SE(G_A, G_B) = 0.

The distance between the two models, $M_A(G_A, \mathbf{p}_A)$ and $M_B(G_B, \mathbf{p}_B)$, as generalized by Sun et al. (1998), can be defined as:

$$d(M_A, M_B) = d_E + \mu d_D + \lambda d_P \tag{28}$$

where

$$d_E(M_A, M_B) = \|g_E(M_A) - g_E(M_B)\|_E \tag{29}$$

$$d_D(M_A, M_B) = \|h_D(M_A) - h_D(M_B)\|_D \tag{30}$$

$$d_P(M_A, M_B) = \|\bar{\mathbf{p}}A - \bar{\mathbf{p}}_B\|_{\bar{G}} \tag{31}$$

where subscript E denotes a prediction/management alternative and its associated prediction space; $\|\cdot\|_E$ is a norm defined in the prediction space; subscript D denotes an observation design and its associated observation space; $h_D(M_A)$ and $h_D(M_B)$ are "observations" based on the same observation design but generated from difference models, M_A and M_B; $\|\cdot\|_D$ is a norm defined in the observation space; $\bar{G}$ is a parameter space having a common overparameterization structure of G_A and G_B; $\bar{\mathbf{p}}_A$ and $\bar{\mathbf{p}}_B$ are spans of $\mathbf{p}_A$ and $\mathbf{p}_B$; $\|\cdot\|_{\bar{G}}$ is a norm defined in $\bar{G}$; μ and λ are weighting coefficients. It is clear that by varying the weighting coefficients, one can emphasize the importance of each distance in the parameter, observation, or prediction/manmanagement space. As a result, this will influence the inverse solution.

12 GENERALIZED INVERSE PROCEDURE

The generalized inverse procedure seeks to minimize the weighted composite objective function as represented by Eq. (28). In this procedure, the unknown model structure (parameter structure) and its corresponding parameter values are determined not only from prior information and observations but also by the accuracy requirement in model applications. Sun et al. (1997, 1998) presented a stepwise regression procedure for a simultaneous estimation of parameter structure and its corresponding parameter values. The procedure starts from a homogeneous parameter structure and gradually increases the complexity of the parameter structure. For a given set of data and a specified model reliability requirement, the method, at each level of complexity, calculates both the least-squares error as well as the parameterization error of using a simpler parameter structure to replace a more complex parameter structure. The method is most general as it considers errors in the parameter, observation, as well as prediction/management space. The established procedure allows one to determine whether a more complex parameter structure is needed or to conclude that data are insufficient to meet the specified model reliability requirement; and hence, additional data are needed.

In this procedure, we form a series of parameter structures of increasing complexity:

$$G_1, G_2, G_3, \ldots, G_m, \ldots$$

where G_1 represents a homogeneous structure, G_2 a two-zone structure, and so forth; G_2 is generated from G_1 by dividing it into two zones and, generally, G_{m+1} is generated from G_m by dividing one of the zones of G_m into two zones. At each level, we calculate the residual error (RE) and the parameter structure error (SE). Specifically, the following steps are involved:

Step 1. Let G_1 be a homogeneous parameter structure, we solve the generalized inverse problem to find $\mathbf{p}_1^*$ and the corresponding residual error RE_1. In general, RE can be found by minimizing a linear combination of the norms in the parameter and observation space. Details can be found in Sun et al. (1998).

Step 2. Divide G_1 into two zones to generate G_2. The method suggested by Sun and Yeh (1995) can be used to optimize simultaneously the zonation pattern and its corresponding parameter values. In this step, we find a model $M_2(G_2, \mathbf{p}_2^*)$ and its residual error RE_2. RE_2 must be smaller than RE_1 because a homogeneous parameter structure is being replaced by a two-zone structure to fit the same set of observations.

Step 3. Calculate the parameter structure error SE_1 by using G_1 to replace G_2. Details with regard to how to calculate SE_1 are presented in Sun et al. (1998).

Step 4. If both SE_1 and RE_2 are large, we continue to increase the parameter structure complexity by finding the optimum three-zone parameter structure M_3 $(G_3, \mathbf{p}_3^*)$ and RE_3.

Step 5. Calculate the parameter structure error SE_2 of using G_2 to replace G_3. If both SE_2 and RE_2 are large, we repeat steps 4 and 5 to obtain $M_4(G_4, \mathbf{p}_4^*)$, RE_4 and SE_3, and so forth. Assume that through this procedure we have found $M_{m+1}(G_{m+1}, \mathbf{p}_{m+1}^*)$, RE_{m+1}, and SE_m.

Step 6. Then consider the following four cases:

1. If both RE_{m+1} and SE_m are large compared to the observation error and accuracy requirement of the prediction/management problem, respectively, increase m by one and repeat step 5.
2. If both RE_{m+1} and SE_m are small, stop and use M_{m+1} as the identified model.
3. If RE_{m+1} is large but SE_m is small, either stop or continue to increase the complexity until RE_{m+1} becomes small;
4. If RE_{m+1} is small but SE_m is large, additional data are required.

In case 1, the identified model cannot satisfy the accuracy requirement of the given model application but, at the same time, the existing data still have the potential to provide more information. Therefore, we increase the parameter structure complexity. In cases 2 and 3, when the complexity of the parameter structure is increased, the prediction/management solution is not significantly improved; thus, we can either accept the identified model or if existing data still contain more information, we can continue to increase the parameter structure complexity. In case 4, the information contained in the existing data is insufficient to identify a reliable model and, thus, additional data are required to be collected.

13 CONCLUSIONS

The development of groundwater simulation models in the early 1970s provided groundwater planners with quantitative techniques for analyzing alternative groundwater pumping or recharge strategies. The accuracy of a simulation model is dependent, to a certain extent, on the accuracy of the inverse solution, which in turn is determined by the quantity and quality of data. The inverse problem is inherently nonunique and unstable. It has been well understood that the number of unknown parameters must be reduced to obtain a unique and stable solution of the inverse problem. The reduction of the number of unknown parameters is achieved by means of parameterization. It also has become apparent that parameterization and its corresponding parameter values are interdependent and must be estimated simultaneously.

A recent advancement made in groundwater modeling has been the development of a generalized inverse procedure. This procedure allows us to analyze the errors in the parameter, observation, and prediction/management space. The requirement of finding the true parameter values in the classical inverse problem is replaced by a weaker requirement. More importantly, it helps us resolve the following two problems: (1) How complex should a groundwater model structure be for a given model application? (2) Are existing data sufficient for developing a reliable model for the stipulated prediction/management objective? The generalized inverse procedure attempts to find the simplest model structure for a given model application. Such a model requires the minimum amount of data to calibrate.

REFERENCES

Akaike, H., Information theory and an extension of the maximum likelihood principle, in B.N. Petrov and F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory, Supplement to Problems of Control and Information Theory*, Akad. Kiado, Budapest, 1972, pp. 267–281.

Anderson, M. P., and W. W. Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*, Academic, San Diego, CA, 1992.

Bard, Y., *Nonlinear Parameter Estimation*, Wiley, New York, 1974.

Bear, J., *Hydraulics of Groundwater*, McGraw-Hill, New York, 1979.

Bedient, P. B., H. S. Rifai, and C. J. Newell, *Ground Water Contamination*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

Bellout, H., Stability result for the inverse transmissivity problem, *J. Math. Anal. Applicat.*, *168*, 13–27, 1992.

Bredehoeft, J. D., and G. F. Pinder, Digital analysis of areal flow in mutiaquifer groundwater systems: A quasi-three dimensional model, *Water Resour. Res.*, *6*(3), 885–888, 1980.

Brutsaert, W., and H. A. Ibrahim, On the first and second linearization of the Boussinesq equation, *J. Am. Soc. Geophys.*, *11*, 549–554, 1966.

Carrera, J., State of the art of the inverse problem applied to flow and solute transport equations, in E. Custodio, A. Gurgui, and J. P. Lobo Ferreira (Eds.), *Groundwater Flow and Quality Modeling*, D. Reidel, Hingham, MA, 1988, pp. 549–583.

Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 3, Application to synthetic and field data, *Water Resour. Res.*, *22*(2), 228–242, 1986.

Chapman, M. J., and K. R. Godfrey, On structual equivalence and identifiability constraint ordering, in E. Walter (Ed.), *Identifiability of Parametric Models*, Pergamon, New York, 1987, pp. 32–39.

Chavent, G., M. Dupuy, and P. Lemonnier, History matching by use of optimal theory, *Soc. Pet. Eng. J.*, *15*(1), 74–86, 1975.

Coats, K. H., J. R. Dempsey, and J. H. Henderson, A new technique for determining reservoir description from filed performance data, *Soc. Pet. Eng. J.*, *10*(1), 66–74, 1970.

Coleman, T. F., A note on New Algorithms for constrained minmax optimization, *Math. Programming*, *15*, 239–242, 1978.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 1. Theory and numerical properties, *Water Resour. Res.*, *13*(2), 318–324, 1977.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 2. Application of statistical analysis, *Water Resour. Res.*, *15*(3), 603–617, 1979.

Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1. Theory, *Water Resour. Res.*, *18*(4), 965–976, 1982.

Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resour. Res.*, *21*(1), 65–72, 1985.

DeWiest, R. J. M., *Geohydrology*, Wiley, New York, 1965.

DiStefano, N., and A. Rath, An identification approach to subsurface hydrological systems, *Water Resour. Res.*, *11*(6), 1005–1012, 1975.

Eagleson, P. S., *Dynamic Hydrology*, McGraw-Hill, New York, 1970.

Emsellem, Y., and G. de Marsily, An automatic solution for the inverse problem, *Water Resour. Res.*, *7*(5), 1264–1283, 1971.

Eppstein, M. J., and D. E. Dougherty, Simultaneous estimation of transmissivity values and zonation, *Water Resour. Res.*, *32*(11), 3321–3336, 1996.

Ezzedine, S., and Y. Rubin, A geostatistical approach to conditional estimation of spatially distributed solute concentration and notes on the use of tracer data in the inverse problem, *Water Resour. Res.*, *32*(4), 853–862, 1987.

Freeze, R. A., A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media, *Water Resour. Res.*, *11*(5), 725–741, 1975.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Haber, R., and H. Unbehauen, Structure identification of nonlinear dynamic system—A survey on Input/Output approaches, *Automatica*, *26*(4), 651–677, 1990.

Hantush, M. S., Nonsteady flow to flowing wells in leaky aquifer, *J. Geophys. Res.*, *64*, 1943–1052, 1959.

Hill, M. C., *A Computer Program (MODFLOWP) for Estimating Parameters of a Transient, Three-Dimensional, Ground-Water Flow Model Using Nonlinear Regression*, Open-File Report 91-484, U.S. Geological Survey, Denver, CO, 1992.

Hoeksema, R. J., and P. K. Kitanidis, Analysis of the spatial structure of properties of selected aquifers, *Water Resour. Res.*, *21*(4), 563–572, 1985a.

Hoeksema, R. J., and P. K. Kitanidis, Comparison of Gaussian conditional mean and Kriging estimation in the geostatistical solution of inverse problem, *Water Resour. Res.*, *21*(6), 825–836, 1985b.

Hyndman, D. W., and S. M. Gorelick, Estimating lithologic and transport properties in three dimensions using seismic and tracer data: The Kesterson aquifer, *Water Resour. Res.*, *32*(9), 2659–2670, 1996.

Kitanidis, P., Quasi-linear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2420, 1995.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res.*, *19*(3), 677–690, 1983.

Koltermann, C. E., and S. M. Gorelick, Heterogeneity in sedimentary: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resour. Res.*, *32*(9), 2617–2658, 1996.

Loaiciga, H. A., and M. A. Marino, The inverse problem for confined flow: Identification and estimation with extension, *Water Resour. Res.*, *23*(1), 92–104, 1987.

Loaiciga, H. A., R. B. Laipnik, P. K. Herdak, and M. A. Marino, Effective hydraulic conductivity of nonstationary aquifers, *Stochast. Hydrol. Hydraul.*, *8*(1), 1–17, 1994.

McDonald, M. G., and A. W. Harbaugh, *A Modular Three-Dimensional Finite Difference Ground-Water Flow Model*, Open-File Report 83-875, U.S. Geological Survey, Denver, CO, 1988.

McLaughlin, D., and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resour. Res.*, *32*(5), 1131–1162, 1996.

Neuman, S. P., Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res.*, *9*(4), 1006–1021, 1973.

Prickett, T. A., and C. O. Lonnquist, *Selected Digital Computer Techniques for Groundwater Resource Evaluation*, Bulletin No. 55, Illinois State Water Survey, Urbana, IL, 1971.

Poeter, E. P., and M. C. Hill, Inverse models: a necessary next step in ground-water modeling, *Ground Water*, *35*(2), 250–260, 1997.

RamaRoa, B. S., M. A. LaVenue, G. de Marsily, and M. G. Marietta, Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 1, Theory and computational experiments, *Water Resour. Res.*, *31*(3), 475–493, 1995.

Rubin, Y., and G. Dagan, Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 1, Theory, *Water Resour. Res.*, *23*(7), 1809–1916, 1992.

Rubin, Y., G. Mavko, and J. Harris, Mapping permeability in heterogeneous aquifers using hydrological and seismic data, *Water Resour. Res.*, *28*(7), 1809–1816, 1992.

Rustem, B., A constrained min-max algorithm for rival models of the same economic system, *Math. Programming*, *53*, 279–295, 1992.

Shah, P. C., G. R. Gavalas, and J. H. Seinfeld, Error analysis in history matching: The optimum level of parameterization, *Soc. Pet. Eng. J.*, *18*(3), 219–228, 1978.

Sun, N-Z., *Inverse Problems in Groundwater Modeling*, Kluwer Academic, Norwell, Mass., 1994a.

Sun, N-Z., *Mathematical Modeling of Groundwater Pollution*, Springer-Verlag, New York, 1994b.

Sun, N-Z., Identification and reduction of model structure for modeling distributed parameter systems, in J. Gottlieb and P. DuChateau (Eds.), *Parameter Identification and Inverse Problems in Hydrology, Geology and Ecology*, Kluwer Academic, Norwell, Mass., 1996.

Sun, N-Z., and W. W-G. Yeh, Identification of parameter structure in groundwater inverse problem, *Water Resour. Res.*, *21*(6), 869–883, 1985.

Sun, N-Z., and W. W-G. Yeh, Coupled inverse problem in groundwater modeling, 1, Sensitivity analysis and parameter identification, *Water Resour. Res.*, *26*(10), 2507–2525, 1990.

Sun, N-Z., and W. W-G. Yeh, A stochastic inverse solution for transient groundwater flow: Parameter identification and reliability analysis, *Water Resour. Res.*, *28*(12), 3269–3280, 1992.

Sun, N-Z., M-C. Jeng, and W. W-G. Yeh, A proposed geological parameterization method for parameter identification in three-dimensional groundwater modeling, *Water Resour. Res.*, *31*(1), 89–102, 1995.

Sun, N-Z., M-C. Jeng, and W. W-G. Yeh, Model structure identification: The generalized inverse problem, in K. W. Watson and Z. Zaporozec (Eds.), *Advances in Ground-Water Hydrology*, American Institute of Hydrology, Tampa, FL, 1997, pp. 130–134).

Sun, N-Z., S. Yang, and W. W-G. Yeh, A proposed stepwise regression method for model structure identification, *Water Resour. Res.*, *34*(10), 2561–2572, 1998.

Todd, D. K., *Groundwater Hydrology*, Wiley, New York, 1980.

Wagner, B. J., and S. M. Gerelick, Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: From data to design, *Water Resour. Res.*, *25*(10), 2211–2225, 1989.

Willis, R., and W. W-G. Yeh, *Groundwater Systems Planning and Management*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

Xiang, Y., J. F. Sykes, and N. R. Thomson, A composite L_1 parameter estimator for model fitting in groundwater flow and solute transport simulation, *Water Resour. Res.*, *29*(6), 1661–1674, 1993.

Yakowitz, S., and L. Duckstein, Instability in aquifer identification: Theory and case studies, *Water Resour. Res.*, *16*(6), 1045-1064, 1980.

Yeh, W. W-G., Review of parameter identification procedure in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(2), 9–108, 1986.

Yeh, W. W-G., Systems analysis in ground-water planning and management, *J. Water Resour. Planning and Mgmt.*, *118*(3), 224–237, 1992.

Yeh, W. W-G., and N-Z. Sun, An extended identifiability in aquifer parameter identification and optimal pumping test design, *Water Resour. Res.*, *20*(12), 1837–1847, 1984.

Yeh, W. W-G., and N-Z. Sun, Variational sensibility analysis, data requirements, and parameter identification in a leaky aquifer system, *Water Resour. Res.*, *26*(9), 1827–1938, 1990.

Yeh, W. W-G., and Y. S. Yoon, A systematic optimization procedure for the identification of inhomogeneous aquifer parameters, in Z. A. Saleen (Ed.), *Advances in Groundwater Hydrology*, American Water Resources Association, Minneapolis, Minnesota, 1976, pp. 72–82.

Yeh, W. W-G., and Y. S. Yoon, Aquifer parameter identification with optimum dimension in parameterization, *Water Resour. Res.*, *17*(3), 664–672, 1981.

Yeh, J. T-C., and J. Zhang, A geostatistical inverse method for variably saturated flow in the vadose zone, *Water Resour. Res.*, *32*(9), 2757–2766, 1996.

Yeh, W. W-G., Y. S. Yoon, and K. S. Lee, Aquifer parameter identification with Kriging and optimum parameterization, *Water Resour. Res.*, *19*(1), 225–233, 1983.

Yoon, Y. S., Yeh, W. W-G. Parameter identification in an inhomogeneous medium with the finite-element method, *Soc. Petrol. Eng. J.*, *16*(4), 217–226, 1976.

Zheng, C., and P. Wang, Parameter structure identification using tabu search and simulated annealing, *Adv. Water Resour.*, *19*(4), 215–224, 1996.

CHAPTER 29

SURFACE RUNOFF GENERATION

KEITH BEVEN

1 INTRODUCTION: DEFINING RUNOFF

There are a number of different definitions of runoff that have been used either explicitly or implicitly in hydrological analyses over the years. In what follows we will use a working definition that *runoff* is that part of the rainfall falling on a catchment area that eventually leaves the catchment as a surface streamflow, whatever the flow pathway that the water has followed on its way to the stream channel. Thus this definition includes both surface and subsurface runoff pathways. Dunne (1978) provides a review of field studies of surface and subsurface runoff generation processes that remains one of the best summaries available.

For a long time, following the work of Robert Horton in the 1930s, storm runoff was often taken to be equivalent to a purely surface runoff process. Horton suggested that the soil surface acted as a separating surface, between fast (surface) storm runoff processes and slow (subsurface) flow processes contributing to *baseflow* (Horton, 1933). This concept still underlies a great deal of hydrological analysis even though we now know that this is often not the case and that much of the runoff in a stream channel seen during a storm event may have followed subsurface flow pathways. C. R. Hursh, working at the same time as Horton, was demonstrating the importance of subsurface flow in storm hydrograph generation in the Coweeta catchments in North Carolina (e.g., Hursh, 1944).

The best basis for the analysis of runoff in a catchment is to allow that there may be a spectrum of surface and subsurface flow velocities and path lengths, which must be expected to change with the state of wetness of the catchment area and with the nature and spatial pattern of a rainfall event. In some conditions or locations, subsurface flow processes may dominate runoff generation; in other conditions or locations (even within the same catchment), surface flow process may dominate. Indeed, in

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

some conditions and locations, subsurface flow may saturate the soil and return to the surface as a *return flow*, the area of saturation also acting as a *dynamic contributing area* for surface runoff due to additional rainfall inputs, that will expand and contract as the catchment wets and dries (Fig. 1). Thus, it is only for convenience that, in what follows, we treat subsurface and surface runoff processes in turn.

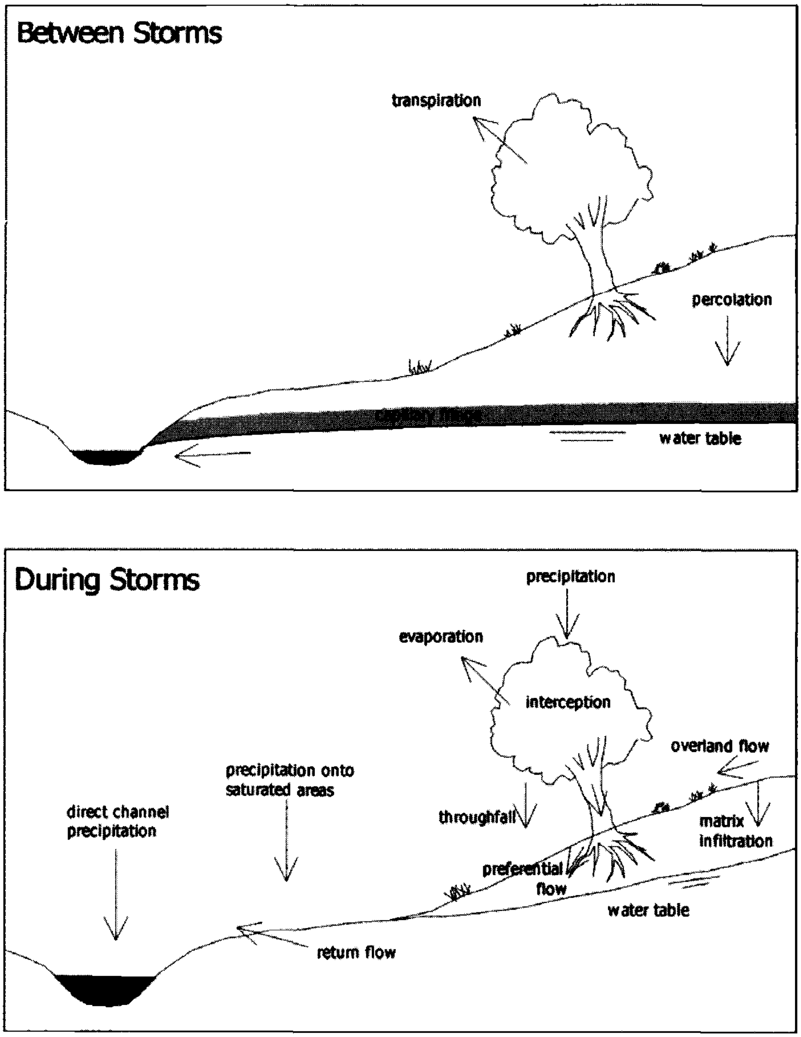


Figure 1 Processses of hillslope response to rainfall (after Beven, 2001).

2 GENERATION OF SUBSURFACE RUNOFF

Role of Soil in Runoff Processes

When rainfall falls on the land surface, in many environments the first thing it hits will be a vegetation canopy. This has the effect of retaining some of the rain on leaf surfaces as *interception* and redistributing the rest down to the ground surface as *throughfall* and *stemflow*. At the ground surface, therefore, the rate of supply of water will no longer be spatially uniform. There may be local concentrations at the base of stems or trunks; there may be other areas that receive lower intensity inputs.

This distribution of intensities falls onto a surface that will have a variable infiltration capacity due to variations in soil properties and initial moisture status. In particular, the properties at the soil surface will be very important in controlling how much of the rainfall can infiltrate into the soil. However, in many environments much of the rainfall input will infiltrate into the soil unless the soil is already completely saturated. It will do so by one of two routes, by direct infiltration into the soil matrix and by infiltration into pathways due to larger structural voids (cracks, root channels, animal or insect burrows, etc). The latter may be very important since the faster flow velocities associated with structural porosity can mean that the water can move rapidly into the soil, bypassing parts of the soil matrix as a *preferential flow* (see, e.g., Beven and Germann, 1982). Some of this water will be absorbed into the soil matrix at depth, some may move rapidly downslope within the preferential flow pathways, and some may be moved rapidly vertically to the local water table. Rates of movement will depend on the input rainfall intensity, the permeability and initial moisture content of the soil matrix, the structural characteristics of the soil, and the depth to the water table. It is well known from soil thin sections, however, that water movement in structural pathways can lead to transport of clay particles that are deposited in thin layers called *cutans*, which can then restrict the infiltration of water into the soil matrix and prolong the preferential flow. It is also known that preferential flow can be important in the transport of contaminants, such as pesticides and herbicides, in runoff since such contaminants are often sorbed to fine particles.

Within the soil matrix, water movement into soil that is not fully saturated will take place primarily vertically as a *wetting front*. The propagation of the wetting front will again depend on the antecedent wetness of the soil, the input intensity, and the matrix hydraulic characteristics. Flow within a continuous soil matrix can usually be described by the *Richards equation*, which is based on the unsaturated form of *Darcy's law*. The Richards equation is difficult to solve for general situations of practical hydrological interest, but there are now very many approximate numerical solution codes available (see Chapter 28). All such solution algorithms will require the specification of the soil hydraulic characteristics, which will depend on the texture and organic matter content of the matrix.

Based on many thousands of measurements, empirical relationships have been developed between the easily measured texture characteristics and the much more difficult to measure soil hydraulic properties (see, e.g., Rawls and Brakensiek, 1989). These empirical regressions are often useful but must be used with care. The estimates obtained in this way are subject to significant estimation error and are also only as good as the original data on which they are based. In this case the measurements were usually based on small soil samples and did not include any effects of the structural porosity. In any case, preferential flows may not be well described by the Richards equation, and it has proven very difficult to develop a comprehensive description of preferential flow with parameters that can easily be estimated in applications.

A useful simple analogy for the movement of water in both matrix and preferential flow pathways is that of the wetting front as a kinematic shock or locally pistonlike front, in which the rate of propagation of the front is given as

$$c = I/\Delta\theta$$

where c is the velocity of the front, I is the local input intensity, and $\Delta\theta$ is an effective change in moisture content across the front. This is an approximation, applicable only where the input rate I does not exceed the infiltration capacity of the soil at any depth, but it is then readily seen that the wave speed c increases with the input intensity and decreases with the change in moisture content. Thus, if $\Delta\theta$ is small, either in a preferential flow pathway (ignoring losses due to sorption into the matrix) or because the soil matrix is already wet, the front may move quickly into the soil. For a low input intensity I, infiltrating into a dry soil with an effectively high $\Delta\theta$, the speed of the wetting front may be very much lower. With a variation of effective intensities at the ground surface, and a variety of effective local $\Delta\theta$ values in different flow pathways, there will be a distribution of wetting front velocities in the soil.

Some of the infiltrating water will be retained in the soil and later evaporated or transpired back to the atmosphere (see Chapter 26), but some of the wetting resulting from a storm rainfall may reach an existing water table, or will induce saturation at the base of the soil profile, or a *perched* zone of saturation above a horizon of lower permeability in the soil profile. The wetting or *recharge* will induce a response in the saturated zone that will ultimately produce some subsurface runoff.

Recharge and Downslope Flow in a Saturated Zone

In fully saturated soil the propagation of the effects of changes in the boundary conditions, such as those due to recharge, is much more rapid than in the unsaturated zone. In shallow subsurface systems, such as where a shallow soil overlays an impermeable rock bed, most of the downslope flow toward stream channels will take place in the saturated zone. Because of the more rapid dissipation of local pressure differences in the saturated zone, a description of flow processes based on Darcy's law is generally more acceptable, even if preferential flow pathways are still contributing to the flow, since those pathways will be subject to similar pressure gradient conditions to the saturated matrix (with the reservation that in large pipe systems the flow may be turbulent and transitional rather than laminar and Darcy's law will not be valid).

Again, in relatively shallow soil systems a kinematic description is a useful analogy for the saturated zone, if we can assume that the local hydraulic gradient is approximately equal to the local slope angle (or even better the local bedslope angle). In this case, the equation of flow is the kinematic wave equation

$$W_x\phi_e\frac{\partial h}{\partial t} = T_h \sin\alpha\frac{\partial W_x h}{\partial x} + r_{x,t}$$

where h is the depth of saturation above the bed, x is distance measured along the slope, W_x, is the width of the slope at point x, α is the local slope angle, r is the recharge rate at point x and time t, and T_h is the integral of the saturated soil hydraulic conductivity function K_h over the depth of saturation h and is called the *transmissivity*; T_h and K_h will be a function of h that may be nonlinear for many soil profiles. The local downslope Darcian velocity of flow (volume flux per unit cross-sectional area) is then given by

$$V = K_h \sin\alpha$$

The mean pore water velocity is then given by

$$V_p = K_h \sin\alpha/\phi$$

where ϕ is the porosity of the soil. This is the mean velocity of the water itself. The kinematic wave velocity is given by

$$V_c = Kh \ \sin\alpha/\phi_c$$

where ϕ_c is an effective storage coefficient for the soil, or effectively in this case, the difference between the soil water content just above the water table and saturation. This last velocity is the rate at which disturbances to the flow are propagated in the direction of flow. The effective storage coefficient will generally be much less than the porosity of the soil, particularly in soils that are near saturation above the water table. Thus the wave speed may be very much faster than the mean pore water velocity, which will be faster than the Darcian velocity (since the porosity must itself be less than 1). The implications of this are that the effects of a recharge to the water table will move downslope much faster than the speed at which the water itself is moving. This will, in general, cause a rise in the subsurface outflow into the downslope stream channel more quickly than the inputs can flow toward that channel. This is one reason why in humid environments subsurface flow processes can make more rapid contributions to storm runoff than has been generally accepted in the past.

Again, Eq. (2), the kinematic wave equation, is an approximation. A fuller description will allow for fully three-dimensional flows in both soil and bedrock, perhaps including flow through bedrock fractures, with time-variable hydraulic gradients (see, e.g., Rasmussen et al., 2000). The same behavior of pressure wave transmission being faster than pore water velocities being in turn faster than Darcian velocities will hold. This is important in understanding the results of tracer experiments.

Equation (2) has been written in a way that allows for the width of the hillslope to vary downslope. It has been known for some time that an important control on the production of subsurface runoff, and of the occurrence of saturated dynamic contributing areas, is the form of hillslopes, in terms of both convergence or divergence in plan, and convexity or concavity in section. Soil water contents will tend to be higher and the saturated zone nearer to the surface in areas that are both convergent and concave. Such areas tend to be found particularly in *zero-order* headwater basins close to the heads of channels or the appearance of springs. These are areas where the soil is most likely to be close to saturation and consequently will show the greatest potential for acting as runoff source areas or *dynamic contributing areas* by either surface or subsurface flow processes.

Hillslope topography is, however, not the only cause of variability in flow rates. There is an increasing appreciation of the role of the geological structure of a catchment in controlling the subsurface flow pathways, even in catchments where there is no deep aquifer. The tracing experiments of Genereux et al. (1993), for example, revealed strong variability in the channel inputs in the Walker Branch catchment, Tennessee, that appeared to result from the bedding structure of the underlying rocks. Fracture systems in the near surface geology can also lead to the concentration of flow in certain locations. The occurrence of local perennial or seasonal springs is an indication that such effects may be important. Deeper fracture systems and flows along fault lines may also have an effect on subsurface flow pathways but are very difficult to study. Usually it is necessary to infer the presence of such flow pathways from the geochemical characteristics of baseflows.

For areas of relatively homogeneous shallow soils, it has been demonstrated that one way of predicting the location of such source areas is by use of the pattern of the topographic index a/s (e.g., Kirkby, 1978; see also O'Loughlin, 1981), which is the ratio of the area draining from upslope through unit contour length at any point in the catchment, to the slope angle at that point. The upslope area a represents the propensity for water to collect at a point; while the slope s represents the ease with which that point will drain. Approximate steady-state theory suggests that the index can be used as an index of hydrological similarity in that, other things being equal, points with similar values of the index should respond in a hydrologically similar way (Fig. 2). The topographic index has been incorporated into the rainfall–runoff model TOPMODEL and land surface parameterization TOPLATS, which aim to predict the dynamics of the surface and subsurface contributing areas and spatial patterns of latent heat flux in a simple way (Beven et al., 1995; Famiglietti et al., 1992). For recent critiques of the success of using the topographic index to represent contributing area dynamics see Beven (1997).

Use of the topographic index assumes that there is a consistent downslope flow of water on the hillslopes, but this is not always the case, particularly in catchment areas that are subject to extended drying periods. In such catchments, the effective subsur-

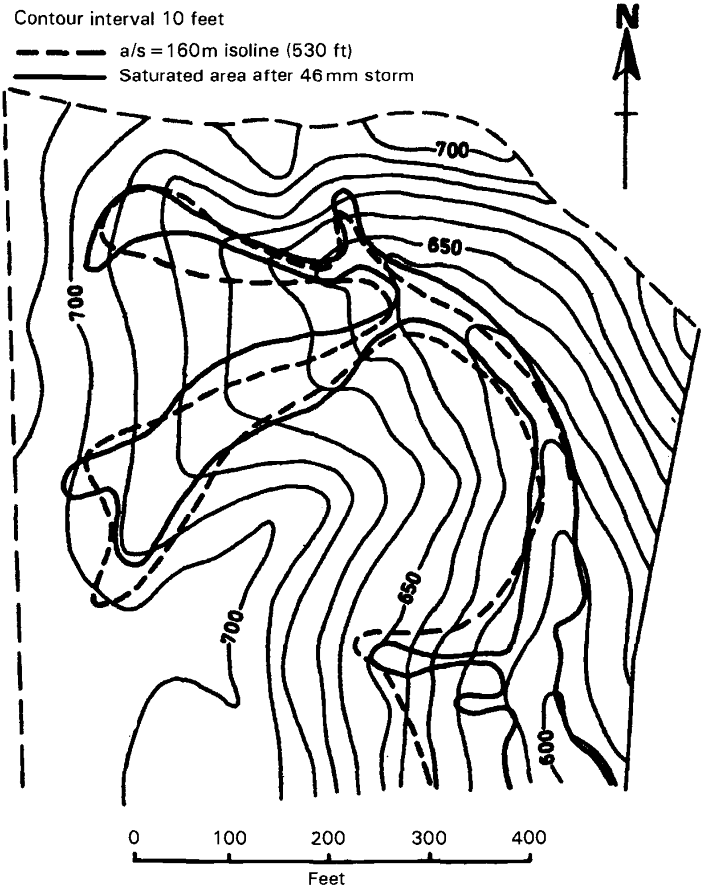


Figure 2 Pattern of the topographic index *a/s* in comparison with measured areas of surface saturation in basin WC-4 Sleepers River, Vermont (after Kirkby, 1978).

face contributing areas to the channel may not normally extend to the catchment divides and the development and connectivity of local saturation zones may be important. As noted earlier, in some soils, *perched* water tables may develop over a permeability break in the profile, resulting in increased downslope flow velocities without the profile being saturated to its base. This will tend to occur first in areas at the base of slopes and in hillslope hollows where the soil is normally wetter prior to an event (e.g., Weyman, 1970).

A final subsurface flow process that can lead to rapid subsurface responses is flow in natural soil pipes or along *percolines* of higher permeability (see, e.g., Beven and

Germann, 1982; McDonnell, 1990). Artificial drainage can have a similar effect, at least under wet conditions (under dry conditions artificial drainage can enhance the storage deficit of the soil prior to a storm and thereby lead to a lower runoff coefficient for an event.

Old and New Water Contributions to Storm Runoff

There have been many studies in the last 30 years that have made use of the natural geochemical characteristics of runoff to give at least an approximate separation of the storm hydrograph into a contribution from the rainstorm itself (the "new" water) and water stored in the catchment prior to the event ("old" water) (see, e.g., Sklash, 1990). Most such studies have used a two-component separation (into old and new water), which requires the assumptions that the geochemical characteristics of the two components are distinctly different and are constant in both space and time. The old water concentration is usually taken as that of the stream water baseflow component, measured prior to the start of a storm. The results, under these assumptions, have often suggested that a large proportion of the storm hydrograph may be made up of old water.

The assumptions, however, have often been questioned, particularly that of the constancy of the old water component. A number of studies have therefore introduced a third component, or *end member*, with the chemical characteristics of the "soil water," as determined from direct sampling. A separation or *end member mixing analysis* (*EMMA*) into three components requires measurements on at least two different tracers and may also be subject to some uncertainty (see, Bazemore et al., 1994; Fig. 3). The results, however, still usually suggest that a large proportion of the hydrograph is made up of water stored in the catchment prior to the event. This proportion would be reduced if it were shown that the new water rapidly takes on the tracer concentration characteristics of soil water, especially for reactive natural tracers such as silica, since it is known that equilibration times for dissolution or desorption can be short relative to storm duration in some circumstances.

Although these results have led to a reevaluation of the Hortonian concept that storm runoff is predominantly provided by rain water running off over the surface of the soil, there is no mystery about the hydrograph showing a large proportion of old water. This requires a displacement of water stored in the catchment prior to an event by the incoming rain water during the event. The simplified kinematic analysis of the different subsurface velocities presented above suggests that the wave speed will be greater than the mean pore water velocity and therefore displacement of old water into the stream would be expected. In addition, at least for moderate storms in humid catchments, the volume of water stored in the profile prior to an event may be much greater than the volume of event water (a 1-m soil profile, with an average of 25% moisture content in the profile, contains the equivalent of 250 mm of water per unit area). Thus, there will often be more than enough water available to be displaced. However, the proportion of old water might be expected to decrease as storm magnitude increases, but very few measurements have been reported for large-magnitude events.

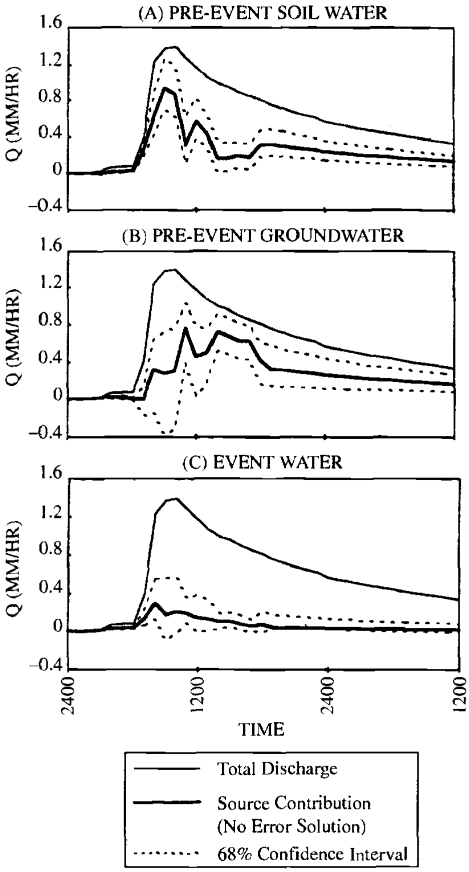


Figure 3 Separation of a stream hydrograph into rainfall, soil water, and groundwater contributions with estimates of the uncertainty associated with each component (after Bazemore et al., 1994).

Special Case of Runoff Generation by Snowmelt

Many hydrological regimes, particularly mountain regimes, are dominated by the spring snowmelt component. Snowmelt has particular characteristics in generating the snowmelt hydrograph. It tends to have only low intensities since even after the snow pack is "ripe" and ready for melt, rates of melt are limited on a daily basis by the energy available to supply the latent heat necessary to convert ice and snow to liquid water. Initially, routing through the snowpack may also diffuse the daily melt signal, and there may be some refreezing of water at night. Another interesting feature of the melt process is that it will have a characteristic spatial pattern since, in general, south facing slopes will melt before north facing slopes (in the Northern Hemisphere) and a low elevation snowpack before a high elevation pack. There may also be spatial variations in melt associated with differences in the storage of snow as a result of drifting during the winter period.

The response of a catchment during the snowmelt period may depend very much on the state of the soil. If the soil is frozen, then it is likely that infiltration rates may be limited and there is a greater chance of the melt generating a downslope surface runoff through the base of the pack. If the soil is unfrozen, then the low intensity of the melt will usually mean that the bulk of the melt will infiltrate into the soil profile. Depending on the weather conditions prior to a pack being established, it is quite possible that in some years the soil surface remains frozen all winter, while in other years the surface is unfrozen at the start of the melt season. The responses expected during melt might then be very different in different years.

Melt rates can be greatly accelerated, if warm rain falls on a ripe snowpack, and in some parts of the world rain on snow events can be a significant cause of flooding. The rain adds both water to the event volume and heat resulting in increased rates of melt. This type of event was involved in the northern California floods of early 1996.

3 GENERATION OF SURFACE RUNOFF

Infiltration Excess Surface Runoff Generation

The classical model of surface runoff generation is by an infiltration excess mechanism in which rainfall intensity exceeds the local infiltration capacity of the soil for a sufficient period of time for any *depression storage* at the soil surface to be satisfied such that downslope flow is initiated. This will not occur where the permeability of the soil is high in comparison with expected rainfall intensities, and even when this is not the case there may be an initial period when all the rainfall infiltrates before bringing the surface to saturation (the *time to ponding*). There have been many infiltration equations proposed, either empirical or based on various approximate solutions to Darcy's law, with the aim of predicting times to ponding, infiltration capacities, and the production of surface runoff. All require the specification of some parameter values that measurements suggest may vary dramatically even within a single soil type. Spatial variability of soil characteristics may therefore be important in the production of surface runoff, since infiltration excess runoff will start first in areas of low infiltration capacity. Where runoff flows downslope onto areas of higher infiltration capacity as run-on, there may be further infiltration of part or all of the water before it reaches a stream channel.

Infiltration capacities of the soil can also be greatly enhanced by the presence of macropores [cracks, root channels, animal burrows see Beven and Germann (1982)] or reduced by the presence of surface crusting resulting from the redistribution of fine particles by raindrop impact (Römkens et al., 1990) such that the properties of the bulk soil matrix alone may not be a good indication of actual infiltration rates.

Saturation Excess Surface Runoff Generation

Surface runoff may be found in areas of high infiltration capacity soils if the rainfall falls on soil that is saturated to the surface or if return flow from the subsurface occurs onto a saturated surface. This may not, in fact, require saturation of the whole soil profile. The buildup of a perched water table, for example, due to a permeability break between horizons in the soil profile, might result in saturation to the surface and the consequent generation of surface runoff. Surface runoff produced in this way has been studied, for example, by Bonell et al. (1981).

Saturation is most likely to occur where upslope contributing areas are large, such as in valley bottoms and hillslope hollows, and where effective hydraulic gradients are small. Convergence of flow lines will tend to increase the likelihood of saturation; steep slopes will tend to decrease it. The topographic index, a/s, discussed above, reflects these counteracting tendencies. High values of the index (high contributing areas, low slope) will indicate, other things being equal, a higher likelihood of saturated conditions occurring; low values of the index indicate the reverse.

Channel Extension

It is not only runoff source areas on the hillslopes that exhibit dynamic extension during storm periods. Many studies have emphasized the role of extension of the ephemeral channel network in the generation of storm runoff (e.g., Hewlett, 1974). The area of the channel itself and the immediately adjacent *riparian* area can, in some catchments in which the storm hydrograph represents a volume equal to only a small proportion of the incident rainfall (a small *runoff coefficient*), be the most important source area of storm runoff.

4 EFFECT OF HETEROGENEITY

Heterogeneity of Hillslope Forms

Topography is an important control on runoff generation in catchments where downslope flows are an important control on the runoff response. The topographic index provides one very simple approach to identify likely runoff production areas. This will be particularly true for saturation excess runoff production or subsurface stormflows where topographic convergence is important. It may also be true for infiltration excess runoff production where a catenary relationship between soil textural properties and topographic position has developed due to translocation of clays and other long-term processes. The topographic index will be limited as a predictor wherever flow lines in the subsurface depart radically from the surface topography, as in fractured systems or deep groundwater systems, or where the soil is relatively dry such that the effective upslope contributing areas for subsurface flow are very small (e.g., Barling et al., 1994).

Heterogeneity of Soil Characteristics

It has also been noted that variability of soil characteristics may have an important control on runoff production for all the suggested mechanisms. This will be true for the variability of soil series at the catchment scale and also for the variability in characteristics found within a soil series that may not show clear spatial patterns. Variability in soil characteristics associated with the occurrence of vegetation may also be important. Dunne et al. (1991), for example, suggest that the infiltration capacities of the soil beneath the plants of a sparse vegetation canopy can be much higher than between the plants, to the extent that depths of overland flow generated during an event may be greatly affected by local infiltration beneath the plants. Similarly in the Tiger Bush area of the HAPEX-SAHEL experiment in Niger, it is thought that surface runoff from bare soil areas may serve to increase the water available to the adjacent stripes of Tiger Bush (Peugeot et al., 1997).

In all these cases, it is clear that the knowledge of mean soil hydrological characteristics may not be sufficient to understand surface runoff production. Rather the distribution of characteristics within a catchment area may be important. Even with detailed information about soil properties, there is some doubt that the nature of runoff production is truly predictable in detail, especially for infiltration excess type mechanisms. The plot studies of Hjelmfelt and Burwell (1984), for example, suggested that measured surface runoff from adjacent plots may be unpredictable while the study of Loague (1990) and Loague and Kyriakidis (1997) on the R-5 catchment at Chickasha, Ohio, revealed the difficulty of modeling spatially heterogeneous infiltration excess runoff production even on a small catchment with detailed soil measurements.

Heterogeneity of Vegetation

The effects of vegetation on the effective rainfall intensities at ground level and on infiltration rates have already been noted. There are also other effects associated with the heterogeneity of vegetation. If surface runoff is generated, then the effective roughness of the surface may be controlled primarily by the surface vegetation. On surfaces covered by grasses, for example, this may lead to very low velocities for surface runoff (velocities as low as 20 m/h have been measured in the field). The vegetation may also protect the surface from erosion by both rainsplash and flowing water.

Heterogeneity of Precipitation Inputs

Whatever the nature of the surface or subsurface runoff generation processes, the amount of runoff generation will predominantly be determined by the forcing of the rainstorm inputs and the state of antecedent wetness of the catchment. Incident precipitation intensities can vary greatly in space, particularly during convective rainstorm events when patterns of intensities depend strongly on the growth and decay of storm cells and the overall movement of the storm (e.g., Smith et al., 1996).

Runoff generation may depend strongly on patterns of rainfall intensity as suggested, for example, by simulations of the Walnut Gulch experimental catchment in Arizona by Goodrich et al. (1994).

5 IMPORTANCE OF RUNOFF IN GRID-SCALE LAND SURFACE MODELING FOR GCMs

In all past general circulation model (GCM) land surface model components, runoff has not been considered to be very important. It has generally been treated simply as an excess of water that magically disappears from the local water balance. In real catchments, of course, runoff does not disappear but may have an effect on the hydrology and energy balance of areas downslope or downstream. Far more effort, computer time, and parameters have been devoted to formulating the controls on the local energy fluxes of latent and sensible heat than the controls on runoff production. There are several good reasons why runoff production requires more attention, and this lack of attention is now starting to be redressed in the development of so-called macroscale hydrological models

Runoff in many environments is a major part of the water balance and in areas where availability of water is a critical control on latent heat fluxes, then estimating correctly the partitioning of the water balance into that part that is runoff and that part that is available for evapotranspiration may be crucial. This may be a more difficult problem than estimating an areal average evapotranspiration flux since, as is clear from the discussion above, runoff generation has an important spatial dimension due to dependencies on patterns of rainfall inputs, patterns of soil characteristics, and the effects of topography. An important reason why it might be important to improve the runoff generation algorithms in GCMs is that there are long-term discharge measurements available for model evaluation in some catchments at a range of scales and in a wide range of climatic conditions. There is also an increasing recognition that transfers across the grid elements of a GCM, by either surface or regional groundwater flows, may contribute to the controls on patterns of inputs of freshwater to the oceans. At least in areas subject to seasonal flooding (such as the Amazon, Nile, Niger, and other large river basins), such transfers may also control the magnitude of latent heat fluxes over extensive areas.

Macroscale representations of runoff generation are still at an early stage and, given the dependencies on complex spatial heterogeneities and antecedent conditions, it is still not clear as to what an appropriate strategy will be for the formulation and parameter identification of a large-scale model.

REFERENCES

Barling, R. D., I. D. Moore, and R. B. Grayson, A quasi-dynamic wetness index for characterising the spatial distributioin of zones of surface saturation and soil water content, *Water Resour. Res.*, *30*, 1029–1044, 1994.

Bazemore, D. E., K. N. Eshleman, and K. J. Hollenbeek, The role of soil water in stormflow generation in a forested headwater catchment: Synthesis of natural tracer and hydrometeric evidence, *J. Hydrol.*, *162*, 47–75, 1994.

Beven, K. J., TOPMODEL: A critique, *Hydrol. Process.*, *11*(9), 1069–1085, 1997.

Beven, K. J., *Rainfall-Runoff Modelling—The Primer*, Wiley, Chichester, 2001.

Beven, K. J., and P. E. Germann, Macropores and water flow in soils, *Water Resour. Res.*, *18*, 1311–1325, 1982.

Beven, K. J., R. Lamb, P. Quinn, R. Romanowicz, and J. Freer, TOPMODEL, in V.P. Singh (Ed.), *Computer Models of Watershed Hydrology*, Water Resource Publications, Co., 1995, pp. 627–668.

Bonell, M., D. A. Gilmour, D. F. Sinclair, Soil hydraulic properties and their effect on surface and subsurface water transfer in a tropical rainforest catchment, *Hydrol. Sci. Bull.*, *16*, 1–18, 1981.

Dunne, T., Field studies of hillslope flow processes, in M. J. Kirkby (Ed.), *Hillslope Hydrology*, Wiley, Chichester, 1978, pp. 227–293.

Dunne, T., W. Zhang, and B. F. Aubrey, Effects of rainfall, vegetation and microtopography on infiltration and runoff, *Water Resour. Res.*, *27*, 2271–2286, 1991.

Famiglietti, J. S., E. F. Wood, M. Sivapalan, and D. J. Thongs, A catchment scale water balance model for FIFE, *J. Geophys. Res.*, *97*(D17), 18997–19007, 1992.

Genereux, D. P., H. F. Hemond, and P. J. Mulholland, Spatial and temporal variability in streamflow generation on the West Fork of Walker Branch Watershed, *J. Hydrol.*, *142*, 137–166, 1993.

Goodrich, D. C., T. J. Schmugge, T. J. Jackson, C. L. Unkrich, T. O. Keefer, R. Parry, L. B. Bach, and S. A. Amer, Runoff simulation sensitivity to remotely sensed initial soil moisture content, *Water Resour. Res.*, *30*, 1393–1405, 1994.

Hewlett, J. D., Comments on letters relating to "Role of subsurface flow in generating surface runoff. 2. Upstream source areas" by R. Allen Freeze, *Water Resour. Res.*, *10*, 605–607, 1974.

Hjelmfelt, A. T., and R. E. Burwell, Spatial variability of runoff, *J. Irrig. Drain. Div. ASCE*, *110*, 46–54, 1984.

Horton, R. E. The role of infiltration in the hydrological cycle, *Trans. Am. Geophys. Union*, *14*, 446–460, 1933.

Hursh, C. R., Subsurface flow, *Trans. Am. Geophys. Union*, *25*, 743–746, 1944.

Kirkby, M. J., Implications for sediment transport, in M. J. Kirkby (Ed.), *Hillslope Hydrology*, Wiley, Chichester, 1978, pp. 325–363.

Loague, K. M., R-5 revisited: 2. Reevaluation of a quasi-physically based rainfall-runoff model with supplemental information, *Water Resour. Res.*, *26*, 973–987, 1990.

Loague, K. M., and P. C. Kyriakidis, Spatial and temporal variability in the R-5 infiltration data set: Déjà vu and rainfall-runoff simulations, *Water Resour. Res.*, *33*, 2883–2896, 1997.

McDonnell, J. J., A rationale for old water discharge through macropores in a steep humid catchment, *Water Resour. Res.*, *26*(11), 2821–2832, 1990.

O'Loughlin, E. M. Saturation regions in catchments and their relations to soil and topographic properties, *J. Hydrol.*, *53*, 229–246, 1981.

Peugeot, C., M. Esteves, S. Galle, J. L. Rajot, and J. P. Vandervaere, Runoff generation processes: Results and analysis of field data collected at the East Central Supersite of the HAPEX-Sabel experiment, *J. Hydrol.*, *188*, 203–223, 1997.

Rasmussen, T. C., R. H. Baldwin, J. F. Dowd, and A. G. Williams, Tracer versus pressure wave velocities through unsaturated saprolite, *Soil Sci. Soc. Am. J.*, *64*, 75–85, 2000.

Rawls, W. J., and D. L. Brakensiek, Estimation of soil hydraulic properties, in H. J. Morel-Seytoux (Ed.), *Unsaturated Flow in Hydrologic Modeling*, Reidel, Dordrecht, 1989, pp. 275–300.

Römkens, M. J. M., S. N. Prasad, and F. D. Whisler, Surface sealing and infiltration, in M. G. Anderson and T. P. Burt (Eds.), *Processs Studies in Hillslope Hydrology*, 1990, pp. 127–172, Wiley, Chichester.

Sklash, M. G., Environmental isotope studies of storm and snowmelt runoff generation, in M. G. Anderson, and T. P. Burt (Eds.), *Process Studies in Hillslope Hydrology*, 1990, pp. 401–436, Wiley, Chichester.

Smith, J. A., M. L. Baeck, M. Steiner, and A. J. Miller, Catastrophic rainfall from an upslope thunderstorm in the central Appalachians, the Rapidan storm of June 27, 1995, *Water Resour. Res.*, *32*, 3099–3113, 1996.

Weyman, D. R., Throughflow on hillsides and its relation to the stream hydrograph, *Bull. Int. Assoc. Sci. Hydrol.*, *15*, 25–33, 1970.

CHAPTER 30

FLOW ROUTING

D. L. FREAD

1 INTRODUCTION

Flow routing is a mathematical method (model) to predict the changing magnitude, speed, and shape of a flood wave at one or more locations along waterways such as rivers, reservoirs, canals, or estuaries. The flood wave can emanate from precipitation runoff (rainfall or snowmelt), reservoir releases (spillway flows or dam failures), and tides (astronomical and/or wind generated).

Flow routing has long been of vital concern and many ways have been developed to predict the characteristic features of a flood wave to improve the transport of water through natural or man-made waterways and to determine necessary actions to protect life and property from the effects of flooding. Commencing with investigations by Newton (1687), Laplace (1776), Poisson (1816), Boussinesq (1871), and culminating in the one-dimensional equations of unsteady flow derived by Barré de Saint-Venant (1871), the theoretical foundation for flow routing was essentially achieved. The original Saint-Venant equations are the conservation of mass equation:

$$\partial(AV)/\partial x + \partial A/\partial t = 0 \tag{1}$$

and the conservation of momentum equation:

$$\partial V/\partial t + V\ \partial V/\partial x + g(\partial h/\partial x + S_f) = 0 \tag{2}$$

in which t is time, x is distance along the longitudinal axis of the waterway, A is cross-sectional area, V is velocity, g is the gravity acceleration constant, and h is the water-surface elevation above a datum; S_f is the friction slope, which may now be

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

evaluated using a steady flow empirical formula such as the Manning equation (Manning, 1889; Chow, 1959), i.e.,

$$Q = \mu A R^{2/3} S_0^{1/2}/n \tag{3}$$

in which $Q = AV$ is discharge or flow, $R = A/P$ is the hydraulic radius, and P is the wetted perimeter of the cross section, S_0 is the channel bottom slope (dimensionless), μ is a units conversion factor, i.e., 1.49 for U.S. units or 1.0 for SI, and n is the Manning roughness (friction) coefficient. Equations (1) and (2) are quasi-linear hyperbolic partial differential equations with two dependent parameters (V and h) and two independent parameters (x and t); A is a known function of h, and S_f is a known function of V and h. Derivations of the Saint-Venant equations can be found in the following references: Stoker (1957), Henderson (1966), Strelkoff (1969), and Liggett (1975).

Due to the mathematical complexity of the Saint-Venant equations (no analytical solution is known), simplifications were necessary to obtain feasible solutions for the salient characteristics of a propagating flood wave. This approach produced a profusion of simplified flow routing models. The simplified flow routing models may be categorized as: (I) purely empirical, (II) storage routing, based on the conservation of mass and an approximate relation between flow and storage, and (III) hydraulic, i.e., based on the conservation of mass and a simplified form of the conservation of momentum equation (2).

Categories I and II are further classified as lumped flow routing techniques in which the flow is computed as a function of time, only at the most downstream location of routing reaches along the waterway. Category III can be classified as distributed flow routing techniques in which flow and depth or water-surface elevation are computed as a function of time at frequent locations within routing reaches along the waterway. During the last two decades dynamic hydraulic distributed flow routing methods based on numerical solutions of the complete Saint-Venant equations have become economically feasible as a result of advances in computing equipment and improved numerical solution techniques. Following is a brief description of some of the more popular storage routing models as well as both simplified and dynamic hydraulic flow routing models.

2 STORAGE ROUTING MODELS

Significant river improvement projects in the early 1900s provided the impetus for development of an array of simplified flow routing methods. These have been termed storage routing models. They are based on the conservation of mass equation (1) written in the following form:

$$\bar{I} - \bar{O} = \Delta \bar{S}/\Delta t \tag{4}$$

in which $\Delta\bar{S}$ is the change in storage within the routing reach during a Δt time increment, $\bar{I} = 0.5[I(t) + I(t + \Delta t)]$, and $\bar{O} = 0.5[O(t) + O(t + \Delta t)]$; the storage $(\bar{S})$ is assumed to be related to inflow $(\bar{I})$ and/or outflow $(\bar{O})$, i.e.,

$$\bar{S} = \bar{K}[\bar{X}\bar{I} + (1 - \bar{X})\bar{O}] \tag{5}$$

in which $\bar{K}$ is a storage constant with dimensions of time, and $\bar{X}$ is a weighting coefficient, $0 \le \bar{X} \le 1$. Storage routing models are limited to typical flood routing applications where the outflow and water-surface elevation relation is essentially single valued, and the waterways are not mild sloping ($S_0 > 0.002$). Thus, backwater effects from tides, significant tributary inflow, and dams or bridges are not considered by these models, nor are they well-suited for rapidly changing unsteady flows such as dam-break flood waves, reservoir power releases, or hurricane storm surges. Generally, storage routing models have two parameters that can be calibrated to effectively reproduce the flood wave speed and its attenuated peak. The calibration requires that most storage routing model applications be limited to where observed inflow–outflow hydrographs exist. When using the observed hydographs to calibrate the routing coefficients, variations in flood wave shapes within the observed data set are not considered, and only the average wave shape is reflected in the fitted routing coefficients.

Reservoir Storage Routing Model

Storage routing models applicable to reservoirs, which have essentially level water-surface profiles, can be developed by assuming $\bar{X}$ to be zero in Eq. (5), i.e., storage is dependent only on outflow. Expressing the term $\Delta\bar{S}/\Delta t$ in Eq. (4) as the product of reservoir surface area (S_a), which is a known function of water-surface elevation (h) and the change of h over a j Δt time step, i.e.,

$$\Delta\bar{S}/\Delta t = 0.5(S_a^j + S_a^{j+1})(h^{j+1} - h^j)/\Delta t^j \tag{6}$$

Now denoting $\bar{O}$ (outflow) as $\bar{Q}$ (discharge), the following reservoir routing model (Fread, 1977) is obtained:

$$0.5(I^j + I^{j+I}) - 0.5(Q^j + Q^{j+1}) - 0.5(S_a^j + S_a^{j+1})(h^{j+1} - h^j)/\Delta t^j = 0 \tag{7}$$

The inflows (I) at times j and $j + 1$ are known from the specified inflow hydrograph; the outflow (Q^j) at time j can be computed from the known water-surface elevation (h^j) and an appropriate spillway discharge equation. The surface area (S_a^j) can be determined from the known value of h^j. The unknowns in the equation consist of h^{j+1}, Q^{j+1}, and S_a^{j+1}; the latter two are known nonlinear functions of h^{j+1}. Hence, Eq. (7) can be solved for h^{j+1} by an iterative method such as Newton–Raphson, i.e.,

$$h_{k+1}^{j+1} = h_k^{j+1} - f(h_k^{j+1})/f'(h_k^{j+1}) \tag{8}$$

in which k is the iteration counter; and $f(h_k^{j+1})$ is the left-hand side of Eq. (7) evaluated with the first estimate for h_k^{j+1}, which for $k = 1$ is either h^j or a linear extrapolated estimate of h^{j+1}; $f'(h_k^{j+1})$; is the derivative of Eq. (7) with respect to h^{j+1}. It can be approximated by using a numerical derivative as follows:

$$f'(h_k^{j+1}) = [f(h_k^{j+1} + \varepsilon) - f(h_k^{j+1} - \varepsilon)]/[(h_k^{j+1} + \varepsilon) - (h_k^{j+1} - \varepsilon)] \tag{9}$$

in which ε is a small value, say 0.1 ft (0.03 m). Using Eq. (8), only one or two iterations are usually required to solve Eq. (7) for h^{j+1}. Initially, the reservoir pool elevation (h^j) must be known to start the computational process. Once h^{j+1} is obtained, Q^{j+1} can be computed from the spillway discharge equation, $Q = f(h_k^{j+1})$.

Level-pool routing is less accurate as the reservoir length increases, as the reservoir mean depth decreases, and as the time of rise of the inflow hydrograph decreases (Fread, 1992). This inaccuracy can have significant economic effects on water control management practices (Sayed and Howard, 1983).

Muskingum Model

A widely used hydrologic flow routing model is the Muskingum model developed by using Eq. (5), with nonzero values for both $\bar{K}$ and $\bar{X}$, for the storage relationship. Substituting this information into Eq. (4), the following is obtained for computing $O(t + \Delta t)$:

$$O(t + \Delta t) = C_1 I(t + \Delta t) + C_2 I(t) + C_3 O(t) \tag{10}$$

where

$$C_0 = \bar{K} - \bar{K}\bar{X} + \Delta t/2 \tag{11}$$

$$C_1 = -(\bar{K}\bar{X} - \Delta t/2)/C_0 \tag{12}$$

$$C_2 = (\bar{K}\bar{X} + \Delta t/2)/C_0 \tag{13}$$

$$C_3 = (\bar{K} - \bar{K}\bar{X} - \Delta t/2)/C_0 \tag{14}$$

and where $C_1 + C_2 + C_3 = 1$ and $\bar{K}/3 \leq \Delta t \leq \bar{K}$ is usually the range for Δt.

Equation (10) is the widely used Muskingum routing model first developed by McCarthy (1938). The parameters $\bar{K}$ and $\bar{X}$ are determined from observed inflow–outflow hydrographs using least squares or its equivalent, the graphical method, or other techniques (Singh and McCann, 1980). Among the many descriptions and variations of the Muskingum model are Chow (1964); Chow et al. (1988), Strupczewski and Kundzewicz (1980), Dooge et al. (1982), and Linsley et al. (1986).

Muskingum–Cunge Model

A significant improvement of the Muskingum model was developed by Cunge (1969) known as the Muskingum–Cunge model. This increased the Muskingum model's accuracy and made it applicable in situations where observed inflow and outflow hydrographs were not available for calibration and enabled it to be changed from a lumped to a distributed flow routing model. Cunge derived Eq. (10) using the assumption of a single-valued $Q(h)$ relation, the classical kinematic wave equation [see Eq. (25)], and applying a four-point implicit finite-difference approximation technique. Equation (10) is rewritten where the flows $I(t)$, $I(t+\Delta t)$, $O(t)$, and $O(t+\Delta t)$ are replaced by Q_i^j, Q_i^{j+1}, Q_{i+1}^j, and Q_{i+1}^{j+1}, respectively, i.e.,

$$Q_{i+1}^{j+1} = C_1 Q_i^{j+1} + C_2 Q_i^j + C_3 Q_{i+1}^j + C_4 \tag{15}$$

$$C_4 = 0.5[q(t) + q(t+\Delta t)]\Delta x\, \Delta t / C_0 \tag{16}$$

Equation (15) has been expanded to include effects of lateral flow (q) along the ΔX routing reach; and where the following expressions for $\bar{K}$ and $\bar{X}$ are determined:

$$\bar{K} = \Delta x / \bar{c} \tag{17}$$

$$\bar{X} = 0.5[1 - \bar{Q}/(\bar{c}\bar{B}S_e\, \Delta x)] \tag{18}$$

where

$$\bar{c} = dQ/dA \tag{19}$$

in which $\bar{c}$ is the kinematic wave speed, Δx is the reach length, and S_e is the energy slope approximated by evaluating S_0 in Eq. (3) for the initial flow condition. The overbar indicates the variable is averaged over the Δx reach and over the Δt time step. Equation (19) may be expressed in an alternative form, i.e.,

$$\bar{c} = K'\bar{Q}/\bar{A} \tag{20}$$

where

$$K' = \tfrac{5}{3} - \tfrac{2}{3}(d\bar{B}/dy)\bar{A}/(\bar{B})^2 \tag{21}$$

in which A is the cross-sectional area, B is the channel width at the water surface, h is the water-surface elevation of the flow, and the Manning equation is used to relate discharge (Q) and depth or water-surface elevation (h). Depending on the cross-section shape, K' may have values in the range $\frac{4}{3} \le K' \le \frac{5}{3}$; the upper value is associated with either a very wide or rectangular channel. Selection of the appropriate time step Δt in secs is given by:

$$\Delta t \le 3600\, T_r/M \tag{22}$$

where T_r is the time of rise in hours of the inflow hydrograph and M is an integer ($10 \le M \le 20$) whose value depends on the extent of variation in the inflow hydrograph. The selection of Δx affects the accuracy of the solution. It is related to Δt and is limited by the following inequality (Jones, 1981):

$$\Delta x \leq 0.5c\,\Delta t[1 + (1 + 1.5Q/(Bc^2 S_0\,\Delta t))^{1/2}] \tag{23}$$

While the Muskingum–Cunge model does not require observed inflow–outflow hydrographs to establish the routing coefficients as required in the Muskingum model, best results are obtained if the wave speed (c) is determined from actual flow data. Also, the model is restricted to applications where backwater is not significant and discharge-water elevation rating curves do not have significant loops and discharge hydrographs are not rapidly changing with time such as dam-break floods. Nonetheless, the Muskingum–Cunge model (Miller and Cunge, 1975; Ponce and Yevjevich, 1978) is a highly versatile simplified routing model.

3 SIMPLIFIED HYDRAULIC ROUTING MODELS

Prior to computers, or more recently the feasible economical availability of such computational resources, the inability to obtain feasible numerical solutions to the complete Saint-Venant equations resulted in the development of several simplified distributed hydraulic routing models. They are based on the mass conservation equation (1) and various simplifications of the momentum equation (2).

Kinematic Wave Model

The most simple type of distributed hydraulic routing model is the kinematic wave model. It is based on the following simplified form of the momentum equation (2):

$$S_f - S_0 = 0 \tag{24}$$

in which S_0 is the bottom slope of the channel (waterway) and a component of the term, $\partial h/\partial x = \partial y/\partial x - S_0$, in which $\partial y/\partial x$ is assumed to be zero. This assumes that the momentum of the unsteady flow is the same as that of steady, uniform flow described by the Manning equation or a similar expression in which discharge is a single-valued function of depth, i.e., $\partial Q/\partial A = dQ/dA = c$. Also, since $\partial A/\partial t = (\partial A/\partial Q)(\partial Q/\partial t)$ and $Q = AV$, Eq. (1) can be expanded into the classical kinematic wave equation, i.e.,

$$\partial Q/\partial t + c\,\partial Q/\partial x = 0 \tag{25}$$

in which the kinematic wave velocity or celerity (c) is defined by Eq. (20).

Solutions for the kinematic wave equation (25) can be achieved using the method of characteristics or directly by finite-difference approximation techniques of either explicit or implicit types (Chow et al., 1988; Hydrologic Engr. Ctr., 1981; Linsley et al., 1986). The kinematic wave equation does not theoretically account for hydrograph (wave) attenuation. It is only through the numerical error associated with the finite-difference solution that attenuation of the hydrograph peak is achieved. Kinematic wave models are limited to applications where single-value, stage-discharge ratings exist—where there are no loop ratings—and where backwater effects are insignificant. Since, in kinematic wave models, flow disturbances can propagate only in the downstream direction, reverse (negative) flows cannot be predicted. Kinematic wave models are appropriately used as components of hydrologic watershed models for overland flow routing of runoff; they are not recommended for channel routing unless the hydograph is very slow rising, the channel slope is moderate to steep, and hydrograph attenuation is quite small. The range of application (with expected modeling errors less than 5%) for kinematic models, including the Muskingum method, is given by the following:

$$T_r S_0^{1.6}/(q_0^{0.2} n^{1.2}) \geq 0.014 \tag{26}$$

in which T_r is the time (in hours) of rise of the wave (hydrograph), i.e., the interval of time from beginning of significant rise to when the peak occurs; S_0 is the bottom slope (in ft/ft), q_0 is the unit-width discharge (Q/B) (in ft^2/s), and n is the Manning roughness coefficient (Fread, 1985, 1992).

Diffusion Wave Model

Another simplified distributed routing model, known as the diffusion wave (zero inertia) model, is based on Eq. (1) along with an approximation of the momentum equation that retains only the last two terms in Eq. (2), i.e.,

$$\partial h/\partial x + S_f = 0 \tag{27}$$

Finite-difference approximation techniques, both explicit and implicit (Strelkoff and Katopodes, 1977), have been used to obtain simultaneous solutions to Eqs. (1) and (27). The diffusion-simplified routing model considers backwater effects; however, its accuracy is deficient for very fast rising hydographs, such as those resulting from dam failures, hurricane storm surges, or rapid reservoir releases, which propagate through mild to flat sloping waterways with medium to small Manning's n. The range of application (with expected modeling errors less than 5%) for the diffusion models, including the Muskingum–Cunge model, is given by the following (Fread, 1992):

$$T_r S_0^{0.7} n^{0.6}/q_0^{0.4} \geq 0.0003 \tag{28}$$

4 DYNAMIC ROUTING MODEL

When the complete Saint-Venant equations [(1) and (2)] are used, the routing model is known as a dynamic routing model. With the advent of high-speed computers, Stoker (1953) and Isaacson et al. (1954, 1956) first attempted to use the complete Saint-Venant equations for routing Ohio River floods. Since then, much effort has been expended on the development of dynamic routing models. Many models have been reported in the literature (Fread, 1985, 1992; Liggett and Cunge, 1975).

Dynamic routing models can be categorized as characteristic or direct methods of solving the Saint-Venant equations. In the characteristic methods, the Saint-Venant equations are first transformed into an equivalent set of four ordinary differential equations that are then approximated with finite differences to obtain solutions. Characteristic methods (Abbott, 1966; Henderson, 1966; Streeter and Wylie, 1967; Baltzer and Lai, 1968) have not proven advantageous over the direct methods for practical flow routing applications.

Direct methods can be classified further as either explicit or implicit. Explicit schemes (Stoker, 1953, 1957; Isaacson et al., 1954; Garrison et al., 1969; Dronkers, 1969; Strelkoff, 1970; Liggett and Cunge, 1975; Veissman et al., 1977; Linsley et al., 1986) transform the differential equations into a set of algebraic equations that are solved sequentially for the unknown flow properties at each cross section at a given time. However, implicit schemes (Preissman, 1961; Amein and Fang, 1970; Strelkoff, 1970; Fread, 1973, 1977, 1978, 1985; Liggett and Cunge, 1975; Cunge et al., 1980; Schaffranek, 1987; Fread and Lewis, 1998; Chow et al., 1988; Barkow, 1990) transform the Saint-Venant equations into a set of algebraic equations that must be solved simultaneously for all Δx computational reaches at a given time; this set of simultaneous equations may be either linear or nonlinear, the latter requiring an iterative solution procedure.

Explicit methods, although simpler in application, are restricted by numerical stability considerations. Stability problems arise when inevitable errors in computational roundoff and those introduced in approximating the partial differential equations via finite differences accumulate to the point that they destroy the usefulness and integrity of the solution, if not the total breakdown of the computations, by creating artificial oscillations of length about $2\Delta x$ in the solution. Due to stability requirements, explicit methods require very small computational time steps on the order of a few seconds or minutes depending on the ratio of the computational reach length (Δx) to the minimum dynamic wave celerity (u), i.e., $\Delta t \leq \Delta x/u$, where $u = V + (gA/B)^{1/2}$. This is known as the Courant condition, and it restricts the time step to less than that required for an infinitesimal disturbance to travel the Δx distance. Such small time steps cause explicit methods to be inefficient in the use of computer time.

Implicit finite-difference techniques, however, have no restrictions on the size of the time step due to mathematical stability; however, numerical convergence (accuracy) considerations require some limitation in time step size (Fread, 1974; Cunge et al., 1980). Implicit techniques are generally preferred over explicit because of their computational efficiency. Therefore, an implicit scheme will be subsequently described in detail herein. Rather than using finite-difference approximation techniques to solve the Saint-Venant equations, finite-element techniques (Gray et al., 1977; DeLong, 1986, 1989) can be used; however, their greater complexity offsets any apparent advantages when compared to a weighted, four-point implicit finite-difference scheme (described later) for solving the one-dimensional flow equations. However, finite-element techniques are often applied to two- and three-dimensional flow computations.

Saint-Venant Equations

A modified and expanded form (Fread, 1988, 1992) of the original one-dimensional Saint-Venant equations [(1) and (2)] consist of the conservation of mass equation, i.e.,

$$\partial Q/\partial x + \partial s_c(A + A_0)/\partial t - q = 0 \tag{29}$$

and the momentum equation, i.e.,

$$\sigma[\partial(s_m Q)/\partial t + \partial(\beta Q^2/A)/\partial x] + gA(\partial h/\partial x + S_f + S_{ec} + S_i) + L + W_f B = 0 \tag{30}$$

where Q is discharge, h is the water-surface elevation, A is the active cross-sectional area of flow, A_0 is the inactive (off-channel storage) cross-sectional area, s_c and s_m are area-weighted and conveyance-weighted sinuosity factors, respectively (DeLong, 1986, 1989), which correct for the departure of a sinuous in-bank channel from the x-axis of the floodplain, x is the longitudinal mean-flow-path distance measured along the center of the waterway (channel and floodplain), t is time, q is the lateral inflow or outflow per lineal distance along the waterway (inflow is positive and outflow is negative), σ is a numerical filter ($0 \leq \sigma \leq 1$, usually $\sigma = 1$) to enable the equations to properly handle mixed subcritical/supercritical flows (Fread et al., 1996) during the numerical solution (see the discussion on subcritical/supercritical mixed flow for more on σ later in this chapter), β is the momentum coefficient for nonuniform velocity distribution within the cross section, g is the gravity acceleration constant, S_f is the boundary friction slope, S_{ec} is the expansion/contraction (large eddy loss) slope, and S_i is the viscous dissipation slope for mud/debris flows.

Friction Slope. The boundary friction slope (S_f) is evaluated by rearranging the Manning Eq. (3) for uniform, steady flow into the following form:

$$S_f = n^2|Q|Q/(\mu^2 A^2 R^{4/3}) = |Q|Q/K^2 \tag{31}$$

in which n is the Manning coefficient of frictional resistance (Chow, 1959; Barnes, 1967; Arcement and Schneider, 1984; Jarrett, 1984; and Fread, 1989), R is the hydraulic radius, μ is a units conversion factor (1.49 for U.S. units and 1.0 for SI), and K is the channel conveyance factor. The absolute value of Q is used to correctly account for the possible occurrence of reverse (negative) flows. The conveyance formulation is preferred (for numerical and accuracy considerations) for composite channels having wide, flat overbanks or floodplains in which K represents the sum of the conveyance of the channel (which is corrected for sinuosity effects by dividing by s_m), and the conveyances of left and right floodplain areas.

When the conveyance factor (K) is used to evaluate S_f, the river channel/valley cross-sectional properties are designated as left floodplain, channel, and right floodplain rather than as a composite channel/valley section. Special orientation for designating left or right is not required as long as consistency is maintained. The conveyance factor is evaluated as follows (Fread and Lewis, 1998):

$$K = K_l + K_c + K_r \tag{32}$$

where:

$$K_l = \frac{\mu}{n_l} A_l R_l^{2/3} \tag{33}$$

$$K_c = \frac{\mu A_c R_c^{2/3}}{n_c s_m^{1/2}} \tag{34}$$

$$K_r = \frac{\mu}{n_r} A_r R_r^{2/3} \tag{35}$$

in which the subscripts l, c, and r designate left floodplain, channel, and right floodplain, respectively.

Sinuosity Factors. The area-weighted and conveyance-weighted sinuosity factors (s_c and s_m, respectively) in Eqs. (29), (30), and (34) represent the ratio(s) of the flow-path distance along a meandering channel to the mean-flow-path distance along the floodplain. Initially, only one sinuosity factor (s_k) is specified as varying only with each Jth depth of flow ($J = 1, 2, \ldots, \hat{J}$, where $\hat{J}$ is the number of user-specified tabular top widths (B) versus h values, which describe the cross-section geometry), but then this is recomputed within the model according to the following relations:

$$s_{cJ} = \sum_{k=2}^{k=J} (\Delta A_{lk} + \Delta A_{ck} s_k + \Delta A_{rk})/(A_{lJ} + A_{cJ} + A_{rJ}) \tag{36}$$

$$s_{mJ} = \sum_{k=2}^{k=J} (\Delta K_{lk} + \Delta K_{ck} s_k + \Delta K_{rk})/(K_{lJ} + K_{cJ} + K_{rJ}) \tag{37}$$

in which $\Delta A = A_{J+1} - A_J$, and s_k represents the sinuosity factor for a differential portion of the flow between the Jth depth and the $J + 1$th depth, and K is the conveyance factor.

Expansion/Contraction Effects. The term S_{ec} is computed as follows:

$$S_{\mathrm{ec}} = k_{\mathrm{ec}}\Delta(Q/A)^2/(2g\,\Delta x) \tag{38}$$

in which k_{ec} is the expansion/contraction coefficient (negative for expansion/positive for contraction), which varies from $-1.0/0.4$ for an abrupt change in section geometry to $-0.3/0.1$ for a very gradual, curvilinear transition between cross sections. The Δ represents the difference in the term $(Q/A)^2$ at two adjacent cross sections separated by a distance Δx. If the flow direction changes from downstream to upstream, k_{ec} can be automatically changed (Fread, 1988).

Large floods such as dam-break-generated floods usually have much greater velocities; it is important, especially for nonuniform channels (Rajar, 1978) to include in the Saint-Venant momentum equation (30) the expansion/contraction losses via the S_{ec} term defined by equation (38). The ratio of expansion/contraction action losses (form losses) to the friction losses can be in the range of $0.01 < S_{\mathrm{ec}}/S_f < 1.0$. The larger ratios occur for very irregular channels with relatively small n values and for flows with large velocities (dam-break floods).

Momentum Correction Coefficient. The momentum correction coefficient (β) for nonuniform velocity distribution across the cross section is (Chow, 1959)

$$\beta = (K_l^2/A_l + K_c^2/A_c + K_r^2/A_r)/[(K_l + K_c + K_r)^2/(A_l + A_c + A_r)] \tag{39}$$

in which K is conveyance, A is wetted area, and the subscripts l, c, and r denote left floodplain, channel, and right floodplain, respectively. When floodplain properties are not separately specified and the total cross section is treated as a composite section, β can be approximated as $1.0 \leq \beta \leq 1.06$ in lieu of Eq. (39). Also, in this case, S_c and S_m are set to unity in lieu of Eqs. (36) and (37).

Lateral Flow Momentum. The term L in Eq. (30) is the momentum effect of lateral flows and has the following form (Strelkoff, 1969): (a) lateral inflow, $L = -qv_x$, where v_x is the velocity of lateral inflow in the x direction of the main channel flow; (b) seepage lateral outflow, $L = -0.5Q/A$; and (c) bulk lateral outflow, $L = -qQ/A$.

Mud or Debris Flows. The friction loss term (S_i) is included (Fread, 1988) in the momentum equation (30) in addition to S_f to account for viscous dissipation effects of non-Newtonian flows such as mud or debris flows. Also, mine tailings dams, where the viscous contents retained by the dam have non-Newtonian properties, are dam-breach flood applications requiring the use of S_i in Eq. (30). This effect becomes significant only when the solids concentration of the flow is greater than about 40% by volume. For concentrations of solids greater than about 50%, the flow behaves more as a landslide and is not governed by the Saint-Venant equations. S_i is evaluated for any non-Newtonian flow as follows (Jin and Fread, 1997):

$$S_i = \frac{\tau_y}{\gamma D}\left[1 + \left(\frac{(b+1)(b+2)Q}{(0.74 + 0.66b)(\tau_y/\kappa)^b DA}\right)^{1/b+0.15}\right] \tag{40}$$

in which γ is the fluid's unit weight, τ_0 is the fluid's yield strength, D is the hydraulic depth (A/B), $b = 1/m$ where m is the exponent of the power function that fits the fluid's stress(τ_s)–strain(dv/dy) properties, and κ is the apparent viscosity or scale factor of the power function, i.e., $\tau_s = \tau_0 + \kappa(dv/dy)^m$. The viscous properties, τ_0 and κ, can be estimated from the solids concentration ratio of the mud flow (O'Brien and Julien, 1984).

Wind Effects. The last term $(W_f B)$ in Eq. (30) represents the resistance effect of wind on the water surface (Fread, 1985, 1992); B is the wetted topwidth of the active flow portion of the cross section; and $W_f = V_r|V_r|c_w$, where the wind velocity relative to the water is $V_r = V_w \cos w + V$, V_w is the velocity of the wind, *positive* (+) if opposing the flow velocity and negative (−) if aiding the flow, w is the acute angle the wind direction makes with the x-axis, V is the velocity of the unsteady flow, and c_w is a wind friction coefficient $(1 \times 10^{-6} \le c_w \le 3 \times 10^{-6})$. This modeling capability can be used to simulate the effect of potential dam overtopping due to wind setup within a reservoir by applying the Saint-Venant equations to the unsteady flow in a reservoir.

Implicit Four-Point, Finite-Difference Approximations

The extended Saint-Venant equations [(29) and (30)] constitute a system of partial differential equations with two independent variables, x and t, and two dependent variables, h and Q; the remaining terms are either functions of x, t, h, and/or Q, or they are constants. The partial differential equations can be solved numerically by approximating each with a finite-difference algebraic equation; then the system of algebraic equations are solved in conformance with prescribed initial and boundary conditions.

Of various implicit, finite-difference solution schemes that have been developed, a four-point scheme first suggested by Issacson et al. (1954, 1956) and first used by Preissmann (1961) and later by Amein and Fang (1970) and then a weighted version by others (Fread, 1974, 1977, 1985, 1988; Cunge et al., 1980) is most advantageous. It is readily used with unequal distance steps, its stability–convergence properties are conveniently modified, and boundary conditions are easily applied.

Space–Time Plane. In the weighted four-point implicit scheme, the continuous x–t region in which solutions of h and Q are sought is represented by a rectangular grid of discrete points as shown in Figure 1. An x–t plane (solution domain) is a convenient means for expressing relationships among the variables. The grid points are determined by the intersection of lines drawn parallel to the x and t axes. Those

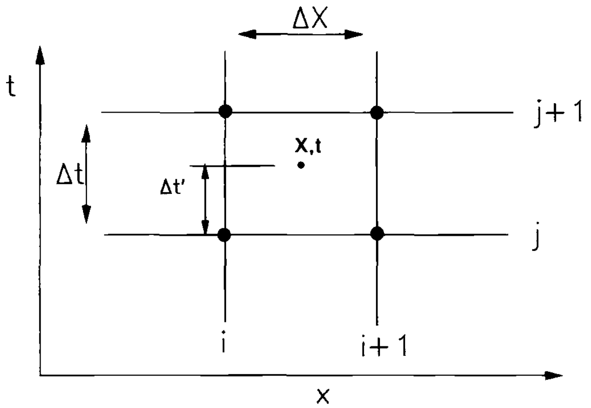


Figure 1 The x–t solution domain for the weighted four-point implicit scheme. See ftp site for color image.

parallel to the t axis represent locations of cross sections; they have a spacing of Δx, which need not be the same between each pair of cross sections. Those parallel to the x axis represent time lines; they have a spacing of Δt, which also need not be the same between successive time lines. Each point in the rectangular network can be identified by a subscript (i), which designates the x position or cross section, and a superscript (j), which designates the particular time line.

Numerical Approximations. The time derivatives are approximated by a forward-difference quotient at point (x', t') centered between the i and $i+1$ lines along the x axis as shown in Figure 1, i.e.,

$$\partial\phi/\partial t = (\phi_i^{j+1} + \phi_{i+1}^{j+1} - \phi_i^j - \phi_{i+1}^j)/2\Delta t_j \tag{41}$$

where ϕ represents any dependent variable or functional quantity (Q, s_c, s_m, A, A_0, q, h). Spatial derivatives are approximated at point (x', t') by a forward-difference quotient located between two adjacent time lines according to weighting factors of θ (the ratio $\Delta t'/\Delta t$ shown in Fig. 1) and $1 - \theta$, i.e.,

$$\partial\phi/\partial x = \theta(\phi_{i+1}^{j+1} - \phi_i^{j+1})/\Delta x_i + (1 - \theta)(\phi_{i+1}^j - \phi_i^j)/\Delta x_i \tag{42}$$

Nonderivative terms are approximated with weighting factors at the same time level [point (x', t)] where the spatial derivatives are evaluated, i.e.,

$$\phi = \theta(\phi_i^{j+1} + \phi_{i+1}^{j+1})/2 + (1 - \theta)(\phi_i^j + \phi_{i+1}^j)/2 \tag{43}$$

The weighted four-point implicit scheme is unconditionally, linearly stable for $\theta \geq 0.5$ (Fread, 1974); however, the sizes of the Δt and Δx computational steps are limited by the accuracy of the assumed linear variations of functions between the grid points in the x–t solution domain. Values of θ greater than 0.5 dampen parasitic oscillations that have wavelengths of about $2\Delta x$ that can grow enough to invalidate or destroy the solution. The θ weighting factor causes some loss of accuracy as it departs from 0.5, a box scheme, and approaches 1.0, a fully implicit scheme. This effect becomes more pronounced as the magnitude of the ratio $(T_r/\Delta t)$ decreases where T_r is the time of rise of the hydrograph (time interval from beginning of significant rise to peak of the hydrograph). Usually, a θ weighting factor of 0.60 is used to minimize the loss of accuracy while avoiding the possibility of weak (pseudo) instability for θ values of 0.5 when frictional effects are minimal.

Selection of* Δt *and* Δx *Computational Parameters. The computational time step (Δt) can be either specified or automatically determined to best suit the most rapidly rising hydrograph occurring within a system of rivers that may contain one or more breaching dams or other dynamic internal boundary conditions. The time step is selected according to the following:

$$\Delta t = T_r/M \tag{44}$$

where T_r is the minimum time of rise (seconds) of any hydrograph that has been specified at upstream boundaries or in the process of being generated at a breaching dam; M is user specified according to the following guidance (Fread, 1993):

$$M = 2.67[1 + \mu' n^{0.9}/(q^{0.1}S_0^{0.45})] \tag{45}$$

in which $\mu' = 3.97$ (3.13 SI units), n is the Manning friction coefficient, q is the peak flow per unit channel width (Q/B), and S_0 is the channel bottom slope; M usually varies within the range, $6 \leq M \leq 40$, with M often assumed to be approximately 20.

The Δx computational distance step can be specified or automatically determined according to the smaller of two criteria (Fread, 1993). The first criterion is

$$\Delta x \leq cT_r/20 \tag{46}$$

in which c is the bulk wave celerity (the celerity or velocity associated with an essential characteristic of the unsteady flow such as the peak of the hydrograph). In most applications, the wave velocity is well approximated as a kinematic wave, and c is estimated as $3/2V$ (V is the flow velocity) or c can be obtained by dividing the distance between two points along the channel by the difference in the times of occurrence of the peak of an observed or computed flow hydrograph at each point. Since c can vary along the channel, and depending upon the extent of this variation, Δx may not be constant along the channel.

The second criterion for selecting Δx is the restriction imposed by rapidly varying cross-sectional changes along the x axis of the waterway. Such expansion/contraction is limited to the following inequality (Samuels, 1985):

$$0.635 < A_{i+1}/A_i < 1.576 \qquad (47)$$

This condition results in the following approximation (Fread, 1988) for the maximum computational distance step:

$$\Delta x \leq L'/N' \qquad (48)$$

where

$$N' = 1 + 2|A_i - A_{i+1}|/\hat{A} \qquad (49)$$

in which L' is the distance between two adjacent (i and $i+1$) cross sections differing from one another by approximately 50% or greater, $\hat{A}$ is the active cross-sectional area, $\hat{A} = A_{i+1}$ if $A_i > A_{i+1}$ (contracting reach) or $\hat{A} = A_i$ if $A_i < A_{i+1}$ (expanding reach), and N' is rounded to the nearest integer value.

Significant changes in the bottom slope of the waterway also require small distance steps in the vicinity of the change. This is required particularly when the flow changes from subcritical to supercritical or conversely along the waterway. Such changes can require computational distance steps in the range of 50 to 200 ft (15 to 63 m).

Automatic Interpolation. It is convenient to automatically provide linearly interpolated cross sections at a user-specified spatial resolution to increase the spatial frequency at which solutions to the Saint-Venant equations are obtained. This is often required for purposes of attaining numerical accuracy/stability when (a) routing very sharp peaked hydrographs such as those generated by breached dams, (b) when adjacent cross sections either expand or contract by more than about 50%, and (c) where mixed flow changes from subcritical to supercritical or vice versa.

Algebraic Routing Equations. Using the finite-difference operators of Eqs. (41) to (43) to replace the derivatives and other variables in Eqs. (29) and (30), the following weighted four-point, implicit, nonlinear, finite-difference algebraic equations are obtained:

$$\theta\left[\frac{Q_{i+1}^{j+1}-Q_i^{j+1}}{\Delta x_i}\right]-\theta q_i^{j+1}+(1-\theta)\left[\frac{Q_{i+1}^{j}-Q_i^{j}}{\Delta x_i}\right]-(1-\theta)q_i^j$$
$$+\left[\frac{s_{c_i}^{j+1}(A+A_0)_i^{j+1}+s_{c_i}^{j+1}(A+A_0)_{i+1}^{j+1}-s_{c_i}^{j}(A+A_0)_i^{j}-s_{c_i}^{j}(A+A_0)_{i+1}^{j}}{2\Delta t_j}\right]=0 \quad (50)$$

$$\sigma\left[\frac{(s_{m_i}Q_i)^{j+1}+(s_{m_i}Q_{i+1})^{j+1}-(s_{m_i}Q_i)^{j}-(s_{m_i}Q_{i+1})^{j}}{2\Delta t_j}\right]$$
$$+\theta\left[\sigma\left(\frac{(\beta Q^2/A)_{i+1}^{j+1}-(\beta Q^2/A)_i^{j+1}}{\Delta x_i}\right)\right.$$
$$\left.+g\bar{A}_i^{j+1}\left(\frac{h_{i+1}^{j+1}-h_i^{j+1}}{\Delta x_i}+\bar{S}_{f_i}^{j+1}+S_{\mathrm{ec}_i}^{j+1}+S_{i_i}^{j+1}\right)+L_i^{j+1}+(W_f\bar{B})_i^{j+1}\right]+(1-\theta)$$
$$(51)$$

$$\left[\sigma\left(\frac{(\beta Q^2/A)_{i+1}^{j}-(\beta Q^2/A)_i^{j}}{\Delta x_i}\right)+g\bar{A}_i^{j}\left(\frac{h_{i+1}^{j}-h_i^{j}}{\Delta x_i}+\bar{S}_{f_i}^{j}+S_{\mathrm{ec}_i}^{j}+S_{i_i}^{j}\right)\right.$$
$$\left.+L_i^j+(W_f\bar{B})_i^j\right]=0$$

where

$$\bar{A}_i=(A_i+A_{i+1})/2 \quad (52)$$
$$\bar{S}_{f_i}=n^2\bar{\theta}_i|\theta_i|/(\mu^2A_i^2\bar{R}_i^{4/3})=\bar{\theta}_i|\theta_i|/\bar{K}_i \quad (53)$$
$$\bar{Q}_i=(Q_i+Q_{i+1})/2 \quad (54)$$
$$\bar{R}_i=\bar{A}_i/\bar{B}_i \quad (55)$$
$$\bar{B}_i=(B_i+B_{i+1})/2 \quad (56)$$
$$\bar{K}_i=(K_i+K_{i+1})/2 \quad (57)$$

The terms L and W_fB are defined in Eq. (30); terms associated with the jth time line are known from initial conditions or previous time-step computations; and μ in Eq. (53) is defined in Eq. (31). The Δx distance between cross sections is measured along the peak flow path through the waterway.

Solution Procedure

The flow equations are expressed in finite-difference form for all Δx_i reaches between the first and last (Nth) cross section ($i = 1, 2, \ldots, N$) along the channel/floodplain and then solved simultaneously for the unknowns (Q and h) at each cross section. In essence, the solution technique determines the unknown quantities (Q and h at all specified cross sections along the waterway) at various times into the future; the solution is advanced from one time to a future time over a finite time interval (time step) of magnitude Δt. Thus, applying Eqs. (50) and (51) recursively to each of the $(N - 1)$ rectangular grids in Figure 1 between the upstream and downstream boundaries, a total of $(2N - 2)$ equations with $2N$ unknowns are formulated. Then, prescribed boundary conditions for subcritical flow [Froude number less than unity, i.e., $\mathrm{Fr} = Q/(A\sqrt{gD}) < 1$], one at the upstream boundary and one at the downstream boundary, provide the two additional and necessary equations required for the system to be determinate. Since disturbances can propagate only in the downstream direction in supercritical flow ($\mathrm{Fr} > 1$), two upstream boundary conditions and no downstream boundary condition are required for the system to be determinate when the flow is supercritical throughout the routing reach. The boundary conditions are described later. Due to the nonlinearity of Eqs. (50) and (51) with respect to Q and h, an iterative, highly efficient quadratic solution technique such as the Newton–Raphson method is frequently used. Other solution techniques linearize Eqs. (50) and (51) via a Taylor series expansion or other means. Convergence of the iterative technique is attained when the difference between successive solutions for each unknown is less than a relatively small prescribed tolerance. Convergence for each unknown at all cross sections is usually attained within about one to five iterations with the majority of solutions obtained within two iterations. A more complete description of the solution method may be found elsewhere (Fread, 1985).

The solution of $2N \times 2N$ simultaneous equations requires an efficient matrix technique for the implicit method to be feasible. One such procedure requiring $38N$ computational operations $(+, -, *, /)$ is a compact, penta-diagonal Gaussian elimination method (Fread, 1971, 1985) that makes use of the banded structure of the coefficient matrix of the system of equations. This is essentially the same as the double sweep elimination method (Liggett and Cunge, 1975; Cunge et al., 1980).

When flow is everywhere and at all times supercritical, the solution technique previously described can be somewhat simplified. Two boundary conditions are required at the upstream boundary and none at the downstream boundary since flow disturbances cannot propagate upstream in supercritical flow. The unknown h and Q at the most upstream cross section are determined from the two boundary equations. Then, cascading from upstream to downstream, Eqs. (50) and (51) are solved for the two unknowns (h_{i+1} and Q_{i+1}) at each cross section by using Newton–Raphson iteration applied recursively to the two nonlinear equations, with $\sigma = 0$ in Eq. (51).

Initial Conditions

Values of water-surface elevation (h) and discharge (Q) for each cross section must be specified initially at time $t = 0$ to obtain solutions to the Saint-Venant equations. Initial conditions may be obtained from any of the following: (a) observations at gaging stations and using interpolated values between gaging stations for intermediate cross sections in large rivers; (b) computed values from a previous unsteady-flow solution (used in real-time flood forecasting); and (c) computed values from a steady-flow backwater solution. The backwater method is most commonly used in which the steady discharge at each cross section is determined by:

$$Q_{i+1} = Q_i + q_i \Delta x_i \qquad i = 1, 2, 3, \ldots, N-1 \tag{58}$$

in which Q_1 is the assumed steady flow at the upstream boundary at time $t = 0$, and q_i is the known average lateral inflow or outflow along each Δx reach at $t = 0$. The water-surface elevations (h_i) are computed according to the following steady-flow simplification of the momentum equation (30):

$$(Q^2/A)_{i+1} - (Q^2/A)_i + g\bar{A}_i(h_{i+1} - h_i + \Delta x_i \bar{S}_{f_i}) = 0 \tag{59}$$

in which $\bar{A}$ and $\bar{S}_f$ are defined by Eqs. (52) and (53), respectively. The computations proceed in the upstream direction ($i = N-1, \ldots 3, 2, 1$) for subcritical flow (they must proceed in the downstream direction for supercritical flow). The starting water-surface elevation (h_N) can be specified or obtained from the appropriate downstream boundary condition for the discharge (Q_N) obtained via Eq. (58). The Newton–Raphson iterative solution method (Fread and Harbaugh, 1971) for a single equation and/or a simple, less efficient, but more stable bisection iterative technique can be applied to Eq. (59) to obtain h_i. Due to friction, small errors in the initial conditions will dampen-out after several computational time steps during the solution of the Saint-Venant equations.

Upstream Boundary

Values for the unknowns at external boundaries (the upstream and downstream extremities of the routing reach) of the channel/floodplain, must be specified to obtain solutions to the Saint-Venant equations. In fact, in most unsteady-flow applications, the unsteady disturbance is introduced at one or both of the external boundaries.

A specified discharge time series (hydrograph) of inflow to the most upstream cross section is used as the upstream boundary condition. The hydrograph should not be affected by downstream flow conditions. This hydrograph may be obtained from the following: (1) historical observations, (2) assumed design hydrograph, or (3) a runoff hydrograph from specified rainfall–runoff model using calibrated or estimated model parameters. The upstream boundary is expressed mathematically as follows:

$$Q_1^{j+1} - Q(t) = 0 \tag{60}$$

in which $Q(t)$ is the specified discharge time series and the subscript indicates the discharge at the first cross section, i.e., the upstream boundary. Equation (60) is used for the upstream boundary if dynamic routing (based on the discretized Saint-Venant equations) commences at this location. However, if the most upstream cross section represents the inlet to an upstream reservoir, a simple routing procedure (reservoir level-pool routing) can be used rather than the considerably more complex dynamic routing if (1) the reservoir is not excessively long and (2) the inflow hydrograph $Q(t)$ is not rapidly changing with time. Level-pool routing errors (Fread, 1992), with a magnitude of less than about 5%, can usually be tolerated.

Downstream Boundary

For subcritical flow, a specified discharge or water-surface elevation time series or a tabular relation between discharge and water-surface elevation (single-valued rating curve) can be used as the downstream boundary condition.

Another downstream boundary condition can be a computed loop-rating curve based on the Manning equation, i.e.,

$$Q_N^{j+1} - \mu/nA_N^{j+1}(R_N^{j+1})^{2/3}(S_{f_N}^j)^{1/2} = 0 \tag{61}$$

The loop is produced by using the friction slope (S_f) rather than the channel bottom slope (S_0) in the Manning equation. The friction slope exceeds the bottom slope during the rising limb of the hydrograph while the reverse is true for the recession limb. The friction slope (S_f) is approximated by using Eq. (30) where L and W_f are assumed to be zero while s_m and β are assumed to be unity (Fread, 1985, 1988, 1992), i.e.,

$$S_{f_N}^j = (Q_N^j - Q_N^{j-1})/(gA_N^j \Delta t^j) - [(Q^2/A)_N^j - (Q^2/A)_{N-1}^j]/(gA_N^j \Delta x_{N-1}) \\ -(h_N^j - h_{N-1}^j)/\Delta x_{N-1} \tag{62}$$

The loop-rating boundary equation allows the unsteady wave to pass the downstream boundary with minimal disturbance by the boundary itself, which is desirable when the routing is terminated at an arbitrary location along the channel/floodplain and not at a location of actual flow control such as a dam or waterfall, or where the flow is affected by downstream backwater conditions produced by tidal action, reservoirs, or tributary inflow.

When the downstream boundary is a stage/discharge relation (rating curve), the flow at the boundary should not be otherwise affected by flow conditions farther downstream. Although there are often some minor effects due to the presence of cross-sectional irregularities downstream of the chosen boundary location, these usually can be neglected unless the irregularity is so pronounced as to cause significant backwater or drawdown effects. Reservoirs, major tributaries, or tidal effects located below the downstream boundary, which cause backwater effects at the boundary, should be avoided. When either of these situations is unavoidable, the routing reach should be extended downstream to the dam in the case of the reservoir or to a location downstream of where the major tributary enters. Sometimes the routing reach may be shortened by moving the downstream boundary to a location farther upstream where backwater effects are negligible.

Internal Boundaries

Often along the channel/floodplain, there are locations such as a dam, bridge, or waterfall (short rapids) where the flow is rapidly varied in space rather than gradually varied. At such locations (internal boundaries), the Saint-Venant equations are not applicable since gradually varied flow is a necessary condition for their derivation. Empirical water elevation-discharge relations such as weir flow are utilized for simulating rapidly varying flow. At internal boundaries, cross sections are specified for the upstream and downstream extremities of the section where rapidly varying flow occurs. The Δx reach containing an internal boundary requires two internal boundary equations since, as with any other Δx reach, two equations equivalent to the Saint-Venant equations are required. One of the required internal boundary equations represents conservation of mass with negligible time-dependent storage, i.e.,

$$Q_i^{j+1} - Q_{i+1}^{j+1} = 0 \tag{63}$$

Dam. The second equation is usually an empirical, rapidly varied flow relation. If the internal boundary represents a dam, the following equation can be used:

$$Q_i^{j+1} - (Q_s + Q_b)^{j+1} = 0 \tag{64}$$

in which Q_s and Q_b are the spillway and dam-breach flow, respectively. In this way, the flows Q_i and Q_{i+1} and the elevations h_i and h_{i+1} are in balance with the other flows and elevations occurring simultaneously throughout the entire flow system, which may consist of additional downstream dams that are treated as additional internal boundary conditions via Eqs. (63) and (64). In fact, this approach can be used to simulate the progression of a dam-break flood through an unlimited number of reservoirs located sequentially along the valley. The downstream dams may also breach if they are sufficiently overtopped. The spillway flow (Q_s) is computed from the following expression:

$$Q_s = c_s L_s (h_i - h_s)^{1.5} + c_g A_g (h_i - h_g)^{0.5} + c_d L_d (h_i - h_d)^{1.5} + Q_t \tag{65}$$

in which c_s is the uncontrolled spillway discharge coefficient, h_s is the uncontrolled spillway crest, c_g is the gated spillway discharge coefficient, h_g is the centerline elevation of the gated spillway, c_d is the discharge coefficient for flow over the crest of the dam, L_s is the spillway length, and Q_t is a constant outflow term that is head independent or it may be a specified discharge time series. The uncontrolled spillway flow or the gated spillway flow can also be represented as a table of head-discharge values. The gate flow may also be specified as a function of time via a known time series for $A_g(t)$. The breach outflow (Q_b) is computed as broad-crested weir flow (Fread, 1977, 1985, 1988, 1992; Fread and Lewis, 1998), i.e.,

$$Q_b = c_v k_s[3.1 b_i(h_i - h_b)^{1.5} + 2.45z(h_i - h_b)^{2.5}] \tag{66}$$

in which c_v is a small correction for velocity of approach, b_i is the instantaneous breach bottom width, h_i is the elevation of the water surface just upstream of the structure, h_b is the elevation of the breach bottom in which h_b is assumed to be a function of time (t_b) from beginning of the breach formation time (τ), z is the side slope of the breach, and k_s is the submergence correction factor due to the downstream tailwater elevation (h_t), i.e.,

$$k_s = 1.0 \qquad h^* \leq 0.67 \tag{6.7}$$

$$k_s = 1.0 - 27.8(h^* - 0.67)^3 \qquad h^* > 0.67 \tag{68}$$

where

$$h^* = (h_t - h_b)/(h_i - h_b) \tag{69}$$

Using a parametric description of the breach, the instantaneous breach bottom width (b_i) starts at a point at the crest of the dam and enlarges at a linear or nonlinear rate over the failure time (τ) until the terminal bottom width (b) is attained and the breach bottom has eroded to the minimum elevation, h_{bm}. The instantaneous bottom elevation of the breach (h_b) is described as a function of time (t_b) according to the following:

$$h_b = h_d - (h_d - h_{\mathrm{bm}})(t_b/\tau)^{\rho} \qquad 0 \leq t_b \leq \tau \tag{70}$$

in which h_d is the elevation of the top of the dam, h_{bm} is the final elevation of the breach bottom, which is usually, but not necessarily, the bottom of the reservoir or outlet channel bottom, t_b is the time since beginning of breach formation, and ρ is the parameter specifying the degree of nonlinearity, e.g., $\rho = 1$ is a linear formation rate, while $\rho = 2$ is a nonlinear quadratic rate; the range for ρ is $1 \leq \rho \leq 4$, with the linear rate usually assumed. The interval of time (τ) required for the breach to form is given by $\tau = 0.3V_r^{0.53}/H_d^{0.9}$ in which $H_d = h_b - h_{\mathrm{bm}}$, V_r is the reservoir volume (acre-ft) from empirical data by Froehlich (1987); the standard error of estimate for τ

is ±0.9h or ±74% of τ (Fread, 1988, 1995). The instantaneous bottom width (b_i) of the breach is given by the following:

$$b_i = b(t_b/\tau)^{\rho} \qquad 0 \le t_b \le \tau \tag{71}$$

in which b is the final width of the breach bottom given by $b = \bar{b} - zH_d$ and $\bar{b} = 9.5k_0(V_r H_d)^{0.25}$ from empirical data by Froehlich (1987) in which $k_0 = 0.7$ for piping and $k_0 = 1.0$ for overtopping; the standard error of estimate for $\bar{b}$ is ±82 ft or ±56% of $\bar{b}$ (Fread, 1988, 1995).

When simulating a dam failure, the actual breach formation can commence when the reservoir water-surface elevation (h) exceeds a user-specified value, h_f. This feature permits the simulation of an overtopping of a dam in which the breach does not form until a sufficient amount of water has passed over the crest of the dam to have eroded away the downstream face of the dam.

If the breach is formed by piping, Eq. (66) is replaced by an orifice equation:

$$Q_b = 4.8A_p(h_i - h_p)^{1/2} \tag{72}$$

where

$$A_p = [b_i + z(h_p - h_b)](h_p - h_b) \tag{73}$$

in which h_p is the specified centerline elevation of the pipe. Each of the terms in Eq. (65) except Q_t may be modified by a submergence correction factor similar to k_s that can be computed by Eqs. (67) to (69), but in Eq. (69) h_b is replaced by h_s, h_g, and h_d, respectively.

Bridge. If the internal boundary represents highway/railway bridges together with their earthen embankments that cross the floodplain, Eqs. (63) and (64) can still be used although Q_s in Eq. (65) is computed by the following contracted bridge flow expression:

$$Q_s = C_b\sqrt{g}A_{i+1}(h_i - h_{i+1})^{0.5} + C_d k_s(h_i - h_c)^{1.5} \tag{74}$$

in which C_b is a coefficient of bridge flow (Chow, 1959), C_d is the coefficient of flow over the crest of the road embankment, h_c is the crest elevation of the embankment, and k_s is similar to Eqs. (67) to (69) except h_b is replaced by h_c. A breach of the embankment is treated the same as with dams.

Levee Overtopping/Floodplain Interactions

Flows that overtop levees located along either or both sides of a main-stem river and/or its tributaries can be treated as lateral flow (q) in Eqs. (29) and (30) where the lateral flow diverted over the levee is computed as broad-crested weir flow. This overtopping flow is corrected for submergence effects if the floodplain water-surface elevation sufficiently exceeds the levee crest elevation. After the flood peak passes, the overtopping flow may reverse its direction when the floodplain water-surface elevation exceeds the river water-surface elevation, thus allowing flow to return to the river. The overtopping broad-crested weir flow is computed according to the following:

$$q = -c_l k_s (h - h_c)^{3/2} \tag{75}$$

where k_s, the submergence correction factor, is computed as in Eqs. (67) to (69) except $h^* = (h_{\mathrm{fp}} - h_c)/(h - h_c)$, in which c_l is the weir discharge coefficient, h_c is the levee-crest elevation, h is the water-surface elevation of the river, and h_{fp} is the water-surface elevation of the floodplain. Flow in the floodplain can affect overtopping flows via the submergence correction factor. Flow may also pass from the waterway to the floodplain through a time-dependent crevasse (breach) in the levee via a breach-flow equation similar to Eq. (66). The floodplain, which is separated from the principal routing channel (river) by the levee, may be treated as (a) a dead-storage area (A_0) in the Saint-Venant equations, in which case Eq. (75) is not relevant, (b) a tributary that receives its inflow as lateral flows (the flows from the river that overtop the levee crest), which are simultaneously dynamically routed along the floodplain, and (c) the flows and water-surface elevations can be computed by using a level-pool routing method particularly if the floodplain is divided into compartments by levees (dikes) or elevated roadways located somewhat perpendicular to the river levee(s).

Supercritical/Subcritical Mixed Flows

Flow can change with either time or distance along the routing reach from supercritical to subcritical while passing through critical flow, or conversely. This "mixed flow" requires special treatment to prevent numerical instabilities in the solution of the Saint-Venant equations. Such a treatment for mixed flows (Fread et al., 1996) is to provide a "local partial inertia" filter, i.e.,

$$\sigma = [1 - (Fr/Fr_c)^m] \tag{76}$$

which multiplies the first two (inertia) terms in the momentum equations [(30) and (51)]. Fr is the Froude number of the flow in any ith Δx reach, and the exponent (m) varies from 1 to 10, with 5 usually preferred, and $0.85 < Fr_c < 0.95$ is the specified range for Fr_c. The filter takes on a value of zero when $\mathrm{Fr} \geq 1$. The local partial inertia filter (σ) avoids numerical difficulties associated with mixed flows while introducing negligible errors, less than about 1 to 2% for all flow conditions.

Flow Through a System of Rivers

A river system consisting of a main-stem river and one or more tributaries is efficiently solved using an iterative relaxation method (Fread, 1973, 1985) in which the flow at the confluence of the main-stem and tributary is treated as the lateral inflow/outflow (q) in Eqs. (29) and (30). This algorithm was extended so as to treat a dendritic system of waterways having nth-order tributaries (Lewis et al., 1996) and further extended to treat a river system that has any bifurcations such as islands, along with or without nth-order tributaries (Jin, et al., 2000). Also, a less versatile direct solution technique can be used (Fread, 1985), wherein three internal boundary equations conserve mass and momentum at each bifurcation or junction confluence. The resulting system of algebraic equations uses a special sparse matrix Gaussian elimination technique for an efficient solution (Fread, 1983).

REFERENCES

Abbott, M. B., *An Introduction to the Method of Characteristics*, American Elsevier, New York, 1966.

Amein, M., and C. S. Fang, Implicit flood routing in natural channels, *J. Hydraul. Div., ASCE, 96*, 2481–2500, 1970.

Arcement, G. J., Jr., and V. R. Schneider, *Guide for Selecting Manning's Roughness Coefficients for Natural Channels and Flood Plains*, Report No. RHWA-TS-84-204, U.S. Geological Survey for Federal Highway Administration, National Technical Information Service, PB84-242585, 1984.

Baltzer, R. A., and C. Lai, Computer simulation of unsteady flow in waterways, *J. Hydraul. Div. ASCE, 94*, (HY4), 1083–1117, 1968.

Barkow, R. L., *UNET One-Dimensional Unsteady Flow Through a Full Network of Open Channels, Users Manual*, Hydrologic Engineering Center, U.S. Army Corps of Engineers, Davis, CA, 1990.

Barnes, Jr., H. H., *Roughness Characteristics of Natural Channels*, Geological Survey Water-Supply Paper 1849, U.S. Government Printing Office, Washington, DC, 1967.

Boussinesq, J., Theory of the liquid intumesence, called a solitary wave or a wave of translation, propagated in a channel of rectangular cross section, *Comp. Rend. Acad. Sci., 72*, 755–759, 1871.

Chow, V. T., *Open-Channel Hydraulics*, McGraw-Hill, New York, 1959.

Chow, V. T., *Handbook of Applied Hydrology*, Sections 7 and 25-II, McGraw-Hill, New York, 1964.

Chow, V. T., D. R. Maidment, and L. W. Mays, *Applied Hydrology*, McGraw-Hill, New York, 1988.

Cunge, J. A., On the subject of a flood propagation computation method (Muskingum method), *J. Hydraul. Res., 7*, (2), 205–230, 1969.

Cunge, J. A., F. M. Holly, Jr., and A. Verway, *Practical Aspects of Computational River Hydraulics*, Pitman, Boston, MA, 1980.

DeLong, L. L., Extension of the unsteady one-dimensional open-channel flow equations for flow simulation in meandering channels with flood plains, in *Selected Papers in Hydrologic Science*, U.S. Geological Survey Water Supply Paper 2220, 1986, pp. 101–105.

DeLong, L. L. Mass conservation: 1-D open channel flow equations, *J. Hydraul. Div., 115*(2), 263–268, 1989.

Dooge, J. C. I., W. G. Strupczewski, and J. J. Napiorkowski, Hydrodynamic derivation of storage parameters of the Muskingum model, *J. Hydrol.*, (54), 371–387, 1982.

Dronkers, J. J., Tidal computations for rivers, coastal areas, and seas, *J. Hydraul. Div., ASCE, 95*(HY1), 29–77, 1969.

Fread, D. L., Discussion of implicit flood routing in natural channels, by M. Amein and C. S. Fang, *J. Hydraul. Div. ASCE, 97*(HY7), 1156–1159, 1971.

Fread, D. L., Technique for implicit dynamic routing in rivers with tributaries, *Water Resour. Res.*, *9*(4), 918–926, 1973.

Fread, D. L., *Numerical Properties of Implicit Four-Point Finite Difference Equations of Unsteady Flow*, HRL-45, NOAA Technical Memo NWS HYDRO-18, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, 1974.

Fread, D. L., The development and testing of a dam-break flood forecasting model, in *Proceedings of Dam-Break Flood Modeling Workshop*, U.S. Water Resources Council, Washington, DC, 1977, 164–197.

Fread, D. L., NWS operational dynamic wave model, in *Verification of Mathematical and Physical Models, Proceedings of 26th Annual Hydr. Div. Specialty Conf.*, American Society of Chemical Engineers, College Park, MD, 1978, pp. 455–464.

Fread, D. L., Computational extensions to implicit routing models, in *Proceedings of the Conference on Frontiers in Hydraulic Engineering*, MIT, Cambridge, MA, 1983, pp. 343–348.

Fread, D. L., Channel routing, in M. G. Anderson and T. P. Burt (Eds.), *Hydrological Forecasting*, Wiley, New York, 1985, pp. 437–503.

Fread, D. L., *The NWS DAMBRK Model: Theoretical Background/User Documentation*, HRL-256, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, 1988.

Fread, D. L., Flood routing and the Manning n, in B. C. Yen, (Ed.), *Proceedings of the International Conference for Centennial of Manning's Formula and Kuichling's Rational Formula*, Charlottesville, VA, 1989, 699–708.

Fread, D. L., Flow routing, in D. Maidment (Ed.), *Handbook of Hydrology* McGraw-Hill, New York, 1992, pp. 10.1–10.36.

Fread, D. L., Selection of Δx and Δt computational steps for four-point implicit non-linear dynamic routing models, in *Proceedings, National Hydraulic Engineering Conference*, American Society of Chemical Engineers, San Francisco, CA, 1993.

Fread, D. L., Dam-breach floods, in V. J. Singh (Ed.), *Hydrology of Disasters*, Kluwer Academic, Boston, 1995, pp. 85–126.

Fread, D. L., and T. E. Harbaugh, Open channel profiles by Newton iteration technique, *J. Hydrol.*, *13*, 79–80, 1971.

Fread, D. L. and J. M. Lewis, NWS FLDWAV Model: Theoretical Description/User Documentation, HAL-406, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, Nov., 1998.

Fread, D. L., M. Jin, and J. M. Lewis, An LPI numerical implicit solution for unsteady mixed-flow simulation, in *Proceedings, North American Water and Environment Congress '96, American Society of Chemical Engineers*, Anaheim, CA, 1996.

Froehlich, D. C., Embankment-dam breach parameters, in *Proceedings of the 1987 National Conference on Hydraulic Engineering*, American Society of Chemical Engineers, New York, 1987, pp. 570–575.

Garrison, J. M., J. P. Granju, and J. T. Price, Unsteady flow simulation in rivers and reservoirs, *J. Hydraul. Div. ASCE*, *95*(HY5), 1559–1576, 1969.

Gray, W. G., G. F. Pinder, and C. A. Brebbia, *Finite Elements in Water Resources*, Pentech Press, London, 1977.

Henderson, F. M., *Open Channel Flow*, Macmillan, New York, 1966, pp. 285–287.

Hydrologic Engineering Center, *HEC-1 Flood Hydrograph Package—Users Manual*, U.S. Army Corps of Engineers, Davis, CA, 1981.

Isaacson, E., J. J. Stoker, and A. Troesch, *Numerical Solution of Flood Prediction and River Regulation Problems*, Report II/III, No. IMM-NYU-205/235, New York University Institute of Mathematics and Science, New York, 1954, 1956.

Jarrett, R. D., Hydraulics of high-gradient streams, *J. Hydraul. Div. ASCE*, *110*(HY11), 1519–1539, 1984.

Jin, M., and D. L. Fread, One-dimensional routing of mud/debris flows using NWS FLDWAV model, in *Proceedings, First International Conference on Debris-Flow Hazards Mitigation: Mechanics, Prediction, and Assessment*, American Society of Chemical Engineers, New York, 1997.

Jin, M., D. Fread, and J. Sylvestre, Channel routing in river networks using NWS FLDWAV model, in 2000 Joint Conference on Water Resources Engineering and Water Resources Planning and Mangement, ASCE Proceedings CD ROM.

Jones, S. B., Choice of space and time steps in the Muskingum-Cunge flood routing method, *Proc. Inst. Civ. Eng.*, Part 2, No. 71, 759–772, 1981.

Laplace, P. S., Recherches sur quelques points due systeme du monde, [Researches on some points of world system], in *Memoirs*, Vol. 9, Acad. Sci., Paris, 1776.

Lewis, J. M., D. L. Fread, and M. Jin, An extended relation technique for modeling unsteady flows in channel networks using the NWS FLDWAV model, in *Proceedings, North American Water and Environment Congress '96*, American Society of Chemical Engineers, Anaheim, CA, 1996.

Liggett, J. A., Basic equations of unsteady flow, in K. Mahmood and V. Yevjevich, (Eds.), *Unsteady Flow in Open Channels*, Water Resources, Fort Collins, CO, 1975, pp. 29–62.

Liggett, J. A., and J. A. Cunge, Numerical methods of solution of the unsteady flow equations, in K. Mahmood and V. Yevjevich (Eds.), *Unsteady Flow in Open Channels*, Vol. I, Water Resources, Fort Collins, CO, 1975, pp. 89–182.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus, *Hydrology for Engineers*, McGraw-Hill, New York, 1986, pp. 502–530.

Manning, R., On the flow of water in open channels and pipes, *Trans. Inst. Civil Eng. Ireland*, *20*, 161–195, 1889.

McCarthy, G. T., The unit hydrograph and flood routing, in *Conf. of the North Atlantic Div.*, U.S. Corps of Engineers, New London, CT, 1938.

Miller, W. A., and J. A. Cunge, Simplified equations of unsteady flow, in K. Mahmood and V. Yevjevih (Eds.), *Unsteady Flow in Open Channels*, Vol. I Water Resources, Fort Collins, CO, 1975, pp. 183–257.

Newton, Sir I., Propositions, Book 2, in *Principia*, Royal Society, London, 1687, pp. 44–46.

O'Brien, J. S., and P. Julien, Physical properties and mechanics of hyper-concentrated sediment flows, in D. S. Bowles (Ed.), *Delineation of Landslide, Flash Flood, and Debris Flow*

Hazards in Utah, General Series UWRL/G-85/03, Utah State University, Utah Water Research Laboratory, Logan, UT, 1984, pp. 260–279.

Poisson, S. D., Memoir on the theory of waves, in *Memoirs*, Vol. 1, Acad. Sci., Paris, 1816, pp. 71–186.

Ponce, V. M., and V. Yevjevich, V. Muskingum-Cunge method with variable parameters, *J. Hydraul. Div. ASCE*, *104*(HY12), 1663–1667, 1978.

Preissmann, A., Propagation of translatory waves in channels and rivers, in *Proc. First Congress of French Assoc. for Computation*, Grenoble, France, 1961, pp. 433–442.

Rajar, R., Mathematical simultion of dam-break flow, *J. Hydraul. Div. ASCE*, *104*(HY7), 1011–1026, 1978.

Saint-Venant, Barré de, Theory of unsteady water flow, with application to river floods and to propagation of tides in river channels, in *Computes rendus*, Vol. 73, Acad. Sci., Paris, France, 1871, 148–154, 237–240. (Translated into English by U.S. Corps of Engineers, No. 49-g, Waterways Experiment Station, Vicksburg, MS, 1949.)

Samuels, P. G., *Models of Open Channel Flow Using Preissmann's Scheme*, Cambridge University Press, Cambridge, 1985, pp. 91–102.

Sayed, I., and D. C. Howard, Application of dynamic backwater modeling to Mactaquac headpond—Saint John River, N.B., in *Proceedings of 6th Canadian Hydrotechnical Conference*, Canadian Society for Civil Engineering, 1983, pp. 203–220.

Schaffranek, R. W., *Flow Model for Open Channel Reach or Network*, Professional Paper No. 1384, U.S. Geological Survey, 1987.

Singh, V. P., and R. C. McCann, Some notes on Muskingum method of flood routing, *J. Hydrol.*, *48*(3), 343–361, 1980.

Stoker, J. J., *Numerical Solution of Flood Prediction and River Regulation Problems; Derivation of Basic Theory and Formulation of Numerical Methods of Attack*, Report I, No. IMM-NYU-200, New York University Institute of Mathematical Science, New York, 1953.

Stoker, J., *Water Waves*, Interscience, New York, 1957, pp. 452–455.

Streeter, V. L., and E. B. Wylie, *Hydraulic Transients*, McGraw-Hill, New York, 1967, pp. 239–259.

Strelkoff, T., The one-dimensional equations of open-channel flow, *J. Hydraul. Div., ASCE*, *95*(HY3), 861–874, 1969.

Strelkoff, T., Numerical solution of Saint-Venant equations, *J. Hydraul. Div. ASCE*, *96*(HY1), 223–252, 1970.

Strelkoff, T., and N. D. Katopodes, Border irrigation hydraulics with zero inertia, *J. Irrig./Drain. Div. ASCE*, *103*, 325–342, 1977.

Strupczewski, W., and Z. Kundzewicz, Translatory characteristics of the Muskingum method of flood routing—a comment, *J. Hydrol.*, *98*, 363–368, 1980.

Viessman, Jr., W., J. W. Knapp, G. L. Lewis, and T. E. Harbaugh, *Introduction to Hydrology*, 2nd ed., Intext Educational Publishers, New York, 1977.

CHAPTER 31

HYDROLOGIC MODELING FOR RUNOFF FORECASTING

HOSHIN GUPTA

1 INTRODUCTION

The problem of forecasting streamflow levels given precipitation data has received the time and attention of a great many hydrologists. Models developed for this purpose have ranged from simple to extremely complex. The simplest ones are based on input–output regression-type relationships, while the most complex ones attempt to represent the detailed water and energy balance physics occurring in the watershed. The complex models are motivated largely by experimental evidence that the subwatershed-scale components of the rainfall–runoff process are strongly nonlinear, time variable, and spatially distributed. However, the processes of aggregation, attenuation, loss, and delay tend to result in an overall watershed response that is far less complex than the point-scale behavior. The effects of subwatershed-scale variability tend to be smoothed and poorly observable (to varying degrees) in the overall watershed-scale response. Thus, while remarkable progress has been made in understanding the physics of how precipitated water moves once it reaches the ground, the level of model complexity required to provide accurate runoff forecasts for any chosen watershed remains unclear. Even less clear is how this complexity varies with climatology, watershed size, and geologic and physiographic characteristics of the landscape.

2 MODELING AND COMPLEXITY

In the absence of such clarity, a wide variety of hydrologic models have found their way into the literature (Singh, 1995). The essential difference in these models is the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

manner in which the underlying processes that transform precipitation into streamflow are conceptualized. The more complex models have been motivated by the scientific pursuit of knowledge and are based on painstaking research into the physics of subwatershed-scale hydrologic processes. Such models attempt, in particular, to account for the spatially and temporally varying nature of watershed inputs (precipitation, solar radiation, etc.), losses (evapotranspiration), and characteristics (topography, permeability, vegetation, etc.). We shall refer to this modeling approach as "physics based." Perhaps the most well-known exponent of this approach is the Systéme Hydrologique Européen (SHE) model (Abbott, 1986). More recent developments are the soil–vegetation–atmosphere–transfer schemes (SVATS) used for climate studies, such as Biosphere Atmosphere Transfer Scheme (BATS) (Dickinson, 1993), Simple Biosphere Model 2 (SiB2) (Randall, 1996), and Variable Infilltration Capacity 2-Layer Model (VIC-2L) (Liang, 1994).

At the other end of the spectrum, the simplest models have been motivated by engineering considerations based on a real need to provide quick and accurate forecasts of streamflow levels in the simplest possible way, particularly wherever human interests are at stake (such as flood-prone locations). Such models attempt to establish direct regression-like relationships between the input and output time series; generally, the streamflow value is regressed on values of precipitation and streamflow at previous times. We shall refer to this modeling approach as "systems theoretic." The most popular systems-theoretic methods have been the ARMAX (auto-regressive moving average with exogenous inputs) (Box, 1976; Salas, 1980) and the ANN (artificial neural network) (Hsu, 1995, 1997).

A third category of models, which are of intermediate complexity, are based on attempts to conceptualize the simplified (lumped) watershed-scale behavior resulting from the integrated effect of the subwatershed-scale hydrologic processes. Such models typically use simple linear and nonlinear tank components (reservoirs) to represent the primary soil moisture zones in the watershed and describe the manner in which moisture exchanges among these stores take place. We shall refer to such models as "conceptual." It is important to note that such models are based on *speculative conjecture* as to how best to partition the watershed into components and how to represent the integrated behavior of each component. This, and the fact that conceptual models are relatively simple to program into a computer, has encouraged a great deal of intellectual experimentation, resulting in a proliferation of conceptual models with widely differing structures. At the simple end, we have methods such as the API (antecedent precipitation index) and UHG (unit hydrograph) which, in a simple manner, partition the watershed response into precipitation excess and infiltration (based on an antecedent soil moisture index) and use linear equations to transform the precipitation excess into streamflow forecasts. Models at the intermediate level include the HEC-1 model (U.S. Army Corps of Engineers, 1973, 1985). At the complex level, we have methods such as the Stanford watershed model (SWM) (Crawford, 1966), the Institute of Hydrology Model (IHDM) (Beven, 1987), the Kineros model (Woolhiser, 1990), the Sacramento soil moisture accounting model (SAC-SMA) (Burnash, 1973), and TOPMODEL (Beven, 1979) that have numerous components. Within the United States, the most widely used of these may well be the API, UHG, and SAC-SMA models because they are extensively used by various regional offices of the U.S. National Weather Service for flood forecasting. Such models are currently being built into more general "modeling systems" such as the advanced hydrologic prediction system (AHPS) of the U.S. National Weather System and the modular modeling system (MMS) of the U.S. Geological Survey. These systems allow the user to build up a complete model by selecting the components from libraries containing several alternative conceptual representations.

Finally, the last half-decade has seen the emergence of a subclass of conceptual models that seek to strike a reasonable and parsimonious balance between the three issues of (a) scientific understanding (physics), (b) speculative conjecture about the nature of integrated watershed-scale processes (conceptualization), and (c) the level of model complexity that can actually be supported by the available watershed response data (i.e., the systems-theoretic issues of observability and identifiability). Examples of such models are the IHACRES [see e.g., Jakeman (1990, 1993)] model and the related HyMod under development at the University of Arizona (Boyle, 2000). For want of a better terminology, we follow Wheater (1993) in referring to these as "hybrid" models.

The three mechanisms of scientific understanding, conceptualization, and data-supportable complexity can be likened to the legs of a stool that must be of proper and complementary length so that the sitting surface is balanced and can perform its intended function (Fig. 1). The key issue in selecting an appropriate model is this intended function. A physics-based model such as SHE and some conceptual models such as TOPMODEL and Kineros may be clearly appropriate for detailed watershed modeling and for testing hypotheses about watershed behavior under perturbed conditions. On the other hand, Hsu et al. (1995) have shown that simple ANN-type systems-theoretic models can give one-step-ahead forecasts that are more accurate than those given by conceptual models, while requiring relatively minor computational resources and being quick and easy to build. However, if the intended

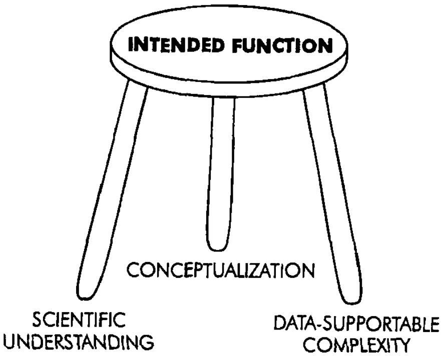


Figure 1 Issues influencing model development and selection.

function is *both* accurate operational streamflow forecasting *as well* as insight into evolving watershed behavior, the emerging evidence suggests that hybrid models such as IHACRES and HyMod, which merge the strengths of the conceptual and systems-theoretic approaches, may prove to be the optimal choice.

3 MODEL PARAMETER ESTIMATION, CALIBRATION, AND EVALUATION

The model selected must be made specific to a watershed by estimating values for its parameters. In the case of physically based models and some conceptual models, approximate values (or ranges of values) for many of the parameters can sometimes be estimated from maps or field measurements. However, because all such models involve conceptualization (simplification and distortion from reality), the parameter estimates obtained in this manner can invariably be improved by calibration to historical input–output data. Certainly, in the case of systems-theoretic models, the only method for inferring structural complexity and parameter values is an automated computer-based identification procedure. Because each of the available systems-theoretic modeling approaches (such as ARMAX and ANN) is generally accompanied by clear procedures for model building and parameter estimation, they will not be described here. The discussion here will focus on parameter estimation for the other three categories of models via a procedure called model calibration.

The model calibration process involves five interrelated components: (a) data set, (b) constraints, (c) measures of closeness, (d) parameter adjustment procedure, and (e) evaluation procedure. Each of these components is discussed in turn.

Data Set

The input–output data set to be used for inferring model parameters must be carefully selected from the historical record to be representative of the behavior of the watershed. Two issues are important here—data quality and data quantity. Data quality has two subissues that must be considered. The first is simply that the data must be checked for accuracy and reliability (i.e., errors in measurement and/or recording). To take a trivial example, if the precipitation records indicate a large storm event but the flow records do not show a response (or vice versa), we might suspect the accuracy of the data. The second subissue is related to data informativeness; i.e., the data must be representative of the important characteristic modes of watershed behavior. For example, if the purpose of the model is flood forecasting, the data must certainly contain several significant storm events. These data will provide information about the parameters related to the partitioning of precipitation into flow components having different recession rates. However, the data must also contain several representative interstorm periods so that information regarding the parameters controlling streamflow recession as well as rates of evaporation loss can be deduced. There have been only a few studies investigating this issue. Gupta (1985) used a theoretical analysis to show that "threshold-type" parameters are best identified when the data are selected to ensure that the model behavior tends to switch across the threshold numerous times; surprisingly, the amount of time spent in each mode of behavior is largely irrelevant. Yapo (1996) studied the reliability of parameter estimates of a conceptual rainfall–runoff model using 40 years of data and clearly demonstrated that the most reliable results are provided by using "wet" years for calibration.

The aforementioned studies also addressed the issue of data quantity (length). Gupta (1985) showed theoretically that a (daily) data set of approximately 3 years length is desirable for model calibration, and that additional amounts of data will provide only marginal gains *unless containing significantly new information.* Yapo (1996) found, however, that for the Leaf River in Mississippi, the SAC-SMA conceptual model requires at least 8 to 10 years of data for reliable calibration results to be obtained, suggesting that the variability of information in a hydrologic data set may extend over approximately a decade.

Having selected the calibration period data set, the next important decision is the selection of an appropriate length for the "buffer" period. A buffer period is a short data segment at the very beginning of the data set for which the measures of closeness (see below) are not computed. The intention is to minimize any potential bias in the calibration procedure caused by uncertain initialization of the model state variables. Because a watershed model tends to average and attenuate inputs, it will also attenuate the impact of initialization errors over time. A buffer period of 90 to 180 days beginning near the end of a long recession and approximately a week or two before the end of the dry season seems to be a good choice.

Constraints

The search for a better (or "best") parameter set is facilitated greatly by specifying upper and lower limits for each of the parameters—this defines the "feasible" region in which an "optimal" parameter set is expected to lie. A useful perspective is to consider this feasible region to be the initial uncertainty in the parameter estimates, based on available prior information. For certain parameters, these upper and lower limits are easily selected based on physical considerations related to the characteristics of the watershed. For parameters such as equation exponents defining the degree of nonlinearity of a transformation, one may only be able to guess at an appropriate range of values. The selection of constraints is rather model dependent and can be quite subjective. However, if the parameter adjustment procedure is sufficiently powerful, the impact of this subjectivity should be minimal.

Measures of Closeness

To select a better (or best) parameter set from the feasible parameter space, we must be able to compare and evaluate, in some manner, the model performance associated with different parameter sets. The indicator of model performance is usually taken to be a comparison between the observed (measured) watershed output time series and the corresponding model simulated quantities. In general, particularly for runoff forecasting models, the indicator representing watershed behavior is the sequence of observed streamflow levels at the watershed outlet, although streamflow level data at gauging stations within the watershed or soil moisture data at specific points within the watershed may sometimes also be available. Other models may have multiple outputs that can serve as indicators of model behavior. In the case of watershed hydrochemical models, we may have data on concentrations of various chemical species, and in the case of watershed scale water-and-energy budget models [e.g., Dickinson (1993) and Schaake (1996)], we may have data on surface soil temperature, emitted short- and long-wave radiation, sensible and latent heat fluxes, etc.

The million-dollar question is: What method should be used to compare the model-simulated streamflow values and the observed streamflow data? A visual comparison of the two time series plotted together on the same graph is intuitively appealing (Fig. 2) but is made difficult by the large number (say n) of time steps at which the simulated streamflow values must be compared to the observed data. If $E_t = \{e_t = s_t - o_t, t = 1, \ldots, n\}$ represents the vector of differences between each simulated flow (s_t) and its corresponding observed data value (o_t), the method of visual comparison will involve adjusting the parameters to simultaneously make each one of the e_t differences as small as possible. Because this approach is subjective, different hydrologists will tend to judge different model-simulated time series (and their associated parameter sets) as being better. Further, while it may be relatively simple to decide that a certain approximate region of the parameter space gives better simulations than some other regions of the parameter space, it can be very difficult to narrow down the choice (see section below on parameter adjustment). The method of visual comparison is also difficult if not impossible to automate.

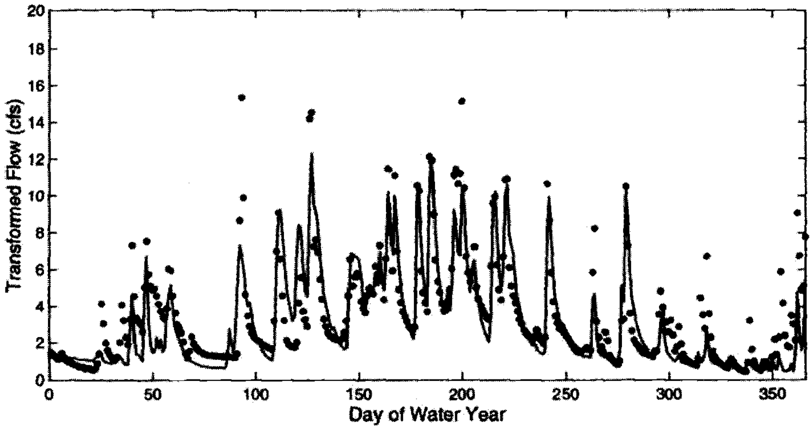


Figure 2 Plot of observed flow (·) vs. model simulated flow (—). Values are in the transformed space (to observe behavior in the full range of flows better), where transformed flow = $[(\text{flow} + 1)\lambda - 1]/\lambda$, and $\lambda = 0.3$.

An alternative to visual comparison is to define a mathematical measure of the "size" of the vector E_t. However, there are an infinite number of ways in which this can be done. The most popular implementation is to compute a scalar measure of the average size of the differences, such as the mean-squared error [MSE = mean(e_t^2, $t = 1, \dots, n$)) or the mean absolute error (MAE = mean($|e_t|$, $t = 1, \dots, n$)]. A number of different such measures, which are commonly called "objective" functions, have been suggested in the literature; Table 1 lists many of the measures used by the U.S. National Weather Service for calibration of its flood forecast models. A related approach is to treat the residuals as though they have stochastic properties and belong to some preassumed probability distribution, usually assumed to be Gaussian. Under this assumption, it is possible to develop maximum-likelihood (ML) measures having theoretical underpinnings; for example, the heteroscedastic maximum-likelihood estimator [HMLE = mean $(w_t^* e_t^2, t = 1, \dots, n)$; $w_t = o_{\text{obs},t}^{2(\lambda-1)} / (\prod_1^n o_{\text{obs},t}^{2(\lambda-1)})^{1/n}$; λ is a parameter to be estimated] criterion developed by Sorooshian (1980) assumes that the residuals are Gaussian, uncorrelated, unbiased, and have

TABLE 1 Objective Functions Used by National Weather Service for Calibration of SAC-SMA Model

Name	Description	Formula	
DRMS	Daily root-mean-squared error	Minimize w.r.t θ	$\sqrt{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]^2}$
TMVOL	Total mean monthly volume-squared error	Minimize w.r.t θ	$\sum_{i-1}^{n\text{month}} \left\{\frac{1}{n\text{day}(i)} \sum_{-1}^{n\text{day}(i)} [s_t - o_t(\theta)]\right\}^2$
ABSERR	Mean absolute error	Minimize w.r.t θ	$\frac{1}{n}\sum_{t-1}^{n} \lvert s_t - o_t(\theta)\rvert$
ABSMAX	Maximum absolute error	Minimize w.r.t θ	$\underset{1\le t\le n}{\text{Max}} \lvert s_t - o_t(\theta)\rvert$
NS	Nash–Sutcliffe measure	Minimize w.r.t θ	$1 - \frac{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]^2}{\frac{1}{n}\sum_{t=1}^{n}(s_t - \bar{s})^2}$
BIAS	Bias (mean daily error)	Minimize w.r.t θ	$\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]$
PDIFF	Peak difference	Minimize w.r.t θ	$\max_{1\le t\le n}\{s_t\} - \max_{1\le t\le n}\{o_t(\theta)\}$
RCOEF	First lag auto-correlation	Minimize w.r.t θ	$\frac{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)][s_{t+1} - o_{t+1}(\theta)]}{\sigma_s \sigma_{0(\theta)}}$
NSC	Number of sign changes	Minimize w.r.t θ	(Count the number of times the sequence of residuals changes sign)

nonhomogenous variance. The advantage of the ML approach is that the validity of the underlying assumptions can be verified by a postcalibration residual analysis. It is important to note that each of these scalar measures defines a different way to gauge the "size" of the error vector, and the minimum value for each will define a simulated streamflow sequence that is "close" to the observed data in a different way. If a certain scalar measure is selected, then it is possible (in principle) to find a single parameter set (or a small region of the feasible space) that minimizes that measure. This makes the model calibration procedure much easier to automate so that the accuracy, speed, and efficiency of a computer can be exploited. Nonetheless, while more efficient than visual comparison, the use of scalar measures is perhaps no less subjective.

Parameter Adjustment Procedure

The parameter adjustment procedure is a directed trial-and-error process by which the parameters are iteratively adjusted to move the model behavior closer to the observed data. The choice of procedure is related to the measure of closeness selected (see above). If the calibration is performed by an expert hydrologist having a great deal of familiarity with the nuances of the model, the method of manual parameter adjustment guided by visual comparison can be extremely effective. However, manual calibration has several drawbacks. First, the procedure requires a great deal of subtlety in evaluating the visual goodness of fit, something that takes time and training to develop. Even to the trained eye, there may appear to exist numerous equally "good" parameter sets that are difficult to distinguish (Beven, 1992; Freer, 1996). Different good parameter sets will appear to match the data well in different ways, and moving from one set to another will trade-off an improvement in matching some parts of the data against deterioration in matching other parts of the data (Gupta, 1998). In practice, the calibration expert can also support the qualitative visual comparison with one or more quantitative scalar measures (e.g., see Table 1). However, this evaluation process still tends to be greatly complicated by the large number of model parameters to be adjusted and their tendency to have interacting and compensating effects on the output. Furthermore, the process can be very time intensive, particularly when the model contains numerous subcomponents and a large number of parameters (e.g., manual calibration of the SAC-SMA model can take several person-days of dedicated effort). These difficulties tend to limit widespread utility of the more complex and sophisticated models.

An alternative to manual parameter adjustment is to use the speed and power of a computer to automatically search the feasible parameter space for "better" solutions. In this approach, the measure of closeness is typically one of the scalar measures of closeness described earlier. A great deal of research has gone into the development of an automatic parameter adjustment procedure that gives satisfactory results while being reliable (effective) and efficient. A satisfactory result is one that gives model simulations similar to those obtained by an expert manual calibration, while resulting in parameter estimates that are conceptually realistic; a reliable procedure is one

TABLE 2 Summary of Five Major Characteristics Complicating the Optimization Problem in CRR Model Calibration

1. Regions of attraction	More than one main convergence region
2. Minor local optima	Many small "pits" in each region
3. Roughness	Rough response surface with discontinuous derivatives
4. Sensitivity	Poor and varying sensitivity of response surface in region of optimum, and nonlinear parameter interaction
5. Shape	Nonconvex response surface with long curved ridges

that consistently provides satisfactory results; and an efficient procedure is one that requires only small amounts of computer time.

The earliest attempts at automatic calibration drew on a class of function optimization techniques called "local search" procedures; examples include the pattern search method (Hooke, 1961), the rotating directions method (Rosenbrock, 1960), the downhill simplex method (Nelder, 1965), and various versions of the Gauss–Newton quadratic approximation method (Luenberger, 1984). It quickly became apparent that such methods were highly unreliable; independent trials of the algorithm initiated from different initial parameter estimates would converge to widely differing solutions. A study by Duan (1993) demonstrated conclusively the reasons for this poor performance; the response surface of the scalar measure being optimized typically has several characteristic properties (see Table 2) that local search

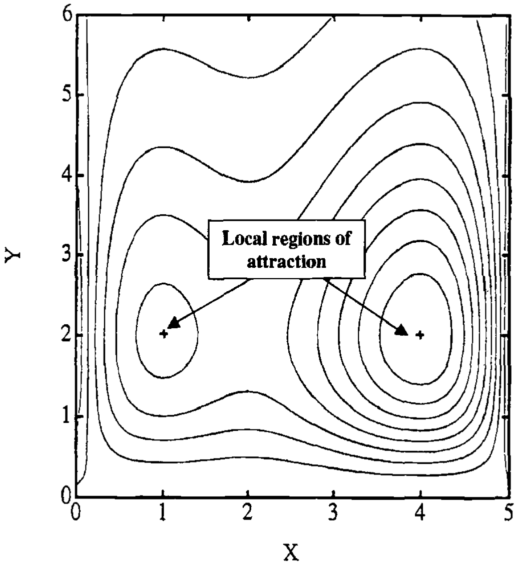


Figure 3 Function response surface showing multiple regions of attraction.

algorithms are not able to handle well. The most important of these are the existence of more than one primary region of attraction (see Fig. 3) as well as large numbers of local optima throughout the feasible space (see Fig. 4). The focus therefore shifted to trying the existing "global search" methods including adaptive random search (Brazil, 1987), the genetic algorithm (Wang, 1991; Tanakamaru, 1995), and the multistart simplex (Duan, 1992; Gan, 1996). The most successful method to date has been the shuffled complex evolution (SCE-UA) method recently developed at the University of Arizona (Duan, 1992, 1994; Sorooshian, 1993), which has proved to be both reliable and relatively efficient (see Fig. 5).

It should be noted that "manual" and "automatic" parameter adjustment approaches have mutually complementary strengths and weaknesses, which suggests the implementation of a hybrid approach that draws on the strengths of each (while minimizing their weaknesses). The strength of the manual approach is its ability, when successful, to provide very satisfying model calibrations because visual comparisons draw on the human ability to perceive patterns that are not easy to detect using numerical techniques. The strength of the automatic approach is that it can very quickly and rapidly find the region(s) of the parameter space that give relatively close matching of the simulated flows and observed data, while manipulating large (even bewildering) numbers of mutually compensating and interacting parameters. The hybrid procedure therefore involves two steps. In the first step, the automatic procedure is used to quickly find several solutions that seem to have similar ability to match the data when measured using one *or more* of the scalar

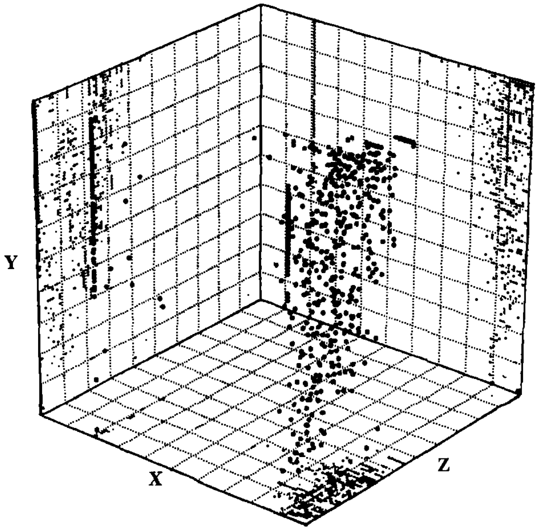


Figure 4 Locations of local optima in three-dimensional parameter subspace; each dot represents optima for a local region.

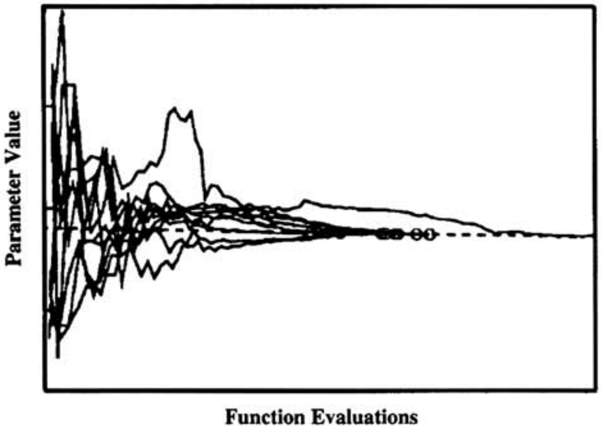


Figure 5 Convergence of model parameter for 10 different trials of the SCE-UA method; dotted line represents the true paramenter value.

numerical measures of closeness described earlier. These then become the starting point for a manual procedure of refinement in which the expertise of the hydrologist can be used to further improve the solution. The computer does what it does better than a human, which is to search through large numbers of options very quickly and reject the unacceptable ones. The human does what humans do better than a computer, which is to use perceptual discrimination to make qualitative distinctions that are difficult to describe mathematically.

Developments of this hybrid approach through the use of multi-objective procedures can be found in the work of Gupta (1998) and Yapo (1998). The multiobjective global optimization strategy, MOCOM (multi-objective complex evolution), is used to identify the set of solutions that provide a "trade-off" in simultaneously minimizing several criteria that measure the goodness of fit of the model to the calibra-

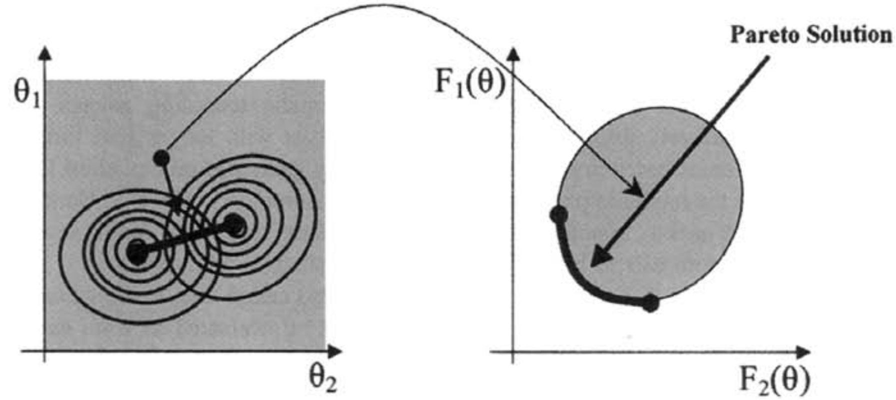


Figure 6 Identification of trade-off solutions using multiobjective optimization strategy.

tion data (Fig. 6). The hydrologist can then use visual means to identify the most perceptually appealing solution(s).

Evaluation Procedure

Once a model has been calibrated by one of the methods outlined above, it is useful to evaluate the result by testing its performance using data not employed for model calibration. For flood forecast models, one might (if possible) select a period of data of comparable length to the calibration period containing several significant storm events. Visual comparison of the observed and simulated outputs for this evaluation period and a check of the goodness-of-fit statistics can reveal any obvious divergence of the model performance from reality. This can also give a reasonable estimate of the approximate forecasting accuracy that can be expected when the model is used for real-time forecasting. Particular attention should be given to any tendencies for the model simulations to be biased at different streamflow levels. Admittedly, the process of model evaluation is somewhat subjective; however, if the simulation performance over the evaluation period is essentially similar to that over the calibration period, the model can then be used for flood forecasting with some understanding of its expected level of performance.

4 FORECASTING AND STATE UPDATING

The calibrated model can be implemented for real-time flood forecasting. The main issue here is that of the "lead time" (duration between time of making the forecast and time of actual occurrence). Clearly, the benefit of a flood forecast lies both in its accuracy *and* in its being available as early as possible before the actual event occurs. With this in mind, the model time step Δt must necessarily be shorter than the time of concentration of the watershed so that the precipitation data available up to time t are used to compute the model-simulated streamflow at time $t + \Delta t$; this is called a one-step-ahead forecast. For small watersheds, this time step Δt may be on the order of only a few hours, while for larger watersheds the time step may be one day or more. To maximize the forecast lead time, it is desirable that the precipitation measurement be either phoned or radioed in to the forecast center within minutes of its occurrence, or even telemetered in by automatic recording gauges and processed immediately through the model. If a forecast with longer lead time is required, it becomes necessary to obtain independently generated precipitation forecasts to feed to the model in place of precipitation measurements. The U.S. National Weather Service uses a "quantitative precipitation forecast" system to enable several time-step-ahead forecasts to be made for many watersheds (Funk, 1991).

A second issue is that of model state updating (also called data assimilation or filtering). Because the accuracy of each forecast can be evaluated as soon as the observed flow for that time step becomes available, this information should, in principle, be useful for adjusting the internal model states to maximize the accuracy of the next forecast. For example, underprediction of the observed flow may indicate that the model storages that represent the wetness of the various watershed components are too dry and should be adjusted accordingly. Kitanidis (1980a, 1980b) rewrote the SAC-SMA model in a state-space form and implemented an extended Kalman filter to enable the model to correct for data errors. Because of the mathematical complexity of reworking a model into a state-space form, state updating has not become widely popular for use with flood forecast models. As the use of the simpler "hybrid" models becomes popular, we can expect to see more exploitation of systems-theoretic methods such as state updating to improve the performance of watershed models.

Finally, it is important to consider the forecasting uncertainty associated with the uncertainty in the model structure and parameter estimates. Some interesting (and somewhat similar) Monte Carlo approaches for representing forecast uncertainty include the generalized likelihood uncertainty estimation (GLUE) method [see, e.g., Beven (1992) and Freer (1996)], the Monte Carlo set membership (MCSM) procedure [see, e.g., Keesman (1990) and van Straten (1991)], and the prediction uncertainty (PU) method [see, e.g., Klepper (1991)]. For example, the GLUE procedure estimates the range of forecast uncertainty by estimating the likelihood associated with the individual forecasts given by different "equifinal" parameter sets in the feasible space.

5 EMERGING DIRECTIONS

It is the thesis of this chapter that the trend in hydrologic modeling for runoff forecasting will be toward a successful marriage of hydrologic science and systems theory, implemented through the coupling of hybrid watershed models with automated procedures for calibration and data assimilation for state updating. We can expect to see clear and rapid progress in all three of these components. Experiments with data from numerous watersheds will help in establishing general guidelines about the level of conceptual detail required to model the dominant watershed responses that are observable in the input–output data. The development of multiobjective calibration procedures (Gupta, 1998; Yapo, 1998) has already begun to merge the strengths of the manual and automated calibration procedures into an effective hybrid calibration method. The simplicity of the hybrid model structures will enable approximate Kalman filtering methods (or other uncertainty estimation methods) to be implemented for improving online forecasts. In addition, radar-based precipitation estimates are already replacing gage-based data and will encourage the development of "distributed" structures but parsimoniously parameterized hybrid watershed models. Finally, because the hybrid modeling approach provides us with a simple functional representation of the watershed, we can also expect progress in understanding how to apply watershed models to ungaged basins.

REFERENCES

Abbott, M. B., J. C. Bathurts, J. A. Cunge, P. E. O'Connell, and J. Rasmussen, An introduction to the European Hydrological System-Systeme Hydrologique Eurorpeen, ASHE@: 2.

Structure of a physically-based, distributed modeling system. *Journal of Hydrology*, *87*, 61–77, 1986.

Beven, K. J., and M. Kirby, A physically based variable contributing area model of basin Hydrology. *Hydrological Sciences Bulletin*, *24*, 43–69, 1979.

Beven, K., A. Calver, and E. Morris, The institute of hydrology distributed model. U.K. Institute of Hydrology Report No. 98, 1987.

Beven, K.J., and A.M. Binley, The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*, 279–298, 1992.

Brazil, L. E., and W. F. Krajweski, Optimization of complex hydrologic models using random search methods. Paper presented at Conference on Engineering Hydrology, Hydraulics Division, American Society of Civil Engineering, Williamsburg, Virginia, Aug. 3–7, 1987.

Box, G. E. P., and G. M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day Inc., San Francisco, CA, 1976.

Boyle, D. P., H. V. Gupta, and S. Sorooshian, Toward improved calibration of hydrological models: combining the strengths of manual and automatic methods, *Water Resources Research*, *36*, 12, 3663–3674.

Burnash, R. J. C., R. L. Ferrell, and R. A. McGuire, *A Generalized Streamflow Simulation System*, Jr. Fed-State River Forecast Center, Sacramento, CA, 1973, 204 pp. 1973.

Crawford, N. H., and R. K. Linsley, Digital Simulation in Hydrology: Stanford Watershed Model IV, Technical Report No. 39, Stanford University Dept. of Civil Engineering, 1966.

Dickinson, R. E., A. Henderson-Sellers, and P. J. Kennedy, Biosphere Atmosphere Transfer Scheme (BATS) Version le as coupled to the NCAR Community Climate Model. NCAR Technical Note, NCAR/TN-387+STR, National Center for Atmospheric Research, Boulder, CO, 1973, 72pp.

Duan, Q., V. K. Gupta, and S. Sorooshian, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*, 4, 1015–1031, 1992.

Duan, Q., V. K. Gupta, and S. Sorooshian, A shuffled complex evolution approach for effective and efficient global minimization, *Journal of Optimization Theory and Applications*, *76*, 3, 501–521, 1993.

Duan, Q., S. Sorooshian, and V. K. Gupta, Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology*, *158*, 265–284, 1994.

Freer, J., A. M. Beven, and B. Ambroise, Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resources Research*, *32*, 7, 2161–2173, 1996.

Funk, T. W., Forecasting techniques utilized by the Forecast Branch of the National-Meteorological-Center during a major convective rainfall event, *Weather and Forecasting*, *6*, 4, 548–564, Dec 1991.

Gan, T. Y. and G. F. Biftu, Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure, *Water Resources Research*, 1996

Gupta, V. K., and S. Sorooshian, The relationship between data and the precision of parameter estimates of hydrologic models, *Journal of Hydrology*, *81*, 57–77, 1985.

Gupta, V. K., S. Sorooshian, and P. O. Yapo, Towards improved calibration of hydrologic models: Multiple and non-commensurable measure of information, *Water Resources Research*, *34*, 4, 751–763, 1998.

Hooke, R., and T. A. Jeeves, Direct search solutions of numerical and statistical problems. *Journal Assoc. Computer Mach.*, *8*, 2, 212–229, 1991.

Hsu, K., H. V. Gupta, and S. Sorooshian, Artificial Neural Network modeling of the rainfall-runoff process, *Water Resources Research*, *31*, 10, 2517–2530, 1995.

Hsu, K., C. Gao, S. Sorooshian, and H. V. Gupta, Precipitation estimation from remotely sensed information using artificial neural networks, *Journal of Applied Meteorology*, *36*, 9, 1176–1190, September, 1997.

Jakeman, A. J., I. G. Littlewood, and P. G. Whitehead, Computation of the instantaneous unit-hydrograph and identifiable component flows with application to 2 small upland catchments, *Journal of Hydrology*, *117*, 1-4, 275–300, September, 1990.

Jakeman, A., and G. Hornberger, How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*, 8, 2637–2649, 1993.

Keesman, K. J., Set theoretic parameter estimation using random scanning and principal component analysis, *Mathematical Computation and Simulation*, 1990.

Kitanidis, P. K., and R. L. Bras, Adaptive filtering through detection of isolated transient errors in rainfall-runoff models, *Water Resources Research*, *16*, 4, 740–748, 1980a.

Kitanidis, P. K., and R. L. Bras, Real-time forecasting with a conceptual hydrological model: 1. Analysis of uncertainty, *Water Resources Research*, *16*, 6, 1025–1033, 1980b.

Klepper, O., H. Scholten, and J. P. G. van de Kamer, Prediction uncertainty in an ecological model of the Oosterschelde Estuary, *Journal of Forecasting*, *10*, 191–209, 1991.

Luenberger, D. G., Introduction to linear and nonlinear programming. Addison-Wesley, Menlo Park, CA, 1984.

Liang, X., and P. D. Lettenmaier, A simple hydrologically based model of land surface water and energy fluxes for general circulation models, *J. of Geophys. Res*, *99*, D7, 14, 415–14, 428, July, 1994.

Nelder, J. A., and R. Mead, A simplex method for function minimization, *Computer Journal*, *7*, 4, 308–313, 1965.

Randall, D. A., S. A. Dazlich, C. Zhang, A. S. Denning, P. J. Sellers, C. J. Tucker, L. Bounoua, S. O. Los, C. O. Justice, and I. Fung, A revised land surface parameterization (SiB2) for GCMs: 3. The greening of the Colorado State University general circulation model, *Journal of Climate*, *9*, 4, 738–763, April, 1996.

Rosenbrock, H. H., An automatic method of finding the greatest or least value of a function, *Computer Journal*, *3*, 175–184, 1960.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, WRP, Littleton, CO, 1960.

Schaake, J. C., V. I. Koren, and Q. Y. Duan, Simple water balance model for estimating runoff at different spatial and temporal scales, *Journal of Geophysical Research*, *101*, D3, 7461–7475, 1996.

Singh, V. P., Computer models of watershed hydrology, LSU Faculty, Water Resources Publication, 1995.

Sorooshian, S., Q. Y. Duan, and V. K. Gupta, Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting model, *Water Resources Research*, *29*, 4, 1185-1194, 1993.

Tanakamaru, H., Parameter estimation for the tank model using global optimization, *Transactions of the Japanese Society of Irrigation, Drainage and Reclamation Engineering*, *178*, 103–112, 1995.

U.S. Army Corps of Engineers, HEC-1 Flood Hydrograph package, Users and Programmers Manuals, HEC Program 723-X6-L2010, 1973.

U.S. Army Corps of Engineers, Hydrologic Engineering Center, HEC-1, Flood Hydrograph Package, Users Manual, September 1981, Rev. January, 1985.

Van Straten, G., and K. J. Keesman, Uncertainty propagation and speculation in projective forecasts of environmental change: A Lake-Eutrophication example, *Journal of Forecasting*, *10*, 163–190, 1991.

Wang, Q. J., The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resources Research*, *27*, 9, 2467–2471, 1991.

Wheater, H. S., S. Tuck, R. C. Ferrier, et al., Hydrological flow paths at the Allt A Mharcaidh Catchment-An analysis of plot and catchment scale observations, *Hydrology Process*, *7*, 4, 359–371, Oct-Dec, 1993.

Woolhiser, D. A., R. E. Smith, and D. C. Goodrich, A kinematic runoff and erosion manual: Documentation and user manual, ARS 77, U.S. Department of Agriculture, 1990.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *Journal of Hydrology*, *181*, 23–48, 1996.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Multi-objective global optimization for hydrologic models, *Journal of Hydrology*, *204*, 83–97, 1998.

CHAPTER 32

STOCHASTIC CHARACTERISTICS AND MODELING OF HYDROCLIMATIC PROCESSES

JOSÉ D. SALAS AND ROGER A. PIELKE, SR.

1 INTRODUCTION

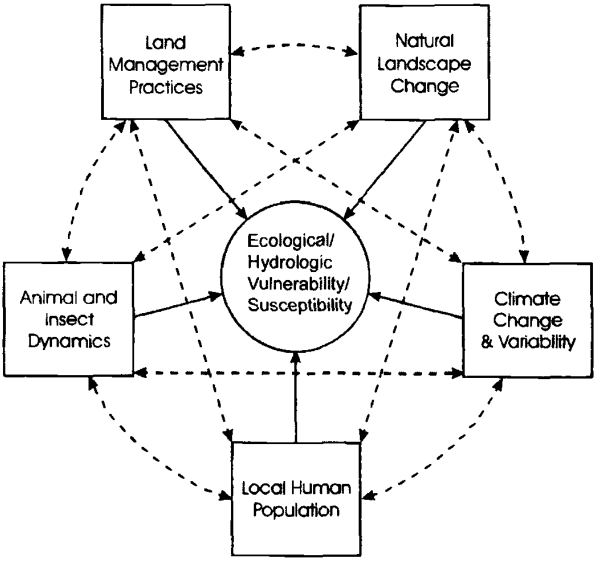
During 2000, the *Economist* (2001) reported that at least 6 of the top 12 loss of life events and 9 of the top 12 insured property losses were associated with hydrologic events. Late in 1999, an estimated 50,000 lives were lost associated with heavy rain along the northern coast of Venezuela. The quality and quantity of potable water is also important (Pielke and Guenni, 1999). As reported in the *Economist* (May 29, 1999, p. 102), while 90% of the world's population has enough water at present, by 2050 more than 40% of the population is estimated as facing a water shortage. The access to safe water is even more serious. In the same article the *Economist* reports that only about 30% of the rural residents of Brazil currently have access to safe water. Vorosmarty et al. (2000) demonstrate that population growth is the much larger threat to global water resources than any of the current generation projections of future climate. Understanding and quantifying the past, present, and future water availability at the global, regional, and local scales are scientifically, socially, and politically important aspects in balancing water supply and water demand.

Predictability of water resources at any scale requires a good understanding of atmospheric, oceans, and land surface processes and their interactions. In addition, land and oceanic biospheric processes play an important role in the global environment. Figure 1 illustrates the suite of environmental stresses that can threaten water resources. As population increases in a watershed, for example, increased clearing of trees and shrubs, as well as habitation within gulleys and ravines, can increase the vulnerability of the local population to flash flooding. This was a major factor in the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

large loss of life in the 1999 flood in Venezuela. Assessing the sensitivity of hydrologic processes to landscape change and vegetation dynamics represents one component of Figure 1. To illustrate the procedure to quantitatively assess sensitivity, Figure 2 shows the change in the total model simulated 210-day (during 1989) precipitation over the central United States (Eastman et al., 2001) associated with: (a) the conversion of the current landscape back to its natural form, (b) the radiative effect of doubled atmospheric carbon dioxide, and (c) the biological effect on vegetation of doubled atmospheric carbon dioxide. A coupled atmospheric–vegetation–soil dynamics model was used. The larger scale atmospheric forcing, however, remains identical for the three experiments and is derived from observed National Center for Environmental Prediction analyses (Kalnay et al., 1996).

This analysis shows the surprising result that both landscape change and the biological effect of carbon dioxide can exert a major effect on precipitation. With landscape change, the natural vegetation in the central United States had larger transpiration associated with greater vegetation coverage, particularly tall grass prairie in the eastern portion of the model domain. The increased transpiration cooled the daytime summer atmosphere, thereby preferentially permitting fewer rain showers in the model. Similarly, the enrichment of the atmosphere with carbon dioxide facilitated greater vegetation growth, such that cooling the daytime


Land Management Practices
Natural Landscape Change
Ecological/ Hydrologic Vulnerability/ Susceptibility
Animal and Insect Dynamics
Climate Change & Variability
Local Human Population

Predictability requires:
- the adequate quantitative understanding of these interactions
- that the feedbacks are not substantially nonlinear.

Figure 1 Use of ecological vulnerability/susceptioility in environmental assessment. (*Adapted from Pielke and Guenni, 1999.*)

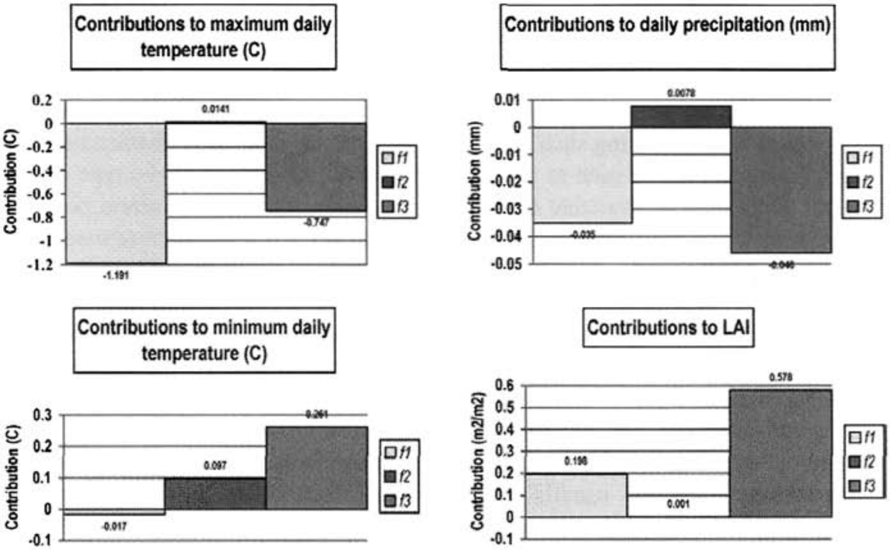


Figure 2 (see color insert) RAMS/GEMTM coupled model results—the seasonal domain-averaged (central Great Plains) for 210 days during the growing season, contributions to maximum daily temperature, minimum daily temperature, precipitation, and leaf area index due to f1 = natural vegetation, f2 = 2XC02 radiation, and f3 = 2xC02 biology. (*Adapted from Eastman et al., 2001*). See ftp site for color image.

atmosphere was also increased for this case. Such experiments illustrate a procedure to assess the sensitivity of hydrologic processes (precipitation in the above example) to environmental change. By assessing the sensitivity of a hydrologic process to the spectrum of environmental stressors, the largest sensitivities can be determined. With this information, social scientists and policy scientists can determine where to most effectively use resources to mitigate or adapt to the environmental threats (Sarewitz et al., 2000).

This brief introduction highlights the importance of the interrelationships and interactions among the various forcing functions of the environment, particularly as they relate to water resources availability and the effect of extremes such as floods and droughts on the environment and on society, and vice versa. Estimating those interactions and effects hinges on the proper characterization of the underlying hydroclimatic processes involved, such as air temperature, precipitation, humidity, snowpack, streamflow, infiltration, soil moisture, sea surface temperature, etc. The rest of this chapter focuses on the characterization and modeling of such processes by using stochastic methods. It is essentially an introduction and overview to two major separate chapters dealing specifically and more in depth with simulation (Salas et al., 2002) and forecasting (Valdes et al., 2002) of hydroclimatic processes particularly precipitation and streamflow.

2 GENERAL CHARACTERISTICS OF HYDROCLIMATIC PROCESSES

Mathematical models are generally used for stochastic simulation and forecasting of hydroclimatic processes. The stochastic characterization of the underlying processes is important in constructing such models. In general, the stochastic characteristics of hydroclimatic processes such as precipitation and runoff depend on the type of data at hand. Data may be available on a continuous time scale or at discrete points in time. For instance, most hydrologic series of practical interest are *discrete time series* defined on hourly, daily, weekly, monthly, bimonthly, quarterly, and annual time intervals. The term *seasonal time series* is often used for series with time intervals that are fractions of a year (usually a month or multiples of a month). Likewise, hourly, daily, weekly, monthly, and seasonal series are often called *periodic-stochastic* series. Hydroclimatic time series may consist of a *single time series* (*univariate series*) or *multiple time series* (*multivariate series*).

Hydroclimatic time series are generally *autocorrelated*. Autocorrelation in some series such as streamflow usually arises from the effect of surface, soil, and groundwater storages that cause the water to remain in the system through subsequent time periods (Salas, 1993). For instance, basins with significant surface storage in the form of lakes, swamps, or glaciers, produce streamflow series that are autocorrelated. Likewise, subsurface storage, especially groundwater storage produces significant autocorrelation in the streamflow series derived from groundwater outflow. Conversely, annual precipitation and annual maximum flows (flood peaks) are usually uncorrelated. Sometimes significant autocorrelation may be the result of trends and/or shifts in the series (Salas and Boes, 1980; Eltahir, 1989). In addition, multiple hydroclimatic series may be *cross-correlated*. For example, the precipitation series at two nearby sites, or the streamflow series of two nearby gaging stations in a river basin are expected to be cross-correlated because the sites are subject to similar climatic and hydrologic events. As the sites considered become farther apart, their cross-correlation decreases. However, because of the effect of some large-scale atmospheric-oceanic phenomena such as El Niño Southern Oscillation (ENSO), significant cross-correlation between sea surface temperature (SST) and streamflow between sites thousands of miles apart can be found (Eltahir, 1996). Furthermore, one would expect a significant cross-correlation between a streamflow time series and the corresponding areal average precipitation series over the same basin.

Hydroclimatic time series are *intermittent* when the variable under consideration takes on nonzero and zero values throughout the length of the record. For instance, the precipitation that is observed in a recording rain gage is an intermittent time series. Likewise, hourly, daily, and weekly rainfall are typically intermittent time series, while monthly and annual rainfall are usually nonintermittent. However, in semiarid and arid regions even monthly and annual precipitation and monthly and annual runoff may be intermittent as well.

Traditionally, certain annual hydroclimatic series have been considered to be *stationary*, although this assumption may be incorrect as a result of large-scale climatic variability, natural disruptions such as a volcanic eruption, and anthropogenic changes such as the effect of reservoir construction on downstream flow, and the effect of landscape changes on some components of the hydrologic cycle. On the other hand, hydroclimatic series defined at time intervals smaller than a year, such as months, generally exhibit distinct *seasonal* (*periodic*) patterns due to the annual revolution of Earth around the sun, which produces the annual cycle in most hydroclimatic processes. Some series of interest to hydrology and water resources, such as daily urban water use, may also exhibit a *weekly pattern* due to variations of demands within a week. Likewise, hourly time series may have a distinct *diurnal pattern* due to the variations of demands within a day. Summer hourly rainfall series or certain water quality constituents related to temperature may also exhibit distinct diurnal patterns due to the daily rotation of Earth that causes variations of net radiation within the day (Obeysekera et al., 1987; Katz and Parlange, 1995). Seasonal patterns of hydroclimatic series translate into statistical characteristics that vary within the year (or within a week or a day as the case may be) such as seasonal or periodic variations in the mean, variance, covariance, and skewness. Removing the seasonality in the mean and in the variance has been generally accomplished by the so-called *seasonal standardization*. This procedure is often referred to in the literature as *deseasonalization*. Unfortunately, this term is a misnomer since it may imply that the residual series is free of seasonality. However, seasonality may still be present in the covariance structure as is generally the case for seasonal streamflow series (Salas, 1993).

Hydroclimatic time series may exhibit trends, shifts or jumps, seasonality, autocorrelation, and non-normality. These attributes of hydroclimatic time series are referred to as *components* (Salas, 1993). In general, natural and human-induced factors may produce gradual and instantaneous trends and shifts (jumps) in hydroclimatic series. For example, a large forest fire in a river basin can immediately affect the runoff, producing a shift in the runoff series, whereas a gradual killing of a forest (e.g., by an insect infestation that takes years for its population to build up) can result in gradual changes or trends in the runoff series. A large volcanic explosion such as the one at Mount St. Helens in 1980 or a large landslide can produce sudden changes in the sediment transport series of a stream. Trends in non-point-source water quality series may be the result of long-term changes in agricultural practices and agricultural land development. Likewise, shifts in certain water quality constituents may be caused by agricultural activities such as sudden changes in the use of certain types of pesticides. Changes in land use and the development of reservoirs and diversion structures may also cause trends and shifts in streamflow series. The current concern about global warming and large-scale climatic variability, such as shifts in the intertropical convergence zone (ICZ) and the effects of large-scale oscillations such as ENSO and the Pacific Decadal Oscillation (PDO), is making hydroclimatologists more aware of the occurrence of trends and shifts in hydroclimatic time series. Figure 3 illustrates the observed swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998 (Landsea et al., 1999). Concerns regarding the effects of such types of sudden shifts observed in some hydroclimatic time series on water resources, the environment, and society have been expressed and documented in the literature (e.g., Kerr, 1992; Taylor, 1999). Statistical techniques are available for detecting,

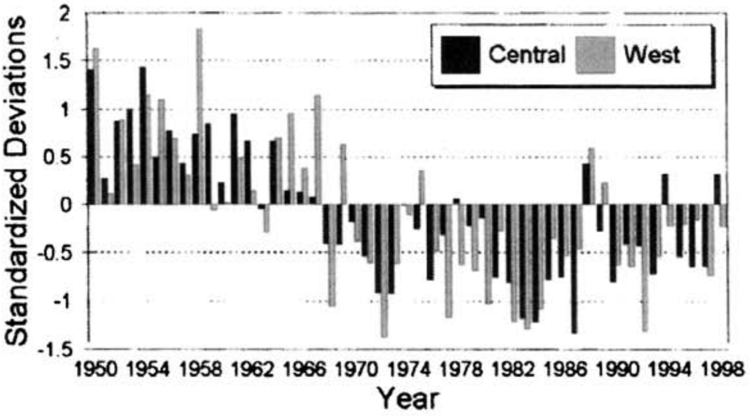


Figure 3 (see color insert) Swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998. (*After Landsea et al., 1999.*) See ftp site for color image.

modeling, and removing trends and shifts from hydroclimatic times series (Helsel and Hirsch, 1992; Salas, 1993; Hipel and McLeod, 1994).

3 STOCHASTIC ANALYSIS AND PROPERTIES OF HYDROCLIMATIC TIME SERIES

Overall Statistical Characteristics

The most commonly used statistical properties for analyzing stationary or nonstationary hydroclimatic time series are the sample mean $\bar{y}$, variance s^2, coefficient of variation cv, skewness coefficient g, lag-k autocorrelation coefficient r_k, and the spectrum $g(f)$. Coefficients of variation of annual flows are typically smaller than one, although they may be close to one or greater in streams in arid and semiarid regions. The coefficients of skewness g of annual flows are typically greater than zero. In some streams, small values of g are found suggesting that annual flows are approximately normally distributed. On the other hand, in some streams of arid and semiarid regions, g can be greater than one.

The lag-k autocorrelation coefficient r_k may be determined as

$$r_k = \frac{c_k}{c_0} \qquad k = 0, 1, 2, \ldots \tag{1a}$$

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \tag{1b}$$

where N is the sample size and k is the time lag. The plot of r_k versus k, i.e., the *correlogram*, may give an idea of the degree of persistence of the underlying time series, and it may be useful for choosing the type of stochastic model that may represent the series. When the correlogram decays rapidly to zero after a few lags, it may be an indication of small *persistence* or *short memory* in the series, while a slow decay of the correlogram is an indication of large persistence or *long memory*. The lag-one serial correlation coefficient r_1 is a simple measure of the degree of time dependence of a series. Generally, r_1 for annual flows is small but positive, although negative r_1's may occur because of sample variability. Large values of r_1 for annual flows can be found for a number of reasons including the effect of natural or man-made surface storage such as lakes, reservoirs, or glaciers, the effect of slow groundwater storage response, and the effect of nonstationarity. The estimators s^2, g, and r_k are biased (downward relative to the corresponding population statistics). Corrections for bias for these estimators have been suggested (Bobee and Robitaille, 1975; Yevjevich, 1972a; Fernandez and Salas, 1990).

In addition, the sample spectrum is another way of studying the variability of hydroclimatic series in the frequency domain (Yevjevich, 1972b). The sample spectrum $g(f_j)$ may be determined as

$$g(f_j) = 2\left[1 + 2\sum_{k=1}^{m} D_k r_k \cos(2\pi f_j k)\right] \qquad f_j = \frac{j}{2m} \quad j = 0, 1, 2, \ldots, m \tag{2}$$

where D_k is a smoothing function and m is the maximum number of lags considered. Figure 4 illustrates the autocorrelation function and the spectrum obtained for the time series of annual PDO indices for the period 1900–1999. The time series shows evidence of low-frequency components, which are manifested in a slow decaying and pseudoperiodic correlogram and a spectrum with visible high values at frequencies near 0.02 and 0.18 cycles per year.

When analyzing several time series jointly, cross-correlations may be important. The cross-correlation coefficient between series $y_t^{(i)}$ and $y_t^{(j)}$, $t = 1, \ldots, N$ for stations i and j, is determined as

$$r_k^{ij} = \frac{c_k^{ij}}{(c_0^{ii} c_0^{jj})^{1/2}} \qquad k = \cdots -2, -1, 0, 1, 2, \ldots \tag{3a}$$

$$c_k^{ij} = \frac{1}{N}\sum_{t=1}^{N-k} (y_{t+k}^{(i)} - \bar{y}^{(i)})(y_t^{(j)} - \bar{y}^{(j)}) \tag{3b}$$

The plot of r_k^{ij} vs. k is the *cross-correlogram*. For n time series, the values of r_k^{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, n$ are elements of the lag-k cross-correlation $n \times n$ matrix $\hat{M}_k$. Figure 5 is a graphical display of the lag-zero cross-correlation matrix obtained for the annual streamflows of 29 stations in the Colorado River system. For reference, station 1 is one of the farthest upstream site while station 29 is the farthest downstream site. The cross-correlation between stations 1 and 29 is large (of the order of 0.9) while the cross-correlation between stations 1 and 27 is small (the

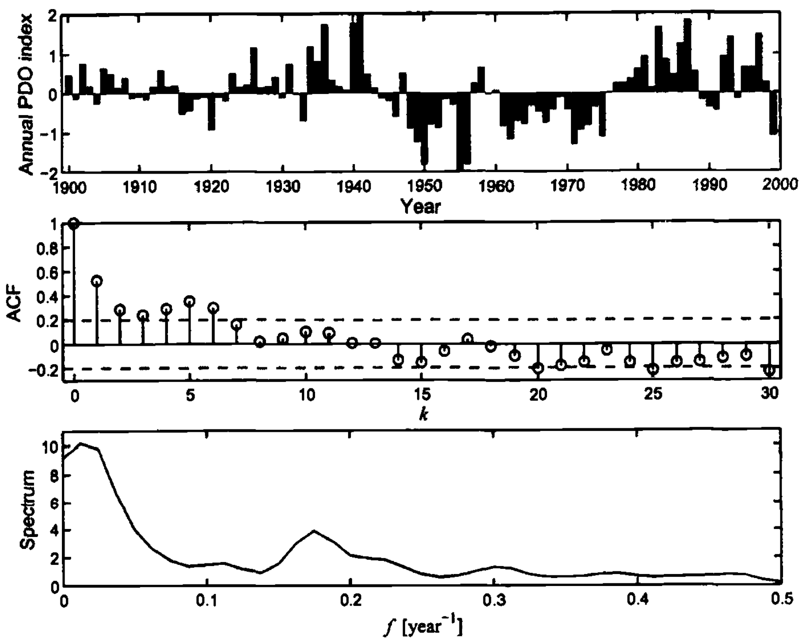


Figure 4 Autocorrelation function and spectrum obtained for the time series of annual PDO indices for the period 1900–1999. The time series shows evidence of low-frequency components. (*From Oli Sveinsson, Ph.D. candidate, CSU.*)

reason being that station 27 is a small tributary of the Colorado River and is very far from station 1).

In modeling hydroclimatic time series such as streamflow for simulation studies of reservoir systems, storage-related stochastic properties such as the range of cumulative departures R_n^*, the rescaled range R_n^{**}, and the Hurst slope K may be particularly important. They have been widely used in the literature as measures of long-term dependence and for comparing alternative models of hydrologic series (Hurst, 1951; Wallis and O'Connell, 1973; Hipel and McLeod, 1994). In particular, Hurst (1951) showed that for a large number of geophysical time series such as streamflow, precipitation, temperature, and tree-ring series, the mean rescaled range $\bar{R}_n^{**}$ (n = sample size) is proportional to n^h with $h > \frac{1}{2}$. The values of h obtained for different series gave a mean of about 0.73 and a standard deviation of 0.09. Theoretical results for normal independent processes and for autoregressive processes (Mandelbrott and Van Ness, 1968) indicated that asymptotically $h = \frac{1}{2}$. The discrepancy between theoretical results stating that $h = \frac{1}{2}$ and Hurst empirical findings suggesting that $h > \frac{1}{2}$ has become known as the *Hurst phenomenon*. However, the estimates of h are transient, meaning they depend on n and as $n \to \infty$, they generally converge to a limiting value, equal to $\frac{1}{2}$ for many time series models (Salas et al.,

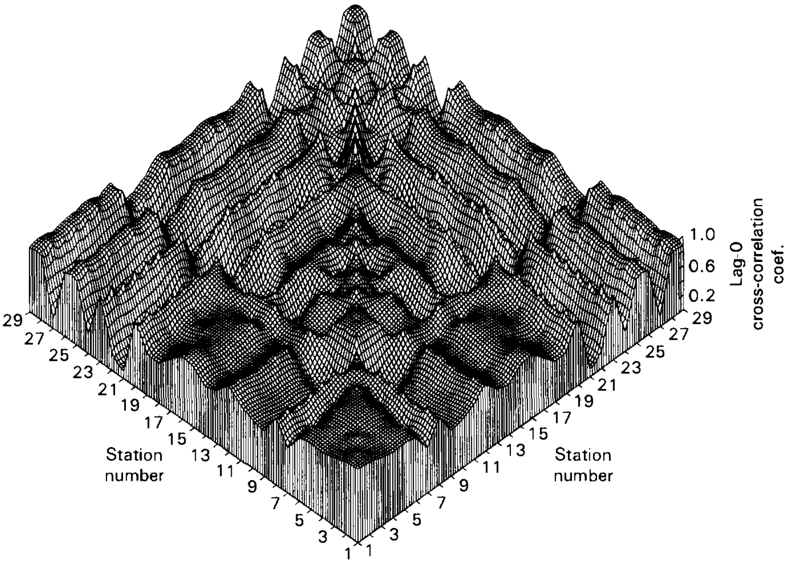


Figure 5 Lag-zero cross-correlation matrix obtained for the annual streamflows of 29 stations in the Colorado River system. For reference station 1 is the furthest upstream site while station 29 is the furthest downstream site.

1979). One interpretation of the Hurst phenomenon has been to associate $h = \frac{1}{2}$ with short memory models possessing short-term dependence structure, and $h > \frac{1}{2}$ with long memory models possessing long-term dependence. A number of models, including the autoregressive moving average (ARMA) processes, can have long-term dependence structure, yet asymptotically they give $h = \frac{1}{2}$. Furthermore, a stationary model with long-term dependence and $h > \frac{1}{2}$ is the fractional ARMA (FARMA) model (refer to Section 4 for definitions of ARMA and FARMA processes). Estimates of h can be useful for comparing the performance of alternative modeling strategies and estimation procedures. Statistical tests to determine whether a given time series exhibits the Hurst effect are also available (Mesa and Poveda, 1993).

Furthermore, drought-related stochastic properties are also important in modeling some hydroclimatic time series such as precipitation and streamflow. Consider a hydrologic time series y_t, $t = 1, \ldots, N$, and a *demand level d* (*crossing level*). Assume that y_t is an annual series and d is a constant (e.g., $d = \alpha\bar{y}$ and $0 < \alpha \leq 1$). A deficit at any given time t occurs when $y_t < d$. A consecutive sequence of deficits (until $y_t > d$ again) may be called a drought, and such a drought can be characterized by its duration L, its magnitude M, and its intensity $I = M/L$ (Yevjevich, 1967). Because a number of droughts can occur in a given hydrologic sample, the maximum drought duration, magnitude, and intensity (in a given sample) have been indicators of the so-called *critical drought* and have been widely used in water resources studies.

Periodic (Seasonal) Statistical Properties

While overall stochastic properties of hydroclimatic time series, such as those previously defined above, may be determined from either annual series or for seasonal series as a whole, specific seasonal (periodic) properties may provide a better picture of the stochastic characteristics of certain hydroclimatic time series that are defined at time intervals smaller than a year such as monthly streamflow data. Let the seasonal time series be represented by $y_{v,\tau}$, $v = 1, \dots, N$; $\tau = 1, \dots, \omega$ in which v is the year, τ is the season, N is the number of years of record, and ω is the number of seasons per year (e.g., $\omega = 12$ for monthly data). Then, for each season τ one can determine a number of statistics such as the seasonal mean $\bar{y}_\tau$, variance s_τ^2, coefficient of variation cv_τ, and skewness coefficient g_τ. Furthermore, the season-to-season correlation coefficient $r_{k,\tau}$ may be estimated by

$$r_{k,\tau} = \frac{c_{k,\tau}}{(c_{0,\tau-k}c_{0,\tau})^{1/2}} \qquad k = 0, 1, 2, \dots; \qquad \tau = 1, \dots, \omega \tag{4a}$$

$$c_{k,\tau} = \frac{1}{N}\sum_{v=1}^{N}(y_{v,\tau} - \bar{y}_\tau)(y_{v,\tau-k} - \bar{y}_{\tau-k}) \tag{4b}$$

For instance, for monthly streamflows $r_{1,4}$ represents the correlation between the flows of the fourth month with those of the third month. Likewise, for multiple seasonal time series, the lag-k seasonal cross-correlation coefficient $r_{k,\tau}^{ij}$ between the seasonal time series $y_{v,\tau}^{(i)}$ and $y_{v,\tau-k}^{(j)}$ for sites i and j, can be determined.

The statistics $\bar{y}_\tau$, s_τ, g_τ, and $r_{k,\tau}$ may be plotted versus time $\tau = 1, \dots, \omega$ to observe whether they exhibit a seasonal pattern. Fitting these statistics by Fourier series is especially effective with weekly and daily data (Salas et al., 1980). Generally, for seasonal streamflow series $\bar{y}_\tau > s_\tau$ although for some streams $\bar{y}_\tau$ may be smaller than s_τ especially during the "low-flow" season. Furthermore, for intermittent streamflow series generally the mean is smaller than the standard deviation, i.e., $\bar{y}_\tau < s_\tau$ throughout the year. Likewise, values of the skewness coefficient g_τ for the dry season are generally larger than those for the wet season indicating that data in the dry season depart more from normality than data in the wet season. Values of the skewness for intermittent hydrologic series are usually larger than skewness for similar nonintermittent series. Seasonal correlations $r_{k,\tau}$ for streamflow during the dry season are generally larger than those for the wet season, and they are significantly different than zero for most of the months. Figure 6 displays $r_{1,\tau}$, i.e., the lag-1 month-to-month correlations, for the monthly streamflows of the referred 29 stations of the Colorado River system. It may be observed that the correlations vary with the month, and with a few exceptions the correlation pattern for the entire system is similar. On the other hand, seasonal correlations for monthly precipitation are generally low or not significantly different from zero for most of the months (Roesner and

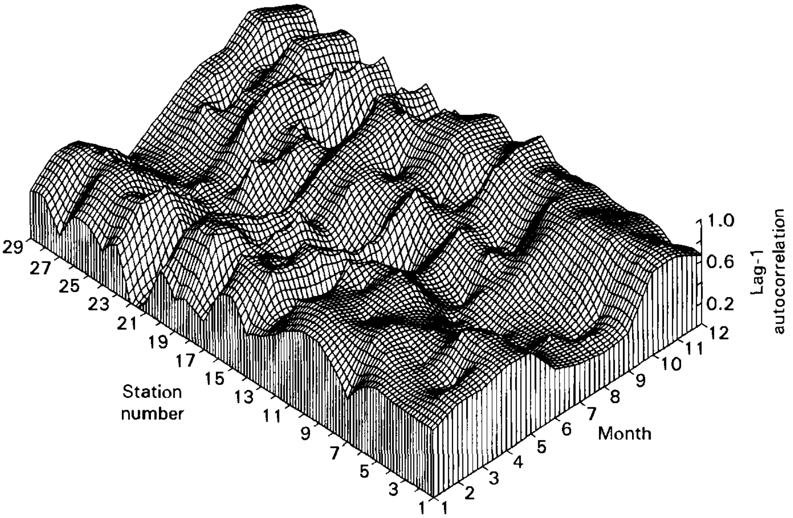


Figure 6 The lag-1 month-to-month correlations, i.e., $r_{1,\tau}$ for the monthly streamflows of 29 stations of the Colorado River system.

Yevjevich, 1966), while for weekly, daily, and hourly precipitation they are generally significant and greater than zero.

Complex, long-term dependence (long memory) of seasonal flows may be evident when the correlations $r_{k,\tau}$ are significant and decay slowly as k increases beyond ω seasons (beyond a year). These correlations are usually small or not significant for many streams, but in river systems such as the Nile River such seasonal correlations may persist for several years. Rivers that exhibit long-term correlation in seasonal flows will exhibit also long-term autocorrelation in the annual flows. In addition, some streamflow hydrographs such as daily and weekly hydrographs may possess directionality (nonreversibility), which means that some of their statistical properties change when direction of time is reversed. This is evident from the typical form of hydrographs in which the rising limb is shorter than the recession limb. In these cases, it is desirable that the mathematical models have such directionality attribute (Fernandez and Salas, 1986).

4 STOCHASTIC MODELS AND MODELING TECHNIQUES

A number of stochastic models and modeling schemes have been developed for simulation and forecasting of hydroclimatic processes. Some of the models are conceptually (physically) based, some others are empirical or transformed or adapted from existing models developed in other fields, while some others have arisen specifically to address some particular features of the process under consideration.

In general models for continuous time processes and models for short time scales such as hourly are more complex than models for larger time scales. Also some of the models have been developed specifically for precipitation while some others are for streamflow. Yet many of them are useful for both and for many other hydroclimatic processes. We will illustrate here as a matter of introduction and subsequent reference, the family of autoregressive and moving average (ARMA) models and extensions and modifications thereof. These models have become quite popular for both simulation and forecasting of many hydroclimatic processes. However, many other stochastic models have been developed, some of them quite different than ARMA models, aimed at the specific process under consideration or the particular features (of the underlying process) one tries to address. For example, for intermittent processes such as daily rainfall, Markov chains and the discrete counterpart of ARMA models, i.e., discrete ARMA (DARMA), are available (e.g., Chang et al., 1984; Guttorp, 1995). Likewise, models with infinite memory such as the Fractional Gaussian noise (e.g., Mandelbrot and Van Ness, 1968) and shifting level models that are capable of simulating sudden shifts (e.g., Salas and Boes, 1980) are available.

Stochastic Models

Stationary Models. The family of ARMA models has been widely used for modeling hydroclimatic processes at various time scales. The ARMA(p, q) model is defined as (Brockwell and Davis, 1991)

$$y_t = \mu + \sum_{j=1}^{p} \phi_j(y_{t-j} - \mu) + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{5a}$$

$$\phi(B)(y_t - \mu) = \theta(B)\varepsilon_t \tag{5b}$$

where μ, the ϕ's, the θ's, and $\sigma^2(\varepsilon)$ are parameters of the model, p is the order of the autoregressive terms, q is the order of the moving average terms, $B^i z_t = z_{t-i}$, and

$$\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p \tag{6a}$$

$$\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q \tag{6b}$$

Particular models derived from (5) are the ARMA($p, 0$) or AR(p) and the ARMA($0, q$) or MA(q) models. In addition, the fractional autoregressive moving average FARMA(p, d, q) model is defined as (Hosking, 1981; Montanari et al., 1997)

$$\phi(B)(1 - B)^d (y_t - \mu) = \theta(B)\varepsilon_t \qquad -0.5 < d < 0.5 \tag{7}$$

This model is capable of representing long-term dependence. The foregoing ARMA and FARMA models are stationary, hence their applications to modeling hydroclimatic time series require that the underlying data be stationary or be converted to stationary by some appropriate transformation.

These models have been generally applied to annual hydroclimatic data. Sometimes they have been applied to seasonal data after seasonal standardization. Likewise, they have been applied to daily data either after seasonal standardization or by separating the year into several seasons and applying different models to the daily series in each season. For example, Parlange et al. (1992) applied physically based concepts to the daily variations of soil moisture and found that it can be described by an AR(1) process. Properties of AR and ARMA models, such as the autocorrelation function, variance, and spectrum, and hydrologic applications may be found in Salas et al. (1980), Loucks et al. (1981), Bras and Rodriguez-Iturbe (1985), Salas (1993), and Hipel and McLeod (1994). Also Chu and Katz (1985) fitted AR and ARMA models to seasonal and monthly Southern Oscillation Index (SOI). Before fitting the models, the annual cycle was removed from the data in order to make them stationary. Xu and Storch (1990) used Principal Oscillation Pattern (POP) analysis to model monthly SOI data. They concluded that their POP scheme was superior over the ARMA scheme. Furthermore, Chu et al. (1995) applied a bivariate AR model for modeling jointly the seasonal SOI and a precipitation index in Florida. The fitted bivariate AR model was then used to forecast precipitation.

Periodic Models. A number of periodic and other nonstationary models such as the family of PARMA, ARIMA, and multiplicative PARMA models has been suggested in the literature for modeling seasonal hydroclimatic processes such as seasonal precipitation and streamflow series (Salas et al., 1980; Loucks et al., 1981; Salas, 1993; Hipel and McLeod, 1994). In particular, the PARMA(p,q) model is defined as

$$y_{v,\tau} = \mu_\tau + \sum_{j=1}^{p} \phi_{j,\tau}(y_{v,\tau-j} - \mu_{\tau-j}) + \varepsilon_{v,\tau} - \sum_{j=1}^{q} \theta_{j,\tau}\varepsilon_{v,\tau-j} \tag{8a}$$

$$\phi_\tau(B)(y_{v,\tau} - \mu_\tau) = \theta_\tau(B)\varepsilon_{v,\tau} \tag{8b}$$

where

$$\phi_\tau(B) = 1 - \phi_{1,\tau}B^1 - \phi_{2,\tau}B^2 - \cdots - \phi_{p,\tau}B^p \tag{9a}$$

$$\theta_\tau(B) = 1 - \theta_{1,\tau}B^1 - \theta_{2,\tau}B^2 - \cdots - \theta_{q,\tau}B^q \tag{9b}$$

and $B^i z_{v,\tau} = z_{v,\tau-i}$. When $q = 0$, the foregoing model becomes the well-known PARMA(p, 0) or PAR(p). More specifically the PAR(1) model (also known as the Thomas–Fiering model) is likely one of the most widely used models in hydrology. In general low-order PARMA models have become popular for modeling seasonal hydroclimatic processes. Physically based or conceptual arguments of the underlying hydrologic cycle of a watershed or river basin justify the applicability of these models. For instance, Salas and Obeysekera (1992) showed that assuming that the precipitation input is an uncorrelated periodic-stochastic process and under some linear reservoir considerations for the groundwater storage, the stochastic model for seasonal streamflow becomes a PARMA(1,1) process. Chu et al. (1995) analyzed time series of seasonal and monthly SOI and fitted AR and ARMA models after the annual cycle was removed from the data. They also used ARMA models with seasonally varying coefficients.

In addition, ARIMA(p, d, q), multiplicative ARMA, and multiplicative ARIMA models have been applied for forecasting hydroclimatic processes (e.g., Salas et al., 1980; Hipel and McLeod, 1994), sampling groundwater levels (Ahn and Salas, 1997), and for detection and estimation of trends in climatological timc series (e.g., Visser and Molenaar, 1995; Zheng and Basher, 1999). Furthermore, simulation of complex processes such as the Nile River monthly flows has been accomplished with multiplicative PARMA models (Salas et al., 1995). Also Lund et al. (1995) provide a general overview of the analysis and modeling of climatological time series having periodic correlation structure. They suggest a test for detection of periodic correlation and applied PARMA models for modeling such series. While the referred stationary and nonstationary models [e.g., models (5) and (8), respectively] are written for single-site or univariate series, their multisite or multivariate counterparts are also available (e.g. Salas, 1993; Hipel and McLeod, 1994).

Stochastic Models for Forecasting

A number of stochastic models have been widely applied for forecasting hydroclimatic processes such as precipitation and streamflow. Many of such models fall in the family of transfer function models. The general transfer function noise (GTFN) model may be written as

$$\gamma(B)(y_t - \mu_y) = \frac{\omega(B)}{\delta(B)}(x_{t-\tau} - \mu_x) + \frac{\theta(B)}{\phi(B)}\varepsilon_t \tag{10}$$

where γ(B), ω(B), δ(B), θ(B), and ϕ(B) are polynomials in B of different orders [similar to those defined in Eq. (6)], x_t is the exogenous variable such as precipitation or ENSO index, the μ's represent the means, τ is the time delay, and ε is the noise term. Some special cases (models) such as the ARMA, ARMAX, unit hydrograph type, multiple linear regression, and the Box–Jenkins transfer function noise models can be derived or simplified from (10). Equation (10) assumes single-site variables, but they are applicable to multisite variables if the variables are vectors and the parameters are matrices. Applications of many of these models can be found in Hipel and McLeod (1994). In addition, forecasting equations based on ARMA, ARMAX, and GTFN models can be written in sequential and recursive forms (e.g., using a Kalman filter). Furthermore, artificial neural networks (ANN) have emerged in the last decade as a useful technique for many modeling applications including forecasting (e.g., Hsu et al., 1995; Govindaraju and Rao, 2000). The application of many of these models, estimation procedures, and ANN algorithms for forecasting precipitation and streamflow are described in some detail in Valdes et al. (2001).

Modeling Schemes

Some specific models have been developed in the hydrologic field to address some unique features related to hydrologic and water resources problems. An example is the so-called *disaggregation* models (e.g., Valencia and Schaake, 1973). The failure of some traditional models such as the PAR(1) model to reproduce annual statistics (or upscale statistics) led to the development of disaggregation techniques. While the main intent of such disaggregation models has been to enable one to generate hydrologic sequences that can reproduce statistics at the annual and seasonal time scales, it has brought a major dimension into the capability of modeling complex hydrologic processes and complex hydrologic systems. Complex systems involve several sites, and the temporal and spatial mass balance requirements, often require the use of *modeling schemes* that may consist of an array of single-site, multisite, and temporal and spatial disaggregation models. While this requirement has been more evident in models constructed for simulation, the same is true for forecasting complex hydrologic systems. Furthermore, in order to facilitate the practical application of stochastic models for simulation of hydrological processes, software packages such as SPIGOT (Grygier and Stedinger, 1990) and SAMS (Salas et al., 2000) have been developed. Still, actual applications of such packages in real-world systems, especially for simulating complex hydrologic processes and complex water resources systems such as the Great Lakes system in North America or the Nile River system in Africa, may not be a straightforward application. Thus adjustments, modifications, additions, etc., may have to be made before a satisfactory or acceptable solution to the problem is attained.

Stochastic Modeling

Stochastic modeling of hydroclimatic processes may involve four major steps: model identification, parameter estimation, model testing, and model verification. By model identification is meant determining a specific model structure and the model order; for example, determining that the model for annual streamflow series is an ARMA(1,1) or determining that the model for daily rainfall is a simple Markov chain. Generally models that belong to the family of ARMA, ARIMA, and transfer function models are amenable for certain identification procedures based on autocorrelation, partial correlation, and cross-correlation analysis (Brockwell and Davis, 1991; Hipel and McLeod, 1994). However, model identification techniques are not available for some models or they are too complex, so instead a model of a certain type and order is applied to the particular hydroclimatic series at hand and its performance is judged by testing and verification. Some hydroclimatic processes such as streamflow and soil moisture have been identified using physically based concepts and arguments (e.g. Salas and Smith, 1981; Parlange et al., 1992; Salas and Obeysekera, 1992).

Once a model is identified, its parameters may be estimated by a number of techniques such as the method of moments, least squares, and maximum likelihood, depending on the particular model and data at hand. Typical method of moments estimation procedures involve matching historical and population (model) first- and second-order statistics, although in some cases some other properties such as skewness and storage and drought related statistics have been used (e.g., Salas et al., 1980; Hipel and McLeod, 1994). In addition, recursive parameter estimation methods and filtering techniques have been used particularly for forecasting problems (e.g., Bras and Rodriguez-Iturbe, 1985). Furthermore, modeling of hydroclimatic time series either for simulation or forecasting generally requires that the underlying series be transformed to approximately normally distributed series (e.g., Salas, 1993; Hipel and McLeod, 1994). Thus parameter estimation is usually made in the transformed domain. Model testing procedures have been well developed for models within the ARMA, ARIMA, ARMAX, and transfer function type of models (e.g., Brockwell and Davis, 1991). Likewise, testing procedures are available for PARMA models (e.g., Salas et al., 1980; Salas, 1993; Hipel and McLeod, 1994). The tests usually involve diagnostic checks to verify whether the model residuals comply with the underlying assumptions of independence and normality (of the residuals). Since many models may comply with such requirements, a model selection criteria based on the Akaike Information Criteria (AIC) is available to discriminate and find a parsimonious model (Brockwell and Davis, 1991). On the other hand, model testing for some other models cannot be done based on analysis of residuals; so instead model testing is based on data generation experiments. In addition, model verification is usually needed beyond testing residuals depending on whether the modeling exercise is geared to simulation or forecasting. For instance, for simulation (data generation) one may like to test whether the model is capable of generating sequences that reproduce a number of storage and drought related historical characteristics. This is usually accomplished by Monte Carlo experiments. On the other hand, model verification for forecasting may involve examining whether the model is capable of estimating the hydrologic process under consideration one or more lead times in advance within a specified error criteria. This may be done by split sampling estimation and testing.

REFERENCES

Ahn, H., and J. D. Salas, Groundwater head sampling based on stochastic analysis, *Water Resour. Res.*, *33*(12), 2769–2780, 1997.

Bobee, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, *11*(6), 851–854, 1975.

Bras, R. L., and Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, MA, 1985.

Brockwell, P. J., and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag, New York, 1991.

Chang, T. J., M. L. Kavvas, and J. W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, *Water Resour. Res.*, *20*(5), 565–580, 1984.

Chu, P. S., and R. W. Katz, Modeling and forecasting the Southern Oscillation: A time-domain approach, *Monthly Weather Rev.*, *113*, 1876–1888, 1985.

Chu, P. S., R. W. Katz, and P. Ding, Modeling and forecasting seasonal precipitation in Florida: A vector time-domain approach, *Int. J. Climatol.*, *15*, 53–64, 1995.

Eastman, J. L., M. B. Coughenour, and R. A. Pielke, The effects of CO_2 and landscape change using a coupled plant and meteorological model, *Global Change Biol.*, 7, 797–815, 2001.

Economist, Catastrophes, March 31, 106, 2001.

Eltahir, E. A. B., A feedback mechanism in annual rainfall in Central Sudan, *J. Hydrol.*, 110, 323–334, 1989.

Eltahir, E. A. B., El Niño and the natural variability in the flow of the Nile River, *Water Resour. Res.*, *32*(1), 131–137, 1996.

Fernandez, B., and J. D. Salas, Periodic gamma autoregressive processes for operational hydrology, *Water Resour. Res.*, *22*(10), 1385–1396, 1986.

Fernandez, B., and J. D. Salas, Gamma-autoregressive models for streamflow simulation, *J. Hydrol. Eng. ASCE*, *116*(11), 1403–1414, 1990.

Govindaraju, R., and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, London, 2000.

Grygier, J. C., and J. R. Stedinger, *SPIGOT, A Synthetic Streamflow Generation Software Package, Technical Description*, Version 2.5, *Cornell University*, School of Civil and Environmental Engineering, Ithaca, NY, 1990.

Guttorp, P., *Stochastic Modeling of Scientific Data*, Chapman & Hall, London, 1995.

Helsel, D. R., and R. M. Hirsch, *Statistical Methods in Water Resources*, Studies in Environmental Science 49, Elsevier, Amsterdam, 1992.

Hipel, K. W., and A. I. McLeod, *Time Series Modeling of Water Resources and Environmental Systems*, Elsevier, Amsterdam, 1994.

Hosking, J. R. M., Fractional differencing, *Biometrika*, *68*, 165–176, 1981.

Hsu, K., H. V. Gupta, and S. Sorooshian, Artificial neural network modeling of the rainfall-runoff process, *Water Resour. Res.*, *31*(10), 2517–2530, 1995.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans. ASCE*, *116*, 770–799, 1951.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, 1996.

Katz, R. W., and M. B. Parlange, Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*(5), 1331–1341, 1995.

Kerr, R. A. Unmasking a shifty climate system, *Research News*, *255*, 1508–1510, 1992.

Landsea, C. W., W. M. Gray, P. W. Mielke Jr., K. J. Berry, and R. K. Taft, June to September rainfall in North Africa: A seasonal forecast for 1999, Atmospheric Science Department, Colorado State University, Fort Collins, CO, available on-line, http://typhoon.atmos.colostate.edu/forecasts/1999/sahel_jun99/,1999.

Loucks, D. P., J. R. Stedinger, and D. Haith, *Water Resources Systems Planning and Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

Lund, R., H. Hurd, P. Bloomfield, and R. Smith, Climatological time series with periodic correlation, *J. Climate*, *8*, 2787–2809, 1995.

Mandelbrot, B. B., and J. W. Van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Rev.*, *10*(4), 422–437, 1968.

Mesa, O. J., and G. Poveda, The Hurst effect: The scale of fluctuation approach, *Water Resour. Res.*, *29*(12), 3995–4002, 1993.

Montanari, A., R. Rosso, and M. S. Taqqu, Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation, *Water Resour. Res.*, *33*(5), 1035–1044, 1997.

Obeysekera, J. T. B., G. Tabios, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

Parlange, M. B., G. G. Katul, R. H. Cuenca, M., L. Kavas, D. R. Nielsen, and M. Mata, Physical basis for a time series model of soil water content, *Water Resour. Res.*, *28*(9), 2437–2446, 1992.

Pielke, Sr., R. A., and L. Guenni, Vulnerability assessment of water resources to changing environmental conditions, *IGBP Newsl.*, *39*, 21–23, 1999.

Roesner, L. A., and V. Yevjevich, Mathematical models for time series of monthly precipitation and monthly runoff, in *Hydrology Papers*, No. 15, Colorado State University, Ft. Collins, CO, 1966.

Salas, J. D., *Analysis and Modeling of Hydrologic Time Series*, in D. R. Maidement (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 19.

Salas, J. D., and D. C. Boes, Shifting level modeling of hydrologic series, *Adv. Water Resour.*, *3*, 59–63, 1980.

Salas, J. D., and J. T. B. Obeysekera, Conceptual basis of seasonal streamflow time series models, *ASCE J. Hydraul. Eng.*, *118*(8), 1186–1194, 1992.

Salas, J. D., and R. A. Smith, Physical bases of stochastic models of annual flows, *Water Resour. Res.*, *17*, 428–430, 1981.

Salas, J. D., D. C. Boes, V. Yevjevich, and G. G. S. Pegram, Hurst phenomenon as a pre-asymptotic behavior, *J. Hydrol.*, *44*(1), 1–15, 1979.

Salas, J. D., J. R. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, Littleton, CO, 1980.

Salas, J. D., N. Saada, and C. H. Chung, Stochastic modeling and simulatio of the Nile River system monthly flows, Colorado State University, Engineering Research Center, Comp. Hydrol. Lab., Technical Report No. 5, Fort Collins, CO, 1995.

Salas, J. D., N. Saada, C. H. Chung, W. L. Lane, and D. K. Frevert, Stochastic analysis, modeling, and simulation (SAMS), Version 2000, User's manual, Colorado State University, Engineering Research Center, Comp. Hydrol. Lab., Technical Report No. 10, Fort Collins, CO, 2000.

Salas, J. D., J. A. Ramirez, P. Burlando, and R. Pielke, Sr., Stochastic simulation of precipitation and streamflow processes, in T. D. Potter and B. Colman (Eds.), *Handbook of Weather, Climate, and Water*, Chapter 33 John Wiley & Sons, Inc., New York, 2002.

Sarewitz, D., R. A. Pielke, Jr., and R. Byerly (Eds.), *Prediction: Science Decision Making and the Future of Nature*, Island Press, Covelo, CA, 2000.

Taylor, K. Rapid climate change, *Am. Sci.*, *87*, 320–326, 1999.

Valdes, J. B., P. Burlando, and J. D. Salas, Stochastic forecasting of precipitation and streamflow processes, in T. D. Potter and B. Colman (Eds.), *Handbook of Weather, Climate, and Water*, Chapter 34 John Wiley & Sons, Inc., New York, 2002.

Valencia, R. D., and J. C. Schaake, Jr., Disaggregation process in stochastic hydrology, *Water Resour. Res.*, *20*(1), 580–585, 1973.

Visser, H., and J. Molenaar, Trend estimation and regression analysis in climatological time series: An application of structural time series models and the Kalman filter, *J. Climate*, *8*, 969–979, 1995.

Vorosmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers, Global water resources: Vulnerability from climate change and population growth, *Science*, *289*, 284–288, 2000.

Wallis, J. R., and P. E. O'Connell, Firm reservoir yield: How reliable are hydrological records, *Hydrol. Sci. Bull* (now *Hydrol. Sci. J.*), *18*, 347–365, 1973.

Xu, J. S., and H. V. Storch, Predicting the state of the Southern Oscillation using principal oscillation pattern analysis, *J. Climate*, *3*, 1316–1329, 1990.

Yevjevich, V., An objective approach to definitions and investigations of continental droughts, in *Hydrology Papers*, Vol. *23*, Colorado State University, Fort Collins, CO, 1967.

Yevjevich, V., Structural analysis of hydrologic time series, in *Hydrology Papers*, Vol. 56, Colorado State University, Fort Collins, CO, 1972a.

Yevjevich, V., *Stochastic Process in Hydrology*, Water Resources Publications, Littleton, CO, 1972b.

Zheng, X., and R. E. Basher, Structural time series models and trend detection in global and regional time series, *J. Climate*, *12*, 2347–2358, 1999.

CHAPTER 33

STOCHASTIC SIMULATION OF PRECIPITATION AND STREAMFLOW PROCESSES

JOSÉ D. SALAS, JORGE A. RAMÍREZ, PAOLO BURLANDO, AND ROGER A. PIELKE, Sr.

Stochastic simulations of hydroclimatic processes such as precipitation and streamflow have become standard tools for analyzing many water-related problems. Simulation signifies "mimicking" the behavior of the underlying process so that realistic representations of it can be made. For this purpose a number of empirical, mathematically/physically based, mathematically/stochastically based, analog/physically based, and physical/laboratory-scale based models and approaches have been proposed and developed in the literature. This chapter emphasizes simulation based on stochastic and probabilistic techniques. Also, the emphasis will be on precipitation and streamflow processes, although many of the methods and models included herein are equally applicable for other hydroclimatic processes as well such as evapotranspiration, soil moisture, surface and groundwater levels, and sea surface temperature.

Stochastic simulation enables one to obtain equally likely sequences of hydroclimatic processes that may occur in the future. They are useful for many water resources problems such as (a) estimating the design capacity of a reservoir system under uncertain streamflows, (b) evaluating the performance of a water resources system in meeting projected water demands under uncertain system's inputs, (c) estimating drought properties, such as drought length and magnitude based on simulated streamflows at key points in the water supply system under consideration, (d) deriving the distribution of the underlying output variable of a groundwater flow equation (e.g., the hydraulic head), given the distribution of the parameters (e.g., the hydraulic conductivity) and boundary conditions, (e) establish-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

ing the uncertainty in travel time and spread of pollutants in porous media as a function of the uncertainty in the parameters of the groundwater contamination transport model, and (f) analyzing the impacts of large-scale climate variability and global climate change on water supply availability and the ensuing planning and operation of water resources projects.

1 STOCHASTIC SIMULATION OF PRECIPITATION

Continuous-Time Precipitation

The theory of *point processes* has been applied for modeling continuous-time precipitation since Le Cam[60] suggested that a Poisson process could model the occurrence of rainfall showers. Let us assume that the number of storms $N(t)$ in a time interval $(0, t)$ arriving to a given point is Poisson distributed with parameter λt (λ = storm arrival rate.) Referring to Figure 1(a), n storms arrived in the interval $(0, t)$ at times $t_1, \ldots, t_n$. The number of storms in any time interval T is also Poisson distributed with parameter λT. Assume further that the rainfall amount R associated with a storm arrival is *white noise* (e.g., R may be gamma distributed) and that $N(t)$ and R are independent. Thus, rainfall amounts $r_1, \ldots, r_n$ correspond to storms occurring at times $t_1, \ldots, t_n$. Such a rainfall generating process has been called *Poisson white noise* (PWN).

The cumulative rainfall in the interval $(0, t)$, $Z(t) = \sum_{j=1}^{N(t)} R_j$ is a *compound Poisson process*. Also the cumulative rainfall over successive nonoverlapping time intervals T, i.e., the discrete-time rainfall process (refer to Fig. 1), is given by $Y_i = Z(iT) - Z(iT - T), i = 1, 2, \ldots$. The basic statistical properties of Y_i assuming that $Z(t)$ is generated by a PWN model has been widely studied.[11,20] Its autocorrelation function $\rho_k(Y)$ is equal to zero for all lags greater than zero, which contradicts actual observations [e.g., $\hat{\rho}_1(Y) = 0.446$ for hourly precipitation at Denver Airport station for the month of June based on the 1948–1983 records.] Despite this shortcoming, the PWN model can be useful for predicting annual precipitation[20] and extreme precipitation events.[10] Instead of assuming that rainfall occurs instantaneously with zero duration one may consider rainfall with random duration D and intensity I, as shown in Figure 1b. This is called the *Poisson rectangular pulse* (PRP) model.[86] A common assumption is that D and I are independent and exponentially distributed. Figure 1b shows a PRP process with n storms in the interval $(0, t)$ occurring at times $t_1, \ldots, t_n$ with associated intensities and durations $(i_1, d_1), \ldots, (i_n, d_n)$. Then, storms may overlap and the aggregated process Y_i becomes autocorrelated. Although the PRP model is better conceptualized than the PWN, it is still limited when applied to rainfall data.[86] Thus, alternative models based on the concept of clusters have been suggested.

Neyman and Scott[69] in modeling the spatial distribution of galaxies originally suggested the concept of *clusters*. Le Cam[60], Kavvas and Delleur,[51] and others[30,82,86–88] applied this concept of space clustering to model continuous-time rainfall. The *Neyman–Scott cluster process* can be described as a two-level mechan-

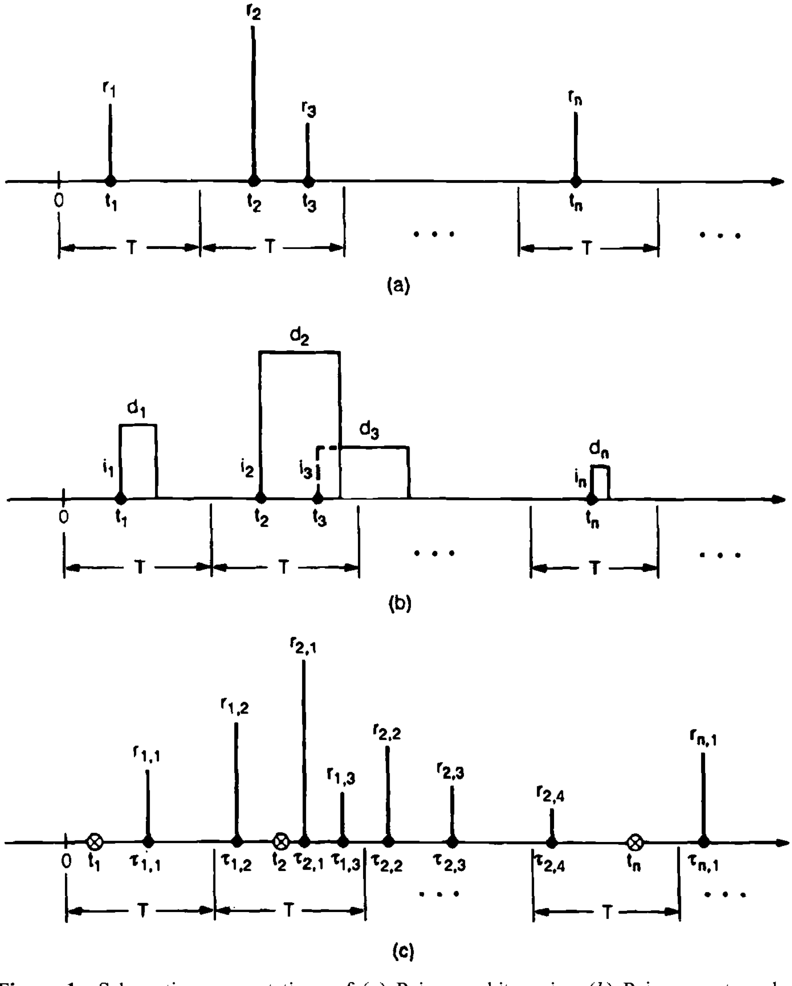


Figure 1 Schematic representations of (*a*) Poisson white noise, (*b*) Poisson rectangular pulse, and (*c*) Neyman–Scott white noise processes (after Salas[93]).

ism for generating rainfall. First, storm-generating mechanisms or simply storms arrive governed by a Poisson process with parameter λt. Figure 1*c* shows that n storms arrive at times $t_1, \ldots, t_n$ in the period $(0, t)$. Then, associated with each storm, there are a number of precipitation bursts that are Poisson or geometrically distributed with parameter ν. Figure 1*c* shows three precipitation bursts associated with the storm that arrived at time t_1. In general m_j precipitation bursts are associated with the storm that arrived at time t_j. In addition, the time of occurrence of bursts, τ, relative to the storm origin t_j may be assumed to be exponentially distributed with parameter β (e.g., in Fig.1*c* the three bursts arising from the first storm are located at times $\tau_{1,1}$, $\tau_{1,2}$, and $\tau_{1,3}$ relative to t_1). Then, if the precipitation burst is described by an instantaneous random precipitation depth R, the resulting precipitation process is known as *Neyman–Scott white noise* (NSWN) while if the precipitation burst is a rectangular pulse the precipitation process is known as *Neyman–Scott rectangular pulse* (NSRP).

Estimation of parameters for Neyman–Scott (NS) models has been a major subject of research in the past two decades.[9,22,30,51,71] The usual estimation has been based on the method of moments, although other approaches have been suggested.[29,51,82] An apparent major estimation problem is that parameters estimated based on data for one level of aggregation, say hourly, may be significantly different from those estimated from data for another level of aggregation, say daily.[11,30,71,86] The problem seems to be that as data are aggregated, information is lost and corresponding second-order statistics do not have enough information to give reliable estimates of the parameters of the generating process (model), and, as a consequence, they become significantly biased with large variance. For example, extensive simulation studies were carried out by Cadavid et al.[11] based on the NSWN model with (known) population parameters: $\lambda = 0.102 \times 10^{-3}$/min, $\beta = 0.00221$/min, $\mu = 24.36$/in, and $1/v = 0.072$ (parameter of the geometric distribution for the cluster size). Hourly and daily series were used to estimate moments (mean, standard deviation, and lag-1 and lag-2 correlation coefficients) from which the parameters were estimated. The results are shown in Table 1. Clearly, despite that the generating mechanism is known (the NSWN), less reliable estimates of parameters are obtained when daily values are used. Estimation based on weighted moments of various time scales in a least-squares fashion is an alternative.[10,22] Also, physical considerations may be useful in setting up constraints in some of the parameters, initializing the estimates to be determined based on statistical considerations, and for comparing the fitted model parameters with some known physical properties.[17] Koepsell and Valdes[54] applied these concepts using the space–time cluster model suggested by Waymire et al.[116] for modeling rainfall in Texas and pointed out the difficulty in estimating the parameters even when using physical considerations.

Besides the class of Poisson processes and Neyman–Scott cluster processes, other types of temporal precipitation models have been suggested such as those based on Cox processes,[103] renewal processes,[7,31] and Barlett–Lewis processes.[40,87] Likewise, alternative space–time multidimensional precipitation models have been developed (e.g., Smith and Krajewski[104]). In addition, all precipitation models based on point and cluster processes proposed up to date are limited in some respects; e.g., they do not include the daily periodicity observed in actual convective rainfall processes.[49,71] Furthermore, Rodriguez-Iturbe et al.[89] and others raised the issue that nonlinear dynamics and chaos may be useful approaches for certain hydrometeorological processes such as rainfall. Finally, excellent reviews of the state of the art in the field have been made[32] and a number of studies pertaining to rainfall analysis, modeling, and predictability have been compiled in special issues of some

TABLE 1 Comparison between Population and Estimated Parameters of NSWN Model Based on Hourly and Daily Values

Parameter (units)	Population Value	Estimated from Hourly Data[a]	Estimated from Daily Data[a]
$\lambda \times 10^3$ (1/min)	0.102	0.103	0.091
$\beta \times 10^3$ (1/min)	2.210	2.300	1.630
μ (1/in)	24.360	23.990	7.010
$1/\nu$	0.072	0.072	0.247

[a] Estimates based on 12 series of size 36,456 for hourly and 12 series of size 1519 for daily. From Cadavid et al.[11]

journals (e.g., *J. Appl. Meteor.*, vol. 32, 1993; *J. Geoph. Res.*, vol. 104, no. D24, 1999).

Hourly, Daily, and Weekly Precipitation

We have seen in the previous section that the models and properties for cumulative precipitation over successive nonoverlapping time periods, i.e., discrete-time precipitation, can be derived from continuous-time precipitation models. However, one may formulate precipitation models directly at hourly, daily, and weekly time scales. In these cases, the theory of *Markov chains* has been widely used in the literature for simulating not only precipitation (in discrete time) but many other hydrologic processes such as streamflow, soil moisture, temperature, solar radiation, and water storage in reservoirs.[7,13,49,85,90]

Consider that $X(t)$ is a discrete valued process that started at time 0 and developed through time, i.e., $t = 0, 1, 2, \ldots$. Then $P[X(t) = x_t | X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}]$ is the probability that the process $X(t) = x_t$ given its entire history. If this probability simplifies to $P[X(t) = x_t \mid X(t-1) = x_{t-1}]$, the process is a *first-order Markov chain* or a *simple Markov chain*. Because $X(t)$ is a discrete valued process, we will use the notation $X(t) = j, j = 1, \ldots, r$ instead of $X(t) = x_t$, where j represents a *state* and r is the number of states; e.g., in modeling daily rainfall one may consider $r = 2$ with $j = 1$ for a dry day (no rain) and $j = 2$ for a wet day. A simple Markov chain is defined by its *transition probability matrix* $P(t)$, a square matrix with elements $p_{ij}(t) = P[X(t) = j | X(t-1) = i]$ for all i, j pairs. Furthermore, $q_j(t) = P[X(t) = j], j = 1, \ldots, r$, is the marginal probability distribution of the chain being at any state j at time t and $q_j(0)$ is the distribution of the initial states. Moreover, if $P(t)$ does not depend on time, the Markov chain is a *homogeneous* or *stationary chain* and, in this case, the notations P and p_{ij} are used. The estimation of some probabilities that are useful for simulation and forecasting of precipitation events are the n-step transition probability $p_{ij}^{(n)}$, the marginal distribution $q_j(t)$ given the distribution $q_j(0)$, and the steady-state probability vector q^*. These probabilities can be determined from well-known relations available in the literature.[39,118]

Estimation for a simple Markov chain amounts to estimating the elements p_{ij} of the transition probability matrix. Common estimation methods include the method of moments and maximum likelihood.[39] To test whether a simple Markov chain is an adequate model for the process under consideration, one can check some of the assumptions of the model and see whether some relevant properties of the precipitation process are reproduced (e.g., compare the probability $p_{ij}^{(n)}$ with that obtained from the observed data, $\hat{p}_{ij}^{(n)}$). Furthermore, the Akaike information criterion has been helpful in selecting the order of Markov chain models.[15,48]

Although in some cases simple Markov chains may be adequate for representing the variability of precipitation, often more complex models may be necessary. For instance, in modeling daily rainfall processes throughout the year, the parameters of the Markov chain may vary with time (e.g., for a two-state Markov chain, the transition probabilities p_{ij} may vary along the year and the estimates can be fitted with trigonometric series to smooth out sample variations[90]). Higher order Markov chains may be necessary in other cases. Chin[15] analyzed daily precipitation records of more than 100 stations across the continental United States and concluded that generally second- and third-order models were preferred for the winter months while the first-order model was better for the summer months. In addition, maximum likelihood for estimating Fourier series coefficients for alternating renewal processes and Markov chains for daily rainfall[90] and mixed models with periodic Markov chains for hourly rainfall (to account for the effect of daily periodicity) have been suggested.[49]

Monthly, Seasonal, and Annual Precipitation

Modeling of precipitation for long time scales such as monthly is generally simpler than for short time scales such as daily, especially because for long time scales the autocorrelation becomes smaller or negligible (except in cases of low frequency[21]). In such cases modeling precipitation at a given site amounts to finding the probability distribution for each month. Generally different distributions will be needed for each month. On the other hand, seasonal precipitation data in semiarid and arid regions may include zero values for some seasons, hence the precipitation is a mixed random variable. Let $X_{v,\tau}$ = precipitation for year v and season τ, and define $P_\tau(0) = P(X_{v,\tau} = 0)$, $\tau = 1, \ldots, \omega$ (ω = number of seasons per year). Then, $F_{X\tau}(x) = P_\tau(0) + [1 - P_\tau(0)]F_{X\tau|X\tau>0}(x)$ is the cumulative distribution function for season τ, in which $F_{X\tau}(x) = P_\tau(X \leq x)$ and $F_{X\tau|X\tau>0}(x) = P(X \leq x|X > 0)$. Thus, prediction of seasonal precipitation requires estimating $P_\tau(0)$ and $F_{X\tau|X\tau>0}(x)$. Several distributions such as the log-normal and log-Pearson have been used for fitting the empirical distribution of seasonal precipitation. For modeling precipitation at several sites, one must consider the intersite cross correlations and the marginal distribution (at each site). For continuous random precipitation, a common modeling approach has been to transform them into normal, then use a lag-0 multivariate model for modeling the transformed precipitation (an approach similar to modeling streamflow as in Section 2). Modeling of annual precipitation is similar to modeling seasonal precipitation, i.e., determining either the marginal distribution $F_X(x)$ or the conditional distribution $F_{X*X>0}(x)$, depending on the particular case at hand. Likewise, modeling of annual precipitation at several sites is generally based on transforming the data into normal and using a multivariate normal model.

2 STOCHASTIC SIMULATION OF STREAMFLOW

If one can develop a stochastic model for streamflow in continuous time, then, in principle, the properties and the models for daily, monthly, and annual streamflow can be obtained. Some attempts have been made for developing models of streamflow processes in continuous time based on physical principles.[20] However, the models of aggregated flows that can be derived from such continuous-time models, become mathematically cumbersome and of limited applicability for operational hydrology.[53] Understanding the rules for upscaling the models and parameters has been a challenging subject for research. Generally most of the models that are available for streamflow simulation in continuous time and short time scales, such as hourly, are based on the transformation of precipitation into runoff by means of physical or conceptual principles. Thus the stochastic characteristics of the precipitation input and of the other relevant processes of the hydrologic cycle of the watershed are transferred into a stochastic streamflow output. Examples of models in this category are represented by SHETRAN[26] and PRMS.[59] SHETRAN simulates "continuous" streamflows along the river network by solving partial differential equations of the physical processes involved while PRMS is a semidistributed conceptual model that simulates hourly and daily streamflows. However, in this section we are mainly concerned with stochastic streamflow models that can be derived explicitly from the physically or conceptually based relations of the underlying hydrologic process of the watershed or directly from the streamflow data.

Continuous Time to Hourly and Daily Streamflow Simulation

The simulation of streamflow on a continuous time scale requires the formulation of a model structure that is capable of reproducing the streamflow fluctuations on a wide dynamical range. As already mentioned, the application of stochastic approaches to continuous time and short time scale streamflow modeling has been limited because of the complex nonlinear relations that characterize the precipitation-streamflow processes at those temporal scales. The early attempts to model hourly and daily streamflows were based on using autoregressive (AR) models after standardization and transformation. However, stochastic models essentially based on process persistence do not properly account the rising limb and recession characteristics that are typical of hourly and daily flow hydrographs. Also shot noise or Markov processes and transfer function models have been proposed for daily flow simulation with some limited success in reproducing the rising limb and recessions.[110]

Nevertheless, interesting work has been done with some success by using conceptual-stochastic models. For instance, Kelman[52] applied a conceptual representation of a watershed considering the effects of direct runoff and surface and groundwater storages. Direct runoff is modeled by a PAR(1) model with an indicator function to produce intermittence and the other components are modeled using linear reservoirs. Kelman's model produced reasonable results for generating daily flows for the Powell River, Tennessee. Also following the approach suggested by Salas and Obeysekera[96] and Claps et al.,[16] Murrone et al.[68] proposed a conceptual-stochastic model for short time runoff. A three-level conceptual runoff component and a stochastic surface runoff model the daily response of the watershed. The base flow is modeled by three linear reservoirs that represent the contribution of deep aquifers with over-year response, aquifers with annual renewal, and subsurface runoff. The surface runoff is regarded as an uncorrelated point process. Modeling rainfall as an independent Poisson process, the above scheme leads to a multiple shot noise streamflow process. The model is effective in reproducing streamflow variability. In addition, intermittent daily streamflow processes have been modeled[25] by combining Kelman's conceptual approach with product models[94] and gamma AR models[28] and by using a three-state Markov chain describing the onset of streamflow and an exponential decay of streamflow recession.[1]

Weekly, Monthly, and Seasonal Streamflow

Single-Site Periodic Models. Stationary stochastic models can be applied for modeling weekly, monthly, and seasonal streamflows after *seasonal standardization*. This approach may be useful when the season-to-season correlations do not vary throughout the year. In general though, models with periodic correlation structure, such as periodic autoregressive (PAR) and periodic autoregressive and moving average (PARMA) are more applicable.[29,92] An example is the PARMA(1,1) model[97]

$$y_{v,\tau} = \mu_\tau + \phi_{1,\tau}(y_{v,\tau-1} - \mu_{\tau-1}) + \varepsilon_{v,\tau} - \theta_{1,\tau}\varepsilon_{v,\tau-1} \qquad (1)$$

where μ_τ, $\phi_{1,\tau}$, $\theta_{1,\tau}$, and $\sigma_\tau(\varepsilon)$ are the model parameters. When the θ's are zeros, model (1) becomes the PARMA(1,0) or PAR(1) model. Low-order PARMA models such as PARMA(1,0) and PARMA(1,1) have been widely used for simulating monthly and weekly flows.[3,18,43,84,92,120]

PARMA models can be derived from physical/conceptual principles. Considering all hydrologic processes and parameters in the watershed varying along the year, it has been shown that seasonal streamflow falls within the family of PARMA models.[96] Alternatively, a constant parameter model with periodic independent residuals was suggested.[16] One of the desirable properties of stochastic models of seasonal streamflows is the preservation of seasonal and annual statistics. However, such dual preservation of statistics has been difficult to get with simple models such as the PAR(1) or PAR(2). For this reason in the 1970s hydrologists turned to the so-called *disaggregation* models (refer to Section 3). The major drawback of such simple PAR models to reproduce seasonal and annual statistics has been the lack of sufficient correlation structure. PARMA models having more flexible correlation structure than PAR models offer the possibility of preserving seasonal and annual statistics. Some hydrologists have argued that PARMA models have too many parameters. Yet, one cannot hope for models such as the PAR(1) to do more than it can, i.e., to reproduce simply the lag-1 month-to-month correlations while failing to reproduce correlations for longer time lags and statistics at higher orders of aggregation. An alternative for reproducing both seasonal and annual statistics is the family of multiplicative models.

Box and Jenkins[5] first suggested multiplicative models. These models have the characteristic of linking the variable $y_{v,\tau}$ with $y_{v,\tau-1}$ and $y_{v-1,\tau}$. McKerchar and Delleur[66] used multiplicative models after differencing the logarithms of the original series for simulating and forecasting monthly streamflow series. Because such multiplicative models do not take into account periodic correlations, differencing was used in an attempt to decrease or eliminate such periodicity. However, they were not able to reproduce the seasonality in the covariance structure and could not establish confidence limits of forecasts with consideration of seasonality. This problem arises because the referred multiplicative model does not include periodic parameters. A model (with periodic parameters) that can overcome the limitations mentioned above is the multiplicative PARMA model.[95] For instance, the multiplicative PARMA(1,1) $\times$ $(1.1)_\omega$ model is written as

$$\begin{aligned} z_{v,\tau} &= \Phi_{1,\tau} z_{v-1,\tau} + \phi_{1,\tau} z_{v,\tau-1} - \Phi_{1,\tau}\phi_{1,\tau} z_{v-1,\tau-1} + \varepsilon_{v,\tau} \\ &\quad - \Theta_{1,\tau}\varepsilon_{v-1,\tau} - \theta_{1,\tau}\varepsilon_{v,\tau-1} + \Theta_{1,\tau}\theta_{1,\tau}\varepsilon_{v-1,\tau-1} \end{aligned} \tag{2}$$

in which $z_{v,\tau} = y_{v,\tau} - \mu_\tau$ and $\Phi_{1,\tau}$, $\Theta_{1,\tau}$, $\phi_{1,\tau}$, $\theta_{1,\tau}$, and $\sigma_\tau(\varepsilon)$ are the model parameters. This model has been applied successfully for simulating the Nile River flows.

A limitation of the foregoing PARMA and multiplicative PARMA models for modeling hydrological time series is the requirement that the underlying series be transformed into normal. An alternative that does not have this requirement is the PGAR(1) model for modeling seasonal flows with periodic correlation structure and periodic gamma marginal distribution.[27] Consider that $y_{v,\tau}$ is a periodic correlated variable with a three-parameter gamma marginal distribution with location λ_τ, scale α_τ, and shape β_τ parameters varying with τ, and $\tau = 1, \ldots, \omega$ (T = number of seasons). Then, the new variable $z_{v,\tau} = y_{v,\tau} - \lambda_\tau$ is a two-parameter gamma that can be represented by $z_{v,\tau} = \phi_\tau z_{v,\tau-1} + (z_{v,\tau-1})^{\delta_\tau} w_{v,\tau}$ where ϕ_τ = periodic autoregressive coefficient, δ_τ = periodic autoregressive exponent, and $w_{v,\tau}$ = noise process. This model has a periodic correlation structure equivalent to that of the PAR(1) process. It has been applied to weekly streamflow series for several rivers in the United States.[27] Results obtained indicated that such PGAR model compares favorably with respect to the normal based models (such as the PAR model after logarithmic transformation) in reproducing the basic statistics usually considered for streamflow simulation. Furthermore, a nonparametric approach for streamflow simulation that is capable of reproducing closely historical distributions has been proposed.[100]

PARMA and PGAR models are less useful for modeling flows in ephemeral streams. In these streams the flows are intermittent, a characteristic that is not represented by the above mentioned models. Instead periodic product models such as $y_{v,\tau} = B_{v,\tau} z_{v,\tau}$ are more realistic,[94] where $B_{v,\tau}$ is a periodic correlated Bernoulli (1,0) process, $z_{v,\tau}$ may be either an uncorrelated or correlated periodic process with a given marginal distribution, and B and z are mutually uncorrelated. Properties and applications of these models for simulating intermittent monthly flows of some ephemeral streams have been reported in the literature.[14,94]

Multisite Periodic Models. In modeling seasonal streamflows at several sites, multivariate PAR and PARMA models are generally used.[6,42,92,97] For example, the multivariate PARMA(1,1) model is

$$Z_{v,\tau} = \Phi_\tau Z_{v,\tau-1} + \underline{\boldsymbol{\varepsilon}}_{v,\tau} - \Theta_\tau \underline{\boldsymbol{\varepsilon}}_{v,\tau-1} \tag{3}$$

in which $Z_{v,\tau} = Y_{v,\tau} - \underline{\boldsymbol{\mu}}_\tau$; $\underline{\boldsymbol{\mu}}_\tau$ is a column parameter vector with elements $\mu_\tau^{(1)}, \ldots, \mu_\tau^{(n)}$. Φ_τ and Θ_τ are $n \times n$ periodic parameter matrices, the noise term $\underline{\boldsymbol{\varepsilon}}_{v,\tau}$ is a column vector normally distributed with $E(\underline{\boldsymbol{\varepsilon}}_{v,\tau}) = \underline{0}$, $E(\underline{\boldsymbol{\varepsilon}}_{v,\tau} \underline{\boldsymbol{\varepsilon}}_{v,\tau}^T) = \Gamma_\tau$, and $E(\underline{\boldsymbol{\varepsilon}}_{v,\tau} \underline{\boldsymbol{\varepsilon}}_{v,\tau-k}^T) = 0$ for $k \neq 0$, and n = number of sites. In addition, it is assumed that $\underline{\boldsymbol{\varepsilon}}_{v,\tau}$ is uncorrelated with $Z_{v,\tau-1}$. Parameter estimation of this model can be made by the method of moments, although the solution is not straightforward. Dropping the moving average term in (3), i.e., $\Theta_\tau = 0$ for all τ's, yields a simpler multivariate PARMA(1,0) or PAR(1) model. This simpler model has been widely used for generating seasonal hydrologic processes. Further simplifications of the foregoing models can be made to facilitate parameter estimation. Assuming that Φ_τ and Θ_τ of Eq. (3) are diagonal matrices, the multivariate PARMA(1,1) model can be decoupled into univariate models for each site. To maintain the cross correlation among sites $\underline{\boldsymbol{\varepsilon}}_{v,\tau}$ is modeled as $\underline{\boldsymbol{\varepsilon}}_{v,\tau} = B_\tau \underline{\boldsymbol{\xi}}_{v,\tau}$ where $E(\underline{\boldsymbol{\xi}}_{v,\tau} \underline{\boldsymbol{\xi}}_{v,\tau}^T) = I$ and $E(\underline{\boldsymbol{\xi}}_{v,\tau} \underline{\boldsymbol{\xi}}_{v,\tau-k}^T) = 0$ for $k \neq 0$. This modeling scheme is a *contemporaneous* PARMA(1,1), or CPARMA(1,1), model. Useful references on this type of models are available in the literature.[42,84,92,97]

Annual Streamflows

Autoregressive (AR) and autoregressive and moving average (ARMA) models have been the most popular models for single site and multisite annual streamflow simulation. Specifically, low-order models have been widely applied for generating annual flow series.[29,42,61,62,72,92]

Single-Site Stationary Models. The AR(1) model is defined as $y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t$. Its autocorrelation function $\rho_k = \phi\rho_{k-1} = \phi^k$ decays exponentially as the time lag k increases. This model has been a prototype of *short memory* models because ρ_k goes to zero relatively fast and as a result $h \to \frac{1}{2}$ rather quickly in $E(R_n^{**}) \sim n^h$ (R_n^{**} = rescaled range of cumulative departures from the sample mean). A more versatile model than the AR(1) is the ARMA(1,1) given by[6,42,97]

$$y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t - \theta\varepsilon_{t-1} \tag{4}$$

Its autocorrelation function $\rho_k = (1 - \phi\theta)(\phi - \theta)(1 - 2\phi\theta + \theta^2)^{-1}\phi^{k-1}$ is more flexible than that of the AR(1) model because it depends on the two parameters ϕ and θ. The ARMA process can represent *long memory* dependence,[72,91] a property that is important for many rivers. AR and ARMA models assume that the underlying series is normally distributed, an assumption that is not always applicable for annual streamflow series. While one can circumvent this assumption by transforming the skewed series into an approximately normal series, a direct approach that does not require a transformation is a viable alternative. The *gamma autoregressive* (GAR) process $y_t = \lambda(1 - \phi) + \phi y_{t-1} + \eta_t$ offers such an alternative where y_t is gamma distributed with parameters λ, α, and β (the location, scale, and shape parameters, respectively), ϕ = autoregressive coefficient, and η_t = noise term. The GAR(1) model has the same autocorrelation function as that of an AR(1) model. Estimation procedures and applications of the GAR model for simulating annual streamflow series can be found in the literature.[28]

AR, ARMA, and GAR models are useful for modeling streamflow processes in perennial rivers, yet they are inadequate for *intermittent processes* such as streamflows in some ephemeral streams. Intermittent processes can be modeled as $y_t = B_t z_t$ where y_t = non-negative intermittent variable, B_t = dependent (1,0) *Bernoulli process*, z_t = positive valued continuous autocorrelated variable, for instance, an AR(1) process, and B_t and z_t are assumed to be mutually uncorrelated. Thus, the resulting product process y_t is intermittent and autoregressive. These models have been applied for modeling short-term rainfall and intermittent flow processes.[7,13,94]

Finally, other type of models, such as fractional Gaussian noise.[64] broken line,[6] shifting level,[93] and FARMA[42] have been proposed for representing certain special properties of annual streamflow time series. For example, the shifting level model has the capability of simulating time series with sudden changes, a property that has been observed in many hydroclimatic processes.

Multisite Stationary Models. Modeling of multiple time series is widely needed in hydrology. Consider the column vector Y_t with elements $y_t^{(1)}, \ldots, y_t^{(n)}$ in which n = the number of series (number of variables) under consideration. The multivariate AR(1) model is defined as[65]

$$Z_t = \Phi Z_{t-1} + \underline{\varepsilon}_t \tag{5}$$

in which $Z_t = Y_t - \underline{\mu}$, $\underline{\mu}$ is a column vector of means $\mu^{(1)}, \ldots, \mu^{(n)}$, $\underline{\varepsilon}_t$ is a column vector of normal noises $\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(n)}$, each with zero mean such that $E(\underline{\varepsilon}_t \underline{\varepsilon}_t^T) = \Gamma$ and $E(\underline{\varepsilon}_t \underline{\varepsilon}_{t-k}^T) = 0$ for $k \neq 0$, and Φ and Γ are $n \times n$ parameter matrices. In addition, it is assumed that $\underline{\varepsilon}_t$ is uncorrelated with Z_{t-1}. Model (5) is a prototype of short-memory models for multiple series and has been widely used in operational hydrology.[29,42,62,92] Likewise, the *multivariate* ARMA(1,1) model can be written as in Eq. (3) except that the parameters Φ and Θ do not depend on time.

Except for low-order multivariate AR models, using the full multivariate ARMA models often leads to complex parameter estimation.[73,92] Thus, model simplifications have been suggested. For instance, a *contemporaneous* ARMA (CARMA) model results if Φ and Θ are diagonal matrices. This concept, which has been advocated by Salas et al.,[92] Stedinger et al.,[106] and Hipel and McLeod[42] can be extended to the general case. A contemporaneous relationship implies that only the dependence of concurrent values of the y's are considered important. Furthermore, the diagonalization of the parameter matrices allows "model decoupling" into component univariate models so that the model parameters do not have to be estimated jointly, and univariate modeling procedures can be employed. Thus, univariate ARMA(p, q) models are fitted at each site where each $\varepsilon_t^{(i)}$, $i = 1, \ldots, n$ is uncorrelated, but are contemporaneously correlated with a variance–covariance matrix Γ. Thus, the parameters, ϕ's and θ's in each model, can be estimated by using univariate estimation procedures and the ε's can be modeled by $\underline{\varepsilon}_t = B\xi_t$ in which $\underline{\xi}$ is normal with $E(\underline{\xi}_t\underline{\xi}_t^T) = I$ and $E(\underline{\xi}_t\underline{\xi}_{t-k}^T) = 0$ for $k \neq 0$. Note that one does not have to consider the same univariate ARMA(p, q) model for each site.

3 TEMPORAL AND SPATIAL DISAGGREGATION MODELS

Disaggregation models, i.e., downscaling models in time and/or space, have been an important part of stochastic hydrology, not only because of our scientific interest in understanding and describing the characteristics of the spatial and temporal variability of hydrological processes, but also because of practical engineering applications. For example, many hydrologic design and operational problems require hourly precipitation data. Because hourly precipitation data are not as commonly available as daily data, a typical problem has been to *downscale* or *disaggregate* daily data into hourly data. Similarly, for simplifying the analysis and modeling of large-scale systems involving a large number of precipitation and streamflow stations, temporal and spatial disaggregation procedures are needed. This section briefly reviews some empirical and mathematical models and procedures for temporal and spatial disaggregation of precipitation and streamflow.

Disaggregation of Precipitation

Generally the disaggregation of station precipitation data defined at a given time interval into precipitation for smaller time intervals has been done empirically.[74] For instance, by using either tables or graphs, one can do disaggregation of 24-h (daily) precipitation into 6-h precipitation. More complete disaggregation schemes has been developed.[41,119]. Hershenhorn and Woolhiser[41] considered daily rainfall amounts and a model to obtain within-the-day magnitudes for the number of storms, amount, duration, and arrival time for each storm. They indicated that simulated rainfall sequences compared well with observed values. Although the foregoing models are innovative, they are not satisfactory, i.e., they are complex and require many transformations of the original data to obtain reasonable results. Another shortcoming is the lack of flexibility in the number of intervals considered.

Another formal disaggregating scheme for short-term rainfall was developed by Cadavid et al.[11] Disaggregation models were developed assuming PWN and NSWN (refer to Section 1) as the underlying rainfall-generating mechanisms. Formulation of the disaggregation algorithm for the PWN model is based on the distribution of the number of arrivals N conditioned on the total precipitation Y in the time interval, the distribution of the white noise term given N and Y, and the distribution of the arrival times conditional on N. The algorithm performs well when using simulated PWN samples. The disaggregation scheme based on the NSWN model is more complex. It performs well on simulated and recorded samples provided that the model parameters used are similar to those controlling the process at the disaggregation scale. The main shortcoming is the incompatibility of parameter estimates at different aggregation levels as pointed out in Section 1. Recently, a rainfall disaggregation based on artificial neural networks has been suggested.[8]

Epstein and Ramírez[24] developed a multiscale, linear regression, statistical climate inversion scheme based on the disaggregation model of Valencia and Schaake[111] given by

$$Y = AX + B\varepsilon \tag{6}$$

where Y is a matrix of downscaled hydroclimatic values (e.g., precipitation), X is a matrix of upscaled hydroclimatic values, A and B are parameter matrices, and ε is a matrix of independent standard normal deviates. All terms in the above equation are functions of time, and the downscaling model is conditioned on time through the temporal evolution of the large-scale field, X. Parameter estimation, based on the method of moments, leads to the preservation of the first- and second-order moments at all levels of aggregation.

Disaggregation of Streamflow Data

The shortcoming of low-order PAR models when applied for simulating seasonal flows in reproducing the annual flow statistics led to the development of disaggregation models such as the Valencia–Schaake model (6). In this model, the modeling and simulation of seasonal flows is accomplished in two or more steps. First the annual flows are modeled so as to reproduce the desired annual statistics [e.g., based on the ARMA(1,1) model]; then synthetic annual flows are generated, which in turn are disaggregated into the seasonal flows by means of Eq. (6). While the variance–covariance properties of the seasonal flow data are preserved and the generated seasonal flows add up to the annual flows, model (6) does not preserve the covariances of the first season of a year and any preceding season. To circumvent this shortcoming, Eq. (6) has been modified as $Y = AX + B\underline{\varepsilon} + CZ$, where C is an additional parameter matrix and Z is a vector of seasonal values from the previous year (usually only the last season of the previous year) for each site.[6] Further refinements and corrections assuming an annual model that reproduces S_{XX} and S_{XZ} has been suggested[57] as well as a scheme that does not depend on the annual model's structure yet reproduces the moments S_{YY}, S_{YX}, and S_{XX}.[105]

The foregoing disaggregation models have too many parameters, a problem that may be significant especially when the number of sites is large and the available historical sample size is small. Lane[56] sets to zero some of the parameters in the above disaggregation model so that

$$Y_\tau = A_\tau X + B_\tau \underline{\varepsilon} + C_\tau Y_{\tau-1} \qquad \tau = 1, \ldots, \omega \tag{7}$$

is a model with fewer parameters. Parameter estimation and appropriate adjustments so that the seasonal values add exactly to the annual values at each site can be found in the literature.[56,97]

The estimation problem can be simplified if the disaggregation is done in steps (stages or cascades) so that the size of the matrices involved and consequently the number of parameters decrease.[6] For instance, annual flows can be disaggregated into monthly flows directly in one step (this is the usual approach), or they can be disaggregated in two or more steps, e.g., into quarterly flows in a first step; then each quarterly flow is further disaggregated into monthly flows in a second step. However, even in the latter approach, considerable size of the matrices will result when the number of seasons and the number of sites are large. Santos and Salas[98] proposed a *stepwise disaggregation scheme* in such a way that at each step the disaggregation is always made into two parts or two seasons. This scheme leads to a maximum parameter matrix size of 2 × 2 for single-site disaggregation and $2n \times 2n$ for multisite. Disaggregation models that reproduce seasonal statistics and the covariance of seasonal flows with annual flows assuming log-normal seasonal and annual flows have been also suggested.[36,107] Also temporal disaggregation based on nonparametric procedures has been proposed.

Although disaggregation has been a major development and a practical tool in stochastic hydrology, still the question of why certain periodic models fail to reproduce annual statistics remained. Thus, more complex models such as PARMA models were suggested and developed in the 1970s and the early 1980s. Their capabilities for reproducing statistical properties beyond the seasons have been explored by Obeysekera and Salas[70] and Bartolini and Salas.[3]

4 TEMPORAL AND SPATIAL AGGREGATION MODELS

As in disaggregation, the *aggregation* (*upscaling*) modeling approach deals with streamflow processes at two or more levels of aggregation or time scales. However, the two concepts are quite different. In disaggregation, the modeling and generation procedure is backward in the sense that one models and generates annual flows first, then monthly, weekly, and daily flows are obtained in successive disaggregation steps. On the other hand, in temporal aggregation, the procedure is forward, i.e., one models and generates daily flows first and then successively weekly, monthly, and annual flows are modeled and generated. The basic premise of the aggregation approach is that the stochastic characteristics at the continuous time scale dictate those at any level of aggregation or time scale. The relationship between statistics at various time scales has been explored in the literature.[50]

The aggregation modeling approach for streamflow processes was developed by Vecchia et al.[112] Assuming that monthly flows follow a PAR(1) or PARMA(1,1) process, it was shown that the resulting model for the annual flows is the stationary ARMA(1,1). The foregoing concepts and results brought into light the structural linkage and compatibility between streamflow models (and their parameters) of various time scales. Streamflow data of the Niger River at Kaulikoro, Africa, were used to illustrate some of the aggregation concepts, especially in relation to reproducing the annual correlation structure when the seasonal flows were modeled by the PAR(1) and PARMA(1,1) models.[70] It was shown that in comparing the parameters and the correlograms of the ARMA(1,1) models of annual flows derived from the models of seasonal flows, the results obtained from the PARMA(1,1) model were significantly better than those obtained from the PAR(1). In addition, the results obtained vary depending on the number of seasons considered in the year (e.g., monthly, quarterly), and better results were obtained, as the number of seasons in the year became smaller.

Bartolini and Salas[3] extended the concept of aggregation to include aggregation not only from seasons to a year but from weeks to months, months to seasons, and seasons to years. For instance, the aggregation of a PARMA(2,1) monthly flows leads to a PARMA(2,2) bimonthly flows; in turn the aggregation of such PARMA(2,2) bimonthly flows gives also a PARMA(2,2) for the quarterly flows. Furthermore, if such quarterly flows are aggregated into annual flows, then the model is the stationary ARMA(2,2). The partial aggregation concepts have been applied to the Niger River seasonal and annual flows, and the results showed the superiority of the PARMA(2,1) and PARMA(2,2) models relative to the other models tested in reproducing the variance–covariance properties of the annual flows. The application of the aggregation concepts for modeling the seasonal and annual flows of the Niger River at various time scales suggest the need for using PARMA models for streamflow modeling and simulation if one would like to reproduce seasonal and annual first- and second-order statistics. The traditional models such as the PAR(1) simply are deficient for modeling flows such as those of the Niger. The usual approach to model seasonal and annual flows has been to use different models for different time scales, disregarding the compatibility among models at various time scales. The aggregation concepts and results discussed in this section point out that such traditional approaches and models for streamflow simulation must be avoided.

Similar reasoning as in temporal aggregation applies for spatial aggregation of streamflow processes. For instance, one can assume a stream network composed of a number of first-, second-, and third-order streams. Consider first modeling the flows at the junction of two first-order streams. Naturally, the model at a site immediately downstream from the junction must be derived from the bivariate model defined at two upstream sites, one in each of the tributaries. In turn, the model of the flows of the second-order stream as it joins the flows of another, say, second-order stream must define the model of the flows immediately downstream from the junction, and so on as the streamflow travels down the stream network. Thus, streamflow models must be compatible in both temporal and spatial scales.

5 SCALING ISSUES AND DOWNSCALING

Understanding, describing, and modeling local, regional, and global climate and its nonlinear interactions with hydrological, biophysical, and biogeochemical processes are currently some of the most challenging problems in the geosciences. It is not only the high spatial and temporal variability of the governing processes and boundary conditions but the wide range of scales over which this variability occurs. Distributed hydrologic models require high-resolution input data. Among all hydrologic variables, precipitation is of paramount importance in the water and energy budgets at the land surface–atmosphere interface, and its accurate representation in hydrologic and atmospheric models is critical. Precipitation is the result of intricately interrelated atmospheric and land surface processes. It has extreme variability over time scales from seconds to years and spatial scales from less than meters to hundreds of kilometers. The sensitivity of hydrologic behavior to the space–time variability of rainfall is the result of nonlinear interactions between precipitation and land surface characteristics controlling the transformation of rainfall into soil water and runoff. Modeling of precipitation requires understanding of the statistical structure of space–time precipitation and understanding of the physical processes governing the evolution of precipitation at a range of space–time scales. Because of scale differences, rainfall downscaling is required for coupling global (or regional) atmospheric models and hydrologic models. In general, downscaling schemes can be grouped in two broad categories, *dynamical* and *statistical* downscaling schemes.

In *dynamic* schemes, climate and land-use change scenarios at regional and local scales are developed using regional and local atmospheric models, for example, the Regional Atmospheric Modeling System (RAMS) of Colorado State University. These models are driven by boundary conditions derived from observations and from the output of global atmospheric models. In this way, the atmospheric model acts as a physically based dynamic interpolator (i.e., physically based downscaling). RAMS has been coupled to a land surface scheme (LEAF-2), to a hydrologic model (TOPMODEL), and to a regional ecosystem model (CENTURY). Thus, dynamical schemes encode multiple, nonlinear and complex local and regional interactions and feedbacks, explicitly.[79,115] However, trying to resolve processes at ever decreasing scales using physically based models rapidly leads to computational inefficiencies and is limited by poor understanding of physical process behavior at small scales. Other atmospheric models such as the NCAR models[33] and the Penn State/NCAR mesoscale model, Version 5,[19] have been used for dynamic downscaling.

In *statistical downscaling*, subgrid temporal and spatial scale details of climatic variability, in particular precipitation, are obtained so that the statistical characteristics of the spatial and temporal variability of hydroclimatic fields is preserved as a function of scale. *Statistical* techniques are commonly based on linear or nonlinear regression, methods from nonlinear dynamics, artificial neural networks, Markov processes, multiplicative random cascade models, etc. One of the limitations of regression approaches is that they are applicable only if a strong relationship between a large-scale parameter and regional and local climate has been identified (often this will not be the case), and they are valid only within the spatial and temporal range of the observations. Although statistical downscaling is computationally efficient, it cannot include the subgrid scale physical feedbacks referred to above explicitly, and it is difficult to couple atmospheric processes with regional ecological and hydrological processes. On the other hand, statistical downscaling based on multiplicative random cascade models can reproduce the scaling features (i.e., scale invariance), the clustering, and intermittency that are characteristic of precipitation fields in space and time with relatively modest computational burden.

Until recently, most of the downscaling methods proposed in the literature only dealt with the spatial variability of the precipitation field while the temporal evolution of the fields is usually described independently of the spatial downscaling. The only temporal correlation structure accounted for is that resulting from the dynamics of the atmospheric model producing the precipitation field at the larger spatial scale or that encoded in the temporal evolution of the observations. Thus, in general, these schemes do not fully and properly account for the temporal correlation structure (i.e., persistence) of the precipitation fields at subgrid scales.

Dynamical Downscaling

Dynamical downscaling can be considered with respect to four basic types of models: one type is strongly dependent on larger-scale numerical weather prediction lateral boundary conditions, bottom boundary conditions, and on initial conditions. A second type has forgotten the initial conditions but is dependent on the observed lateral and bottom boundary conditions. A third type is where a large-scale model is run that is only forced with surface boundary conditions, and the output used to downscale with a regional model. A fourth type is when a true global climate model (with coupled ocean, atmosphere, continental sea ice, landscape processes, etc.) is used to provide lateral boundary conditions to a regional model. This is the Intergovernmental Panel on Climate Change (IPCC) type of downscaling except only a limited set of Earth system forcings (e.g., the radiative effect of CO_2, solar insolation) is included in the IPCC approach. To summarize with examples (IC = initial conditions; LBC = lateral boundary conditions, and BBC = bottom boundary conditions; with the recognition that BBC includes bottom interfacial fluxes): type 1 ETA[4] uses observed IC, LBC, and BBC; type 2 PIRCS[34,35] uses observed LBC and BBC; type 3 ClimRAMS forced by CCM3 integrated with observed SSTS[102] uses observed BBC; and type 4 Earth system global model downscaled using a regional model.[45] Observational constraints on the solution become less as we move from type 1 to 4. Thus forecast skill will diminish from type 1 to 4.

With respect to current generation models, such as atmospheric–ocean global circulation models (AOGCMs), neither AOGCMs nor the regional ones (type 4 models) include all of the significant human effects on the climate system. The combined effects of land-use change, the biogeochemical effect on the atmosphere, e.g., due to increased CO_2, and, e.g., the microphysical effect of pollution aerosols have not yet been included in these models. Thus the existing model runs should only be interpreted as sensitivity experiments, not forecasts, projections, or even scenarios.[80]

In addition, with respect to dynamic downscaling, as currently applied, there is not a feedback upscale to the AOGCM from the regional model, even if all of the significant large-scale (GCM scale) human-caused disturbances were included. The AOGCM also has a spatial resolution that is inadequate to properly define the lateral boundary conditions of the regional model. Anthes and Warner[2] show that the lateral boundary conditions are the dominant forcing of regional atmospheric models as associated with propagating features in the polar westerlies. With numerical weather prediction (type 1 and 2 models), the observations used in the analysis to initialize a model retain a component of realism even when degraded to the coarser model resolution of a global model. This realism persists for a period of up to a week or so, when used as lateral boundary conditions for a regional numerical weather prediction model. This is not true with the AOGCMs where data do not exist to influence the predictions. A regional model cannot reinsert model skill, when it is so dependent on lateral boundary conditions, no matter how good the regional model.

Statistical Downscaling

The output of mesoscale atmospheric models such as RAMS or the observations data such as those from the NEXRAD (next generation radar) network are usually at grid sizes that are larger [e.g., $O(10^3$ to $10^4)$ m] than those associated with distributed hydrologic models (e.g., in the order of 10^0 m). The land surface system responds to excitations from the atmosphere, e.g., precipitation, and feeds back moisture into the atmosphere, e.g., through evapotranspiration and latent heat flux. The excitations and responses are spatially heterogeneous over a broad range of scales, including subgrid scales [e.g., $< O(10^4)$m]. This subgrid spatial variability has a significant impact on the magnitude and distribution of upscale and downscale land surface fluxes whose interaction is nonlinear. Accounting for this space–time heterogeneity is important for hydrologic modeling and for describing land surface–atmosphere interactions.[23,78,83] In addition, the statistical downscaling, besides requiring that the AOGCMs are accurate predictions of the future, also require that the statistical equations used for downscaling remain invariant under changed regional atmospheric and land surface conditions. There is no way to test this hypothesis. In fact, it is unlikely to be valid since the regional climate is not passive to larger-scale climate conditions but is expected to change over time and feedback to the larger scales. More details of this concern regarding downscaling have been reported.[81]

Regression Schemes. The relationships between large-scale and local-scale climatic fields can be established by regression-based schemes. The most direct way of downscaling is by direct interpolation. This method is easy to apply and effective for smoothly varying fields such as sea level pressure or temperature but not appropriate for nonsmooth intermittent fields such as precipitation. Some examples on regression schemes are (1) a method (based on principal component analysis, canonical correlation, and regression analysis) called climatological projection by model statistics to relate general circulation model (GCM) grid point free atmosphere statistics to important surface observations;[47] (2) use of interannual variations in climate to derive, through conventional regression analysis, statistical relationships between large-scale climate variations and local values of temperature and precipitation;[117] (3) a multiscale, linear regression, statistical climate inversion scheme that preserves all first- and second-order moments across scales (spatial downscaling of precipitation and temperature are applied in the context of impact assessment studies associated with global climate variability[24]); and (4) methods for conditional stochastic generation of rain fields used for disaggregation when constrained to large-scale values that are given by some outer sources.[58, 67]

Scale Invariance Scheme. Self-similarity or scale invariance is a kind of symmetry observed in nature. *Simple scaling* is a scaling in the probability distributions. Letting $R(\,)$ be the rainfall intensity field, this property is expressed as $R(\lambda A) \overset{\text{dist}}{=} \lambda^{-\theta} R(A)$, which indicates that the probability density function of the rescaled variable $R(\lambda A)$ is equal to that of the original variable $R(A)$ except for a factor that is a function of the length scale ratio λ and the scaling exponent θ. Simple scaling translates into two properties: (1) a log-log linearity between statistical moments of order n and length scale λ, i.e., $E[R_\lambda^n] = \lambda^{\theta(n)} E[R_1^n]$ and (2) a linear dependence on n of the slope $\theta(n)$ of that log-log linear relationship, i.e., $\theta(n) = n\theta$. Simple scaling is associated with additive (linear) processes, and unique scaling exponents θ are related to unique fractal dimensions. For example, Figures 2 and 3 are the scaling plots for the NEXRAD rainfall scan of July 4, 1997, for the central United States that is produced at grid scales of 2 km×2 km.[46] They show that the precipitation data for this date and region follow a simple scaling. It is often found that property (1) holds but the slope function $\theta(n)$ is nonlinear, a structure called *multiscaling* or *multifractal*.[37] Multifractal scaling behavior (i.e., scale-invariant

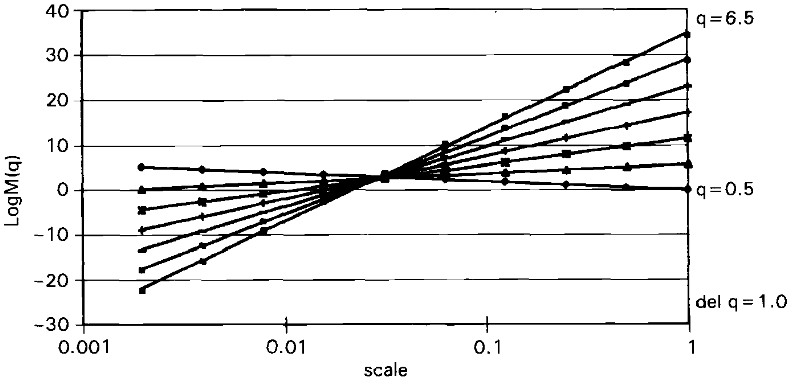


Figure 2 Marginal moments of precipitation over the central United States on July 4, 1997, from NEXRAD scan at 2 km×2 km grid size (from Kang and Ramirez[46]).

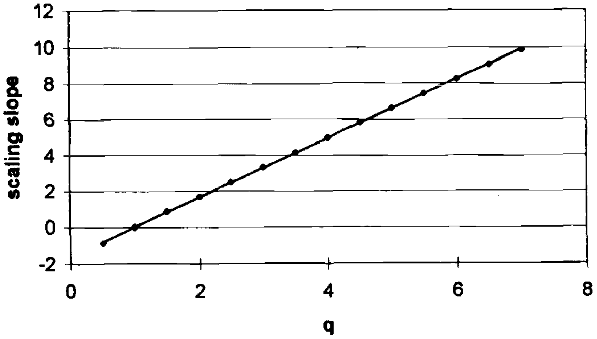


Figure 3 Slope of marginal moments log-log scaling function of NEXRAD scan of precipitation over the central United States on July 4, 1997, at 2 km×2 km grid size (from Kang and Ramirez[46]).

behavior) of the scaling exponents has been found in the spatial distribution of rainfall[37,75,109] and in the temporal distribution of rainfall.[12] Also both the scale invariance and intermittency of precipitation may be exploited to develop parsimonious stochastic models of rain.[63]

Multiplicative random cascades have been used to generate fractal fields that emulate the spectrum of scaling exponents of observed rainfall. Cascade generators are chosen according to the scaling spectra they produce. Notably, models such as the *universal multifractals*,[109] the β-model[38] or the log-Poisson model[101] were proposed. For illustration, we will discuss below random cascade models.[38,46,75]

Discrete random cascades distribute mass on successive regular subdivisions of a d-dimensional cube. A schematic of this process is shown in Figure 4. The initial cube, with length scale L_0, is subdivided at each level into b equal parts, where $b \geq 2^d$ is the branching number. The ith subcube after n levels of subdivision is denoted Δ_n^i (there are $i = 1, \ldots, b^n$ subcubes at level n). The length scale of the subcube Δ_n^i is denoted L_n, and the dimensionless spatial scale is defined as $\lambda_n = L_n/L_0 = b^{-n/d}$. The distribution of mass through different levels on the cubes occurs as follows. First the initial cube (at level $n = 0$) is assigned a nonrandom density R_0, i.e., an initial mass $R_0 L_0^d$. The subcubes Δ_1^i, $i = 1, \ldots, b$ after the first subdivision (at level $n = 1$) are assigned the density $R_0 W_1(i)$, i.e., mass $R_0 L_1^d W_1(i)$, where W are independent and identically distributed (iid) random variables—the *cascade generator*. This multiplicative process continues through all n levels of the cascade, so that the mass in subcube Δ_n^i is

$$\mu_n(\Delta_n^i) = R_0 L_n^d \prod_{j=1}^{n} W_j(i) = R_0 L_0^d b^n \prod_{j=1}^{n} W_j(i) \qquad \text{for } i = 1, 2, \ldots, b^n$$

where $E[W] = 1$; thus mass is on the average conserved at all levels in the random cascade.

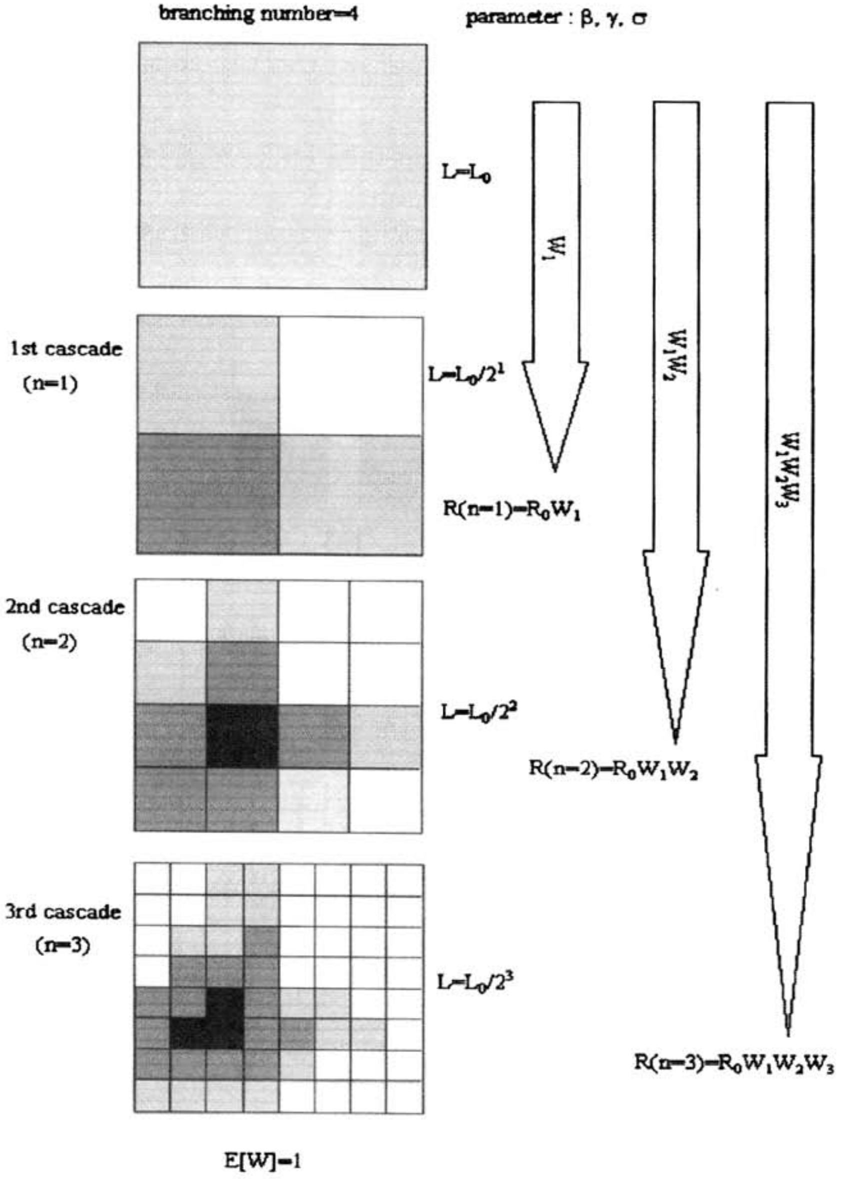
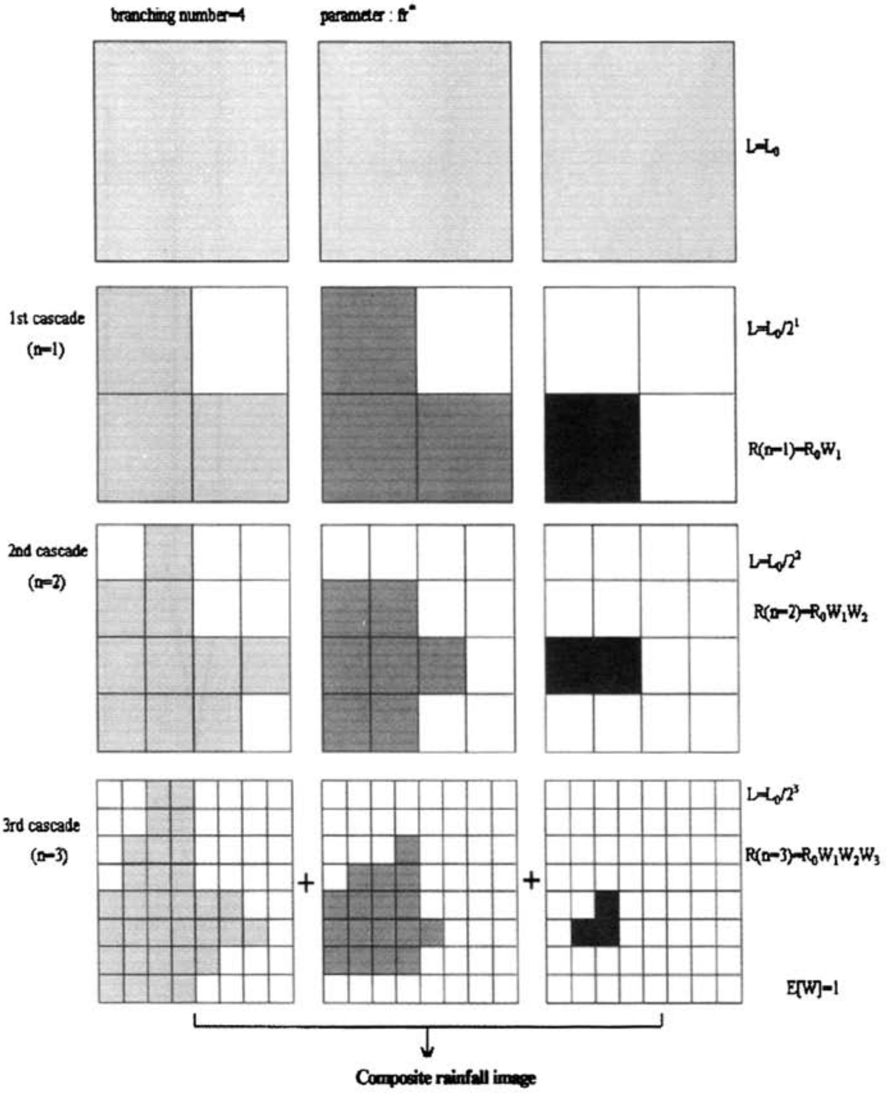


Figure 4 Schematic diagram for a two-dimensional discrete random cascade models: (*a*) Discrete random cascade model (from Kang and Ramirez[46]).



(b) nonparametric hierarchical random cascade model

Figure 4 Schematic diagram for a two-dimensional discrete random cascade models: (*b*) nonparametric random cascade model (from Kang and Ramirez[46]).

The cascade limit mass μ_∞ is obtained as $n \to \infty$, and it is considered degenerate if the total mass is zero with probability 1. Nondegeneracy depends on the distribution of W, and it requires that the condition $E[W] = 1$ be satisfied. The limit mass in a subcube $\mu_\infty(\Delta_n^i)$ satisfies a recursion relation:[75] $\mu_\infty(\Delta_n^i) = \mu_n(\Delta_n^i)Z_\infty(i)$, for $i = 1, \ldots, b^n$ where Z_∞ are iid random variables, distributed as $Z_\infty = \mu_\infty(\Delta_0)/$

$\mu_0(\Delta_0) = \mu_\infty(\Delta_0)/R_0L_0^d$ for all i, n. The cascade limit mass $\mu_\infty(\Delta_n^i)$ is given by the product of a large-scale low-frequency component $\mu_n(\Delta_n^i)$, and a subgrid subgrid (i.e., subcube) scale high-frequency component $Z_\infty(i)$. The latter term represents subgrid-scale variability at each level of cascade development.

Random cascades exhibit *moment scaling* behavior from which properties of the cascade generator W can be estimated. Sample spatial moments are defined as $M_n(q) = \sum_{i=1}^{b^n} \mu_\infty^q(\Delta_n^i)$, where $q =$ moment order (for $q = 0$, only nonzero limit masses are included in the sum). For large n, the sample moments should converge to the ensemble moments, but since they diverge to infinity or converge to zero as $n \rightarrow \infty$, the rate of convergence/divergence of the moments with scale is considered instead. In a random cascade, the ensemble moments are shown to be a log-log linear function of the scale λ_n. The slope of this scaling relationship is known as the Mandelbrot–Kahane–Peyriere (MKP) function: $\chi_b(q) = 1 - q + \log_b E[W^q]$. The MKP function contains important information about the distribution of the cascade generator W and thus characterizes the scaling properties of rainfall. Similarly, the slope of the sample moment scaling relationship can be defined as $\tau(q) = \lim_{\lambda_n \rightarrow 0}[\log M_n(q)/ - \log \lambda_n]$. For large n (as $\lambda_n \rightarrow 0$) and for a specific range of q, slopes of the moment scaling relationships for sample and ensemble moments converge, i.e., $\tau(q) = d\chi_b(q)$. In data analysis, the scaling of the sample moments is used to estimate the $\tau(q)$ function and the distribution of the cascade generator, from which parameters of the cascade model can be inferred.

For intermittent temporal and spatial rainfall data, it is desirable that $P(W = 0)$ be positive. For this purpose an intermittency model for the cascade generator W is written as $W = BY$, where B is an intermittency generator of the so-called β-model and Y is a strictly positive random variable. The β model divides the domain into rainy and nonrainy fractions based on the following probabilities: $P(B = 0) = 1 - b^{-\beta}$ and $P(B = b^\beta) = b^{-\beta}$, where β is a parameter and $E[B] = 1$. The β model does not allow for variability in the positive part of the large-scale component of the limit mass $\mu_n(\Delta_n^i)$ (at every level n it assumes the nonrandom value $R_0L_ndb^{\beta n}$). Variability in the positive part of the limit mass is obtained from the second element in the composite generator Y. The distribution of Y is arbitrary, but it has to be positive and $E[Y] = 1$. For rainfall modeling, good results have been obtained with Y log-normal.[75] Consider $Y = b^{\gamma+\sigma X}$, where X is a normal $N(0, 1)$ random variable. The condition $E[Y] = 1$ gives $Y = b^{-\sigma^2 \ln b/2+\sigma X}$, where $\sigma^2 =$ variance of Y. Then W is distributed as $P(W = 0) = 1 - b^{-\beta}$ and $P(W = b^\beta Y = b^{\beta-\sigma^2 \ln b/2+\sigma X}) = b^{-\beta}$ with parameters β and σ^2.

The MKP function of W is $\chi_b(q) = (\beta - 1)(q - 1) + (\frac{1}{2})\sigma^2 \ln b(q^2 - q)$. MKP functions must be convex and $\chi_b(1) = 0$.[75] Also for the cascade to be nondegenerate it is required that $\chi_b^{(1)} < 0$. For large σ, this requirement may not be met, leading to degenerate cascades. The range of moments q in nondegenerate cascades is given by $[0, q_c/2)$ when $q_c \leq 2$, and at least in the closed interval $[0,1]$ when $q_c > 2$. The latter condition implies that $\beta < 1 - \sigma^2 \ln b$ in nondegenerate cascades.[75] Figure 5 shows the relationship between the parameters of the random cascade model and the large-scale forcing, here described in terms of the large-scale average rainfall intensity, for the NEXRAD data.[46]

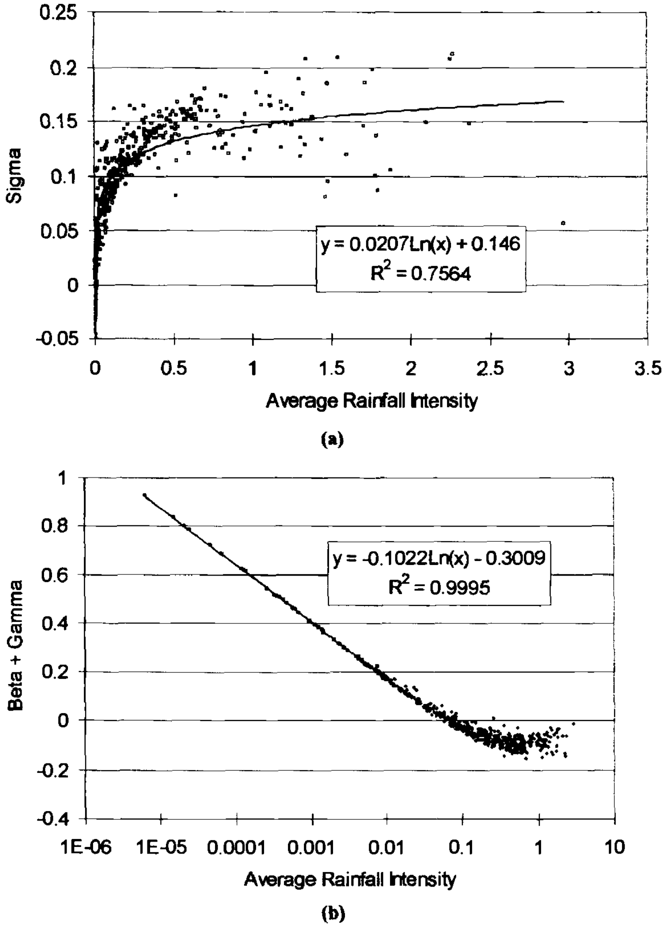
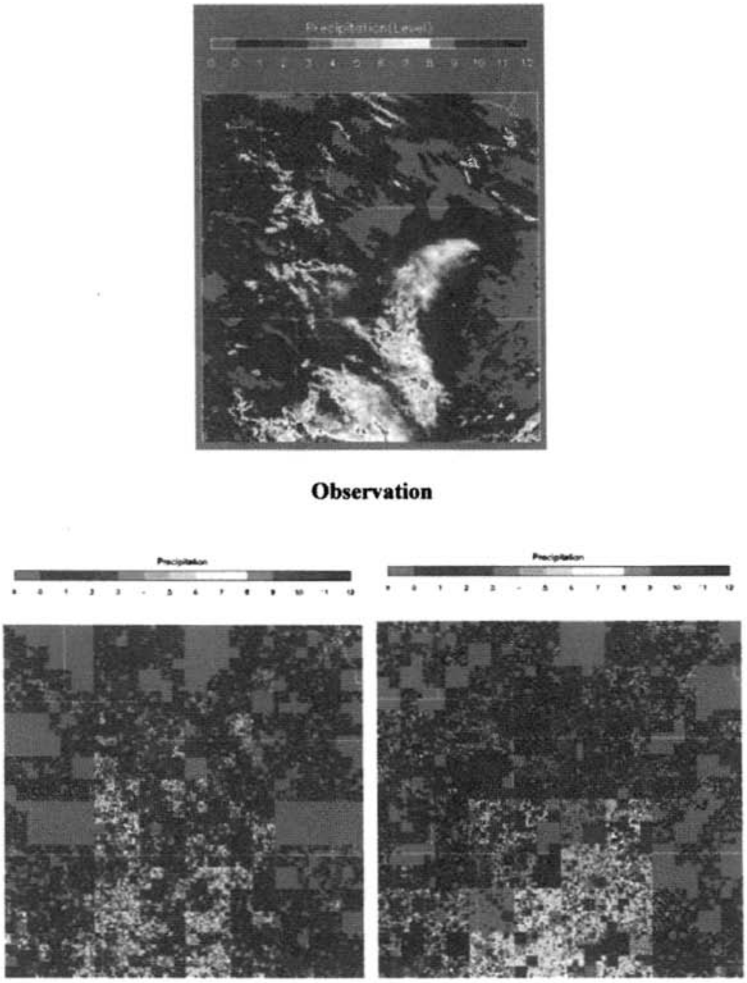


Figure 5 (*a*) Sigma vs. average rainfall intensity and (*b*) Beta + Gamma vs. average rainfall intensity (from Kang and Ramirez[46]). See ftp site for color image.

Because of the nondegeneracy condition on β, the β model is limited in the range of fractional rainy area values that it can describe. To deal with this limitation and to improve the preservation of the clustering structure, a simple pragmatic modification of the β model has been proposed, and the resulting scheme is referred to as a *nonparametric hierarchical* scheme. This procedure performs spatial downscaling by hierarchically applying a sequence of random cascade models to a set of precipitation fields defined in terms of intensity classes of the observations. Figure 6 shows a comparison of observed and simulated precipitation fields, using the β-lognormal random cascade model and the nonparametric hierarchical scheme. Figures 7



Observation

β-lognormal model **Nonparametric hierarchical model**

Figure 6 (see color insert) Comparison of observed and downscaled rainfall fields (July 6, 1997) (from Kang and Ramirez[46]). See ftp site for color image.

to 9 show a comparison of the ability of the respective models to reproduce important characteristics of the precipitation field. Table 2 presents a comparison of some summary statistics.

Further Remarks

The parameters of the scaling characteristics of the precipitation field must be related to some physically meaningful and measurable variable characterizing the large-

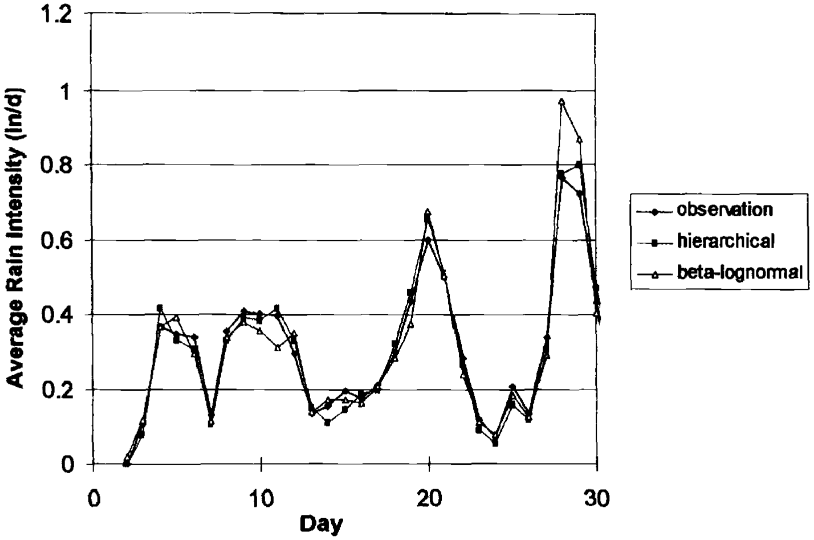


Figure 7 Time series of average rainfall intensity (NEXRAD, July 1997) (from Kang and Ramirez[46]). See ftp site for color image.

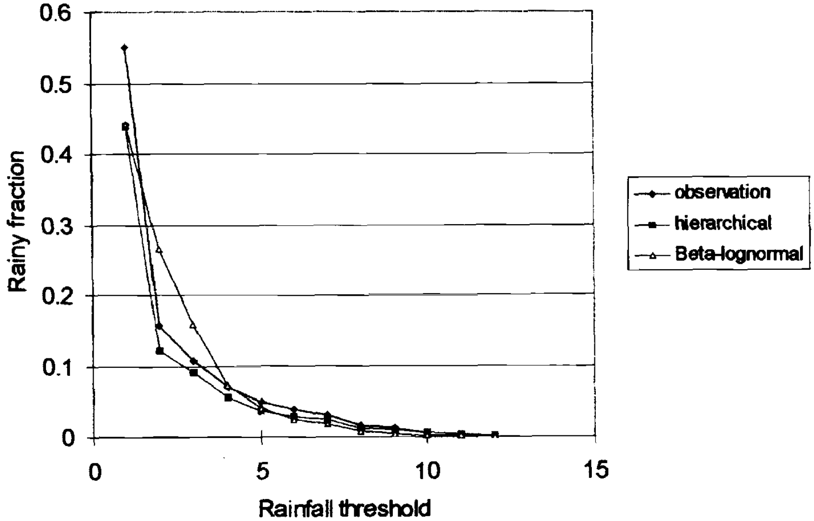


Figure 8 Rainy fraction as a function of rainfall threshold (NEXRAD, July 6, 1997) (from Kand and Ramirez[46]). See ftp site for color image.

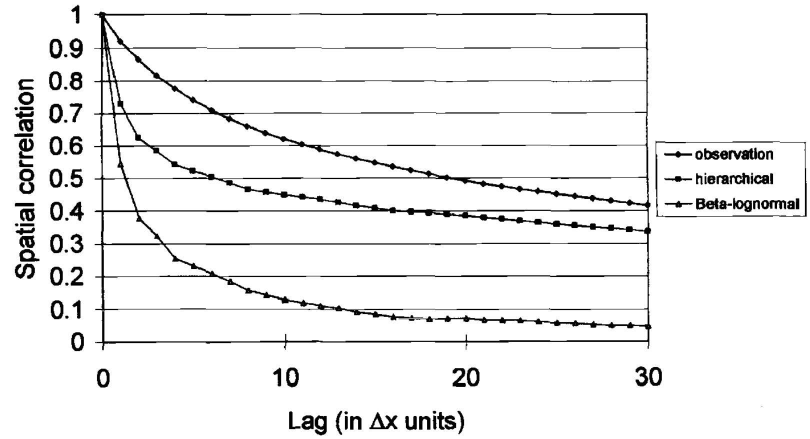


Figure 9 Comparison of correlation function for monthly rainfall (from Kang and Ramirez[46]). See ftp site for color image.

scale environment. Being able to parameterize the scaling characteristics of precipitation as a function of such variables is a prerequisite for implementing of downscaling methodologies based on random cascades. Perica and Foufoula-Georgiou[77] introduced a spatial downscaling scheme that is able to statistically reproduce the spatial heterogeneity of observed precipitation fields at subgrid scales while being conditioned on large-scale averages and physical properties. They computed multiscale standardized fluctuations using an orthogonal Haar wavelet decomposition and found that, at least for the range of scales of their analysis, these fluctuations exhibited normality and simple scaling. They also found that the scale-independent parameter H characterizing the simple scaling behavior of the standardized fluctuations was strongly dependent on the convective instability of the prestorm environment, namely on the convective available potential energy. The utility of the model in reproducing the small-scale statistical variability of precipitation as well as the fraction of area covered by rain at all subgrid scales was demonstrated,[77] and the relationship between H and the convective available potential energy of the prestorm

TABLE 2 Comparison of Basic Statistics for Observed (July 1997) and Simulated Precipitation Fields

	Observation	Nonparametric Superposition	Log-normal
Mean	0.296	0.295	0.300
Standard deviation	0.0257	0.0221	0.544
Skewness coefficient	1.174	1.185	16.78
Kurtosis	8.624	6.207	654.6

From Kang and Ramirez.[46]

environment established.[77] On the other hand, the relationship between the β-lognormal random cascade model parameters and the mean of the large-scale precipitation intensity was also observed and established.[46,75]

Most downscaling methodologies proposed in the literature only deal with the spatial variability of the precipitation field. The temporal evolution of the fields is usually described independently of the spatial downscaling, so that these schemes do not properly account for the temporal correlation structure, i.e., persistence, of the precipitation fields at subgrid scales. Recently, the linkage between the spatial and temporal scaling of precipitation fields has been explicitly addressed.[12,75,113,114] Over and Gupta[75] propose a model for space–time description of rainfall distributions based on multiplicative random cascades with independent weights in space, which are time varying according to an imposed structure. Carsteanu and Foufoula-Georgiou[12] argue that space and time variations of rainfall are necessarily connected. They postulate and experimentally verify a Taylor-like hypothesis stating that the power law variation for the moments is the same in time and space. Venugopal et al.[114] found that for spatial scales of 2 to 30 km and for temporal scales of 10 min to several hours, the evolution of precipitation remained statistically invariant under a transformation of the type $t \sim L^z$, where z is a so-called dynamic scaling exponent. That is, they found that the space–time organization of rainfall fields is scale-invariant and that its characteristics can be obtained by a simple renormalization of the space and time coordinates as implied by the $t \sim L^z$ transformation. They used the above results to develop a space–time precipitation downscaling scheme that is capable of preserving not only the spatial correlation of precipitation but also the temporal correlation at subgrid scales.

Finally, Seed et al.[99] have modeled the space–time behavior of radar precipitation using a multiplicative bounded (multifractal) cascade, each level of which was linked to the same level at the next time step via a different ARMA(1,1) model. Also Pegram and Clothier,[76] developed the so-called *string-of-beads* model in which power-law filtering of Gaussian random fields in space and time is used to capture the correlation structure of the rainfall process. Two autoregressive models, one at the image scale, the other at the pixel scale, drive the string-of-beads model. The spatial power-law filtering then ensures that the generated fields scale correctly in space and time.

REFERENCES

1. Aksoy, H., and M. Bayazit, A model for daily flows of intermittent streams, *Hydrol. Proc.*, *14*(10), 1725–1744, 2000.
2. Anthes, R. A., and T. T. Warner, Development of hydrodynamic models suitable for air pollution and other mesometeorological studies, *Monthly Weather Rev.*, *106*, 1045–1078, 1978.
3. Bartolini, P., and J. D. Salas, Modeling of streamflow processes at different time scales, *Water Resour. Res.*, *29*(8), 2573–2588, 1993.

4. Black, T. J., The new NMC mesoscale eta model: Description and forecast examples, *Weather Forecast.*, *9*, 265–278, 1994.
5. Box, G. E. P., and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco, 1976.
6. Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, MA, 1985.
7. Buishand, T. A., Some remarks on the use of daily rainfall models, *J. Hydrol.*, *36*, 295–308, 1977.
8. Burian, S. J., S. R. Durrans, S. Tomic, R. L. Pimentel, and C. N. Wai, Rainfall disaggregation using artificial neural networks, *ASCE J. Hydrol. Eng.*, *5*(3), 299–307, 2000.
9. Burlando, P., and R. Rosso, Comment on "Parameter estimation and sensitivity analysis for the modified Bartlett-Lewis Rectangular pulses model of rainfall by Islam et al., *J. Geophys. Res.*, vol. 95, no. D3, 1990, p. 2093–2100," *J. Geophys. Res.*, *96*(D5), 9391–9395, 1991.
10. Burlando, P., and R. Rosso, Stochastic models of temporal rainfall: Reproducibility, estimation and prediction of extreme events, in J. Marco Segura, R. Harboe, and J. D. Salas (Eds.), *Stochastic Hydrology and its Use in Water Resources Systems Simulation and Optimization*, Kluwer Academic Publishers, The Netherlands, 1993, pp. 137–173.
11. Cadavid, L. G., J. D. Salas, and D. C. Boes, Disaggregation of short-term precipitation records, in *Water Resources Papers*, Vol. 106, Colorado State University, Fort Collins, CO, 1992.
12. Carsteanu, A., and E. Foufoula-Georgiou, Assessing dependence among weights in a multiplicative cascade model of temporal rainfall, *J. Geophy. Res.*, *101*(D21), *26*, 363–26, 370, 1996.
13. Chang, T. J., M. L. Kavvas, and J. W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, *Water Resour. Res.*, *20*, 565–580, 1984.
14. Chebaane, M., J. D. Salas, and D. C. Boes, Product periodic autoregressive processes for modeling intermittent monthly streamflows, *Water Resour. Res.*, *32*(5), 1513–1518, 1995.
15. Chin, E. H., Modeling daily precipitation occurrence process with Markov chain, *Water Resour. Res.*, *13*(6), 949–956, 1977.
16. Claps, P., F. Rossi, and C. Vitale, Conceptual-stochastic modeling of seasonal runoff using autoregressive moving average models and different scales of aggregation, *Water Resour. Res.*, *29*(8), 2545–2559, 1993.
17. Cowpertwait, P. S. P., and P. E. O'Connell, A regionalized Neyman-Scott model of rainfall with convective and stratiform cells, *Hydrol. Earth Syst. Sci.*, *1*, 71–80, 1997.
18. Delleur, J. W., and M. L. Kavvas, Stochastic models for monthly rainfall forecasting and synthetic generation, *J. Appl. Meteor.*, *17*, 1528–1536, 1978.
19. Dudhia, J., A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and the simulation of an Atlantic cyclone and cold front, *Monthly Weather Rev.*, *121*, 1493–1513, 1993.
20. Eagleson, P., Climate, soil and vegetation, 2, The distribution of annual precipitation derived from observed storm sequences, *Water Resour. Res.*, *14*, 713–721, 1978.
21. Eltahir, E. A. B., A feedback mechanism in annual rainfall in Central Sudan, *J. Hydrol.*, *110*, 323–334, 1989.

22. Entekhabi, D., I. Rodriguez-Iturbe, and P. S. Eagleson, Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulse model: Parameter estimation validation, *Water Resour. Res.*, *25*(2), 295–302, 1989.
23. Entekhabi, D., and P. S. Eagleson, Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability, *J. Climate*, *2*(8), 816–831, 1989.
24. Epstein, D., and J. A. Ramírez, Spatial disaggregation for studies of climatic hydrologic sensitivity, *ASCE J. Hydr. Div.*, *120*(12), 1449–1467, 1994.
25. Evora, N. D., and J. R. Rousselle, Hybrid stochastic model for daily flows simulation in semiarid climates, *ASCE J. Hydrol.*, *5*(1), 33–42, 2000.
26. Ewen, J., G. Parkin, and P. E. O'Connell, SHETRAN: Distributed River Basin flow and transport modeling system, *ASCE J. Hydrol. Eng.*, *5*(3), 250–258, 2000.
27. Fernandez, B., and J. D. Salas, Periodic gamma autoregressive processes for operational hydrology, *Water Resour. Res.*, *22*(10), 1385–1396, 1986.
28. Fernandez, B., and J. D. Salas, Gamma-autoregressive models for streamflow simulation, *J. Hydr. Eng. ASCE*, *116*(11), 1403–1414, 1990.
29. Fiering, M. B., and B. B. Jackson, *Synthetic Streamflows*, Water Resources Monograph 1, American Geophysical Union (AGU), Washington, DC, 1971.
30. Foufoula-Georgiou, E., and P. Guttorp, Compatibility of continuous rainfall occurrence models with discrete rainfall observations, *Water Resour. Res.*, *22*, 1316–1322, 1986.
31. Foufoula-Georgiou, E., and D. P. Lettenmaier, A Markov renewal model of rainfall occurrences, *Water Resour. Res.*, *23*(5), 875–884, 1987.
32. Foufoula-Georgiou, E., and W. Krajewski, Recent advances in rainfall modelling, estimation and forecasting, *Rev. Geophys.*, Suppl., 1125–1137, July 1995.
33. Giorgi, F., and L. O. Mearns, Approaches to the simulation of regional climate change: A review, *Rev. Geophys.*, *29*(2), 191–216, 1991.
34. Giorgi, F., M. R. Marinucci, and G. T. Bates, Development of a second-generation regional climate model (RegCM2). Part I: Boundary-layer and radiative transfer processes, *Monthly Weather Rev.*, *121*, 2794–2813, 1993a.
35. Giorgi, F., M. R. Marinucci, and G. T. Bates, Development of a second-generation regional climate model (RegCM2). Part II: Convective processes and assimilation of lateral boundary conditions, *Monthly Weather Rev.*, *121*, 2814–2832, 1993b.
36. Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, *24*, 1574–1584, 1988.
37. Gupta, V. K., and E. Waymire, Multiscaling properties of spatial rainfall and river flow distributions, *J. Geoph. Res.*, *95*(D3), 1999–2009, 1990.
38. Gupta, V. K., and E. Waymire, A statistical analysis of mesoscale rainfall as a random cascade, *J. Appl. Meteor.*, *12*(2), 251–267, 1993.
39. Guttorp, P., *Stochastic Modeling of Scientific Data*, Chapman Hall, London, 1995.
40. Gyasi-Agyei, Y., and G. R. Willgoose, A hybrid model for point rainfall modeling, *Water Resour. Res.*, *33*(7), 1699–1706, 1997.
41. Hershenhorn, J., and D. A. Woolhiser, Disaggregation of daily rainfall, *J. Hydrol.*, *95*, 299–322, 1987.
42. Hipel, K. W., and A. I. McLeod, *Time Series Modeling of Water Resources and Environmental Systems*, Elsevier, Amsterdam, 1994.

43. Hirsch, R. M., Synthetic hydrology and water supply reliability, *Water Resour. Res.*, *15*(6), 1603–1615, 1979.

44. Hosking, J. R. M., Fractional differencing, *Biometrika*, *68*, 165–176, 1981.

45. Intergovernmental Panel on Climate Change (IPCC), *Summary for Policymakers*, report of Working Group I of the IPCC, available on-line, http://www.ipcc.ch/, 2001.

46. Kang, B., and J. A. Ramirez, Comparative study of the statistical features of random cascade models for spatial rainfall downscaling, in J. A. Ramirez (Ed.), *Proc. AGU Hydrol. Days 2001*, Hydrology Days Publications, Fort Collins, CO, 2001, pp. 151–164.

47. Karl, T. R., W. C. Wang, M. E. Schlesinger, R. W. Knight, and D. Portman, A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics. *J. Climate*, *3*, 1053–1079, 1990.

48. Katz, R. W., On some criteria for estimating the order of a Markov chain, *Technometrics*, *23*(3), 243–249, 1981.

49. Katz, R. W., and M. B. Parlange, Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*, 1331–1341, 1995.

50. Kavvas, M. L., L. J. Cote, and J. W. Delleur, Time resolution of the hydrologic time series models, *J. Hydrol.*, *32*, 347–361, 1977.

51. Kavvas, M. L., and J. W. Delleur, A stochastic cluster model of daily rainfall sequences, *Water Resour. Res.*, *17*(4), 1151–1160, 1981.

52. Kelman, J, A stochastic model for daily streamflow, *J. Hydrol.*, *47*, 235–249, 1980.

53. Koch, R. W., A stochastic streamflow model based on physical principles, *Water Resour. Res.*, *21*(4), 545–553, 1985.

54. Koepsell, R. W., and J. B. Valdes, Multidimensional rainfall parameter estimation from a sparse network, *ASCE J. Hydr. Eng.*, *117*(7), 832–850, 1991.

55. Krajewski, W. F., and J. A. Smith, Sampling properties of parameter estimators for a storm field rainfall model, *Water Resour. Res.*, *25*(9), 2067–2075, 1989.

56. Lane, W. L., *Applied Stochastic Techniques (Last Computer Package), User Manual*, Division of Planning Tech. Services, Bureau of Reclamation, Denver, CO, 1979.

57. Lane, W. L., Corrected parameters estimates for disaggregation schemes, in V. P. Singh (Ed.), *Statistical Analysis of Rainfall and Runoff*, Water Resources Publications (WRP), Littleton, CO, 1982.

58. Lanza, L. G., A conditional simulation Model of intermittent rain fields, *Hydrol. Earth Sys. Sci.* *4*(1), 173–183, 2000.

59. Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon, *Precipitation-Runoff-Modelling-System—User's Manual*, USGS Water Resour. Invest. Report, U.S. Geological Survey, 83-4238, 1983.

60. Le Cam, L. A., A stochastic description of precipitation, in J. Newman (Ed.), *Proc. IV Berkeley Symp. on Math., Statis. & Prob.*, University of Calif. Press, Berkeley, 1961, pp. 165–186.

61. Lettenmaier, D. P., and S. J. Burges, Operational assessment of hydrologic models of long-term persistence, *Water Resour. Res.*, *13*(1), 113–124, 1977.

62. Loucks, D. P., J. R., Stedinger, and D. Haith, *Water Resources Systems Planning and Analysis*, Prentice Hall, Englewood Clifts, NJ, 1981.

63. Lovejoy, S., and B. B. Mandelbrot, Fractal properties of rain and a fractal model, *Tellus*, *37A*, 209–232, 1985.

64. Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises: Part 1, Averages and variances, *Water Resour. Res.*, *5*(1), 228–241, 1969.

65. Matalas, N. C., Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, *3*(4), 937–945, 1967.

66. McKerchar, A. I. and J. W. Delleur, Application of seasonal parametric stochastic models to monthly flow data, *Water Resour. Res.*, *10*, 246–255, 1974.

67. Mellor, D., The modified turning bands (MTB) model for space-time rainfall. I. Model definition and properties, *J. Hydrol.*, *175*(1–4), 113–127, 1996.

68. Murrone, F., F. Rossi, and P. Claps, Conceptually-based shot noise modeling of streamflows at short time interval, *Stochast Hydrol. Hydraul.*, *11*(6), 483–510, 1997.

69. Neyman, J., and E. L. Scott, Statistical approach to problems of cosmology, *J. R. Stat. Soc. Ser. B*, *20*(1), 1–43, 1958.

70. Obeysekera, J. T. B., and J. D. Salas, Modeling of aggregated hydrologic series, *J. Hydrol.*, *86*, 197–219, 1986.

71. Obeysekera, J. T. B., G. Tabios, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

72. O'Connell, P. E., A simple stochastic modeling of Hurst's law, in *1971 Warsaw Symp. in Mathematical Models in Hydrology*, International Association of Hydrologic Sciences, Pub. vol. 100, No. 1, 1974, pp. 169–187.

73. O'Connell, P. E. Stochastic modeling of long-term persistence in streamflow sequences, Ph.D. dissertation, Imperial College of Science and Technology, University of London, England, 1974.

74. Ormsbee, L. E., Rainfall disaggregation model for continuous hydrologic modeling, *ASCE J. Hydraul. Eng.*, *115*(94), 507–525, 1989.

75. Over, T. M., and V. J. Gupta, A space-time theory of mesoscale rainfall using random cascades, *J. Geophys. Res.*, *101*(D21), 26319–26331, 1996.

76. Pegram, G. G. S., and A. N. Clothier, High resolution space-time modeling of rainfall: The string of beads model, WRC Report no. 752/1/99, report to the Water Research Commission, Pretoria, South Africa, 1999.

77. Perica, S. E., and E. Foufoula-Georgiou, Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions, *J. Geophys. Res. Atmos.* *101*(D21), 26347–26361, 1996.

78. Pielke, R. A., and R. Avissar, Influence of landscape structure on local and regional climate, *Landscape Ecol.*, *4*, 133–155, 1990.

79. Pielke, R. A., W. R. Cotton, R. L. Walko, C. J. Tremback, M. E. Nicholls, M. D. Moran, D. A. Wesley, T. J. Lee, and J. H. Copeland, A comprehensive meteorological modeling system—RAMS, *Meteor. Atmos. Phys.*, *49*, 69–91, 1992.

80. Pielke, Sr., R. A., Overlooked issues in the U.S. national climate and IPCC assessments, Preprints, in *11th Symp. on Global Change Studies, 80th AMS Annual Meeting*, Long Beach, CA, January 9–14, 2000, pp. 32–35.

81. Pielke, Sr., R. A., and L. Guenni, Vulnerability assessment of water resources to changing environmental conditions, *IGBP Newslett.*, *39*, 21–23, 1999.

82. Ramirez, J. A., and R. L. Bras, Conditional distributions of Neyman-Scott models for storm arrivals and their use in irrigation control, *Water Resour. Res.*, *21*, 317–330, 1985.

83. Ramírez, J. A., and S. Senarath, A statistical-Dynamical parameterization of canopy interception and land surface-atmosphere interactions, *J. Climate*, *13*, 4050–4063, 2000.

84. Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J. C. Rassam, and B. Bobee, Estimation and validation of contemporaneous PARMA models for streamflow simulation, *Water Resour. Res.*, *32*(10), 3151–3160, 1996.

85. Richardson, C. W., and D. A. Wright, *WGEN: A Model for Generating Daily Weather Variables*, U.S. Department of Agriculture, Agriculture Research Service, ARS-8, August, 1984.

86. Rodriguez–Iturbe, I., V. K. Gupta, and E. Waymire, Scale considerations in the modeling of temporal rainfall, *Water Resour. Res.*, *20*(11), 1611–1619, 1984.

87. Rodriguez–Iturbe, I., D. R. Cox, and V. Isham, Some models for rainfall based on stochastic point processes, *Proc. R. Soc. Lond. Ser. A*, *410*, 269–288, 1987.

88. Rodriguez–Iturbe, I., B. Febres de Power, and J. B. Valdes, Rectangular pulses point process models for rainfall: Analysis of empirical data, *J. Geophys. Res.*, *92*(D8), 9645–9656
1987.

89. Rodriguez–Iturbe, I., B. Febres de Power, M. B. Sharifi, and K. Georgakakos, Chaos in rainfall, *Water Resour. Res.*, *25*(7), 1667–1675, 1989.

90. Roldan, J., and D. A. Woolhiser, Stochastic daily precipitation models: 1. A comparison of occurrence processes, *Water Resour. Res.*, *18*(5), 1451–1459, 1982.

91. Salas, J. D., D. C. Boes, V. Yevjevich, and G. G. S. Pegram, Hurst phenomenon as a pre-asymptotic behavior, *J. Hydrol.*, *44*(1), 1–15, 1979.

92. Salas, J. D., J. R. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, Littleton, CO, 1980.

93. Salas, J. D., and D. C. Boes, Shifting level modelling of hydrologic series, *Adv. Water Resour.*, *3*, 59–63, 1980.

94. Salas, J. D., and M. Chebaane, Stochastic modeling of monthly flows in streams of arid regions, in *Proc. Intern. Symp. HY&IR Div. ASCE*, San Diego, CA, 1990, pp. 749–755.

95. Salas, J. D., and M. W. Abdelmohsen, Determining streamflow drought statistics by stochastic simulation, in *Proc. U.S.-PRC Bilateral Symp. on Droughts and Arid-Region Hydrology*, Tucson, AZ, September 16–20, 1991, U.S. Geological Survey Open–File Report No. 91–244, 1991.

96. Salas, J. D., and J. T. B. Obeysekera, Conceptual basis of seasonal streamflow time series models, *ASCE J. Hydraul. Eng.*, *118*(8), 1186–1194, 1992.

97. Salas, J. D., Analysis and modeling of hydrologic time series, in D. R. Maidment (Ed.) *Handbook of Hydrology*, McGraw-Hill Book, New York, 1993, Chapter 19.

98. Santos, E., and J. D. Salas, Stepwise disaggregation scheme for synthetic hydrology, *ASCE J. Hydr. Eng.*, *118*(5), 765–784, 1992.

99. Seed, AW, R Srikanthan, and M. Menabde, A space and time model for design storm rainfall, *J. Geophy. Res.*, accepted for publication.

100. Sharma, A., D. G. Tarboton, and U. Lall, Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308, 1997.

101. She, Z. S., and E. C. Waymire, Quantized energy cascade and log-Poisson statistics in fully developed turbulence, *Phys. Rev. Lett.*, *74*(2), 1995.

102. Shukla, J., J. Anderson, D. Baumherner, C. Brankovic, Y. Chang, E. Kalnay, L. Marx, T. Palmer, D. Paolino, J. Ploshay, S. Schubert, D. Straus, M. Suarez, and J. Tribbia, Dynamical seasonal prediction, *Bull. Am. Meteor. Soc*, *81*, 2593–2606, 2000.

103. Smith, J. A., and A. F. Karr, A point process model of summer season rainfall occurrences, *Water Resour. Res.*, *19*(1), 95–103, 1983.

104. Smith, J. A., and W. F. Krajewski, Statistical modeling of space–time rainfall using radar and rain gage observations, *Water Resour. Res.*, *23*(10), 1893–1900, 1987.

105. Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, *20*(1), 47–56, 1984.

106. Stedinger, J. R., D. P. Lettenmaier, and R. M. Vogel, Multisite ARMA(1,1) and disaggregation models for annual streamflow generation, *Water Resour. Res.*, *21*, 497–509, 1985a.

107. Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, *21*(5), 665–675, 1985b.

108. Tarboton, D. G., A. Sharma, and U. Lall, Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*, 107–119, 1998.

109. Tessier, Y., S. Lovejoy, and D. Schertzer, Universal multifractals: Theory and observations for rain and clouds, *J. Appl. Meteorol.*, *32*(2), 223–250, 1993.

110. Treiber, B., and E. J. Plate, A stochastic model for the simulation of daily flows, *Hydrol. Sci. Bull.*, *22*(1), 175–192, 1977.

111. Valencia, D. R., and J. C. Schaake, Jr., Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*(3), 580–585, 1973.

112. Vecchia, A., J. T. B. Obeysekera, J. D. Salas, and D. C. Boes, Aggregation and estimation for low-order periodic ARMA models, *Water Resour. Res.*, *19*(5), 1297–1306, 1983.

113. Venugopal, V., and E. Foufoula-Georgiou, Energy decomposition of rainfall in the time-frequency-scale domain using wavelet packets, *J. Hydrol.*, *187*, 3–27, 1996.

114. Venugopal, V., E. Foufoula-Georgiou, and V. Sapozhnikov, Evidence of dynamic scaling in space-time rainfall, *J. Geophys. Res.*, *104*(D24), 31599–31610, 1999.

115. Walko, R. L., L. Band, J. Baron, T. G. Kittel, R. Lammers, T. J. Lee, D. Ojima, R. A. Pielke, Sr., C. Taylor, C. Tague, C. J. Tremback, and P. L. Vidale, Coupled atmosphere-biophysics-hydrology models for environmental modeling, *J. Appl. Meteor.*, *39*, 931–944, 2000.

116. Waymire, E., V. K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-β scale, *Water Resour. Res.*, *20*(10), 1453–1465, 1984.

117. Wigley, T. M. L., P. D. Jones, K. R. Briffa, and G. Smith, Obtaining sub-grid-scale information from coarse resolution general circulation model output, *J. Geophys. Res.*, *95*(D2), 1943–1953, 1990.

118. Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, Academic, San Diego, CA, 1995.

119. Woolhiser, D. A., and H. B. Osborn, A stochastic model of dimensionless thunderstorm rainfall, *Water Resour. Res.*, *21*(4), 511–522, 1985.

120. Yevjevich, V., *Stochastic Processes in Hydrology*, Water Resouces Publications, Littleton, CO, 1972.

CHAPTER 34

STOCHASTIC FORECASTING OF PRECIPITATION AND STREAMFLOW PROCESSES

JUAN B. VALDÉS, PAOLO BURLANDO, AND JOSÉ D. SALAS

1 INTRODUCTION

Over the past two decades, considerable research has been carried out in hydrology on developing mathematical tools and approaches for short- and long-term precipitation and streamflow forecasting. The forecasts may be concerned with flood warning, flood control, water quality control, navigation, energy production, and irrigation. Hydrologic *forecasting* signifies estimating the time of occurrence and the magnitude of a hydrological event before its actual occurrence (e.g., estimating daily streamflow with days or weeks in advance), i.e., an estimate of the future states of the hydrological phenomena is obtained in *real-time*. The adjective real-time is often used to reinforce the distinction between forecasting (the estimation of future hydrologic events based on the currently available data) and simulation, sometimes called long-term prediction (the estimation of equally likely scenarios of hydrologic events without necessarily conditioning on real-time data). In short, forecasting is generally used for operational and management purposes while simulation is used for design and planning purposes.

Forecasting of hydrological processes is an important tool for many water resources management and operational problems. For example, rainfall and streamflow forecasting hours, days, weeks, and months in advance (depending on the particular case at hand) are important for many flood warning, evacuation, and mitigation plans and actions. The U.S. National Weather Service (NWS) routinely issues precipitation forecasts (throughout the year) for all the U.S. territories and flow forecasts at key control points of the stream network systems in the United

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

States. Forecasting the number of hurricanes of certain strengths that may occur in the following year (Gray et al., 1994) and forecasting the path and the intensity of an ongoing hurricane, have been regular activities of the National Oceanic and Atmospheric Administration's (NOAA's) Hurricane Center. From the hydrologic and water resources perspectives, forecasting hurricanes has many implications, particularly as they relate to the occurrence of floods. In systems involving reservoirs, hurricane forecasts are useful for planning and implementing special operating rules to cope with impending floods. In small river systems that may be subject to flash floods, forecasting rainfall and streamflow a few hours in advance may be critical for implementing emergency actions such as alerting and warning the public. On the other hand, in large systems, such as the Mississippi River in the United States or the Paraná River in Argentina, flood occurrences may develop through several weeks and months. In these cases rainfall and flow forecasts are usually needed with lead times of weeks and months. Also in river systems where spring and summer runoff occurs from snowmelt, forecasts are usually needed weeks and months in advance for planning water supply and hydropower systems operations and for preparing for possible snowmelt floods. In such cases, determining the current amount of snow pack in the system and snow properties is of outmost importance. The development of reservoir operating rules and the real-time operation of reservoir systems may require hourly, daily, weekly, monthly, and yearly forecasts depending of the particular case at hand. Forecasts of rainfall, snowfall, snow pack, soil moisture, evaporation, streamflow, reservoir levels, river levels, and groundwater heads are generally needed in most cases of practical interest.

Forecasting of hydrologic processes has been developed using similar approaches as for simulation, although many models and techniques are unique either for simulation or forecasting. This chapter emphasizes forecasting based on stochastic and probabilistic techniques. Also, the emphasis will be on precipitation and streamflow processes, although many of the methods and models included herein are equally applicable for other hydroclimatic processes as well as evapotranspiration, soil moisture, surface and groundwater levels, and sea surface temperature.

In developing precipitation and streamflow forecasting models, one must be aware of the large uncertainty in the model parameters because of inadequate historical data of the relevant processes under consideration. Furthermore the model parameters may be expected to change slowly/rapidly with time, but the exact nature of the change is not predictable. In such cases, it is highly desirable to develop a model that has self-learning capabilities, so that it can adapt itself to the current situation (Brown and Hwang, 1997). For this purpose, filters have been formulated in the literature under the assumption that dynamic system parameters and input/measurement error statistics are known. This is not the case for precipitation and streamflow forecasting and additional estimation techniques are necessary. The sequential estimation procedure, known as the *Kalman filter*, is optimal under such conditions. However, if the actual values of system coefficients and covariances are different from those used in state estimation, then the filter is suboptimal: State estimates may contain more errors than is necessary and, in some cases, diverge from the neighborhood of the true values. State estimates could be improved by simultaneously estimating the uncertain parameters and the statistics. This additional information may be used to adapt the filter gains and model coefficients to the measurements. Adaptive filters may perform as well as optimal filters in the limit (Stengel, 1986).

Nonstationary characteristics are conventionally assumed to arise from the presence of one or more integrators in the stochastic part of the signal generation process. This applies in those cases where the model of the underlying time series data can only be characterized adequately by parameters, which vary over time in some significant manner. In all these situations the Kalman filter provides information on the possible nature of these parametric variations. Other statistical tools that are used for short- and long-term forecasting of precipitation and streamflows include methods based on regression models, autoregressive integrated moving average (ARIMA) models, ARMAX models, transfer function noise (TFN) models, and models based in artificial neural networks (ANN). In the next section a brief description of the Kalman filter will be made because of the ample use of this technique in hydroclimatic forecasting and because many of the above models can be used in conjunction with the Kalman filter. Subsequent sections will include many of the referred models and techniques for precipitation and streamflow forecasting.

2 ADAPTIVE PREDICTION: THE KALMAN FILTER

Since its introduction the Kalman filter has become a powerful tool in the fields of estimation and control theory (Kalman, 1960; Kalman and Bucy, 1961). As systems become more complex and noise becomes present in both input and output variables, it is then necessary to search for statistical solutions that can take advantage of past performance and adjust future forecasts accordingly. It is viewed as a complementary tool to the mathematical modeling of the rainfall–runoff process rather than a substitute because the knowledge of the underlying mechanism of the hydrologic process is essential for a successful implementation of the filter. The main purpose of this section is to present an introductory view of the Kalman filter rather than a thorough theoretical explanation of the statistical properties of the filter. For a successful application of the filter to real-time forecasting of hydroclimatic variables, the main hypothesis and limitations of the filter must be understood.

There are three different types of estimation problems depending on how the observations are used:

- *Filtering* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_t\}$ are used for filtering to obtain an estimate $\mathbf{x}_{t|t}$ of the state of the system $\mathbf{x}_t$.
- *Smoothing* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_t\ z_{t+1}\}$ are used for smoothing to obtain an estimate $\mathbf{x}_{t|t+1}$ of the state of the system $\mathbf{x}_t$.
- *Prediction* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_{t-1}\}$ are used in prediction to obtain an estimate $\mathbf{x}_{t|t-1}$ of the state of the system $\mathbf{x}_t$.

For a detailed discussion on the topic the reader is referred to specialized books (e.g., Brown and Hwang, 1996). This reference also includes the software for some applications. Recursive algorithms are ideal for estimation of time-varying parameters. Modifications based on stochastic modeling of the parameter variations lead naturally to the development of the Kalman filter and the estimation of time-varying states in stochastic dynamic systems. Kalman considerably extended the state-estimation and filter theory of time-varying parameters or states so as to handle the analysis of nonstationary time series and provide a natural approach to the analysis of time series data that are assumed to be generated from stochastic state-space equations.

When modeling a system that evolves through time, specifically a stochastic process that is defined in discrete time, one would like to put the system in a *state-space* form or in the so-called *state of the system* vector $\mathbf{x}_t$. (Most linear models can be put into state-space form; nonlinear models can be linearized by using Taylor series expansion to reformulate them in state-space form.) If future values of the state of the system, $\mathbf{x}_{t+s}$, $s = 1, 2, \ldots$, can be modeled using knowledge of x_t (i.e., $\mathbf{x}_t$ contains all the required information about the previous values $\mathbf{x}_{t-s}$, $s = 1, 2, \ldots$), we obtain what is called a Markovian system. The best description of $\mathbf{x}_t$ using $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots$ can be modeled as

$$\mathbf{x}_t = \Phi(\mathbf{x}_{t-1}, t-1) + \Gamma(\mathbf{w}_t, t) \tag{1}$$

Equation (1) is called the *state equation* of the system, where $\Phi(\cdot)$ is the transition function, $\Gamma(\cdot)$ is the noise transition function, $\mathbf{w}_t$ is the vector of system noises that describes the part of $\mathbf{x}_t$ that is not explained by $\mathbf{x}_s$, $s < t$, and is assumed independent of $\mathbf{x}_s$ and $\mathbf{w}_s$ for $s < t$. When $\Phi(\cdot)$ and $\Gamma(\cdot)$ do not vary with time dependence, the system is referred to as *stationary*.

In most applications, the state of the system, $\mathbf{x}_t$, is not directly observed but rather measured in an observation vector $\mathbf{z}_t$, which is a function of $\mathbf{x}_t$, corrupted by measurement noise $\mathbf{v}_t$. This may be written as:

$$\mathbf{z}_t = \mathbf{H}_t(\mathbf{x}_t, t) + \mathbf{v}_t \tag{2}$$

This equation is called the *observation equation of the filter*, and Eqs. (1) and (2) together constitute the heart of the Kalman filter; they may represent linear or nonlinear systems. The filtering problem is to estimate $\mathbf{x}_t$ from the observations $\mathbf{z}_1, \ldots, \mathbf{z}_t$, which are corrupted by measurement noises. If both the system and the observations are assumed linear, Eqs. (1) and (2) will have the following form:

$$\mathbf{x}_t = \Phi_t\mathbf{x}_{t-1} + \Gamma_t\mathbf{w}_t \tag{3}$$

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t \tag{4}$$

where Φ_t, Γ_t, and $\mathbf{H}_t$ are known matrices.

The stochastic properties of the system and measurement noises have to be defined in order to apply the Kalman filter. Also the properties of the initial state of the system have to be defined. The main challenges in the application of the filter to forecasting of precipitation and streamflow are to define the appropriate matrices and other components of the filter in terms of the hydrologic variables and to estimate them from available data and knowledge of the physical process. In subsequent sections the application of various statistical techniques including the Kalman filter for forecasting hydroclimatic processes, particularly precipitation and streamflow, is presented.

3 STOCHASTIC PRECIPITATION FORECASTING

Precipitation forecasting is of great significance for water resources management and flood protection, although it is not an easy task. Rainfall forecasts based on the analysis of the temporal and spatial evolution of the meteorological phenomena would be always desirable. Considerable progress has been made in this respect by using numerical weather prediction approaches and general circulation models. However, information from such sources is not always available in operational form. In this situation rainfall forecasts can be made based on the persistence characteristics of current and past rainfall measurements, even though the accuracy of such forecasts may suffer because of the lack of the physical aspects involved in the precipitation phenomena. A number of examples of precipitation forecasting by statistical, stochastic, or probabilistic techniques can be found in the literature. They include regression techniques, Markov chains, ARMA-type methods, probability-function-based approaches, and artificial neural networks (ANNs). All of these approaches have been used for short-term and for mid- and long-term forecasting, as illustrated below.

Short-Term Forecasting

Quantitative precipitation forecasting, often denoted as QPF, is one of the major tasks in flood forecasting. It has been demonstrated that QPF allows extending the lead time of flood forecasts and improving the accuracy of flood estimates for a given forecast lead time (Brath et al., 1988). Although research in the field of numerical weather prediction has achieved significant progress in recent years (see, e.g., Bougeault et al., 2000), forecasting techniques based on stochastic and statistical modeling are useful especially for operational purposes and in the context of mesoscale basins, which are characterized by rapid response time. However, because of the complexity of the rainfall phenomena, which exhibits significant spatial and temporal variability, nonstationarity, and nonlinearity, especially on small scales, rainfall forecasting by stochastic approaches involves a challenging feat and experience.

Early attempts to forecast rainfall were formulated as statistical black-box models used for storm tracking. For instance, Phanartzis (1979) developed a simple model for forecasting the direction of storm movement based on the cross-correlation of rainfall measured at a network of rain gages. A similar approach was developed by

Nguyen et al. (1978) to be used with radar storm tracking signals. A more sophisticated storm tracking statistical procedure based on Kalman filter was proposed by Johnson and Bras (1980). Also, French and Krajewski (1994) and French et al. (1994) used the Kalman filter for state updating and incorporation of uncertainty in a two-dimensional physically based model and surface meteorological observations. Furthermore, Sugimoto et al. (2001) also used the extended Kalman filter as a state estimator to update the model parameter of the conceptual model with new radar data and with forecasts from a numerical weather prediction model.

Other authors, such as Lardet and Obled (1994), generated scenarios of rainfall duration and volume by probability functions conditioned on past rainfall. Statistical methods based on classification trees were also used for QPF (Carter and Elsner, 1997; Carter et al., 2000). In other applications physically based model structures are combined with stochastic components to account for the uncertainties associated with model hypotheses and structure (Jinno et al., 1993). Kawamura et al. (1996, 1997) added a Gaussian white noise in time and space to an advection-diffusion model of space–time rainfall, to consider a certain degree of error and uncertainty inherent in rainfall modeling.

Other approaches try to overcome the intrinsic limitation of persistence-based methods for predicting rainfall, due to the short decorrelation time of the precipitation process, which has been shown to be of the order of approximately 20 min (Zawadzki, 1987). Four stochastically based approaches for forecasting short-term precipitation are presented below.

Point Process Models. The models based on point processes perform satisfactorily with respect to reproducing the cluster dependence properties of observed rainfall (Entekhabi et al., 1989) and related extreme properties (Burlando and Rosso, 1993). However, the formulation required for real-time forecasting is very complex. Ramirez and Bras (1985) developed an algorithm for forecasting storm arrivals assuming the Neyman–Scott white-noise model as the underlying rainfall-generating mechanism. They derived the general expressions for the distribution functions of the time to the next storm event, conditioned on part of the immediate rainfall history, and applied the algorithm for irrigation scheduling. French et al. (1992a) developed a real-time forecasting scheme based on the space–time model of Rodriguez-Iturbe and Eagleson (1987). The forecasting model consists of a single distributed state-space equation, which is used to derive the conditional mean and the conditional covariance of rainfall intensity. Updating of the rainfall field in real time is carried out by representing the model structure as a distributed parameter Kalman filter. While some work has been done in using point and cluster processes for real-time forecasting of precipitation, their development has been limited to research studies.

Regression-Based Methods. A good example of how rainfall forecasting based on statistical methods is useful for operational purposes is the U.S. National Weather Service's centralized statistical quantitative precipitation forecasts (Antolik, 2000). The statistical forecast is based on multiple linear regression (Glahn and

Lowry, 1972; Lowry and Glahn, 1976), where the rainfall amount over a given time interval is predicted as a function of meteorological variables, both observed and computed by numerical weather models. Despite the relative simplicity of the model, it often outperforms physically based methods and more complex techniques, depending on the proper identification of the predictors. The use of regression methods though is more common in long-term forecasting.

Markov Chains Approach. The theory of Markov chains has been suggested for short- and long-term forecasting of rainfall. For example, Bertoni et al. (1992) used a first-order Markov chain for real-time forecasting of rainfall for a few hours lead time, which in turn was used for flood forecasting. Historical rainfall data were classified in states that divide the range of rainfall variation into sequences of nonoverlapping intervals. The transition probabilities were estimated as $p_{ij} = n_{ij}/\sum_{j=1}^{r} n_{ij}$, $(i, j = 1, \ldots, r)$, where r is the number of states, and n_{ij} is the number of transitions from state i to j, which is computed from historical observations on a seasonal basis. The p_{ij} values are elements of the transition probability matrix, which is then used to estimate (forecast) the m-step (ahead) transition probability $p_{ij}^{(m)}$ on the basis of the incoming observations (i.e., the present state) and the given conditional nonexceedence probability. The selection of an appropriate nonexceeding probability is key in achieving acceptable rainfall forecasts. Yu and Yang (1997) adopted a similar approach and further analyzed the role played by the choice of the nonexceeding probability with respect to forecast accuracy. In addition to seasonal dependence, the nonexceeding probability strongly depends on storm profile, being considerably different in the raising limb than in the recession limb of the hyetographs.

Dahale and Puranik (2000) applied a six-state simple Markov chain to forecast 5-day spatial rainfall persistence of summer monsoons over the Indian region. Fraedrich and Müller (1983) used a five-state simple Markov chain, and Miller and Leslie (1984) adopted a four-state second-order model to predict rainfall probabilities from past weather states. One must note that high forecast skills are generally obtained for short lead times, and they significantly decrease with increasing lead times. Johnson and Bras (1980) combined forecasts of the mean rainfall rate throughout the event at each gage with the modeling of a random residual component based on a Markovian model. The choice of the optimal order of a Markov chain also plays a role in forecast accuracy. Akaike information criterion and Bayes information criterion can be used for this purpose (e.g., Tong, 1975; Katz, 1981; Gregory et al., 1992).

ARMA Models. Trotta et al. (1977), and Labadie et al. (1981) showed that ARMA and transfer function models can be used for modeling rainfall persistence. They used an autoregressive transfer function model for short-term rainfall forecasting for the purpose of improving the control of a sewer system. The model uses parameters estimated from historical data at the beginning of the storm event, when information of the ongoing event is still poor. As the storm progresses, the parameters are progressively tuned to reflect the increasing real-time information. This is done in the estimation step by including weighting factors in a least-squares algorithm to account differently for the historical information and current rainfall event information. Obeysekera et al. (1987) showed that certain point process models widely applied for modeling short-term rainfall, such as the Poisson rectangular pulse (PRP) and the Neyman Scott rectangular pulse (NSRP), possess correlation structures like those of ARMA(1,1) and ARMA(2,2) models, respectively. Thus, in principle, ARMA models could be used for simulation and forecasting of short-term rainfall processes. Because ARMA models are stationary, and the underlying variable is normally distributed, their application to real-time forecasting of short-term precipitation, such as hourly and daily rainfall, requires certain procedures to be followed to take into account such requirements. Burlando et al. (1993) used the ARMA(2,2) model given as

$$Z_t = \sum_{j=1}^{2} \phi_j Z_{t-j} + \varepsilon_t - \sum_{j=1}^{2} \theta_j \varepsilon_{t-j} \tag{5}$$

where $Z_t = X_t - \mu$, X_t represents hourly rainfall, μ is the mean of X_t, ϕ_j and θ_j are the autoregressive and moving average coefficients, respectively, and ε_t is a normally distributed noise with mean zero and variance σ_ε^2.

Nonstationarity of the rainfall throughout the year was accounted for either by seasonal estimation of parameters based on the analysis of the continuous data set or by event-based parameter estimation carried out only on extracted nonzero rainfall events. In the latter approach a different parameter set was determined for each storm event considered. To account for the nonlinearity that characterizes storm precipitation events, some modifications were necessary for the estimation of the ARMA model (5) as shown schematically in Figure 1. Specifically, data are first transformed to account for non-normality by means of the Box–Cox transformation (Box and Cox, 1964), and the estimation of the model parameters is performed by an iterative adaptive least-squares technique. Thus, the data used for estimation are only those available as the storm event evolves through time, implicitly assuming a local stationarity. While the results based on the continuous data set were not satisfactory, the event-based application provided satisfactory results. Figure 2 shows an example of the forecast accuracy. A noticeable problem, however, is the one-hour phase shift that characterizes most of the forecasted events. Toth et al. (2000) obtained similar results by slightly modifying the procedure introduced by Burlando et al. (1993), in that the event-based parameter estimation was carried out on the basis of a moving window of fixed length rather than on the complete event data sequence. Transformation of data was also relaxed because forecast applications based on ARMA models do not require the data to be Gaussian, i.e., ARMA models provide the best linear prediction even for non-Gaussian data (Brockwell and Davis, 1991).

The temporal phase shift exhibited by forecasts obtained by the univariate ARMA model can be partially explained by the error induced by storm movement. One can ameliorate this effect by selecting additional data measured at other neighboring stations (e.g., by using the cross-correlation between the rainfall at the station of

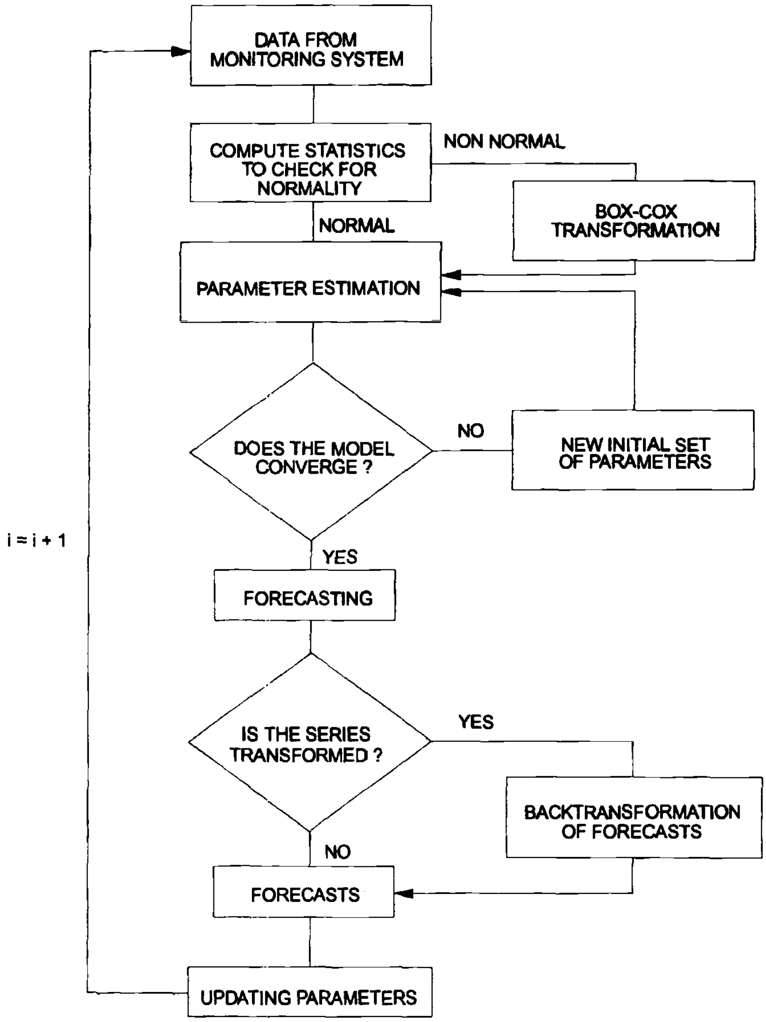

DATA FROM MONITORING SYSTEM
COMPUTE STATISTICS TO CHECK FOR NORMALITY
NON NORMAL
BOX-COX TRANSFORMATION
NORMAL
PARAMETER ESTIMATION
DOES THE MODEL CONVERGE ?
NO
NEW INITIAL SET OF PARAMETERS
YES
FORECASTING
IS THE SERIES TRANSFORMED ?
YES
BACKTRANSFORMATION OF FORECASTS
NO
FORECASTS
UPDATING PARAMETERS
i = i + 1

Figure 1 Flowchart of the event-based ARMA forecasting procedure (from Burlando et al., 1993).

interest, i.e., the station where the forecast is issued, and those at the other stations) and reduce the error associated with the phase shift. Using a multivariate integrated ARMA (MARIMA) forecasting scheme (Montanari et al., 1994; Burlando et al., 1996) can do this. Montanari et al. (1994) suggested that a multivariate scheme could remarkably improve the forecasts when the rain gages to be used in forecasting

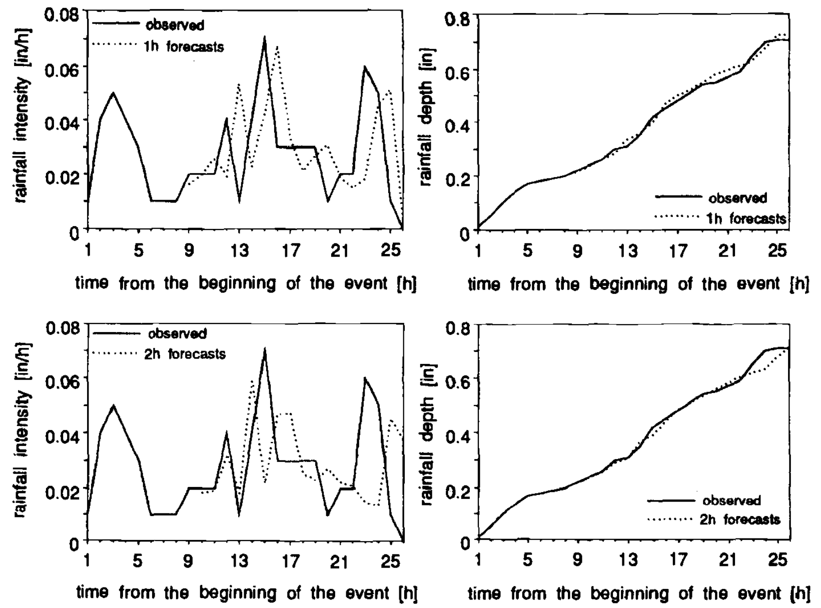


Figure 2 Example of 1- and 2-h rainfall forecasts for the event of October 14, 1960, Denver, Colorado, obtained by means of an ARMA(2,2) process (from Burlando et al., 1993).

are selected adequately. Burlando et al. (1996) showed that the estimation of a Lagrangian space–time correlation of the moving storm could be made using storm maps recorded by weather radar, which provide the direction and the speed of the storm movement. Storm tracking can thus be applied to actual events to select those stations that are characterized by the highest Lagrangian cross-correlation of observed precipitation, and therefore are the best suitable for application with the multivariate model. The parameters of the multivariate model are thus estimated using only observed rainfall at the selected stations throughout the current event.

Specifically, the MARIMA model estimates the future occurrences of a time series as a linear combination of (a) past occurrences of the underlying time series and of time series which are cross-correlated to it—i.e., the autoregressive component—and of (b) the present and past occurrences of a random white-noise component—i.e., the moving average component. The MARIMA model can be expressed as

$$Z_t = \sum_{i=1}^{p} \Phi_i Z_{t-i} + \sum_{j=0}^{q} \Theta_j \varepsilon_{t-j} \tag{6}$$

where p and q are the autoregressive and the moving average order respectively, $\mathbf{Z}_t = (\mathbf{I} - \mathbf{B})^d \mathbf{X}_t$, $\mathbf{X}_t$ is the rainfall intensity, $\mathbf{I}$ is the identity matrix, $\mathbf{B}$ is the backward operator, d is the differencing order of the model, and ε_t is a normally distributed noise term. Both $\mathbf{Z}_t$ and $\mathbf{X}_t$ are n-dimension column vectors ($n =$ number of series), and Φ and Θ are the $n \times n$ autoregressive and moving average parameters matrices of the model. The number of parameters in (6) becomes large as the orders p and q increase. This is a major limitation in analytical tractability and parameter estimation especially in those cases where a limited number of observations are available. Accordingly, the values of p and q, as well as the number of series n, should be selected as a compromise between the conflicting needs of the process descriptiveness and of mathematical tractability.

Burlando et al. (1996) explored the suitability of the MARIMA(1,1,0) model for a catchment in northern Italy. Parameter estimation was carried out on individual events, as in Burlando et al. (1993), and using the method of moments as

$$\Phi = \mathbf{M}_1 \mathbf{M}_0^{-1} \tag{7a}$$

$$\Theta\Theta^T = \mathbf{M}_0 - \mathbf{M}_1 \mathbf{M}_0^{-1} \mathbf{M}_1^T \tag{7b}$$

where $\mathbf{M}_0$ and $\mathbf{M}_1$ denote the lag-0 and lag-1 covariances, respectively. The identification of the pair of stations was carried out either on the basis of historical cross-correlations or from the analysis carried out in real time from radar maps. The latter provided the basis for the analysis of the kinematic characteristics of the storm, so allowing the identification of a (first) *lead station*, located downwind the (second) forecasting *station*. The lead station is taken as a reference station for the second station is selected among those located along the direction of the storm movement that is identified from the radar maps. The MARIMA(1,1,0) was thus estimated using rain gage data observed at the selected stations, and rainfall forecasts were issued at each station as a function of the current and past occurrences observed at the station itself and at the lead station. Satisfactory results were obtained as reported in Burlando et al. (1996).

Artificial Neural Networks. An alternative route to the foregoing stochastic forecasting techniques is the use of artificial neural networks. These are essentially data processing systems that can reproduce by learning the relationships between a pair of one- or multidimensional data sets. An artificial neural network (ANN) is made of many simple nonlinear units that mimic the human neurons. These collect the input from a single or multiple sources producing an output according to a predefined nonlinear function. In a sense an ANN is a sort of a transfer function model that appears to be suitable to tackle the problem of rainfall forecasting.

Use of ANNs for the purpose of weather-related quantities started in the early 1990s. French et al. (1992b) developed a neural network to forecast rainfall intensity fields in time and space, which were generated by a modified version of the stochastic rainfall simulation model proposed by Rodriguez-Iturbe and Eagleson (1987). The network with input, hidden, and output layers was trained using the back-propagation technique on a regular grid domain to test the ability of the ANN to investigate the role of the number of hidden nodes on its performance. The model skill was tested based on a varying number of training sets and the rainfall fields generated by the stochastic model. Real-time learning and off-line learning were additionally tested. Kuligowksi and Barros (1998) applied a combination of precipitation data from a number of rain gages and wind directions to forecast rainfall amounts for a target location and a lead time of 6 h. Specifically, rainfall observations at rain gages in a region of radius 300 km centered on the target location, upper level winds at three radiosonde locations, and wind direction data from a number of levels were combined to build the training set of the ANN.

More recently, Luk et al. (2000) adopted ANNs to forecast short-term rainfall for an urban catchment, aiming at the investigation of the effect of temporal and spatial information on short-term rainfall forecasting. The forecast accuracy of ANNs was evaluated for different configurations of lag orders and number of spatial inputs based on historical rainfall patterns. They concluded that the most accurate predictions depend on the identification of an optimum number of spatial inputs, and that the network with lower lag consistently produced better performance. An interesting application of ANNs has been recently shown by Toth et al. (2000), who provided for a real case study a comparison of ANNs performance with respect to real-time prediction based on ARMA models and a nonparametric nearest-neighbor technique. Multilayer feed-forward network architectures were tested against the one-layer scheme in order to determine the optimal network configuration, both in the case of split-sample application and adaptive calibration. As one may expect, better performances were obtained for the split-sample application, which makes use of larger training sets, whereas the adaptive calibration gives worse results for short lead times. Compared to other forecasting techniques ANNs was slightly superior in the overall performance due to their ability to account for the nonlinearities that characterize temporal rainfall. Grecu and Krajewski (2000) reported another interesting application of back-propagation neural network (BPNN) for rainfall forecasting. In this case, rainfall amounts were not directly modeled by means of the BPNN, but this was used to model one component of the statistical radar-based quantitative precipitation forecast procedure.

Mid- and Long-Term Forecasting

If quantitative short-term forecasting is useful for flood forecasting, mid- and long-term forecasting plays a major role in the management of water resources. Agriculture and water supply, among other water uses, can significantly benefit from the availability of forecasts of rainfall amounts that can be expected over a time horizon of a month or a season. This issue is particularly relevant for complex systems that strongly depend on joint management of surface and groundwater resources. Forecasting at the mid- and long-term scales involves problems that are similar to the one already observed for smaller time scales. Nonstationarity, nonlinearity as well as the identification of the correct predictors guided the development of methods.

Whereas regression methods and, more recently, artificial neural networks have been extensively used for the purpose, a few other approaches can be found in the literature to forecast rainfall on mid- and long-term time scales. A truncated normal distribution is, for instance, the basis of the formulation of a nonstationary multisite model of rainfall that Sansó and Guenni (2000) show to capture the year-to-year variability and suggest to be suitable for short-term forecasting as well. Stone et al. (1996) and de Jager et al. (1998) used a simple probabilistic rainfall forecasting technique that is based on the identification of lag relationships between the values of the Southern Oscillation Index (SOI)—which can be considered as representative of the phase of the El Niño Southern Oscillation (ENSO) cycle—and future rainfall. Probability distributions for the subsequent 3 months are thus derived conditioned on the state of the SOI. Sharma (2000) introduced a nonparametric probabilistic model for forecasting rainfall with 3 to 24 months of lead times. Specifically, nonparametric kernel methods (e.g., Scott, 1992) for probability density function (PDF) estimation are used to express the conditional probability density function. Then, probabilistic forecasts are made by resampling from the rainfall probability density conditioned on the current value of the associated predictor set. An interesting feature of this approach is that the shape of the PDF is directly built from the data, and this leads to forecasts that are expected to resemble the characteristics of the sample and therefore reproduce the variability of observed rainfall.

Regression-based techniques have been extensively used for predicting seasonal rainfall. The increased availability of predictor variables, like ENSO, in near real time, by means of either observations or numerical weather prediction has increased the applicability of regression-based models and the Kalman filter (e.g., Liu et al., 1998). Makarau and Jury (1997) forecasted summer rainfall in Zimbabwe based on a set of climatic predictors by means of a multivariate linear regression model in a forward stepwise approach. Fairly simple models including up to five predictors produced jack-knife skill test correlations of about 80 to 85% for a lead time of 2 to 3 months. A similar approach was used by Jury (1998) to forecast seasonal rainfall and other climatic variables for the KwalaZulu-Natal region in southern Africa, also obtaining a forecast skill of about 76% for rainfall and about two thirds of the variance in the other cases. Francis and Renwick (1998) focused on predicting seasonal (either 1 month or one 3-month season) rainfall anomalies.

Similarly, Thapliyal (1997) carried out a comparison of forecast models based on the correlation between predictors and predictands and a dynamic stochastic transfer model to predict monsoon rainfall in India. The dynamic stochastic transfer model corresponds essentially to an ARIMA model structure, the orders of which are estimated against observations. It should be observed that a critical issue in using regression-based techniques is the stability of the selected predictor and the robustness of the model describing its temporal evolution. Finally, ARMA models, which have been extensively applied to forecast streamflows, have also been used to simulate rather than to forecast mid- and long-term rainfall. Another application of ARIMA models for forecasting monthly rainfall series with the purpose of providing an input for flow forecasting in the management of water resources systems can be found in Delleur and Kavvas (1978).

As in the case of short-term forecasts, artificial neural networks have been proposed for forecasting seasonal rainfall. ANNs has been found to be useful for forecasting the behavior of complex and highly dynamic systems such as the

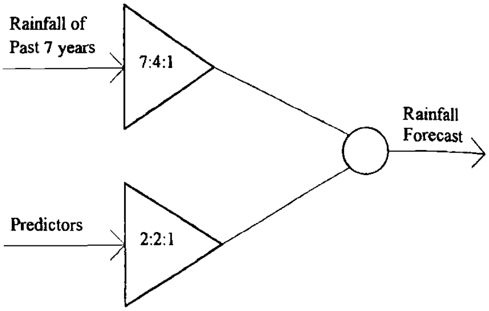


Figure 3 Architecture of a hierarchical artificial neural network combining deterministic information from historical data and stochastic component represented by the event predictors for seasonal forecasting of monsoon rainfall (from Navone and Ceccatto, 1994).

monsoon rainfall. Simple deterministic neural networks show, however, a limited robustness in terms of forecast skill, so that more complex networks are often introduced. An example of a simple four-layer input, two hidden layers, and one neuron output network for forecasting Indian monsoon rainfall can be found in Sahai et al. (2000). On the other hand, Navone and Ceccatto (1994), proposed an interesting but complex application of ANNs, as a nonlinear method to correlate preseason predictors to rainfall data and as an algorithm for reconstructing the rainfall time series dynamics. Accordingly, they implemented a hierarchical neural network, which is sketched in Figure 3. The network trained to correlate predictors and the network trained to learn the time series dynamics are combined by connecting their output units to a new neuron, which is then used to issue the forecasts. The authors refer about an improved forecast skill due to the hierarchical approach, especially in forecasting large anomalies. The performance of ANNs can be reduced if the parameters used to train and forecast are correlated. Guhathakurta et al. (1999) used principal component analysis—as suggested by Hsieh and Tang (1998)—to transform the original variables into a new set of uncorrelated variables. These are then used to train and issue forecasts by means of a three-layer, five-input, three hidden nodes in one single hidden layer ANN. The output from such ANN and the output of a simple deterministic ANN using the untransformed parameter set were then used each as input to a simple two-layer ANN without any hidden layer, which produced rainfall forecasts. The final hybrid model increased the overall forecast skill from about 40 to 80%.

4 STOCHASTIC STREAMFLOW FORECASTING

Over the past two decades, considerable research has been carried out in hydrology for developing stochastic models for short- and long-term forecasting of river flows.

The form of these models generally follows the ARMA, ARMAX, and transfer function type of models (Box and Jenkins, 1976), with the last two proving to be more reliable for multiple forecasting periods (Burns and McBean, 1985; Awwad and Valdés, 1992). After defining the mathematical model, usually in a state-space form, the Kalman filter is widely used as a powerful tool for obtaining optimal hydrologic forecasts and updates of the states (e.g., Chiu, 1978; O'Connell, 1980; Wood and O'Connell, 1985).

Implementation of real-time stochastic models in large-scale hydrologic systems has been thoroughly discussed by Wood (1985) and Awwad and Valdés (1992). Furthermore, adaptive filters have also been used in combination with conceptual hydrologic models for streamflow forecasting since the late 1970s. Bras and collaborators at MIT, developed techniques to represent the National Weather Service River Forecasting System (NWSRFS) land component in a state-space form. Some of the research results are discussed in the next section. Alternative techniques such as stepwise regression and transfer function models have also been used. Later in this chapter the application of ANN in streamflow forecasting will be presented. In addition, probabilistic, interval, forecasts using the standard conceptual rainfall–runoff models for short-term forecasting has been developed. The most widely used is the ESP (extended streamflow prediction, now called ensemble streamflow prediction), which is described later in this chapter.

Short-Term Forecasting of Streamflows

A number of investigators have evaluated the benefits of streamflow forecast using the Kalman filter. For example, Georgakakos (1986) studied the performance of a hydrometeorological model for streamflow forecasting using 6-h data for Bird Creek, a 2344 km^2 catchment in Oklahoma. A precipitation forecasting model was developed and coupled to a modified version of the U.S. National Weather Service rainfall–runoff model to produce the streamflow forecast. Variables such as soil moisture storages and streamflow were updated through a Kalman filter. Eight 2-month forecast periods were examined. The results showed that the nonupdating forecasting model produced forecasts where the time-to-peak discharges were very different from those observed. The forecasts using updating techniques showed significant improvements.

Other applications in streamflow forecasting include the work of Takasao and Shiiba (1984) and Takasao et al. (1989) who developed a simple nonlinear streamflow forecasting model and applied to the Haze River, a 370-km^2 basin in Japan. Their work shows model performance for a flood in September 1965 with and without updating. As expected, the forecast errors of the model without updating are larger. The deterministic model NAMS11/MIKE11 developed by the Danish Hydraulic Institute (DHI) uses a state variable updating procedure based on the Kalman filter for their conceptual rainfall–runoff model NAM (Refsgaard, 1997). Ahsan and O'Connor (1994) have expressed that the full capabilities of the Kalman filtering are not completely utilized since the predictions are expected to match the observed flows, which are considered to be noise free and that the filter will be more fully utilized in the future when remote sensing becomes more predominant.

Awwad and Valdés (1992) proposed an adaptive evaluation/forecasting algorithm for hydrologic forecasting and presented two multisite hydrologic forecasting approaches suitable for real-time applications. Their model is based on past and present flow rates, with the upstream inflows treated as exogenous inputs to the models. They applied the model to the Fraser River, Canada. In their original application Awwad and Valdés (1992) did not use precipitation terms, and even though their models performed very well in the one- and two-steps-ahead forecast error deteriorated rapidly. The authors later extended the adaptive evaluation/forecasting algorithm to include precipitation inputs, upstream inflows forecasted/evaluated with uncertainty, and deterministic reservoir releases in the stochastic models (Awwad et al., 1994). This approach was adopted because it has a relatively simple dynamic structure in a black-box form and is calibrated online as additional information becomes available. The inclusion of precipitation information considerably benefited the multiple-period forecasting ability of the stochastic models.

The general form of the ARMAX models used in Awwad et al. (1994) followed the well-known state-space form of the Kalman filter presented above where optimal forecasts and updates of the states were obtained using the Kalman filter. Two other filters in the form of the Kalman filter, referred to as the parameter space and the noise-space filters, are used in parallel with the state-space filter to update the model parameters and noise statistics online along with the states. This adaptive estimation technique using parallel filtering does not require preassigned values for the Kalman filter coefficients and noise statistics, which are usually unknown in real-world applications. Other applications in short-term forecasting include: use of an ARMAX model to do predictions on the Fraser River in Canada (Ngan and Russell, 1986), use of the Kalman filter to estimate the parameters of a PARMA model with application to the Saugeen River in Ontario, Canada (Jimenez et al., 1989), and use of the ARMAX model with Kalman filter for short-term flow forecasting of snowmelt runoff in the Rio Grande at the Del Norte station (Haltiner and Salas, 1988).

As stated before in section 3 ANNs have been widely used for a number of hydrologic problems including forecasting of precipitation and streamflow. A vast literature already exists on the subject. For example, the *ASCE J. Hydrol. Engr.* (vol. 5, no. 2, 2000) is a dedicated issue on the subject. It includes the articles "Artificial Neural Networks in Hydrology I: Preliminary Concepts" and "Artificial Neural Networks in Hydrology II: Hydrologic Application," co-authored by the ASCE Task Committee on Applications of Artificial Neural Networks in Hydrology. The second article includes a review of various applications of ANNs on short-term and long-term flow forecasting. In addition, the book *Artificial Neural Networks in Hydrology* (Govindaraju and Rao, 2000) includes some chapters specifically on flow forecasting (e.g., Gupta et al., 2000; Salas et al., 2000; Deo and Thirumalaiah, 2000). The work by Gupta et al. (2000) discusses in some detail the training of ANNs based on multilayer feedforward neural networks (MFNNs), which are most commonly used for streamflow forecasting. It also presents some results illustrating some applications. The study by Salas et al. (2000) discuss some very basic concepts underlying ANNs, gives a simple detailed example, and two applications for daily and monthly streamflow forecasting. Finally, the work by Deo and Thirumalaiah (2000) includes the application of ANNs for real-time forecasting of daily flows and daily river stages.

Long-Term Forecasting of Streamflows

Forecasting Models Using Climatic Precursors. Long-term forecasting of hydrologic variables requires climate forecasts. As mentioned in the section on long-term precipitation forecasting, the increase in predictive skills of the models to forecast climatic anomalies based on the ENSO phenomena have provided renewed impetus for hydrologic forecasting.

There have been a significant number of contributions to hydrologic forecasting using climatic precursors. For example, the hydrologic forecasting model proposed by Liu et al. (1997, 1998) used multiple ENSO forecasts to produce forecasts of seasonal precipitation, streamflow, and other variables. This is essentially a combined data fusion and forecasting system that incorporates ENSO forecasts, persistence-based forecasts, and up-to-date observations. The system assimilates past observations, hindcasts, and projects with both multiple model outputs and persistence into its merged forecast. The error bounds on the forecast are also propagated. Liu and co-workers applied the approach to forecast droughts in Texas (Liu et al., 1997) and precipitation in the Equatorial Pacific and streamflows in Colombia (Liu et al., 1998). An example of these applications is given in Figure 4. Other applications include Berri and Flamenco (1999) who used a regression model with climatic precursors to forecast seasonal volumes in the Rio Diamante and Chiew et al. (1998) who studied the relationship between ENSO and rainfall, drought and flows in Australia, and its potential for forecasting. For example, they found that spring runoff in southeast Australia may be predicted several months in advance. Other examples are the work of Guetter and Georgakakos (1996) on the relationship between Iowa River flows and ENSO and of Kayha and Dracup (1993, 1994) on the Southwest flows.

An interesting example of the use of seasonal climate outlooks of expected air temperature and precipitation probabilities in hydrologic forecasting is found in the work of Croley and colleagues at the NOAA GREL (Croley, 1996, 1997; Croley and Kunkel, 1996). In their work historical meteorology record segments are used with hydrological and other models to simulate hydrological scenarios. The historical meteorological records are weighted to be compatible with NOAA's climate outlooks. An example of the use of the procedure is shown in Figure 5. Hamlet and Lettenmaier (1999) proposed a method to incorporate both ENSO and PDO signals in the long-term forecasting of streamflows in the Columbia River. The climate forecasts, classified in six categories as a function of ENSO and PDO, are used to generate climatic scenarios that serve as input to a hydrologic model.

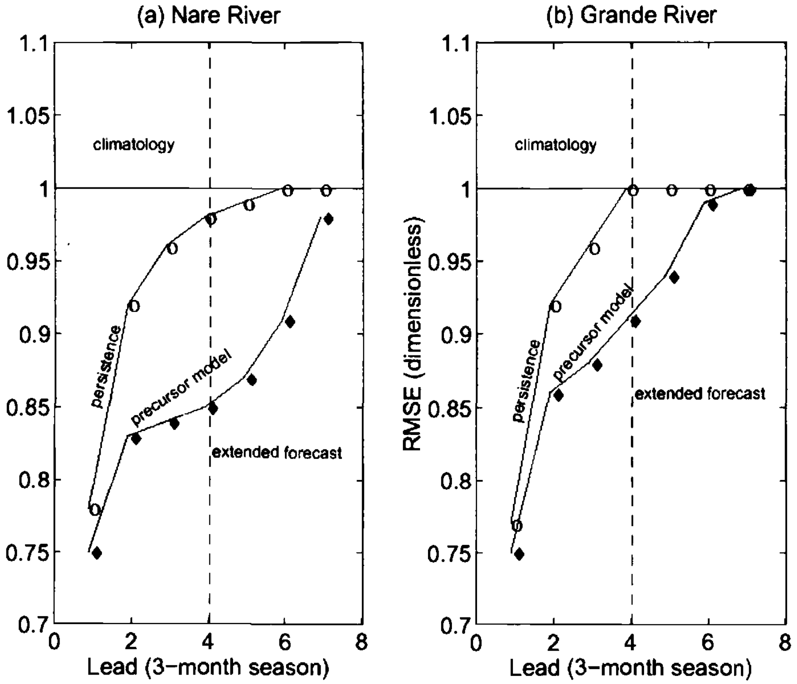


Figure 4 Performance of long-term streamflow forecasting models using climatic precursors (from Liu et al., 1998).

Extended/Ensemble Streamflow Prediction (ESP). The extended streamflow prediction (ESP) was proposed by the National Weather Service (Twedt et al., 1977; Day, 1985) as a procedure to obtain probable runoff scenarios based on simulations runs of a conceptual simulation model that uses hydroclimatic series, either historic or synthetic, and using the current soil conditions. The hydroclimatic series, usually temperature and precipitation, represent historic or synthetic time series that are likely to happen in the period of forecast. The results are then used to obtain a probability distribution of future runoff and, if desired, a point estimate, usually the sample average for a particular lead time. Figure 6 shows an example for the Rex River in Washington (Lettenmaier and Wood, 1993). The ESP approach has been widely used in the United States and other countries as a component of their forecast systems, and it is usually combined with a reservoir operation model that uses the probabilistic outputs of the ESP approach. For example, as part of the NWSRFS the ESP approach is being utilized in the operation of the Columbia River (260,000 mi^2 and a mean annual discharge of 198 million acre-feet; von der Heydt et al., 1994).

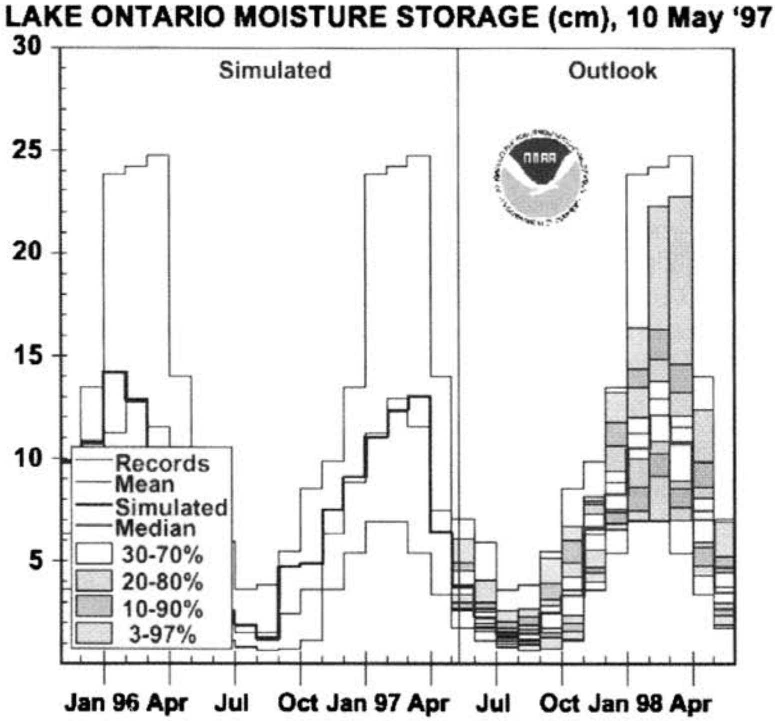


Figure 5 Example of hydrologic forecasts using climate outlooks (from Croley, 1997, 2000). See ftp site for color image.

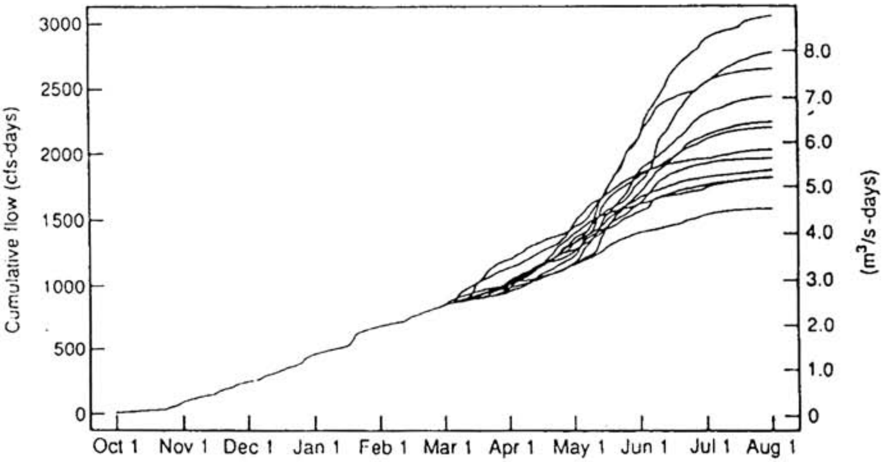


Figure 6 Example of ensemble streamflow prediction application (taken from Lettenmaier and Wood, 1993).

REFERENCES

Ahsan, M., and K. M. O'Connor, A simple non-linear rainfall-runoff model with a variable gain factor, *J. Hydrol.*, *155*, 151–183, 1994.

Antolik, M. S., An overview of the National Weather Service's Centralized Statistical Quantitative Precipitation Forecasts, *J. Hydrol.*, *239*, 306–337, 2000.

Awwad, H. M., and J. B. Valdes, Adaptive parameter-estimation for multisite hydrologic forecasting, *J. Hydraul. Eng. ASCE*, *118*(9), 1201–1221, 1992.

Awwad, H. M., J. B. Valdes, and P. J. Restrepo, Streamflow forecasting for Han River Basin, Korea, *ASCE J. Water Resour. Pl.*, *120*(5), 651–673, 1994.

Berri, G. J., and E. Flamenco, Seasonal volume forecast in the Diamante River, Argentina based on El Niño observations and predictions, *Water Resour. Res.*, *35*(12), 3803–3810, 1999.

Bertoni, J. C., C. E. Tucci, and R. T. Clarke, Rainfall-based real-time flood forecasting, *J. Hydrol.*, *131*, 313–339, 1992.

Bougeault, P., P. Binder, A. Buzzi, R. Dirks, R. Houze, J. Kuettner, R. B. Smith, R. Steinacker, and H. Volkert, The MAP special observing period, *Bull. Am. Met. Soc.*, *82*(3), 433–462, 2000.

Box, G. E. P., and D. R. Cox, An analysis of transformations, *J. R. Statist. Soc. B*, *26*, 211–252, 1964.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day Press, San Francisco, 1976.

Brath, A., P. Burlando, and R. Rosso, Sensitivity analysis of real-time flood forecasting to on-line rainfall predictions, in F. Siccadi and R. L. Bras (Eds.), *Selected Papers of the Workshop on Natural Disasters in European Mediterranean Countries*, Colombella, Perugia, Italy, 1988, pp. 469–488.

Brockwell, P. J., and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag, New York, 1991.

Brown, G. R., and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, 3rd ed., Wiley, New York, 1996.

Burlando, P., and R. Rosso, Stochastic models of temporal rainfall: Reproducibility, estimation and prediction of extreme events, in J. Marco-Segura, R. Harboe, and J. D. Salas (Eds.), *Stochastic Hydrology and Its Use in Water Resources Systems Simulation and Optimization*, Kluwer, Dordrecht, 1993, pp. 137–173.

Burlando, P., R. Rosso, L. Cadavid, and J. D. Salas, Forecasting of short-term rainfall using ARMA models, *J. Hydrol.*, *144*, 193–211, 1993.

Burlando, P., A. Montanari, and R. Ranzi, Forecasting of storm rainfall by combined use of radar, rain gages and linear models, *Atmos. Res.*, *42*, 199–216, 1996.

Burn, D. H., and E. A. McBean, River flow forecasting model for Sturgeon river, *J. Hydraul. Eng. ASCE*, *111*(2), 316–333, 1985.

Carter, M. M., and J. B. Elsner, A statistical method for forecasting rainfall over Puerto Rico, weather and forecasting, *12*(3), 515–525, 1997.

Carter, M. M., J. B. Elsner, and S. P. Bennett, A quantitative precipitation forecast experiment for Puerto Rico, *J. Hydrol.*, *239*, 162–178, 2000.

Chiew, F. H. S., T. C. Piechota, J. A. Dracup, and T. A. McMahon, El Niño/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting, *J. Hydrol.*, *204*, 138–149, 1998.

Chiu, C. L. (Ed.), *Applications of Kalman Filter to Hydrology, Hydraulics and Water Resources*, University of Pittsburgh Press, Pittsburgh, PA, 1978.

Croley II, T. E., *Using Meteorology Probability Forecasts in Operational Hydrology*, American Society of Civil Engineer (ASCE) Press, 2000.

Croley II, T. E., Using NOAA's new climate outlooks in operational hydrology, *ASCE J. Hydrol. Eng.*, *1*(3), 93–102, 1996.

Croley II, T. E., Water resource predictions from meteorological probability forecasts, in D. Rosbjerg et al., *Proceedings of the Sustainability of Water Resources Under Increasing Uncertainty*, IAHS Publication 240, IAHS Press, Institute of Hydrology, Wallingford, Oxfordshire, 1997, pp. 301–309.

Croley II, T. E., and K. Kunkel, Application of the new NWS climate outlook in operational hydrology, in *Proceedings of the Thirteenth Conference on Probability and Statistics in Atmospheric Sciences*, American Meteorological Society, San Francisco, CA, 1996, pp. 231–238.

Dahale, S. D., and P. V. Puranik, Climatology and predictability of the spatial coverage of 5-day rainfall over Indian subdivisions, *Int. J. Climatol.*, *20*(4), 443–453, 2000.

Day, G., Extended streamflow forecasting using NWSRFS, *ASCE J. Water Res. Planning Mgmt.*, *111*(2), 157–170, 1985.

de Jager, J. M., A. B. Potgieter, and W. J. van den Berg, Framework for forecasting the extent and severity of drought in maize in the Free State Province of South Africa, *Agric. Syst.*, *57*(3), 351–365, 1998.

Delleur, J. W., and M. L. Kavvas, Stochastic models for monthly rainfall forecasting and synthetic generation, *J. Appl. Meteorol.*, *17*, 1528–1536, 1978.

Deo, M. C., and K. Thirumalaiah, Real time forecasting using neural networks, in R. S. Govindaraju and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 53–72.

Entekhabi, D., I. Rodriguez-Iturbe, and P. S. Eagleson, Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulse model: Parameter estimation validation, *Water Resour. Res.*, *25*(2), 295–302, 1989.

Fraedrich, K., and K. Müller, On single station forecasting: Sunshine and rainfall Markov chains, *Beitr. Phys. Atmos.*, *56*, 108–134, 1983.

Francis, R. I. C. C., and J. A. Renwick, A regression-based assessment of the predictability of New Zealand climate anomalies, *Theor. Appl. Climatol.*, *60*, 21–36, 1998.

French, M. N., R. L. Bras, and W. F. Krajewski, A Monte-Carlo study of rainfall forecasting with a stochastic model, *Stochast. Hydrol. Hydraul.*, *6*(1), 27–45, 1992a.

French, M. N., W. F. Krajewski, and R. R. Cuykendall, Rainfall forecasting in space and time using a neural network, *J. Hydrol.*, *137*, 1–31, 1992b.

French, M. N., and W. F. Krajewski, A model for real-time quantitative rainfall forecasting using remote sensing. 1. Formulation, *Water Resour. Res.*, *30*(4), 1075–1083, 1994.

French, M. N., H. Andrieu, and W. F. Krajewski, A model for real-time quantitative rainfall forecasting using remote sensing. 1. Formulation, *Water Resour. Res.*, *30*(4), 1085–1097, 1994.

Georgakakos, K. P., A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 2. Case studies, *Water Resour. Res.*, *22*(13), 2096–2106, 1986.

Glahn, H. R., and D. A. Lowry, The use of model output statistics (MOS) in objective weather forecasting, *J. Appl. Meteorol.*, *11*, 1203–1211, 1972.

Govindaraju, R. S., and A. R. Rao, *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000.

Gray, W. M., W. L. Christofer, P. W. Mielke, and K. R. J. Berry, Predicting Atlantic Basin seasonal tropical cyclone activity by 1 June, *Weather Forecast.*, *9*, 103–115, 1994.

Grecu, M., and W. Krajewski, Simulation study of the effects of model uncertainty in variational assimilation of radar data on rainfall forecasting, *J. Hydrol.*, *239*(1–4), 85–96, 2000.

Gregory, J. M., T. M. Wigley, and P. D. Jones, Determining and interpreting the order of a two-state Markov chain: Application to models of daily precipitation, *Water Resour. Res.*, *28*(5), 1443–1446, 1992.

Gupta, H. V., K. Hsu, and S. Sorooshian, Effective and efficient modeling for streamflow forecasting, in R. S. Govindaraju, and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 7–22.

Guetter, A. K., and K. P. Georgakakos, Are the El Niño and La Niña predictors of the Iowa River seasonal flow, *J. Appl. Meteorol.*, *35*, 690–705, 1996.

Guhathakurta, P., M. Rajeevan, and V. Thapliyal, Long range forecasting Indian summer monsoon rainfall by a hybrid principal component neural network model, *Meteorol. Atmos. Phys.*, *71*(3–4), 255–266, 1999.

Halinter, J. P., and J. D. Salas, Short-term forecasting of snowmelt runoff using ARMAX models, *Water Resour. Bull.*, *24*(5), 1083–1089, 1988.

Hamlet, A. F., and D. P. Lettenmaier, Columbia River streamflow forecasting based on ENSO and PDO climate signals, *ASCE J. Water Resour. Planning Mgmt.*, *125*(6), 333–341, 1999.

Hsieh, W. W., and B. Tang, Applying neural network models to prediction and data analysis in meteorology and oceanography, *Bull. Am. Meteorol. Soc.*, *79*, 1855–1870, 1998.

Jimenez, C., A. I. McLeod, and K. W. Hipel, Kalman filter estimation for periodic autoregressive-moving average models, *Stochastic Hydrol. Hydraul.*, 227–240, 1989.

Jinno, K., A. Kawamura, R. Berndtsson, M. Larson, and J. Niemczynowicz, Real-time rainfall prediction at small space-time scales using a 2-dimensional stochastic advection-diffusion model, *Water Resour. Res.*, *29*(5), 1489–1504, 1993.

Johnson, E. R., and R. L. Bras, Multivariate short-term rainfall prediction, *Water Resour. Res.*, *16*(1), 173–185, 1980.

Jury, M. R., Statistical analysis and prediction of KwaZulu-Natal climate, *Theor. Appl. Climatol.*, *60*(1–4), 1–10, 1998.

Kalman, R. E., A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng. Ser. D*, *82*, 35–45, 1960.

Kalman, R. E., and R. S. Bucy, New results in linear filtering and prediction theory, *Trans. ASME J. Basic Eng. Ser. D*, *83*, 95–107, 1961.

Kayha, E., and J. A. Dracup, US streamflow patterns in relation to the El Niño Southern Oscillation, *Water Resour. Res.*, *29*, 2491–2503, 1993.

Kayha, E., and J. A. Dracup, The influences of type 1 El Niño and La Niña events on streamflows in the Pacific Southwest of the United States, *J. Climatol.*, *7*, 965–976, 1994.

Katz, R. W., On some criteria for estimating the order of a Markov chain, *Technometrics*, *23*(3), 243–249, 1981.

Kawamura, A., K. Jinno, R. Berndtsson, and T. Furukawa, Parameterization of rain cell properties using an advection-diffusion model and rain gage data, *Atmos. Res.*, *42*, 67–73, 1996.

Kawamura, A., K. Jinno, R. Berndtsson, and T. Furukawa, Real-time tracking of convective rainfall properties using a two-dimensional advection-diffusion model, *J. Hydrol.*, *203*(1–4), 109–118, 1997.

Kuligowski, R. J., and A. P. Barros, Experiments in short-term precipitation forecasting using artificial neural networks, *Monthly Weather Rev.*, *126*, 470–482, 1998.

Labadie, J. W., R. C. Lazaro, and D. M. Morrow, Worth of short-term rainfall forecasting for combined sewer overflow control, *Water Resour. Res.*, *17*(5), 1489–1497, 1981.

Lardet, P., and C. Obled, Real-time flood forecasting using a stochastic rainfall generator, *J. Hydrol.*, *162*(3–4), 391–408, 1994.

Lettenmaier, D. P., and E. F. Wood, Hydrologic forecasting, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993.

Liu, Z., J. B. Valdés, and D. Entekhabi, Merged forecasts of drought index anomalies along the Gulf Coast in the US using multiple precursors, *Exper. Long-Lead Forecast Bull.*, *6*(2), 9–11, 1997.

Liu, Z., J. B. Valdés, and D. Entekhabi, Merging and error analysis of regional hydrometeorologic anomaly forecasts conditioned on climate precursors, *Water Resour. Res.*, *34*(8), 1959–1969, 1998.

Lowry, D. A., and H. R. Glahn, An operational model for forecasting probability of precipitation—PEATMOS PoP, *Monthly Weather Rev.*, *104*, 221–232, 1976.

Luk, K. C., J. E. Ball, and A. Sharma, A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting, *J. Hydrol.*, *227*(1–4), 56–65, 2000.

Makarau, A., and M. R. Jury, Predictability of Zimbabwe summer rainfall, *Int. J. Climatol.*, *17*(13), 1421–1432, 1997.

Miller, A. J., and L. M. Leslie, Short-term single station forecasting of precipitation, *Monthly Weather Rev.*, *112*, 1198–1205, 1984.

Montanari, A., P. Burlando, and R. Rosso, Forecasting of short-term rainfall using multivariate ARMA models, *Annal. Geophys.*, *12*(sp. issue), C325–C409, 1994 (abstract).

Navone, H. D., and H. A. Ceccatto, Predicting Indian monsoon rainfall—a neural network approach, *Climate Dynam.*, *10*(6–7), 305–312, 1994.

Ngan, P., and S. O. Russell, Example of flow forecasting with Kalman filter, *ASCE J. Hydraul. Eng.*, *112*(9), 818–832, 1986.

Nguyen, V. T. V., M. B. McPherson, and J. Rousselle, *Urban Water Resource Research Program*, Technical Memo 35, American Society of Chemical Engineers, New York, 1978.

Obeysekera, J. T. B., G. Q. Tabios III, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

O'Connell, P. E. (Ed.), *Real Time Hydrological Forecasting and Control*, Institute of Hydrology, Wallingford, England, 1980.

Phanartzis, C. A., Rainfall prediction, Progress Report Wastewater Program, City and County of San Francisco, CA, 1979.

Ramirez, J. A., and R. L. Bras, Conditional distributions of Neyman-Scott models for storm arrivals and their use in irrigation control, *Water Resour. Res.*, *21*, 317–330, 1985.

Refsgaard, J. C., Validation and intercomparison of different updating procedures for real-time forecasting, *Nordic Hydrol.*, *28*, 65–84, 1997.

Rodriguez-Iturbe, I., and P. S. Eagleson, Mathematical models of rainstorm events in space and time, *Water Resour. Res.*, *23*(1), 181–190, 1987.

Sahai, A. K., M. K. Soman, and V. Satyan, All India summer monsoon rainfall prediction using an artificial neural network, *Climate Dynam.*, *16*(4), 291–302, 2000.

Salas, J. D., M. Markus, and A. S. Tokar, Streamflow forecasting based on artificial neural networks, in R. S. Govindaraju, and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 23–52.

Sansó, B., and L. Guenni, A nonstationary multisite model for rainfall, *J. Am. Statist. Assoc.*, *95*(452), 1089–1100, 2000.

Schaake, J. C., and L. Larson, A strategy for ensemble streamflow prediction (ESP), *Proceedings of the American Meteorological Society Annual Meeting*, Phoenix, AZ, Vol. J104-J105, 1997.

Scott, D. W., *Multivariate density estimation: Theory, practice and visualisation*. Probability and Mathematical Statistics Series, Wiley, New York, 1992.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—a non parametric probabilistic forecast model, *J. Hydrol.*, *239*(1–4), 249–258, 2000.

Stengel, R. F., *Stochastic Optimal Control Theory and Application*, Wiley, New York, 1986.

Stone, R. C., G. L. Hammer, and T. Marcussen, Prediction of global rainfall using phases of the Southern Oscillation index, *Nature*, *384*, 252–255, 1996.

Sugimoto, S., E. Nakakita, and S. Ikebuchi, A stochastic approach to short-term rainfall prediction using a physically based conceptual rainfall model, *J. Hydrol.*, *242*(1–2), 137–155, 2001.

Takasao, T., and M. Shiiba, Development of techniques for on-line forecasting of rainfall and flood runoff, *Natural Disaster Sci.*, *6*(2), 83–112, 1984.

Takasao, T., M. Shiiba, and K. Takara, Stochastic state-space techniques for flood runoff forecasting, in *Proc. Pacific Int. Seminar on Water Resour. Systems*, Tomanu, Japan, 1989, 117–132.

Thapliyal, V., Preliminary and final long range forecast for seasonal monsoon rainfall over India, *J. Arid Environ.*, *36*(3), 385–403, 1997.

Tong, H., Determination of the order of a Markov chain by Akaike's information criterion, *J. Appl. Prob.*, *12*, 488–497, 1975.

Toth, E., A. Brath, and A. Montanari, Comparison of short-term rainfall prediction models for real-time flood forecasting, *ASCE J. Hydrol.*, *239*(1–4), 132–147, 2000.

Trotta, P. D., J. W. Labadie, and N. S. Grigg, Automatic control strategies for urban stormwater, *ASCE J. Hydraul. Div.*, *103*(HY12), 1977.

Twedt, T. M., J. C. Schaake, Jr., and E. L. Peck, National Weather Service extended streamflow prediction, *Proc. 45th Western Snow Conference*, Albuquerque, NM, April 1977, pp. 52–57.

von der Heydt, L., L. E. Brazil, and K. Jawed, ABPA realtime hydrologic forecasting for the Columbia River Basin, in *Proceedings of the 21st Annual Conference of the ASCE Water Resources Planning and Management Division*, Denver, CO, 1994.

Wood, E.F. and P.E. O'Connell, "Real Time Forecasting" in *Hydrological Forecasting*, M. G. Anderson and T. P. Burt (Eds), John Wiley & Sons Inc., 1985, 505–558.

Yu, P. S., and T. C. Yang, A probability-based renewal rainfall model for flow forecasting, *Nat. Hazards*, *15*(1), 51–70, 1997.

Zawadzki, I., Fractal structure and exponential decorrelation in rain, *J. Geophys. Res.*, *92*(D8), 9586–9590, 1987.

CHAPTER 35

REMOTE SENSING AND GEOGRAPHICAL INFORMATION SYSTEMS APPLICATIONS IN HYDROLOGY

EDWIN T. ENGMAN AND NANDISH MATTIKALLI

1 INTRODUCTION

Remote sensing and associated image-processing technology provide access to spatial and temporal hydrologic information from watershed to global scales. Advances in sensor and imaging technology are increasing the capability of remote sensing for specific hydrologic application.

There are two general areas where remote sensing can be used in hydrologic modeling: (1) determining watershed geometry, drainage network, and other map-type information for distributed hydrologic models and for empirical flood peak, annual runoff, or low-flow equations; and (2) providing input data such as snow cover or precipitation, diagnostic variables such as soil moisture or surface temperature, or model parameters such as delineated land-use classes used to define runoff coefficients. In this review, the latter is addressed. The various uses of remote sensing to provide input data and diagnostic variables for hydrologic models are treated as they are used to measure the different hydrologic variables or processes, e.g., precipitation, snow, or evaporation. Each of these hydrologic variables or processes are discussed individually with the emphasis on how remote sensing is being used, and not on the technology as far as sensor details and specific instruments are concerned. More details can be found in two recent books on this general subject (Engman and Gurney, 1982; Schultz and Engman, 2000).

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

Finally, the current developments and hydrologic applications of integrated geographical information systems (GIS) technology are presented. Management and efficient utilization of large spatial data volumes is going to be one of the major challenges of the coming decades. GIS have the capability to efficiently store, manipulate, retrieve, and analyze spatially referenced data. This is the primary reason why GIS are becoming popular among the hydrological community to develop new types of hydrological models and to modify existing models to incorporate widely available spatial data.

2 PRECIPITATION

Recognizing the practical limitations of rain gages for measuring spatially averaged rainfall over large areas and inaccessible areas, hydrologists have increasingly turned to remote sensing as a means for quantifying the precipitation input, especially in areas where there are few surface gages. Because the fundamental approach to measuring rainfall and snow are different with respect to remote sensing, snow is discussed separately.

Direct measurement of rainfall from satellites for operational purposes has not been generally feasible because the presence of clouds prevents observation of the precipitation directly with visible, near-infrared and thermal-infrared sensors. However, improved analysis of rainfall can be achieved by combining satellite and conventional gage data. Satellite data are most useful in providing information on the spatial distribution of potential rain-producing clouds, and gage data are most useful for accurate point measurements. Although ground-based radar, which is a remote-sensing technique, has advanced to an operational stage for locating regions of heavy rain and for estimating rainfall rates, it will not be discussed in this chapter.

Useful data can be derived from satellites used primarily for meteorological purposes, including polar orbiters such as the NOAA-N series and the Defense Meteorological Satellite Program, and from geostationary satellites such as *GOES* (Geostationary Orbiting Environmental Satellite), *GMS*, and *Meteosat*. However, their visible and infrared images can provide information only about the cloud tops rather than cloud bases or interiors. Since these satellites provide frequent observations (even at night with thermal sensors), the characteristics of potentially precipitating clouds and the rates of changes in cloud area and shape can be observed. From these observations, estimates of rainfall can be made that relate cloud characteristics to instantaneous rainfall rates and cumulative rainfall over time. For example, Strubing and Schultz (1983) have developed a runoff regression model that is based on Barrett's (1970) indexing technique. The cloud area and temperature are the satellite variables used to develop a temperature-weighted cloud cover index. This index is then transformed linearly to mean monthly runoff. Rott et al. (1986) also developed a daily runoff model using *Meteosat* data for a cloud index. Schultz (1994) has demonstrated the use of the infrared channel from *Meteosat* to estimate monthly rainfall volumes using a modified Arkin approach (Papadakis et al., 1992). The monthly rainfall data were transformed into monthly runoff volumes for the 16,000 km^2 Tano basin in West Africa using a model based on a series of nonlinear reservoirs. The results were reasonably good and certainly adequate for water resources planning. For the practicing hydrologist, satellite rainfall methods are most valuable when there are no or very few surface gages for measuring rainfall.

Tsonis et al. (1996) investigated the ability of visible and infrared satellite data to produce rainfall estimates for input to the National Weather Service river forecast model that had been calibrated with rain gage data. They found good correlations with gage data for areas over 10,000 km^2. In a companion study, Guetter et al. (1996), using the satellite-derived rainfall estimates produced streamflow and soil moisture estimates using the river forecast model. They concluded that flow simulation accuracy is sensitive to basin scale with better results being produced from larger basins. Derived soil moisture estimates were similar to those simulated with gaged data for the surface layer but lower for the deep soil moisture.

3 SNOW HYDROLOGY

Snow is a form of precipitation; in hydrology it is treated somewhat differently because of the lag between when it falls and when it produces runoff and groundwater recharge, and is involved in other hydrologic processes. Remote sensing is a valuable tool for obtaining snow data for predicting snowmelt runoff as well as climate studies. Nearly all regions of the electromagnetic spectrum provide useful information about the snowpack. Depending on the need, one may like to know the areal extent of the snow, its water equivalent, or the "condition" or grain size, density, and presence of liquid water within the snowpack. Although no single region of the spectrum provides all these properties, techniques have been developed to provide all of the properties to some degree or other.

The water content of snow can be measured from low elevation aircraft carrying sensitive γ-radiation sensors. This approach is limited to low aircraft altitudes (approximately 150 m) because the atmosphere attenuates a significant portion of the γ radiation. Currently, this operational program covers over 1400 flight lines annually in the United States and Canada. This method is effective for measuring snow cover in open plains, but is less effective in more hilly terrain or when there is extensive forest cover. Use of satellite data for snow mapping has become operational in several regions of the world. Currently, the National Oceanic and Atmospheric Administration (NOAA) develops snow cover maps for about 3000 river basins in North America of which approximately 300 are mapped according to elevation for use in streamflow forecasting (Carroll, 1990). NOAA also produces regional and global maps of mean monthly snow cover.

Microwave remote sensing offers great promise for future applications to snow hydrology. This is because the microwave data can provide information on the snowpack properties of most interest to hydrologists, i.e., snow cover area, snow water equivalent (or depth), and the presence of liquid water in the snowpack, which signals the onset of melt. With the availability of satellite microwave data Scanning

Multichannel Microwave Radiometer (SSMR) and Special Sensor Microwave/ Imager (SSM/I), algorithms have been developed for estimating snow water equivalent for dry snow and mapping the depth and global extent of snow cover (Chang et al., 1987). The passive microwave systems are limited by their interaction with other media such as forest areas, although a method to correct for the absorption of the snow signal by forest cover has been developed (Chang et al., 1991). The spatial resolution attainable by the passive satellite systems is also a limitation but Rango et al. (1989) have shown that that reasonable snow water equivalent estimates can be made on basins smaller than 10,000 km^2.

Active microwave remote sensing also has the potential to provide important information about the snowpack at very high resolution with synthetic aperture radar (SAR) (Stiles et al., 1981; Rott, 1986). Unfortunately, analysis of radar data is more complex than passive microwave data and, until very recently, there have been no orbiting SAR systems for collecting snow data. In spite of that, aircraft and shuttle SAR measurements have shown that SAR can discriminate between snow and glaciers from other targets and discriminate between wet and dry snow (Shi and Dozier, 1992, 1995).

Snowmelt runoff procedures that use remote sensing can be grouped into empirical approaches and modeling. Early use of remote sensing focused on empirical relationships between snow cover area or percent snow cover and monthly or accumulated runoff. These simple relationships work very well for some applications, particularly in data-sparse regions of the world. The snowmelt runoff model (SRM) (Martinec et al., 1983) was specifically developed for using remote sensing of snow cover by elevation zone as the primary input variable. Although SRM uses a simple degree-day melt model, it applies the model to the different elevation zones to account for the areal distribution of the snow. SRM has been extensively tested on basins of different sizes and regions of the world. Although SRM is a degree-day model that uses only snow cover as remote-sensing derived input, this model has been recently modified to include a simple snowmelt energy budget algorithm (Kustas et al., 1994). This model has been tested against lysimeter data and suggests that the radiation-based snowmelt factor may improve runoff predictions at the basin scale.

4 SOIL MOISTURE

Recent advances in remote sensing have shown that soil moisture can be measured by a variety of techniques. However, only microwave technology has demonstrated a quantitative ability to measure soil moisture under a variety of topographic and vegetation cover conditions so that it could be extended to routine measurements from a satellite system.

The major factor inhibiting widespread use of remotely sensed soil moisture data in hydrology is the lack of data sets and optimal satellite systems. For the most part, scientists have been restricted to data from short-duration aircraft campaigns or analysis of the SMMR and SSM/I passive microwave satellites. Although the available passive systems do not have the optimum wave lengths for soil moisture, research has demonstrated that in areas of sparse vegetation a valuable estimate can be obtained (Owe et al., 1988). Historical data from the SSMR passive microwave system is more valuable than the SSM/I data because it had a C-band radiometer, which is a better instrument for soil moisture (Owe et al., 1992); however, its period of record is limited to 1982 to 1987. In both cases the footprint is rather large, varying from about 25 km for the SSM/I to about 150 km for the C-band SMMR.

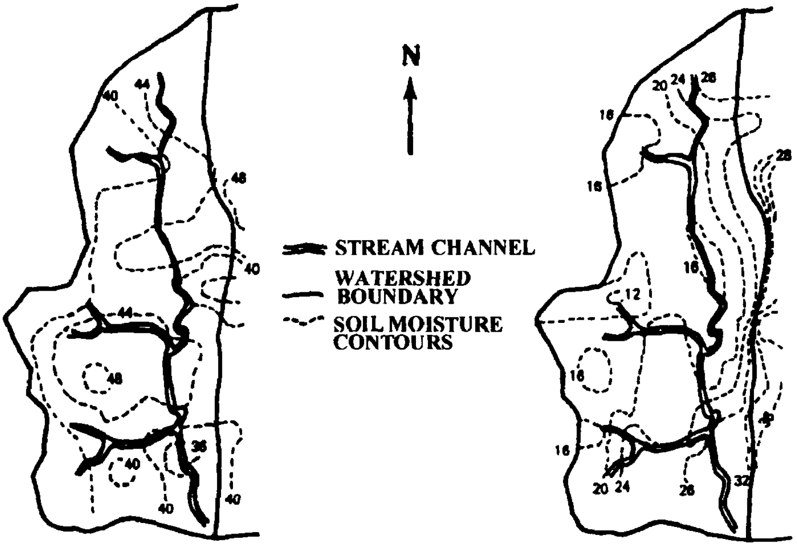
The SAR systems offer perhaps the best opportunity to measure soil moisture routinely over the next few years. Currently, the European Resources Satellite (*ERS-1*) C-band and Japan Environmental Resources Satellite (*JERS-1*) L-band SARs and the Canadian *RADARSAT* (also C-band) are operational. Although it is believed that an L-band system would be optimum for soil moisture, the preliminary results from the *ERS-1* C-band radar demonstrate its capability as a soil moisture instrument. One main drawback to the existing SAR systems is that there are no existing algorithms for the routine determination of soil moisture from single-frequency, single-polarization radars. A second limitation comes from their long period between repeat passes; for the most part it is 35 to 46 days, although the *RADARSAT* has 3-day capability for much of the globe in a SCANSAR (wide swath, 500 km) mode.

There continues to be speculation about the potential value for soil moisture as an input variable in hydrologic models, either to establish the initial conditions for simulating storm runoff or as a descriptor of hydrologic processes. To date there has been more promise than substance, but initial progress is beginning to appear as some of the aircraft experimental data become available.

Aircraft data taken during the Fist ISLSCP Field Experiment (FIFE) campaign were used to map the spatial pattern of soil moisture resulting from drainage and ET in a 37.7-ha watershed (Wang et al., 1989). These patterns, shown in Figure 1, were seen to match the results of a simple slab model and identified the region contributing base flow to the channel (Engman et al., 1989). Attempts to use passive microwave measurements in a small watershed showed good correlation with the ground data and may yield a reliable technique for calibrating the model (Wood et al., 1993). Even the relatively low-resolution passive data can improve the water budget calculations of a small basin (Lin et al., 1994). Goodrich et al. (1994) studied the prestorm soil moisture at various scales of basin runoff. They concluded that initial values were important but that the resolution of the final remote-sensing product was not a limitation.

The value of remotely sensed soil moisture data in a semidistributed hydrology model was demonstrated using data from the 1992 Washita microwave experiment. Initializing the surface soil moisture fields with the Electronically Steered Thinned Array Radiometer (ESTAR) L-band microwave data produced more accurate model predictions of soil moisture changes and absolute values than those produced from the model initialized with streamflow data (O'Neill and Hsu, 1997).

The feasibility of synthesizing distributed fields of remotely sensed soil moisture by the four-dimensional data assimilation applied to a hydrological model, TOPLATS, has been explored (Houser et al., 1998) with several alternative assimilation schemes. The synthetic soil moisture fields were assimilated from remote-



28 MAY 1987

4 JUNE 1987

Figure 1 Temporal and spatial patterns of soil moisture in a small drainage basin illustrating the drying pattern (after Wang et al., 1989).

sensing soil moisture data and the output of a soil–vegetation–atmosphere scheme. The spatially distributed hydrology model's descriptive ability was improved with the assimilation of the soil moisture data.

5 EVAPOTRANSPIRATION

In general, remote-sensing techniques cannot measure evaporation or evapotranspiration directly. However, remote sensing does have two potentially very important roles in estimating evapotranspiration. First, remotely sensed measurements offer methods for extending point measurements or empirical relationships, such as the Thornthwaite (1948), Penman (1948), and Jensen and Haise (1963) methods, to much larger areas, including those areas where measured meteorological data may be sparse. Secondly, remotely sensed measurements may be used to measure variables in the energy and moisture balance models of evapotranspiration. Although there has been progress made in the direct remote sensing of the atmospheric parameters that affect evapotranspiration, such as the Rahman LIDAR, this is essentially a ground-based, point measurement and will not be covered in this report.

The question of how to use the spatial nature of remote-sensing data to extrapolate point evapotranspiration measurements to a more regional scale has been addressed in several ways. Using the temperature sounders on the meteorological satellites in a linear regression model, Davis and Tarpley (1983) estimated shelter temperatures with an error of about 2 K for clear or partly cloudy conditions. Price (1982) used thermal data from the Heat Capacity Mapping Mission (HCMM) to estimate regional-scale evapotranspiration rates, which were found to be comparable to pan evaporation data. Jackson (1985) and Gash (1987) have proposed an analytical framework for relating the horizontal changes in evaporation to horizontal changes in surface temperature. Kustas et al. (1990) demonstrated these concepts for an agricultural area under clear sky conditions. Humes et al. (1994) has proposed a simple model using remotely sensed surface temperatures and reflectances for extrapolating energy fluxes from a point to a regional scale; however, other than for clear sky conditions, variations in incoming solar radiation, meteorological conditions, and surface roughness limit this approach.

Several variables related to the energy balance equation can be measured by remote sensing and simple meteorological measurements. Generally, the latent heat term is determined as the residual of the other terms in the energy balance. Incoming solar radiation can be estimated from satellite observations of cloud cover, primarily from geosynchronous satellites (Brakke and Kanemasu, 1981; Tarpley, 1979). Pinker and Laszlo (1992) have proposed a model that infers incoming short-wave fluxes and surface albedos from *GOES* data. Pinker et al. (1994) used this model to demonstrate that incoming shortwave radiation can be measured quite accurately, even under variable cloud conditions, at the basin scale.

For clear sky conditions, the surface albedo may be estimated by measurements covering the entire visible and near-infrared waveband, while empirical relations using narrow spectral bands can be used to determine albedo over heterogeneous surfaces (Jackson, 1985; Brest and Goward, 1987). Although albedo estimated this way is not the true hemispherical albedo, lack of directional data or simple models make this correction not feasible under most applications.

Surface temperature can be estimated from measurements in thermal infrared wavelengths, that is, the 10.5- to 12.5-μm waveband, either assuming a surface emissivity (close to unity for natural surfaces) or having measured values of the surface emissivity. Surface temperatures can be used to estimate the outgoing long-wave radiation term in the net radiation equation (Kustas et al., 1994).

The soil heat flux term can be estimated with remote-sensing measurements. A simplified approach defines the ratio of soil heat flux to net radiation in terms of vegetation cover, which, in turn, is determined from visible and near-infrared measurements (Clothier et al., 1986; Choudhury et al., 1987; Kustas and Daughtry, 1990). The diurnal effects (Owe and van de Grind, 1990) and influence of soil moisture (Brutsaert, 1982) are assumed to be secondary for large areas (Kustas et al., 1994).

The sensible heat flux can be estimated using several approaches, including the bulk resistance approach proposed by Monteith (1973) and similarity principles for the unstable boundary layer (Brutsaert and Sugita, 1992), where the surface temperatures are measured by remote sensing. These approaches have met with varying degrees of success (Hall et al., 1992; Brutsaert and Sugita, 1992; Brutsaert et al., 1993; Kustas et al., 1994).

Additional approaches for estimating ET from remote sensing data are being explored. Ottle et al. (1989) have shown how satellite-derived surface temperatures can be used to estimate ET and soil moisture in a model that has been modified to use these data. Mauser (1990) has shown how multitemporal Système Probatoire pour l'Observation de la Terre (*SPOT*) and Thematic Mapper (TM) data to derive plant parameters for estimating ET in a GIS-based model. Later, Mauser (1996) used Advanced Very High Resolution Radiometer (AVHRR) thermal data to validate an actual ET mesoscale model by comparing them to the surface temperature distributions. Soares et al. (1988) demonstrated how thermal infrared and C-band radar could be used to estimate bare soil evaporation. Choudhury et al. (1994) have shown strong relationships between evaporation coefficients and vegetative indices.

6 RUNOFF

One of the first applications of remote-sensing data in hydrologic models used *Landsat* data to determine both urban and rural land use for estimating runoff coefficients (Jackson et al., 1976). Land use is an important characteristic of the runoff process that affects infiltration, erosion, and evapotranspiration. Distributed models, in particular, need specific data on land use and its location within the basin. Most of the work on adapting remote sensing to hydrologic modeling has involved the Soil Conservation Service (SCS) runoff curve number model (U.S. Department of Agriculture, 1972) for which remote-sensing data are used as a substitute for land cover maps obtained by conventional means (Jackson et al., 1977; Bondelid et al., 1982).

In remote-sensing applications, one seldom duplicates detailed land-use statistics exactly. For example, a study by the Corps of Engineers (Rango et al., 1983) estimated that an individual pixel may be incorrectly classified about one-third of the time. However, by aggregating land use over a significant area, the misclassification of land use can be reduced to about 2%, which is too small to affect the runoff coefficient or the resulting flood statistics.

Studies have shown (Jackson et al., 1977) that for planning studies the *Landsat* approach is cost effective. The authors estimated that the cost benefits were on the order to 2.5 to 1 and can be as high as 6 to 1, in favor of the *Landsat* approach. These benefits increase for larger basins or for multiple basins in the same general hydrological area. Mettel et al. (1994) demonstrated the recomputation of Probable Maximum Flood (PMFs) for the Au Sable River using HEC-1 and updated and detailed land-use data from Landsat TM resulted in 90% cost cuts in upgrading dams and spillways in the basin.

7 WATER AND ENERGY BALANCE MODELS

In recent work, Dubayah and Lettenmaier (1997) have attempted to maximize the use of remote-sensing data as drivers for a large-scale coupled water and energy balance model. They used the VIC-3L model (Liang et al., 1994) applied to the

Arkansas–Red River basins in the Southern Great Plains in the United States. There were two objectives to this research: (1) to develop and test a land surface hydrologic model capable of using remote-sensing data and (2) to develop and test algorithms for generating data from remote-sensing measurements. Remote-sensing data were obtained from *GOES* (solar radiation), AVHRR (downwelling long-wave radiation, air temperature, surface humidity, and vegetation), Normalized Difference Vegetation Index (NDVI), Leaf Area Index (LAI) (canopy interception, and canopy resistance).

The VIC-3L model was used to simulate the water and energy fluxes for the month of June, 1987. The model was first used with ground-based data (from 26 meteorological stations) and then with the remote-sensing-derived data. Comparison of the results yielded differences of as much as 40% in net radiation, 15% in latent heat, and 100% in sensible heat. For such studies, the results from the ground-based data are not necessarily correct; it is not known whether the remote-sensing or the ground-based data give the correct results. Thus it really could not be determined if the remote-sensing data resulted in an improvement or a degradation in the water and energy fluxes. However, the remote-sensing simulations did provide a spatial pattern that appears to provide more information about the distribution of the fluxes than do the ground-based measurements.

8 GEOGRAPHICAL INFORMATION SYSTEMS

Geographical information systems provide appropriate methods for efficient storage, retrieval, manipulation, analysis, and display of large volumes of spatially referenced data. Accordingly, GIS consist of four basic components: data input and editing, storage of geographic databases, data analysis and spatial modeling, and data visualization and presentation (Fig. 2). The data may be collected from fieldwork, extraction of map data, air photo interpretation, and interpretation and classification of remotely sensed images. Data input may be carried out by manual digitization or computer-assisted semiautomatic methods. The data are organized into a series of spatially co-registered layers, with each layer relating to a particular theme or a set of layers relating to temporal variation of a theme.

Data input and structuring is one of the most time-consuming and expensive tasks in the creation of a GIS. Remotely sensed data can be put to the best use if they are incorporated in GIS. A GIS, therefore, when combined with up-to-date data from a remote-sensing system, can assist in the automation of several operations (e.g., interpretation, change detection, map revisions). The hydrological system is a dynamic entity; the information stored in a GIS is only a static representation of the real world and therefore has to be updated for the temporal coverage (i.e., a third dimension) on a regular basis. Remotely sensed satellite data offer an excellent input in this context to provide repetitive, synoptic, and accurate information of the changes of a watershed.

Integration of GIS with hydrological models is necessary to better explain the complexity of hydrological processes arising from spatial heterogeneity of inPut

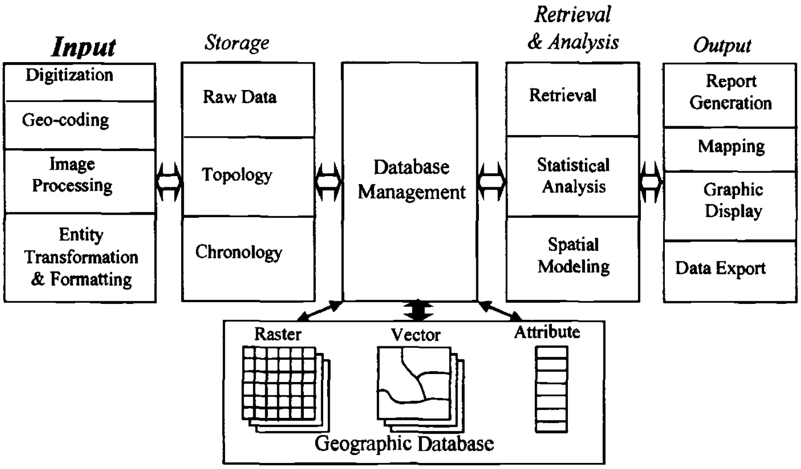


Figure 2 Submodules of a GIS for input, storage, retrieval, analysis, modeling, and presentation of spatial and nonspatial data.

parameters such as topography, soil types and characteristics, vegetation, and antecedent soil conditions. A GIS can be employed to supply physical and hydrological parameters to a hydrological model. Model simulation results can be analyzed using a GIS. This requires a continuous flow of information to and from both the GIS and hydrological model. One of the useful possibilities of linking GIS with hydrological modeling is the capability of dynamic spatial visualization of the model simulation results, including user interaction. Real-time or near-real-time visualization of simulated hydrological processes could greatly improve existing analysis of simulation results.

Integration of Remotely Sensed Data into GIS

Remote sensing can be incorporated into the system in a variety of ways: as a measure of land use, impervious surfaces, for providing initial conditions for flood forecasting, and for monitoring flooded areas (Neumann et al., 1990). The GIS allows for the combining of other spatial data forms such as topography, soils maps as hydrologic variables such as rainfall distributions, or soil moisture. This approach was demonstrated by Kouwen et al. (1993) where their grouped response unit (GRU) included satellite-based land use and lies within a computational element that may be either a sub-basin or and area of uniform meteorological forcing. In HYDROTEL, Fortin and Bernier (1991) propose combining *SPOT* DEM (digital elevation model) data with satellite-derived land use and soils mapping data to define homogeneous hydrologic units (HHU). In a study of the impact of land-use change on the Mosel River Basin, Ott et al. (1991) and Schultz (1993) have defined hydrologically similar units (HSU) by DEM data, soils maps, and satellite-derived land use. They also used satellite data to determine a vegetation index (NDVI) and a leaf water content index (WCI), which are combined to delineate areas where a subsurface supply of water is available to vegetation. The distribution of microwave remotely sensed near-surface (0–5-cm deep) soil moisture was analyzed to identify areas of high soil moisture gradients (Mattikalli et al., 1998). This analysis showed a direct correlation between soil moisture dynamics and soil texture. Soil moisture data were employed in a hydrological model linked to a GIS, to predict subsurface hydraulic conductivity.

Remote-sensing systems use raster format for collection and acquisition of data. Many of the commonly employed GIS systems (e.g., Arc/Info) mainly use the vector format to store data layers. In this format, data are collected as points, lines, and polygons, where each structure holds information for a specific region (Fig. 3). Both the vector and raster structures have advantages and difficulties that are well described in the literature (e.g., Peuquet, 1984; Burrough, 1990), yet their fundamental differences make the integration a complicated task (Piwowar et al., 1990). In the recent past, many commercial GIS have been adapted to offer raster image display and handling capabilities (e.g., Arc/Info Version 6.0 or later), and several others offer both raster and vector capabilities (e.g., GRASS). The integration of remotely sensed data with GIS data occurs naturally in a raster GIS because data structures are approximately the same for both sources. In a vector system, the integration requires more effort, and several technical problems need to be overcome for the true integration. Important problems in the integration are the raster/vector dichotomy, generalization, and accuracy of digital information (Piwowar et al., 1990; Lunetta et al., 1991). Although the raster/vector dichotomy is a major impediment for a true integration, a significant advancement has been made to resolve the issue (e.g., McKeown, 1987; Conese et al., 1992; van der Laan, 1992; Westmoreland and Stow, 1992). These studies have employed a variety of approaches including use of quadtrees, object-oriented methods, knowledge-based systems, expert systems, artificial intelligence, etc. to achieve the task of true integration (Fritsch, 1992; Molenaar and Janssen, 1992). Examples of some commercially available systems

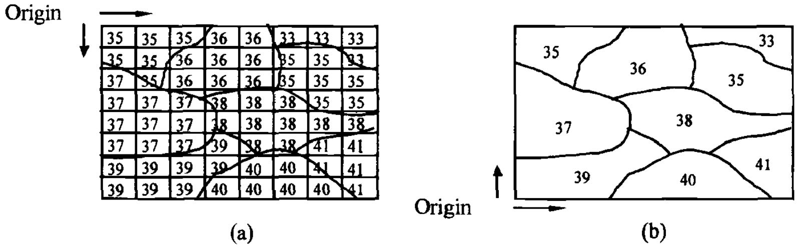


Figure 3 Representation of spatial data in a GIS: (a) raster-formatted data consists of a sequence of orderly placed pixels (or picture elements) and (b) vector-formatted data consists of polygon entities to represent features.

that have some integration capabilities include GRASS, Arc/Info Version 6.0 onwards, ERDAS IMAGINE, and PCI.

Integration of raster and vector data types requires an efficient raster-to-vector (and/or vice versa) conversion routine. Mattikalli et al. (1995) developed a methodology for the separate but equal type of integration, in which the key process is a raster-to-vector (and vice versa) conversion. The procedure makes use of some built-in routines commonly available in most vector GIS, and some intermediate data formats, viz. lattice and SVF (single variable file). This approach has been employed by Mattikalli (1995) to integrate remotely sensed satellite data derived from both fine- and coarse-resolution sensors with digitized map data.

GIS and Digital Terrain Models

Digital terrain modeling is one of the strong areas where GIS has been widely utilized in hydrology. Digital elevation model (DEM) data are employed to derive watershed characteristics such as slope, aspect, curvature, drainage network structure (e.g., Fairfield and Leymarie, 1991), hydrologic response units (HRUs), and also to delineate watershed boundaries (Band, 1986; Jensson and Domingue, 1988; Schultz, 1994). Lozar (1992) delineated drainage paths and watersheds of the entire Earth based on the 5-arc-minute DEM of Earth's land surface. Remotely sensed information can be integrated with DEM for a variety of hydrologic applications. For example, Dubayah (1992) employed a DEM, *Landsat* TM data, and a radiative transfer algorithm to model spatial variability of net solar radiation at fine spatial resolution. Also, remotely sensed products are employed in conjunction with a DEM to produce realistic perspective views of a watershed that aid visualization and understanding of spatial and temporal variability of hydrological parameters (Gugan and Dowman, 1988).

GIS and Hydrologic Models

Watershed database development usually is the first important stage in a hydrologic modeling study. Remotely sensed data might be employed to generate thematic maps and also to serve as map basis when no other reliable data are available. *Landsat* TM and *SPOT* images data are suitable for production of digital map at scales ranging from 1 : 50,000 to 1 : 100,000 (Welch et al., 1985; Swann et al., 1988; Gugan and Dowman, 1988; Konecny et al., 1988). Base maps, produced from remote sensing and integrated within a GIS, hold promise in terms of greater reliability, i.e., lower meta-uncertainty (uncertainty about uncertainty) for map information because errors are known and tracked throughout the map generation process. Overlaying, merging, and performing map calculations are key GIS features often used in many hydrological applications. Schultz (1993) presents an example in which soil water storage information was derived by merging plant root depth data (derived from land-use classification of *Landsat* image) and soil porosity data (derived from digitized soil maps).

Historically, runoff modeling at the river basin scale has lumped rainfall, infiltration, and other hydraulic parameters to apply everywhere in the basin. With the advent of distributed modeling, a basin is subdivided into computational elements at a smaller scale. A distributed simulation model allows a user to simulate spatially variable parameters without lumping. However, setting up such a model with spatially distributed data and parameters is a time-consuming and laborious task. If a GIS is integrated with the model, these chores become much easier and often transparent to the user. An additional advantage of integrating distributed numerical models with a GIS includes calculation and display of runoff flow depths across watershed sub-basins.

The runoff curve number (CN) approach (USDA, 1972) to rainfall–runoff modeling is appealing for an integrated remote-sensing and GIS environment. This approach estimates volume of direct runoff (Q) in terms of volume of rainfall (P) and potential maximum storage (S), which is derived from the CN, a coefficient that is directly related to watershed land use, land management, and soil properties. Since land use can be routinely monitored using remote sensing, it is possible to analyze the effects of land-use changes (e.g., urbanization) on watershed runoff. Figure 4 shows various stages of computation of this approach implemented within a GIS. Mattikalli et al. (1996) employed Arc/Info to store various input parameters as thematic layers and generated flood hydrographs in a predominantly rural watershed. This approach has also been used to generate single-event flood hydrographs and synthetic flood frequency curves (Muzik and Chang, 1993).

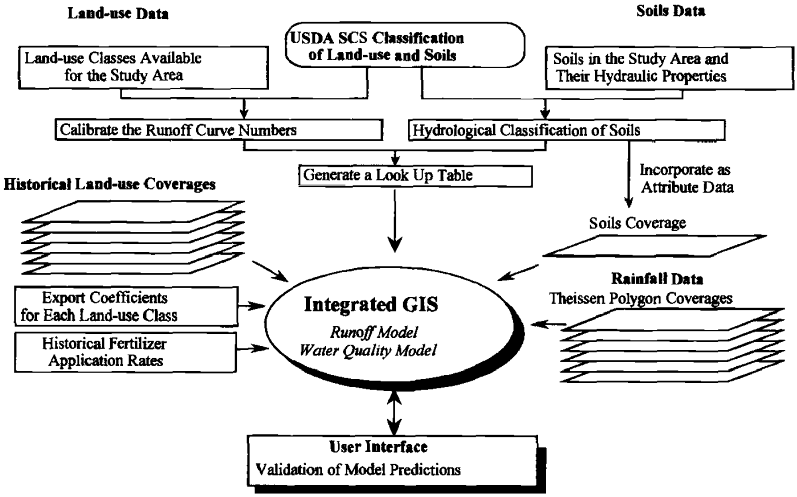

Land-use Data
Soils Data
USDA SCS Classification of Land-use and Soils
Land-use Classes Available for the Study Area
Soils in the Study Area and Their Hydraulic Properties
Calibrate the Runoff Curve Numbers
Hydrological Classification of Soils
Generate a Look Up Table
Historical Land-use Coverages
Incorporate as Attribute Data
Soils Coverage
Rainfall Data
Theissen Polygon Coverages
Export Coefficients for Each Land-use Class
Historical Fertilizer Application Rates
Integrated GIS
Runoff Model
Water Quality Model
User Interface
Validation of Model Predictions

Figure 4 Schematic diagram of a GIS approach for prediction of river discharge using the SCS curve numbers and water quality using the export coefficient model (Mattikalli et al., 1996).

In urban watersheds, the spatial analysis capabilities of a GIS can be used for hydrological analysis. Watershed attributes such as soils information (infiltration rates, hydraulic conductivity, and storage capacities), surface characteristics (pervious, impervious, slope, roughness), geometry and dimensions of flow planes, routing lengths (overland, gutter, and sewer), and geometry and characteristics of routing segments can be efficiently stored and utilized for urban runoff calculations. Most of the earlier studies have used GISs to derive parameters of lumped models. For example, Johnson (1989) used GIS for the generation of input data for a digital map-based modeling system that supports lumped parameter models such as unit hydrograph, time–area, and cascade of reservoirs. The advent of distributed modeling and powerful GIS allows modelers to simulate spatially variable parameters. To date, several hydrological models such TOPMODEL and CREAMS have been integrated to operate within GIS environments (Chairat and Delleur, 1993; Romanowicz et al., 1993). Moeller (1991) used GIS to determine input parameters for the HEC-1 model, Sircar et al. (1991) used a GIS to determine time–area curves. Djokic and Maidment (1991) used Arc/Info with the rational method to determine inlet and pipe capacity of an urban storm sewer system. Kim and Ventura (1993) used a GIS to manage and manipulate the land-use data for modeling the non-point-source pollution of an urban basin using an empirical urban water quality model. Greene and Cruise (1996) employed Arc/Info GIS to derive urban watershed feature attributes (location coordinates, parameters of runoff generating polygons, gutters and storm drains) for input into hydrologic modeling procedures to estimate runoff. Vieux (1991) developed a method for modeling direct surface runoff using a combination of the finite-element method and GIS. Schultz (1994) presents three different examples on hydrological modeling using remote sensing in the framework of ILWIS and Arc/Info GIS. These examples demonstrate merging of *Landsat* TM and *Meteosat* geostationary image products and ancillary data (viz. DEM and its derived products) stored in a GIS for rainfall/runoff modeling and water balance parameter computation at 30 m, 5 km, and at HRU spatial scales. Mattikalli et al. (1996) employed the runoff curve number (CN) approach to compute direct runoff depth and its spatial and temporal variations based on historic remotely sensed data within a GIS framework.

Water quality modeling applications using remote sensing and GIS have concentrated mainly on non-point source (NPS) pollution. To date, several water quality models (AGNPS, ANSWERS, USLE, export coefficient model, etc.) have been interfaced with GIS. The spatially distributed agricultural non-point-source (AGNPS) model integrated with GIS (Srinivasan and Engel, 1994) allows modelers to handle each point source, pesticide, and channel information in a decision support system, WATERSHEDSS (Water, Soil, and Hydro-Environmental Decision Support System) (Osmond et al., 1997). Using such a system, one can determine critical areas within a watershed and evaluate effects of alternative land treatment scenarios on water quality. Mattikalli et al. (1996) implemented an export coefficient model within a vector-based GIS to quantify spatial and temporal changes of total nitrogen loading in surface water as a response to changes in watershed land use, management, and fertilizer application rates. Although this method is based on empirical export coefficients derived from the literature, more accurate coefficients can be derived by inverse solution to a physical based model.

Management and modeling of groundwater and its quality have also been explored (e.g., Maidment, 1993; Merchant, 1994). In the majority of studies, spatial models designed to evaluate groundwater vulnerability for contamination have been implemented in GIS. However, these approaches have not employed data derived from remote sensing, probably because of the specific nature of the input parameters. The models need to be adapted to incorporate remotely sensed products and then implemented within a GIS.

Monitoring and/or prediction soil erosion computed using the universal soil loss equation (USLE) is another application of integrated GIS (e.g., Pelletier, 1985). Slope steepness (S) and slope length (L) factors are derived using DEM, and rainfall factors are assigned using the triangular irregular network (TIN) structure for the rainfall gaging stations. Erosion control practice and land-use/land-cover (or cropping management) factors are estimated using *Landsat* (Multi-Spectal Scanner (MSS) and TM) and SPOT sensor data via land-use/land-cover classification and associated land management information (Jurgens and Fander, 1993). In the revised USLE (Renard et al., 1991), the L factor has been modified for influence of profile convexity/concavity using segmentation of irregular slopes of a complex terrain. Mitasova et al. (1996) integrated regularized spline with tension for computation of S and L factors and used a unit stream power and directional derivative approach for modeling spatial distribution of areas with topographic potential for erosion or deposition.

GIS and Spatial Analysis

The synoptic nature of remote-sensing data offers an excellent opportunity to identify spatial characteristics of land surface changes and other hydrologic variables. Analysis of spatial variability is performed using different techniques including Monte Carlo methods (Fisher, 1991) and geostatistical techniques such as semivariograms and kriging (Oliver and Webster, 1990).

Three- and Four-Dimensional GIS

To date, applications have recognized the importance of change of hydrologic processes or input model parameters over time. Modern GIS have capabilities of traditional two-dimensional (2D) GIS to perform spatial analysis as well as the ability to handle and visualize third dimension (such as depth) and time as a fourth dimension (Fisher, 1993). Three-dimensional (3D) GIS are suitable for many applications in hydrology such as predictive hydrogeological modeling (Raper, 1989). Three-dimensional GIS lend themselves to the iterative process of modeling as well as the evolutionary nature of site characterization and remediation. Although most current 3D GIS provide some solutions for complex subsurface processes, they are still at the visualization stage rather than true modeling or interpretation. No one system yet meets all the needs of an ideal modeling environment, hence integration between multiple systems is desired. Also, four-dimensional (4D) GIS do not adequately represent the temporal dimension (Langran, 1992) because no GIS currently adequately handles chronology. We typically illustrate the effects of temporal change as slices of time for discrete intervals, but we need to show dynamic change over continuous time. Although many GIS can generate a 3D diagram, no commercial system has 3D geometry and topology such that disparate databases can be integrated in three dimensions as well as they are in two dimensions. The ultimate solution would be able to handle change in time as well as change in space. An ideal GIS handling time as a fourth dimension (4D GIS) will have chronology treated much like topology; before and after taking on the same importance as left and right in 2D space or above and below in 3D space. Such 4D GIS would be of immense value for a number of research areas in hydrology including soil moisture modeling, groundwater modeling, etc. because of their inherent four-dimensional nature.

9 ARY AND CONCLUSIONS

Continuing high spatial resolution data from the *Landsat* and *SPOT* satellites, passive microwave data from the special sensor/microwave imager (SSM/I) and continuing meteorological satellite coverage from the *NOAA*, *GOES*, *GMS*, and *Meteosat* series all mean that the remotely sensed techniques can continue to be employed and expanded upon. New sensors, particularly in the microwave region, promise great potential for hydrologic applications. There are several satellites, such as *ERS-1/2* launched by the European Space Agency, the *J-ERS-1* launched by the Japanese, and *RADARSAT* launched by the Canadians that will provide useful data for hydrologists. All carry single-polarization, single-frequency SARs. An additional satellite being planned that will have considerable hydrologic interest is the Tropical Rainfall Measurement Mission (TRMM) (Simpson et al., 1988).

The EOS (Earth Observational System) (Butler, 1988), and its counterpart European and Japanese platforms, will lead to considerable advances in the understanding of all the earth sciences, including hydrology. The EOS instruments of most interest to hydrologists would include the MODIS and AMSR; the latter is a microwave instrument with a C-band radiometer that should provide interesting data of the land surface moisture conditions. EOS also includes the organization of the data and of other earth science in an information system where time series of all the data will be readily available is also important. This data system will allow many types of data to be used simultaneously to calibrate or be assimilated into numerical models.

Future progress in the hydrological sciences will depend a great deal upon the availability of adequate data for model development and validation. Remote sensing can and should play a pivotal role in this progress. Without it, it is very possible that future progress in the hydrological sciences will be severely retarded if not completely stopped. With it, hydrological sciences should be able to advance rapidly and to successfully address some of the previously intractable problems. An issue that must be addressed is the modification or development of new models to specifically use remote-sensing data. For the most part, existing models have not been developed to effectively use remote-sensing data. A second point is that ground-based data are frequently available at shorter time intervals than remote-sensing data. This becomes important when simulating processes that are driven by the diurnal cycle.

Another very important issue that needs to be addressed by the hydrologic and remote-sensing communities is validation. We should not automatically assume that ground-based measurements provide the "truth." Ground-based data have an inherent weakness in that they are point measurements being applied to large, inhomogeneous areas. There is a need to develop innovative approaches to validate not only the remote-sensing-derived products but also the application of water and energy balance models to large regions.

While remote-sensing systems generate large volumes of valuable spatial data, GIS offer an appropriate technology not only for efficient storage and retrieval of spatially referenced data but also for data manipulation and spatial analysis required in distributed hydrologic modeling. With the advent of an EOS suite of platforms and sensors, it is expected that the volume of data being received will require the use of fully integrated spatial information systems supported by knowledge-based techniques in all facets of data handling.

Future development of GIS application is controlled by the state of technology, and therefore assessing the probable developments is a difficult task. At the present time, the current level of integrated applications utilizes an environment largely free of the logistical considerations of data transfer between remote sensing and GIS. Over the next few years, more efforts need to be focused on the fundamental aspects of integration such as data generalization and accuracy specification. Several problems of a true integration of remote sensing and GIS could probably be solved by recognition that GIS and remote-sensing systems process and manage spatial information at different levels of representation. Ultimately, GIS and remote sensing should be viewed as one entity that will be concerned with handling and analyzing spatial hydrologic data. The unification of these technologies will lead to a synergistic integration of spatial data handling, and the final system would have more application capabilities than just the sum of the two.

REFERENCES

Band, L. E., Topographic partition of watersheds with digital elevation models, *Water Resour. Res.*, *22*(1), 15–24, 1986.

Barrett, E. C., The estimation of monthly rainfall from satellite data, *Monthly Weather Rev.*, *98*, 322–327, 1970.

Bondelid, T. R., T. J. Jackson, and R. H. McCuen, Estimating runoff curve numbers using remote sensing data, in *Proceedings of the International Symposium on Rainfall-Runoff Modeling, Applied Modeling in Catchment Hydrology*, Water Resources Publications, Littleton, CO, 1982, pp. 519–528.

Brakke, T. W., and E. T. Kanemasu, Insolation estimation from satellite measurements of reflected radiation, *Remote Sensing Environ.*, *11*, 157–167, 1981.

Brest, C. L., and S. N. Goward, Deriving surface albedo measurements from narrow band satellite data, *Int. J. Remote Sensing*, *8*, 351–367, 1987.

Brutsaert, W. H., *Evaporation into the Atmosphere: Theory, History and Application*, Reidel, Boston, MA, 1982.

Brutsaert, W. H., and M. Sugita, Regional surface fluxes from satellite-derived surface temperatures (AVHRR) and radiosonde profiles, *Boundary-layer Meteorol.*, *58*, 355–366, 1992.

Brutsaert, W. H., A. Y. Hsu, and T. J. Schmugge, Parameterization of surface heat fluxes above forest with satellite thermal sensing and boundary-layer soundings, *J. Appl. Meteorol.*, *32*(5), 910–917, 1993.

Burrough, P. A., *Principles of Geographical Information Systems for Land Resources Assessment*, Clarendon, Oxford, 1990.

Butler, D., *From Pattern to Process: The Strategy of the Earth Observing System*, NASA, Washington, DC, 1988.

Carroll, T. R., Airborne and satellite data used to map snow cover operationally in the U.S. and Canada, in *Proceedings of the International Symposium on Remote Sensing and Water Resources*, Enschede, The Netherlands, 1990, pp. 147–155.

Conese, C., G. Maracchi, F. Maselli, M. Romani, and L. Bottai, Integration of remotely sensed data into a GIS for the assessment of land suitability, *EARSeL Adv. Remote Sensing*, *1*, 173–179, 1992.

Chang, A., J. Foster, and D. K. Hall, NIMBUS-7 derived global snow cover parameters, *Ann. Glaciol.*, *9*, 39–44, 1987.

Chang, A. T. C., J. L. Foster, and A. Rango, Utilization of surface cover composition to improve the microwave determination of snow water equivalent in a mountainous basin, *Int. J. Remote Sensing*, *12*, 2311–2319, 1991.

Chairat, S., and J. W. Delleur, Integrating a physically based hydrological model with GRASS, in *HydroGIS 93*, IASH Publ. No. 211, 1993, pp. 143–150.

Choudhury, B. J., S. B. Idso, and R. J. Reginato, Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared-temperature based energy balance equation, *Agric. Forest Meterol.*, *39*, 283–297, 1987.

Choudhury, B. J., N. U. Ahmed, S. B. Idso, R. J. Reginato, and C. S. T. Daughtry, Relations between evaporation coefficients and vegetation indices studied by model simulations, *Rem. Sens. Environ.*, *50*(1), 1–17, 1994.

Clothier, B. E., K. L. Clawson, P. J. Pinter, Jr., M. S. Moran, R. J. Reginato, and R. D. Jackson, *Agric. Forest Meteorol.*, *37*, 75–88, 1986.

Davis, P. A., and J. D. Tarpley, Estimation of shelter temperatures from operational satellite sounder data, *J. Climatol. Appl. Meteorol.*, *22*, 369–376, 1983.

Djokic, D., and D. R. Maidment, Terrain analysis for stormwater modeling, *HP*, *5*(1), 115–124, 1991.

Dubayah, R., Estimating net solar radiation using Landsat Thematic Mapper and digital elevation data, *Water Resour. Res.*, *28*(9), 2469–2484, 1992.

Dubayah, R., and D. Lettenmaier, Combining remote sensing and hydrologic modeling for applied water and energy balance studies, paper presented at NASA EOS Interdisciplinary Working Group Meeting, San Diego, CA, 1997.

Engman, E. T., and R. J. Gurney, *Remote Sensing in Hydrology*, Chapman & Hall, London, 1982.

Engman, E. T., G. Angus, and W. P. Kustas, Relationship between the Hydrologic balance of a small watershed and remotely sensed soil moisture, in *Proceedings of the IAHS Third International Assembly,* IAHS Publ. No. 186, Baltimore, 1989, pp. 75–84.

Fairfield, J., and P. Leymarie, Drainage networks from grid digital elevation models, *Water Resourc. Res., 27,* 709–711, 1991.

Fisher, T. R., Integrating three-dimensional geoscientific information system (GSIS) technologies for groundwater and contaminant modeling, in *HydroGIS 93,* IAHS Publ. No. 211, 1993, pp. 235–241.

Fisher, T. R., and R. Q. Wales, Rational splines and multi-dimensional geologic modeling, in R. Pflug and J. W. Harbaugh (eds.), *Three Dimensional Computer Graphics in Modeling Geologic Struuctures and Simulating Process,* Springer-Verlag, Heidelberg, 1991, pp. 17–28.

Fortin, J.-P., and M. Bernier, Processing remotely sensed data to derive useful input data for hydrotel hydrologic model, in *Proc. IGARS,* Houston, TX, 1991.

Fritsch, D., Analysis of remote sensing data in geographical information systems, *EARSeL Adv. Remote Sensing, 1,* 60–65, 1992.

Gash, J. H. C., An analytical framework for extrapolating evaporation measurements by remote sensing surface temperature, *Int. J. Remote Sensing, 8*(8), 1245–1249, 1987.

Goodrich, D. C., T. J. Schmugge, T. J. Jackson, C. L. Unkrich, T. O. Keefer, R. Parry, L. B. Bach, and S. A. Amer, Runoff simulation sensitivity to remotely sensed initial soil water content, *Water Resour. Res., 30*(5), 1393–1405, 1994.

Greene, R. G., and J. F. Cruise, Development of a geographic information system for urban watershed analysis, *Photogrametric Engineering and Remote Sensing, 62*(7), 863–870, 1996.

Guetter, A. K., K. P. Georgakakos, and A. A. Tsonis, Hydrologic applications of satellite data, Part 2. Flow simulations and soil water estimates, *J. Geophys. Res. Atmos., 101*(D21), 26527–26538, 1996.

Gugan, D. J., and I. J. Dowman, Accuracy and completeness of topographic mapping from SPOT imagery, *Photogram. Rec., 12*(72), 787–796, 1988.

Hall, F. G., K. F. Humerick, S. J. Goetz, P. J. Sellers, and J. E. Nickerson, Satellite remote sensing of surface energy balance: success, failures and unresolved issues, in FIFE (First ISLSCP Field Experiment) Special Issue *J. Geophys. Res., 97*(D17), 19061–19090, 1992.

Houser, P. R., W. J. Shuttleworth, H. V. Gupta, J. S. Famiglietti, K. H. Syed, and D. C. Goodrich, Integration of soil moisture remote sensing and hydrologic modeling using data assimilation, *Water Resour. Res., 34*(12), 3405–3420, 1998.

Humes, K. S., W. P. Kustas, and M. S. Moran, Use of remote sensing and reference site measurements to estimate instantaneous surface energy balance components over a semiarid rangeland watershed, *Water Resour. Res., 30*(5), 1363–1373, 1994.

Jackson, R. D., Estimating evapotranspiration at local and regional scales, *IEEE Trans. Geosci. Remote Sensing, GE-73,* 1086–1095, 1985.

Jackson, T. J., R. M. Ragan, and R. P. Shubinski, Flood frequency studies on ungaged urban watersheds using remotely sensed data, in *Proceedings of the National Symposium on Urban Hydrology, Hydraulics and Sediment Control,* University of Kentucky, Lexington, KY, 1976, pp. 31–39.

Jackson, T. J., R. M. Ragan, and W. N. Fitch, Test of Landsat-based urban hydrologic modeling, *ASCE J. Water Resourc. Planning Mgmt. Div., 103* (No. WR1), 141–158, 1977 (Proc. Papers 12950).

Jensen, S. K., and J. O. Domingue, Extracting topographic structure from digital elevation data for geographical information system analysis, *Photogram. Eng. Remote Sensing*, *54*, 1593–1600, 1988.

Jensen, M. E., and H. R. Haise, Estimating evapotranspiration from solar radiation, *Proc. Am. Soc. Civil Eng., J. Irrig. Drain. Div.*, *89*, 15–41, 1963.

Johnson, L. E., MAPHYD—A digital map based hydrologic modeling system, *Photogrametric Engineering and Remote Sensing*, *55*(6), 911–917, 1989.

Jurgens, C., and M. Fander, Soil erosion assessment and simulation by means of SGEOS and ancillary digital data, *Int. J. of Remote Sensing*, *14*(15), 2847–2855, 1993.

Kim, K., and S. Ventura, Large-scale modeling of urban nonpoint source pollution using a geographical information system, *Photogrametric Engineering and Remote Sensing*, *59*(10), 1539–1544, 1993.

Konecny, G., K. Jacobsen, P. Lohmann, and W. Muller, Comparison of high resolution satellite imagery, in *Proceedings of the 16th Congress of the Int. Soc. Photogrametry and Remote Sensing*, Kyoto, Japan, B9/IV, 1988, pp. 226–237.

Kouwen, N., E. D. Soulis, A. Pietroniro, J. Donald, and R. A. Harrington, Grouped response units for distributed hydrologic modeling, *J. Water Resourc. Planning Mgmt.*, *119*(3), 289–305, 1993.

Kustas, W. P., M. S. Moran, R. D. Jackson, L. W. Gay, L. F. W. Duell, K. E. Kunkel, and A. D. Matthias, Instantaneous and daily values of the surface energy balance over agricultural fields using remote sensing and a reference field in an arid environment, *Remote Sensing Environ.*, *32*, 125–141, 1990.

Kustas, W. P., M. S. Moran, K. S. Humes, D. I. Stannard, P. J. Pinter, L. E. Hipps, E. Swiatek, and D. C. Goodrich, Surface energy balance estimates at local and regional scales using optical remote sensing from an aircraft platform and atmospheric data collected over semiarid rangelends, *Water Resourc. Res.*, *30*(5), 1241–1259, 1994.

Langran, G., *Time in GIS*, Taylor and Francis, New York, 1992.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, A simple hydrologically based model on land surface water and energy fluxes for general circulation models, *J. Geophys. Res.*, *99*, 14415–14428, 1994.

Lin, D.-S., E. F. Wood, J. S. Famiglietti, and M. Mancini, Impact of microwave derived soil moisture on hydrologic simulations using a spatially distributed water balance model, in *Proceedings of the Sixth International Symposium on Physical Measurements and Signatures in Remote Sensing*, Val d'Isere, France, 1994.

Lozar, R. C., Global Climate Management by Watershed Basin Units, Construction Engineering Research Laboratory, U.S. Army Corps of Engineers, Champaign, IL, 1992.

Maidment, D. R., GIS and hydrologic modeling, in M. Goodchild, B. Parks, and L. Steyaert (Eds.), *Environmental Modeling with GIS*, Oxford University Press, New York, 1993, pp. 147–167.

Martinec, J., A. Rango, and E. Major, *The Snowmelt-Runoff Model (SRM) User's Manual*, NASA Ref. Publ. 1100, National Aeronautics and Space Administration, Washington, DC, 1983.

Mattikalli, N. M., Integration of remotely sensed raster data with vector based geographical information system for land-use change detection, *Int. J. Remote Sensing*, *16*(15), 2813–2828, 1995.

Mattikalli, N. M., B. J. Devereux, and K. S. Richards, Integration of remotely sensed satellite images with a geographical information system, *Comput. Geosci.*, *21*(8), 947–956, 1995.

Mattikalli, N. M., B. J. Devereux, and K. S. Richards, Prediction of river discharge and surface water quality using an integrated geographical information system approach, *Int. J. Remote Sensing*, *17*(4), 683–701, 1996.

Mattikalli, N. M., E. T. Engman, T. J. Jackson, and L. R. Ahuja, Microwave remote sensing of temporal variations of brightness temperature and near-surface soil water content during a watershed-scale field experiment, and its application to the estimation of soil physical properties, *Water Resour. Res.*, *34*(9), 2289–2299, 1998.

Mauser, W., Modeling the spatial variability of soil moisture and evapotranspiration with remote sensing data, in *Proceedings of the IAH Symposium on Remote Sensing and Water Resources*, Enschede, 1990, pp. 249–260.

Mauser, W., Mesoscale modeling of evapotranspiration using remote sensing data, *Proc. Europto Series*, Int. Soc. for Photo. Optical Engineering Vol. 2959, Taormina, Italy, 1996, pp. 108–117.

McKeown, D., The role of artificial intelligence in the integration of remotely sensed data with geographic information systems, *IEEE Trans. Geosci. Remote Sensing*, *25*, 330–348, 1987.

Merchant, J. W., GIS-based groundwater pollution hazard assessment: a critical review of the DRASTIC model, *Photogrametric Engineering and Remote Sensing*, *60*(9), 1117–1127, 1994.

Mettel, C., D. McGraw, and S. Strater, Money saving model, *Civil Eng.*, *64*(1), 55–56, 1994.

Mitasova, H., J. Hofierka, M. Zlocha, and L. Iverson, Modeling topographic potential for erosion and deposition using GIS, *Int. J. of Geographical Information Systems*, *10*(5), 629–641, 1996.

Moeller, R. A., Application of a geographic information system to hydrologic modeling using HEC-1, in D. B. Stafford (Ed.), *Civil Engineering Applications of Remote Sensing and GIS*, American Society of Civil Engineers, 1991, pp. 269–277.

Molenaar, M., and L. L. F. Janssen, Integrated processing of remotely sensed and geographic data for land inventory purposes, *EARSeL Adv. Remote Sensing*, *1*, 113–121, 1992.

Monteith, J. L., *Principles of Environmental Physics*, Edward Arnold, London, 1973.

Muzik, I., and C. Chang, *Flood Simulation Assisted by a GIS*, Int. Assoc. Hydrological Sciences Publication No. 211, 1993, pp. 531–539.

Neumann, P., W. Fett, and G. A. Schultz, A Geographic Information System as data base for distributcd hydrological models, in *Proceedings of the International Symposium on Remote Sensing and Water Resources*, Enschede, The Netherlands, August 1990, pp. 781–791.

O'Neill, P. E., and A. Y. Hsu, The impact of microwave-derived surface soil moisture on watershed hydrological modeling, in *1997 Research and Technology Report*, NASA/GSFC, 1997.

Osmond, D. L., R. W. Gannon, J. A. Gale, D. E. Line, C. B. Knott, K. A. Phillips, M. H. Turner, M. A. Foster, D. E. Lehning, S. W. Coffey, and J. Spooner, WATERSHEDSS: A decision support system for watershed-scale nonpoint source water quality problems, *J. Am. Water Resour. Assoc.*, *33*(2), 327–341, 1997.

Ott, M., Z. Su, A. H. Schumann, and G. A. Schultz, Development of a distributed hydrological model for flood forecasting and impact assessment of landuse change in the international Mosel basin, Int. Assoc. Hydrological Sciences Publication No. 201, 1991.

Ottle, C. D., D. Vidal-Madjar, and G. Girard, Remote sensing applications to hydrological modeling, *J. Hydrol.*, *105*, 369–384, 1989.

Owe, M., and A. A. van de Grind, Daily surface moisture model for large area semi-arid land application with limited climate data, *J. Hydrol.*, *121*, 119–132, 1990.

Owe, M., A. T. C. Chang, and R. E. Golus, Estimating surface soil moisture from satellite microwave measurements and satellite derived vegetation index, *Remote Sensing Environ.*, *24*, 331–345, 1988.

Owe, M., A. A. van de Grind, and A. T. C. Chang, Surface soil moisture and satellite microwave observations in semiarid southern Africa, *Water Resour. Res.*, *28*(3), 829–839, 1992.

Papadakis, I., J. Napiorkowski, and G. A. Schultz, Monthly runoff generation by nonlinear models using multi-spectral and multi-temporal images, in *Remote Sensing Oceanogr. Hydrol. Agric. Adv. Space Res.*, *13*(5), 181–186, 1992.

Pelletier, R. E., Evaluating non-point pollution using remotely sensed data in soil erosion models, *Journal of Soil and Water Conservation*, *40*, 332–335, 1985.

Penman, H. L., Natural evaporation from open water, bare soil and grass, *Proc. R. Soc. A*, *193*, 129–145, 1948.

Peuquet, D. J., A conceptual framework and comparison of spatial data models, *Cartographica*, *21*(4), 66–113, 1984.

Pinker, R. T., and I. Laszlo, Modeling surface solar irradiation for satellite applications on global scale, *J. Appl. Meteorol.*, *31*, 194–211, 1992.

Pinker, R. T., W. P. Kustas, I. Laszlo, M. S. Moran, and A. R. Huete, Basin-scale irradiance estimates in semi-arid regions using GOES-7, *Water Resourc. Res.*, *30*(5), 1375–1386, 1994.

Piwowar, J. M., E. F. LeDrew, and D. J. Dudycha, Integration of spatial data in vector and raster formats in a geographic information system environment, *Int. J. Geogr. Inf. Syst.*, *4*, 429–444, 1990.

Price, J. C., Estimation of regional scale evapotranspiration through analysis of satellite thermal-infrared data, *IEEE Trans. Geosci. Remote Sensing*, *GE-20*, 286–292, 1982.

Rango, A., A. Feldman, T. S. George III, and R. M. Ragan, Effective use of Landsat data in hydrologic models, *Water Resour. Bull.*, *19*, 165–174, 1983.

Rango, A., J. Martinec, A. T. C. Chang, J. L. Foster, and V. F. van Katwijk, Average areal water content of snow in a mountainous basin using microwave and visible satellite data, *IEEE Trans. Geosci. Remote Sensing*, *27*, 740–745, 1989.

Raper, J. F., The three-dimensional geoscientific mapping and modeling system: A concept design, in J. F. Raper (Ed.), *Three Dimensional Applications in Geographic Information Systems*, Taylor and Francis, London, 1989, 11–19.

Renard, G. K., G. R. Foster, G. A. Weesies, and J. P. Porter, RUSLE—Revised universal soil loss equation, *Journal of Soil and Water Conservation*, *46*, 30–33, 1991.

Romanowicz, R., K. Beven, and J. Freer, *TOPMODEL as an Application Module within WIS*, IAHS Publication No. 211, 1993, pp. 211–223.

Rott, H., *Prospects of Active Remote Sensing for Snow Hydrology, Hydrologic Applications of Space Technology*, IAHS Publ. No. 160, 1986, pp. 215–233.

Rott, H., J. Aschbacher, and K. G. Lenhart, Study of river runoff prediction based on satellite data, European Space Agency Final Report No. 5376, 1986.

Shi, J. C., and J. Dozier, Radar response to snow wetness, in *Proc. International Geoscience and Remote Sensing 92*, CH3041-1, 1992, pp. 927–929.

Shi, J. C., and J. Dozier, Inferring snow wetness using C-band data from SIR-C's polarimetric synthetic aperture radar, *IEEE Trans. Geosci. Remote Sensing*, *33*(4), 905–914, 1995.

Schultz, G. A., Hydrological modeling based on remote sensing information, *Adv. Space Res.*, *13*(5), 149–166, 1993.

Schultz, G. A., Meso-scale modeling of runoff and water balances using remote sensing and other GIS data, *Hydrological Sciences Journal*, *39*(2), 121–142, 1994.

Schultz, G. A., and E. T. Engman (Eds.), *Remote Sensing in Hydrology and Water Management*, Springer, Berlin, 2000.

Simpson, J., R. F. Adler, and G. R. North, A proposed Tropical Rainfall Measuring Mission (TRMM) satellite, *Bull. Am. Meteorol. Soc.*, *69*, 278–295, 1988.

Sircar, J. K., R. M. Ragan, E. T. Engman, and R. A. Fink, A GIS based geomorphic approach for the computation of time-area curves. in D. B. Stafford (Ed.), *Civil Engineering Applications of Remote Sensing and GIS*, American Society of Chemical Engineers, New York, 1991, pp. 287–296.

Soares, J. V., R. Bernard, O. Toconet, D. Vidal-Madjar, and A. Weill, Estimation of bare soil evaporation from microwave measurements, *J. Hydrol.*, *99*, 281–296, 1988.

Srinivasan, R., and B. A. Engel, A spatial decision support system for assessing agricultural nonpoint source pollution, *WRB*, *30*(3), 441–462, 1994.

Stiles, W. H., F. T. Ulaby, and A. Rango, Microwave measurements of snowpack properties, *Nordic Hydrol.*, *12*, 143–166, 1981.

Strubing, G., and A. Schultz, Estimation of monthly river runoff data on the basis of satellite imagery, in *Proc. Hamburg Symposium*, Int. Assoc. Hydrological Sciences Publication No. 145, 1983, pp. 491–498.

Swann, R., D. Hawkins, A. Westwell-Roper, and W. Johnstone, The potential for automated mapping from geocoded digital image data, *Photogram. Eng. Remote Sensing*, *54*(2), 187–193, 1988.

Tarpley, J. D., Estimating incident solar radiation at the surface from geostationary satellite data, *J. Appl. Meteorol.*, *18*, 1172–1181, 1979.

Thornthwaite, C. W., An approach toward a rational classification of climates, *Geophys. Rev.*, *38*, 55–94, 1948.

Tsonis, A. A., G. N. Triantafyllou, and K. P. Georgakakos, Hydrologic applications of satellite data for rainfall estimation, *J. Geophys. Res. Atmos.*, *101*(D21), 26517–26525, 1996.

U.S. Department of Agriculture. Soil conservation service, in *National Engineering Handbook, Section 4, HYDROLOGY,* U.S. Government Printing Office, Washington, DC, 1972.

van der Laan, F. B., Integration of remote sensing in a raster and vector GIS environment, *EARSeL Adv. Remote Sensing*, *1*, 71–80, 1992.

Vieux, B. E., Geographic information systems and non-point source water quality and quantity modelling, *Hydrol. Process.*, *5*, 101d 113, 1991.

Wang, J. R., J. C. Shiue, T. J. Schmugge, and E. T. Engman, Mapping soil moisture with L-band radiometric measurements, *Remote Sensing Environ.*, *27*, 305–312, 1989.

Welch, R., T. R. Jordan, and M. Ehlers, Comparative evaluations of the geodetic accuracy and cartographic potential of Landsat-4/-5 TM image data, *Photogram. Eng. Remote Sensing*, *51*(9), 1249–1262, 1985.

Westmoreland, S., and D. A. Stow, Category identification of changed land-use polygons in an integrated image processing/geographic information system, *Photogramm. Eng. Remote Sensing*, *58*, 1593–1599, 1992.

Wood, E. F., D.-S. Lin, M. Mancini, D. Thongs, P. A. Troch, T. J. Jackson, J. S. Famiglietti, and E. T. Engman, Intercomparisons between passive and active microwave remote sensing, and hydrological modeling for soil moisture, *Adv. Space Res.*, *13*(5), 167–176, 1993.

CHAPTER 36

FLOODS

STEVEN JENNINGS AND EVE GRUNTFEST

1 INTRODUCTION

The interface between humans and hydrologic features across Earth's surface has helped shape human culture. From the earliest agricultural, complex societies established along some of the great rivers of the world to the bustling seaports of today, humans have gained from the myriad advantages of living in proximity to water. Fertile soil, ease of transportation, and availability of resources (both materials and energy) have allowed for the development of complex material and intellectual cultures. The relationship between water and humans also brings a great deal of risk. Flooding is one of these risks. The impact of floods on humans has been evident from *Genesis* to tonight's evening news. Early Mesopotamian maps may have been drawn to facilitate the reestablishment of property lines after flooding. While the impacts of flooding on humans have been positive in the case of fertile floodplains that support much of the world's agricultural productivity, there is the potential for a great deal of negative impact (Brown, 1984; Clark et al., 1985). Losses of life and property have focused the efforts of scientists, engineers, and government agencies on the prediction, control, and mitigation of floods and flood damage.

In spite of efforts to deal with flooding problems, monetary losses continue to rise at an alarming rate. In Venezuela in December 1999, two weeks of heavy rain resulted on December 15th in flash floods laden with soil, vegetation, and debris. Damages of US$3.2 billion, or 3.3% of the country's gross domestic product were reported. At least 20,000 people were killed. Generally, the number of lives lost due to flooding remains high. However, improvements in flood warnings particularly for major large-scale storms such as cyclones and typhoons have had dramatic effects. The severe 1991 cyclone in Bangladesh resulted in 140,000 dead and property losses

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

of US$2.0 billion. A cyclone of similar intensity in 1993 resulted only in the loss of 126 lives. The early warnings and cyclone shelters accounted for the major improvement (*www.ndndr.org*). China has also witnessed a reduction in the number of lives lost to floods. While 3000 people died in 1998 floods, the 1998 floods were as great as those of 1931 and 1954 where the loss of lives was 145,000 and 33,000, respectively (*www.ndndr.org*). In October 1998 hurricane Mitch was the worst in the eastern Caribbean since 1780 when a hurricane killed 22,000 people. The death toll from Mitch is reported as 11,000. More than 3 million people were left homeless or were severely affected (*www.ncdc.noaa.gov/ol/reports/mitch/mitch.html*).

Floods take a variety of forms with the interplay of several factors leading to the inundation of normally dry land. Wohl (2000) identifies four primary challenges in reducing escalating flood damages. These are (1) estimating flood magnitude for a given recurrence interval, (2) accurately forecasting floods based on rapidly evolving weather conditions, (3) effectively operating flood-warning and evacuation procedures, and (4) establishing and enforcing land-zoning regulations. This chapter first discusses the contexts and causes of flooding, the first two points addressed by Wohl (2000). The second topic is the complex human responses to floods, points 3 and 4 of Wohl (2000). Many of these topics are illustrated with examples of floods.

2 DEFINITION OF FLOODS

Streams are linear water features that flow under the impetus of gravity. The amount of water contained in a stream is usually regulated by contributions of groundwater and surface runoff to the stream channel (Zaslavsky and Sinai, 1981; Knighton, 1998). Much of the time water in a stream flows within the confines of its channel. When inputs of water increase sufficiently, stream discharge leaves the stream channel and covers all or parts of the adjacent floodplain. Since the floodplain surface is usually a virtually flat surface and near the elevation of the stream channel, water can easily spread over the floodplain once water exceeds the elevation of the stream's banks. Most floods develop over a period of days or months as discharge increases gradually (Hirschboeck, 1987, 1988). Flash floods by contrast occur suddenly with little warning and are of short duration. Semiarid and arid areas are likely to experience flash floods (Reid and Frostick, 1987; Hassan, 1990). Flooding is not always associated directly with stream channels. Flooding occurs any time when water covers a surface that is normally not under water. Flooding can occur in coastal areas, low lying areas with poor drainage, or locations with inadequate urban drainage systems.

3 FACTORS THAT LEAD TO FLOODING

Floods have a multitude of causes. Some causes are related to what would be considered natural processes that would occur whether humans are present or not. Many causes have been affected by human activities. In some cases the severity of floods and the types of damage are a direct result of agriculture, urbanization, and the areas selected for development. In all cases flooding is related to increased discharge in stream channels.

Saturated Soil

Much of Earth's surface is covered by a weathered cover of regolith. Whether forming a true soil with well-developed horizons or a weakly developed detrital cover, the regolith is composed of a mix of mineral particles, organic fragments, and pore space. Commonly, much of the pore space is filled with air and to a lesser extent water. When large amounts of precipitation are received in a region, the pore space fills with water as the input of water from precipitation exceeds the output of water from the soil column to the water table. Decreases in infiltration lead to increases in runoff. The lag time between the precipitation event and the arrival of water to stream channels decreases significantly when soil saturation occurs. As a result, peak discharge increases significantly and the likelihood of overbank flow is high (Smith and Ward, 1998). Spatially, soil saturation may occur over large-scale basins, which leads to flooding in large areas. The peak discharge flows downstream and becomes concentrated in higher order streams causing flooding. In many cases saturation follows a period of high amounts of precipitation over a prolonged time period, possibly weeks or months (Wolman and Gerson, 1978; Ward and Robinson, 1990).

Basin Characteristics

Surface characteristics influence infiltration and runoff rates (Roberts, 1989; Kuhnle et al., 1996). Impervious surfaces such as exposed bedrock or a paved road accelerate surface runoff, thus decreasing lag time between the precipitation event and entrance of water into a nearby channel. Urbanized areas, therefore, with large percentages of impervious surface such as roofs, streets, and parking lots coupled with an engineered drainage system designed to move water quickly to stream channels greatly increase the chances that some flooding will occur after a significant precipitation event (Wolman, 1967; Hammer, 1972; Roberts, 1989; Newson, 1992). Conversely, rural areas with large areas of soil, natural vegetation, and the potential for a faster infiltration rate are less likely to have significant flooding resulting from a single precipitation event. Removal of as much as half the forest cover and a decrease of marsh land along the Yangtze River in China has led to increased flooding. Half a billion people, or 45% of China's total population, reside on the banks or floodplains of the Yangtze and the area produces about 42% of China's gross domestic product. In 1998, 79.6 million people in three Chinese provinces were affected by repeat flooding on the Yangtze. The floods killed more than 3000 people. Fourteen million people were evacuated and 21 million were made homeless (*Weather.ou.edu/spark/AMON/v2_n3/News/DR_980819China12.html*).

Topography

Topography will influence the rate at which precipitation will be incorporated as stream discharge (Patton, 1988). Steep, rocky canyon walls have low infiltration rates as well as a great deal of potential gravitational energy that leads to the concentration of discharge during a short period of time (Strahler, 1964). Alluvial plains usually have a much longer lag time between a precipitation event and the introduction of runoff water into a stream channel. When land cover on steeper slopes is affected by perturbations such as wild fire or building-related oversteepening of slopes, the likelihood of mass movement events is greatly increased. These events are usually related to unstable regolith on steep slopes, which is susceptible to failure when sufficient precipitation is received. For example, see Figure 1.

High Amounts of Precipitation

Flooding is created by the delivery of larger than normal amounts of runoff into stream channels (Smith and Ward, 1998, p. 67). Periods of above-average precipitation lead to floods. In some cases seasonal variability leads to great fluctuations in

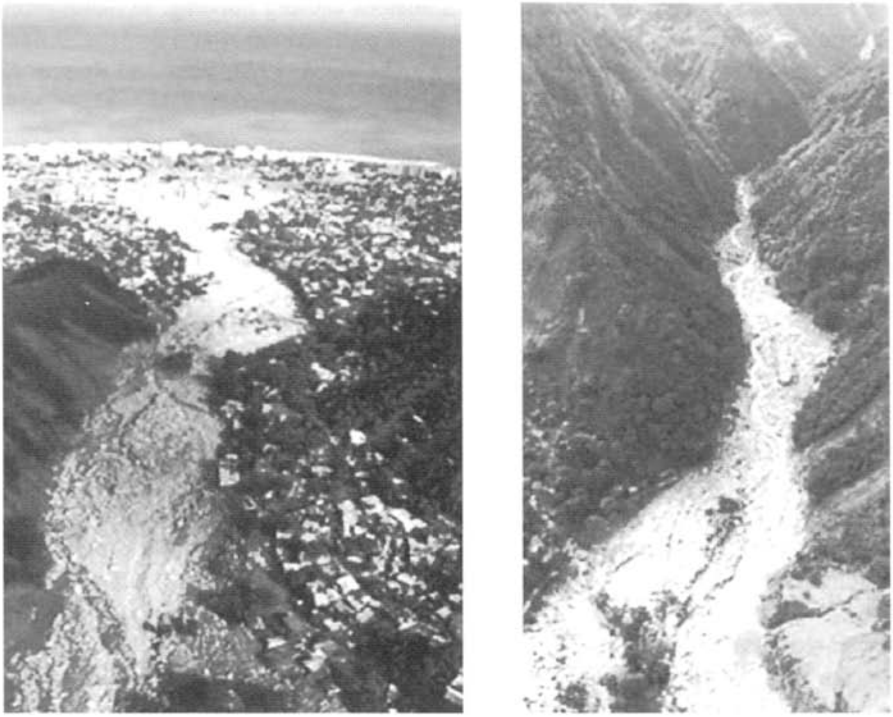


Figure 1 (see color insert) Quebrada San Julián upstream of Caraballeda showing evidence of recent debris flows and flash floods. Note the high slope angles, large numbers of debris flow scars, and abundance of new alluvium and colluvium in the channel bed and fan surface. See ftp site for color image.

stream discharge. Wet–dry subtropical or monsoonal climates with distinctive seasons of precipitation lead to fluctuations from dry stream channels to potential flooding events. These cyclical events are related to large-scale atmospheric circulation patterns that operate through an annual or longer period. In the midlatitudes, the annual migration of subtropical high pressures and the polar front lead to distinct precipitation patterns. In the tropics, monsoonal flow can lead to large precipitation events (Milne, 1986). On longer time scales El Niño and La Niña events are persistent over several years and can lead to wet or dry conditions over large areas of Earth's surface from the Equator to the midlatitudes (Waylen and Caviedes, 1987; Pearce, 1988; Ely, et al., 1994).

Extended Wet Periods

In many cases flooding is caused by the reception of precipitation over an extended time period, on the order of weeks to months, that leads to the saturation of soils in a large-scale region (Rodda, 1970b; Smith and Ward, 1998). This saturation leads to increased runoff at a time when streams are at capacity (Ward and Robinson, 1990). Additional water introduced to stream channels cannot be conveyed in the channel but is spread across the floodplain. Wet periods are related to synoptic conditions such as the position of the polar front that delivers cyclonic storms in quick succession. Poleward migration of subtropical air masses over continental areas such as the Mississippi River Basin help to supply large amounts of water to be precipitated by frontal activity. For example, see Figure 2. In some locations rainfall may fall on snow-covered or frozen ground (Thomas and Lamke, 1962). These waters are unavailable to the hydrologic cycle as long as they remain in a solid form. In the case of the former, rainfall may accelerate the introduction of water into the stream network as snowmelt augments the precipitation already being received (Kattelman, 1990; Naef and Bezzola, 1990; Caine, 1995). The latter will greatly decrease the infiltration capacity of the soil causing most of the precipitation to quickly enter the stream network (Horton, 1933).

Decaying Tropical Cyclones

Some of the largest precipitation amounts received as the result of a single meteorological event have been associated with the movement of tropical cyclones (e.g., hurricanes, cyclones, and typhoons) poleward and over continents. These powerful cyclonic storms carry large amounts of warm moist air over land surfaces. While wind speeds associated with these storms decrease quickly after landfall, these decaying storms are capable of delivering precipitation over wide areas during a relatively short period of time, on the order of days to weeks. In some cases cyclonic storms associated with the polar front may exacerbate conditions by introducing a lifting mechanism that leads to increased condensation and precipitation. The relatively low-lying coastal plain of eastern North America is especially susceptible to damage from these types of storms (Bailey and Patterson, 1975; Hirschboeck, 1988). For example, see Figure 3. In 1998 hurricane Mitch produced as much as 50 to 75

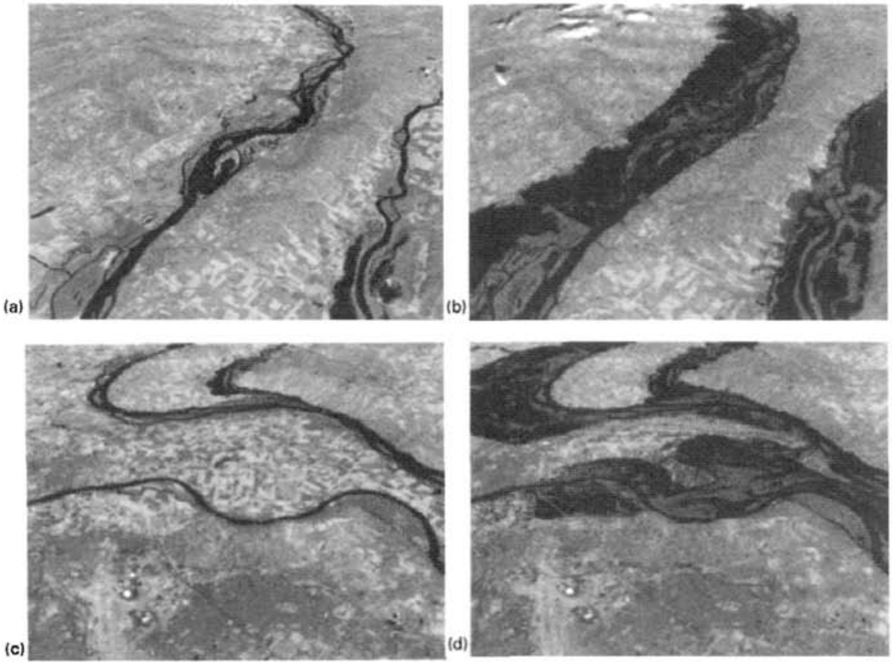


Figure 2 (see color insert) (4 panels): These scenes show various sections of the Mississippi River near St. Louis before and just after the 1993 floods, which peaked in late July/early August. The images show the area as seen by the LandSat Thematic Mapper (TM) instrument. The short-wave infrared (TM band 5), infrared (TM band 4), and visible green (TM band 2) channels are displayed in the images as red, green, and blue, respectively. In this combination, barren and/or recently cultivated land appears red to pink, vegetation appears green, water is dark blue, and artificial structures of concrete and asphalt appear dark gray or black. Reddish areas in the scenes during the flood show where water had started to recede, leaving barren land. See ftp site for color image.

inches of precipitation in some areas of Central America. At least 11,000 deaths were associated with hurricane Mitch and more than 3 million people were left homeless or were severely affected (*www.ncdc.noaa.gov/ol/reports/mitch/mitch.html*).

Intense Thunderstorms

Thunderstorms are usually intense, short-lived storms that produce high winds, hail, and heavy rainfall. These storms can be caused by convection in moist tropical air masses over continental surfaces or fast-moving cold fronts that displace those moist air masses (Hirschboeck, 1987). When these storms develop over mountainous areas where the precipitation is concentrated by the topography the potential for large, catastrophic floods is great (Hall, 1981). For example, see Figure 4. The eastern slope of the Rocky Mountains and the southwestern deserts of North America are

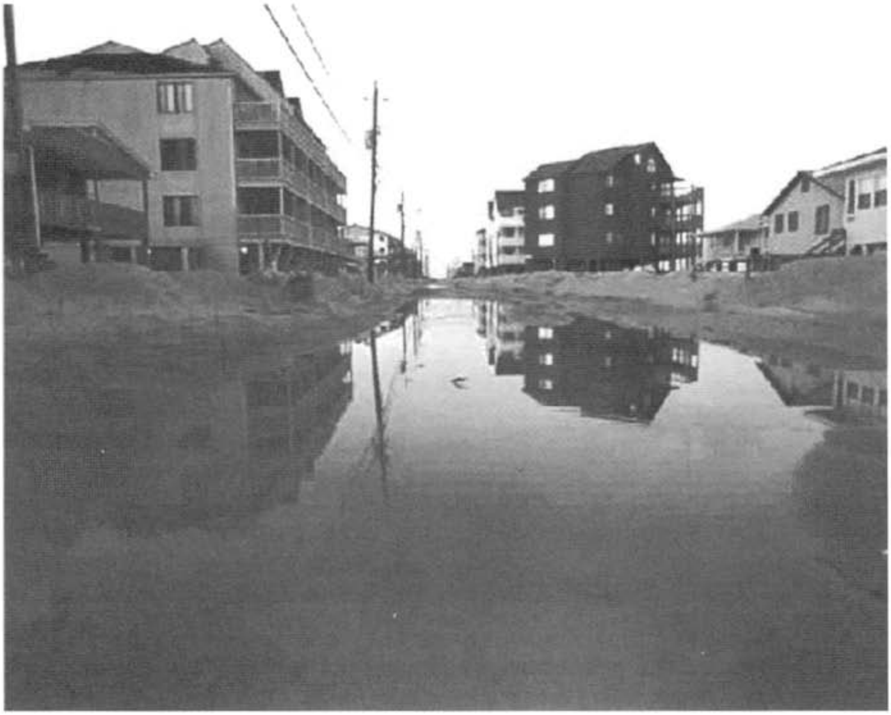


Figure 3 Water and sand washed inland to make travel difficult in North Topsail Island, North Carolina, after hurricane Fran. See ftp site for color image.

common locations for the development of thunderstorms. As moist air encounters higher elevations in these locations, it is forced to rise. Unstable atmospheric conditions are created as mountain slopes heat and in turn heat the atmosphere. Adiabatic cooling causes condensation and the development of large cumulonimbus clouds that can reach the upper altitudes of the troposphere. Sometimes there is little movement associated with a thunderstorm or thunderstorm complex, with respect to the ground; heavy precipitation concentrated in a small geographical area can have catastrophic results.

Quick Snowmelt

The storage of water in the form of snow temporarily removes that water from the hydrologic cycle. In many cases this sequestration of water is short term. Snow accumulates during winter especially at higher elevations and latitudes. With the onset of warmer spring and summer conditions, snowmelt supplies water to streams. A typical early warming may mean that snowmelt may be accelerated with large amounts of runoff entering stream channels. Mountain ranges in mid-latitude coastal regions such as the Coast Ranges and Sierra Nevada of California receive a signifi-

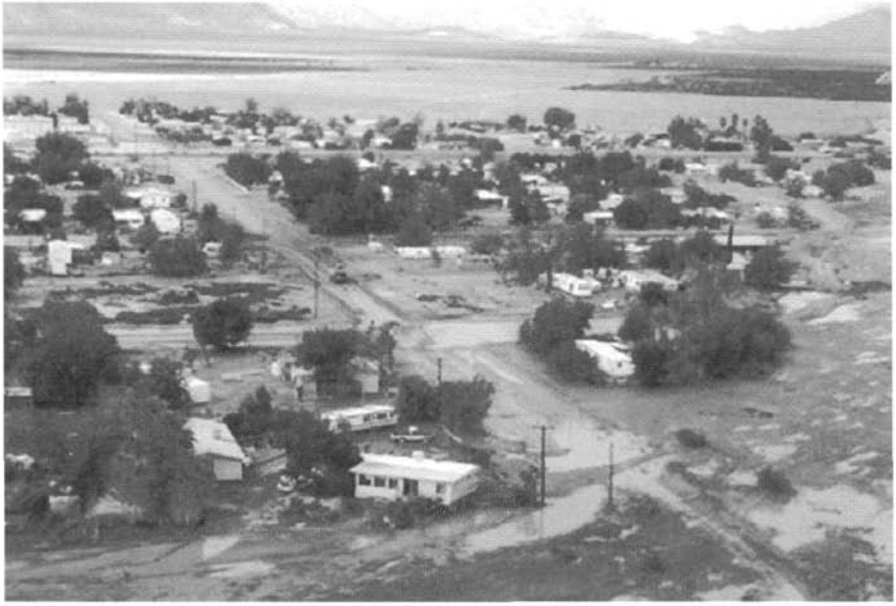


Figure 4 Arizona flash flood, Wenden, Arizona. This community was flooded twice in late October 2000 when waters from Centennial Wash swept into the town. (Photo courtesy of U.S. Small Business Administration.) See ftp site for color image.

cant portion of their annual precipitation in the form of snow. It is possible for warm early spring rains to fall on the snowpack, causing much faster runoff than normal (Bolt et al., 1975; Church, 1988). Another source of snowmelt is the subsurface introduction of heat from volcanic activity. Large volcanoes can be high enough to support permanent snow and ice cover. High temperatures associated with volcanic activity lead to the instantaneous melting of snow and ice. The melt water is commonly mixed with pyroclastic debris to form lahars (Smith, 1996).

Failure of Flood Control Structures

A variety of humanly constructed structures are used in an effort to limit the extent and severity of flooding (Gregory, 1995). Dams and levees are common flood control structures designed to contain water within designated areas (Brookes, 1985, 1988). These structures can fail because of construction errors, poor design, and overtopping by water (Biswas and Chatterjee, 1971; Costa, 1988). Flood control structures can fail because of the failure of a key component. For example, a spillway that erodes away has the potential to lead to the catastrophic failure of the entire dam as the water cuts downward. Sound structures may fail when the water retained by the structure exceeds the height of the structure. Large precipitation events or the displacement of water in a reservoir have the potential to send water flowing over the flood control structure (Kiersch, 1964). This may lead to the failure of the structure through erosion. Flooding may be exacerbated by these structures since a feature such as a levee tends to raise the stream level well above the floodplain. When a levee fails, a large amount of fluvial energy is concentrated through that break and a great deal of damage can occur near the break.

Cyclonic Storm Created Surges

In low-lying coastal locations a temporary increase in sea level associated with the approach and landfall of storms with significantly high winds and low central pressure can cause significant damage. Sea level rises in response to low pressure as it passes over ocean surfaces. Additionally, the upper portion of the water column is pushed into waves by the high winds. Storm surges can be more than 5 m above the normal high tide (Rappaport, 1994). In some areas such as bays coupled with low-lying deltas, like the Bay of Bengal and the mouth of the Ganges, where storm energy is concentrated, storm surges can reach high levels causing significant flooding (Frank and Husain, 1971; Murty and Neralla, 1992). Barrier islands are also susceptible to flooding by storm surges. Development on barrier islands along the southeastern coast of North America has led to rising property damage related to storms.

Mass Movement Events

A variety of mass movement events, while strictly not fluvial events, behave in a similar way to floods (Carson, 1976). The gravitationally fueled downhill movement of poorly consolidated regolith results from the introduction of meteoric water that adds weight and decreases hillslope cohesion. These events can do significant damage. Several types of mass movement events are composed of a larger percentage of sediments than a typical stream. Events such as mudflows, or lahars, commonly may approach the viscosity and velocity of streams. Valleys can be filled with fine-grained sediments as the deposits dewater following the initial surge of water and sediment. A variety of factors lead to mass movement events. The removal of plant cover by fire may expose soil surfaces so that infiltration rates may increase and lead to the accumulation of water along failure planes in the regolith. In areas with a subtropical wet–dry climate, such as the Mediterranean climate type, the burning of plant cover during the dry season and a subsequent wet season before the reestablishment of plant cover leads to mass wasting events (Rice et al., 1969; Campbell, 1975).

Human Responses to Flooding

There are no accurate estimates of the population in the world's floodplains. Even in the United States, only broad estimates are available, but the trends to increased vulnerability are clear. In 1955 U.S. floodplains had 10 million occupants. Thirty years later the number doubled to 20 million and by the mid-1990s about 12% of the national population lived in areas of periodic inundation. One sixth of the nation's floodplains are urbanized, and they contain more than 20,000 communities susceptible to flooding. Half of these communities have been developed since the early 1970s (Burby, 1985; Montz and Gruntfest, 1986; Alexander, 1993).

Many of the people at risk do not understand the potential consequences of the hazards they face. In the United States, flood damages exceed $2 billion annually. Only 20 to 30% of eligible structures are insured against flooding. Federal and state disaster assistance accounts for most of the difference. In the United States, almost two-thirds of the residential flood losses result from events that occur once every 1 to 10 years, even though the 100-year floodplain regulation is standard (Alexander, 1993).

In the United States, floods tend to be repetitive phenomena. From 1972 to 1979, 1900 communities were declared disaster areas by the federal government more than once, 351 were inundated at least three times, 46 at least four times and 4 at least five times. As of 1993, the United States was said to spend $9 billion a year on flood control and $300 million on flood forecasting (Alexander, 1993; Conrad, 1998).

Definitions of Structural and Nonstructural Measures

Adjustments to floods can be broadly classified into structural and nonstructural measures. Nonstructural approaches involve adjustment to human activity to accommodate the flood hazard (White, 1964; James, 1975; White, 1974) whereas structural methods are based on flood abatement or the protection of human settlement and activities against the ravages of inundation.

Structural change involves modification to the built environment to minimize or eliminate flood damage directly or flood channel construction changes. For example, see Figure 5. Structural measures are expensive. They may give the illusion of security but the record shows otherwise (Alexander, 1993). The security can be temporary. A flood can occur that is bigger than the design of the channel or levee, and changing priorities in flood control projects that require higher reservoir levels for recreation or water supply can diminish the efficacy of structural measures (Williams, 1998).

The failure of structural flood control works poses a significant threat to the lives of the people who live downstream from a massive structural project such as a dam. More than 2000 people died in 1969 in Italy when the Vaiont Dam collapsed (Blaikie et al., 1994). Because of stringent engineering standards and a system of inspections, the United States has seen few major failures. However, many structures are at the end of their design lives of 50, 75, or 100 years.

Structural flood control is still the dominant idea in many parts of the world. Following the 1927 Mississippi River floods, when river levees collapsed and 200 people died, 700,000 were displaced, and more than 135,000 buildings were damaged (Moore and Moore, 1989), the Army Corps of Engineers did not abandon its dream of controlling all floods. Rather, it proposed building large dams upstream to reduce flood peaks to the capacity of the floodway between the levees (Williams, 1998).

Until the 1970s, most flood loss reduction efforts involved structural solutions. Although nonstructural measures were discussed as alternatives, they were rarely implemented. The shift from mostly structural to mixed structural/nonstructural

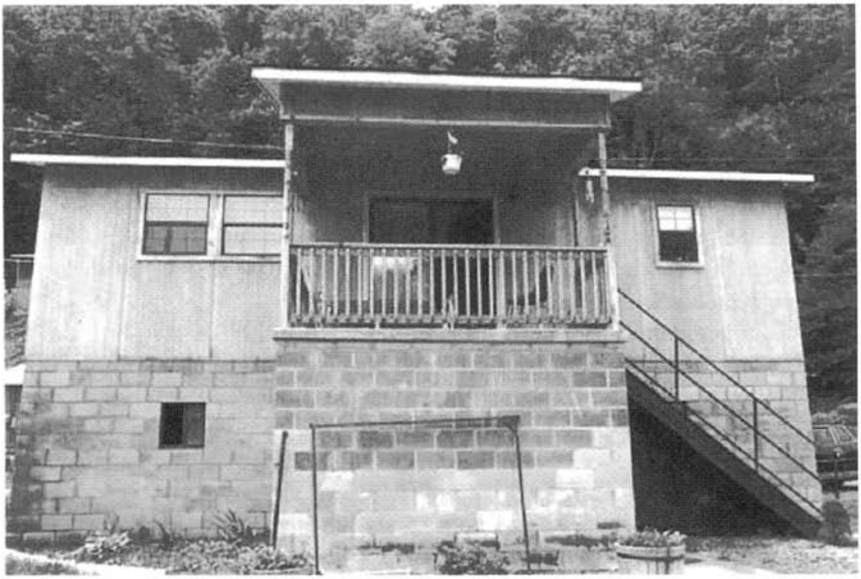


Figure 5 Elevated home in West Virginia is a mitigation success story. Risk is greatly reduced to homes elevated before a flood. See ftp site for color image.

measures began in the 1970s and continues today. The mix of adjustments varies for each situation. In Europe almost all measures that are taken have elements of combined structural and nonstructural measures. There has also been a move to be antistructural. Some dikes are being removed in favor of nonstructural or more environmentally sensitive techniques (Smith and Ward, 1998).

Nonstructural measures include floodproofing, land-use planning, soil bioengineering, warning systems, preflood mitigation efforts, and insurance. The simplest nonstructural measure is to accept the loss. Another nonstructural measure is to provide postflood relief. Protection of floodplain residents and users, and the supply of relief when they suffer damage, are forms of hidden subsidy (Alexander, 1993). This category includes aid provided by the Red Cross, voluntary organizations, and governmental agencies.

Nonstructural measures include flood insurance and land-use management, acquisition and relocation, floodproofing, preflood mitigation preparedness, outdoor warning systems, and soil bioengineering,

4 DISCUSSION OF NONSTRUCTURAL MEASURES

Flood Insurance, Floodplain Mapping, and Land-Use Ordinances

In 1968 the U.S. National Flood Insurance Program (NFIP) was launched. It made affordable insurance available to residents in flood-prone areas. In 1999 more than

18,000 communities belonged to the program. Participating local governments require developers to meet minimum standards designed to avoid damages that might be inflicted by a catastrophic 100-year flood. The program also requires property owners to purchase flood insurance to receive a federally insured mortgage (Myers, 1996). Flood insurance is a means for placing some of the burden of losses onto the people who take (or make) the risk, namely the floodplain users and residents (Alexander, 1993). Communities can participate in a Community Rating System, established by the Federal Emergency Management Agency (FEMA), that allows them to show innovative strategies to reduce flood losses in return for lower insurance premiums for floodplain residents.

Before a community can participate in the flood insurance program, the flood hazard must be recognized, assessed, and mapped. These assessments include flood history, cost and types of past flood damages, maps of the limits of the 100-year flood (or other designated flood) on a topographic map, compilations of profiles and cross sections of the river to show the levels of past floods, and compilations of flood frequency curves and locally representative hydrographs.

FEMA works with the state and community governments to identify their flood hazard areas and publishes a Flood Hazard Boundary Map of those areas. When a community joins the NFIP, it must require permits for all construction or other development in these areas and ensure that the construction materials and methods used will minimize flood damage. However, there is not careful monitoring to be sure that reducing flood hazard in a particular area does not increase flood potential elsewhere. Often, the problems are just shifted to different locales. In return the federal government makes subsidized flood insurance available to those whose structures were in the flood hazard area prior to issuance of the flood maps. All others are eligible for flood insurance at actuarial rates. FEMA issues a Flood Insurance Rate Map after the Flood Insurance Study of risk zones and elevations has been prepared (*http://floodplain.org/Jan32.htm*).

Acquisition and Relocation

The most effective measure to reduce losses is to keep the floodplains free of development. However, in many river valleys in the world, it is too late for that option. One of the most promising strategies for reducing flood losses is the public acquisition of developed land susceptible to flooding (Conrad, 1998; *www.fema.gov/mit/homsups.htm*). The authorization for U.S. federal cost sharing for relocation is more than 30 years old. However, only recently have communities, tired by chronic flooding, taken advantage of funding packages and relocated. In one case, the entire town of Valmeyer, Illinois, was relocated. The town had a long history of floods. In 1943, 1944, and 1947 unusually high levels of the Mississippi caused flooding in the nearby bottomlands affecting Valmeyer. After the 1947 floods, the U.S. Army Corps of Engineers raised the levees protecting the reach of the floodplain to 47 ft. On August 1, 1993, the flood overtopped the levees inundating Valmeyer, prompting its ultimate relocation. Since 1993 nearly 20,000 properties in 36 states and one territory have been bought out and over 25,000 families have moved from floodplains (*http://www.nwf.org/nwf/pubs/higherground/intro.html*).

Floodproofing

Floodproofing is a range of adjustments aimed at reducing flood damages to a structure or to the contents of buildings. There are three categories: (1) raising or moving the structure; (2) constructing barriers to stop floodwater from entering a building; and (3) wet flood proofing (U.S. Army Corps of Engineers, 1997).

Detection and Response Warning Systems

New technological advances in stream and rain gage networks and the increased regional floodplain management efforts have led to the adoption of thousands of local flood-warning systems. Many are simple detection systems and do not provide any mechanism for alerting the population at risk. In the United States until the 1990s warning or detection systems were planned and administered primarily at the local level.

Since then, the federal government including the Bureau of Reclamation, the U.S. Army Corps of Engineers, the National Oceanic and Atmospheric Administration, and the Federal Emergency Management Agency have actively participated in the installation and maintenance of detection and warning systems. Many systems are still managed by regional or local entities, but the percentage of federal dollars has increased substantially. Standards have also been established to help make the systems more compatible across regions (U.S. Department of Commerce, 1997).

An automated integrated network of stream and rain gages is being used in more than 1000 communities in the United States to help provide lead time for floods. Most of the systems are developed through collaborative efforts of many agencies. These ALERT systems (automated local evaluation in real time) have performed many functions other than flood warning, including helping in water supply decision making, fire weather forecasting, pollution monitoring, and providing data for river recreationists (Gruntfest and Huber, 1991). The availability of real-time data on the Internet also has increased interest in these monitoring systems (Gruntfest and Weber, 1998). The State of Arizona is developing a network for flood warning throughout the state. More than 30 agencies and communities are working together on the comprehensive ALERT system (*http://www.alertsystems.org/saas/*).

Warning systems may be nothing more than "cheap payoffs of the raingods." Too often communities install rain gage/stream gage monitoring systems without a plan for getting the warning message disseminated. A warning system is only necessary once poor land-use decisions have been made, allowing people to settle in harm's way. Many of the systems being built are not being adequately maintained to be reliable (Gruntfest and Huber, 1991; Parker and Fordham, 1996). Public education encouraging people to heed environmental cues is also being used. It is particularly difficult to provide adequate lead times for flash floods. Some communities do have drills to test the reliability and completeness of their systems to be sure the systems will operate when the conditions warrant.

As of 2001 a combination of factors increase the likelihood that automated detection systems may become more popular and more valuable. More powerful, less expensive computers, and World Wide Web access provide opportunities for inexpensive real-time weather data. While real-time stream and rain gage networks may be originally installed for flash flood forecasting, many agencies and users find the data useful for alternative purposes.

Soil Bioengineering

Anchored plantings along stream banks serve as the basis for this technique. Soil bioengineering and biotechnical engineering are cost-effective and environmentally compatible ways to protect slopes against surficial erosion and shallow mass movement. These approaches provide alternatives to structural channel "improvements." They raise questions about the notion of why engineers ever considered that concrete-lined channels should be considered "improved" (Gray and Sotir, 1996). Generally, bioengineering solutions must also include a strategy to carry floodwaters away.

The bioengineering technique is gaining support throughout the United States and Europe. It is less expensive to install and less expensive to maintain as well. The broader adoption of soil bioengineering may radically alter floodplain management.

Combined Structural and Nonstructural Measures to Reduce Flood Losses

From the first attempts to reduce flood losses in the United States, structural measures were preferred for three main reasons: (1) their benefits appeared to be relatively easy to measure, (2) they did not require extensive and politically controversial land-use planning, and, (3) the federal cost-sharing agreements encouraged communities to select the most expensive engineering projects. These reasons were supported by a faith in the technology of structural measures to protect people and property from floods.

The record now shows that in spite of massive expenditures, flood losses have continued to rise. Since the 1960s, especially in the United States, there has been a call for a shift from primarily structural measures to control floods to nonstructural measures (Galloway, 1994; Larson, 1996; Williams, 1998). Land-use control is one of the most effective ways of reducing flood hazards. Statutes, ordinances, regulations, and compulsory purchases can be employed and relocation can be subsidized. A floodway left undeveloped through the city can become beautiful public open space.

5 CONCLUSION

Floods are generally caused by the combination of large amounts of precipitation and basin topography. For example, saturation of soils caused by large amounts of precipitation can lead to flooding. Urbanization of a drainage basin increases the amount of runoff reaching a channel and decreases the lag time between a precipitation event and peak flow. A variety of weather events lead to flooding, including extended wet periods, decaying tropical cyclones, intense thunderstorms, and quick snowmelt. In some cases humanly constructed structures designed to prevent flooding collapse, causing flooding or accentuating flooding. In low-lying coastal areas storm surges may cause significant flooding. Mass movement events are similar to flooding, although the proportion of sediments to water is larger than an alluvial flood with the outcome just as disastrous.

Humans respond to flooding in a variety of ways. Broadly defined these fall into two categories, structural and nonstructural measures. Structural measures include dams and dikes. Through time the efficacy of structural features has been questioned, and there has been a shift from purely structural approaches to controlling floods to a mix of structural and nonstructural flood mitigation strategies. Nonstructural measures include flood insurance, floodplain mapping, and land-use ordinances, acquisition and relocation, floodproofing, detection and response warning systems, soil bioengineering, and combined structural and nonstructural measures to reduce flood losses. Some progress is being made in addressing the hazards associated with flooding. The reduction of flood impacts continues at great expense, but vulnerability will continue to rise as long as more people build in floodplains, increasing the risk of catastrophic floods. Even the best warnings will not eliminate the risks increasingly being taken around the globe.

References

Alexander, D., *Natural Disasters*, Chapman and Hall, New York, 1993, 632 p.

Bailey, J. F. and J. L. Patterson, *Hurricane Agnes Rainfall and Floods, June–July 1972*, USGS Professional Paper 924.

Biswas, A. K. and S. Chatterjee, Dam disasters: an assessment, *Engineering Journal*, *54*, 3–8, 1971.

Blaikie P., T. Cannon, I. Davis, and B. Wisner, *At Risk Natural Hazards, Peoples Vulnerability and Disasters*, Routledge, New York, 1994, 282 p.

Bolt, B. A., W. L. Horn, G. A. Macdonald, and R. F. Scott, *Geological Hazards*, Verlag, Berlin, 1975.

Brookes, A., River channelization: traditional engineering methods, physical consequences and alternative practices, *Progress in Physical Geography*, *9*, 44–73, 1985.

Brookes, A., *Channelized Rivers: Perspectives for Environmental Management*, John Wiley & Sons, Chichester, 1988.

Brown, L. R., Conserving soils, in L. R. Brown (Ed.), *State of the World*, Norton, New York, 1984, pp. 53–75.

Burby, R. J., *Flood Plain Land Use Management: A National Assessment*, Westview Press, Boulder Colorado, 1985.

Caine, N., Snowpack influences on geomorphic processes in Green Lakes valley, Colorado Front Range, *Geographical Journal*, *161*, 55–68, 1995.

Campbell, R. H., Soil slips, debris flows and rainstorms in the Santa Monica Mountains and vicinity, southern California, *US Geological Survey, Professional Paper 851*, 1975.

Carson, M. A., Mass-wasting, slope development and climate, in E. Derbyshire (Ed.), *Geomorphology and Climate*, John Wiley & Sons, Chichester, 1976, pp. 101–136.

Church, M., Floods in cold climates, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 205–229.

Clark, E. H., J. A. Haverkamp, and W. Chapman, *Eroding Soils: The Off-Farm Impacts*, The Conservation Foundation, Washington, DC, 1985.

Conrad, D., *Higher Ground: Voluntary Property Buyouts in the Nation's Floodplains, A Common Ground Solution Serving People at Risk, Taxpayers and the Environment*, National Wildlife Federation, Washington, DC, 1998.

Costa, John E., Floods from Dam Failures, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 439–463.

Ely, L. L., Y. Enzel, and D. R. Cayan, Anomalous North Pacific atmospheric circulation and large winter floods in the southwestern United States, *Journal of Climate*, *7*, 977–987, 1994.

Frank, N. L. and S. A. Husain, The deadliest tropical cyclone in history? *Bulletin of American Meteorological Society*, *52*(6), 438–444, 1971.

Galloway, G., *A Blueprint for Change, Sharing the Challenge: Floodplain Management into the 21st Century*, report of the Interagency Floodplain Management Review Committee to the Administration Floodplain Management Task Force, Washington, DC, June 1994.

Gray, D. H. and R. B. Sotir, *Biotechnical and Soil Bioengineering Slope Stabilization: A Practical Guide for Erosion Control*, Wiley, New York, 1996, 378 p.

Gregory, K. J., Human activity and paleohydrology, in K. H. Gregory, L. Starkel, and V. R. Baker (Eds.), *Global Continental Palaeohydrology*, Wiley-Interscience, Chichester, 1995, pp. 151–172.

Gruntfest, E. and C. J. Huber, Toward a comprehensive national assessment of flash flooding in the United States, *Episodes*, *14*(1), 26–34, 1991.

Gruntfest, E. and M. Weber, Internet and emergency management prospects for the future, *International Journal of Mass Emergencies and Disasters*, *16*(1), 1998.

Hall, A. J., *Flash Flood Forecasting*, Operational Hydrology Report No. 18, WMO, Geneva, 1981.

Hammer, T. R., Stream channel enlargement due to urbanization, *Water Resources Research*, *8*, 1530–1540, 1972.

Hassan, M. A., Observations of desert flood bores, *Earth Surface Processes and Landforms*, *15*, 481–485, 1990.

Hirschboeck, K. K., Catastrophic flooding and atmospheric circulation anomalies, in L. Mayer and D. Nash (Eds.), *Catastrophic Flooding*, Unwin Hyman, London, 1987, pp. 23–56.

Hirschboeck, K. K., Flood hydroclimatology, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 27–49.

Horton, R. E., The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union*, *14*, 446–460, 1933.

Kattelman, R., Floods in the high Sierra Nevada, California, USA, in R. O. Sinniger and M. Monbaron (Eds.), *Hydrology in Mountainous Regions II. Artificial Reservoirs, Water and Slopes*, IAHS Publ. No. 205, 1990, pp. 311–317.

Kiersch, G. A., The Vaiont River disaster, *Civil Engineering*, *34*, 32–39, 1964.

Knighton, David, *Fluvial Forms and Processes: A New Perspective*, Arnold, London, 1998.

Kuhnle, R. A., R. L. Binger, G. R. Foster, and E. H. Grissinger, Effect of land use changes on sediment transport in Goodwin Creek, *Water Resources Research, 32*, 3189–3196, 1996.

Larson, L., Lessons drawn from the 1993 flood, *Forum for Applied Research and Public Policy*, Fall, No. 3, 1996, pp. 102–104, 1996.

Milne, A., *Flood Shock*, Sutton, Gloucester, 1986.

Montz, B. E. and E. Gruntfest, Changes in American urban floodplain occupancy since 1958: the experiences of nine cities, *Applied Geography, 6*, 325–338, 1986.

Moore, J. W. and D. P. Moore, *The Army Corps of Engineers and the Evolution of Federal Floodplain Management Policy*, Institute of Behavioral Science, Boulder, CO, special publication No. 30, 1989.

Murty, T. S. and V. R. Neralla, On recurvature of tropical cyclones and the storm surge problem in Bangladesh, *Natural Hazards*, *6*(3), 275–279, 1992.

Myers, M. F., Midwest floods channel reforms, *Forum for Applied Research and Public Policy, 11*, Fall, No. 3, 1996, pp. 88–97.

Naef, F. and G. R. Bezzola, Hydrology and morphological consequences of the 1987 flood event in the upper Reuss valley, in R. O. Sinniger and M. Monbaron (Eds.), *Hydrology in Mountainous Regions II. Artificial Reservoirs, Water and Slopes*, IAHS Publ. No. 205, 1990, pp. 339–346.

Newson, M., *Land, Water and Development: River Basin Systems and their Sustainable Management*, Routledge, London, 1992.

Parker, D. and M. Fordham, Evaluation of flood forecasting, warning and response systems in the European Union, *Water Resources Management*, *10*(4), 1996, pp. 279–302.

Patton, Peter C., Drainage basin morphometry and floods, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 51–64.

Pearce, F., Cool oceans caused floods in Bangladesh and Sudan, *New Scientist*, *8*, 31, 1988.

Rappaport, E. N., Hurricane Andrew, *Weather*, *49*(2), 51–60, 1994.

Reid, I. and L. E. Frostick, Flow dynamics and suspended sediment properties in arid zone flash floods, *Hydrologic Processes*, *1*, 239–253, 1987.

Rice, R. M., E. S. Corbett, and R. G. Bailey, Soil slips related vegetation, topography and soil in southern California, *Water Resources Research*, *5*, 647–659, 1969.

Roberts, C. R., Flood frequency and urban-induced channel change: Some British examples, in K. Beven and P. Carling (Eds.), *Floods: Hydrological, Sedimentological and Geomorphological Implications*, John Wiley & Sons, Chichester, 1989, pp. 57–82.

Rodda, J. C., Rainfall excesses in the United Kingdom, *Transactions of the Institute of British Geography, 49*, 49–60, 1970.

Smith, K. and R. C. Ward, *Floods: Physical Processes and Human Impacts*, John Wiley, New York, 1998.

Smith, K., *Environmental Hazards: Assessing Risk and Reducing Disaster*, 2nd ed., Routledge, London, 1996.

Strahler, Alan N., Quantitative geomorphology of drainage basins and channel networks, in V. T. Chow (Ed.), *Handbook of Applied Hydrology*, McGraw-Hill, New York, 1964, pp. 4-40–4-74.

Thomas, C. A. and R. D. Lamke, Floods of February 1962 in southern Idaho and northeastern Nevada, *US Geological Survey Circular*, *467*, 30, 1962.

U.S. Army Corps of Engineers, *Flood Proofing Techniques, Programs and References*, Available from Corps at CECW-PF, 20 Massachusetts Avenue, NW, Washington, DC, 1997, 26 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrology, *Automated Local Flood Warning Systems Handbook*, Weather Service Hydrology Handbook No. 2, February, Silver Spring, MD, 1997, unpaged.

Ward, R. C. and M. Robinson, *Principles of Hydrology*, 3rd ed, McGraw-Hill, Maidenhead, 1990.

Waylen, P. R. and C. N. Caviedes, El Nino and annual floods in coastal Peru, in L. Mayer and D. Nash (Eds.), *Catastrophic Flooding*, Unwin Hyman, 1987, pp. 57–77.

Williams, P., Inviting trouble downstream, *Civil Engineering*, February, 1998, pp. 50–53.

Wohl, Ellen (Ed.), *Inland Floods*, Cambridge University Press, 2000.

Wolman, M. G., A cycle of sedimentation and erosion in urban river channels, *Geografiska Annaler, 49A*, 385–395, 1967.

Wolman M. G. and R. Gerson, Relative scales of time and effectiveness of climate in watershed geomorphology, *Earth Surf. Proc.*, *3*, 189–208, 1978.

Zaslavsky, D. and G. Sinai, Surface hydrology: I-Explanation of phenomena; II-Distribution of raindrops; III-Causes of lateral flow; IV-Flow in sloping layered soil; V-In-surface transient flow, *Journal of Hydraulics Division American Society of Civil Engineers*, *107(HY1)*, 1–93, 1981.

Selected Web Pages Related to Nonstructural Measures

http://www.alertsystems.org/saas/ALERT User Group

http://FEMA.gov U.S. Federal Emergency Management Agency

http://ceres.ca.ca.gov/ – State of California water resources agency

http://www.usace.army.mil/inet/functions/cw/cwfpms/fpms.htm U.S. Army Corps of Engineers with emphasis on floodplain management activities

http://www.nwf.org/nwf/pubs/higherground/intro.html National Wildlife Federation site for manuscript Higher Ground

http://member.aol.com/damsafety/homepage.htm Association of State Dam Safety Officials

http://www.ci.fort-collins.co.us/csafety/oem/index.htm Comprehensive emergency preparedness homepage from Ft Collins, Colorado an excellent reference.

http://web.uccs.edu/geogenvs/work/Eve/Beyond%20Flood%20Detection%20Final.html.

www.ncdc.noaa.gov/ol/reports/mitch/mitch.html – NOAA Website about Hurricane Mitch

SECTION 3

SOCIETAL IMPACTS

CHAPTER 37

CLIMATE AND SOCIETY

MICHAEL H. GLANTZ

At the turn of the twentieth century, scholars who wrote about the interplay between climate and society did so based on their perceptions of climate as a boundary constraint for the development prospects of a society. Perceptions of climate were used as an excuse to dominate societies in Africa, Asia, and Latin America. As a result, climate–society studies soon became viewed as a colonial ploy to control populations in developing areas in the tropics. Perhaps the most cited book in this regard was written by Ellsworth Huntington, *Climate and Civilization*, published in 1915. In his view, inhabitants of the tropics were destined to accept lower levels of economic and social development because their climate setting was not conducive to lively (i.e., productive) human activity or an aggressive work ethic. According to Huntington, tropical climate was the main culprit causing people in the tropics to be less productive than people in temperate regions. Huntington argued that the temperate climate has an energizing effect on humans. With the growing belief that such an argument was racist in intent, Huntington's work was challenged, and discussion of the various ways in which climate might influence human behavior was stifled for decades, notwithstanding a few notable exceptions. One such exception is entitled *Climate and the Energy of Nations* (Markham, 1944) in which Markham referred to the "air-conditioning revolution," a revolution based on the development and spread of a new technology into the tropical areas. Markham asserted that technology brings islands of temperate-zone climate into the tropics, thereby generating a more aggressive work ethic.

Following the end of World War II and the onset of the Cold War between Soviet-style communism and Western capitalism and democracy, attention of governments turned to Cold War conflicts, avoidance of nuclear war, searches for allies, and decolonization. The major Cold War nations were in a competition to show that *their* approach to economic development was the only way for the newly indepen-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

dent countries to follow. A main stated objective was their intent to assist these countries to become food secure based on the nation's resources. Consideration of climate was making its way back into the discussions of economic development in developing countries. Once again interest was raised with regard to climatic constraints to economic development in tropical countries.

In the 1950s and 1960s, attention focused on decolonization and political development of the newly independent states (e.g., Pye, 1966). In the mid-1970s, a World Bank report about the economic prospects for developing countries—*The Tropics and Economic Development: A Provocative Inquiry into the Poverty of Nations*—hinted at the economic, social, and political problems caused by climate variability from one year to the next. Its author (Kamarck, 1976) noted that recurrent droughts in northeast Brazil are a chronic constraint on the region's economic development prospects. His reference to interannual climate variability was brief and unelaborated. However, climate as a boundary constraint was starting to give way to climate as something that societies might be able to forecast and cope with, at least in its extremes.

In the 1970s, attention focused on how the vagaries of weather exposed hundreds of millions of people to hunger and, depending on the socioeconomic situation in a particular country, to famine as well (e.g., Glantz, 1976; Sen, 1981). Thus, there was a growing number of examples of the notion that climate was not really a boundary constraint to the level of development that a people or culture could expect to attain. This notion began to give way to the belief that variability in climate, from one year to the next or one decade to the next, could be coped with so as to soften the impacts of climate variability and weather extremes on agriculture and livestock and, more generally, on the productivity of the land's surface (e.g., Glantz, 1977; Hare, 1977).

Recall that the 1970s was a disruptive decade with respect to climate: 5 years of drought in the West African Sahel (Glantz, 1976); failure of the Soviet harvest and subsequent large-scale, low-cost grain purchases by the Soviet Union in the early 1970s (Trager, 1975); the global food crisis (Brown and Eckholm, 1974); talk of a possible return to an ice age (e.g., Ponte, 1976; Weather Conspiracy, 1974); the Ethiopian famine (Wolde Mariam, 1988); drought-related coups in sub-Saharan Africa; drought in the wheat-producing Canadian prairie provinces (Glantz, 1977); the first drop in global fish catches since the end of World War II (Brown and Eckholm, 1974), and so forth.

A devastating 5-year drought from 1968 to 1973 in the West African Sahel and its associated death and environmental destruction in the region drew attention to the impacts on household and village responses to prolonged, multiyear droughts. Widespread droughts around the globe in 1972–1973, famines in West Africa and Ethiopia, blamed for the most part on an El Niño event, along with the drop in fish landings, prompted the U.N. Secretary General to convene a series of UN-sponsored world conferences on food (1974), population (1974), human settlements (1976), water (1977), desertification (1977), climate (1979) and technology (1979).

Thus, toward the middle of the 1970s, at least five new major climate-related scientific issues emerged: the effect of chlorofluorocarbons (CFCs) on the ozone layer in the stratosphere, talk of an impending Ice Age suddenly shifted to talk of a human-induced global warming, acid rain, desertification, and El Niño. Each of these issues raised interest in climate–society interactions to higher levels among researchers in different disciplines, government agencies, economic sectors, the media, and the public. Societies around the globe responded (and continue to respond) in different ways to each of these climate-related issues. For example, desertification is an environmental issue that is of great concern to African countries.

North Americans, however, refused to accept the view that desertification could occur in the U.S. West as a result of mismanagement of the land's surface, while noting that desertification was the plight of poor developing countries in Africa. The term *desertification* first appeared in a report on the destruction of dry forests in central Africa by a French forester (Aubreville, 1949). Since then, the concept of desertification has been expanded to include such land degradation processes as soil erosion, wind deflation, soil salinization, water logging, livestock overgrazing, and soil trampling. While many of these processes were exposed during the prolonged drought in the West African Sahel and then labeled as desertification, it is not difficult to show that similar processes also take place in the U.S. West.

The acid rain issue was addressed in the United States with the implementation by the U.S. Congress of a decade-long national assessment called NAPAP (National Acid Precipitation Assessment Program). Stratospheric ozone depletion was addressed globally in the 1980s with the development of international legal instruments culminating in the Montreal Protocol of 1987 and, later, amendments to it (Benedick, 1998).

It was in the early 1970s, 1972–1973 to be exact, that an El Niño event (defined briefly as an invasion of warm water from the Western Pacific into the central and eastern equatorial Pacific Ocean) attracted global attention. An event in 1982–1983, the biggest in a century until that time, captured the full attention of the scientific community and various governments as a natural phenomenon that spawned hazards around the globe. Such hazards included, but were not limited to, droughts, floods, frosts, fires and food shortages, famine, and disease. Ever since the mid-1970s, research funding of El Niño–related research has been growing along with international interest in the phenomenon and its societal and environmental impacts. The extraordinary El Niño event of 1997–1998 helped to make El Niño and its cold counterpart, La Niña, household words throughout much of the world. Only at the end of the twentieth century did La Niña events become of serious interest to the El Niño research and forecasting communities (Glantz, 2002). This belated interest is even more surprising given the scientific observation that tropical storms and hurricanes in the Atlantic Basin and in the Gulf of Mexico tend to increase in number during La Niña events and drop in number during El Niño events.

Global warming is an environmental issue that arose out of discussions and governmental and scientific concerns about the possibility of a global cooling. It was first suggested in 1896 by Swedish chemist Arrhenius (1896, 1908) that the burning of coal by human societies would add enough extra carbon dioxide into the air to eventually heat up Earth's atmosphere by a few degrees Celsius. This issue was revisited in the 1930s by Callendar (1938), who thought that a human-induced global warming of the atmosphere could stave off the imminent recurrence of an ice age. The issue was again revisited in the 1950s when global warming was looked at in neutral terms, as an experiment that societies were performing on the chemistry of the atmosphere, for which the outcome is unknown (Revelle and Suess, 1957).

It was not until the mid-1970s that human-induced global warming began to be viewed as an adverse event for future generations of human societies and ecosystems that might not be able to adapt to the rate of warming expected to occur. The cause of the warming was attributed to the increasing amounts of greenhouse gases (CO_2, CFCs, CH_4, NO_x, collectively referred to as GHGs) being emitted into the atmosphere as a result of human activities. Carbon dioxide is a product of the burning of fossil fuels, and its amount in the atmosphere has been rising since the onset of the Industrial Revolution in the late 1700s. Tropical deforestation also contributes carbon dioxide to the atmosphere. Tropical forests have served as sinks for carbon dioxide, pulling it out of the air and storing it. When trees are felled, decompose or burned, the stored carbon is emitted into the air.

Chlorofluorocarbons (CFCs), a greenhouse gas as well as a stratospheric "ozone eater," are man-made chemicals first discovered in the 1920s for use as a refrigerant. Methane resulting from livestock rearing (e.g., cattle, pigs) and from the increasing number of landfills is another greenhouse gas. Nitrous oxides are used by farmers in fertilizers and have been widely applied to agricultural lands around the globe in increasing amounts since the end of World War II. Of these major greenhouse gases, carbon dioxide is seen at the main culprit in the global warming debate.

Current scientific research suggests that the level of climate change that might be expected (at current rates of greenhouse gas emissions) is on the order of 1.5 to 4.5°C by the end of the twenty-first century (IPCC, 1990, 1996, 2001). Concerned with the prospects of a changing global climate, many nations have come together to call for a technical assessment of the state of the science through the Intergovernmental Panel on Climate Change (IPCC).

The degree of warming, however, is dependent on numerous factors: the rate at which GHGs continue to be emitted into the atmosphere, the shift by societies to alternative energy sources, the rate of tropical deforestation, the residence time of GHGs in the atmosphere (several of these gases will remain in the atmosphere for decades to centuries), the development of methods to sequester carbon (i.e., taking it from the atmosphere and binding it in some way in Earth's land surface, vegetation, or oceans), and so forth. Some degree of global warming is inevitable, given the residence time of the GHGs already emitted into the atmosphere. This means that societies around the globe, from local to national, must attempt to ascertain how a warmer global climate regime might affect regional and local climates. Will there be more extreme climate events (such as droughts, floods, frosts, fires) or fewer? These societies must also seriously consider nationally, as well as collectively in cooperation with other countries, the most effective way(s) to
cope with the potentially adverse impacts of some degree of human-induced global warming.

Coping mechanisms for climate change likely to occur decades in the future can be divided into three categories: preventive, mitigative and adaptive measures. *Preventive* measures are designed to prevent the increased buildup of GHGs in the atmosphere. *Mitigative* measures depend on an improved understanding of how global warming might affect local climates worldwide and are designed to improve societies' ability to respond to changes, some of which can be anticipated with some degree of reliability. *Adaptive* measures are used to refer to society pursuing a "business as usual" strategy, not seeking to control GHG emissions, allowing global warming to occur, and responding to any of the impacts of climate change as they might appear. Today, adaptation now encompasses mitigation as well (Smith et al., 1996).

As we enter a new millennium, it appears that national governments have shifted their concern from global climate change to local and regional climate impacts, and from climate change alone to climate change AND climate variability on various time scales, from the seasons to years to decades. Reinforcing this interest has been the fact that climate-related disasters in the 1990s have been the most costly since records have been kept.

More specifically, the climate-related events of the 1990s merit special attention. Devastating hurricanes, such as Hurricane Andrew (1992), Hurricanes Mitch and Georges (1998), Hurricane Floyd (1999), floods in Europe (1993 and 1995), floods in Bangladesh (1998), devastating floods in China (1998), flash flooding and mud slides in Venezuela (1999), droughts in southern Africa (1991–1992), a prolonged El Niño event (1991–1995), the destructive ice storm in northeastern North America (1998), and a second El Niño of the century (1997–1998), drought and famine in Ethiopia (2000), among many other climate-related problems, have heightened public, media, and policymaking interest in climate to levels never before seen. While several of these events might have been expected to occur, they were for the most part surprising in their timing or intensity.

As a result of this new-found international concern, there has been an increasing number of studies on climate-related impacts and on how societies have been affected by (or coped with) those impacts. There has also been an apparent realization that an improved understanding of climate variability and climate extremes can help societies to better cope with climate change several decades in the future.

Recent hurricanes, ice storms, droughts, floods, and intense El Niño events have proven that all countries, regardless of their level of economic development or type of political system, are subject to the adverse effects of climate variability. Despite programs designed to "droughtproof" or "floodproof" a country, recent studies have shown that industrialized countries are no more immune to the impacts of climate from one year to the next than are developing countries. A major difference, however, is that in the industrialized countries, governments are in a relatively better position (economically) than developing ones to address those impacts and their long-term implications for economic development prospects.

Climate modelers and other climate researchers have come to realize that there are likely to be more surprises with regard to the behavior of the global climate system. While there are aspects of the system that are somewhat predictable, other aspects are not. A climate-related surprise can be defined as a gap between one's expectations about climate's behavior and what the climate system actually does.

Climate-related surprise is not a black-and-white condition. People are hardly ever either totally surprised or never surprised. There are shades of surprise with regard to human responses to the same climate-related event. They can be hardly surprised, mildly surprised, somewhat surprised, very surprised, extremely surprised or totally surprised (NB: each of these examples was taken from the scientific literature). Myers (1995, p. 358) introduced the interesting notion of "semisurprised." Thus, surprise may best be described in "fuzzy" terms with the degree of surprise dependent on several intervening variables such as personal experience, core beliefs, expectations, or knowledge about a phenomenon or about a geographic location.

One could argue that there are knowable as well as unknowable surprises (Streets and Glantz, 2000). Knowable refers to the fact that some climate surprises can be anticipated (Myers, 1995). For example, certain parts of the globe are drought prone. It is known that drought will likely recur. What is not known is exactly when it will take place, how long it will last, how intense it will be, or where its most devastating impacts are likely to occur. El Niño is in this category. While we have now come to expect these events to recur, we do not know when that will happen or what it will be like. The uncertainty then cascades down the "impacts chain," and as we speculate about likely impacts of an El Niño, the degree of uncertainty will increase.

Take, for example, the 1997–1998 El Niño. Even with the best monitoring and observing system in the world focused on minute changes in various aspects of the tropical Pacific Ocean, forecasters and modelers were unable to predict the onset of one of the biggest El Niño events in the past 100 years. Nor were they able to predict the course of development of that event. They were better than in earlier times, however, at predicting some of its impacts on societies in certain parts of the globe, especially those where the influences of changes in the sea surface temperatures in the tropical Pacific are known to be strong.

Societies (and their scientists) are on a learning curve with regard to the various ways that climate variability and climate change might affect climate-related human activities. They must avoid becoming complacent as a result of a belief that they fully understand atmospheric processes or their impacts. They must accept that there will be climate surprises in the future, even if the global climate does not change. They must learn from past experiences on how best to cope with the vagaries of climate (Glantz, 1988).

Many countries now realize that climate-related problems do not stop at international boundaries. There are many transboundary issues that demand regional (if not international) cooperation, given that countries share river basins, inland seas, airsheds, the global atmosphere as well as the onslaught and impacts of extreme meteorological events such as droughts, floods, and tropical storms.

While climate-related anomalies cannot be prevented, societal preparation for, and response to, their adverse impacts can be improved through better knowledge of the direct and indirect ways in which atmospheric processes interact with human activities and ecological processes. The enhancement of such knowledge will lead to better forecasts as well as better computer modeling of the interactions among land, sea, and air. A society forewarned of climate-related hazards is forearmed to cope with those hazards more effectively.

REFERENCES

Arrhenius, S., *Worlds in the Making*, Harper & Brothers, New York, 1908.

Arrhenius, S., On the influence of carbonic acid in the air upon the temperature of the ground, *Philos. Mag.*, 41, 237276, 1896.

Aubreville, A., *Climate, Forests and Desertification in Tropical Africa*, Sociéte d'Editions Géographiques, Maritimes et Coloniales, 1949.

Benedick, R.E., *Ozone Diplomacy: New Directions in Safeguarding the Planet*, Harvard University Press, Cambridge, Massachusetts, Enlarged Edition, 1998.

Brown, L., and E.P. Eckholm, *By Bread Alone*, Praeger Press, New York, 1974.

Callendar, G.S., The artificial production of carbon dioxide and its influence on temperatures, *Q. J. Roy. Meteor. Soc.*, 64, 223237, 1938.

Glantz, M.H. (ed.), *La Niña and Its Impacts: Facts and Speculation*. United Nations University Press, Tokyo, Japan, 2002.

Glantz, M.H. (ed.), *Societal Responses to Regional Climatic Change: Forecasting by Analogy*, Westview Special Study, Boulder, Colorado, 1988.

Glantz, M.H. (ed.), *Desertification: Environmental Degradation in and around Arid Lands*, Westview Press, Boulder, Colorado, 1977.

Glantz, M.H. (ed.), *The Politics of Natural Disaster: The Case of the Sahel Drought*, Praeger Press, New York, 1976.

Hare, K., Connections between climate and desertification, *Environ. Conserv., 4* (2), 82, 1977.

Huntington, E., *Civilization and Climate*, Yale University Press, New Haven, Connecticut, 1915, rev, ed. 1924, reprinted by University Press of the Pacific, 2001.

IPCC (Intergovernmental Panel on Climate Change), *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, Contribution of Working Group II to Third Assessment Report, Cambridge University Press, Cambridge, UK, 2001.

IPCC, *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to Second Assessment Report, Cambridge University Press, Cambridge, UK, 1996.

IPCC, *Climate Change: The IPCC Scientific Assessment*, Cambridge University Press, Cambridge, UK, 1990.

Kamarck, A.M., *The Tropics and Economic Development: A Provocative Inquiry into the Poverty of Nations*, The Johns Hopkins University Press, Baltimore, Maryland, 1976.

Markham, S.F., *Climate and the Energy of Nations*, Oxford University Press, London, 1944.

Myers, N., Environmental unknowns, *Science*, 269, 358360, 1995.

Ponte, L., *The Cooling*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

Pye, L., *Aspects of Political Development*, Little, Brown and Co., Boston, Massachusetts, 1966.

Revelle, R.R., and H.E. Suess, Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO_2 during the past decades, *Tellus*, 9, 1827, 1957.

Sen, A., *Poverty and Famines: An Essay on Entitlement and Deprivation*, Oxford University Press, Oxford, UK, 1981.

Smith, J.B., N. Bhatti, G. Menzhulin, R. Benioff, M.I. Bodyko, M. Campos, B. Jallow, and F. Rijsberman (eds.), *Adapting to Climate Change: An International Perspective*, Springer-Verlag, New York, 1996.

Streets, D.G., and M.H. Glantz, Exploring the concept of climate surprise, *Global Environ. Chang.*, 10, 97107, 2000.

Weather Conspiracy: The Coming of the New Ice Age, Ballentine Books, New York, compiled by The Impact Team, 1977.

Wolde Mariam, M., *Rural Vulnerability to Famine in Ethiopia, 195777*, Vikas Publishing House, New Delhi, 1984.

CHAPTER 38

HOUSEHOLD FOOD SECURITY AND COPING WITH CLIMATIC VARIABILITY IN DEVELOPING COUNTRIES

THOMAS E. DOWNING AND YOLANDE STOWELL

1 INTRODUCTION

Climate affects food security in two distinct ways. Primarily, climate in association with soils, terrain, and vegetation is a resource that influences potential agricultural production. Agricultural productivity, in turn, provides income to individuals and households, leads to investment in infrastructure, and fuels the regional economy.

Climate also includes the hazards of drought, flood, windstorms, hail, and temperature extremes. Such climatic hazards lead to direct losses of income, infrastructure, and even lives. Indirect effects of hazards include aversion to investment because of high risk, a lack of infrastructure, and stagnation of the regional economy. Drought is the leading climatic hazard for vulnerable populations in most developing countries, although flood risk is critical in South Asia and China, and in recent years to some parts of Central America.

Conceptions of food security have evolved over the past few decades, and it has increasingly been recognized that "much more was involved in food security than just climate" (Glantz, 1997). Since the United Nations (UN) World Food Conference in 1974, there have been three major shifts in food security thinking: changing the scale from global to household and individual, broadening the focus from food alone to long-term resilience of livelihoods, and diversifying from objective indicators such as target levels of consumption to more subjective perceptions of security (Maxwell, 1994). (See Box 1 for a summary of food security definitions.)

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

Box 1 Definitions of Food Security

Food security is defined in its most basic form as access by all people at all times to the food needed for a healthy life. Achieving food security has three dimensions; first, it is necessary to ensure a safe and nutritionally adequate food supply both at the national level and at the household level. Second, it is necessary to have a reasonable degree of stability in the supply of food both from one year to the other and during the year. Third, and most critical, is the need to ensure that each household has physical, social, and economic access to enough food to meet its needs. This means that each household must have the knowledge and ability to produce or procure the food that it needs on a sustainable basis. In this context, properly balanced diets that supply all necessary nutrients and energy without leading to overconsumption or waste should be encouraged. It is also important to encourage the proper distribution of food within the household, among its members.

The right to an adequate standard of living, including food, is recognized in the *Universal Declaration of Human Rights*. Food security should be a fundamental objective of development policy as well as a measure of its success. Household food insecurity affects a wide cross section of the population in both rural and urban areas. The food-insecure socioeconomic groups may include: farmers, many of them women, with limited access to natural resources and inputs; landless laborers; rural artisans; temporary workers; homeless people; the elderly; refugees and displaced persons; immigrants; indigenous people; small-scale fishermen and forest dwellers; pastoralists; female-headed households; unemployed or underemployed people; isolated rural communities; and the urban poor. Increasing the productivity and incomes of these diverse groups requires adopting multiple policy instruments and striking a balance between short-term and long-term benefits. The choice of policies must be attuned to the characteristics of a country's food security problem, the nature of the food-insecure population, resource availability and infrastructural and institutional capabilities at all levels of government and communities. Breast-feeding is the most secure means of assuring the food security of infants and should be promoted and protected through appropriate policies and programs.

Source: International Conference on Nutrition, Plan of Action, Rome, Italy, 11 December 1992, from *www.brown.edu/Departments/World_Hunger_Program/hungerweb/intro/food_security.html*; compiled by: Nancy B. Leidenfrost, National Program Leader, Extension Service, USDA.

Four epochs can be distinguished in the history of food security of developing countries. Precolonial societies were dominated by rural, *self-provisioning* economies with various forms of organization generally based on ethnic affiliation. Trade linked remote areas with overseas territories, but production was primarily agricultural—cultivation, livestock rearing, hunting, and gathering. Resilience in the face of drought depended on the ability to store food, either within the household or among kin.

Colonialism sundered traditional land tenure and governance. New forms of vulnerability were created in the transition from self-provisioning to a mixture of local and national governance. The *political economy of colonialism* determined access to land and to famine relief in the case of a drought.

With independence, the state continued to dominate economic systems. However, *weak national economies and political systems* were often unable to respond to famines or indeed to ameliorate widespread impoverishment. The catastrophic famines of the 1970s illustrate the enhanced vulnerability resulting from international and national political conflicts, terms of trade that failed to promote development, and hindrances in information.

Current vulnerability might be labeled *interdependence*. National states no longer dominate in famine early warning systems and food interventions. Market forces predominate in determining access to resources. There is some sense of progress in the ability of international aid organizations to monitor and prevent famines. However, many conditions of food insecurity persist in endemic poverty and in countries and regions isolated for political reasons (see Box 2).

Box 2. World Food Security

The World Food Summit of 1996 reviewed the state of world food security. Gains in agricultural productivity and economic growth over the last 30 years has led to an 18% growth in world per capita food supply. Average per capita dietary energy supply (DES) has grown from 2440 to 2720 calories/capita/day from 1969–1971 to 1990–1992. Aggregate figures, however, do not show the full picture, and hunger persists. In the period 1990 to 1992 an estimated 840 million people remained chronically undernourished, having access to less than 2700 calories per day. In addition, large numbers suffer from micronutrient deficiencies caused by dietary inadequacies, an example being the estimated 1.6 billion suffering from iodine deficiency. In absolute terms food aid has declined, while the increasing number and complexity of emergencies has resulted in a growing proportion (from 30 to 50% in two decades) of total food aid being targeted relief and development food aid.

Source: FAO, Technical Background Documents, World Food Summit, Rome 13–17 November, 1996; *http://www.fao.org/wfs/final/e/list-e.htm.*

In this section three case studies set the scene for further discussions of vulnerability and food security from a household perspective. These case studies indicate the range of situations regarding household food security and coping with climatic variability. In addition, recent research on climate prediction and its role in alleviating food security is summarized.

2 CASE STUDIES OF VULNERABILITY AND COPING

Changing Vulnerability in Central and Eastern Kenya

The failure of the first rains in 1984, following poor second rains in 1983, triggered a serious food crisis in central and eastern Kenya. The drought illustrates diverse impacts and responses in different environments.

The study area comprises six districts in central and eastern Kenya: Kiambu, Murang'a, Embu, Machakos, and Kitui, spanning five agroecological zones (Table 1) [see Downing (1988) and Downing et al. (1989*b*), for details]. The highland zones (I and II) are suitable for coffee and tea as cash crops, along with maize as the staple food crop. Farm sizes are small due to pressure from population growth. In the middle zone (III) maize is the dominant crop, although farms are also relatively small. In the drier zones (IV and V), farms are larger, with more livestock and grazing. Maize is still common, but millet and sorghum are more suitable. The arid ranching zone (VI) is dominated by extensive land uses.

Several long-term trends have affected food security in the region:

- Maize, although more drought prone, has replaced sorghum and millet.
- Farmers have less variety in their cropping systems, partly due to smaller holdings and the reduced availability of common lands.
- Markets for crop inputs, trade, and employment have penetrated the region and are well connected with Nairobi and the coast.

Food production in the six districts declined in 1984. Maize and bean harvests for the 1984 long rains were less than 20% of the harvest from the 1985 long rains. The effect of the drought on livestock was equally severe. Of 565 households surveyed in the study area, over a third had slaughtered, sold, or lost cattle, and a quarter had slaughtered, sold, or lost goats. The average number of cattle per household declined from over 4 in April 1984 to 2.4 in January 1985. The largest declines were in the lower agroclimatic zones. However, famine was averted.

The sources of household income changed (Table 2). Whereas 60% of the households usually have income from sales of farm produce, only 40% did so during the drought. With less food produced, there was less surplus after household consumption. Income from casual labor and businesses decreased as well. A general recession in the rural economy lowered casual employment opportunities and earnings. Income from remittances and permanent employment tended to increase or remain at their usual levels. Although livestock sales increased, the market collapsed and little income was realized.

The principal response by the government was to import yellow maize. Food prices in local markets increased during the drought, but less than would have been the case in the absence of abundant yellow maize imports. Prices of white maize and beans doubled between early 1984 and late 1984, while the imported yellow maize prices remained about the same. In most markets, food was available; complete dearth of food for extended periods was rare.

Households generally survived the food crisis by purchasing their food: participation in the monetary economy reduced their vulnerability to drought. Households are largely self-sufficient in average years with current farming practices and land holdings. A slight surplus can be produced in the humid zones (II), where maize grows well. The larger holdings in the semiarid zones (V) also produce a surplus in years of average to good weather.

Drought affects each zone, with increasing intensity from zones II to V. In a severe drought, production is half of household requirements in zone II, while there is little or no production in zones IV and V. Neither improvements in yields nor adoption of present drought-resistant crops will significantly improve self-sufficiency in drought years. In good and average years, production can be dramatically increased with fertilizers and other inputs. Households in the humid zones could be largely self-sufficient in drought years. But in the drier zones, little improvement in production can be expected in moderate or severe droughts.

In contrast, storage of surplus food is a potentially effective drought-coping strategy, depending on good production during above-average years. However, food storage has a cost. In central Kenya, average food storage in the 1980s was equivalent to 40 to 100 days of consumption, and twice that in eastern Kenya, which is drier and more prone to drought. These low levels of food storage indicate a preference to sell surplus food in order to pay for school fees, medical expenses, and for food not grown on the farm.

Rangeland Management in Botswana

Drought occurs, on average, once every 7 years in Botswana. As elsewhere in Africa, drought has been a common cause of food insecurity. Two studies highlight the confluence of drought and the political economy.

Was the 1979 to 1987 drought more severe than preceding dry periods? Years after the 1979 to 1987 drought in Botswana, the government had not withdrawn relief to many areas and had in some areas expanded relief efforts. Solway (1994) argues that non-meteorological factors are critical to the failure of much of rural Botswana to return to "normal" after 1987. The differential effects of drought—the distribution of both benefits and disadvantages among various classes, races, and among men, women, and children—cannot be sought solely in meteorological factors.

The drought made legitimate a shift of dependency from the extended family to the state and subsequently a greater dependency on the state. Traditional patterns of food security, which permitted semi-independent production on the part of the poorer majority, were significantly eroded during the drought. They have not been revived. In the case of Kalahari villagers, access to draft animals through a chain of entitlements based on kinship relations was replaced by state social security introduced during the drought and then maintained. This shift in dependency was favored by the rural elite, who saw an opportunity to consolidate wealth with the commodification of agriculture and privatization of production occurring in rural Botswana. This shift was also supported by the government, which viewed traditional patron–

TABLE 1 Agroclimatic Zones in Kenya[a]

Zone	R/E_0 (%)	Growing Season (days)	Dry Matter (mt/ha)	Population Density (km^2)	Major Crops	
					Food	Commercial
I	>80	365	>30	333	Beans, maize, potato	Coffee, dairy, tea
II	65–80	290–365	20–30	468	Beans, maize, potato	Dairy, coffee
III	50–65	235–290	12–20	275	Beans, cassava, cow peas, green grams, maize, pigeon peas, sorghum, millet	Cotton, fruit, tobacco
IV	40–50	180–235	7–12	111	Beans, cassava, cow peas, green grams, maize, pigeon peas, sorghum, millet	Cotton, fruit, sunflower, tobacco
V	25–40	110–180	3–7	36	Millet, sorghum	Cotton, livestock, sisal, sunflower
VI	<25	<110	<3	<5		Livestock, sisal

[a] R/E_0 is the ratio of rainfall to potential evaporation. Population density is for 1979.

Source: Sombroek et al. (1982), Downing, Lezberg et al. (1989b) and Jaetzold and Schmidt (1983).

client relations as "backward" and favored "an ideal of individualized nuclear family production units which functioned independently of one another but in conjunction with the state" (Solway, 1994, p. 491). It brought the rural sector in line with modern systems of taxation, land registration, and government regulation.

With the introduction of "welfare" and the loss of access to the local means of production, poorer rural residents were no longer able to farm, and overall rural agricultural production dropped, despite the fact that bumper harvests were reported in several villages during the study period. Falling production statistics provided further "evidence" for the severity of the drought and justified the further expansion of government relief programs.

In Solway's conceptualization, the drought was a revelatory crisis, arising from "structural contradictions" between traditional and modern, bringing latent societal tensions to the surface, and providing a context for the accelerated change of the bases of social reproduction. These structural contradictions were simultaneously

TABLE 2 Sources of Household Income during the 1984 Drought in Kenya

	Agrochemical Zone					
Income Source	I	II	III	IV	V	Total
Farm produce	90[a]	71	50	67	50	62
	98	113	68	28	28	65
Livestock	14	17	14	19	57	29
	100	118	164	200	107	88
Agricultural casual labor	27	25	31	35	31	31
	85	96	94	97	68	87
Nonagricultural casual labor	20	10	18	25	37	23
	100	100	89	84	81	87
Businesses	10	14	15	12	22	15
	70	57	100	125	91	93
Remittances	35	25	41	46	41	39
	86	100	90	109	95	97
Permanent employment	24	24	32	29	35	30
	92	79	106	107	83	97

[a]The first number in each cell is the percent of households for whom the income source is a usual source of income. The second number is the ratio (in percent) of households who had the income source during the 1984 drought (April to December 1984) compared to the number of households who had it as a usual source. Thus over all of the zones, 62% of the households usually have farm produce as a source of (cash) income. During the drought months, only 65% of these households had income from their farm produce.

Source: Anyango et al. (1989).

revealed and concealed during the drought; the discourse of crisis allowed them to be hidden from view by claims that "exceptional" circumstances prevailed. Discourse here served a dual purpose: Not only did it obscure deeper processes at work, it also impelled and made legitimate innovation with normative codes and regulatory, management, and institutional frameworks.

Other research in the Kalahari area of Botswana, investigating the effects of the 1975 tribal grazing lands policy (TGLP), also notes the importance of the social and political dimensions to food security (Thomas and Sporton, 1997b). Studies focused on three areas: study area 1 in eastern Kalahari with tree and bush savannah, study area 2 in Ncojane with dryer bush savannah, and study area 3 in Tshabong, the driest area of arid shrub savannah. It was found that the areas with the most flexible approaches to the TGLP were best able to overcome impacts of environmental variability, notably the drought of 1994 to 1995. The adaptive strategies listed in Table 3 were only possible if some elements of the TGLP, such as fencing of ranches, were not enforced. Study area 2, therefore, enjoyed better livelihood security than study area 1, where TGLP was applied most rigorously, despite the fact that on ecological grounds area 1 offered better opportunities.

The style of management and relationship between ranch lessees and workers also has a major effect on food security. Study area 2 had a more participatory, flexible management style, which minimized hardship (see Box 3).

TABLE 3 Adaptive Strategies Employed in Study Area 2

By livestock holders/ranch lessees
Fences dropped between ranches/paddocks to increase grazing range
Temporary removal of livestock to distant cattleposts
Pooling of resources with neighboring lessees
By ranch residents/other rural groups
Settlement on (temporarily) abandoned ranches
Gathering (and hunting) across neighboring ranches
Move residence to ranch nearest to service center, where drought relief may be available

Source: Thomas and Sporton 1997b.

Box 3. Perceptions of Food Security

Interview from study area 1 (male in his fifties):

"I don't get any food rations from the owner. I just wake up hungry every morning. I help myself live by milking the cows and drinking the milk. I don't know why the owner of the ranch is treating me this way. I don't know if he hates me because of my ethnic group, because I'm a Moswara... When I was a young man I used to be satisfied every day... But now that I am working for someone I don't eat at all."

Interview from other study area 2 (male in his fifties):

"I came here to look after the ranch-owner's cattle. I'm not paid for doing the work. He doesn't pay me, he just buys me food. Not being paid is not a big problem. I'm OK as things are. When I need something I just tell the ranch owner and he buys it for me...The owner is a good man I'm happy to work for him."

Source: Thomas and Sporton (1997a).

Displaced Populations in Sudan

Sudan has a large population of displaced people who have left their original villages and moved elsewhere, often to urban settlements. The Commission of Displaced (COD) estimates that 3.5 million people had been displaced by 1991, of which 0.6 to 1.5 million were in the greater Khartoum area (Kuch, 1993). The major factors forcing the migrations were civil war in the south, which erupted during the dry summer of 1983, and the 1984 drought that resulted in the worst famine to hit northern Sudan in the past 100 years (Kuch, 1993). The underlying causes for displacement go well beyond escaping the impact of drought. A mixture of socioeconomic and environmental pressures is responsible, partly resulting from the

TABLE 4 Crisis-Induced Migration Model (CIM)

1.	Normality (pre-migration situation)
2.	Emergency situation
3.	Local reaction chain
4.	Migration
5.	Sanctuary phase (seeking refuge of a temporary nature)
6.	Settlement phase
7.	Return phase

Source: Elnur et al. (1993, p. 50); developed by the UN Emergency Unit.

dominant development model that for years ignored the traditional sector (Elnur et al., 1993).

Most studies of food security and survival strategies focus on the premigration situation and regard the act of migration itself as the final coping strategy (see Dagnew 1995, p.109). To investigate the food security strategies of displaced populations, it is necessary to look at stages beyond migration (see Table 4). It is important to note that where war is a predominant forcing factor the expected premigration sequence of events may not be followed. The speed and forced nature of this type of migration increases the vulnerability of the displaced as they are likely to have lost all assets, leaving nothing with which to start their postmigration life (Elnur et al., 1993).

The displaced populations are not a homogeneous group, which is reflected in the diversity of the survival strategies adopted (see Table 5). These strategies are a function of many socioeconomic variables such as the predisplacement economic activity of the people. For example, many of those displaced from the south were cattle herders or subsistence farmers with few professional skills. The sex, number, and age of household members are also major determinants of livelihood strategies; 35% of displaced families are female headed (Kuch, 1993). Even within groups with the same basic means of livelihood, tremendous differences can occur; poorer families may have trouble obtaining credit to enable them to diversify their activities.

A study by Kuch (1993) also investigated the food situation of the displaced in Khartoum and found them to be highly vulnerable. People without access to land cannot produce their own food, making them heavily dependent on the market economy. Monetary income is, therefore, crucial to food security in these situations, and people even resort to illegal methods, such as the sale of alcohol, which is punishable with lashes and fines and up to a year imprisonment (Kuch, 1993). Inability to purchase even a single meal a day was frequently experienced. Malnutrition among those under 5 years old was found to double within a year in most settlements. Rations introduced by the Sudan Council of Churches' (SCC) Primary Health Care Program (PHCP) to supplement their diets had to be extended to entire families because sharing of the rations meant that the health status of the under-5-year-olds often did not improve (Kuch, 1993).

TABLE 5 Responses Adopted by Displaced Southern Sudanese to Secure Their Livelihood

Seeking shelter with relatives already residing in planned residential areas
Working as guardians at construction sites
Departure of young men for mechanized agricultural schemes
Retail selling of different consumer items
Selling of rationed supply items
Child labor at markets
Domestic work of women and children, mainly washing and cleaning other people's houses
Working as daily workers at construction sites and industries
Production and distribution of local alcoholic beverages
Depending on government and nongovernmental organisations (NGOs) aid
Begging and collecting of rubbish
Robbery and other illicit activities, including prostitution

Source: Elnur et al. (1993, p. 51).

Government support for food-insecure migrants comes in the form of the "essential commodity distribution card" system, which subsidizes products such as sorghum, sugar, tea leaves, vegetable oil, batteries, matches, and washing soap. For the first 9 months of the scheme, unregistered quarters, including displaced settlements, were not given cards. Even when the displaced were given cards, their purchasing power limited them because, despite subsidies, the commodities' prices remained more than many could afford (Kuch, 1993). Relief food was distributed by agencies under government surveillance.

Many migrants will not achieve truly sustainable livelihoods while in camps and settlements. They await the return to their home area as the only solution to their survival problems (Yath, 1993). Further work is needed in order to assist policy decisions such as planning what type of food distribution (e.g., commercial or relief systems) is best in settlements. Also, more attention needs to be directed at assisting people to develop and expand their own coping strategies.

3 APPROACHES TO COPING, CAPACITY, AND VULNERABILITY

The three case studies highlight diverse situations and ways in which households have coped with climatic variations and other stresses. A rich literature now exists on coping strategies, capacity to withstand shocks and stress, and more generally vulnerability. A few observations on vulnerability frameworks are now presented as an introduction to the following section, which specifically addresses the intersection of vulnerability and climate prediction.

Vulnerability can be defined as the "degree of loss resulting from a potentially damaging phenomenon" (UNDHA, 1992, p. 63) or "the insecurity of the well-being of individuals, households, or communities in the face of a changing environment" (Moser, 1996, p. 2).

Critical questions are: What determines the relationship between a hazard and its effects? Who is vulnerable and why? These questions require a broader analysis of vulnerability. This amplification of vulnerability stems from the literature on development and livelihood security [e.g., see Chambers (1989) and Dow and Downing (1995)] rather than the more circumscribed work on disaster relief. As such, it begins to place vulnerability in the wider structures of political ecology.

Particular vulnerabilities are the conjuncture of social, economic, and political structures. Anderson and Woodrow (1989) chart vulnerability and capability related to physical/material resources, social/organizational relations, and motivational/attitudinal aspects. Bohle et al. (1994) suggest a tripartite causal structure of vulnerability (Fig. 1) based on the human ecology of production, expanded entitlements in market exchanges, and the political economy of accumulation and class processes. Vulnerability per se is best viewed as "an aggregate measure of human welfare that

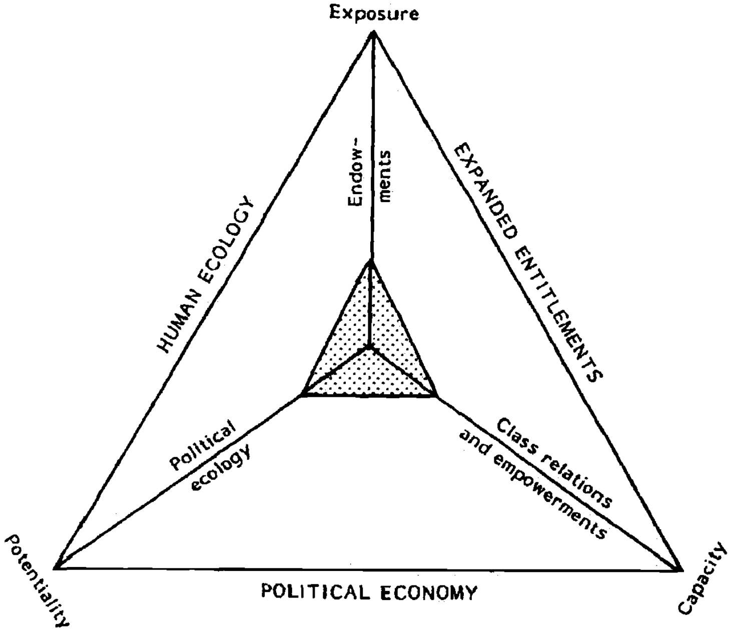


Figure 1 Three dimensions of vulnerability. The fundamental processes that determine vulnerability are implied in the conjuncture of the human ecology of production, exchange entitlement, and political economy. Vulnerable groups can be located in different sectors of the triangle. For instance, subsistence farmers would be more dependent on their land and labor resources than on market exchanges. The destitute and refugees are closely tied to the political economy of aid. The urban poor are dependent on what they can earn in informal markets (after Bohle et al., 1994).

integrates environmental, social, economic and political exposure to a range of harmful perturbations" (Bohle et al., 1994, pp. 37–38). In one of the fullest treatments of vulnerability and disasters, Blaikie et al. (1994) regard vulnerability as a product of such characteristics as ethnicity, religion, caste membership, gender, and age that influence access to power and resources (Fig. 2). One application of the concept of vulnerability is in the U.S. Famine Early Warning System, which monitors food crises in Africa (Fig. 3).

Regardless of the nuance of vulnerability frameworks, key concepts are:

- Vulnerability is a relative measure. The analyst, whether the vulnerable themselves, external aid workers, or various societies that include both the vulnerable and interventionists, must define what is a critical level of vulnerability.
- Everyone is vulnerable, although their vulnerability differs in its causal structure, its evolution, and the severity of the likely consequences.
- Vulnerability relates to the consequences of a perturbation, rather than its agent. Thus people are vulnerable to loss of life, livelihood, assets, and income rather than to specific agents of disaster, such as floods, windstorms or technological hazards. This focuses vulnerability on the social systems rather than the nature of the hazard itself.
- The locus of vulnerability is the individual related to social structures of household, community, society, and world system. Places can only be ascribed a vulnerability ranking in the context of the people who occupy them.

These concepts of vulnerability shift the focus of vulnerability away from a single hazard to the characteristics of the social system. Vulnerability is explicitly a *social* phenomenon, a threat to a human value system. Places and ecosystems can only be termed vulnerable if we ascribe human value to them.

Vulnerability changes over time, incorporating social responses as well as recurrences of hazardous events. Bohle et al. (1994) captures the dynamic nature of vulnerability (Fig. 4). In this illustration, vulnerability begins to increase at the end of the first year, reaching a crisis at 30 months. Here the outcome of the crisis is uncertain. In a resilient society with appropriate interventions, recovery and mitigation can bring vulnerability back down to baseline (or lower) levels. Unmitigated, or in conjunction with another event such as civil strife following drought, the crisis may become a disaster. Or, some groups and communities may continue in crisis, on the edge of disaster. Social groups vary in the structure of their vulnerability. For example, the rural landless (without nonagricultural incomes) are typically more sensitive to food shortages, with less on-farm storage and buffering capacity than smallholders. Thus, the trajectories shown here may be sharper and the outcome different for different groups, even in the same region.

Specific groups of vulnerable peoples can be defined. While the precise boundaries of vulnerability vary between cultures and environments, the common catalog often starts with the characteristics of individuals:

PROGRESSION OF VULNERABILITY

ROOT CAUSE	⇒	*DYNAMIC PRESSURES*	⇒	**UNSAFE CONDITIONS**	⇒	*DISASTERS*	⇐	HAZARDS
Limited access to		**Lack of**		**Fragile physical emnvironment**				**Earthquake**
• Resources		• Institutions		• Dangerous locations		**RISK**		
• Structures		• Training		• Unprotected structures				**Wind storm**
• Power		• Skills				(equals)		
		• Investment		**Fragile local economy**				**Flooding**
Ideologies		• Markets		• Livelihoods at risk		**HAZARD**		
• Political systems		• Press freedom		• Low income				**Volcano**
• Economic		• Civil society				(plus)		
				Vulnerable society				**Landslide**
		Micro-forces		• Groups at risk		**VULNERABILITY**		**Drought**
		• Population		• Little capacity to cope				
		• Urbanization						**Virus and pest**
		• Arms expenditure		**Public actions**				
		• Debt repayment		• Lack of preparedness				**Heatwave**
		• Deforestation		• Endemic disease				
		• Soil degradation						

Figure 2 Structure of vulnerability and disasters. Dynamic pressures are processes that translate social, political, and economic structures in relation to specific types of hazards into particular forms of insecurity. Regional or global pressures such as rapid population growth, urbanization, war, foreign debts, epidemic disease, export promotion, etc. have effects on the local or regional level. Some of these pressures have a universal character, others are specific to a certain region or society. Unsafe conditions reflect situations and circumstances specific to a region and time, in conjunction with a particular group of people. They reflect specific forms of vulnerability related to specific hazards (Blaikie et al., 1994).

Level of Vilnerability	Conditions of vulnerability	Typical Coping Strategies and/or Behaviours	Interventions to Consider
SLIGHTLY VULNERABLE	**Maintaining or Accumulating Assets**	**Assets/resources/wealth:** either accumulating additional assets/resources/wealth or only minimal net change (normal "belt-tightening" or seasonal variations) in assets, resorces or welth over a season/year, i.e., coping to minimize risk	**Development Programs**
	Maintaining Preferred Production Strategy	**Production Strategy:** any changes in production strategy are largely volitional for perceived gain, and not stress related	
MODERATELY VULNERABLE	**Drawing-down Assets**	**Assets/resources/wealth:** coping measures include drawing down or liquidating less important assets, husbanding resources, minimizing rate of expenditure of wealth, unseasonable "belt-tightening" (e.g., drawing down food stores, reducing amount of food consumed, sale of goats or sheep)	**Mitigation and/or Development: Asset Support** (release food price-stabilization stocks, sell animal fodder at "social prices", community grain bank, etc.)
	Maintaining Preferred Production Strategy	**Production strategy:** only minor stress-related change in overall production/income strategy (e.g., minor changes in cropping/planting practices, modest gathering of wild food, inter-household transfers and loans, etc.)	
HIGHLY VULNERABLE	**Depleting Assets**	**Assets/resources/wealth:** liquidating the more important investment, but not yet "production" assets (e.g., salw of cattle, sale of bicycle, sale of professionals such as jewelry)	**Mitigation and/or Relief: Income and Asset Support** (Food-for-work, Cash-for-work, etc.)
	Disrupting Preferred Production Strategy	**Production Strategy:** coping measures being used have a significantly costly or disruptive character to the usual/preferred household and individual life-styles, to the environment, etc. (e.g., time-consuming wage labor, selling firewood, farming marginal land, labor migration of young adults, borrowing from merchants at high interest rates)	
EXTREMELY VULNERABLE	**Liquidating Means of Production**	**Assets/resources/wealth:** liquidating "production" resources (e.g., sale of planting seed, hoes, oxen, land, prime breeding animals, whole herds)	**Relief and/or *Mitigation*: Nutrition, Income and Asset Support** (food relief, seed packs, etc.)
or			
AT-RISK	**Abandoning Preferred Production Strategy**	**Production strategy:** Seeding non-traditional sources of income, employent, or production that preclude continuing with preferred/usual ones (e.g., migration of whole families)	
FAMINE	**Destitute**	**Coping Strategies Exhausted:** no significant assets, resources, or wealth; no income/production	**Emergency Relief** (food, shelter, medicine)

Figure 3 Vulnerability Matrix for the U.S. Famine Early Warning System. Vulnerability is portrayed as a progression from slightly vulnerable to famine. At each level of vulnerability, households pursue different strategies—from production to survival. Consequently, different forms of intervention are warranted for different levels of vulnerability, from targeted development assistance to supporting coping strategies and ultimately emergency food relief [U.S. Famine Early Warning System (FEWS), 1992].

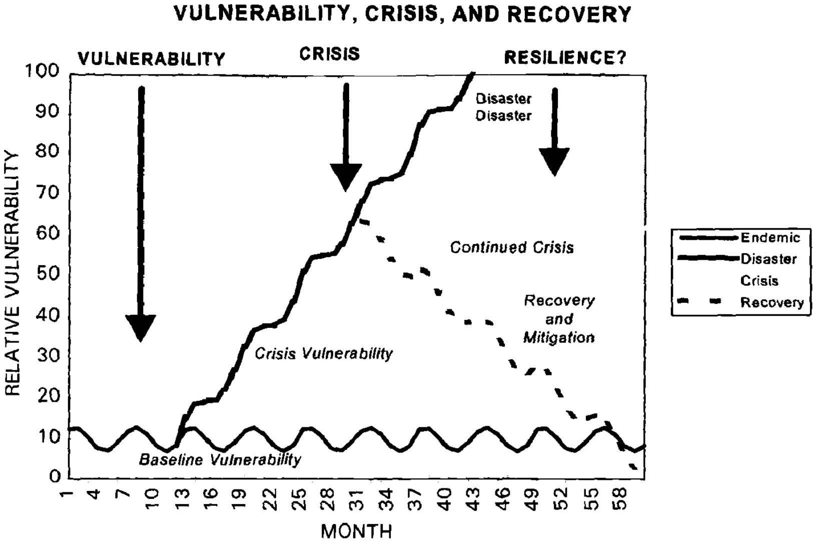


Figure 4 Dynamic vulnerability during a food crisis (after Bohle et al., 1994).

- Women, especially those with special nutritional needs during and after pregnancy
- Children, who are less resilient in terms of nutrition or who may already be malnourished
- Elderly, who may suffer from a lack of mobility and less mental awareness
- Disabled and disease stricken, who have special needs and require routine assistance for survival

At the household level, vulnerability may be delineated by socioeconomic class and means of securing a livelihood: In rural areas:

- Smallholder agriculturalists may be resource poor with limited access to land and labor, in marginal lands, with varying degrees of empowerment and access to emergency and development assistance.
- Pastoralists often have little empowerment to development resources, yet operate in regions with pronounced climatic hazards. However, they often attract international assistance during a disaster.
- Landless laborers relying on casual employment are often at the margin of poverty with little ability to accumulate savings or to invest in more productive activities.
- Destitute peoples have been forced out of productive activities, often because of ill health and old age in addition to being impoverished through natural

disasters and other causes. Where the destitute migrate to urban centers, they may have more opportunities for assistance and work, although this depends on the nature of the receiving society.

In urban areas:

- Unemployed destitute in urban areas may be incorporated into social welfare systems (often informal), but suffer significantly in times of disaster if the numbers become too large and if relief fails to target their pressing needs.
- Underemployed poor people, comparable to landless laborers, are on the margins of survival. A slow deterioration in the economy can affect this group, often leading to a major but largely hidden crisis.
- Refugees are the most visible vulnerable population, usually swelling in numbers after a disaster. They may also be vulnerable to further hazards, for instance, while attempting to return to their homes and occupations or in camps with inadequate protection against floods, heat, and frost, among other hazards. Yet, this group tends to benefit from its visibility and various formal channels of assistance.

Finally, community characteristics can be enumerated that place vulnerable social groups at risk from specific types of hazards. Such vulnerability includes, for example:

- Building location, design, and standards that determine the resilience of homes, workplaces, and community meeting places
- Occupancy patterns and who is in substandard buildings, when, for how long, etc.
- Transport and mobility, for example, the pathways between home and work may cross hazardous regions and access to safe areas such as cyclone shelters in Bangladesh
- Health, water, power, and communication infrastructure that sustain life as well as provide channels for relief assistance.

4 COPING AND CLIMATE PREDICTION

How does the broad range of vulnerability and capacity in Africa relate to emerging skills in climate prediction? Recent developments in seasonal forecasting, especially for the tropics (e.g., Chen et al., 1995), have drawn attention to the opportunity for appropriating such forecast information into drought management systems and other natural resource operations (e.g., Gibberd et al., 1996) (Table 6). Africa is one of the potential beneficiaries of such improved forecasts.

The wide range of effects that climate, and particularly drought, can have is discussed in Glantz's work on El Niño (Glantz, 1996, pp. 145–148). It is emphasized

TABLE 6 Qualitative Assessment of Current Status of Long-Lead Climate Forecasts

	Operational Status Depending on Region			
	Untried	Experimental	Pre-operational	Operational
Multiyear (climate)	Most areas	Global/ hemispheric-scale	—	—
Seasonal	Some equatorial and high-latitude areas	Many areas	Certain promising areas including southern Africa	Parts of the United States, Australia, and a few other suitable areas
Within season	—	Many areas including southern Africa	Some developed economies	Certain well-researched areas (United States, Europe)

Source: Gibberd et al., (1996).

that "weather" does not only affect crop yield but also land quality, on-farm storage, labor migration, rates of urbanization and rural population growth, use of inputs such as fertilizer, farm income, farmers' skill and experience, and so forth. The utility of reliable long-range forecasts, therefore, could be enormous, not just for earlier warning of need for emergency aid but also for ongoing food security (Table 7). Policymakers and farmers alike should benefit.

Seasonal forecasts are already being used in some parts of Africa, for example, in predicting maize yields in Zimbabwe (Cane et al., 1994). For agriculture and water resource management the benefits could be quite extensive, altering the entire basis of economic planning in Africa. Most farmers would benefit from seasonal forecasts, although lead time and reliability will be important issues (Table 8). Several studies have analyzed costs of El Niño events, for example, and predict that considerable savings could be made if accurate warnings of the onset of the phenomenon could be used. The 1991–1992 El Niño-related drought in southern Africa was estimated to cost the U.S. government $800 million in responses to the phenomena (Farmer 1997).

5 CONCLUSIONS

The three case studies focus on drought and associated famines in various parts of Africa over the past decade or so. Droughts in 1983–1984 and 1991–1992 were both described as unusual or the worst to affect the subcontinent in the twentieth century (Rook, 1997). It is remarkable that, despite serious reductions in harvests, widespread hunger was averted, at least in 1991–1992, a fact largely attributed to the

TABLE 7 Users and Potential Applications of Climate Forecasts

	Potential Application of Forecast			
Type of User	Multiyear Forecasts	Seasonal Forecasts	Within Season	Benefits
Commercial producers	Capital and land investment	Acreage planted; planting dates; crop/variety selection; water management	Water management; application of inputs; harvest dates	Increased certainty and reduced risk; improved financial viability; long-term survival; enhancement of comparative advantage
Subsistence producers	Limited, possible diversification and off-farm savings	Planting dates; crop/variety selection	Limited	Improved food security in poor years; improved marketable surpluses in good years
Agricultural support services	Plant and capital investment; research and development priorities; location decisions; production strategies	Product selection; sales forecasts; pricing policy	Adjustments to marketing strategy	Improved financial viability; ability to respond better to farmers' requirements; recovery from drought
Agricultural extension services	Promotion of drought mitigation strategies; development of improved extension advice	Preparation of climate-specific extension advice to subsistence and smallholder producers	Specific adjustments to earlier extension messages and advice	Better extension service to subsistence and smallholder producers

Source: After Gibberd et al. (1996).

TABLE 8 Utility and Requirements for Seasonal Forecasts for African Agriculturists

	Farm Type				Required Lead Time	Required Precision/ Accuracy[a]
Weather	Subsistence	Transitional	Commercial.	Decision	(months)	(%)
Drought	✓	✓	✓	Plant or not plant; choice of crops and tillage; contingency plans for livestock and water	3	90
Overall quality of the rainy season	✓	✓	✓	Choice of crops, crop varieties and tillage; irrigation planning to use impounded water efficiently; arrange seasonal credit	3	80
Onset of planting rains	✓	✓	✓	Timing of field operations; expectation of yield where correlated to planting date	0.5–1	80
Nature of early rains	✓	✓	✓	Whether to risk dry planting, depending on frontal, widespread rainfall or convective, isolated, discontinuous rainfall	0.5	80
Beginning of midseason drought		✓	✓	Choice of variety and planting date	3	80
Length of midseason drought		✓	✓	Choice of crop	3	60
Severity of midseason drought		✓	✓	Preparing to divert grain crops to fodder	3	60
End of rainy season	✓	✓	✓	Timing of harvest operations; possibility of late catchcrops; planning postharvest tillage	2	80

(continued)

TABLE 8 (*continued*)

Weather	Farm Type			Decision	Required Lead Time (months)	Required Precision/ Accuracy[a] (%)
	Subsistence	Transitional	Commercial.			
Amount of winter rains		✓	✓	Plan summer crops for optimum winter cereal crop; possibility of other winter crops	6	80
Distribution of winter rains		✓	✓	Level of inputs to invest in winter crop	1	60
First frost date	✓	✓	✓	Planting date for late planted crops; cut-off planting date for frost-sensitive crops	6	80
Last frost date			✓	Date of winter cereal planting to avoid frost at anthesis; planning spring plantings under irrigation	6	80
Frost frequency over winter			✓	Preparedness for frost on winter horticultural crops	1	40
Dry season severity	✓	✓	✓	Off-season farm capital development; livestock management	1	40
Dry season length	✓	✓	✓	Disposal of crop residues; fodder rationing to livestock; livestock mobility and sales	1	40
Above-normal summer temperatures			✓	Precautions in dairying and horticulture	3	6
Below-normal winter temperatures			✓	Precautions in small stock and horticulture	3	40

[a] Precision/accuracy: 100% completely reliable, 0% same as no forecast.

Source: Based on Gibberd et al. (1996).

rapid responses instigated by regional early warning systems (e.g., Betsill et al., 1997).

The effects of climate, and in particular climatic hazards, depend very much on the socioeconomic vulnerability of the population. The use of climate prediction is also related to vulnerability, both in terms of direct effect and in terms of how easy it is for a given group to access climate predictions and respond to them. Building institutional capacity to provide medium-term climate forecasts to enhance adaptive resource management in Africa would be a major step forward both in achieving present development aims and in preparing for climate fluctuations and change. Research on how to disseminate information and ensure it is applicable at the household level is also crucial.

Recently, increased attention has been paid to the El Niño Southern Oscillation (ENSO) phenomenon and its links with climatic hazards throughout the world, such as droughts in Africa (Glantz, 1997; Wolde-Georgis, 1997). A great deal is being done on the physical side of climate prediction; however, there is a "major gap in the application of research findings" (Glantz, 1997). Improved communications between the climate community and communities dealing with food security, such as famine early warning systems, are needed. For example, forecasters need to be explicit about the spatial and temporal resolution and confidence limits of predictions (Farmer, 1997).

However, an effective climate forecast and use system is not in itself sufficient to bring about increased food security. The forecasting process tends to focus on natural determinants of famine and tends to distract attention from other factors that shape societal and household vulnerability. "There is frequently a danger that forecasting can become an end in itself, detached from many of the social processes that give rise to hunger and starvation" (Tapscott, 1997).

Would better forecasts have altered the outcomes in Kenya, Botswana, and Sudan? The answer depends on how the political economy would have adapted to widespread dissemination of forecasts (and other data on climate and production). Good forecasts, in each case, would not by themselves have been sufficient to ensure early responses, to bolster sustainable livelihoods, and to prevent vulnerable populations from being displaced.

There is a need to ensure that agricultural development occurs in such a way that longer-term sustainability, and not short-term production maximization, is the aim (Rook, 1997). Climate predictions can help in this aim, but other social, economic and political factors must also be considered.

REFERENCES

Anderson, M. B., and P. J. Woodrow, *Risingfrom the Ashes: Development Strategies in Times of Disaster*, Westview, Boulder, CO, 1989.

Anyango, G. J., T. E. Downing, et al., Drought vulnerability in central and eastern Kenya, in T. E. Downing, K. W. Gitu, and C. M. Kamau (Eds.), *Coping with Drought in Kenya: Local and National Strategies*, Lynne Rienner, Boulder, CO, 1989, pp. 169–210.

Bakhit, A. H., Low-salary government employees in Khartoum: Strategies and mechanisms of survival of the urban poor, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, FL, Breitenbach, 1993, pp. 81–96.

Blakie, O., T. Cannon, et al., *At Risk: Natural Hazards, People's Vulnerability, and Disasters*, Routledge, London, 1994.

Bohle, H. G., T. E. Downing, et al., Climate change and social vulnerability: Towards a sociology and geography of food insecurity, *Global Environ. Change*, *4*(1), 37–48, 1994.

Cane, M. A., G. Eshel, et al., Forecasting Zimbabwean maize yield using eastern equatorial Pacific sea surface temperature, *Nature*, *370*, 204–205, 1994.

Chambers, R., Vulnerability, coping and policy, *IDS Bull.*, *20*(2), 1–7, 1989.

Chen, D., S. E. Zebiak, et al., An improved procedure for El Niño forecasting: Implications for predictability, *Science*, *269*, 1699–1702, 1995.

Dagnew, E., Differential socio-economic impact of food shortages and household coping strategies: A case study of Wolaita District in southern Ethiopia, *Afr. Devel.*, *20*(1), 89–124, 1995.

Dow, K., and T. E. Downing, Vulnerability research: Where things stand, *Human Dimensions Q. 1*(3), 3–5, 1995.

Downing, T. E., *Climatic Variability, Food Security and Smallholder Agriculturalists in Six Districts of Central and Eastern Kenya*, Department of Geography, Clark University, Worcester, MA, 1988, p. 262.

Downing, T. E., K. W. Gitu, and C. M. Kamau (Eds.), *Coping with Drought in Kenya*. Lynne Rienner, Boulder, CO, 1989a.

Downing, T. E., S. Lezberg, et al., Population change and environment in central and eastern Kenya from 1969 to 1979, *Environ. Conservation*, 1989b.

Elnur, I., F. Elrasheed, et al., Some aspects of survival: strategies among the Southern Sudan displaced people in Greater Khartoum, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahm (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food–Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, pp. 45–58.

Farmer, G., What does the famine early warning community need from the ENSO research community? *Internet J. Afr. Stud.* (2), available on-line, 1997.

Gibberd, V., J. Rook, et al., *Drought Risk Management in Southern Africa: The Potential of Long Lead Climate Forecasts for Improved Drought Management*, Chatham Maritime, Natural Resources Institute, 1996.

Glantz, M. H., *Currents of Change: El Niño's Impact on Climate and Society*, Cambridge University Press, Cambridge, 1996.

Glantz, M. H. Summary: ENSO/FEWS discussions, *Internet J. of Afr. Stud.* (2), available on-line, 1997.

Glantz, M.H., M. Betsill, and K. Crandall, *Food Security in Southern Africa: Assessing the Use and Value of ENSO Informationl*, NOAA Project Report, ESIG/NCAR, Boulder, CO, 1997.

International, C. o. N., *Plan of Action. N. P. L.*, compiled by N. B. Leidenfrost, Extension Service, USDA, Rome, Italy, available on-line, www.brown.edu/Departments/World_Hunger_Program/hungerweb/intro/food_security.html, 1992.

Jaetzold, R., and H. Schmidt, *Farm Management Handbook of Kenya*, Ministry of Agriculture, Nairobi, 1983.

Kuch, P. J., Food situation among the displaced Sudanese in Khartoum State, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, 33–44.

Maxwell, S., *Food Security: A Post-Modern Perspective*, Institute of Development Studies, 1994.

Moser, C. O. N., *Confronting Crisis: A Summary of Household Responses to Poverty and Vulnerability in Four Poor Urban Communities*, World Bank, Washington, D.C., 1996.

Rook, J. M. The SADC regional early warning system: Experience gained and lessons learnt from the 1991–92 Southern Africa drought, *Internet J. Afr. Stud.* (2), available on-line, http://www.dir.ucar.edu/esig/enso, 1997.

Solway, J. S., Drought as a "Revelatory crisis": An exploration of shifting entitlements and hierarchies in the Kalahari, Botswana, *Devel. Change*, *25*: 471–495, 1994.

Sombroek, W. G., H. M. H. Braun, et al., *Exploratory Soil Map and Agroclimatic Zone Map of Kenya*, Kenya Soil Survey, Nairobi, 1982.

Tapscott, C., Is a better forecast the answer to better food security? To better early warning? To better famine prevention? *Internet J. Afr. Stud.* (2), available on-line, 1997.

Thomas, D. A. G., and D. Sporton, Understanding the dynamics of social and environmental variability, *Appl. Geogr.*, *17*(1), 11–27, 1997a.

Thomas, D. A. G.. and D. Sporton *Environmental Change and Poverty in Kalahari Pastoral Systems: Full Report of Research Activities and Results*, available on-line, http://www.shef.ac.uk/uni/academic/I-M/idry/Esrcreport.html, 1997b.

United Nations Department of Humanitarian Affairs (UNDHA), *Glossary: Internationally Agreed Glossary of Basic Terms Related to Disaster Management*, Geneva, UNDHA, 1992.

Wolde-Georgis, T., El Niño and drought early warning in Ethiopia, *Internet J. Afr. Stud.* (2), available on-line, 1997.

Yath, Y. A., Dinka migrants in Khartoum: Coping strategies in the face of economic and political hardships. The example of the Suq El Markazi squatter settlement, Sudan, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, pp. 59–80.

CHAPTER 39

DROUGHT IN THE U.S. GREAT PLAINS

DONALD A. WILHITE

1 INTRODUCTION

Drought, a normal feature of the climate for virtually all portions of the United States, is one of the defining characteristics of the Great Plains region. Early maps referred to this region as the Great American Desert, a belief attributed to the explorations of Zebulon Pike in the early 1800s (Brown, 1948). The region's past is firmly rooted in the drought of the 1890s and, in particular, the Dust Bowl years of the 1930s (Hurt, 1981). More recently, droughts have occurred at regular intervals, affecting all portions of a region that stretches from Texas and New Mexico northward through the Dakotas and Montana into the Prairie Provinces of Alberta, Saskatchewan, and Manitoba. In reality, it is rare for drought not to occur in the region each year, a fact that has forced considerable adjustments from a predominantly agricultural economy. Irrigation development and many other technological adjustments in the post-1930s era have improved the resilience of the region to the ravages of drought, but drought continues to produce devastating and widespread impacts.

The purpose of this chapter is to discuss drought in the context of the Great Plains. In that process, I will review some of the basic concepts of drought. A grasp of these concepts is essential to understanding the history and impacts of drought in the Great Plains and the region's continuing vulnerability to this insidious natural hazard. Current and future attempts to lessen drought impacts in the region are also discussed.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

2 CONCEPT OF DROUGHT: DEFINITION AND TYPES

Drought is the consequence of a natural reduction in the amount of precipitation received over an extended period of time, usually a season or more in length, although other climatic factors (such as high temperatures, high winds, and low relative humidity) are often associated with it in many regions of the world and can significantly aggravate the severity of the event (Wilhite, 1992, 2000). High winds and low relative humidity aggravated the effects of the drought of the 1930s in the Great Plains. Drought is also related to the timing (i.e., principal season of occurrence, delays in the start of the rainy season, occurrence of rains in relation to principal crop growth stages) and the effectiveness (i.e., rainfall intensity, number of rainfall events) of the rains. Thus, each drought is unique in its climatic characteristics and impacts. Likewise, society is changing in response to increasing and shifting population, new technologies, government policies, and social behavior. These factors alter vulnerability, a fact that will be discussed in greater detail later in this chapter.

Drought is a temporary aberration that occurs in high- and low-rainfall regions (Wilhite, 1992). Although droughts are commonly associated with the Great Plains and other semiarid regions, it is difficult for many people to visualize drought occurring in more humid regions such as the eastern United States, Southeast Asia, Brazil, or western Europe. This fact emphasizes both the regional and relative nature of drought.

Drought differs from other natural hazards (e.g., floods, tropical cyclones, and earthquakes) in several ways. First, drought is a slow-onset natural hazard. It is often referred to as a creeping phenomenon (Tannehill, 1947). The effects of drought accumulate slowly over a considerable period of time and may linger for years after the termination of the event. As a result, the onset and end of drought are difficult to determine. Even today, with more sophisticated monitoring technology, climatologists struggle to recognize the onset of drought, and scientists and policymakers continue to debate the basis (i.e., criteria) for declaring an end to a drought. Second, the absence of a precise and universally accepted definition of drought adds to the confusion about whether or not a drought exists and, if it does, its degree of severity. Realistically, definitions of drought must be region-specific and application (or impact) specific. Wilhite and Glantz (1985) analyzed more than 150 definitions in their classification study, and many more definitions exist. Although the definitions are numerous, many do not adequately define drought in meaningful terms for scientists or policymakers. Third, drought impacts are nonstructural and spread over a larger geographical area than are damages that result from other natural hazards. For example, a recent analysis of drought occurrence by the U.S. National Drought Mitigation Center for the 48 contiguous states in the United States demonstrated that severe and extreme drought affected more than 25% of the country in 27 of the past 100 years.

Because drought affects virtually all regions of the world and many economic and social sectors, scores of definitions exist. Impacts are complex, vary on spatial and temporal scales, and depend on the societal context of drought. The impacts of drought in the Great Plains will differ from those experienced in the southeast, northeast, or far western portions of the United States. As a result, it is not possible to formulate a definition of drought that is universally acceptable in each of these settings. Wilhite and Glantz (1985) concluded that definitions of drought should reflect a regional bias since water supply is largely a function of climatic regime.

Drought has been grouped by type as follows: meteorological, agricultural, hydrological, and socioeconomic (Wilhite and Glantz, 1985). Meteorological (or climatological) drought is expressed solely on the basis of the degree of dryness (often in comparison to some normal or average amount) and the duration of the dry period. The *Encyclopedia of Climate and Weather* (Schneider, 1996) defines drought as an extended period—a season, a year, or several years—of deficient rainfall relative to the statistical multiyear mean for a region. This definition identifies two critical components that must be accounted for in a viable definition—intensity and duration. Meteorological drought definitions must be considered as region specific since the atmospheric conditions that result in deficiencies of precipitation are climate regime dependent. In the Great Plains, the distribution of precipitation is seasonal: Approximately 70% of the precipitation in this region occurs during the 6-month period from April to September. Definitions that differentiate meteorological drought on the basis of the number of days with precipitation less than some specified threshold (e.g., for Britain, 15 days, none of which received as much as 0.25 mm of precipitation; British Rainfall Organization, 1936) rather than the magnitude of the deficiency over some period of time would be inappropriate for the Great Plains.

Agricultural drought links various characteristics of meteorological drought to agricultural impacts, focusing on precipitation shortages, differences between actual and potential evapotranspiration (ET), soil water deficits, and so forth. Rosenberg (1980) defined agricultural drought as a climatic excursion involving a shortage of precipitation sufficient to adversely affect crop production or range productivity. A definition of agricultural drought should account for the variable susceptibility of crops at different stages of crop development. The impacts of drought are crop specific because the most weather-sensitive phenological stages vary between crops. Planting dates and maturation periods also vary between crops and locations. A period of high-temperature stress that occurs in association with dry conditions may coincide with a critical weather-sensitive growth stage for one crop while missing a critical stage for another crop. Agricultural planning can often reduce the risk of drought impact on crops by altering the crop, genotype, planting date, and cultivation practices. Considerable progress has been made in the Great Plains and elsewhere in applying various types of adaptive strategies to reducing the impacts of drought on crop and rangeland (Rosenberg, 1980, 1986).

Agricultural droughts usually take 3 months or more to develop, but this time period can vary considerably, depending on the timing of the initiation of the precipitation deficiency. For example, in the Great Plains a significant dry period during the winter season may have few, if any, impacts for many locales. However, if this deficiency continues into the growing season, the impacts may magnify quickly since low precipitation during the autumn and winter season results in low soil moisture recharge rates, leading to deficient soil moisture at spring planting. Although the region's agriculture is usually the first to feel the effects of drought, a prolonged dry period can result in significant disruptions in other sectors, particularly water-based transportation, energy production, municipal water supplies, and recreation-based businesses. Also, a short-lived drought of less than 3 months can have serious impacts on crop yields if it occurs during the critical crop growth stages and is accompanied by high temperatures.

Hydrological droughts are associated more with the effects of periods of precipitation shortfall on surface or subsurface water supply (i.e., streamflow, reservoir and lake levels, groundwater) than with precipitation shortfalls directly (Dracup et al., 1980; Klemês, 1987). Hydrological droughts are usually out of phase or lag the occurrence of meteorological and agricultural droughts. Meteorological droughts result from precipitation deficiencies; agricultural droughts are largely the result of soil moisture deficiencies. More time elapses before precipitation deficiencies are detected in other components of the hydrological system (e.g., reservoirs, groundwater). As a result, impacts from hydrological drought are out of phase with those in other economic sectors. Water in hydrological storage systems (e.g., reservoirs, rivers) is often used for multiple and competing purposes (e.g., power generation, flood control, irrigation, recreation), further complicating the sequence and quantification of impacts. Competition for water in these storage systems escalates during drought, and conflicts between water users increase significantly. For example, changing water-use patterns and the series of drought years that occurred in the Missouri River basin between 1987 and 1992 resulted in significant conflicts between upstream and downstream water users.

Finally, socioeconomic drought associates the supply and demand of some economic good or service with elements of meteorological, hydrological, and agricultural drought. For example, the supply of some economic good (e.g., water, hay, hydroelectric power) is highly sensitive to the vagaries of weather. In most instances, the demand for that good is increasing as a result of increasing population and/or per capita consumption. Therefore, drought could be defined as occurring when the demand for that good exceeds supply as a result of a weather-related supply shortfall (Sandford, 1979). This concept of drought supports the strong symbiosis that exists between drought and human activities. Thus, the incidence of drought could increase because of a change in the frequency of the physical event, a change in societal vulnerability to water shortages, or both. For example, poor land-use practices such as overgrazing can decrease animal carrying capacity and increase soil erosion, which exacerbates the impacts of and vulnerability to future droughts. Overdrafts of groundwater, such as have been occurring in the southern portions of the Ogallala aquifer, will affect future vulnerability to drought in the Great Plains since access to groundwater is a primary adaptive strategy employed to alleviate the effects of drought.

3 DROUGHT CLIMATOLOGY OF THE GREAT PLAINS

Historical climate records for the Great Plains provide only a brief snapshot of the drought climatology for the region. For most portions of the region, climatic records cover the period since about 1900, and at only a few locations. To learn more about the occurrence and patterns of drought before that time, tree-ring data can be used to reconstruct the drought history of the region. These data provide insights into past climates extending back many centuries. Figure 1 is adapted from the work of Weakly (1965) for western Nebraska. His work was based largely on tree rings from Red Cedar for a 748-year period, from 1210 to 1958. The results of his study showed the occurrence of 21 drought periods of 5 years or more duration. The most remarkable of these drought periods was from 1276 to 1313, a period of 38 years. However, the average duration of the droughts was 12.8 years. A similar study was conducted in northern and southern Texas by Stahle and Cleaveland (1988) for the period 1698 to 1980. This analysis revealed numerous drought events during this nearly 300-year study period. The most severe individual drought years before 1900 for north Texas were 1772, 1790, 1805, 1855, 1872, and 1887. The years 1917, 1925, 1939, and 1956 were the driest years of the twentieth century. For south Texas, individual drought years before 1900 were 1790, 1805, 1855, 1857, and 1887, and the driest years of the twentieth century occurred in 1917, 1925, 1956, 1967, and 1971. Other tree-ring studies in the region confirm the recurrent nature of single and multiyear droughts in the region (Stockton and Meko, 1975, 1983).

Figure 2 provides a historical perspective of the percent area of the Great Plains in severe to extreme drought, according to the Palmer Drought Severity Index (PDSI) (Palmer, 1965) from 1895 to 1995. The PDSI is a meteorological drought index that integrates many variables in a water balance accounting procedure. This index is calculated routinely for each of the climate divisions in the United States. PDSI values commonly range from + 4.0 (extreme wetness) to − 4.0 (extreme drought), although values above and below these levels are often computed. For example, during August 1977, PDSI values reached − 7.0 in parts of the upper Midwest. For the Great Plains region, Figure 2 illustrates three interesting characteristics. First, the percent area in drought is highly variable from year to year. The peak drought year was 1934 in which 95% of the region was experiencing severe to extreme drought; other severe drought years were 1936 and 1956, with 90 and 80%, respec-

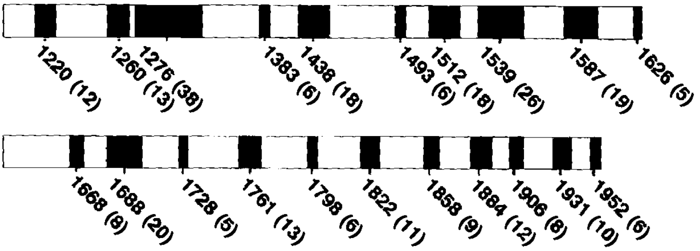


Figure 1 Periods of drought in western Nebraska, five or more years in duration, 1200–1960. Periods of drought shown in black. Numbers in parentheses following year indicate length of drought period. Average duration of drought: 12.8 years (adapted from Weakly, 1965).

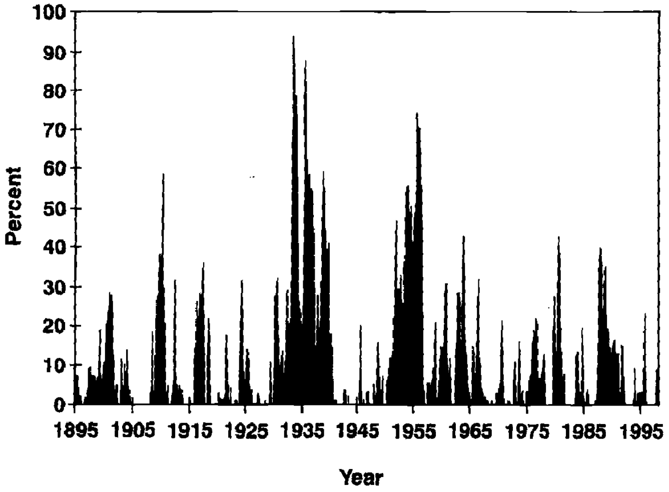


Figure 2 Percent area of the Great Plains experiencing severe to extreme drought, 1895–1995.

tively, of the region in severe and extreme drought. Second, it is rare for severe drought not to be occurring at some location in the region in every single year during 1895 to 1995. Third, clusters of drought years, while rare, are particularly noticeable in the 1930s, mid-1950s, late 1970s, and late 1980s to early 1990s. Multiyear droughts are important because impacts magnify as drought continues into a second and subsequent years. As surface and subsurface water supplies are gradually depleted, more economic sectors are affected. It often takes years for these systems (i.e., reservoirs, groundwater) to recover following an extended drought episode.

Figure 3 illustrates the percent area of three river basins in the Great Plains region in severe to extreme drought during the period 1895 to 1995. The river basins shown are the Missouri, Arkansas–White–Red, and Rio Grande. These three basins encompass most of the area of the Great Plains. The drought climatology of these three basins displays characteristics similar to those shown in Figure 2. However, there are two important differences. First, largely because of the spatial characteristics of drought, it is more common for drought to affect nearly 100% of these basins. This is especially noticeable for the Arkansas–White–Red and Rio Grande basins. This information is important to planners and water supply managers. Second, the pattern of drought differs from one portion of the region to another. For example, the 1930s drought was of much greater duration in the Missouri Basin than in the Rio Grande. By contrast, the 1950s drought was of greater duration and intensity in the Rio Grande basin. This illustrates the regional nature of drought and the fact that planning must be based on the drought of record for the area of interest. Tree-ring data can help scientists reconstruct the long-term climatic history of the region.

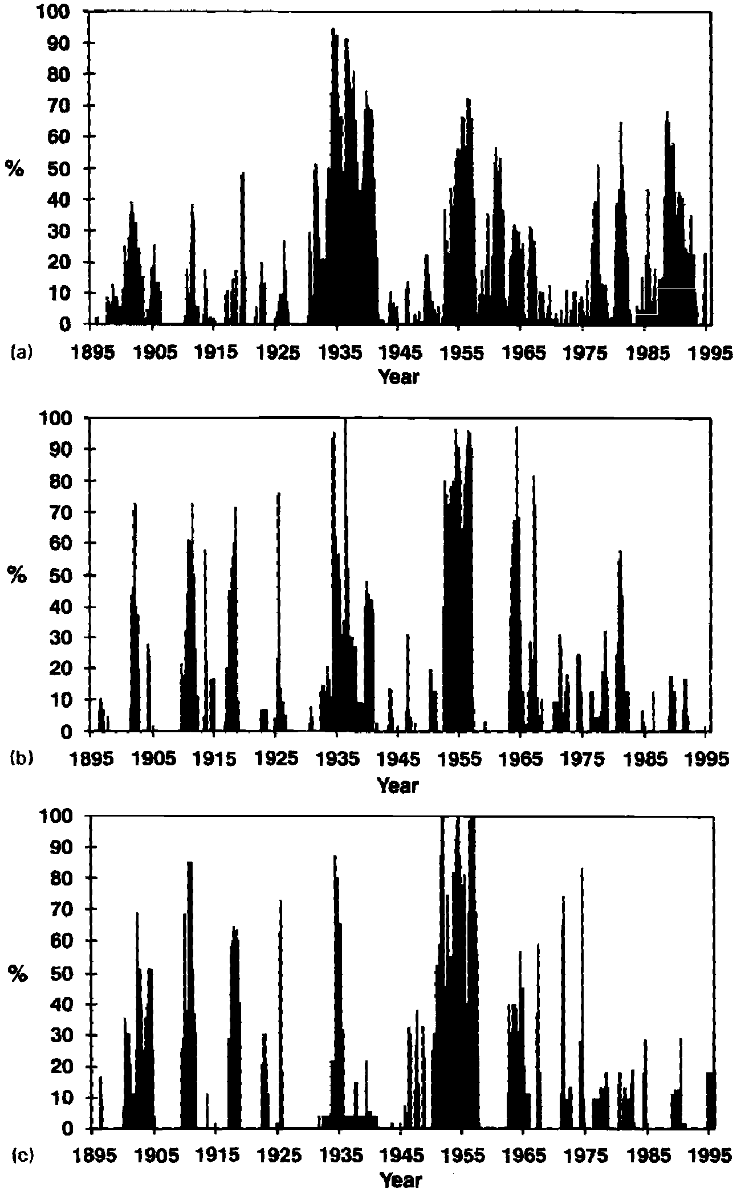


Figure 3 Percent area of (a) Missouri Basin, (b) Arkansas–White–Red Basin, and (c) Rio Grande Basin in severe and extreme drought (i.e., ≤ -3.0) during the period 1895–1995.

4 IMPACTS OF DROUGHT

Drought produces a complex web of impacts that not only ripple through many sectors of the economy but may be experienced well outside the affected region, extending even to the global scale. This complexity is largely caused by the dependence of so many sectors on water for producing goods and providing services. Agricultural production in the Great Plains is of critical importance to food production in the United States. Substantial drought-related production losses not only affect food supplies and prices in this country but also have serious implications for the many nations that depend on U.S. grain exports to offset domestic supply shortfalls.

Impacts from drought are commonly classified as direct or indirect. Reduced crop, rangeland, and forest productivity; increased fire hazard; reduced water levels; increased livestock and wildlife mortality rates; and damage to wildlife and fish habitat are a few examples of direct impacts. The consequences of these impacts illustrate indirect impacts. For example, a reduction in crop, rangeland, and forest productivity may result in reduced income for farmers and agribusiness, increased prices for food and timber, unemployment, reduced government tax revenues because of decreased expenditures, increased crime, foreclosures on bank loans to farmers and businesses, migration, and disaster relief programs. Direct or primary impacts are usually of a biophysical nature. Conceptually speaking, the more removed the impact from the cause, the more complex the link to the cause.

Because of the number of affected groups and sectors associated with drought, the geographic size of the area affected, and the difficulties in quantifying environmental damages and personal hardships, the precise determination of the financial costs of drought is a formidable challenge. The economic costs and losses associated with drought are highly variable from year to year. These costs and losses are also quite variable from one drought year to another in the same place, depending on timing, intensity, and spatial extent of the droughts.

The impacts of drought are commonly classified as economic, environmental, and social. Table 1 presents a comprehensive list of the impacts associated with drought. This list represents the experiences of the Great Plains and many other drought-prone areas of the world. Although drought produces impacts that are regionally distinct, there are many similarities in the types of impacts experienced from one region to another. Many economic impacts occur in broad agricultural and agriculturally related sectors, including forestry and fisheries, because of the reliance of these sectors on surface and subsurface water supplies. In addition to obvious losses in yields in both crop and livestock production, drought is associated with increases in insect infestations, plant disease, and wind erosion. Droughts also bring increased problems with insects and diseases to forests and reduce growth. The incidence of forest and range fires increases substantially during extended droughts, which in turn places both human and wildlife populations at higher levels of risk.

Income loss is another indicator used in assessing the impacts of drought because so many sectors are affected. Reduced income for farmers has a ripple effect, as their ability to purchase goods and services is limited. Thus, many retailers experience

TABLE 1 Classification of Drought-Related Impacts (Costs and Losses)

Problem Sectors	Impacts
Economic	Loss from crop production
	Annual and prerennial crop losses; damage to crop quality
	Reduced productivity of cropland (wind erosion, etc.)
	Insect infestation
	Plant disease
	Wildlife damage to crops
	Loss from dairy and livestock production
	Reduced productivity of rangeland
	Forced reduction of foundation stock
	Closure/limitation of public lands to grazing
	High cost/unavailability of water for livestock
	High cost/unavailability of feed for livestock
	High livestock mortality rates
	Increased predation
	Range fires
	Loss from timber production
	Forest fires
	Tree disease
	Inset infestation
	Impaired productivity of forest land
	Loss from fishery production
	Damage to fish habitat
	Loss of young fish due to decreased flows
	Loss of national economic growth, retardation of economic development
	Income loss for farmers and others directly affected
	Loss of farmers through bankruptcy
	Loss to recreational and tourism industry
	Loss to manufacturers and sellers of recreational equipment
	Increased energy demand and reduced supply because of drought-related power curtailments
	Costs to energy industry and consumers associated with substituting more expensive fuels (oil) for hydroelectric power
	Loss to industries directly dependent on agricultural production (e.g., machinery and fertilizer manufacturers, food processors, etc.)
	Decline in food production/disrupted food supply
	Increase in food prices
	Increased importation of food (higher costs)
	Disruption of water supplies
	Unemployment from drought-related production declines
	Strain on financial institutions (foreclosures, greater credit risk, capital shortfalls, etc.)
	Revenue losses to federal, state, and local governments (from reduced tax base)
	Deters capital investment, expansion
	Dislocation of businesses

(*Continued*)

TABLE 1 ***(continued)***

Problem Sectors	Impacts
	Revenues to water supply firms
	Revenue shortfalls
	Windfall profits
	Loss from impaired navigability of streams, rivers and canals
	Cost of water transport or transfer
	Cost of new or supplemental water resource development
Environmental	Damage to animal species
	Reduction and degradation of fish and wildlife habitat
	Lack of feed and drinking water
	Disease
	Increased vulnerability to predation (e.g., from species concentration near water)
	Loss of biodiversity
	Wind and water erosion of soils
	Reservoir and lake drawdown
	Damage to plant species
	Water quality effects (e.g., salt concentration, increased water temperature, pH, dissolved oxygen)
	Air quality effects (dust, pollutants)
	Visual and landscape quality (dust, vegetative cover, etc.)
	Increased fire hazard
	Estuarine impacts; changes in salinity levels, reduced flushing
Social	Increased groundwater depletion (mining), land subsidence
	Loss of wetlands
	Loss of cultural sites
	Insect infestation
	Food shortages (decreased nutritional level, malnutrition, famine)
	Loss of human life (e.g., food shortages, heat)
	Public safety from forest and range fires
	Conflicts between water users, public policy conflicts
	Increased anxiety
	Loss of aesthetic values
	Health-related low flow problems (e.g., diminished sewage flows, increased pollutant concentrations, etc.)
	Recognition of institutional constraints on water use
	Inequality in the distribution of drought impacts/relief
	Decreased quality of life in rural areas
	Increased poverty
	Reduced quality of life, changes in life-style
	Social unrest, civil strife
	Population migration (rural to urban areas)
	Reevaluation of social values
	Increased data/information needs, coordination of dissemination activities
	Loss of confidence in government officials
	Recreational impacts

[a] Input from working groups was used to modify a table from Wilhite (1992).

significant reductions in sales. This leads to unemployment, increased credit risk for financial institutions, capital shortfalls, and loss of tax revenue for local, state, and federal government. The recreation and tourism industries are also affected because of less discretionary income. Prices for food, energy, and other products increase as supplies are reduced. In some cases, local supply shortfalls for certain goods will result in the importation of these goods from outside the stricken region. Reduced water supply impairs the navigability of rivers and results in increased transportation costs because products must be transported by rail or truck. Hydropower production is also significantly reduced. For example, hydropower generation was 25 to 40% below average for large sections of the country in 1988, resulting in serious revenue losses for the industry (Wilhite, 1993a).

Environmental losses are the result of damages to plant and animal species, wildlife habitat, and air and water quality; forest and range fires; degradation of landscape quality; loss of biodiversity; and soil erosion. Some of the effects are short term and conditions quickly return to normal following the end of the drought. Other environmental effects linger for some time or may even become permanent. Wildlife habitat, for example, may be degraded through the loss of wetlands, lakes, and vegetation. However, many species will eventually recover from this temporary aberration. The degradation of landscape quality, including increased soil erosion, may lead to a more permanent loss of biological productivity of the landscape. Although environmental losses are difficult to quantify, growing public awareness and concern for environmental quality has forced public officials to focus greater attention and resources on these effects.

Social impacts mainly involve public safety, health, conflicts between water users, reduced quality of life, and inequities in the distribution of impacts and disaster relief. Many of the impacts specified as economic and environmental have social components as well. Population out-migration was a significant problem in the Great Plains in response to the 1930s drought and continues to be a major problem in many countries.

As with all natural hazards, the economic impacts of drought are highly variable within and between economic sectors and geographic regions, producing a complex assortment of winners and losers with the occurrence of each disaster. For example, decreases in agricultural production result in enormous negative financial impacts on farmers in drought-affected areas, at times leading to foreclosure. This decreased production also leads to higher grain, vegetable, and fruit prices. These price increases have a negative impact on all consumers as food prices increase. However, farmers outside the drought-affected area with normal or above-normal production or those with significant grain in storage reap the benefits of these higher prices. Similar examples of winners and losers could be given for other economic sectors as well.

5 DROUGHT MANAGEMENT

Although drought is a natural hazard, the term *drought management* implies that human intervention can reduce vulnerability and impacts (Wilhite, 1993b). To be successful in this endeavor, many disciplines must work together in tackling the complex issues associated with detecting, responding to, and preparing for the inevitability of future events. To improve our management of drought requires that we view drought as having both a natural component and a social component. In other words, the risk associated with drought in the Great Plains or any region is a product of both its exposure to the event (i.e., probability of occurrence at various severity levels) and the vulnerability of society to the event. The natural event (i.e., meteorological drought) is a result of the occurrence of persistent large-scale disruptions in the global circulation pattern of the atmosphere. Exposure to drought varies spatially and there is little, if anything, that we can do to alter the occurrence of meteorological drought. As previously discussed, the Great Plains has historically had a very high incidence of drought. Certainly, there is no reason to believe that this incidence will diminish in the future. In fact, output from general circulation models suggest that the interiors of midlatitude continents are likely to become drier as a result of increasing concentrations of greenhouse gases in the atmosphere. This would result not only from increasing temperatures and associated increases in evapotranspiration but also from possible changes in the amount, seasonal distribution, and effectiveness of precipitation. The result may be a net loss in soil moisture during the growing season. If these projections are correct, the incidence of drought in the Great Plains may increase.

Vulnerability, on the other hand, is determined by factors such as population, demographic characteristics, technology, policy, and social behavior. These factors change over time, and thus vulnerability is likely to increase or decrease in response to these changes. Subsequent droughts in the same region will have different effects, even if they are identical in intensity, duration, and spatial characteristics, because societal characteristics have changed.

Much has been done to lessen societal vulnerability to drought in the Great Plains. The widespread adoption of irrigation, conservation tillage practices, soil evaporation reduction, snow management, and irrigation scheduling have all proved effective in stabilizing agricultural production in a region exposed to the vagaries of weather. However, we must continue to implement new mitigation techniques and preparedness strategies in the face of drought. Recent droughts illustrate our continuing vulnerability to extended periods of water shortage. In 1988, drought affected nearly 40% of the nation and resulted in nearly $16 billion in agricultural losses (Riebsame, et al. 1991). In the Great Plains, this drought had serious impacts on spring wheat yields, reducing yields by 54% (Riebsame, et al. 1991). In 1989, winter wheat and sorghum yields were reduced substantially in parts of the central Great Plains. In 1996, drought in the Southwest and southern Great Plains states resulted in substantive agricultural losses, increased incidence of forest and range fires, municipal water supply problems, and losses in recreation and tourism. In Texas alone, the 1996 drought losses were estimated to be $6.5 billion (Boyd, 1996). The Federal Emergency Management Agency (FEMA, 1995) recently estimated annual losses resulting from drought in the United States at $6 to $8 billion.

Drought planning is one of the mechanisms being employed by many states in the Great Plains and nationwide to reduce the economic losses and personal hardships

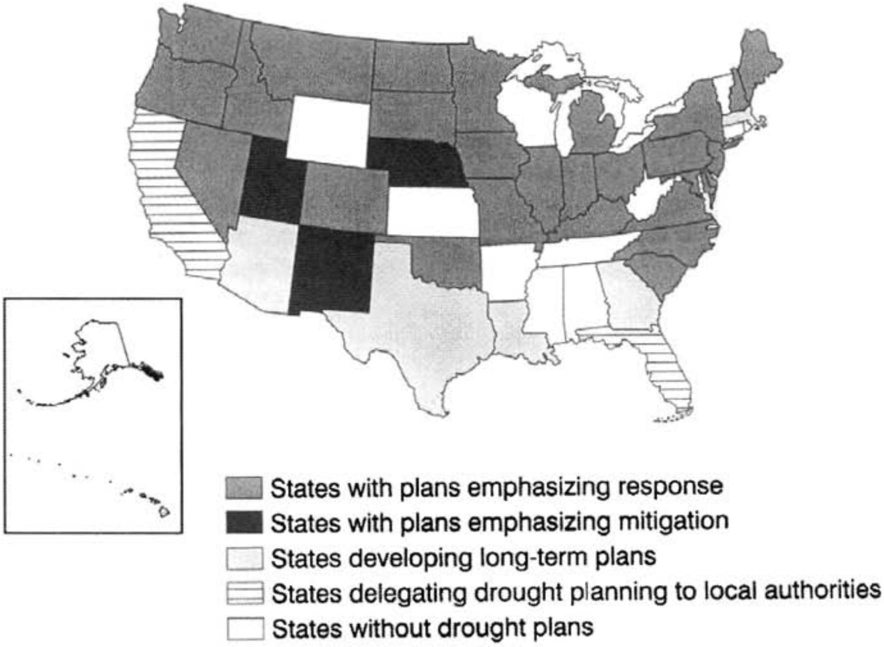

States with plans emphasizing response
States with plans emphasizing mitigation
States developing long-term plans
States delegating drought planning to local authorities
States without drought plans

Figure 4 Status of state drought plans as of January 2001.

associated with drought. The number of states in the United States with drought plans has grown from 3 in 1982 to 27 in 1997 (Fig. 4) (Wilhite, 1997a). In addition to the states that now have plans, six states (Alabama, Louisiana, Texas, New Mexico, Arizona, and Pennsylvania) are at various stages of plan development or have expressed intent to develop a plan. In the U.S. portion of the Great Plains region, all states except Kansas and Wyoming have developed plans. Alberta has also undertaken some initial steps in drought plan development.

The basic goal of state drought plans is to improve the effectiveness of state response and preparedness efforts. This is accomplished by improving monitoring and early warning, impact and vulnerability assessment, and preparedness, response, recovery, and mitigation programs (Wilhite, 1997b). These plans are also directed at improving coordination and building partnerships within agencies of state government and between state, local, and federal governments. The growth in the number of states with drought plans suggests an increased concern about the potential impact of extended water shortages and an attempt to address those concerns through planning. However, more attention needs to be placed on mitigation, defined as short- and long-term actions, programs, or policies implemented in advance that reduce the degree of risk to people, property, and productive capacity (Wilhite, 1997b).

In 1997, the Western Drought Coordination Council (WDCC) was formed under the auspices of the Western Governors' Association (WGA) as a result of a memorandum of understanding between WGA and key federal agencies (Departments of Agriculture, Interior, and Commerce, FEMA, and the Small Business Administration). This activity represents an important attempt to build regional partnerships between local, state, federal, and tribal governments to reduce the impacts of future drought events in the western states through greater attention to planning and mitigation. The Great Plains states are actively participating in the WDCC. These types of improved institutional arrangements, in combination with the existence of drought plans and the application of new technologies, are an important new trend in mitigating the effects of drought in the Great Plains states and elsewhere (Wilhite, 1997c).

6 SUMMARY

Drought is a complex, recurrent, and insidious natural hazard that has historically resulted in significant impacts in the Great Plains. Its impacts are far-reaching and may linger for months or even years beyond the termination of the event. The economic, social, and environmental impacts of drought result from complex interactions between physical and social systems, and they are difficult to quantify. Scientists and policymakers must understand the characteristics of drought and appreciate the magnitude and complexity of impacts in order for viable assessment and response strategies to be established. The aim of these strategies is to reduce societal vulnerability to periods of water shortages.

Drought inflicts considerable pain and hardship on society. The impacts of contemporary droughts in the Great Plains have demonstrated this fact repeatedly over the past several decades. Drought illustrates, in innumerable ways, the vulnerability of economic, social, political, and environmental systems to a variable climate. It also illustrates the dependencies that exist between systems, reinforcing the need for improved coordination within and between levels of government.

Extended periods of normal or benign weather conceal the vulnerability of societies to climate variability, but drought exposes these sensitivities. Projected changes in climate because of increased concentrations of carbon dioxide and other atmospheric trace gases suggest a possible increase in the frequency and intensity of severe drought in the Great Plains region. In a region where the incidence of drought is already high, any increase in drought frequency will place even greater pressure on the region's already limited water supplies. It is critical for us to assess our exposure and vulnerabilities to drought and take the actions necessary to reduce risk through enhanced mitigation and preparedness.

REFERENCES

Boyd, J., *Southwest Farmers Battle Record Drought*, United Press International, May 30, 1996.

British Rainfall Organization, *British Rainfall*, Air Ministry, Meteorological Office, cited in World Meteorological Organization, *Drought and Agriculture*, Technical Note 138, Geneva, Switzerland, 1936.

Brown, R. H., *Historical Geography of the United States*, Harcourt, Brace, and World, New York, 1948.

Dracup, J. A., K. S. Lee, and E. G. Paulson, Jr., On the definition of droughts, *Water Resourc. Res.*, *16*(2), 297–302, 1980.

Federal Emergency Management Agency (FEMA), *National Mitigation Strategy: Partnerships for Building Safer Communities*, FEMA, Washington, DC, 1995.

Hurt, R. D., *An Agricultural and Social History: The Dust Bowl*, Nelson-Hall, Chicago, IL, 1981.

Klemês, V., Drought prediction: A hydrological perspective, in D. A. Wilhite and W. E. Easterling (Eds.), *Planning for Drought: Toward a Reduction of Societal Vulnerability*, Westview, Boulder, CO, 1987, Chapter 7.

Palmer, W. D., *Meteorological Drought*, Research Paper No. 45, U.S. Weather Bureau, Washington, DC, 1965.

Riebsame, W. E., S. A. Changnon, and T. R. Karl, *Drought and Natural Resources Management in the United States*, Westview, Boulder, CO, 1991.

Rosenberg, N. J. (Ed.), *Drought in the Great Plains: Research on Impacts and Strategies*, Water Resources Publications, Littleton, CO, 1980.

Rosenberg, N. J., Adaptations to adversity: Agriculture, climate and the Great Plains of North America, *Great Plains Q.*, *6*, 202–217, 1986.

Sandford, S., Towards a definition of drought, in M. T. Hinchey, (Ed.), *Botswana Drought Symposium*, Botswana Society, Gaborone, Botswana, 1979.

Schneider, S. H. (Ed.), *Encyclopedia of Climate and Weather*, Oxford University Press, New York, 1996.

Stahle, D. W., and M. K. Cleaveland, Texas drought history reconstructed and analyzed from 1698 to 1980, *J. Climate*, *1*, 59–74, 1988.

Stockton, C. W., and D. M. Meko, A long-term history of drought occurrence in the western United States as inferred from tree rings, *Weatherwise*, December 1975. pp. 245–249.

Stockton, C. W., and D. M. Meko, Drought recurrence in the Great Plains as reconstructed from long-term tree-ring records, Journal of Climate and Applied Meteorology. vol. 22, pp. 17–29.

Tannehill, I. R., *Drought: Its Causes and Effects*, Princeton University Press, Princeton, NJ, 1947.

Weakly, H. D., Recurrence of drought in the Great Plains during the last 700 years, *Agric. Eng.*, February 1965, Vol. 46, p. 85.

Wilhite, D. A., Drought, in *Encyclopedia of Earth System Science*, Vol. 2, Academic, San Diego, CA, 1992, pp. 81–92.

Wilhite, D. A., Understanding the phenomenon of drought, *Hydro-Review*, *12*, 136–148, 1993a.

Wilhite, D. A. (Ed.), *Drought Assessment, Management, and Planning: Theory and Case Studies*, Kluwer Academic, Boston, MA, 1993b.

Wilhite, D. A., State actions to mitigate drought: Lessons learned, *J. Am. Water Res. Assoc.*, *33*(5): 961–968, 1997a.

Wilhite, D. A., Improving drought management in the West, Report to the Western Water Policy Review Advisory Commission, 1997b.

Wilhite, D. A., Responding to drought: Common threads from the past, visions for the future, *J. Am. Water Res. Assoc.*, *33*(5): 951–959, 1997c.

Wilhite, D. A. and M. H. Glantz, Understanding the drought phenomenon: The role of definitions, *Water Int.*, *10*, 111–120, 1985.

Wilhite, D. A. (editor), 2000, Draught: A Global Assessment. Routledge Publishers. London, England.

CHAPTER 40

FLOODS ON THE MISSISSIPPI RIVER SYSTEM OF THE UNITED STATES

STANLEY A. CHANGNON

1 INTRODUCTION

For the past 150 years a titanic struggle has been underway between the human occupants of the Mississippi River system and its floods. A flood of some type occurs somewhere in this giant basin each year, and every 2 to 10 years a massive flood encompasses a fourth or more of the 3.2 million km^2 basin. Humans first tried to control the floods with structures: channel straightening, levees, dams, and reservoirs; but after 100 years and the expenditure of billions of dollars, losses to property and lives continued to grow. Efforts since the 1950s to encourage land-use changes in flood-prone areas and the use of flood insurance often have been thwarted by government relief payments to flood victims, plus a continuing human desire to ignore the threat and reside in floodplains. Only 10% of the residents of flooded areas in the massive floods on the Mississippi and Missouri Rivers in 1993 and on the Ohio River in 1997 had flood insurance. Ever growing urban sprawl in the floodplains, intense use of the rivers for shipping, dense surface transportation networks in the floodplains, and major cities and industrial complexes built along major rivers help keep the basin highly vulnerable to today's floods.

The river system brings enormous economic value to the United States, but interests in protecting and enhancing the natural environment of the river are commonly in direct conflict with economic interests, and flood mitigation is often caught in the middle of the debate about how to manage the river system to satisfy all interests. The massive 1993 flood losses and responses, costing $18 billion, brought about needed changes in crop and flood insurance. The huge and costly

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

infrastructure built to control the major rivers for floods and navigation is aging and is beginning to need replacement. This offers an opportunity for more correctly handling flood mitigation to satisfy the complex mix of economic, human, and environmental interests (Shabman, 1994). However, if the past is a predictor of the future, this seems unlikely, but we do know one thing about the future—major record-setting floods will continue to occur.

Thirteen lessons emanating from the recent struggles between humans and floods include:

1. After massive expenditures to control flooding, flooding in the Mississippi River system is still the most costly natural hazard of the region.
2. Floods occurring at various scales, from localized flash floods to enormous basin-scale floods cannot be controlled, although existing control works help to reduce losses.
3. Forecasting and warning of floods have improved, leading to a reduction in lives lost, but people still are killed due to failure to understand the risks or to receive warnings.
4. Surface and river transportation systems suffer major damages and costly delays.
5. Communities suffer costly damages to water and sewage treatment plants and huge costs for postflood cleanup, and many flood-prone communities consciously chose to not construct adequate flood protection systems.
6. Flooding in suburban areas is increasing, due to construction in questionable areas, a lack of control on residential and commercial developments with inadequate stormwater control systems, and a lack of regional plans for managing floodwaters.
7. Millions of individuals continue to risk flood losses by failing to purchase flood insurance, relying on government relief, and simply not appreciating their risk.
8. Environmental impacts of floods are mixed—the worst relate to soil erosion and river pollution by chemicals—but flooding generally helps floodplain ecosystems.
9. Agricultural impacts from floods can be enormous, particularly if the floods occur during the growing season.
10. Government relief for flood victims and communities remains a major response that often acts to discourage doing the right things in floodplains.
11. Rivers will reclaim their floodplains in extreme floods, and it is wise to work with the river and not against it in deciding where and how to build levees and in rebuilding other control structures such as the aging lock-and-dam system.
12. Future use of the floodplains of the Mississippi River system will require a careful balance between controlling for natural variations for navigation and flood protection purposes or for benefits to the ecosystem.

13. In sum, the nation's policy philosophy about flood mitigation must change: The nation needs to move beyond reliance on political responses and solutions to inappropriate uses of floodplains. Individuals must assume responsibility for their locational decisions, not the government, and future government policies must stand firm over time.

To understand the floods of this huge river system, and the impacts the floods create, requires background information about the physical and human setting of the river system. This setting has its roots in the basin's physical formation and configuration and the history of settlement and ensuing development of the basin. Today's impacts of flooding integrate many decades of massive, costly efforts to mitigate flooding in the Mississippi River system.

Physical Setting

The flow of the Mississippi River ranks as the world's third largest behind the Amazon and Congo. The outflow of the river into the Gulf of Mexico averages 173,600 m^3/s and represents 5% of all the freshwater discharged into the oceans of the world (Tarbuck and Lutgens, 1984). The variability of precipitation falling in the basin is large, reflected in a 1-year record low flow of 75,000 m^3/s, and a record 1-year high of 600,000 m^3/s, more than three times the long-term average. The basin occupies all or part of 34 states and 41% of the contiguous United States, sprawling over 3.2 million km^2. Headwaters exist in the Rocky Mountains of the west, the Appalachian Mountains of the east, and the forests of the northern United States. (Fig. 1).

The enormous basin embraces five major climatic regions including humid continental, semiarid steppe, and wet subtropical climates. Four very different physiographic regions are found in the basin including the world's most productive soils and seven very different vegetative regions (White et al., 1979). This diversification in the physical setting greatly affects the type and frequency of floods that occur. Millions of years ago, the river's delta began forming near Cairo, Illinois (Fig. 1), and subsequently advanced 1600 km to the south. New Orleans, the river's major port city, rests where ocean waters existed only 5000 years ago. The enormity of the basin's erosion and sediment transport of the river is shown by the fact that the river deposits 750 million tons of sand, silt, and clay annually into the Gulf of Mexico (Tarbuck and Lutgens, 1984). The sediment load, resulting from the record 1993 flood, led to a discharge of 2.1 billion tons (Bhowmik, 1996).

The basin is so large that it embraces two other enormous river basins—those of the Ohio and Missouri Rivers, and the Mississippi alone is so huge that it is divided hydrologically at St. Louis into the upper and lower basins. Due to climatic differences and the basin's enormity, there has never been a flood that encompassed the entire Mississippi basin. The net effect of the basin's physical situation (size and climatic factors creating floods) is that floods only occur at a given time on one, or infrequently two, of these large basins like the Ohio, the Upper Mississippi, or the

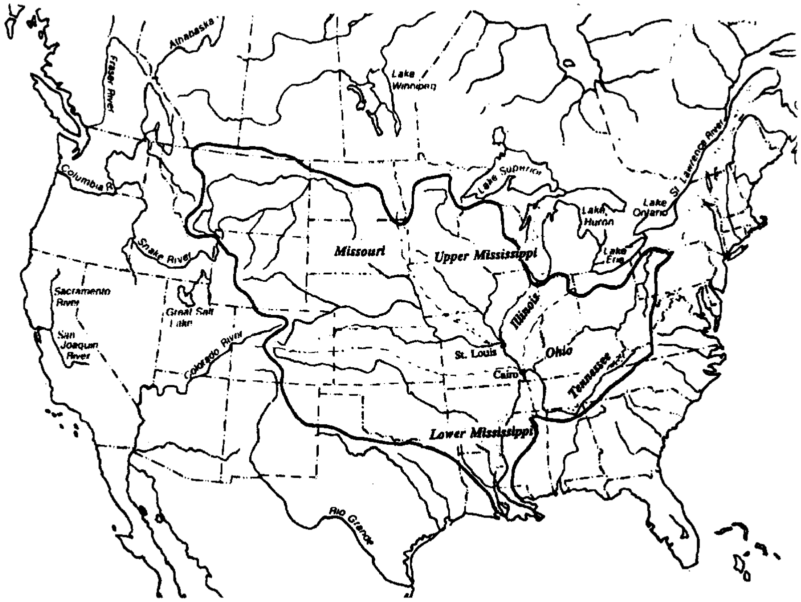


Figure 1 Mississippi River System and its main rivers.

Missouri. The areas most prone to large-scale simultaneous floods are the Ohio River and Lower Mississippi.

Human Setting

Today's human setting is a result of the basin's settlement patterns and the ensuing human development and use of the land and the river system. The original 13 American states claimed lands extending west to the Mississippi River well before any settlers had moved westward from the nation's east coast. As the eighteenth century ended, rapid settlement of the area began by westward movement into the Ohio River basin. The Louisiana Purchase of lands west of the Mississippi River from France for $15 million in 1803 doubled the size of the United States, surely ranking as one of the greatest bargains in history, and encompassed the entire Mississippi basin west of the river including the Missouri, Red, and Arkansas rivers.

This acquisition brought a rapid influx of settlers up the Mississippi River past New Orleans, and the formation of a major port at St. Louis. This invasion and settlement of the American center largely occurred during the first 75 years of the nineteenth century. The basin's estimated population in 1800 was 0.2 million and by 1890 it was 28 million. The Mississippi River and its tributaries were the "avenues" of settlement. They rapidly became the avenues of commerce to handle the flow of goods in and products out. This led to the development of steamboats, and river transport became the first major commercial use of the river, forever establishing navigation as a major priority for the Mississippi River system (Interagency Committee, 1994). In the 1900s, Congress directed the U.S. Army Corps of Engineers to dredge a channel 6 ft deep from the mouth of the river to Minneapolis, and by 1950, a system of 29 locks and dams had been built along the Upper Mississippi from St. Louis to Minneapolis (Keating, 1971). In 1945 Congress authorized development of a 9-foot navigation channel for navigation on the Missouri River from St Louis to Sioux City, Iowa, with construction of six locks and dams (Interagency Committee, 1994).

Initial settlement involved farming in the fertile floodplains of the basin, followed by settlement of the uplands, which were largely prairies. Forests were cut to satisfy demands for wood of growing cities and to access new farmlands, an action that changed the landscape and made it more flood prone and erosion prone. Farming in the humid basins of the Mississippi and Ohio Rivers required extensive drainage works to eliminate the swamplike prairies, and this also changed the basin's water balance and further enhanced the movement of water and flooding. Farming in the more arid High Plains suffered from lack of water and ultimately led to widespread irrigation using river waters and groundwater. Congress passed the Reclamation Act of 1902 to aid irrigation in the West, and by 1990 more than 30,000 km^2 of the Missouri basin were being irrigated, further changing the region's water balance. As farming grew, towns developed along the rivers and the major port cities grew ever larger to handle the commerce of the basin.

By 1880 a settlement pattern involving intensive farming, transportation networks (by then roads and railroads as well as river transport), and cities with industry had emerged. The occurrence of floods within this now well-developed human setting brought chaos—people drowned, crops were washed away, and property destroyed, particularly along the floodplains of the great Mississippi. Action to address the floods of the Mississippi and its tributaries became a major issue for all levels, from local to federal governments.

2 EFFORTS TO CONTROL FLOODING: 1851 TO PRESENT

Levees and Warnings

Extensive floods in 1849 and 1850 (followed by an all-time record flood in 1858) led to action in Washington. Congress established the Delta Survey in 1851 to address the design and construction of works to control floods *and* to aid navigation on the Lower Mississippi River. The river's massive 1858 flood gave the survey team of army engineers a benchmark to work from. The Delta Survey team recommended a "levee-only" policy in 1861, a policy followed well into the twentieth century. The levee-only approach was to be done primarily to protect cities and communities along the river's main course. The U.S. Department of the Army was given the primary responsibility for addressing the problem, and its engineers launched a program to control flooding.

In 1871 Congress directed the secretary of the Army to establish a network of river-stage gages along the Mississippi and Ohio Rivers. The Weather Service, established in 1870 as an arm of the Army's Signal Service, formed in 1873 a River and Flood Service, which began collecting the stream-level values and rainfall amounts (by telegraph) to make flood warnings, issued in Washington and transmitted by telegraph to district offices. These were issued for major flooding developing in the Lower Mississippi River basin, which was then the most flood-prone part of the entire Mississippi basin (Morrill, 1897). Flood prediction remained an ever-improving art based on empirical relationships of past rainfall conditions and river stages until the 1950s, when the science of rainfall forecasting had advanced. Such advances had led to improvements in flood forecasting on the Mississippi and its headwaters.

Most federal attention to flood control in the nineteenth century went to the construction of levees along the Lower Mississippi, then an easily flooded alluvial valley. In 1879 Congress established the Mississippi River Commission to survey the *entire* river system and to develop plans for navigation and flood control on all main river channels, reflecting growing federal responsibility for control of flooding. By this time, privately funded flood damage reduction measures, and mainly levees of varying types, were being built along parts of the Mississippi River. (Fig. 2).

As the nineteenth century ended, a system of major levees was being erected along the Lower Mississippi without any central planning or direction, and flooding continued there. Devastating floods in 1903 and 1912 led to ever more public

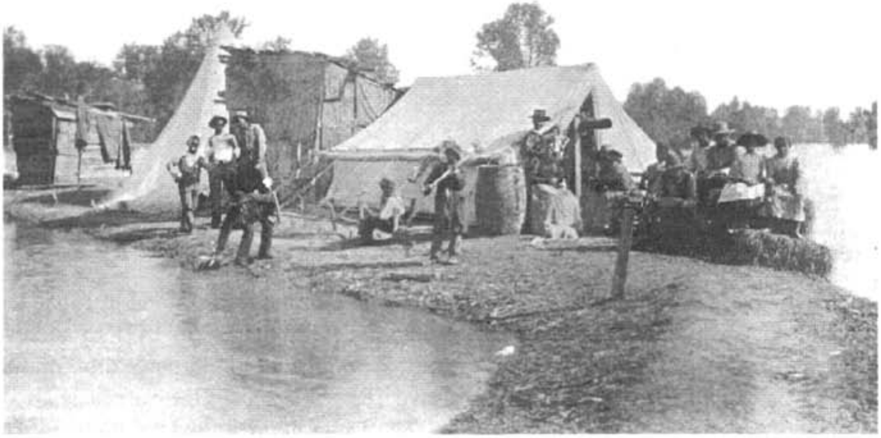


Figure 2 Refugees of the 1893 flood along the Lower Mississippi River. They survived by living on one of the many dirt levees built during the 1866–1890 era. Unfortunately, this scene has been repeated several times during the twentieth century.

pressure for government action against floods. Congress passed a Flood Control Act in 1917, calling for levee construction based on cost sharing with local districts, one of the first government–private sector partnerships.

Levees continued to be built in the basin, and in 1927 the Mississippi River Commission proudly announced that the levee system of the Lower Mississippi was ready to withstand the worst of floods. Two months later nature refuted their claim. An enormous spring flood developed on the Ohio River and spread into the Lower Mississippi River basin. It overwhelmed the massive levee system built during 1870 to 1926. The flood killed 246 persons and drove 600,000 from their homes as it spread over a 200-km wide swath extending from Cairo, Illinois, southward for 1500 km (Keating, 1971), covering large parts of Kentucky, Tennessee, Mississippi, Arkansas, and Louisiana.

Federal Preeminence

The shocking enormity of the 1927 flood should have revealed the fallacy of the levee-only policy, but the failure was blamed on God and the private district levees. This led Washington politicians to enact the Flood Control Act of 1928. It called for control of flooding on the *entire* Lower Mississippi River system. Major floods recurred in 1936 and 1937 (Smith, 1937), and these again led to more Congressional activity, actions that clearly established the federal government as being primarily *responsible for planning and accomplishing flood control across the nation.*

Fortunately, by the 1940s the human learning curve had led to a much better understanding of floods and how to manage floodplains. This included recognition that flood control works such as reservoirs were needed, and that these should serve multiple purposes including flood control, river navigation, water supply, and in later years, recreational needs (Wright, 1996). Hence, between 1936 and 1952, Congress spent $11 billion for flood control projects, primarily designed to store water. The Corps of Engineers built 76 reservoirs in the Upper Mississippi River basin and 49 on the Missouri River basin, and the Bureau of Reclamation built 22 flood control reservoirs in the Missouri basin. The New Deal government committed sizable funds to flood control in the Mississippi River basin and among its efforts was the establishment of the Tennessee Valley Authority in 1933.

Another policy shift in flood mitigation had evolved over time and eventually brought a development of a proper balance in the sharing for flood mitigation activities involving the states, the federal government (in charge), and local entities such as communities and flood control districts. A regional-scale approach to flood control also emerged under the leadership of U.S. Department of Agriculture's (USDA's) Soil Conservation Service in the 1940s. It had three regional components: (1) land treatment at the farm/local level (such as terracing), (2) upstream watersheds with flow retardation and channel stabilization, and (3) the standard downstream flood control measures (levees, pumps to remove water protected by levees, and major reservoirs). Efforts to improve flood mitigation were evolving.

Billions of dollars had been spent between 1851 and 1950 to structurally control flooding on the Mississippi River system, but it had not succeeded in abating flood losses. The words of a well-known river sage, Mark Twain, issued 40 years before flood experts and Congress recognized their mistakes over 100 years, are relevant. In his *Life on the Mississippi* published in 1896, Twain wrote,

> Ten thousand river commissions, with the minds of the world at their back, cannot tame that lawless stream, cannot curb it or confine it, cannot say to it, "Go here," or "Go there," and make it obey; cannot save a shore which it has sentenced; cannot bar its path with an obstruction which it will not tear down, dance over, and laugh at.

Flood mitigation had long been a national goal, and by 1940 had become a federal responsibility. By then flood mitigation embraced a complex of various federal and state agencies, each with a different mission and with each often addressing varying constituents with conflicting views about how to manage the rivers and handle flooding.

These groups included farming interests, agribusiness, river transportation, hydroelectric power generation, irrigation, and recreation interests.

Changing Approaches for Handling the Flood Hazard

National recognition emerged in the 1950s that the structural approach to flood control was inadequate (White, 1958). This led to the development of a new thrust based on floodplain management through altering land use in floodplains and use of flood insurance, or "working with the river." The National Flood Insurance Program was established by Congress in 1968. Emerging environmental concerns in the 1960s led to the National Environmental Protection Act (NEPA) in 1968, bringing environmental quality into the objectives of water and floodplain management. This new "nonstructural" comprehensive approach to mitigating flood losses has evolved over the past 30 years but not yet solved the problems—flood losses continued to grow (National Science Foundation, 1980). Major floods on the Mississippi system occurred in 1965, 1973, and 1982–1983, and flash floods killed 236 in 1 hour at Rapid City, South Dakota, in 1972, and 139 in 2 hours in Colorado in 1976. As a result, many new projects for dealing with flash floods emerged.

Relief payments to flood victims, becoming an alternative to mitigation since 1970, became an ever-increasing way to address flood damages. Special relief payment legislation was issued by Congress after each major flood during the past 20 years. This mixed approach, relief and nonstructural, has essentially replaced the expensive flood control construction program of the 1851 to 1950 period. The enormous relief costs of the 1993 flood finally brought this budget-busting problem to the forefront, leading to badly needed changes in crop insurance and flood insurance programs (Changnon, 1996b).

Government policies relating to the struggle between humans and nature in the Mississippi River system have changed immensely over the past 150 years, but none have managed to solve the flood problem in the Mississippi River system.

3 IMPACTS FROM FLOODING

The failure of 150 years of national policies to solve the problems of flooding in the Mississippi River system was clearly illustrated by the impacts of three recent floods.

- A major summer (June and July) 1993 flood affected the Upper Mississippi and Lower Missouri Rivers (Changnon, 1996a), leading to $18 billion in losses and responses.
- In July 1996 a record-setting 43 cm rainstorm in 24 h occurred in Illinois with a large flash flood covering 15,000 km^2 in a matter of hours, engulfing a third of Chicago's suburban area in the headwaters of the Illinois River, a tributary of the Mississippi (Fig. 1).
- In 1997, a massive early spring flood occurred along the Ohio River, inundating many areas considered flood proof in five states.

Assessment of the impacts from recent flooding on the Mississippi River system drew heavily on events with these three recent, yet different types of floods. Major impacts were found in four broad sectors: (1) economic impacts, (2) environmental effects, (3) impacts to and responses by government at the local, state, and federal levels, and (4) social disruption.

Overview of Floods

The 1993 flood rated physically as the worst on the Upper Mississippi and Missouri Rivers, and the losses and costs of responding to the flood made it the most expensive flood on the Mississippi River system. Major losses included 52 dead, the highest since the floods of 1951, 56,000 homes damaged, 8.5 million farm acres either unplanted or unharvested, crop losses equating to $1.4 billion in corn and soybeans, and more than $1.9 billion losses to transportation systems including $920 million to the barge industry (20% of the year's revenue). The total losses amounted to an estimated $18 billion. Congress ultimately authorized $6.2 billion in aid. Insured crop losses totaled $1.6 billion, and the government paid out $301 million in flood insurance payments. State governments spent an estimated $1 billion in flood-related costs (Changnon, 1996a).

The large flash flood in July 1996 set record flow records on local streams and the Illinois River. It inundated parts of 18 large suburban communities of Chicago, causing enormous damage, disrupting transportation to Chicago, rural crop losses, and killing 5 people. Costs of the damages and responses to this flood exceed $0.6 billion (Changnon et al., 1997).

The late winter/early spring flood of 1997, due to heavy prolonged rains falling on frozen ground, did not set records. It was the most severe flood along the Ohio River since the floods of 1936, a 51-year period during which many thought that the area had become essentially flood proof. There were 24 killed with 83,000 homes damaged, and extensive damages to floodplain properties totaling $1.6 billion.

Economic Impacts

The economic impacts of the floods of 1993, 1996, and 1997 involved losses to individuals in and near flooded communities, to floodplain farmers, and to Midwest businesses and industries. Business losses affected regional sales, agricultural production, utilities, manufacturing, transportation, tourism, and recreation. However, the 1993 and 1997 floods also produced some winners in the agriculture, business, and transportation sectors.

Estimates of losses and costs of responding to the 1993 flood varied from $12 to $15.7 billion, and when the railroad and barge losses of $1.3 billion are added, the grand total is $18.1 billion, making it the nation's second worst weather disaster behind hurricane Andrew. The 1996 flash flood costs were $0.6 billion, and the estimates of costs of the 1997 flood are at $1.6 billion and growing.

The 1993 flood inundated vast amounts of valuable farmland representing about 4% of the Corn Belt's planted acreage. Lands lost to crop production due to the 1993 flooding included 2.5 million acres of corn and 1.97 million acres of soybeans. Agricultural losses amounting to $8.9 billion exceeded the losses of all other sectors. National corn production for 1993 was 31% less than for 1992, largely due to the flood. The effects of the 1997 flood on agriculture were less, since it occurred before planting had begun. Many nonflooded Midwestern farmers in 1993 came out "winners," as the flood caused grain prices to rise.

One of the greatest problems caused by all three floods was the curtailment of transportation. The 1993 flood became an absolute barrier to cross-river train and vehicular traffic, paralyzing transportation along 500 miles of the Mississippi River for 6 weeks. River-based barges were halted for nearly 2 months as about 1000 miles of navigable rivers were ultimately closed to navigation. With losses in excess of $1.9 billion, damages to the region's surface transportation systems were the worst in history. Barge movement along the Ohio River was halted for 5 weeks during the 1997 flood, a loss of $600 million. Revenues lost by navigation interests during the 2-month shutdown in 1993 amounted to $320 million.

The nation's major east–west railroads interconnect the top three rail hubs at Chicago, Kansas City, and St. Louis, but unfortunately cross through the badly flooded 1993 areas of Missouri, Iowa, and Illinois. Major washouts of track as well as bridge closings occurred in all floods and total damages for the railroads amounted to $240 million in 1993, and $38 million in 1996. Nearly 3000 long-distance trains had to be re-routed on longer, circuitous routes around the flood-affected areas. The damages and losses suffered by Midwest railroads ranked the 1993 flood as the worst natural disaster ever experienced by the railroad industry.

The public sector of surface transportation was heavily damaged in all three floods. Approaches to highway bridges were flooded and damaged. Road and highway closings during the floods were mainly concentrated along the major rivers, and in 1993, 3200 miles of roads were closed. The 1997 flood closed 75 highways in Ohio for at least 10 days. Severe erosion and washouts affected state and interstate highways, and hundreds of county roads, and damages amounted to $434 million in 1993, $78 million in 1996, and $185 million in 1997. Numerous floodplain businesses and industries were flooded or their operations were severely curtailed. Facilities were damaged, and production either stopped or slowed down greatly. Severe limitations on the transportation systems interrupted incoming and outgoing raw materials and manufactured products, producing loss of revenue and work stoppages. About 1900 businesses reported closings due to flooding in 1993, and as a result of the closures and commuting problems, 20,000 persons became unemployed.

Environmental Impacts

Many impacts on the environment were difficult to measure, some remain unmeasured, and many are tertiary and will take many more years before they are fully evident. The 1993 flood sizably altered the natural ecosystem of the Upper Mississippi and Missouri Rivers and their floodplains, changing many environmental conditions forever. The 1997 flooding also did enormous environmental damage in the floodplains in Kentucky, Ohio, Indiana, and Illinois. The 1996 flood did only minimal environmental damage with soil erosion and pollution of surface waters being the biggest impacts. Along with the flooding and related excessive erosion came further erosion and extensive silting to the floodplains and their wetlands. Although the two big floods (1993 and 1997) damaged some trees and plants, they generally provided a windfall for most plant and animal species, especially fish populations. Prolonged immersion of the nonfarmed portions of the floodplains had deleterious effects on certain trees. When levee breaks suddenly inundated vast areas, some wildlife already isolated by the flood drowned. Populations of certain insect pests were altered, at least on a 1-year time scale.

With more than 1000 levees failing in the Midwest in 1993, turbid and sediment-laden water moved out of the river into the newly exposed floodplains. Many backwater lakes along the Mississippi and Missouri Rivers had already lost between 70 and 100% of their capacities and lost substantially more volume after receiving sizable quantities of sediment during the 1993 flood. These excessive silts and sand in the floodplains smothered vegetation and compromised large areas of productive farmland.

Floodwaters with high flow rates resulted in much-above-normal amounts of eroded sediments and agricultural chemicals in the rivers, and these impacts from all three floods were sizable on water quality. The daily load of atrazine passing in the Mississippi River near Cairo, Illinois, in July 1993 was 12,000 lb per day, four times higher than during any previous year. A large percentage of the eroded herbicides and nitrates entering Midwestern rivers was carried into the Gulf and had a major impact on the ecosystem of the Gulf shore area. Another water quality problem in 1993, 1996, and 1997 along the rivers was large amounts of raw sewage, bacteria, viruses, and parasites carried by the floodwaters. Health officials were concerned that the organisms in the water could cause hepatitis, cholera, typhoid, or gastrointestinal illnesses; therefore, thousands of persons living and working along the river were inoculated to prevent disease outbreaks and, fortunately, waterborne diseases were minimal in all three floods.

Many parts of the ecosystems in and around the flooded rivers of 1993 and 1997 derived benefits from the floods. Large river–floodplain ecosystems in the river system have adapted to exploit seasonal flooding. A major problem in the 1993 flood related to Zebra mussels, which were inadvertently introduced in 1986 to the Great Lakes. The mussels entered the Illinois River from Lake Michigan via the canal system at Chicago in 1991 to 1992 and established themselves in the Upper Illinois River. These mussels released their larvae as the 1993 flood was occurring, and the floodwaters transported huge numbers of the larvae into the lower Illinois River and downstream into the Mississippi, moving laterally into many floodplain lakes and up many tributaries, and into industrial and municipal treatment plants. The Zebra mussel has prospered in its newly colonized habitats, adding greatly to the cost of water treatment and plant maintenance, jeopardizing the survival of native mollusks, and even altering river food webs by filtering detritus, suspended sediment, and the contaminants associated with these particles. This flood-induced spread of Zebra mussels was truly an "environmental disaster."

Impacts to Government

Government entities at all levels from local to federal levels experienced severe impacts due to the flooding. Many government activities fell within the broad definition of "responses," but many others were more "impacts." In 1993, 532 counties were identified as federal disaster areas; 11 counties in 1996; and in 1997, 79 counties were similarly declared. The federal government ultimately paid $6.2 billion for flood aid, insurance, and loans in 1993, and the total for 1997 is estimated at over $0.4 billion. Certain state agencies were heavily involved in flood and water monitoring, in emergency services, levee repair (National Guard units), water quality assessments, and in measuring the losses, representing a severe impact on state budgets, and the flooded states in 1993 spent an estimated $730 million on aid and rebuilding costs.

Many communities lost their water treatment plants for several weeks, making it extremely difficult and expensive to provide potable water, and many communities had severe or total losses of their sewage treatment plants. Mud-covered, flooded streets and city facilities required costly cleanup efforts, and the net result of the urban problems left many communities broke. Flood-fighting efforts at mid-sized river communities such as Quincy, Illinois, in 1993 cost $0.5 million, and $0.3 million in Aurora, Illinois, in 1996. Several flooded towns in 1993 considered relocation, and five of the badly flooded communities have subsequently relocated to higher ground.

Federal flood policies were impacted. The magnitude and damages of all three floods raised fundamental questions about the nation's floodplain management approach and the utility of the flood insurance program. In all three floods, the percent of those with floodplain insurance and damaged property was 10 percent or less. Excessive levee damages affected various governmental bodies. The levee system along the 1993 flooded rivers included 229 federal levees (39 damaged), 268 nonfederal levees (164 damaged), and 1079 private levees (879 damaged). The rebuilding of these levees has represented substantial costs to the federal government, to state governments, and to numerous local flood protection districts. Severely questioned were the benefits and effects of the development by the Corps of Engineers of the lock-and-dam system and the levee system. The Corps of Engineers calculated that the flood protection works (reservoirs and levees) on the Upper Mississippi had actually prevented an additional $4.9 billion in damages in 1993. Environmentalists countered, arguing that had the floodplains largely been left in their natural state, the two floods would have been of lesser magnitude and the damages due to unwise occupancy of the floodplains would have been negligible.

Social Disruption

The descriptions of the environmental effects, the sizable and pervasive economic impacts, and the complex maze of governmental actions due to the three recent floods all lead to the same obvious conclusion: There were considerable impacts on society in the flooded areas. Fatalities in the floods were a relatively small number considering the magnitude of the floods, reflecting improvements in flood prediction and warning. Fighting the flood was one of the major efforts of the 1993 and 1997 floods. The massive efforts involved thousands of persons residing in the threatened floodplains, volunteers, and hundreds of National Guard troops. Thousands of people were evacuated from their homes along the Mississippi in 1993, along tributaries of the Illinois River in 1996, and along the Ohio in 1997.

Anxiety among flood victims was high for long periods due to the initial fear of being flooded, the actual flooding and damages to personal property, and finally the exhaustive cleanup and restoration process. Loss of residence, or fear of its loss, was a primary cause of stress, along with loss of primary services, including protracted outages of power, water, and sewage treatment in communities and farms along the flooded rivers. Social disruption from the flood is most startling when viewed through the following numbers: of 94,000 persons evacuated from their residences in the summer of 1993, 45,000 were homeless at the end of November 1993, and 3000 were still homeless in June 1994. Furthermore, 61,000 Midwestern homes were seriously damaged, of which 60% were a total loss. The 1996 flash flood led to evacuations of 13,000, and 35,000 homes were flooded. The 1997 flood led to evacuations of 28,000 persons and flooding of 83,000 homes.

4 LESSONS

Assessment of the three recent floods led to the identification of major issues and common lessons about floods and their mitigation. All the issues and lessons learned for the 1993 flood have been defined in detail (Changnon, 1996b).

Floods Exceeding Past Experience and Design Extremes Continue to Occur

These extreme events caused unusual effects on riverine systems, extreme damage to "containment" structures, unexpected social and economic impacts, and assessments of the "cause" of the events came under scrutiny. Many system failures due to the floods were no one's fault—the design values were simply exceeded by conditions never or very infrequently experienced since river records have been kept.

Several scientific and technical actions are needed to improve understanding, mitigation, and response to extreme flooding. They include: (1) development of plans for data collection during and after floods, (2) development and installation of better instruments to measure floods and river flows, (3) development of hydrologic models for floods, and (4) timely collection of flood data before it disappears.

Major Unexpected Impacts Occurred

- *Unique Impacts to All Forms of Transportation* The nation's surface transportation systems, particularly the railroads and highway systems, experienced unusual and extensive damages from these three floods. The barge industry and shippers who depend on commercial navigation should seek improved river forecasting models and flood-monitoring systems. Approaches to many critical highway bridges need to be rebuilt to higher levels.
- *Structural Damage Exceeds Expectations* These floods with record rains and river levels offer lessons and information for engineers and structural experts about how to design structures more effectively to withstand flood extremes and to improve building codes. Current damage estimation techniques are inadequate. Data from the floods should be used to develop better guidelines for estimating flood damage.
- *River–Floodplain Ecosystems were Surprise Beneficiaries* Major floods, regardless of the human alterations in the floodplains, enhance river–floodplain ecosystems.
- *Human Actions Create Major Unexpected Environmental Problems* Human activities have hurt river ecosystems in many ways, and floods facilitate pest invasions and help create environmental disasters. The potential impacts of the nutrients and herbicides swept into the Gulf of Mexico in 1993 and 1997 need monitoring.
- *Unusual and Unplanned Adjustments and Responses Occurred* In extreme events, unexpected major impacts occurred and some existing governmental systems responded quickly and effectively. Ingenuity and resources are important ingredients in responding to extreme flooding.
- *Hopes for Restoring River Habitats in the Aftermath of the 1993 Flood Look Glum* The Corps of Engineers' annual budgets show sizable growth in funds for construction ($803 million in FY96, $1031 million in FY97, and to $1393

million in FY98), whereas funds for the Environmental Management Program of the Upper Mississippi declined from $19.5 million in FY96 to $12 million in FY98. Structural needs continue to overwhelm environmental concerns (Vanderpool, 1997).

Systems for Monitoring and Predicting Flood Conditions Were Inadequate or Failed

Existing systems for flood monitoring and flood forecasting are still inadequate. The National Weather Service needs better quantitative precipitation forecasts for periods 2 to 7 days ahead, and needs to more effectively integrate its new radar network to implement better flash flood warning techniques. The inadequacy of river monitoring equipment in the Mississippi River system calls for major improvements. Basin hydrologic models used for forecasting need revisions. There is a need for a layered geographical information system (GIS) for every river mile to allow better damage estimates as floods develop.

Flood Information Was Often Incomplete, Incorrect, or Not Timely

- *Loss Values* Data on flood conditions and losses were typically poor and generally inaccurate (often on the low side), and estimates remained highly inaccurate for considerable time after the floods. Means for obtaining more accurate near real-time data on conditions and losses should be developed to improve planning for in-flood adjustments and for relief and restoration activities.
- *Forecasts by Government Agencies* The operational hydrologic models used for flood predictions on all time scales need major improvements. Interactions between forecasters and hydrologists need improvement with a clarification of responsibilities. Methods used to estimate regional and national effects of large-scale wet and dry weather conditions on crop yields are inadequate and make sizable errors in growing season flood situations.
- *Confusion over Government Relief* Near real-time estimates of flood losses and predictions of the flood's size, both physically and economically, were underestimated in 1993 and 1997. They reflect the lack of real-time information about the magnitude of the flood and its impacts, plus poor outlooks about the growth of damages. The government should improve its means for acquiring information on impacts and work to remove or clarify overlapping responsibilities between agencies for handling relief aid for problem areas such as home reconstruction and levee rebuilding.
- *Public Understanding about Floods* There is widespread misunderstanding about floods and their frequency. A flood-related educational program would bring rewards in understanding forecasts, warnings, and description of terms used by scientists and engineers, plus clearer recognition of the risks related to living and farming in floodplains. Government officials need to realize how

affected citizens get disaster-related information (largely via TV) and utilize the broadcast media more effectively to disseminate information. The media has become the major source of information to victims and others interested in this form of natural hazard.

Many Approaches to Mitigate Flood Damages Failed But Some Succeeded

Government Flood Mitigation Policies Failed

Past structural and nonstructural approaches to flood mitigation have not worked, and major past efforts to improve U.S. flood policies have not succeeded. Only 10% of those flood damaged had flood insurance. The floods re-enforced the need to make improvements in floodplain use policies and the federal insurance programs. The considerable failure of the levee systems in 1993 and 1997 revealed that not all levees, particularly agricultural protection levees, can be built in a cost-effective manner to withstand floods. However, many past investments in flood control structures had utility.

Floods Produce Benefits

The major theme of these three floods was extensive losses. However, most weather events, including extremes such as droughts and floods, produce winners as well as losers. The 1993, 1996, and 1997 floods were not exceptions. Scientists and engineers benefited from new knowledge about floods, certain environmental problems are receiving needed attention, most aspects of the river–floodplain ecosystem benefited, inadequate federal policies gained public and political awareness and improved policies resulted, damaged often aged or inadequate facilities are being replaced by better facilities and equipment, and many farmers and businesses in nonflood areas benefited financially. The 1993 flood and its uniqueness produced major changes in flood policy including changes in the National Flood Insurance Program Act and the Federal Crop Insurance Program.

5 SUMMARY

Have the lessons taught by recent floods on the Mississippi River system been learned? The new (1994) law relating to flood insurance has not changed coverage purchased in areas flooded in 1996 and 1997. Obviously, floodplain residents continue to rely, as in the past, on "relief" as their "insurance" against floods and other hazards. Changes in the crop insurance laws in 1994 have led to increased purchases with less reliance on relief payments for flooding. Participation by at-risk populations will determine the extent to which future floods (certain to occur) will be damaging. Hopefully, the public will assume more responsibility for their actions.

Will changing government policies relating to reducing federal expenditures and focusing on more responsibility in the states and private sector help or hurt flood mitigation? In an era when cutting back government spending seems to be what the voter sees as prudent fiscal policy, are the cutbacks going to reduce investments in flood mitigation measures that may save inhabitants of the Mississippi River system substantial losses in the longer term? The resolution of this issue depends on public involvement and political wisdom and will. It remains to be seen whether government and the general populace will act on the lessons learned by the recent severe floods.

REFERENCES

Bhowmik, N., Physical effects: A changed landscape, in *The Great Flood of 1993,* Westview, Boulder, CO, 1996, pp. 101–131.

Changnon, S. A., W. C. Ackermann, G. F. White, and L. Ivens, *A Plan for Research on Floods and Their Mitigation in the United States*, Contract Report 302, Illinois State Water Survey, Champaign, IL, 1983.

Changnon, S. A., Losers and winners: A summary of the flood's impacts, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996a, pp. 276–299.

Changnon, S. A., The lessons from the flood, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996b, pp. 300–320.

Changnon, S. A., J. Angel, D. Changnon, F. A. Huff, P. Merzlock, P. Silberberg, and N. Westcott, *The Record Floods of July 1996 in Northeastern Illinois*, Miscellaneous Report 345, Illinois State Water Survey, Champaign, IL, 1997.

Faber, S., and C. Hunt, River management post-1993: The choice is ours, *Water Resour. Update*, *95*, 21–25, 1994.

Interagency Floodplain Management Review Committee, (IFMRC) *Sharing the Challenge: Floodplain Management into the 21st Century*, IFMRC, Washington, DC, 1994.

Keating, B., *The Mighty Mississippi*, National Geographic Society, Washington, DC, 1971.

Morrill, P., *Floods of the Mississippi River*, Bulletin E, Weather Bureau, Department of Agriculture, Washington, DC, 1897.

National Science Foundation (NSF), *A Report on Flood Hazard Mitigation*, NSF, Washington, DC, 1980.

National Weather Service, *The Great Flood of 1993*, National Disaster Survey Report, NOAA, Washington, DC, 1994.

Shabman, L., Responding to the 1993 flood: The restoration option, *Water Resourc. Update*, *95*, 26–30, 1994.

Smith, W. D., *The Flood of January 1937 in the Ohio and Lower Mississippi River Basins*, Illinois Department of Public Works and Buildings, Chicago, IL, 1937.

Tarbuck, E. J., and F. K. Lutgens, *The Earth.* C. E. Merrill, Columbus, OH, 1984.

Vanderpool, G., River restoration in jeopardy, *Mississippi Monitor*, *1*, 1–5, 1997.

White, G. F., *Changes in Urban Occupancy of Flood Plains in the United States*, Research Paper 57, Department of Geography, University of Chicago, Chicago, IL, 1958.

White, C. L., E. J. Foscue, and T. McKnight, *Regional Geography of Anglo-America*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Wright, J. M., Effects of the flood on national policy: Some achievements, major challenges remain, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996, pp 245–275.

CHAPTER 41

DROUGHT IN NORTHWEST AFRICA

WILL SWEARINGEN, ABDELLATIF BENCHERIFA

1 INTRODUCTION

The northwest African countries of Morocco, Algeria, and Tunisia frequently experience drought*. This region, commonly called the Maghreb, is situated between the Mediterranean and the Sahara Desert on the southern margins of midlatitude storm systems. As a result, both the timing and total amounts of rainfall are extremely irregular. Precipitation levels are generally insufficient for reliable or prosperous rain-fed agriculture in most of the region. Reduced rainfall is caused, among other factors, by the cold Canary Current off the region's western shores, which induces atmospheric stability and decreases the potential for rainfall. High-pressure ridges periodically develop offshore during the autumn–spring rainy season, barring access to moisture-bearing storms. If these high-pressure ridges persist for extended periods, drought results.

Drought is the leading natural hazard in the region and occurs frequently in all three countries. For example, during the twentieth century, Morocco has averaged 1 year of agricultural drought every 3 to 4 years. Unfortunately, there is no detectable periodicity. Each of these Maghreb countries experiences roughly the same frequency of drought. However, drought in one country is often not correlated with drought in the other two countries. For example, in 1988, Morocco had the largest cereal harvest in its entire history (a record since surpassed) while Tunisia suffered its worst harvest in over 40 years owing to drought.

*"Drought" as used here refers primarily to agricultural drought and is assessed through the use of cereal production statistics. This research was supported, in part, by grants from the Human Dimensions of Global Change initiative of the National Science Foundation and the Program in Science and Technology Cooperation of the U.S. Agency for International Development.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

Drought has major socioeconomic significance in northwest Africa because rain-fed (nonirrigated) cereal cultivation occupies a predominant place in the region's agriculture. Since pre-Roman times, this region has specialized in production of cereal crops, mainly wheat and barley, though maize and other cereals (including oats, sorghum, millet, rye, and rice) are also cultivated. Cereal crops account for approximately 85% of the region's cropland and are primarily produced by rain-fed methods. Wheat and barley are the mainstays of the national diets in the region and are consumed primarily as bread and cous cous.

Drought in this region sharply reduces both cereal acreage and yields, causing total production to plummet. This poses a food security threat, particularly if drought continues through a second year. Typically, during a drought year, food shortages develop, cereal imports rise dramatically, herds perish or are slaughtered for lack of forage, many farmers temporarily abandon their land and migrate to the cities, and soil erosion and desertification increase. Finally, the affected country's economy suffers a recession.

Good cereal harvests in northwest Africa require adequate rainfall during *both* the planting season (normally from October to December) and subsequent growing season (which extends until harvesting between April and June, depending on the region). Poor harvests or crop failure can result from rainfall shortages during either season. Given the potential for extreme interannual variability in precipitation levels, 400 mm annual average precipitation is normally considered the threshold for viable rain-fed cereal production in northwest Africa (Bencherifa, 1988b). However, the temporal distribution of rainfall is just as critical as the total amount. For example, if the entire winter precipitation falls during a few intense cloudbursts, most will disappear as runoff and be unavailable for crop use. Thus, regardless of the total amount of rain, drought conditions will probably develop.

2 INCREASING VULNERABILITY TO DROUGHT

Since the earliest historical times, drought has been a major hazard in northwest Africa. Historical surveys of drought and other natural calamities have determined that there were 49 major drought-related famines in Morocco during the period from the late ninth century to the early 1900s (Bois, 1957) and at least 26 such episodes in Tunisia from around AD100 to the late 1800s (Bois, 1944).

While the drought hazard has perhaps always existed in northwest Africa, this hazard has been increasing during the present century (Swearingen, 1992, 1994, 1996a). It has been increasing primarily due to two key processes: (1) expansion of cereal cultivation to drought-prone rangeland and (2) reduction of fallow. During the colonial period, these processes were fostered by large-scale land expropriation, by the dislodging of peasants to marginal lands, by a cereal policy offering high crop prices and other incentives, by agricultural mechanization, which facilitated the mining of marginal areas during periods of higher-than-normal rainfall, and by population pressure associated with rapid population growth. Other significant factors during this period include the gradual loss of peasant ability to stockpile grain as insurance against drought and the progressive substitution of wheat for drought-resistant barley. Since independence, populations in northwest Africa have continued to multiply at rapid rates. High population growth rates, along with neglect of cereal production, gradually precipitated a food security crisis by the early 1980s (Swearingen, 1987b, 1996b). To counter this crisis, all three countries have been making concerted efforts to boost their cereal production. Unfortunately, the policies adopted are further promoting cultivation of drought-prone rangeland and reduction of fallow.

The Colonial Period

Algeria became a French colony in 1830, Tunisia a French protectorate in 1881, and Morocco a French and Spanish protectorate in 1912. The colonial period lasted until 1956 in Tunisia and Morocco, and 1962 in Algeria. In all three countries, colonization introduced major changes. The net effect of these changes was a gradual increase in the drought hazard.

Prior to the colonial period, agriculture in Northwest Africa consisted of an extensive system of dry-land cereal cultivation and animal husbandry, with irrigated orchards and gardens surrounding most urban centers and many villages and pastoral nomadism practiced in the desert regions (Bencherifa, 1986, 1988a; Swearingen, 1987a). Most land was communally owned. Rain-fed landholdings were concentrated in higher rainfall areas with above 400 mm rainfall per year. Landholdings were usually dispersed to provide for equity and to help counter the risk of crop failure. Each peasant farmed several dispersed plots. Surplus grain from bountiful harvests was stockpiled to cover crop failures during drought years. Some stockpiling also occurred at the national level during precolonial times. For example, many of the Alawi sultans in Morocco maintained large granaries as a hedge against drought and famine (Meyers, 1981). Fallowing (periodically letting cropland lie idle instead of cultivating it) was widely practiced. Fallowing both replenished soil moisture and helped to restore soil fertility. Low population pressure gave the arable expanses a relatively underutilized appearance. In addition, lower-rainfall areas were used only for seasonal stockraising.

French colonial planners viewed northwest Africa as ideal for large-scale French settlement. In all three countries, colonization dislodged peasants from much of the best land. Europeans acquired roughly 30% of Algeria's arable land (or 2.7 million hectares), nearly 20% of Tunisia's land (or 800,000 hectares), and 15% of Morocco's land (or 1 million hectares).

Exacerbating the effect of European colonization was land concentration by native large landowners. During the colonial period, indigenous landowners allied with the French were able to amass sizable landholdings in all three countries. In Algeria, some 25,000 native Algerians acquired a total of nearly 2.8 million hectares, somewhat over 30% of the country's arable land (Pfeifer, 1985). In Morocco, 7500 Moroccan landowners acquired 1.6 million hectares, or 24% of the arable total (Swearingen, 1987a). And in Tunisia, 7200 Tunisians acquired 630,000 hectares—15% (Sethom, 1985).

Land concentration during the colonial period had two important consequences: First, as land was expropriated, peasants became concentrated on a diminished amount of land. This reduced peasants' ability to let part of their land lie fallow (Bencherifa and Johnson, 1990, 1991). Reduction of fallow significantly increased the potential for drought. The primary purpose of fallowing in semiarid regions like northwest Africa is to allow soil moisture to accumulate (WMO, 1975). Approximately 20 to 25% of the precipitation falling during the fallow rainy season (roughly October to April) is retained in the soil. Thus, fallowing substantially boosts the available water supply for subsequent crop use. In low-rainfall areas, this soil moisture component is often the critical difference between a successful harvest and drought. With the reduction of fallowing, this buffer was lost, and vulnerability to drought increased. In addition, excessive land-use pressure caused soil fertility to decline. The resulting impoverishment of their land made it increasingly difficult for peasants to stockpile grain as a hedge against drought.

Second, large masses of peasants were dislodged to marginal land that was not sufficiently attractive for colonization. The marginal areas were commonly characterized by poor soils, unfavorable slope, or deficient rainfall. Previously, most of this land had been used only for livestock grazing. Once under plow, it became prone to soil erosion and desertification. Unfortunately, it also became more vulnerable to drought (Bencherifa, 1996).

While land concentration was taking place during the colonial period, other significant changes were occurring. Health measures introduced by the French caused native death rates to plunge. Northwest Africa's population expanded dramatically, with roughly a fivefold increase during colonial times. This population explosion (combined with the expropriation of between a third and a half of the arable land by Europeans and indigenous large landowners) intensified pressure on remaining agricultural resources. Fallow was further reduced, peasant landholdings became increasingly fragmented, soil fertility continued to decline, and more marginal land was put under cultivation.

Colonial agricultural policy, per se, also played a major role in deepening northwest Africa's vulnerability to drought. Between roughly 1915 and 1928, colonial authorities in all three countries had a mandate from the *métropole* to substantially boost cereal production for France. The architects of this mandate were convinced that France's *Afrique du Nord* had been a bountiful breadbasket for Rome during classical times, and that France could restore this land to its former productivity (Swearingen, 1987a). Various subsidies and bonuses were offered to encourage cereal cultivation, especially cultivation by mechanized means. High market prices were also offered, particularly for wheat. Agricultural mechanization and high crop prices enabled marginal areas to be profitably cultivated during higher-than-normal rainfall periods. Although Europeans and native large landowners were the primary beneficiaries of the subsidies and bonuses, lucrative crop prices also encouraged peasant farmers to significantly expand their cereal acreage. Cereal acreage in the region increased dramatically.

Contributing to northwest Africa's vulnerability to drought was the fact that the colonial policy favored wheat production over barley. Previously, barley had been the predominant native cereal. However, wheat now became predominant, and consumer tastes changed to prefer this cereal. With the varieties at the time, the critical rainfall limits for barley were some 30% less than those for wheat. In addition, barley ripens and can be harvested significantly earlier than wheat; therefore, it is less vulnerable to the untimely onset of summer drought conditions. In short, by substituting wheat for barley, the colonial wheat policy increased the potential for significant drought impacts.

Since Independence

Since independence, each of the northwest African countries has pursued a different development strategy. Algeria, emerging from a traumatic colonial experience and devastating war of independence in 1962, has attempted to achieve economic independence through a comprehensive program of industrialization. Morocco, since independence in 1956, has emphasized export agriculture, investing heavily in irrigated production of citrus and market vegetables. Tunisia has adopted the most balanced development strategy since its independence in 1956: it has invested in export agriculture and has also encouraged export-led industrialization by multinational firms.

All three countries recovered ownership of colonial landholdings and have engaged in limited land reform. However, much of the former colonial land passed into the hands of more prosperous native landowners. Furthermore, most of the large landholdings acquired by native landowners during the colonial period were never subject to land reform.

For at least two decades following independence, the northwest African countries seriously neglected domestic food production. By the early 1980s, all three countries were experiencing a food security crisis. Key symptoms of this crisis were declining per capita cereal production, alarming, ever-growing levels of cereal imports; heavy foreign indebtedness related to these imports, and massive food subsidy programs.

By the early 1980s, Algeria was importing approximately two-thirds of its cereal supply, Tunisia was importing nearly half, and Morocco was importing over a third (FAO, various years). In each country, a significant percentage of the population was having difficulty meeting daily food needs. The political implications of this crisis became clear by 1981, when Morocco experienced a bloody food-related riot. Similar food-related riots erupted in Morocco and Tunisia in 1984 and in Algeria in 1988.

Since the early to mid-1980s, all three countries have been undertaking major agricultural reforms (Swearingen, 1996b). The overriding objective is to increase dry-land cereal production. Reforms include privatization of the state agricultural sectors to improve efficiency and promotion of modern seed varieties and fertilizers. In terms of drought enhancement, however, the most significant reforms involve changes in crop prices, especially in Morocco, promotion of agricultural mechanization, and a "new lands" policy in Algeria.

Since independence, northwest African governments have maintained tight control over producer prices of basic food crops. Prices for these crops, cereals in particular, were held artificially low until the 1980s. Indeed, for much of this period, crop prices were only about a fourth of what they would have been without government intervention (Cleaver, 1982). Government rationale was that low crop prices would enable them to provide cheap food to their urban populations, helping to keep wages low and thereby assisting industrialization and other urban development initiatives. An ulterior motive behind the cheap food strategy was to help prevent social unrest among the growing ranks of the urban poor. Unfortunately, low crop prices acted as a major disincentive to farmers, creating a vicious spiral of declining production.

Beginning in the late-1970s, fixed producer prices for cereals and other basic food crops were gradually raised. In Algeria and Tunisia, these prices approached world market levels by the mid-1980s, helping to stimulate cereal production and extension of cultivation to previously uncultivated rangeland areas. However, in Morocco, changes in pricing policy were far more dramatic. Here, the government boosted producer prices of barley and wheat to nearly *twice* world market levels. The stimulus effect has been remarkable and has led to a major expansion. Average annual cereal acreage during the 1980 to 1984 period was slightly over 4.4 million hectares. However, during the 1985 to 1989 period, it expanded to 5.2 million hectares—an increase of over 15% (FAO, various years). This increase has come primarily through the extension of cereal cultivation to marginal rangeland and reduction of fallow. Both of these processes are increasing vulnerability to drought.

Government efforts in all three northwest African countries to promote mechanization have facilitated the extension of cereal cultivation to drought-prone rangeland. The tractor and disc plow have colonized large stretches of rangeland in all three countries. In southern Tunisia, for example, roads created by oil exploration crews have enabled mechanized farmers to penetrate regions that previously were accessible only to pastoral nomads. Similar penetration of previously remote grazing lands has also occurred in the other two countries. Some of these new lands normally receive as little as 200 mm of annual rainfall. Their poor soils can sustain cultivation for a few years, as long as higher-than-normal rainfall prevails. However, the return of normal (reduced) rainfall forces their abandonment. Desertification quickly advances in the abandoned areas.

In Algeria, cultivation of marginal lands has actually become official policy. In 1983, Algeria's government passed legislation that established an ambitious homesteading program. The overriding purpose of this program is to encourage Algerian citizens to maximize the agricultural potential of the country through development of public domain land that has not previously been cultivated.

The government views the program as a way to expand the agricultural resource base, increase the food supply, combat peasant exodus to the cities, and counterbalance excessive urban development along the country's northern coast. The goal is to put approximately 800,000 hectares of new land into production. About half of this land will be in the Saharan zone and involves small (less than 3-hectare) irrigated plots. However, the other half, some 400,000 hectares, involves larger dry-land allotments in the country's high plateau region. Virtually all new "cropland" in this region is low-rainfall steppeland suitable only for stockraising. The homesteading program, then, will significantly increase the proportion of Algeria's cropland in drought-prone areas.

The homesteading program, however, is only part of Algeria's current new lands scheme. In 1984, the Algerian government initiated a comprehensive agricultural plan that includes the goal of putting *2 million* hectares of new land into production. Two-fifths of this new land is to come from the homesteading program. The other three-fifths, or 1.2 million hectares, will come from reduction of fallow in the traditional crop rotation system. This major reduction of fallow, for reasons previously discussed, will substantially increase the risk of drought in Algeria.

3 FIELD RESEARCH TO ASSESS LINKAGES BETWEEN HUMAN ACTIVITIES AND DROUGHT

To help assess the linkages between human activities and drought in northwest Africa as well as the region's increasing vulnerability to drought, the authors organized an extensive field research project in Morocco during the early 1990s. A project team led by one of the authors (Bencherifa) intensively interviewed farmers about farming practices and drought-coping strategies. These interviews were conducted in three different regions of Morocco: the Chaouia (a subhumid region), the northeast, usually referred to as Maroc oriental (a semiarid region), and the Chichaoua (an arid region).

The research team surveyed a total of 335 households or production units. The survey consisted of a series of six questionnaire interviews, which were administered orally to the same household units over a period of nearly 2 years, from 1992 to 1994. The interview protocol was intended to capture information about dynamic responses of producers to specific climate conditions, including both drought and abundant rainfall. By coincidence, the survey covered periods of highly variable weather, including drought and higher-than-normal rainfall years.

In all three study areas, it quickly became clear that rainfall variability is well accounted for in production strategies. The historical backgrounds of the communities in these study areas provide, in effect, a reservoir of memories that allows drought to be regarded as a normal rather than exceptional event. Farmers expect periodic drought and plan for it in their production strategies.

However, the survey also revealed that traditional drought-coping strategies have been losing their effectiveness. Namely, the strategies have been weakened by increasing population pressure and more intensive use of agricultural land. In addition, increasing inequalities between producers (a result of the unequal penetration of market-oriented farming practices) have increased the vulnerability of the poorest farming households to the impacts of drought.

From this extensive field survey, the research team was able to make several generalizations, which can be extrapolated to northwest Africa as a whole:

1. *Vulnerability to Drought Is Related to a Variety of Agronomic Factors* Agronomic factors that help determine vulnerability to drought include the

following: (a) The actual time of plowing and planting: These operations need to be keyed to the timing of the first fall rains in the October to November period. Farmers need to make basic decisions about when to plant, which entails risks. Farmers who achieve optimal timing in planting are less likely to be impacted by drought than farmers who plant either too early or too late. (b) The specific crops grown: Barley is the most drought resistant cereal crop—thus its domination in arid and semiarid conditions. Hard and soft wheat are more sensitive to shortfalls of precipitation. (c) The preceding land use: Fallow helps to mitigate the effects of drought because of the accumulation of soil moisture in fallowed fields. (d) The amount and type of labor inputs: Labor inputs allocated for preparation of soils (most importantly, animal traction versus modern machinery) have a major influence on vulnerability to drought. (e) The type of soil: Heavy soils are excellent agronomically when rainfall is above average. However, light soils have advantages during years of below-average rainfall.

2. *Fallow Has a Critical Role within the Agropastoral System* Fallowing is universally recognized by farmers to be a major determinant in increasing crop output. During the fallow year (as previously noted), fields not only accumulate soil moisture but also nitrogen, leading to increased yields when these fields are again cultivated. In addition, fallow is an essential part of livestock production. This is because farmers obtain fodder both from stubble remaining from the previous year's cultivation as well as from weeds that grow on fallowed fields. When fallow disappears from the agropastoral system due to demographic pressure, farmers become more vulnerable to drought. In all three regions, fallow has been steadily decreasing, as is generally true throughout northwest Africa.
3. *Livestock Raising Is a Basic Drought-Coping Strategy* Livestock play a key role in farmer survival strategies. Despite environmental and demographic differences, most of northwest Africa is characterized by a combination of animal herding and cultivation. However, agropastoral systems differ considerably in herd composition. Herds typically vary from cattle, to combined cattle and sheep, to sheep and goats. Differences are due both to the availability of arable land versus rangeland (in part, the result of different levels of population pressure) and the quality of the rangeland. Because of its comparatively high income-generating potential, livestock production is regarded as the key way to maximize farm incomes during years of adequate or abundant rainfall. However, this integration of livestock production and cultivation increasingly is becoming dysfunctional in the case of multiyear droughts. This is because it relies heavily on the use of by-products from cultivation for animal fodder (hay, straw, stubble, and weeds from fields in fallow). Available fodder has largely disappeared from most farms following a single year of drought. Because of population pressure and the conversion of higher-quality rangeland to cultivation, northwest Africa's agropastoral system has become increasingly vulnerable to drought.

4. *Farmers Employ a Traditional Suite of Other Strategies to Cope with Drought* These include the following: (a) Grain and animal fodder are stockpiled using a variety of traditional storage systems, including conical stacks of hay and underground grain storage pits. Stockpiling grain and fodder is an effective way to buffer drought's impacts, particularly if it does not continue for more than a single year. (b) Farmers reduce their herd size to a level that can be sustained through the drought. However, even in severe drought conditions, they attempt to maintain a small herd of breeding stock. This core herd allows a new start once rainy conditions return. (c) Farmers often adopt a relatively mobile, pastoral-nomadic stockraising pattern to seek grazing resources elsewhere if they run out of fodder on their own farms. (d) Farmers take advantage of rainfall whenever it occurs. If the normal cereal crops cannot be planted in fall or early winter because of drought, and if late-season rain occurs, they plant late crops such as chickpeas or lentils.

   During the early 1990s in Morocco, farmers progressively adopted *new drought-coping strategies*, which can be found generally in northwest Africa. These included the following: (a) Farmers almost universally adopted mechanization wherever possible. When farm income did not allow them to purchase their own farm machinery, they hired plowing services. Mechanization of plowing, in particular, has become a general drought-coping strategy because it allows both for rapid planting following the first rains as well as for quick response to rainfall. For example, in case of late spring rain after previously planted crops have failed, modern farm machinery allows for rapid replanting. Mechanization also dramatically increases farm output during favorable rainfall years by allowing farmers to maximize the cultivated area. This increased production can be stockpiled as a hedge against future drought. However, mechanization also has increased the negative impacts of drought. In short, it has helped agricultural production become a "high risks, high rewards" game. (b) Farmers have adopted fertilizer use as a way to increase production during good years to stockpile grain and fodder in preparation for future drought. (c) Farmers have adopted intensive livestock raising in stables as a way to increase farm income. (d) Wherever possible, farmers have attempted to develop irrigation through digging of new wells, use of motor pumping, and use of diversion devices to concentrate runoff to their plots. (e) Farm families increasingly rely on off-farm income to supplement their farming resources. This strategy included temporary migration to urban areas by one or more family members during drought years. In the severe 3-year drought in the Chichaoua in Morocco during the early 1990s, around 80% of a typical family's resources came from outside the farm.

5. *Socioeconomic Impacts of Drought Are Related to Its Duration* The duration of drought is a major determinant of its socioeconomic impacts. One year of drought following a normal rainfall year has far fewer negative impacts at the household level than is commonly assumed. This is because of the effectiveness of traditional drought-coping strategies, including stockpiling of grain

and fodder during higher-than-normal rainfall years. The general calculation among the farmers surveyed is that "good years cover the bad years." Only when drought lasts more than a single year do its effects generally become critical at the household level. When drought lasts more than a single year, stockpiles of grain and fodder become exhausted, throwing household economies into crisis and threatening starvation for both livestock and people.

4 CONCLUSION

A highly vulnerable agricultural system has emerged in northwest Africa owing to historical, demographic, economic, technological, and policy-related factors. In all three countries, increasing population pressure and higher market demand have exerted pressure on local natural resources, resulting in the extension of cultivation to lower-rainfall areas and the reduction of fallowing. Data collected during extensive field research in Morocco suggest that most northwest African farmers can successfully endure only a single year of meteorological drought without significant hardship. If drought continues through a second year or longer, socioeconomic impacts become critical, even devastating.

REFERENCES

AID. Morocco: Country development strategy statement (FYs 1987–1991). Annex C: The agricultural sector in Morocco: A description, unpublished report, United States Agency for International Development, Washington, DC, February 1986.

Bencherifa, A., Agropastoral systems in Morocco. Cultural ecology of tradition and change, Ph.D thesis, Clark University, Worcester, MA, 1986.

Bencherifa, A., Agropastorale Organisations for men im Atlantischen Marokko, *Die Erde, 119*, 1–13, 1988a.

Bencherifa, A., Le Monde rural marocain, in T. Agoumy and A. Bencherifa (Eds.), *La Grande Encyclopédie du Maroc: Géographie Humaine*, GEP, Cremona, Italy, 1988b

Bencherifa, A., Is sedentarization of pastoral nomads causing desertification? The case of the Beni Guil of Eastern Morocco, in W. Swearingen and A. Bencherifa (Eds.), *The North African Environment at Risk*, Westview Boulder, CO, 1996, pp. 117–130.

Bencherifa, A., and D. Johnson, Adaptation and intensification issues in pastoral systems: Observations in Morocco, in J. G. Galaty and D. Johnson (Eds.), *The World of Pastoralism*, Guilford, New York, 1990, pp 394–416.

Bencherifa, A., and D. Johnson, Resource management changes in the Middle Atlas Mountains: From extensive pastoralism to intensive cash production, *Mountain Res. Devel. 3*, 183–194, 1991.

Bois, C., Années de disette, années d'abondance: Sécheresses et pluies en Tunisie de 648 à 1881, *Rev. lEtude Calamite´s, 21*, 3–26, 1944.

Bois, C., Années de disette, années d'abondance: Sécheresses et pluies au Maroc, *Rev. lEtude Calamités, 26~~3~~5* , 33–71, 1957.

Cleaver, K., *The Agricultural Development Experience of Algeria, Morocco and Tunisia: A Comparison of Strategies for Growth*, Staff working paper 552, World Bank, Washington, DC, 1982.

Food and Agriculture Organization (FAO), *Production Yearbook*, FAO, Rome, various years.

Meyers, A.R., Famine relief and imperial policy in early modern Morocco: The political functions of public health, *Am. J. Public Health, 71*, 1266–1273, 1981.

Morocco, *Consommation et Dépenses des Ménages 198485. Premiers Re ′sultats*, Vol. 1: *Rapport de Synthése*, Direction de la Statistique, Rabat, 1988.

Pfeifer, K., *Agrarian Reform under State Capitalism in Algeria*, Westview, Boulder, CO, 1985.

Sethom, H., L'Action des pouvoirs publics sur les paysages et l'économie rurale dans la Tunisie indépendante, in P. R. Baduel et al. (eds), *Etats, Terrotoires et Terroires auMaghreb*, Centre National de la Recherche Scientifique, Paris, France, 1985, pp. 98–113.

Swearingen, W., *Moroccan Mirages: Agrarian Dreams and Deceptions, 19121986* , Princeton University Press, Princeton, 1987a.

Swearingen, W., Morocco's agricultural crisis in I. W. Zartman (Ed.), *The Political Economy of Morocco*, Praeger, New York, 1987b, pp. 159–172.

Swearingen, W., Drought hazard in Morocco, *Geogr. Rev. 82*, 401–412, 1992.

Swearingen, W., Northwest Africa, in M. H. Glantz (Ed.), *Drought Follows the Plow*, Cambridge University Press, Cambridge, 1994, pp. 117–133.

Swearingen, W., Is drought increasing in Northwest Africa? A historical analysis, in W. Swearingen and A. Bencherifa (Eds.), *The North African Environment at Risk*, Westview, Boulder, CO, 1996a, pp. 17–34.

Swearingen, W., Agricultural reform in Northwest Africa: Economic necessity and environmental dilemmas, in D. Vandewalle (Ed.), *North Africa: Development and Reform in a Changing Global Economy*, St Martins', New York, 1996.

World Meteorological Organization (WMO), *Drought and Agriculture*, Technical Note No. 138, WMO, Geneva.

World Bank, *Social Indicators of Development*, Johns Hopkins University Press, Baltimore, 1989.

CHAPTER 42

HURRICANE AS AN EXTREME METEOROLOGICAL EVENT

ROGER A. PIELKE, JR. AND ROGER A. PIELKE, SR.

1 INTRODUCTION: UNDERSTANDING SOCIETAL RESPONSES TO EXTREME WEATHER EVENTS

In the 1970s, many decision makers became increasingly interested in climate because of numerous weather-related impacts around the world. Events that helped to stimulate this interest included the failed Peruvian anchovy harvest in 1972 and 1973, the 1968 to 1973 drought in the African Sahel, a severe winter freeze in 1972 in the Soviet Union, and in 1974 floods, drought, and early frost in the U.S. Midwest. In 1977, winter in the eastern United States was the coldest ever recorded and summer was one of the three hottest in a century. As a consequence of these extreme events and their impacts, decision makers began paying more attention to the relation of weather and climate to human affairs.

Understanding societal responses to weather and climate requires an understanding of the terms *weather* and *climate*. The 1979 World Climate Conference adopted the following definitions of weather and climate:

> *Weather* is associated with the complete state of the atmosphere at a particular instant in time, and with the evolution of this state through the generation, growth and decay of individual disturbances.
>
> *Climate* is the synthesis of weather events over the whole of a period statistically long enough to establish its statistical ensemble properties (mean value, variances, probabilities of extreme events, etc.) and is largely independent of any instantaneous state.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

Climate refers to more than "average weather" (Gibbs, 1987). Climate is, in statistical terminology, the distribution of weather events and their component properties (e.g., rainfall) over some period of time, typically a few months to thousands of years. In general, climate statistics are based on actual (e.g., weather station) or proxy (e.g., ice core) records of weather observations. Such a record of weather events can be used to create a frequency distribution that will have a central tendency, which can be expressed as an average, but it will also have a variance (i.e., spread around an average). Often, variability is more important to decision makers than the average state (Katz and Brown, 1992).

How society thinks about "extreme" weather is, of course, related to what is defined as "normal" weather. What, then, is a "normal" weather event? There are different ways to define normal weather. Of course, it is possible to argue that on planet Earth all weather events are in some sense normal; however, such a definition has little practical utility for decision makers. One way to refine the concept is to define normal weather events as those events that occur within a certain range within a distribution, such as, for instance, all events that fall within one standard deviation of the mean. In practice, historical records of various lengths and reliabilities have been collected around the world for temperature, precipitation, storm events, and others. When data is available, such a statistical definition lends itself to equating normal weather with "expected" weather, where expectations are set according to the amount of the distribution defined as normal. For example, about 68% of all events fall within one standard deviation of the mean of a bell-shaped distribution.

A change in the statistical distribution of a weather variable—such as that associated with a change in climate—is troubling because decision makers may no longer expect that the future will resemble the past. For the insurance industry, as well as other decision makers who rely on actuarial information, such a possibility of a changing climate is particularly troubling. A climate change is thus a variation or change in the shape or location (e.g., mean) of a distribution of discrete events (Katz, 1993).

"Extreme" weather events can simply be defined as those not normal, however normal is chosen to be defined. For instance, if normal weather events are those that occur within 2 standard deviations of the mean, then about 5% of all events will be classified as extreme.

While it is possible to classify hurricanes as either "normal" or "extreme" in this manner, the simple fact is that for most communities any landfalling hurricane would qualify as an extreme event because of their rarity at particular locations along the coast.

From the standpoint of those human activities sensitive to hurricane impacts, it is often the case that decisions are made and decision processes established based on some set of expectations about what future weather or climate will be like. Building codes, land-use regulations, insurance rates, disaster contingency funds are each an example of decisions that are dependent upon an expectation of the frequency and magnitude of future normal and extreme events.

In short, decision makers typically establish policies based upon an expectation of normal weather. Yet for most coastal communities normal weather has historically (or at least over the time of a human memory) meant no hurricanes! Consequently, people are often surprised when a hurricane does strike and then overwhelms response capabilities. Because decision makers do not always consider the possibility of extreme weather, when such events occur, they often reveal society's vulnerabilities and sometimes lead to human disaster. A fundamental challenge facing society is to incorporate information about weather and climate risks into decision making in order to take advantage of normal weather and to prepare for the extreme. The degree to which society exploits normal weather and reduces its vulnerabilities to extreme weather is a function of how society organizes itself in the face of what is known about various typical and extreme weather events. The challenge is made more difficult by variability at all measurable time scales in the underlying climate, and hence in the frequency, magnitude, and location of various weather events. And, of course, decisions that have a weather or climate component also are laden with all of the political, practical, and social factors that influence policy.

2 HURRICANES DEFINED

One of the most powerful natural phenomena on the face of Earth, the hurricane is a member of a broader class of phenomena called cyclones.* The term *cyclone* refers to any weather system that circulates in a counterclockwise direction in the Northern Hemisphere and in a clockwise direction in the Southern Hemisphere. "Tropical cyclones" typically form over ocean waters of the tropics. The tropics are the area on Earth's surface between the Tropic of Capricorn and the Tropic of Cancer, 23° 27″ south and north of the equator, respectively. Extratropical cyclones, for comparison, form as a result of the temperature contrast between the colder air at higher latitudes and warmer air closer to the equator. Extratropical storms form over both the ocean and land.

Tropical cyclones have been given different names depending on their region of origin. In the western north Pacific, they are called typhoons, while in the Bay of Bengal they are referred to as severe cyclonic storms of hurricane intensity. In the Atlantic, Gulf of Mexico, Caribbean, and Pacific north of the equator and east of the international dateline they are hurricanes. Evidence of tropical cyclones has been documented in a variety of other geographic locations including Europe and North Africa at earlier geologic times (Ager, 1993). Figure 1 shows the tracks of all tropical cyclones with winds greater than 39 mph for the 10-year period 1979 to 1988.

The meteorological community uses a number of terms to classify the various stages in the life cycle of tropical cyclones. The following are definitions of tropical cyclones used in the Atlantic Ocean basin (Pielke and Pielke, 1997):

*This chapter considers hurricanes as an extreme meteorological event. It first discusses the physical aspects of hurricanes, including their development and impacts on ocean and land. It then overviews societal impacts.

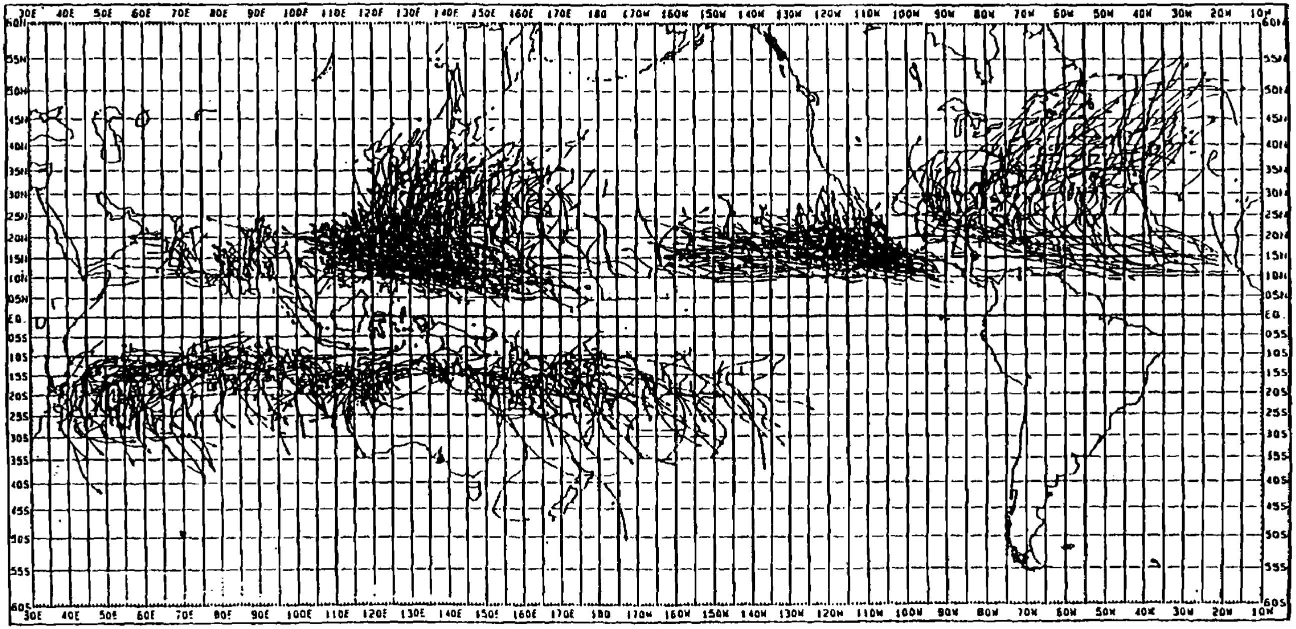


Figure 1 Tracks of all tropical cyclones with winds greater than 39 mph for a 10-year period (Neumann, 1993).

Tropical low A surface low-pressure system in the tropical latitudes.

Tropical disturbance A tropical low and an associated cluster of thunderstorms that has, at most, only a weak surface wind circulation.

Tropical depression A tropical low with a wind circulation of sustained 1-min surface winds of less than 34 knots (kt) [39 miles per hour (mph), 18 meters per second [m/s] circulating around the center of the low]. [A knot (i.e., a nautical mile per hour) equals about 1.15 mph. A nautical mile is the length of 1 min of arc of latitude.]

Tropical storm A tropical cyclone with maximum sustained surface winds of 34 to less than 64 kt (39 to 74 mph, 18 to 33 m/s).

Hurricane A tropical cyclone with maximum sustained surface winds of 64 kt (74 mph, 33 m/s) or greater. (In the Pacific Ocean west of the international date line, hurricanes are called typhoons. They are the same phenomenon.)

3 HURRICANES IN NORTH AMERICAN HISTORY

The word *hurricane* derives from the Spanish *huracán*, itself derived from the dialects of indigenous peoples of the Caribbean and Latin America (Dunn and Miller, 1964). 'Hunraken' was the name of the Mayan storm god, and 'Huraken' was the god of thunder and lightning for the Quiche of southern Guatemala (Henry et al., 1994). The Tainos and Caribe tribes of the Caribbean called their God of Evil by the name Huracan. Other indigenous dialects included words such as *aracan*, *urican*, and *hurivanvucan* to refer to "Big Wind." The deification of the hurricane and the connection of indigenous referents with evil and violence is an indication that hurricanes had a significant impact on the lives of many peoples of the Caribbean and Latin America.

The historical record of documented hurricane events begins with the European conquest of North America. Columbus, in his four voyages to North America, experienced direct contact with an Atlantic hurricane only in his fourth voyage. Meteorological historian David Ludlam notes that Columbus' good fortune in his first voyage leads one to wonder "what the course of history in the West Indies might have been if, in the autumn of 1492, a full-blown tropical storm had dashed the frail craft of the Admiral's fleet to the bottom of the sea or flung them shipwreck on some tiny cay" (Ludlam, 1963, p. 1). Others did not experience such good fortune. Shakespeare's play, *The Tempest*, was loosely based on reports of a 1609 hurricane near Bermuda that sunk the vessel *Sea Venture* and stranded the passengers, including John Rolf, future husband of Pocahontas, on the island for 10 months. This storm's movement was among the first successfully anticipated by the colonists. During the course of the storm's trek through the Caribbean, a skipper in the Royal Navy cautioned the British fleet to move out of the storm's path, based on his experience with the movement of past hurricanes. During the 1700s and 1800s numerous coastal locations were struck by severe hurricanes. Charleston (South Carolina), New Orleans (Louisiana), and Boston (Massachusetts) were particularly hard hit a number of times. In 1772 in the West Indies, teenaged Alexander Hamilton wrote about a hurricane's impact for a local newspaper. His writing caught the attention of the local gentry who then raised money to send him to the mainland colonies to further his education, thus setting the stage for his political career.

Tropical storms were once named after the particular "saint's day" that fell nearest the hurricane event (Tannehill, 1952). For instance, "Hurricane Santa Ana" hit Puerto Rico on 26 July 1825 (see Rodriguez, 1997). Today, tropical cyclones are "named" when they reach tropical storm strength. According to one explanation, this practice dates to the 1950s, following the publication of George R. Stewart's *Storm*, a book that featured a forecaster who named storms (Williams, 1992). Another explanation has the origin of the hurricane naming convention beginning with a military radio operator who, during World War II, ended each hurricane warning singing "Every little breeze seems to whisper Louise," prompting the naming of a particular hurricane Louise (Henry et al., 1994). Whatever the origin, the practice caught on because it proved useful in identifying different storms that existed simultaneously. The personification of the extreme event was also found to be a valuable practice by the various user communities. Until 1979, tropical storms were given only women's names in English. In 1979 forecasters began to use men's, French, and Spanish names as well. The repeating, 6-year list of names assigned to tropical cyclones in the Atlantic was put together by the World Meteorological Organization. It can be found at the National Hurricane Center's website at *http://www.nhc.noaa.gov/names.html.* Hurricanes that cause significant damage or are particularly memorable, such as Andrew (1992), Camille (1969), or Gilbert (1988), are retired and those names are not used again. Table 1 lists retired hurricanes through 1995 and notes death and damages associated with each.

4 GEOGRAPHIC AND SEASONAL DISTRIBUTION: ORIGIN

Typically, in the Atlantic Ocean basin tropical storms and hurricanes develop over warm water between around 10°N to 35°N, generally, during the summer and fall. During an average year about 16 tropical cyclones develop in the eastern Pacific and approximately 10 in the Atlantic including the Gulf of Mexico and Caribbean Sea (Neumann, 1993). During the period of record, tropical cyclones fail to develop south of the equator in the Western Hemisphere east of 130 W because of one or more of the following factors: the relatively cold ocean temperature, typically strong winds in the upper troposphere, or the absence of an initiation area for tropical low-pressure systems with an associated cluster of thunderstorms (Gray, 1968).* Elsewhere these storms develop in the Indian Ocean, western Pacific, and eastern Pacific

*McAdie and Rappaport (1991), however, discussed the formation of a weak tropical cyclone in the south Atlantic west of tropical Africa in 1991.

TABLE 1 "Retired" Atlantic Hurricane Names through 1994

Year	Name	Location	U.S. Costs (1990$) and Total Casualties, etc.
1954	Carol	Louisiana, Mississippi, and Alabama	$2.37 billion, 60 deaths
1954	Hazel	Antilles, North and South Carolina	$144 billion, 1000 deaths
1955	Connie	North Carolina	25 deaths
1955	Diane	Mid-Atlantic and Northeast U.S.	$4.20 billion, 184 deaths
1955	Ione	North Carolina	$444 million
1955	Janet	Lesser Antilles, Belize, and Mexico	538 deaths
1957	Audrey	Louisiana and North Texas	$696 million, 550 deaths
1960	Donna	Bahamas, Florida, and eastern U.S.	$1.82 billion, 364 deaths
1961	Carla	Texas	$1.93 billion, 46 deaths
1963	Flora	Haiti and Cuba	8000 deaths
1964	Cleo	Lesser Antilles, Haiti, Cuba, southeast Florida	$595 million, 213 deaths
1964	Dora	Northeast Florida	$1.16 billion
1964	Hilda	Louisiana	$579 million, 304 deaths
1965	Betsy	Bahamas, southeast Florida, southeast Louisiana	$6.46 billion, 75 deaths
1966	Inez	Lesser Antilles, Hispaniola, Cuba, Florida Keys, Mexico	1000 deaths
1967	Beulah	Antilles, Mexico, South Texas	$844 million; most tornadoes, 115, ever associated with a hurricane
1969	Camille	Louisiana, Mississippi, and Alabama	$5.24 billion, 256 deaths
1970	Celia	South Texas	$1.56 billion
1972	Agnes	Florida, northeast U.S.	$5.24 billion, 122 deaths
1975	Eloise	Antilles, northwest Florida, and Alabama	$1.08 billion
1979	David	Lesser Antilles, Hispaniola, Florida, and eastern U.S.	$487 million, 2000 deaths
1988	Joan	Curacao, Venezuela, Columbia, and Nicaragua	216 deaths; crossed into Pacific and was renamed Miriam
1989	Hugo	Antilles and South Carolina	$7.16 billion, 56 deaths
1990	Diana	Mexico	96 deaths
1990	Klaus	Martinique	
1991	Bob	North Carolina and northeast U.S.	$1.5 billion
1992	Andrew	Bahamas, South Florida, and Louisiana	> $25 billion
1995	Luis	Leeward Islands	$2.5 billion, 16 deaths
1995	Marilyn	Virgin Islands	$1.5 billion, 8 deaths
1995	Opal	Mexico, Florida	$3 billion, 59 deaths
1995	Roxanne	Mexico	$1.5 billion, 14 deaths

After Pielke and Pielke (1997).

TABLE 2 Saffir/Simpson Hurricane Scale

	Central Pressure		Winds	Surge	
Category	(mbars)	(inches)	(mph)	(ft)	Damage
1	⩾980	⩾28.94	74–95	4–5	Minimal
2	965–979	28.50–28.91	96–110	6–8	Moderate
3	945–964	27.91–28.47	111–130	9–12	Extensive
4	920–944	27.17–27.88	131–155	13–18	Extreme
5	<920	<27.17	>155	>18	Catastrophic

See Pielke and Pielke (1997, p. 17).

north of the equator (Fig. 1). The western north Pacific is the most active area with an annual average of more than 26 tropical cyclones. Globally, there are about 84 tropical cyclones each year with an annual average of 45 that reach hurricane strength (Neumann, 1993).

Hurricanes are classified by their damage potential according to a scale developed in the 1970s by Robert Simpson, a meteorologist and then-director of the National Hurricane Center, and Herbert Saffir, a consulting engineer in Dade County, Florida (Simpson and Riehl, 1981). The Saffir/Simpson scale was developed by the National Weather Service to give public officials information on the magnitude of a storm in progress and is now widely used by producers and users of hurricane forecasts. The scale has five categories, with category 1 representing the least intense hurricane and category 5 the most intense. Table 2 shows the Saffir/Simpson scale and the corresponding criteria for classification.

5 HURRICANE IMPACTS ON OCEAN AND LAND

When a hurricane forms, it poses a significant danger to society. The importance and danger of tropical cyclones differ between land and water. Over the oceans, the human activities and assets at risk are primarily oil rigs, shipping, and air traffic. On land, particularly along the coast, cities, towns, and industrial activities become threatened. Hurricanes also have ecological and geological impacts.

Ocean Impacts

Winds of hurricane speed over the ocean can create monstrous waves. For example, in 1995, the cruise ship *Queen Elizabeth II* was rocked by a 70-ft (21-m) wave caused by distant hurricane Luis. The sea near a hurricane is chaotic, and an extreme hazard to shipping can occur in response to wave motion moving in many directions.

For comparison, strong winds, of course, also occur in winter storms over the open ocean. The risk to shipping and other activities from wave action, however, is generally less serious in such storms for two reasons. First, the wind blows primarily in one direction in a given sector of a winter storm. Hence the waves move in concert with the wind. A ship can thus orient itself to minimize the effect of the waves. In a hurricane, winds change direction rapidly around the eye. The result is a chaotic sea with swells and waves propagating in a myriad of directions. A ship cannot simply steer into the running sea to reduce its risk since there is no one direction from which the waves come. Large waves also superimpose on top of each other, producing enormous swells.

Land Impacts at the Coast and a Short Distance Inland

At the coast, the major impacts of either a landfalling hurricane or one paralleling the coast are:

- Storm surge
- Winds
- Rainfall
- Tornadoes

Of these weather features, the storm surge has accounted for over 90% of the deaths in a hurricane. In recent years, and particularly in the aftermath of hurricane Andrew, more attention has been paid to the effects of hurricane winds.

Storm Surge

"Storm surge" refers to a rapid rise of sea level that occurs as a storm approaches a coastline. This is in addition to changes in variations in sea level due to tides. Thus, a storm surge causes greatest inundation at high tide. A very strong hurricane may produce a storm surge of 20 ft (6 m), of which about 3 ft (1 m) is due to the lower atmospheric pressure at the center of a hurricane. The remaining storm surge is due to: (i) the piling up of water at the coast, generated by the strong onshore winds and (ii) a decreased ocean depth near the coast, which steepens the surge. A common misconception is that the lower pressure at the center of a storm is the primary cause of the storm surge.

At landfall, storm surge is highest in the front right quadrant of a westward-moving tropical cyclone (in the Northern Hemisphere), where the onshore winds are the strongest. It is also large where ocean bottom bathymetry focuses the wave energy (e.g., as in a narrowing embayment). Peak storm surge from a landfalling cyclone increases with greater wind speeds and the areal extent of the storm's maximum winds, out to about 30 miles (48 km).

Storm surge also occurs when a storm parallels the coast without making landfall. The storm surge will precede the passage of the storm's center when winds blow onshore preceding passage of the eye. Similarly, the surge will lag the storm's center when the hurricane is moving such that onshore winds follow the passage of the eye.

Offshore winds that are associated with a storm can produce a negative surge, as the sea level is lowered by the strong winds blowing out from the coast.

Storm surge is estimated to generally diminish in depth by 1 to 2 ft (0.3 to 0.6 m) for every mile (1.6 km) that it moves inland. Even if the inland elevation were only 4 to 6 ft (1.2 to 1.8 m) above mean sea level, a storm surge of 20 ft (6 m) might typically reach no more than 7 to 10 miles (11 to 16 km) inland. Thus, the most destructive effect of the storm surge hazard is on beaches and offshore islands.

Storm Surge Hazards. A storm surge can be deadly. In 1900, up to 12,000 deaths occurred in Galveston, Texas, primarily as a result of the storm surge that was associated with a Gulf of Mexico hurricane. In 1957, a storm surge was the major cause of death for 390 people in Louisiana. The storm surge, associated with hurricane Audrey, was over 12 ft (3.5 m) in depth and extended as far inland as 25 miles (40 km) in this particularly low-lying region. In September 1928, the waters of Lake Okeechobee, FL driven by hurricane winds, overflowed the banks of the lake and were the main cause of more than 1800 deaths.

Areas to be evacuated due to storm surge in the case of hurricane landfall are determined through a model developed by the National Weather Service (NWS) called SLOSH (sea, lake, and overland surges from hurricanes; Jarvinen and Lawrence, 1985). The SLOSH model is used to define flood-prone areas in 31 "SLOSH basins" along the U.S. Gulf of Mexico and Atlantic coasts (Fig. 2). Determination of storm surge vulnerabilities is the result of an interagency and intergovernmental process funded by the National Oceanic and Atmospheric Administration (NOAA), the Federal Emergency Management Agency (FEMA), Army Corps of Engineers, and various state and local governments (BTFFDR, 1995). From development through application the SLOSH process for a particular location takes about 2 years. Because coastlines are constantly changing due to human and natural forces, the SLOSH process is an ongoing challenge.

Winds

The strong winds of a hurricane can produce considerable structural damage and risk to life from flying debris, even inland from the coast. The damage caused by hurricane Andrew was predominantly due to wind. Although winds reduce after landfall, as the central pressure increases, and the intensity of the storm lessens, destructive winds can still occur far inland.

The damage from winds is proportional to the energy of the airflow, i.e., to the velocity squared; thus, a wind of 100 mph is four times as effective at causing damage as a wind of 50 mph. Maximum gusts, of course, are even stronger than reported sustained winds (which are measured in the United States by averaging wind speed over 1 min). In a hurricane over the open ocean at about 36 ft (11 m) a gust averaged over 2 s is generally about 25% greater than the 1 min average. For flat grassland, the 2-s speed is around 35% larger, while in woods or cities, this measure of gust speeds is 65% greater. Thus a 1 min average wind of 100 mph would be expected to have gusts to 125 mph over the ocean and 165 mph over a forest.

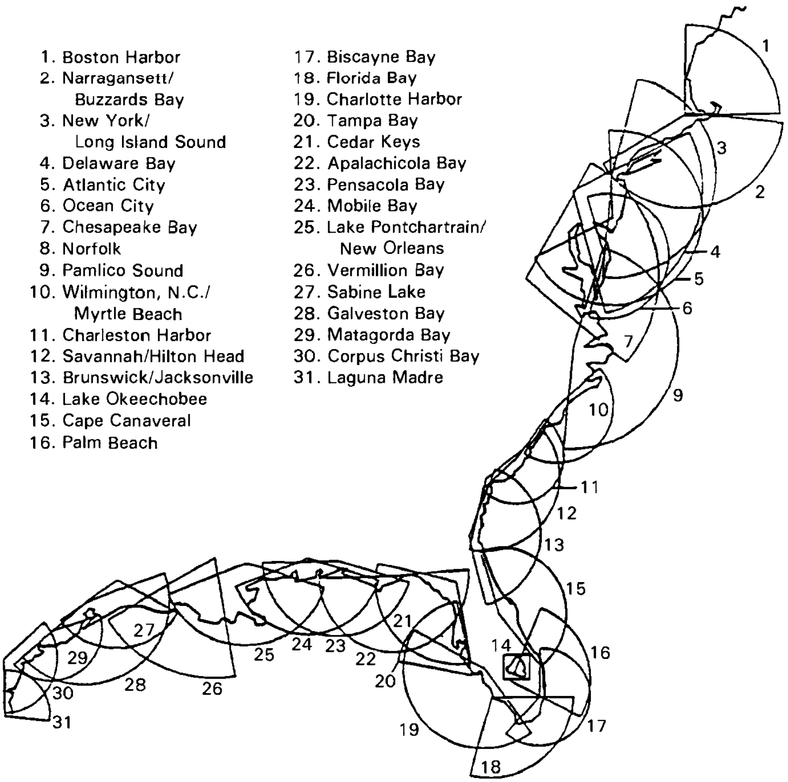


Figure 2 The 31 SLOSH basins along the U.S. Gulf and Atlantic coasts.

Rainfall. Rainfall from hurricanes is beneficial to agriculture, such as the rains from hurricane Dolly (1995) in southern Texas and northeastern Mexico that relieved a drought (Rippey, 1997; cf. Sugg, 1967). Even relatively weak tropical-like disturbances can result in extreme rainfall, as seen, for example, over coastal Texas in September 1979 in which upwards of 19 inches (483 mm) of rain inundated the area over a period of several days (Bosart, 1984). Occasionally, for reasons not completely understood, rainfall is light in the vicinity of hurricanes. Hurricane Inez in 1966, for instance, resulted in only a few drops of rain in Miami for several hours when the center was south and south-southwest of Miami and at its closest point to the city. At the time, Miami was under the storm and, normally, torrential rains would have been expected. As a result of the absence of rain, the strong winds blew salt spray many miles inland, causing severe damage to vegetation from salt accumulation. Homestead Air Force Base, south of Miami and closer to the path traveled by the hurricane's center, received only 0.62 inches (15.7 mm) of rain during the entire storm.

Tornadoes. Tornadoes are also a threat from tropical cyclones. Much of the damage of Andrew was associated with tornadic vortices whose wind speeds were added onto the large-scale hurricane winds (Black and Wakimoto, 1994). These rapidly rotating small-scale vortices are spawned in squalls, usually in the front right quadrant of the storm with respect to the storm's track.

Wind damage and tornadoes also can occur well inland associated with tropical cyclones. In 1959, hurricane Gracie caused 12 deaths in central Virginia 24 h after landfall on the South Carolina coast. Hurricane Hugo in 1989 caused significant damage in Charlotte, North Carolina, after landfall.

Inland Impacts

Inland, away from the coast, the largest threat to life and property occurs as a result of flash flooding and large-scale riverine flooding from excessive rainfall. Particularly dangerous are tropical cyclones whose rainfall is initially light and benign after landfall only to erupt a couple of days later into torrential downpours when the environment becomes favorable for precipitation of the large quantities of tropical moisture that have moved inland with the storm.

A particularly extreme example of such a system is hurricane Camille of 1969. After killing 139 people along the Gulf coast on August 17, the storm rapidly weakened after moving inland across Mississippi, into Tennessee and Kentucky. There was relatively little concern expressed by the National Weather Service and certainly no hint of the tragedy that was to happen on the night of August 19, 1969, in central Virginia. The 24-h and 12-h precipitation forecasts for the area, for example, indicated that only slightly more than 2 inches (50 mm) were expected. In fact, a deluge occurred in one part of Virginia as the remnants of Camille began to rejuvenate through interaction with a cold front and when the associated moist tropical air was lifted by the mountains. The rainfall of almost 30 inches (760 mm) in 6 h liquefied soils on the mountainous slopes and flooded drainage basins, burying and drowning 109 individuals. As a result of this tragedy, a radar site was installed in southern Virginia. One of the justifications of the new National Weather Service U.S. Doppler radar network (the WSR-D-88 system) is to detect heavy rainfall events.

Such excessive rains well inland from landfalling tropical cyclones should be expected occasionally as occurred over Georgia associated with tropical storm Alberto in 1994. The environment of a storm is a localized region of the atmosphere that is enriched with water vapor, well in excess of even the average tropical environment. After landfall, this rich reservoir of moisture moves inland and can be copiously precipitated when it is lifted through a mechanism such as a mountain barrier and/or ascent over a weather front. Hurricane Agnes in 1972, for instance, produced enormous rainfalls over large areas of the middle Atlantic states because of strong large-scale atmospheric lifting and the movement of the moist air up and over the Appalachian mountains, resulting in disaster.

Even snowfall has been reported to be associated with the inland portion of a hurricane circulation. In 1963, hurricane Ginny left more than 14 inches (36 cm) of snow in northern Maine as the hurricane moved into Nova Scotia with winds of around 100 mph (45 m/s).

Societal Impacts

When they strike the U.S. coast, hurricanes cost lives and dollars and disrupt communities. Category 3, 4, and 5 storms—intense hurricanes—are responsible for more than 80% of hurricane-related damages. Loss of life, however, occurs from storms of various intensities. Due largely to better warning systems, hurricane-related loss of life has decreased dramatically in the twentieth century (NRC 1989). Yet, in spite of reduced hurricane-related casualties "-a large death toll in a U.S. hurricane is still possible. The decreased death totals in recent years may be as much a result of lack of major hurricanes striking the most vulnerable areas as they are of any fail-proof forecasting, warning, and observing systems" (Hebert et al., 1993, p. 14).

While loss of life has decreased, the economic and social costs of hurricanes are large and rising. A rough calculation shows that annual losses to hurricanes have been in the billions of dollars. In the United States alone, after adjusting for inflation, tropical cyclones were responsible for an annual average of $1.6 billion for the period 1950 to 1989, $2.2 billion over 1950 to 1995, and $6.2 billion over 1989 to 1995 (Hebert et al., 1996). For a comparison, China suffered an average $1.3 billion (unadjusted) in damages related to typhoons over the period 1986 to 1994 (World Meteorological Organization, various years). Significant tropical cyclone damages are also experienced by other countries including those in East Asia (including Japan, China, and Korea) and Southeast Asia, those along the Indian Ocean (including Australia, Madagascar, and the southeast African coast), islands of the Caribbean, and in Central America (including Mexico). While a full accounting of global damages has yet to be documented and made accessible, it is surely in the tens of billions of dollars annually. Other estimates range to $15 billion annually (e.g., Southern, 1992).

Experts have estimated that tropical cyclones result in approximately 12,000 to 23,000 deaths worldwide (Southern, 1992; Smith, 1992; Bryant, 1991). Tropical cyclones have been responsible for a number of the largest losses of life due to a natural disaster. For instance, in April 1991, a cyclone made landfall in Bangladesh resulting in the loss of more than 140,000 lives and disrupting more than 10 million people (and leading to $2 billion in damages; Southern, 1992). A similar storm resulted in the loss of more than 250,000 lives November 1970. China, India, Thailand, and the Philippines have also seen loss of life in the thousands in recent years.

While the hurricane threat to the U.S. Atlantic and Gulf coasts has been widely recognized, it has only been in recent years, following hurricane Andrew, that many public and private decision makers have sought to better understand the economic and social magnitude of the threat.

One study has sought to "normalize" U.S. hurricane damages to assess the impact that past storms would have had in 1995 (Pielke and Landsea, 1997). The study adjusted past damages to account for changes in population, inflation, and wealth. The study found a total of $366 billion in losses over the period 1925 to 1995, or about $5 billion annually. Interestingly, the normalized data show a trend of *decreasing* losses from the 1940s through the early 1990s, which is contrary to the non-normalized data (Figs. 3 and 4). This highlights the good fortune experienced by the U.S. with respect to hurricane landfalls in recent decades (Landsea et al., 1996).

6 CONCLUSION

Tropical cyclones affect hundreds of millions of people every year around the world. While the rainfall produced by these storms often provides valuable societal and environmental benefits, these storms also have the potential to inflict great harm and suffering. Recent history suggests that communities in the Atlantic basin have been fortunate in recent decades, due to an extended period of relatively fewer hurricanes (Landsea et al., 1996). However, simply because hurricanes have been depressed in

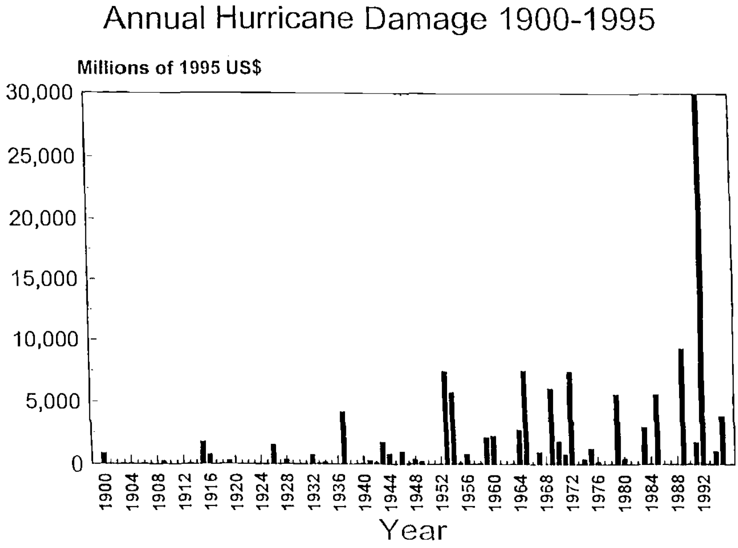


Figure 3 Inflation adjusted hurricane damages of the 20th century. (Pielke and Landsea, 1997).

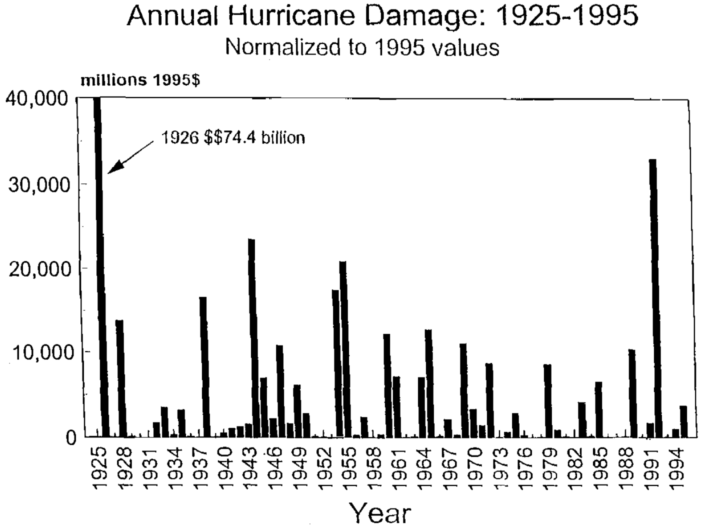


Figure 4 Hurricane damages adjusted for inflation, wealth, and population 1925 to 1995 (Pielke and Landsea, 1997).

recent decades does not eliminate the possibility of large impacts, as shown by hurricane Andrew, which occurred during the quietest 4-year period of hurricane activity since 1950. The $30 billion hurricane Andrew was the costliest tropical cyclone ever (Landsea et al., 1996; Pielke and Pielke, 1997).

Tropical cyclones occur every year around the world. In this most basic sense, they are "normal" climatological events on planet Earth. But from a human perspective, even a weak tropical cyclone can be an "extreme" occurrence. The challenge of effectively reducing societal vulnerability to hurricanes is made more difficult by the relative infrequency with which storms affect particular communities. Consider that the last major hurricane to strike Dade County, Florida, prior to Andrew was in 1950! Thus, one important step any decision maker should take is to understand the risks and potential consequences of choices made in tropical cyclone-prone regions. Damaging losses associated with tropical cyclones can never be eliminated, but with close attention to those factors that increase our vulnerability—where we live, how we live, etc.—we can hope to live in greater harmony with one of nature's most powerful forces.

REFERENCES

Ager, D., *The New Catastrophism*. Cambridge University Press, Cambridge, 1993.

Anthes, R. A., *Tropical Cyclones: Their Evolution, Structure and Effects*, American Meteorological Society, Boston, MA, 1982.

Bipartisan Task Force on Funding Disaster Relief BTFFDR. *Federal Disaster Assistance: Report of the Senate Task Force on Funding Disaster Relief*, No. 104-4, U.S. Government Printing Office, Washington, DC, 1995.

Black, P. G., and R. M. Wakimoto, Damage survey of hurricane Andrew and its relationship to the eyewall, *Bull. Am. Meteor. Soc.*, *75*, 189–200, 1994.

Bosart, L. F., The Texas coastal rainstorm of 17–21 September 1979: An example of synoptic-mesoscale interaction, *Monthly Weather Rev.*, *112*, 1108–1133, 1984.

Bryant, E. A., *Natural Hazards*. Cambridge University Press, Cambridge, 1991.

Dunn, G. E., and B. I. Miller, *Atlantic Hurricanes*, Louisiana State University Press., Baton Rouge, LA, 1964.

Elsberry, R. L., W. M. Frank, G. J. Holland, J. D. Jarrell, and R. L. Southern, A global view of tropical cyclones, based largely on materials prepared for the International Workshop on Tropical Cyclones, Bangkok, Thailand, November 25–December 5, 1985, Office of Naval Research, Marine Meteorology Program, Robert F. Abbey, Director, 1987.

Gibbs, W. J., Defining climate. *WMO Bull.*, *36*, 290–296, 1987.

Gray, W. M., A global view of the origin of tropical disturbance and storms, *Monthly Weather Rev.*, **96**, 669–700, 1968.

Hebert, P. J., J. D. Jarrell, and M. Mayfield, *The Deadliest Costliest, and Most Intense United States Hurricanes of This Century*, NOAA NWS NHC-31, 1993.

Hebert, P. J., J. D. Jarrell, and M. Mayfield, *The Deadliest, Costliest, and Most Intense United States Hurricanes of This Century (and Other Frequently Requested Hurricane Facts)*, NOAA Technical Memorandum NWS TPC-1, National Hurricane Center, Miami, FL, February 1996.

Henry, J. A., K. M. Portier, and J. Coyne, *The Climate and Weather of Florida*, Pineapple Press, Sarasota, FL, 1994.

Jarvinen, B. R., and M. B. Lawrence, An evaluation of the SLOSH storm-surge mode, *Bull. Am. Meteor. Soc.*, *66*, 1408–1411, 1985.

Katz, R. W., and B. G. Brown, Extreme events in a changing climate: Variability is more important than averages, *Climatic Change*, *21*, 289–302, 1992.

Katz, R. W., Towards a statistical paradigm for climate change, *Climate Res.* **2**, 167–175, 1993.

Landsea, C. W., N. Nicholls, W. M. Gray and L. A. Avila, Quiet early 1990s continues trend of fewer intense Atlantic hurricanes, *Geophys. Res. Lett.*, *23*, 1697–1700, 1996.

Ludlam, D. M., *Early American Hurricanes*: 1492–1870, American Meteorological Society, Boston, MA, 1963.

McAdie, C. J., and E. N. Rappaport, *Diagnostic Report of the National Hurricane Center*, Vol. 4, No. 1, NOAA, National Hurricane Center, Coral Gables, FL, 1991.

Neumann, C. J., Global overview, in *Global Guide to Tropical Cyclone Forecasting*, World Meteorological Organization (WMO) Technical Document, WMO/TD NO. 560, Tropical Cyclone Programme, Report No. TCP-31, WMO, Geneva, Switzerland, Chapter 1, 1993.

Neumann, C. J., B. R. Jarvinen, and A. C. Pike, *Tropical Cyclones of the North Atlantic Ocean, 18711986* , 3rd rev., NOAA Historical Climatology Series 6-2, NCDC; Asheville, NC, 1987.

National Research Council (NRC), *Opportunities to Improve Marine Forecasting*, National Academy Press, Washington, DC, 1989.

Pielke, Jr., R. A., and C. W. Landsea, Normalized hurricane damages in the United States 1929–1995, *Weather Forecast.*, 1997.

Pielke, Jr. R. A., and R. A. Pielke, *Hurricanes: Their Nature and Impacts on Society*, J Wiley, New York, 1997.

Rippey, B., Weatherwatch—August 1996, *Weatherwise*, *49*, 51–53, 1997.

Rodriquez, H., A socioeconomic analysis of hurricanes in Puerto Rico: An overview of disaster mitigation and preparedness, in H. F. Diaz and R. S. Pulwarty (Eds.), *Hurricanes*, Springer-Verlag, Berlin, 1997, pp. 121–143.

Simpson, R. H., and H. Riehl, *The Hurricane and Its Impact*, Louisiana State University Press, Baton Rouge, LA, 1981.

Smith, K., *Environmental Hazards: Assessing Risks and Reducing Disaster*, Routledge, London, 1992.

Southern, R. L., Savage impact of recent catastrophic tropical cyclones emphasizes urgent need to enhance warning/response and mitigation systems in the Asia/Pacific region, unpublished.

Sugg, A. L., Economic aspects of hurricanes, *Monthly Weather Rev.*, *95*, 143–146, 1967.

Tannehill, I. R., *Hurricanes: Their Nature and History*, 8th ed., Princeton University Press, Princeton, NJ, 1952.

Williams, J., *The Weather Book*. Vintage Books, New York, 1992.

CHAPTER 43

EL NIÑO IN AUSTRALIA

NEVILLE NICHOLLS

1 INTRODUCTION

Before the 1972–1973 El Niño episode, understanding of the impacts of the El Niño–Southern Oscillation (ENSO) on Australia was limited. Studies in the 1970s and 1980s documented its effects, but the 1982–1983 event still caught the country by surprise. By the El Niño events of the early 1990s, a routine seasonal climate prediction service, based on the earlier work on the ENSO, had been established.

2 EL NIÑO–SOUTHERN OSCILLATION EFFECT ON AUSTRALIAN CLIMATE

Australian droughts generally accompany El Niño episodes (e.g., Allan, 1991). Figure 1 illustrates the relationship between widespread Australian drought and low values of the Southern Oscillation Index (the SOI, a simple measure of the ENSO, is the standardized difference in surface atmospheric pressure between Tahiti and Darwin), by comparing time series of the percentage of Australia with annual rainfall in the lowest decile with annual averages of the SOI. The figure also indicates that years with little of the country in drought tend to have large positive SOI values, ie., La Niña episodes.

Figure 1 only uses data from 1950, for clarity. The relationship between the SOI and drought, however, is evident in data throughout the twentieth century. Prior to the late nineteenth century there are insufficient data to allow a strict, quantitative comparison of widespread Australian droughts with the ENSO. Nicholls (1988) examined reports of the governors of the colony of New South Wales to the colonial secretary of the British Government in London for references to drought in the early

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

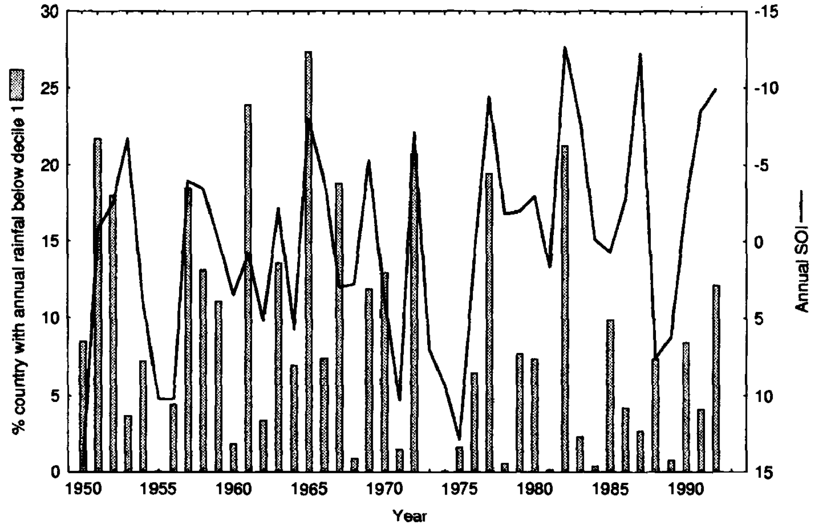


Figure 1 Annual mean SOI (full line) and the percentage of Australia with annual rainfall below the first decile, i.e., in drought, (broken line).

years of the colony and found that the coincidence of El Niño events and Australian droughts has existed from, at least, the start of European colonization in 1788.

The ENSO also enhances Australian rainfall variability, as it does wherever it impacts on climate (Nicholls et al., 1997). Also, many Australian droughts tend to last about a year because El Niño and La Niña episodes both tend to last about 12 months and this sets the time scale of Australian rainfall fluctuations (Nicholls, 1991). The link with the ENSO is most consistent with east and north Australian rainfall (e.g., Pittock, 1975; McBride and Nicholls, 1983; Ropelewski and Halpert, 1987, 1989).

3 DISCOVERY OF EFFECT OF EL NIÑO–SOUTHERN OSCILLATION ON AUSTRALIA

India suffered a severe drought and famine during 1877. Sir Henry Blanford, the director of the Indian Meteorological Service, noted the very high atmospheric pressures over Asia at the time and requested pressure information from other meteorologists around the world. Sir Charles Todd, the South Australian government observer noted that pressures were also high during 1877 over Australia, and much of the country suffered from drought that year. Todd compared earlier droughts and concluded that Indian and Australian droughts usually coincided. This observation has since been confirmed (e.g., Williams et al., 1986) and forms part of the suite of climate linkages we now call the Southern Oscillation (SO).

When Sir Gilbert Walker named and documented the SO in the early decades of the twentieth century, its close relationship with Australian rainfall quickly became apparent (e.g., Bliss and Walker, 1932). Walker's work suggested that north Australian summer rainfall could be predicted with an index of the SO. Quayle (1910, 1929) suggested that rainfall farther south could be predicted in the same way. After that, a trickle of studies discussed the relationship between the SO and Australian climate, up to the mid-1970s, when the worldwide attention on El Niño led to a resurgence of interest among Australian meteorologists.

By the early- 1980s attention had turned to the possible use of the ENSO in prediction. Work on the physical cause of the phenomenon had commenced, and several studies describing patterns and relationships between the ENSO, sea surface temperature, and Australian climate had been published (e.g., Pittock, 1975; Streten, 1981; Coughlan, 1979). Some of the lag relationships suggested by Quayle and others had been validated and extended using new data (Nicholls and Woodcock, 1981; McBride and Nicholls, 1983). New relationships indicating that seasonal temperature, wet-season onset, and even seasonal tropical cyclone activity also were predictable, through the ENSO, had been uncovered (Nicholls, 1978, 1979; Nicholls et al., 1982). The recognition in mid-1982 that a major El Niño episode was under way led to cautious statements regarding possible implications for Australian rainfall through the remainder of 1982, based on this work (Nicholls, 1983). The Bureau's National Climate Centre began preparing and issuing regular monthly "Seasonal Climate Outlooks" in 1989, based on the SOI. These provide forecasts of 3-month rainfall anomalies, across the country.

Variables other than seasonal rainfall appear to be predictable through the use of the ENSO. For instance, Stone et al. (1996) suggest that seasonal frost forecasts could be feasible in eastern Australia. Nicholls and Kariko (1993) and Suppiah and Hennessy (1996) found that rainfall events and intensity were related to the ENSO. Whetton et al. (1990) and Allan et al. (1996) documented relationships of the ENSO with streamflow variations.

4 ECOLOGICAL IMPACTS OF EL NIÑO–SOUTHERN OSCILLATION

The widespread effect of the ENSO on Australian climate variations suggests that there should be strong responses to the phenomenon in Australian biota, including crops. There is ample evidence that the high variability the ENSO imposes on Australian climate impacts the wildlife and vegetation; populations of many Australian animals are maintained at levels well below the carrying capacity of the good years. Populations typically increase dramatically during periods of good rainfall and fall during the frequent droughts. Some adaptations to variable precipitation observed in the Australian native biota are described by Nicholls (1989, 1991). Each adaptation allows opportunistic use of good conditions, thereby producing rapid increases in populations when a drought breaks.

Red Kangaroo

Australia's largest herbivore, the red kangaroo, inhabits the open arid and semiarid plains that cover most of the continent. It shows no seasonal pattern of reproduction but breeds opportunistically in response to good conditions by producing young in rapid succession. Under prolonged drought the kangaroos stop breeding. Drought-breaking rains trigger an immediate hormonal response. The females return rapidly to breeding and may be found with young in the pouch after 60 days. In favorable environmental conditions females become sexually mature when 15 to 20 months old. Drought delays the onset of sexual maturity and after 2 years of drought a population may include females aged 3 years or more that have never produced young. After rain these animals come into breeding condition almost immediately. The life-history strategy of the red kangaroo is clearly adapted to highly variable rainfall.

Green Turtles

The number of green turtles observed nesting around northern Australia varies widely from year to year, and these interannual fluctuations are in phase at widely separated rookeries, with large numbers of turtles breeding 2 years after major El Niño episodes (Limpus and Nicholls, 1988). Preparation for breeding commences well over a year before oviposition. Atmospheric or oceanic anomalies associated with El Niño (perhaps increased availability of food due to the reduced number of tropical cyclones during El Niño) triggers the turtles to commence breeding. The relationship with El Niño provides a means for predicting, a long way in advance, the approximate numbers of turtles breeding. Such a prediction is potentially useful in sea turtle management in areas where eggs, courting turtles, or nesting females are harvested.

Australian Encephalitis

Australian encephalitis (AE) is a severe, often fatal, viral illness transmitted to humans by mosquitoes and influenced by the ENSO (Nicholls, 1986). Since 1917 when the clinical symptoms of the illness were first diagnosed, there have been only 7 years when cases of AE were observed in southeastern Australia. Cases occur between January and April and follow widespread flooding over several seasons. Flooding leads to increased mosquito numbers by increasing the numbers of breeding sites and host populations (birds, marsupials). The probability that AE will be diagnosed in southeastern Australia is predictable from the SOI in the previous spring (September–November). The relationship is sufficiently strong to allow health authorities to increase surveillance and prophylactic action, in years when the SOI, during spring, is very high.

Banana Prawns

The prawn season in the Gulf of Carpentaria extends from March to June. Prawn catch is related to the amount of rainfall. Mature prawns require a saline environment. With the advent of heavy wet-season rainfall, salinity in the rivers where the prawns breed is lowered, and they are forced to migrate offshore. Once offshore the prawns may be harvested. If wet-season rainfall is very low, then fewer prawns leave the rivers, leading to a low catch. The relationship between the SOI and rainfall in this region means that there is a significant relationship between the SOI in November and the subsequent prawn catch (Love, 1987). Low values of the SOI lead to a lower than normal catch.

Waterfowl

The number of ducks shot on the opening day of annual waterfowl season in southeast Australia is correlated with the SOI some 2 year prior to the season (Norman and Nicholls, 1991). Apparently, heavy rains (which often follow an El Niño–related drought) result in widespread floods. These fill ephemeral wetlands, leading to enhanced waterfowl breeding (few species breed on permanent waters). In the following year, as the post-El Niño floods recede, the waterfowl congregate in the permanent wetlands, thereby leading to inflated numbers. An increased harvest ensues in the next open season. So, an El Niño, with low SOI, tends to be followed about 2 years later by an increased duck harvest.

Australian Birds

Some other adaptations of Australian birds that can be linked to the unpredictable environment (in turn caused by the ENSO) are nomadism, irregular and seasonal breeding initiated by sudden falls of rain, variations in clutch and multiple broods dependent on the rainfall, precocious breeding, and the habit of older offspring of assisting the breeding male in raising siblings of subsequent broods. All these behavior patterns contribute to an opportunistic life-history strategy. Population increases can be very rapid after drought, just as with the red kangaroo.

Vegetation

Vegetation also is linked to the high variability caused by ENSO. The following are just some of the characteristics of Australian vegetation that may be, at least in part, attributable to the ENSO influence on the climate: absence of succulents, establishment dependent on extended wet periods, drought tolerance/avoidance, diverse life-history strategies, and fire resistance/dependence (Nicholls, 1991).

These are just a few examples of the many and varied adaptations of Australian biota to the highly variable rainfall found in some members of most groups of Australian plants and animals. There appears to be a consensus among ecologists that much of the Australian flora and fauna is adapted to a highly variable rainfall that, in turn, is caused by the ENSO, and that this adaptation is more complete than in other areas of the globe.

5 EL NIÑO–SOUTHERN OSCILLATION AND VEGETATION CHANGES

Since the native Australian vegetation was adapted to the climate rhythms and variability induced by the ENSO, it is not surprising that the introduction of plants and animals not so adapted led to rapid changes in vegetation (Nicholls, 1991). The best known of these changes is probably the area now known as the Pilliga Scrub in northern New South Wales (Rolls, 1981; Austin and Williams, 1988). Much of this area of 400,000 ha was open grassy country with only about eight large trees per hectare when Europeans arrived in the 1830s. Frequent burning by Aboriginals, and grazing by indigenous marsupials, restricted the opportunities for trees and shrubs to establish. Fire germinated the seed of the trees and shrubs, but rat kangaroos ate many of the resulting seedlings before they could establish.

The introduction of sheep reduced the numbers of rat kangaroos, by destroying their cover and their food. A severe drought during the major El Niño of 1877–1978 further reduced the numbers of indigenous marsupials. The following year, a major La Niña event, was very wet. The few large trees seeded well and when stock owners burnt to destroy grasses with seeds that got into their sheep's wool, seedlings came up thickly, unhindered by the grasses that would usually compete with them for space. This time there were no rat kangaroos to eat the seedlings either and the trees grew unchecked.

Over the next decade there were several further periods of establishment, again synchronized with El Niño–La Niña oscillations. The European rabbit, also an enthusiastic eater of seedlings, arrived in the area in the late 1880s and prevented further establishment until myxomatosis in 1951 reduced the rabbit population. The first successful release of myxomatosis occurred in 1950. Earlier releases of the disease had not led to widespread establishment. The extensive rains and flooding in 1950, associated with a major La Niña, contributed to the successful establishment of the disease by providing ideal breeding conditions for the insects that spread it.

In 1917 the Forestry Commission stopped burning in the Pilliga and by 1950 large amounts of forest litter had accumulated. So had decades of seed production. The forest dried in El Niño event of 1951, following good growth during La Niña of 1950, and a major fire started in November 1951. In the absence of rat kangaroos and rabbits, the new growth induced by the fire had nothing to stop it.

In less than a century Europeans had unintentionally transformed the area from grazing land into the dense Pilliga Scrub supporting sustained timber harvesting. The ENSO phenomenon played a critical role in this transformation. McKeon et al. (1990) cite other examples where the extreme climate events associated with both extremes of the ENSO resulted in major long-term vegetation degradation. In western Queensland there was a rapid increase in the sheep population during the above-average rainfall years of the early 1890s. Major El Niño events between 1899 and 1902 resulted in very low rainfall and a rapid drop in animal numbers. Heavy utilization of edible grasses and shrubs during this drought led to a spread of inedible plants and carrying capacities seem to have been permanently reduced. In the subtropical grasslands of southern coastal Queensland, rapid change in species composition to bunch spear grass appears to have resulted from overgrazing with sheep during El Niño–related drought of 1881–1882. More recently, low beef prices in the mid-1970s led to increased stocking rates in Queensland. These years were wet, the result of the 1973–1975 La Niña, but attempts to maintain the high stocking rates into the 1980s with their drier, El Niño conditions have led to pasture degradation, species changes, and soil erosion.

6 IMPACTS OF EL NIÑO–SOUTHERN OSCILLATION ON AUSTRALIAN CROPS

Not surprisingly, given its effects on Australian climate, the ENSO has a major impact on crop yields. Figure 2 shows time series of wheat yields, averaged across Australia, and the SOI. The year-to-year differences in the two variables are plotted, to remove the effects of trends and changes such as the introduction of new cultivars. The relationship is clear—negative values of the SOI lead to widespread drought (Fig. 1), which leads to low crop yields (Nicholls, 1985).

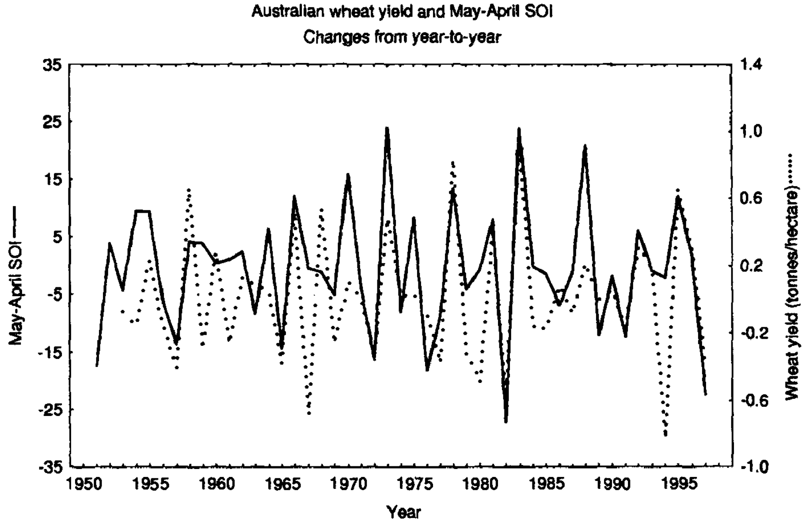


Figure 2 Annual mean SOI (full line) and the yield (tonnes per hectare) of wheat, averaged across the country. Differences in SOI and yield plotted, to remove long-term variations such as the effects of changes in cultivars.

Rimmington and Nicholls (1993) demonstrated that wheat yields in all states were correlated with values of the SOI from before and near the sowing date, which therefore can provide skillful yield forecasts of Australia's major crop. These forecasts would be available several months before harvest starts, require little data, and are quick and simple to prepare. Strong negative relationships also exist with the SOI in the year before the crop is planted, i.e., an El Niño episode is often followed by good crops the following year. This partly reflects the biennial nature of the ENSO, but may also reflect a tendency for a drop in pests in the droughts associated with negative SOI values. This would amplify any response of the crops to good rains in the following year.

Hammer et al. (1996) examined the value of ENSO-based forecasting methodologies to wheat crop management in northern Australia, by examining decisions on nitrogen fertilizer and cultivar maturity using simulation analyses of specific production scenarios. The average profit and risk of making a loss were calculated for the possible range of fixed (i.e., the same each year) and tactical (i.e., varying depending on the ENSO-based seasonal forecast) strategies. Significant increases in profit (up to 20%) and/or reduction in risk (up to 35%) of making a loss were associated with the tactical (forecast-based) strategies. The skill in seasonal rainfall and frost predictions, based on the ENSO, generated the value from using tactical management. This study demonstrated that the skill obtainable in Australia was sufficient to justify, on economic grounds, their use in crop management. Presumably, these forecasts could also be useful in drought management decision making, for instance, in determination of appropriate stocking rates on pastoral properties (McKeon et al., 1990).

REFERENCES

Allan, R. J., Australasia, in M. Glantz, R. Katz, and N. Nicholls (Eds.), *Teleconnections Linking Worldwide Climate Anomalies*, Cambridge University Press, Cambridge, 1991, pp. 73–120.

Allan, R. J., G. S. Beard, A. Close, A. L. Herczeg, P. D. Jones, and H. J. Simpson, *Mean Sea Level Pressure Indices of the El Niño–Southern Oscillation: Relevance to Stream Discharge in South-eastern Australia*, Divisional Report 96/1, *CSIRO Division of Water Resources*, Canberra, Australia, 1996.

Austin, M. P., and O. B. Williams, Influence of climate and community composition on the population demography of pasture species in semi-arid Australia, *Vegetatio*, *77*, 43–49, 1988.

Bliss, E. W., and G. T. Walker, World weather V, *Mem. R. Meteorolo. Soc.*, *4*, 52–84, 1932.

Coughlan, M. J., Recent variations in annual-mean maximum temperatures over Australia, *Q. J. R. Meteorol. Soc.*, *105*, 707–719, 1979.

Hammer, G. L., D. P. Holzworth, and R. Stone, The value of skill in seasonal climate forecasting to wheat crop management in a region with high climatic variability, *Aust. J. Agric. Res.*, *47*, 717–737, 1996.

Limpus, C. J., and N. Nicholls, The Southern Oscillation regulates the annual numbers of green turtles (*Chelonia mydas*) breeding around northern Australia, *Austral. J. Wildlife Res.*, *15*, 157–161, 1988.

Love, G., Banana prawns and the Southern Oscillation Index, *Austral. Meteorol. Mag.*, *35*, 47–49, 1987.

McBride, J. L., and N. Nicholls, Seasonal relationships between Australian rainfall and the Southern Oscillation, *Monthly Weather Rev.*, *111*, 1998–2004, 1983.

McKeon, G. M., K. A. Day, S. M. Howden, J. J. Mott, D. M. Orr, W. J. Scattini, and E. J. Weston, Management of pastoral production in northern Australian savannas, *J. Biogeog.*, *17*, 355–372, 1990.

Nicholls, N., A possible method for predicting seasonal tropical cyclone activity in the Australian region, *Monthly Weather Rev.*, *107*, 1221–1224, 1978.

Nicholls, N., A simple air-sea interaction model, *Q. J. R. Meteorol. Soc.*, *105*, 93–105, 1979.

Nicholls, N., Predictability of the 1982 Australian drought, *Search*, *14*, 154–155, 1983.

Nicholls, N., Impact of the Southern Oscillation on Australian crops, *J. Climatol.*, *5*, 553–560, 1985.

Nicholls, N., A method for predicting Murray Valley Encephalitis in southeast Australia using the Southern Oscillation, *Austral. J. Exper. Biol. Med. Sci.*, *64*, 587–594, 1986.

Nicholls, N., More on early ENSOs: Evidence from Australian documentary sources, *Bull. Am. Meteorol. Soc.*, *69*, 4–6, 1988.

Nicholls, N., How old is ENSO? *Climatic Change*, *14*, 111–115, 1989.

Nicholls, N., The El Niño–Southern Oscillation and Australian vegetation, *Vegetatio*, *91*, 23–36, 1991.

Nicholls, N., and A. P. Kariko, East Australian rainfall events: Interannual variations, trends, and relationships with the Southern Oscillation, *J. Climate*, *6*, 1141–1152, 1993.

Nicholls, N., and F. Woodcock, Verification of an empirical long-range weather forecasting technique, *Q. J. R. Meteorol. Soc.*, *107*, 973–976, 1981.

Nicholls, N., J. L. McBride, and R. J. Ormerod, On predicting the onset of the Australian wet season at Darwin, *Monthly Weather Rev.*, *110*, 14–17, 1982.

Nicholls, N., W. Drosdowsky, and B. Lavery, Australian rainfall variability and change, *Weather*, 1997.

Norman, F. I., and N. Nicholls, The Southern oscillation and variations in waterfowl abundance in southeastern Australia, *Austral. J. Ecol.*, *16*, 485–490, 1991.

Pittock, A. B., Climatic change and the patterns of variation in Australian rainfall, *Search*, *6*, 498–504, 1975.

Quayle, E. T., *On the Possibility of Forecasting the Approximate Winter Rainfall for Northern Victoria*, Bulletin No. 5, Commonwealth Bureau of Meteorology, Melbourne, March 1910.

Quayle, E. T., Long range rainfall forecasting from tropical (Darwin) air pressures, *Proc. R. Soc. Victoria*, *41*, 160–164, 1929.

Rimmington, G. M., and N. Nicholls, *Austral. J. Agric. Res.*, *44*, 625–632, 1993.

Rolls, E. C., *A Million Wild Acres*, Nelson, Melbourne, 1981.

Ropelewski, C. F., and M. S. Halpert, Global and regional scale precipitation patterns associated with the El Niño–Southern Oscillation, *Monthly Weather Rev.*, *115*, 1606–1626, 1987.

Ropelewski, C. F., and M. S. Halpert, Precipitation patterns associated with the high index phase of the Southern Oscillation, *J. Climate*, *2*, 268–284, 1989.

Stone, R., N. Nicholls, and G. Hammer, Frost in northeast Australia: Trends and influences of phases of the Southern Oscillation, *J. Climate*, *9*, 1896–1909, 1996.

Streten, N. A., Southern Hemisphere sea surface temperature variability and apparent associations with Australian rainfall, *J. Geophys. Res.*, *86*, 485–497, 1981.

Suppiah, R., and K. J. Hennessy, Trends in the intensity and frequency of heavy rainfall in tropical Australia and links with the Southern Oscillation, *Austral. Meteorol. Mag.*, *45*, 1–18, 1996.

Whetton, P., D. Adamson, and M. Williams, Rainfall and river flow variability in Africa, Australia and East Asia linked to El Niño–Southern Oscillation events, *Geol. Soc. Austral. Symp. Proc.*, *1*, 71–82, 1990.

Williams, M. A. J., D. A. Adamson, and J. T. Baxter, Late Quaternary environments in the Nile and Darling basins, *Austral. Geogr. Stud.*, *24*, 128–144, 1986.

CHAPTER 44

BIOLOGICAL AND SOCIETAL IMPACTS OF CLIMATE VARIABILITY: AN EXAMPLE FROM PERUVIAN FISHERIES

KENNETH BROAD

1 INTRODUCTION

The collapse of the massive industrial anchovy fishery in Peru in 1973 brought El Niño to the world's attention. However, small-scale (artisanal) fishermen from northern Peru and southern Ecuador were aware of this phenomenon long before this widely publicized event. They realized at least a century ago that every few years, around Christmas time, a warm water current appeared close to their desert shores; they named this current "El Niño", or "the boy child," after the baby Jesus. The first time the term El Niño appeared was in a Peruvian newspaper in 1891, as a result of what we now know was a very strong event. Historical reconstruction (Quinn et al., 1987) of data from ship's logs, fish landings, bird populations, crop yields, among other indicators, have documented El Niño events back several hundred years, and others have argued, using proxy evidence, that El Niño has recurred over many millennium (Rodbell et al., 1999).

After intensive studies and data accumulation, it was discovered that El Niño was not only an oceanographic phenomenon related to the western coast of South America but a complex interaction between the ocean (sea surface temperature changes in the equatorial Pacific) and atmosphere (sea-level pressure changes in the western equatorial Pacific)—hence called the El Niño–Southern Oscillation (ENSO). ENSO is now known to impact climate patterns around the globe. Nonetheless, the Peruvian coast remains one of the areas most consistently and directly impacted by this

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

recurrent event. ENSO also has a less common, cold phase, sometimes referred to as La Niña. Except for scientists who study ENSO, most Peruvians do not differentiate La Niña from "normal" conditions. (This chapter uses the term ENSO to refer to El Niño or ENSO warm events.)

While most associate ENSO with negative impacts on flora and fauna, this cyclical climate event, which thus far has presented itself differently with each occurrence in terms of its strength, duration, and intensity, has a range of both negative and positive ecological impacts. These environmental changes trigger varied socioeconomic and political reactions that in turn may alter aspects of that society as a whole over time. In addition, society undergoes change. A changed society will then react differently to the next ENSO event. This makes planning for mid to long-term climate variability a challenging task for governments and individuals. In the last decade and a half, much effort has been put into better understanding and predicting regional climate variability associated with ENSO events, and forecasts are now being used by governments and individuals in many parts of the world in their planning efforts.

This section discusses some of the key impacts of ENSO on the fisheries sector of Peru, including a brief description of what ENSO is and the evolution of scientific interest in this phenomenon; a brief overview of the Peruvian fisheries sector, with examples of impacts from the 1997–1998 ENSO event; and policy implications of ENSO-related climate forecasts.

2 WHAT IS ENSO?

The El Niño–Southern Oscillation (ENSO) is a coupled atmospheric–oceanic phenomenon that has global manifestations and recurs approximately every 2 to 10 years. The atmospheric component of ENSO is the Southern Oscillation, an interannual seesawing of sea-level atmospheric pressure anomalies between northern Australia (Darwin) and the southeast Pacific (near Tahiti). There is both a "warm phase" (El Niño) and a "cold phase" (La Niña). The warm phase involves an extensive warming of the upper ocean along the central and eastern equatorial Pacific and a depression of the thermocline (the boundary that separates the warmer mixed upper layer of the ocean from the cold abyss) in the eastern tropical Pacific. The cold phase involves a cooling of the upper ocean and a rise of the thermocline toward the ocean's surface in the eastern tropical Pacific.

During an ENSO warm phase, as the thermocline deepens, wind-driven coastal upwelling off the shores of South America carries warmer water than usual to the surface. Coastal seasurface temperature anomalies as high as 10°C have been recorded off Peru (Sharp and McLain, 1993).

The Southern Oscillation leads to a cyclic increase and decrease in the strength of the Southern Hemisphere (southeast) trade winds. These winds are strongest during the oceanic cold phase, when sea-level atmospheric pressure in northern Australia, normally low, is anomalously lower, while that in the southeast Pacific, normally high, is anomalously higher. During the oceanic warm phases, when the sea-level pressure in northern Australia is anomalously high and that in the southeast Pacific is anomalously low, the trade winds weaken, and in extreme cases even blow westerly. There is positive interaction (e.g., feedback) between the ocean and atmosphere, as increased sea surface temperatures increase atmospheric pressure (Philander, 1990). For a comprehensive overview of the observations and mechanisms of the 1997–1998 ENSO see McPhaden (1999).

Dramatic shifts of flora and fauna in waters off southern Colombia and Ecuador to Peru and northern Chile are linked to ENSO events (Arntz et al., 1985). In severe events, the increased ocean temperatures and reduced concentrations of phytoplankton negatively impact some pelagic species, such as the commercially important anchovy and sardines (see Figure 1). Tropical species of fish, however, may extend their ranges to the south and closer to shore, as warmer waters appear along the Peruvian coast. For an overview of the 1997 to 1998 event on biogeochemical cycles and on the use of satellite technology during this event, see Chavez et al. (1998), McPhaden (1999), and Carr and Broad (2000).

The workings of ENSO were first investigated in the mid-1960s by a researcher named Jacob Bjerknes (1966). Bjerknes linked the oceanic process off the Peruvian coast with the seesaw in atmospheric pressure between the western and central equatorial Pacific (i.e., the Southern Oscillation). A growing interest in the possible global connection of climatic events led to the establishment of the World Climate

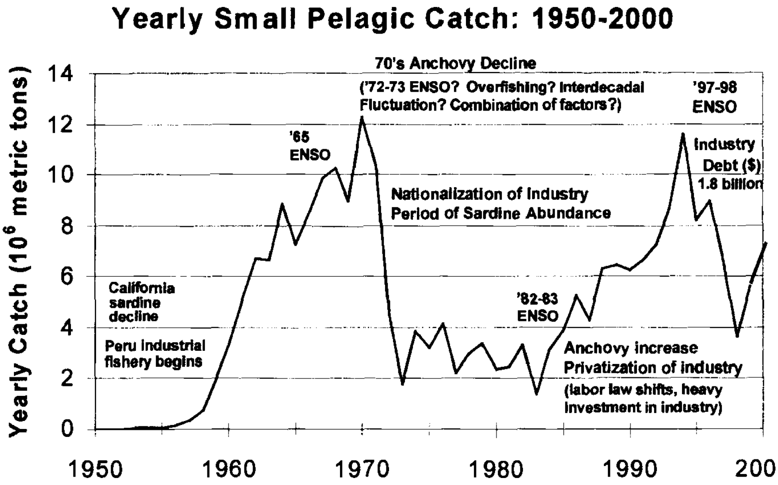


Figure 1 Peru annual small pelagic catch 1950–2000 (includes anchovy, sardine and mackerel). (Sources: 1950–1990: Csirke et al., "La ordenación y planificación pesquera y la reactivación del sector pesquero en el Perú" Rome: Jun 92. 1991–1996: Perú: desembarque de recursos marítimos, según especie 90-96 INEI: Jul 97. 1997–1999: *Statistical Reference Book*, No. 13: FEO Proceedings from 1999 FEO Annual Conference, Hong Kong, April 8–9, 1999. Paris: Fishmeal Exporters Organization.)

Research Programme (WCRP) under the auspices of the World Meteorological Organization (WMO), whose mandate resulted in national programs of various types around the world. Scientists were setting up an international experiment as part of CUEA, Coastal Upwelling Ecosystems Analysis, off the Peruvian coast when the 1972 El Niño event took place. Interest in forecasting El Niño was seen as socially relevant because forecasting the strength and biological productivity of coastal upwelling processes could be seen as more than an academic exercise (Glantz, 1979). It was suggested that knowledge gained about coastal upwelling and natural factors that inhibit it could be applied to address national economic development issues.

El Niño's impacts on Peruvian fisheries (especially the surface dwelling anchovy) became widely publicized during the 1972–1973 event, which has been blamed, along with overfishing and recruitment failure, for the collapse of that fishery. At that time about one-third of Peru's foreign exchange earnings were derived from the export of anchovy-based fishmeal. There are increasing claims, however, that the rapid decline in catch in 1973 began prior to the ENSO event, and that it was associated with natural fluctuations in abundance of small pelagic stocks. This is based on mounting evidence that anchovy (genus *Engraulins*) and sardine (genus *Sardinops*) populations fluctuate on multiyear or decadal scales as well as interannual timescales. Furthermore, there are indications of basin-wide synchrony in fluctuations of small pelagics (Bakun, 1996; Kawasaki et al., 1991; Lluch-Belda et al., 1992; Sharp and Csirke, 1983; Sharp and McLain, 1993).

The extraordinary 1982–1983 El Niño was the catalyst to expanding government and scientific interest in developing an El Niño forecast capability. In 1985, WCRP launched the multinational 10-year TOGA (Tropical Ocean Global Atmosphere) program that resulted in the recognition of ENSO as a key aspect in the interannual variability of the global climate system. The data collected from this project by researchers from as many as 40 countries aided the development of coupled ocean–atmosphere general circulation models. These models are intended to produce routine seasonal to interannual climate forecasts. Despite intensive worldwide coordination and effort invested in the physical understanding of ENSO phenomenon, only recently have researchers begun to address the socioeconomic factors. The rise in scientific understanding eventually translates into policy-making decisions and economic adjustments that have social consequences at all levels of society.

The ENSO event of 1997–1998 further heightened interest in the phenomenon and in the forecasting of it. Thanks to scientific concern and large-scale media coverage, numerous workshops and conferences, and the evolution of global communications technologies such as the Internet, societies around the globe became aware of El Niño and ENSO forecasts.

Peru, in particular, closely monitored ENSO information, and Peruvians generated their own forecasts of how the event would evolve. Peru has been a key member in the Comisión Permanente del Pacífico Sur (CPPS), which, in response to the 1972–1973 ENSO event, created a group called ERFEN (Estudios Regionales del Fenomeno El Niño). CPPS has committees in each of its member countries—Colombia, Ecuador, Peru, and Chile. Through this organization, national data from cruises and observational stations are shared, allowing more comprehensive study of this regional phenomenon. In Peru, for example, several institutions, including the government oceanographic agency, the navy, the national meteorological service, the Peruvian Geophysical Institute, civil defense, and the private sector were involved in studies, conferences, and workshops attempting to enhance the understanding and preparation for the impacts of this event.

The Peruvian fishing sector, given its dramatic experiences with past events, was among the first groups to become concerned about the possibility of an extreme event in 1997–1998. Peru co-sponsored, along with the U.S. National Oceanic and Atmospheric Administration (NOAA), workshops bringing together international groups of scientists to discuss the matter, and to provide more detail about the event and its possible impacts. These workshops brought to light both the potential gains and difficulties in using forecasts for planning effective mitigative action in the fishing sector.

3 PERUVIAN FISHERIES SECTOR

In 1996, Peru was second to China in terms of the volume of fish landings and accounted for more than 10% of the world's catch. In 1997, this sector generated more than 2% of Peru's gross domestic product and consistently accounts for approximately 16% of all export products (second only to mining products). Bordered to the north by Ecuador and the south by Chile, the Peruvian coastline stretches 3100 km (01°01′48″ and 18°21′03″ south latitude), and has a continental shelf of 87,200 km^2. Its stark coastal desert is dotted with over 70 fishing ports ranging in population from a few hundred fishermen and their families to several hundred thousand persons. The fishing sector as a whole employs about 80,000 persons, out of a total population of over 24 million.

With the exception of ENSO years, the arid coastal climate is stable, as a result of the relatively cold coastal sea surface temperatures and high barometric pressure. There is evidence of reliance on living marine resources dating back to 7000 B.C., and continuing through the Moche, Chimbu, Nazca, and Paracas cultures, as well as during the Spanish colonial period up to the present. The focus, however, has changed from subsistence fishing for local consumption toward supplying a growing internal and international market. This makes fishermen not only reliant on local conditions and supply, which are impacted by the changing environment, but also vulnerable to fluctuations in global market prices and direct and indirect consumer preferences.

Peruvian coastal waters are home to over 107 commercial species (pelagic, demersal, and benthonic), of which 73 are fish, 11 crustacean, 16 mollusks, 2 echinoderms, and 5 algaes. Some of these species are classified as overexploited, while others are considered underutilized. Fluctuations in abundance of these species are a response to the environmental variability (often ENSO related) in combination with fish population dynamics, fishing pressure, habitat destruction, and pollution. Often, it is impossible to determine the precise reason for a population's variation

TABLE 1 Impacts of ENSO Warm Events on Common Marine Species

Marine Resource	ENSO (warm event)
Pelagics (e.g., primarily anchovy)	Start of event
	Schools concentrate near coast (easier to catch)
	Event strengthens
	Schools go deeper/migrate south
	Increased natural mortality
	Reduced reproduction/recruitment
Demersal (e.g., hake)	Go deeper/migrate south (harder to catch)
	Decreased natural mortality
Littoral (e.g., scallops, shrimp, octopus)	Population increases
Littoral (e.g., mussels, crabs)	Population decrease
Seabirds, marine mammals	Population decrease

Adapted from Ñiquen, M. et al., (1999). Efectos del Fenómeno El Niño 1997–1998 Sobre Los Principales Recursos Pelágicos en la Costa Peruana, In J. Tarazona and E. Castillo (Eds.), *Rev. peru. boil.* Vol. Extraordinario: 85–96.

because of the many factors in operation, in addition to the relative lack of knowledge of the life cycles of many species. What is clear is that some species are favored by the ENSO-related warm waters, while others are harmed. Table 1 summarizes some of the more consistent impacts on different species during an ENSO event.

Almost as varied as these marine organisms are those who harvest them. There are many types of fishers who use a range of equipment and techniques to gather the living marine resources. These fishers range from shore-based breathhold divers who use only a mask, fins, and speargun to shoot the large groupers (*Ephinephelus labriformis*) to the crews of the 800-ton industrial purse seine ships who use spotter planes and satellite images to search for schools of anchovy (*Engraulis ringens*) and sardine (*Sardinops sagax sagax*). Fishermen must register legally as artisanal (small-scale) or industrial, based on techniques, target species, and the size of their vessel. ENSO impacts these groups differently.

ENSO events shift the spatial availability and relative abundance of the species; a given event may benefit one member and harm another.

4 ARTISANAL SUBSECTOR

The artisanal subsector consists of more than 40,000 small-scale producers operating about 7000 vessels and can be characterized by the use of relatively rudimentary technology that has changed little over the last several decades. Artisanal fishermen use diving apparatus, nets, longlines, hook and line, and collect algaes and mollusks in the intertidal zones. Historically, they have occupied the lower socioeconomic strata of society, have been in general self-employed, and have had limited political influence because of poor organization as an economic subsector. Thus, their voice is manifested primarily through voting power.

Divers

Some divers simply hold their breath while hunting and gathering, while others use surface-supplied compressors to collect oysters and scallops, sea urchins, octopus and to spear a variety of finfish along the rocky shores in relatively shallow water (30 m or less). Their boats or rafts are usually rowed out to the fishing ground, powered by small outboards, by sail, or the fishermen simply dive from shore. The catch is then sold to middlemen back at the home port, where it is transported in refrigerated trucks throughout the country or shipped overseas to markets in the United States, Europe, and Asia.

Divers are generally adapted to the changing conditions of the water temperature brought on by ENSO. A moderate ENSO warm event, which warms the sea temperature, actually allows the divers to stay in the water longer without getting cold. As the water gets too warm, however, some species move into deeper waters, which lures the divers to follow them. Diving at deeper depths, combined with warm water that allows one to comfortably remain in the water for longer periods, can lead to an increase in incidence of decompression sickness, i.e., "the bends".

In extreme events such as 1982–1983, the water got so hot (more than 9°C above normal along the northern Peru coast) that many species of shellfish just died, while other species moved to depths outside the range of the divers. As diving equipment improves, and market demands for high-quality shellfish increase, divers will likely continue to push their depths. Unfortunately, training and emergency facilities are not on par with the increase in diving activity. This may be exacerbated, once awareness of an impending strong event occurs, as divers try to squeeze in as much time as possible in the water with the hope of "getting what you can" before the ENSO-related conditions deteriorate.

During warm events, however, some species such as octopus (*Octopus spp.*) and scallops (*Argopecten purpuratus*) grow at faster rates, which can permit increased harvesting at sustainable levels. Again, once the temperatures get too warm, however, these species can also perish. The temporary abundance of these commercially valuable species draws people from other occupations and areas of the country to begin diving for these marine resources. Most do not have proper training, which can result in increased diving accidents.

Net and Longline Fishermen

Net fishermen also fish for a range of finfish close to shore in small vessels (5 to 7 m), using gill nets that sometimes stretch over a kilometer in length. Once they have set their nets and are waiting for a couple hours to retrieve them, they will often fish with hook and line for bottom dwelling fish such as flounder (*Paralychthys adspersus*).

Longline fishermen tend to use larger vessels (10 to 15 m in length with inboard diesel motors), targeting shark, mahi-mahi (*Coriphaena hippurus*), and swordfish (*Xiphias gladius*). They often spend days at sea and go out as far as 80 miles offshore, depending on where the optimal water temperature and currents are found.

This group of fishermen is impacted by changes in the water temperature and depth of the thermocline, as tropical species move in closer to the Peruvian coast, making them more accessible to the fishermen. Mahi-mahi, for instance, feed on the eggs of flying fish (*Cypselurus heterurus*), and the flying fish are one of the first species to migrate with the warmer waters toward the coast, luring the mahi-mahi with them. During the early phases of the 1997 event, the abundance of mahi-mahi stretched down into northern Chilean waters, initially providing a steady source of income for the small-scale fishermen. However, as this ENSO began to increase in strength, an overabundance of these fish flooded the markets (both for internal consumption and for export), leading to a drop in prices. At one point, some fishermen stopped going to sea as the prices for their catch fell to very low levels ($1 per kilo); it was not worth their expenditures on fuel, ice, and materials.

The largest vessel of the artisanal fleet (approximately 30 gross registered metric tons) uses purse seine nets and is dedicated to the capture of anchovy, which is sold to the fishmeal plants. This group has a relative advantage during some ENSO events, as the anchovy move closer to shore in search of the nutrient-rich upwelled water, because the industrial fleet is not permitted to fish within 5 miles of the shore. This can lead to conflict and informal negotiations between the two sectors. With the total disappearance of anchovy during extreme ENSO events such as that of 1997–1998, some of these fishermen have modified their boats with permission and some minor subsidization from the government enabling them to trawl for langostinos, among other species that migrated down from Ecuador.

Many artisanal fishermen live in remote coastal rural villages and lack the infrastructure, such as ice machines and refrigeration systems, that enables them to store their products until the market situation improves. This makes them reliant on middlemen for the sale of their product, and often for ice, fuel, and other fishing supplies. This reliance is exacerbated during ENSO, as the temperature of the water, as well as of the air, can be several degrees Celsius above normal. Because their catches tend to spoil much faster due to the heat, they are more desperate to offload and sell their catch as soon as they reach shore.

The tendency for increased spoilage is problematic down the production chain. While it is difficult to blame directly on ENSO, there is a tendency for a sharp increase in gastrointestinal problems during the warm weather, again, due in part to a lack of refrigeration in many parts of the country, and the accelerated growth of bacteria because of adverse climatic conditions. An extreme example of the apparent linkage between ENSO and human health is the Pan-American cholera pandemic that began off the coast of northern Peru and spread across the continent in 1991. Facilitated by the 1992 ENSO, cholera destroyed the Peruvian market for artisanal fish and rapidly spread throughout South America. (Epstein, 1993). *Vibrio cholera*, which has been isolated from phyto- and zooplankton, is directly influenced by changes in water temperature and chemistry.

Another, perhaps equally important impact results from the increase in precipitation along the northern coast of Peru, where normally arid coastal communities can be inundated by the torrential rains. While fishermen may have an abundance of products to sell, the roads and bridges may be washed out by these rains and swollen rivers, and, thus, there is no way to get their products to market. Again, as most fishing ports lack refrigeration, the products spoil due to the interruption in transportation.

Additionally, the sea level along the coast of Peru during an ENSO warm event may rise as much as 30 cm. The sea-level height is increased as a result of the eastward shift in ocean currents and the thermal expansion of the water that results from the increased ocean surface temperatures. This, in turn, can result in many more days with dangerous storm surges as waves from storms of "typical severity" have a stronger impact on the coast. ENSO-related storms, therefore, make it too risky for the small boats to navigate out of port.

5 INDUSTRIAL SUBSECTOR

The industrial subsector is characterized by a purse seine fleet of about 750 vessels with an average capacity of 225 metric tons, employing about 26,000 fishermen and plant workers. Most of these vessels target small pelagic species (primarily anchovy), of which more than 90% is processed into fishmeal, a flourlike substance high in protein that is then used throughout the world primarily as an animal feed supplement and in aquaculture farming. The majority of nonmanagement laborers are fishing fleet, fishmeal plant, and cannery workers. A significant number also work in associated industries such as net making and repair, engine repair, and shipping services. Most of the women employed in this industry work in the canneries. Throughout the period of commercial fishing, the industrial subsector has wielded strong political influence through lobbying activities and through explicit "places at the table" within the policy-making process.

ENSO warm events directly impact the anchovy stocks (there is a southern stock, which is shared with fishermen in northern Chile, and a larger north-central stock). During the onset of a warm event, the environmental conditions make the anchovy particularly vulnerable, as the pockets of cool, nutrient-rich water are reduced in area and are located near the coast. During the onset of an ENSO event, because the anchovy population becomes more concentrated near shore, they become easier to catch. The initial rise in catch is followed by a sharp decline, as the woresening conditions cause the fish to move deeper (out of the range of the nets) in search of food, and/or migrate southward in search of cooler waters. It is a combination of this spatial shift and fishing pressure that can rapidly reduce the short to midterm availability of stocks. In addition, during warm events, the reproduction and growth cycles of the anchovy are delayed and/or stunted, and in extreme cases, the fish may die because of poor environmental conditions. The impact of the reduction in anchovy population propagates through the food chain as, for example, and seabirds and marine mammals lose their food supply and are forced to dive deeper for food or to migrate southward. Widespread death of sea birds (e.g., guano birds) due to starvation is not uncommon during extreme ENSO events.

The short- and long-term impacts on the short-lived pelagic fish stocks are influenced not only by the current climate and fishing conditions but also by the state of the fish stock prior to the onset of an event, predation by other species, and other variables that are not directly ENSO related. For instance, if the stock is in good shape prior to an event, the odds increase that the stock will recover in a relatively short time period. The timing of the ENSO event may also influence the impacts. If the event peaks during the summer, as opposed to the winter, the absolute sea surface temperatures will be considerably higher, making a significant difference between slowed recruitment and reproduction versus actual survival of the organisms. However, in the case of extraordinary events like those of 1982–1983 and 1997–1998, the prior state of the stocks may be irrelevant and overwhelmed by the magnitude of change in environmental conditions.

The history of this fishery is one that illuminates the impact of ENSO events as well as the ensuing sociopolitical changes these climate shifts may spur. Prior to fishing, the anchovy stocks supported massive bird populations, whose excrement—guano—was the most effective natural fertilizer at the time. Fueled by the increased post–World War II demand for fishmeal and the collapse of the California sardine fishery (also blamed on overfishing combined with ENSO effects) (Radovich, 1981), the state-promoted Peruvian industrial fishing sector boom began in the mid-1950s and lasted until the early 1970s. In the context of weak regulations and technological advances, its catch increased to more than 12 million metric tons (primarily anchovy) by 1972. The ENSO-influenced anchovy collapse in 1973, coupled with political change in the country, led to a nationalization of the fishery, resulting in massive layoffs and a restructuring of the industry. During this epoch, the fishermen's labor union was much more influential and active, fighting against the nationalization of the sector in 1973 and then against denationalization in 1976 (Radovich, 1981).

In the 1970s and 1980s, the exploitation of sardines slowly replaced that of anchovies in the upwelling zone. The intense 1982–1983 ENSO event aided this regime shift, further reducing anchovy catches from an historic low of 118,168 metric tons in 1983 to 24,818 metric tons in 1984 tons. Sardine catches increased from 1,172,000 in 1983 to a peak of 3,398,000.

During the 1990s, however, anchovy reclaimed the ecological niche from sardine and returned to its primacy in commercial importance. Landings over the years up to the 1997–1998 ENSO event averaged more than 5,958,000tons. The 1997–1998 event, however, led to a decline in catch from over 8 million metric tons in 1996 to over 6.6 million metric tons in 1997 and less than 4 million metric tons in 1998, and almost 6 million metric tons in 1999.

During ENSO warm events, some of the fish migrate southward to relatively cooler waters. Fishmeal plants in the southern section of the country increase catch (and profits), while the plants to the north face difficult times. Some firms have adapted by building plants in the north, central, and southern parts of the coast to take advantage of the spatial fluctuations in resources. Others have decided only to operate fleets, which they can then move up and down the coast as the movement of the resource dictates. Other firms own both plants and fleets, which allows them to better hedge the supply in the face of uncertainty. A larger scale economic response to the high variability of the fish stocks by the largest fishing firms is diversification into canned fish products, agriculture, mining, and other industries.

The few years preceding the 1997–1998 ENSO event were characterized by increasing catches combined with high fishmeal prices that spurred large investments in new boats and plants (as occurred in the years prior to the 1972–1973 event), financed by loans from private banks. Overcapitalization coincided with the onset of the 1997–1998 event and despite a relatively rapid recovery of the anchovy stocks, low fishmeal prices have greatly reduced the financial recovery of the industry. In terms of labor, unlike in the early 1970s, union power has at present virtually disappeared from the fishing sector, having been replaced by the large fraction of short-term contract (as opposed to permanent) workers. The government, following an extreme neoliberal philosophy has not been willing to subsidize the industry.

Many of Peru's banks have major investments in the industrial fishing sector (at the end of 1996, the industry as a whole had borrowed over US$270 million and by the start of 1999 the industry was alleged to owe the banks approximately US$1.5 billion). As ENSO events begin to evolve, banks must make decisions on whether to make new loans. Once an event is underway, they will often refinance loans on terms based upon expectations of the upcoming quarter's catch. The banks respond to a range of information, including rumors and media coverage. Some banks hired scientists from the government agencies as consultants and co-sponsored ENSO forums to evaluate the event and its impact on the fishing sector. Despite the early awareness of the 1997–1998 event, the recent history of overinvestment in vessels and plants led to major defaults on loans by virtually all the fishing firms. This exacerbated the countries' economic crisis by adding to the unpaid debts of the agriculture and mining industry, and several large banks were forced to close down or merge with other banks to survive. By the end of 1999, the catch had returned to fairly good levels. However, the fishing firms' economic situation could not be remedied and the industry remained in a crisis. At the time of the writing of this chapter, the government is discussing the need for a major restructuring of the industry, including reducing the fleet size and number of plants, and implementing additional regulatory mechanisms such as individual transferable quotas.

Regulations are made by the Ministry of Fisheries. In theory, its decisions are informed by the recommendations of the board of directors of the governmental scientific organization in charge of fisheries and oceanographic studies. Regulatory mechanisms include species as well as minimum size restrictions, closed seasons (*vedas*), spatial as well as gear restrictions, and statistical reporting. These are not consistently enforced as regulators face difficult decisions during the various stages of ENSO. During the 1997–1998 event, the Ministry of Fisheries implemented several short additional fishing bans to protect some resources (though they were heavily contested by the industry) and actually passed a special decree granting licenses to allow fishermen to temporarily fish jack mackerel (*Trachurus murphi*) and chub mackerel (*Scomber japonicus*) with sardine nets (smaller size mesh than normally allowed) during this event. Thus, there is a constant trade-off between trying to conserve the species, preserve jobs, and appease the politically powerful industrial interests. It is in this context that ENSO forecasts can become a key factor in public-sector policy formation.

6 POLICY IMPLICATIONS OF CLIMATE INFORMATION

In the previous sections, we identified some of the positive and negative socioeconomic effects and adaptations by members of the Peruvian fishing sector to ENSO events. At the artisanal level, this may include gear switching or migration to take advantage of changing marine resources. At the industrial level, many firms have adapted by diversifying, including moving into canning as well as fishmeal, but also into other industries. Some industrial fishermen and plant workers have second jobs, or small shops, often run out of their homes by family members. This helps to carry them through the leaner fishing seasons, as there are no laws that guarantee a minimum wage or labor security during poor fishing periods. Banks respond by altering their loan policy (usually in unfavourable ways) and by refinancing existing loans. Scientific institutions adapt by increasing their monitoring efforts, as well as using the event itself to lobby for more funding from the central government. Regulatory agencies may increase vigilance, change gear restrictions, or in some cases alter quotas.

Secondary impacts on the sector include the increased spoilage and health effects induced by the high air temperatures and transportation problems caused by intense rains. On a macroeconomic level, ENSO events favourably affect the growing, marketing, and sale of soy bean products in other parts of the world. Soymeal is the main competitor with fishmeal as an animal feed supplement. Buyers chose between these two meals, based on their relative prices.

Given the range of impacts and adaptations that are made in response to climate variability, combined with the substantial improvement in understanding and predicting the ENSO phenomenon, one can imagine more efficient political and private-sector decision-making possibilities. Observations from the 1997–1998 ENSO highlight some of the challenges in policy formation based on climate forecast and information. For a thorough treatment of this point using the case of fisheries and food security, see Broad et al. (2002) and Broad and Agrawala (2000). These difficulties stem from two primary sources: (1) scientific uncertainty and (2) societal constraints on the use of climate information.

Scientific Uncertainty

One of the primary constraints is that climate forecasters have relatively poor skill in predicting sea surface temperature (SST) and thermocline depth near coastal areas due to the steep gradients in oceanographic as well as bathymetric properties, compared to the open ocean areas. Thus, it may be difficult for regulators to plan their closed seasons and scientific cruises with much precision because of this temporal and spatial uncertainty. One of the factors preventing the more accurate prediction of ENSO impacts on the coastal ecosystem of Peru is the lack of regional scale (e.g., coastal) observations. Enhanced observations, combined with the targeted training and further model development, would aid in the monitoring and prediction of changes in coastal environmental conditions. Further, even if there were perfect climate forecasts (an impossibility), there remains the difficult step of incorporating such information into fisheries models (Glantz, 1979).

In addition to climate there are numerous variables that impact fish populations, including fishing pressure. Year-to-year and multitimescale variability in stock abundance during non-ENSO years is still not well understood (Bakun, 1996; Sharp and Csirke, 1983; Kawasaki et al., 1991; Lluch-Belda et al., 1992). One example of how this uncertainty is played out in real decision making is highlighted by the two options that were being debated during the 1997–1998 ENSO event. On the one hand, at the start of the event, conservationists called for a halt to all fishing in order to preserve as many anchovy as possible in the hope of preventing a future total collapse of the fish population. A complete ban on fishing, of course, has massive social implications, including widespread unemployment, loss of export dollars, and ensuing political unrest. This regulatory move may be rationalized by claiming that it is an attempt to ensure resources for current as well as for future generations. On the other hand, as it became clear that an extremely warm event was underway, some argued, not without reason, that fishing regulations should be removed and the fishermen should be allowed to catch what they can before all the fish disappear for "natural" reasons. Again, an arguably "rational" position, albeit self-serving, based on past experience during extreme ENSO events.

To decide the best option between these two possibilities (or a combination thereof), there is a need for more reliable forecasts as well as improved fisheries population dynamics models. Only recently have meteorologists, biological oceanographers, and fisheries biologists begun to join forces on these problems. During the 1997–1998 event, the Peruvian authorities followed a more politically viable middle ground, by exercising what can be considered "adaptive management" practices, i.e., adjusting their standard regulatory measures based on increased observations and sampling, allowing them to set their closed fishing periods with greater flexibility in response to a changing environment and political pressure.

Societal Constraints

Members of the fishing sector have different, often conflicting, goals. For instance, most industry members want to maximize their short-term profits; regulators want to conserve the resource for future generations, but also want to keep their jobs (as they recognize the power of the industry to get them fired); and banks want positive returns on their investment. These conflicting goals lead to the "political" use of the scientific information such as climate forecasts. For example, during the 1997–1998 ENSO event, those industrial fishing firms that had many outstanding loans tried to coerce some scientists to claim that it was going to be a strong ENSO with a very negative impact on the fisheries. Their logic was that the banks would then be willing to refinance under more favorable conditions or temporarily suspend interest payments. In contrast, firms waiting for loans pressured scientists to downplay the potential severity of the event in order not to scare the bankers out of making the loans for fear of a collapse and several years of low catches. In addition, two of the largest firms had begun issuing bonds immediately prior to the event and were afraid of scaring off potential investors with doomsday predictions of a fish stock collapse.

Attempts to sway economic decisions based on expectations of the evolution of the ENSO event were played out in the media, which also capitalized on the sensational aspect of a looming "disaster." The media also served as the venue for competing forecasts issued by different Peruvian scientific institutions, each one of which was vying to be the voice of authority on the event. A successful forecast could help them to get funds from the state to further research and monitor the event. With public servant salaries relatively low, individual scientists were tempted by industry's incentives—often negotiated under the table—to interpret uncertain information for the benefit of industry objectives.

Another constraint on the management of the Peruvian fisheries in light of the 1997–1998 ENSO event results from geopolitical border politics. Peru shares the southern anchovy stock with Chile, and during ENSO warm events, the stock tends to migrate further southward into northern Chilean waters. As the Peruvian quota was reached in the south, there was extreme pressure put on the regulators *not* to halt fishing in the south, the argument being that Peru should not stop fishing, since the fish were just going to go to their Chilean competitors. After much heated debate, and being accused of antinationalist tendencies, the Peruvian authorities did ban fishing in the south for a period of time, although it was only after increasing the original quota.

When considering the potential impact of forecasts, an additional factor that arises is that some groups may be more susceptible to negative impacts of climate forecasts and *mis*forecasts than others. For example, a small-scale fisherman who normally fishes close to shore and has little personal savings may receive a forecast of an ENSO event and decide to sell his gillnet and buy a longline net in anticipation of the arrival of tropical species. If the event does not occur with the anticipated intensity, however, he is left with useless gear that no one will want to buy. An example from the industrial sector may be that, if owners have prior information that fishing may be poor in the upcoming months, they may fire plant workers in advance in order to reduce their potential losses. For a discussion of who may benefit from climate forecasts at the expense of others, see Pfaff et al. (1999).

These scenarios suggest that the way climate information is disseminated plays a critical role in how society makes use of it. A forecast provider should be aware of issues of equity when making a forecast publicly accessible. Merely putting information on the Internet, for instance, may allow access to only a limited few in Peru, such as the industrial owners and managers. In contrast, many of the rural fishing villages receive information via short-wave radio, which presents a different dissemination challenge for forecast providers. Groups also have differing capacities to understand probabilistic forecasts, thus necessitating that varied types of training and education accompany the distribution of these forecasts.

We have entered a new phase in meteorology, where operational climate forecasts are being generated, and much effort is being put into disseminating this information in the United States and abroad. If any country stands to gain from this information, it is Peru, a nation directly impacted by ENSO-driven interannual climate variability. Maximizing the societal value of this information, however, necessitates communication and cooperation by those generating the predictions, distributing the forecasts, and the end users of this information. Only this approach can assure the equitable distribution of information in the proper form, with the proper training, to various decision makers at different socioeconomic strata of society.

REFERENCES

Arntz, W., A. Landa, and J. Tarazona, *Boletín: "El Niño": Su Impacto en la Fauna Marina* (Volumen Extraordinario), IMARPE, Callao, 1985.

Bakun, A., *Patterns in the Ocean: Ocean Processes and Marine Population Dynamics*, California Sea Grant College System, 1996.

Barber, R. T., and F. P. Chavez, Biological consequences of El Niño, *Science*, *222*, 1203–1210, 1983.

Bjerknes, J. A possible response of the atmospheric Hadley Circulation to equatorial anomalies of ocean temperature, *Tellus*, *8*, 820–829, 1966.

Bjerknes, J., Atmospheric teleconnections from the equatorial, *Monthly Weather Rev.*, *97*, 164–172, 1969.

Broad K., and S. Agrawala, The Ethiopia food crisis: Uses and limits of climate forecasts, *Science*, *289* (8), 1693–1694, 2000.

Broad, K., et al., Effective and equitable dissemination of seasonal-to-interannual climate forecasts: Policy implications from the Peruvian fishery during El Niño 1997–98, *Climatic Change*, 1–24, 2002.

Carr, M.-E., and K. Broad, Satellites, society, and the Peruvian fisheries during the 1997-98 El Nino, in D. Halpern (Ed.), *Satellites, Oceanography and Society*, Elsevier Science B.V., Amsterdam, 2000, pp. 171–191.

Chavez, F. P. et al., Biological-physical coupling in the central equatorial Pacific during the onset of the 1997-98 El Niño, *Geophys. Res. Lett.*, *25*, 3543–3546, 1998.

Epstein, P. R., Algal blooms in the spread and persistence of cholera, *BioSystems*, *31*, 209–221, 1993.

Glantz, M. H., Science, politics, and economics of the Peruvian anchoveta fishery, *Marine Policy*, 201–210, 1979.

Kawasaki, T., S. Tanaka, Y. Toba, and Taiguchi, *Long-Term Variability of Pelagic Fish Populations and Their Environment: Proceedings of the International Symposium*, Pergamon, 1991.

Lluch-Belda, D., et al., Sardine and anchovy regime fluctuations of abundance in four regions of the world oceans: A workshop report, *Fish. Oceanogr. 1*, 339–347.

McPhaden, M. J., Genesis and evolution of the 1997–98 El Niño, *Science*, *283*, 950–954, 1999.

Pfaff et al., Who benefits from climate forecasts? *Nature*, *397* (25), 645–646, 1999.

Philander, S. G., *El Niño, La Niña, and the Southern Oscillation*, Academic, New York, 1990.

Quinn, W. H., V. T. Neal, et al., El Niño occurrences over the past four and a half centuries, *J. Geophys. Res.*, *92* (14), 1449–1461, 1997.

Radovich, J., The collapse of the California sardine fishery: What have we learned? in M. H. Glantz and J. D. Thompson (Eds.), *Resource Management and Environmental Uncertainty: Lessons from Coastal Upwelling Fisheries*, Vol. 11, Wiley, New York, 1981, pp. 107–137.

Rodbell, D. T., et al., An ~ 15,000-year record of El Niño–driven alluviation in Southwest Ecuador, *Science*, *283*, 516–520, 1999.

Sharp, G. D., and J. Csirke, *Proceedings of the Expert Consultation to Examine Changes in Abundance and Species Composition of Neritic Fish Resources*, Food and Agriculture Administration, San Jose, Costa Rica, 1983.

Sharp, G. D., and D. R. McLain, Fisheries, El Niño–Southern Oscillation, and upper-ocean temperature records: An Eastern Pacific example, *Oceanography*, *6*, 13–22, 1993.

CHAPTER 45

DROUGHT IN SOUTH AFRICA

COLEEN VOGEL

1 INTRODUCTION

Drought is a recurrent phenomenon in Africa with occurrences noted over several centuries, varying in spatial extent and severity (Glantz, 2001; Nicholson, 1978; 1989). The impacts of drought are diverse occurring at several scales including at household, national, and regional levels. When droughts are coupled with poor mitigation strategies and lack of preparedness, accentuated periods of hardship can occur particularly for poor, vulnerable, urban, and rural areas (e.g., Glantz and Katz, 1985; Vogel, 1995; Davies, 1996; Downing et al., 1996). Dry spells, heavy rains, disruption to commercial farming, depletion of grain reserves and increases in staple food prices are some of the factors contributing to food insecurity in Southern Africa during 2002 (World Food Programme, 2002). Drought is therefore a multifaceted phenomenon, the consequences of which are the result of complex human and biophysical interrelationships. In this chapter these interactive dimensions of drought are traced for South Africa.

2 BIOPHYSICAL DIMENSIONS OF DROUGHT IN SOUTH AFRICA

Several factors interact to produce the climate and weather of South Africa (for good overviews see Tyson, 1986; Mason and Jury, 1997; Lindesay, 1998; Tyson and Preston-Whyte, 2000). The physical location of the country in the subtropics shapes climate and weather (Fig. 1). The local orography and landscape features of the country (most of the country is situated on a plateau approximately 1500 m above mean sea level) and sea surface temperatures (SSTs) of the cooler Atlantic and warmer Indian Ocean also modulate and influence local weather and climate.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

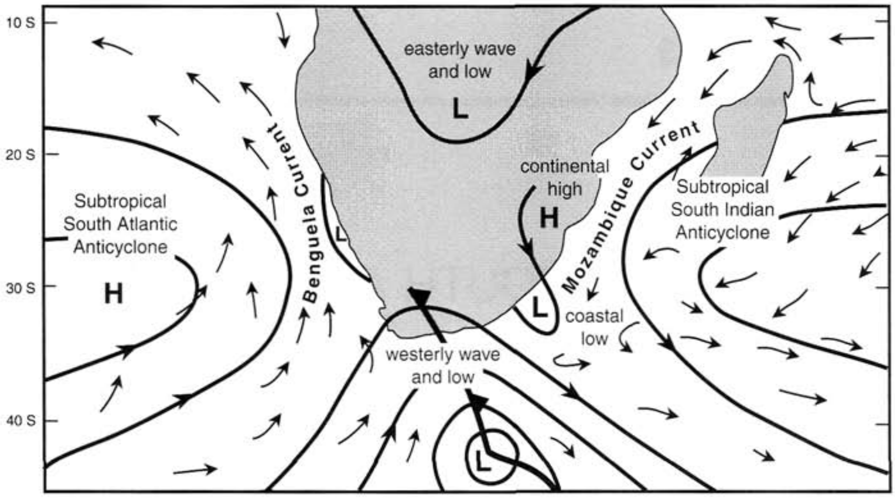


Figure 1 Some important features and controls of the atmospheric circulation over southern Africa (modified after Tyson and Preston-Whyte, 2000, with kind permission, Oxford University Press, Cape Town).

The subtropical high-pressure belt (e.g., the South Atlantic High and the South Indian Anticyclone) is usually associated with dry, stable atmospheric conditions that can prevail over much of the country for several days of the year (Tyson and Preston-Whyte, 2000). Tropical easterlies also affect southern Africa throughout the year (Tyson and Preston-Whyte, 2000). Sources of atmospheric moisture from the warmer tropical regions interact with colder drier air from the southwest to provide the setting for weather producing systems. The interaction between warm, moist easterly air and drier, colder westerly air produce convergence zones, cloud bands, and rainy spells for days across the country. The variation and location of these cloud-band convergence zones over the southern Africa–Indian sector in summer has important implications for rainfall over the country (Tyson and Preston-Whyte, 2000; Harrison, 1984a).

Wider-scale interactions between oceans (sea surface temperatures) and the atmosphere also influence the variability of rainfall (Jury, 1995; Jury et al., 1996; Mason and Jury, 1997), both locally and on a larger scale. Changes in the sea surface temperatures that are linked to atmospheric pressure fields and circulation have been shown to influence local weather and climate (Tyson and Preston-Whyte, 2000). Warmer and colder phases of water in the Pacific Ocean, for example, may influence the climate of the country by enhancing or suppressing local circulation systems. As will be shown below, these changes in circulation can become pronounced (e.g., during an El Niño or La Niña) and can produce periods of extremes in temperature and rainfall over the country (Lindesay, 1998).

The combination of the biophysical features outlined above configures the country into arid and semi-arid areas in the west of the country, becoming wetter as one

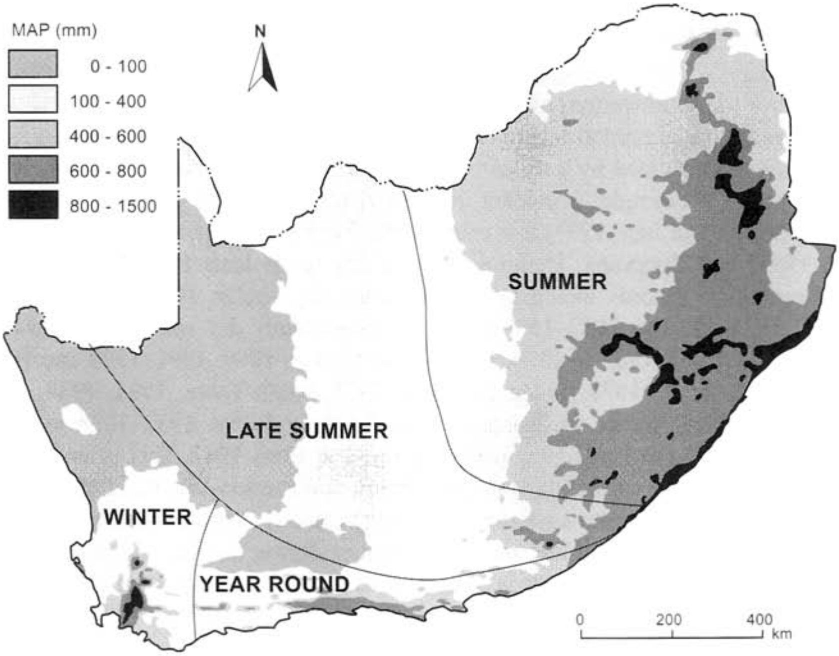


Figure 2 Spatial distribution of mean annual rainfall in South Africa (after Fox and Rowntree, 2000, with kind permission, Oxford University Press, Cape Town).

progresses eastward (Fig. 2). The surface aridity of much of the country is a function of the variability of the rainfall* as well as the relationship between evaporation and precipitation (Lindesay, 1998; Tyson, 1986). Evaporation usually exceeds precipitation over almost all of southern Africa with the deficit being greatest in the west, particularly in western South Africa and neighboring Namibia and Botswana. As a consequence of the topography and rainfall distribution, the natural availability of water is very unevenly distributed across the country with more than 60% of the flow arising from only 20% of the land area (Department of Water Affairs and Forestry, 1997; Schulze et al., 2001). Large bulk water transfer schemes have been established to ensure water supply from these wetter to drier areas (Abrams, 1997).

3 RAINFALL VARIABILITY

Rainfall in southern Africa varies temporally with annual, seasonal, and daily variations in the amount of rainfall received. Southern Africa, for example, experiences a high level of intraannual and interannual rainfall variability (Tyson, 1986). Rainfall occurs in the summer months across the central and northeastern parts of the country

* Precipitation usually includes the deposition of water in solid or liquid form including rain, dew, snow, hail (Goudie, 1994). In this chapter rainfall will be the main focus of the discussion.

with a winter rainfall area predominating to the southwestern part of the country. Summer rainfall is usually characterized by frequent thunderstorms over much of the central parts of the country, and winter precipitation is usually the result of frontal rains over the southwestern areas.

Assessments of rainfall records for South Africa indicate that interannual rainfall has been characterized by a series of wetter and drier years. An approximately 18-year cycle, for example, is evident over much of the summer rainfall area (Tyson, 1986; Mason and Jury, 1997) that extends into Zimbabwe (Torrance, 1972; Ngara et al., 1983) and Botswana. Definite wet and dry spells have been identified from meteorological records including the following dry spells: 1905–1906 to 1915–1916; 1925–1926 to 1932–1933 (the most consistently dry spell), 1944–1945 to 1952–1953; 1962–1963 to 1970–1971; 1980–1981 to 1990; 1991–1992 and 1994–1995 (Tyson et al., 1975; Tyson and Dyer, 1978, 1980; Tyson, 1981, 1984, 1986; Lindesay, 1998). Wetter spells occurred in the early 1920s, 1933–1934 to 1943–1944, 1953–1954 to 1961–1962 and 1971–1972 to 1980–1981. This latter wet spell was also the most persistently wet spell (Tyson and Preston-Whyte, 2000). Though drier spells tend to exhibit a greater spatial homogeneity than wet spells, the impression should not be gained that rainfall anomalies during the wet and dry spells are spatially uniform. Rather there is marked interannual and spatial variability in the distribution of wetter and drier conditions (Tyson, 1986; Lindesay, 1998). Available evidence suggests this pattern of wet and dry years has generally persisted since at least 1840 (Vogel, 1989).

4 CAUSES OF DROUGHTS IN SOUTH AFRICA

Having described the pattern of rainfall-producing systems in the country, attention now focuses on some of the causes of periods of insufficient rainfall and droughts. The forcing mechanisms for droughts are well documented for the southern African region (see, e.g., Tyson, 1986; Mason and Jury, 1997; Lindesay, 1998; Tyson and Preston-Whyte, 2000). Much of the recent research in the country has focused on the prolonged droughts of the last two decades (namely the 1980s and 1990s) Tyson and Dyer, 1978; Dent et al., 1987; Jury and Levey, 1993; Alexander, 1995; Mason and Jury, 1997; Lindesay, 1998; Landman and Mason, 1999).

Rainfall variability for the country has been associated with atmospheric circulation configurations and interchanges in easterly and westerly circulations, the interactions between tropical and temperature systems, and the variation in pressure patterns over Marion and Gough Island (summarized in Tyson, 1986; Lindsay, 1998). Prolonged heat waves and droughts are linked, in most cases to the predominance of prevailing anticyclonic circulation over the country. Longer periods, such as extended wet spells, for example, are usually caused by an invigoration of tropically induced circulation disturbances forced by tropical easterlies whereas the extended dry spells usually occur with an expansion and increase of westerly disturbances. In the latter case, summers become drier in the summer rainfall region (Tyson and Lindesay, 1992). Aspects of this circulation patterning during wet and dry spells have been shown over long time scales, possibly during late Interglacial and mid-Holocene warming phases (Partridge, 1997).

Much of the work on seeking causal mechanisms for rainfall variability has focused on the relationship between rainfall and the circulation parameters within the region, such as changes in the intensity and positioning of pressure systems (Tyson, 1981, 1984; Miron and Lindesay, 1983; Taljaard, 1986, 1989). Adjustment in factors such as cloud bands (important conduits of energy and momentum into the subcontinent) and ocean temperature (Walker and Lindesay, 1989; Jury, 1995) have been shown to also induce rainfall-related changes over the subcontinent and more specifically over South Africa (Harrison, 1984a, b, c; Lindesay, 1988; Mason, 1990).

More recent work has also increasingly focused on the global-scale forcing mechanisms of rainfall and linkages to atmospheric circulation changes across the Southern Hemisphere. The *Southern Oscillation* (the climatic oscillation between warm and cold periods in the tropical Pacific) and *sea surface temperature* (SST) in the South Atlantic and South Indian Oceans, have become key foci of climate research in recent years (e.g., Tyson, 1986; Lindesay, 1988, 1998; Harrison, 1986; Jury et al., 1996; Mason and Jury, 1997; Landman and Mason, 1999). Aspects of these, relating specifically to drought, are highlighted below.

The Southern Oscillation is a major forcing mode of the interannual circulation variations over southern Africa (Lindesay, 1988; Allan et al., 1996). One mechanism by which the Southern Oscillation signal is transmitted to southern Africa is via the Indian Ocean Walker Circulation. High- and low-phase years of this circulation (known as El Niño Southern Oscillation, ENSO) have been shown, furthermore, to influence rainfall over southern Africa (e.g., Lindesay, 1988, 1998). Briefly, low-phase years of the El Niño Southern Oscillation are accompanied by reductions in heat release and convection over tropical southern Africa caused by adjustments in the Walker Circulation and a retarded number of tropical-temperate troughs across South Africa. Fewer cloud bands occur and rainfall is diminished (Harrison, 1986). The reverse essentially occurs for high-phase or wetter years (Fig. 3). This general association between the Southern Oscillation phase changes and South African rainfall has been shown to exist in both present and preinstrumental rainfall records in South Africa from at least 1820 (Lindesay and Vogel, 1990).

An integral part of the Southern Oscillation is the role that sea surface temperatures play in modulating the occurrence of low- and high-phase changes and associated atmospheric circulation interactions. Although not explicitly direct, relationships between sea surface temperature anomalies, atmospheric circulation, and rainfall have been shown to exist in several instances (Glantz et al., 1987; Ogallo et al., 1988; Nicholson and Entekhabi, 1986; Mason, 1990; Walker, 1990) including relationships between above-average sea surface temperatures in the Benguela area and dry years such as 1982–1983 (Walker et al., 1984; Philander, 1986).

ENSO warm events have been associated with drought, resulting in a variety of impacts over much of southern Africa (e.g., Ogallo, 1987; Enfield, 1989; Cane et al., 1994). The 1982–1983 ENSO event, for example, served to exacerbate the prevailing dry conditions in much of the subcontinent (Bhalotra, 1985; Dent et al., 1987; Taljaard, 1989). The rainfall-producing systems of the subcontinent were displaced

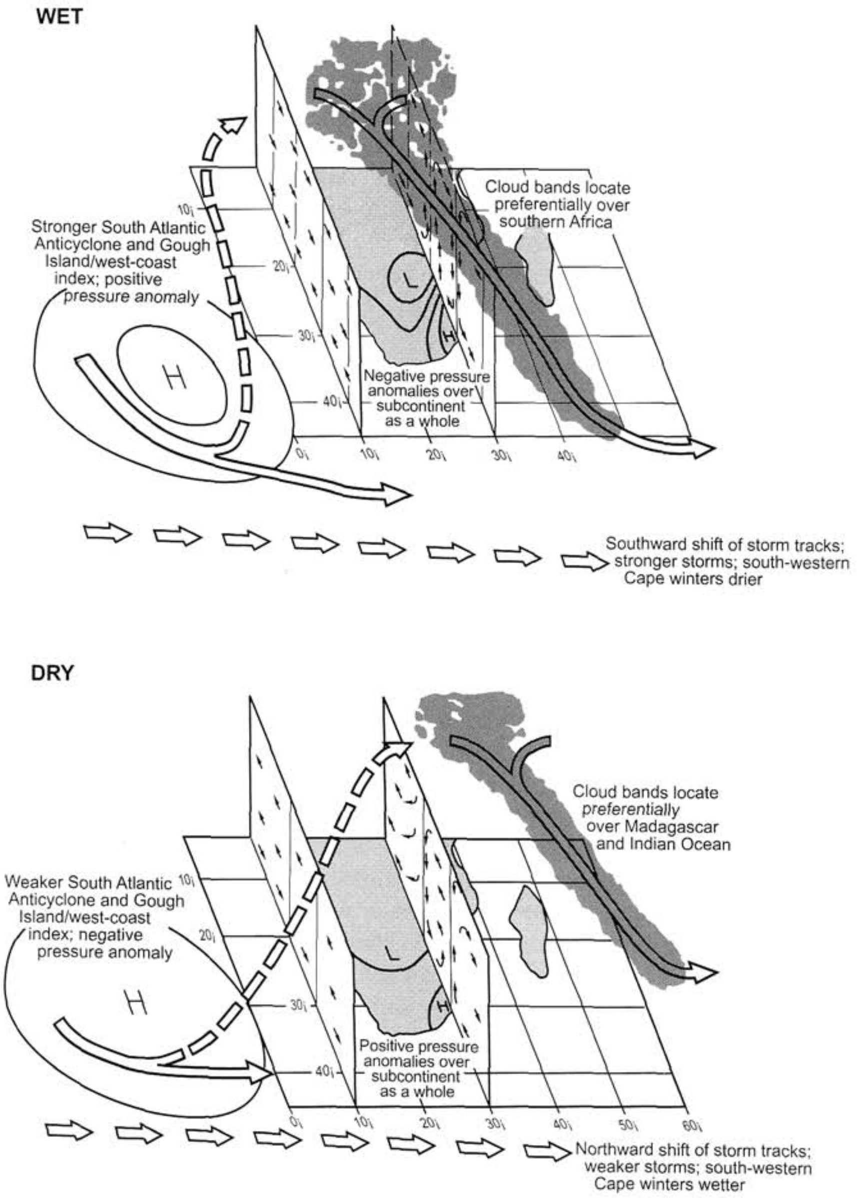

WET
Stronger South Atlantic Anticyclone and Gough Island/west-coast index; positive pressure anomaly
Cloud bands locate preferentially over southern Africa
Negative pressure anomalies over subcontinent as a whole
H
L
H
10°
20°
30°
40°
0°
10°
20°
30°
40°
Southward shift of storm tracks; stronger storms; south-western Cape winters drier
DRY
Weaker South Atlantic Anticyclone and Gough Island/west-coast index; negative pressure anomaly
Cloud bands locate preferentially over Madagascar and Indian Ocean
Positive pressure anomalies over subcontinent as a whole
H
L
H
10°
20°
30°
40°
0°
10°
20°
30°
40°
50°
60°
Northward shift of storm tracks; weaker storms; south-western Cape winters wetter

Figure 3 Model of the circulation changes over southern Africa during wet and dry spells (after Tyson and Preston-Whyte, 2000, with permission, Oxford University Press, Cape Town).

eastward during this time (Fig. 3) (Harrison, 1983; Tyson, 1986; Lindesay, 1988; Muller and Tyson, 1988).

The recent ENSO event of 1997–1998 was a very powerful one with anomalously high SSTs and, thus, indications for a possible drought (of similar magnitude to that of 1982–1983); interventions were planned, both locally and elsewhere (see Thomson et al., 1998; Mason et al., 1999; and NOAA-OGP, 1999). The El Niño impact was reduced in some parts of southern Africa and appears to have been modulated by temperatures in the Indian and Arabian Sea (Landman and Mason, 1999). Despite the strong linkage between ENSO and rainfall in parts of southern Africa, it must be remembered that not all droughts are associated with ENSO. There thus remain areas of uncertainty and predictive skill in seasonal forecasting for the country because of the complex interactions between the oceans, atmosphere, and land (Mason et al., 1994, 1996; Landman and Mason, 1999).

In summary, current research on the causes of droughts in South Africa indicates that rainfall over the country is influenced by a number of interactive mechanisms, details of which are still being examined (Mason and Jury, 1997; Landman and Mason, 1999; Mason and Tyson, 2000). The interaction between tropically and extratropically sourced weather systems, when combined with upper-air tropospheric dynamics and convectively unstable air masses, usually results in high possibilities of rainfall. Extreme drought events in southern Africa are the result of a number of atmospheric circulation interactions (Tyson, 1986; Lindesay, 1998) with some droughts and periods of reduced rainfall, for example, connected with ENSO events.

5 CLASSIFYING DROUGHTS

With this background, the focus narrows to consider the character of droughts and drought impacts in South Africa in more detail. Defining what is meant by drought is not an easy task. Droughts can be classified as meteorological, agricultural, hydrological, or sociological (Wilhite and Glantz, 1985; Dent et al., 1987; Erasmus, 1987; Bruwer, 1989, 1990). Drought, moreover, is also a relative rather than an absolute condition. Droughts can be described as being either an agricultural drought (a condition where soil moisture is depleted such that yields are considerably reduced) or a hydrological drought (actual water supply being less than the minimum required for normal operations) (Wilhite and Glantz, 1985).

In South Africa, drought is broadly defined as occurring when 75% or less of normal precipitation is received (Laing, 1992), being classed as severe if it extends over two consecutive seasons. Other indices such as the Palmer Drought Severity Index (PDSI), include inputs of soil moisture, runoff, evaporation, and temperature (Zucchini and Adamson, 1984; Erasmus, 1987; Wilhite, 1987; Bruwer, 1990). Classification of a dry spell and/or drought depends on the definition, criteria, and statistical methods used.

The timing of rainfall is also critical in determining the progression and consequences of a drought situation. Although most of the country, for example, experienced good rainfalls in 1987–1988 and 1988–1989, the individual rain events were short, torrential downbursts. As these incidences show, timing and adequate rainfall, balanced against evaporation and infiltration, determine the amount of rainfall available for surface and groundwater usage. Opportune rainfall is also essential for crop growth, and several improvements in determining drought severity and crop response have been developed including the Agricultural Catchments Research Unit (ACRU) model (Schulze, 1984, 2000; Schulze et al., 2001) and the Crop Environment Resource Synthesis (CERES-MAIZE) model (Du Pisani, 1987).

Using such methods, together with other descriptive criteria including rainfall over three seasons, grassland condition, availability of water for stock, stock condition/deaths, availability and volume of fodder purchased (Bruwer, 1989, 1990), one is able to divide the country into areas that are more or less drought prone. Records for a 30-year period, for example (1936 to 1986), show that 27% of the country has been declared a disaster drought area for more than 50% of the time (Bruwer, 1989, 1990).

Periods of drought, usually spanning at least 2 or more seasons, occurred in the late 1920s and early 1930s, much of the 1960s and 1980s, and more recently in the early and mid-1990s. These drought periods have brought with them and have compounded a variety of problems (including access to water in rural poor areas and water availability and use in agricultural and industrial sectors). Droughts have also highlighted several prevailing factors that predispose certain areas and groups to heightened drought risk (e.g. poor water infrastructure, land degradation, globalization) (Vogel, 1994; Scoones et al., 1996; Benson and Clay, 2000; Dilley, 2000).

6 IMPACTS OF DROUGHTS IN SOUTH AFRICA

Drought impacts, however, are not the result only of insufficient rainfall or searing temperature. In most cases, drought impacts are the outcome of the interaction of a number of social and other human factors that can heighten the "vulnerability" of communities and various exposure units (e.g., vegetation) and reduce "resilience" of society and ecosystems to the natural hazard (Dilley, 2000; Vogel et al., 2000). As a result of these components of drought, a number of impacts are recorded.

The scale of these impacts also varies and can be tracked at various levels (e.g., regional, national, community, and household) of agricultural production. For example, production declined as a result of the 1980s and 1990s droughts in southern Africa. Harvest failures of between 30 and 80% below-normal across the Southern African Development Community (SADC) region were recorded. Cereal production in the SADC countries dropped to less than 50% of the annual requirement in 1992, and the cost of imported food to the region rose to approximately $4 billion (Hulme, 1996). Drought related to the 1982–1983 El Niño cost nearly US$1 billion in direct damages with an estimated US$350 million spent on famine relief (1983 prices) in southern Africa (International Federation of Red Cross and Red Crescent Societies, 1999). The economic loss to Africa's agricultural sector in the early 1990s drought was estimated at US$7 billion (1992 prices)–an estimated 20 times the value of 1993 World Bank loans to sub-Saharan agriculture (International Federation of Red Cross and Red Crescent Societies, 1999). More recently, the combined influence of drought, floods, economic instability, and HIV/AIDS threatened the food security of millions in southern Africa (WFP, 2002).

At a national level, drought ripples across sectors and impacts on a range of activities. On a national scale, drought in South Africa results in a reduction in the yield of the maize crop with yields falling to below 1 ton per hectare (Fig. 4). The outward effects triggered by drought range from its impact on agriculture's contribution to the gross domestic product (GDP) to a host of other impacts such as food supply, employment opportunities, and a number of forward and backward linkages associated with the agricultural sector (Ballard, 1986; Van Zyl et al., 1987; Van Zyl and Nel, 1988). The declining agricultural yields associated with the droughts of the 1990s, for example, negatively affected GDP growth by between 0.5 and 2% (Mather and Adelzadeh, 1997).

The human consequences and across sectors and groups, are difficult to quantify accurately. For the agricultural sector the occurrence of drought, together with changes in agricultural policy, provision of farmer loans and other economic factors, can combine to heighten the impacts on the sector. During the recent severe drought of the early 1990s, for example, it was estimated that 50,000 jobs would be lost in the agricultural sector (with a further 20,000 in related sectors) and about 250,000 in total (families included) would be affected (AFRA, 1992; Adams, 1993; Van Zyl, 1993). Crop failures occurred for both commercial and smaller-scale farmers (Adams, 1993) and water levels in several of the major dams were less than two-thirds their normal capacity (Fig. 5). Faulty and poorly maintained water infrastructure further aggravated the precarious water situation.

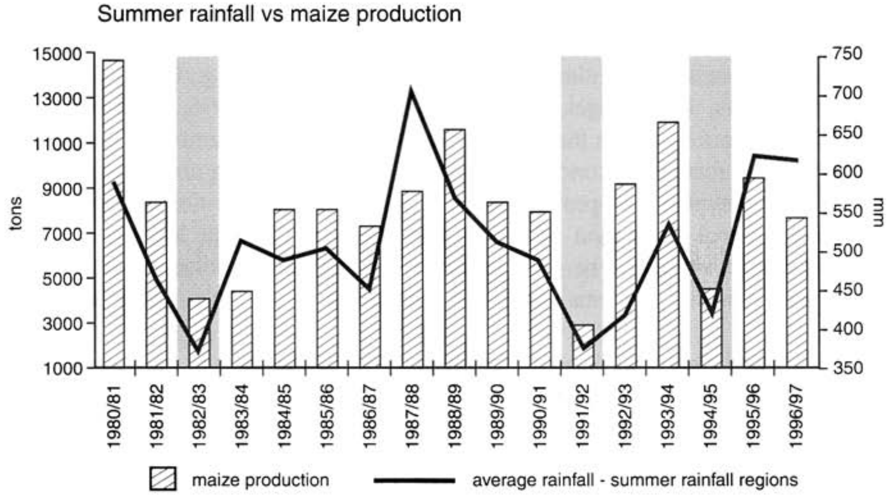


Figure 4 Summer rainfall versus maize production (modified after McClintock, 1997, with kind permission, UBS Warburg, formerly SBC Warburg, Johannesburg).

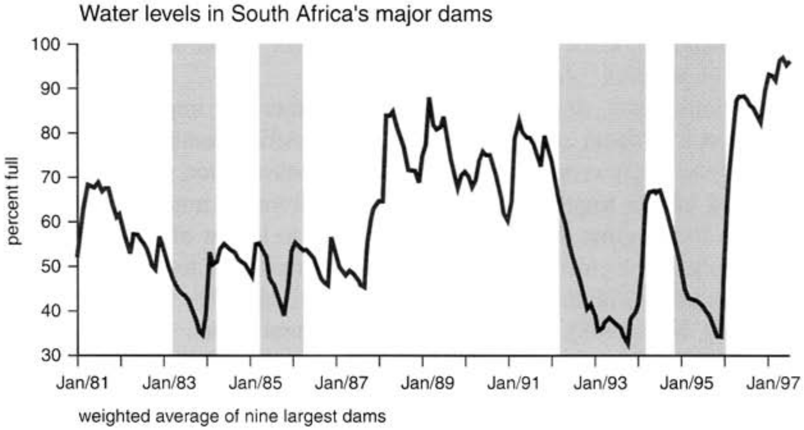


Figure 5 Water levels in South Africa's major dams (modified after McClintock, 1997, with kind permission, UBS Warburg, formerly SBC Warburg, Johannesburg).

Drought is not, therefore, the fundamental problem in sub-Saharan Africa. Drought needs to be viewed together with a host of other hazards and realities: including HIV/AIDS, violence and conflict, growing disparities between rich and poor, failing economies, struggles over land, water, and poverty. Drought indeed often merely uncovers the African development crisis and allows glimpses of harsh daily realities. At a local scale, the impacts of droughts are often hidden "costs" escaping detailed quantification. These include the stripping of household assets used to procure a livelihood (stock and crops are reduced, the price of water and of basic food supplies often increases, retrenchments occur), household income, and social dislocation and disruption of local livelihood (e.g., Bratton, 1987; Adams, 1993; Vogel, 1995; Scoones et al., 1996).

Several assessments of vulnerability to droughts in Africa (e.g., Glantz and Katz, 1985; Chambers, 1989; Vogel, 1995; Jallow, 1995; Davies, 1996; Downing et al., 1996) have therefore shown that droughts act together with a number of underlying factors to exacerbate local conditions. In southern Africa, during non-drought years, the baseline prevalence of problems related to inadequate nutrition is "normally" low (International Federation of Red Cross and Red Crescent Societies, 1999). Drought cannot, therefore, be seen as the "cause" of such problems, but rather, it exacerbates existing problems associated with poverty (e.g., Abrams et al., 1992; AFRA, 1992).

Few detailed studies of drought impacts, coping responses, and mitigation at the rural-poor household level have been undertaken in South Africa (e.g., Freeman, 1984; AFRA, 1992; Vogel, 1995). Most assessments show that it is "access" and "entitlements" to resources that usually determine the magnitude of impact. Rural communities that farm and depend on the land for a livelihood often require access to irrigation, access to markets to sell stock, financial resources to procure farming equipment, etc. Failure to procure these resources, together with a severe dry spell, can result in inexorable difficulties. During the drought of the early 1990s, for example, large numbers of cattle died in several rural areas [an estimated 500,000 cattle in the former Transkei, a former "independent homeland" in the eastern part of the country (Adams, 1993)]. Such losses of cattle do not only result in less meat and milk but also severely constrain the limited household incomes derived from cattle sales. Underpinning these circumstances is the complex history of the country, which has had a major influence on who farms, owns land, and can obtain access to the resources mentioned above (Lipton et al., 1996).

7 DROUGHT MANAGEMENT AND POLICY INITIATIVES

Droughts, as shown here, are a regular feature in the tapestry of South African history but have been traditionally managed from an agricultural and conservation perspective (Union of South Africa, 1923). This focus has, however, been expanded during the past decade to include a wider group of affected communities and stakeholders. During the droughts of the 1990s, the impact of the drought on rural populations, for example, was actively monitored by various task forces that emerged from a National Consultative Committee on Drought as a result of reports of severe impacts on rural communities in the country, particularly those that had been relocated during the years of apartheid. The activities of the Drought Forum raised the profile and plight of the rural poor during droughts in South Africa and ushered in a change in drought policy (Abrams et al., 1992; Adams, 1993).

Building on the experiences of the 1990s drought, a strong mitigation focus for droughts has been fostered and is contained in the new disaster management policy, which includes a focus on the biophysical resources of the country as well as concentrating on mitigating the host of other factors that exacerbate drought impacts (*White Paper on Disaster Management*, 1999 and Bill, (forthcoming 2002)). The national policy on disasters, including drought, calls for a more proactive response to future droughts in the country. The growing official awareness and policy efforts are envisaged to lead to greater coordination with FEWS (Famine Early Warning Systems) in the wider SADC region and hence improve drought management, both locally and in the region.

Past and present experiences with the vagaries of weather and climate are prompting concerted efforts to improve preparedness for droughts. Collaborative efforts by both forecasters, atmospheric modelers and users of climate outlooks and forecasts have resulted in the formation of various climate forums that have been established throughout Africa, through consensus to improve the quality of the forecasts for the forthcoming seasons. Combining their knowledge of climate conditions in southern Africa, these experts provide users with a consensus, probabilistic assessment of the upcoming rainy season (NOAA-OGP, 1999). The South African Weather Services is actively involved in the forum and through its Research Group for Seasonal Climate Studies, three monthly mean rainfall and temperature forecasts for the country, regularly updated, are produced and issued by the Long-term Operational Group Information Centre (LOGIC). The science of forecasting seasonal rainfall and temperature for southern Africa and South Africa is well developed (e.g., Joubert et al., 1996; Mason et al., 1996; Joubert and Hewitson, 1997; Mason and Joubert, 1997; Mason, 1997; Mason and Jury, 1997). The overwhelming need, however, still remains for integrated science that incorporates the human and physical dimensions of climate variability and change so that effective mitigation strategies can be initiated and implemented.

Despite these advances, several groups, particularly the rural poor in Southern Africa, remain food insecure. One solution to avoid such situations is to ensure that forecasts are made more accessible. Others, however, argue that much more is required. The eradication of food emergencies and in cases, famines, requires more than technical capacity. Substantial political will, at national and international levels, more than has been evident to date, is needed (Devereux, 2000).

8 CONCLUSIONS

Drought is endemic to the southern African region. Changes in sea surface temperatures together with realignments of pressure systems can, in some cases, trigger a severe drought period such as occurred in the early 1980s and the 1990s. The impacts of such droughts usually result in severe constraints on food production at regional, national, and local levels impacting on GDP, commercial food supply, and water availability. Reductions in water aggravate such problems creating ripple effects that touch on several industries, activities, and communities.

Drought, however, has many facets and it is often the poorly documented cases, such as household vulnerability to drought in poor rural and urban areas, that require much more careful research. Research and understanding of the multidimensional nature of drought, the complex coping strategies in the face of drought, and mitigation strategies are required if effective mitigation and management of droughts is to occur. Research that has been conducted and available vulnerability assessments indicate that drought often unveils many "everyday" realities that are rooted in other factors such as poverty, development, and the complex interlinkages among economic, social, political, and environmental issues. Land degradation, urban and periurban growth, and the impact of HIV/AIDS are some of the factors that heighten vulnerability to the vagaries of weather and climate in Africa. The response to such problems therefore has to be a multifaceted one in South Africa and elsewhere in Africa.

REFERENCES

Abrams, L., Drought policy—water issues, The African Water Page, http://www.africanwater.org.drghtwater.htm 1997.

Abrams, L., R. Short, and J. Evans, Root cause and relief restraint report, National Consultative Forum on Drought, Secretarial and Ops Room, Johannesburg, October 8, 1992.

Adams, L., A rural voice, strategies for drought relief, *Indicator S. Afr.*, *10*(4), 41–46, 1993.

Alexander, W. J. R., Floods, droughts and climate change, *S. Afr. J. Sci.*, 91, 403–408.

Allan, R. J., J. A. Lindesay, and D. Parker, *El Nino Southern Oscillation and Climatic Variability*, CSIRO Publishing, Melbourne, Australia, 1996.

Association for Rural Advancement (AFRA), *Drought Relief and Rural Communities*, Special Report No. 9, AFRA, Pietermaritzburg, 1992.

Ballard, C., Drought and economic disasters: South Africa in the 1980s, *J. Interdiscipl. Hist.*, *17*, 359–378, 1986.

Benson, C., and E. J. Clay, Developing countries and the economic impacts of natural disasters, in Kreimer, A., and Arnold, M. (Eds.), *Managing Risk in Emerging Economies*, The World Bank, Washington, D.C., 11–21, 2000.

Bhalotra, Y. P. R., *The Drought of 1981–1985 in Botswana*, Department of Meteorological Services, Ministry of Works and Communications, Gaborone, Botswana, 1985.

Bratton, M., Drought, food and social organization of small farmers in Zimbabwe, in M. Glantz (Ed.), *Drought and Hunger in Africa: Denying Famine a Future*, Cambridge University Press, Cambridge, 1987, pp. 213–244.

Bruwer, J. J., Drought policy in the Republic of South Africa, Part I, *Drought Network News*, *1*(3), 14–16, University of Nebraska, Lincoln, USA, 1989.

Bruwer, J. J., Drought policy in the Republic of South Africa, in A. L. Du Pisani, (Ed.), *Proceedings of the SARCCUS Workshop on Drought*, June, 1989, SARCCŪS, Pretoria, 1990, pp. 23–38.

Cane, M. A., G. Eshel, and R. W. Buckland, Forecasting Zimbabwean maize yield using eastern equatorial pacific sea surface temperature, *Nature*, *370*, 204–205, 1994.

Chambers, R., Vulnerability, coping and policy in *IDS Bulletin, 20, 2: Vulnerability: How the Poor Cope*, Institute of Development Studies IDS, University of Sussex, Brighton, England, 1989, pp. 1–7 .

Davies, S., *Adaptable Livelihoods: Coping with Food Insecurity in the Malian Sahel*, Macmillan, London, 1996.

Dent, M. C., R. E. Schulze, H. M. M. Wills, and S. D. Lynch, Spatial and temporal analysis of the recent drought in the summer rainfall region of Southern Africa, *Water SA*, *13*, 37–42, 1987.

Department of Water Affairs and Forestry, *Overview of Water Resources, Availability and Utilization in South Africa*, Department of Water Affairs and Forestry, Pretoria, 1997.

Dilley, M., Climate Change and Disasters, in Kreimer, A., and Arnold, M. (Eds.), *Managing Disaster Risk in Emerging Economies*, The World Bank, Washington D.C., 45–50, 2000.

Downing, T., M. Watts, and H. Bohle, Climate change and food insecurity: Toward a sociology and geography of vulnerability, in T. E. Downing, (Ed.), *Climate Change and World Food Security*, Nato ASI Series, Global Environmental Change, 1996, pp. 183–206.

Du Pisani, A. L., The CERES-MAIZE model as a potential tool for drought assessment in South Africa, *Water SA*, *13*, 159–164, 1987.

Enfield, D. B., El Nino, past and present, *Rev. Geophys*. *27*, 159–187, 1989.

Erasmus, J. F., Drought monitoring: Using rainfall DECILES as a drought index, *Pixels and Bytes*, *5*, 42–48, 1987.

Fox, R., and K. Rowntree (Eds.), *The Geography of South Africa in a Changing World*, Oxford University Press, Cape Town, 2000.

Freeman, C., Drought and agricultural decline in Bophuthatswana, in South African Research Services, (Eds.), *South African Review II*, Ravan Press, Johannesburg, 1984, pp. 284–289.

Glantz, M. H. (Ed.), *Once Burned, Twice Shy? Lessons Learned from the 199798 El Nin˜o*, UNEP, NCAR, United Nations University, WMO, ISOR, United Nations University, Japan, 2001.

Glantz, M. H., and R. W. Katz, Drought as a constraint to development in sub-Saharan Africa, *Ambio*, *14*, 334–339, 1985.

Glantz, M. H., R. Katz, and M. Krenz, (Eds.), *The Societal Impacts Associated with the 1982–83 Worldwide Climate Anomalies*, Environmental and Societal Impacts Group, United Nations Publications Office, New York, 1987.

Goudie, A. (Ed.), *The Encyclopedic Dictionary of Physical Geography*, 2nd ed., Blackwell, Oxford, 1994.

Harrison, M. S. J., The Southern Oscillation, zonal equatorial circulation cells and South African rainfall, in *Preprints of the First International Conference on Southern Hemisphere Meteorology*, American Meterological Society, Boston, 1983, pp. 302–305.

Harrison, M. S. J., A generalized classification of South African summer rain-bearing synoptic systems, *J. Climatol.*, *4*, 547–560, 1984a.

Harrison, M. S. J., The annual rainfall cycle over the central interior of South Africa, *S. Afr. Geograph. J.*, *66*, 46–64, 1984b.

Harrison, M. S. J., Comparison of rainfall time series over South Africa generated from real data and through principals component analysis, *J. Climatol. 4*, 561–564, 1984c.

Harrison, M. S. J., A synoptic climatology of South African rainfall variations, Ph.D. thesis, University of the Witwatersand, 1986.

Hulme, M. (Ed.), Climate Change and Southern Africa: An exploration of some potential impacts and implications in the SADC region, Report commissioned by the WWF International and co-ordinated by the Climate Research Unit, UEA, Norwich, United Kingdom, 1996.

International Federation of Red Cross and Red Crescent Societies, (IFRCRCS), *World Disasters Report*, IFRCRCS, Geneva, Switzerland, 1999

Jallow, S. S., Identification of and response to drought by local communities in Fulladu West District, The Gambia, *Singapore J. Trop. Geogr.*, *16*, 22–41, 1995.

Joubert, A. M., and B. Hewitson, Simulating present and future climates of southern Africa using General Circulation Models, *Prog. Phys. Geogr.*, *21*, 51–78, 1997.

Joubert, A. M., S. J. Mason, and J. S. Galpin, Droughts over southern Africa in a doubled-CO_2 climate, *Int. J. Climatol. 16*, 1149–1156, 1996.

Jury, M. R., A review of research on ocean-atmosphere interactions and South Afrian climate variability, *S. Afr. J. Sci.*, *91*, 289–294, 1995.

Jury, M. R., and K. M. Levey, The climatology and characteristics of drought in the Eastern Cape of South Africa, *Int. J. Climatol.*, *13*, 629–641, 1993.

Jury, M. R., B. M. R. Pathack, C. J de W. Rautenbach, and J. van Heerden, Drought over South Africa and Indian Ocean SST: Statistical and GCM results, *Global Atmos. Ocean Syst. 4*, 47–63, 1996.

Laing, M., Drought update 1991–1992, South Africa, *Drought Network News*, *4*(2), 15–17, University of Nebraska, Lincoln, USA, 1992.

Landman, W. A., and S. J. Mason, Change in the association between Indian Ocean sea-surface temperatures and summer rainfall over South Africa and Namibia, *Int. J. Climatol.*, *19*, 1477–1492, 1999.

Lindesay, J. A., Southern African rainfall, the Southern Oscillation and a Southern Hemisphere semi-annual cycle, *J. Climatol.*, *8*, 17–30, 1988.

Lindesay, J. A., Present climates of southern Africa, in J. E., Hobbs, J. A. Lindesay, and H. A Bridgman (Eds.), *Climates of the Southern Continents Present, Past and Future*, Wiley, Chichester, 1998.

Lindesay, J. A., and C. H. Vogel, Historical evidence for Southern Oscillation–southern African rainfall relationships, *Int. J. Climatol.*, *10*, 679–689, 1990.

Lipton, M., F. Ellis, and M. Lipton, *Land, Labour and Livelihoods in rural South Africa*, Vol. 2: *Kwa Zulu Natal and Northern Province*, Indicator Press, University of Natal, South Africa, 1996.

Mason, S. J., Temporal variability of sea surface temperatures around southern Africa: A possible forcing mechanism for the eighteen-year rainfall oscillation? *S. Afr. J. Sci.*, *86*, 243–252, 1990.

Mason, S. J., Review of recent developments in seasonal forecasting of rainfall, *Water SA*, *23*, 57–62, 1997.

Mason, S. J., L. Goddard, N. E. Graham, Yulaeva, L. Sun, and P. A. Arkin, The IRI seasonal climate prediction system and the 1997/98 El Nino event, *Bull. Am. Meteorol. Soc.*, *80*, 1853–1973, 1999.

Mason, S. J., and M. R. Jury, Climatic variability and change over southern Africa: a reflection on underlying processes, *Prog. Phys. Geogr.*, *21*(1), 23–50, 1997.

Mason, S. J., and A. M. Joubert, Simulated changes in extreme rainfall over southern Africa, *Int. J. Climatol.*, *17*, 291–301, 1997.

Mason, S. J, A. M. Joubert, Cosijn, C. and S. J. Crimp, Review of the current state of seasonal forecasting techniques with applicability to Southern Africa, *Water SA*, *22*(3), 203–209, 1996.

Mason, S. J., J. A. Lindesay, and P. D. Tyson, Simulating drought over southern Africa using sea surface temperature variations, *Water SA*, *20*, 15–21, 1994.

Mason, S. J., and P. D. Tyson, The occurrence and predictability of droughts over southern Africa, in D. Wilhite, (Ed.), *Drought*, Vol. 1: *A Global Assessment, Routledge Hazards and Disasters Series*, Routledge, 2000, pp. 112–134.

Mather, C., and A. Adelzadeh, Macroeconomic strategies, agriculture and rural poverty in post-apartheid South Africa, paper presented for the Land and Agricultural Policy Centre, Johannesburg, 1997.

McClintock, M., *El Nino, Cloud on the Horizon, Economics, SBC Warburg*, Swiss Bank Corporation, Johannesburg, South Africa, 1997.

Miron, O., and J. A. Lindesay, A note on changes in airflow patterns between wet and dry spells over South Africa, 1963 to 1979, *S. Afr. Geogr. J.*, *65*, 141–147, 1983.

Muller, M. J., and P. D. Tyson, Winter rainfall over the interior of South Africa during extreme dry years, *S. Afr. Geogr. J.*, *70*, 20–30, 1988.

National Oceanic and Atmospheric Administration, Office of Global Programs (NOAA-OGP), U.S. Department of Commerce, An experiment in the application of climate forecasts: NOAA-OGP activities related to the 1997-98 El Nino event, Washington, D.C., NDAA/OGP, 1999.

Ngara, T., D. L. McNaughton, and S. Lineham, Seasonal rainfall fluctuations in Zimbabwe, *Zimbabwe Agri. J.*, *80*, 149–150, 1983.

Nicholson, S. E., Climatic variations in the Sahel and other African regions during the past five centuries, *J. Arid Environ.*, *1*, 3–24, 1978.

Nicholson, S. E., African drought: Characteristics, causal theories and global teleconnections, in A. Berger, R. E. Dickinson, and J. W. Kidson (Eds.), *Understanding Climate Change, Geophysical Monograph*, *52*, American Geophysical Union, Washington, DC, 1989, pp. 79–100.

Nicholson, S. E., and D. Entekhabi, The quasi-periodic behaviour of rainfall variability in Africa and its relationship to the Southern Oscillation, *Arch. Meteorl. Geophys. Bioklimatol. Ser. A.*, *34*, 311–348, 1986.

Ogallo, L. J., Impacts of the 1982–83 ENSO event on eastern and southern Africa, in M. H. Glantz, R. Katz, and M. E. Krenz (Eds.), *The Societal Impacts Associated with the 198283 Worldwide Climate Anomalies*, United Nations Publications Office, New York, pp. 55–61, 1987.

Ogallo, L. J., J. E. Janowaik, and M. S. Halpert, Teleconnections between seasonal rainfall over East Africa and global sea surface temperature anomalies, *J. Meteorol. Soc. J.*, *66*, 807–821, 1988.

Partridge, T. C., Cainozoic environmental change in southern Africa, with special emphasis on the last 200 000 years, *Progr. Phys. Geogr.*, *21*(1), 3–22, 1997.

Philander, S. G. H., Unusual conditions in the tropical Atlantic Ocean in 1984, *Nature*, *322*, 236–238, 1986.

Schulze, R. E., Hydrological simulation as a tool for agricultural drought assessment, *Water SA*, *10*, 55–62, 1984.

Schulze, R. E., Modelling hydrological responses to land use and climate change: A southern African perspective, *Ambio*, *29*, 12–22, 2000.

Schulze, R., J. Meigh, and M. Horan, Present and future vulnerability of eastern and southern Africa's hydrology and water resources, *S. Afr. J. Sci.*, *97*, 150–160.

Scoones, I., C. Chibudud, S. Chikara, P. Jeranyama, D. Machaka, W. Machanja, B. Mavedzenge, B. Mombeshpra, M. Mudhara, C. Mudsiwo, F. Murimbarimba, and B. Zirera, *Hazards and Opportunities: Farming Livelihoods in Dryland Africa: Lessons from Zimbabwe*, Zed Press, London, 1996.

Taljaard, J. J., Change of rainfall distribution and circulation patterns over southern Africa in summer, *J. Climatol.*, *6*, 579–592, 1986.

Taljaard, J. J., *Climate and Circulation Anomalies in the South African Region during the Dry Summer of 19821983* , South African Weather Bureau Technical Paper No. 21, Weather Bureau, Pretoria, 1989.

Thomson, A., P. Jenden, and E. Clay, Information, risk and preparedness: Responses to the 1997 El Nino event, Research report, DFID, ESCOR No AG1215, SOS SAHEL, London, 1998.

Torrance, J. D., Malawi, Rhodesia and Zambia, in J. F. Griffiths (Ed.), *Climates of Africa, World Survey of Climatology*, Vol. 10, Elsevier, Amsterdam, 1972, pp. 409–460.

Tyson, P. D., Atmospheric circulation variations and the occurrence of extended wet and dry spells over southern Africa, *J. Climatol.*, *1*, 115–130, 1981.

Tyson, P. D., The atmospheric modulation of extended wet and dry spells over South Africa, 1958–1978, *J. Climatol. 4*, 621–635, 1984.

Tyson, P., *Climate Change and Variability in Southern Africa*, Oxford University Press, Cape Town, 1986.

Tyson, P. D., and T. G. J. Dyer, The predicted above-normal rainfall of the seventies and the likelihood of droughts in the eighties in South Africa, *S. Afr. J. Sci.*, *74*, 372–377, 1978.

Tyson, P. D., and T. G. J. Dyer, 1980: The likelihood of droughts in the eighties in South Africa, *S. Afr. J. Sci.*, *76*, 340–341, 1980.

Tyson, P. D., T. G. J. Dyer, and M. N. Mametse, Secular changes in South African rainfall: 1880 to 1972, *Q. J. Roy. Meteorol. Soc.*, *101*, 817–833, 1975.

Tyson, P. D., and J. A. Lindesay, The climate of the last 2000 years in southern Africa, *Holocene*, *2*, 271–278, 1992.

Tyson, P. D., and R. A. Preston-Whyte, *The Weather and Climate of Southern Africa*, Oxford University Press, Cape Town, 2000.

Union of South Africa, *Final Report of the Drought Investigation Commission*, UG 49–23, Government Printers, Cape Town, 1923.

Van Zyl, J., The last straw: Drought and the economy, *Indicator SA*, *10*(4), 47–51, 1993.

Van Zyl, J., and H. J. G. Nel, The role of the maize industry in the South African economy, *Agrekon*, *27*, 10–16, 1988.

Van Zyl, J., A. Van der Vyver, and J. A. Groenewald, The influence of drought and general economic effects on agriculture: A macro-analysis, *Agrekon*, *26*, 8–12, 1987.

Vogel, C., A documentary-derived chronology for southern Africa, 1820–1900, *Climate Change*, *14*, 291–306, 1989.

Vogel, C. H., People and drought in South Africa: reaction and mitigation, in T. Binns (Ed.), *People and Environment in Africa*, Wiley, London, 1995, pp. 249–256.

Vogel, C. H., M. Laing, and K. Monnik, Drought in South Africa, with special reference to the 1980–1994 period, in D. Wilhite (Ed.), *Drought*, Vol. 1: *A Global Assessment, Routledge Hazards and Disasters Series*, Routledge, London, 2000, pp. 348–366.

Walker, N., Links between South African summer rainfall and temperature variability of the Agulhas and Benguela currents systems, *J. Geophys. Res.*, *95*, 3297–3319, 1990.

Walker, N., J. Taunton-Clark, and J. Pugh, Sea temperatures off the South African west coast as indicators of Benguela warm events, *S. Afr. J. Sci.*, *80*, 72–77, 1984.

Walker, N. D., and J. A. Lindesay, Preliminary observations of oceanic influences on the February–March 1988 floods in central South Africa. *S. Afr. J. Sci.*, *85*, 164–169, 1989.

Wilhite, D. A., The role of government in planning drought: Where do we go from here? in D. A. Wilhite, W. E. Easterling, and D. A. Woods (Eds.), *Planning for Drought: Toward a Reduction of Societal Vulnerability*, Westview Press, Boulder, CO, 1987, pp. 425–444.

Wilhite, D. A., and M. H. Glantz, Understanding the drought phenomenon: The role of definitions, *Water Int.*, *10*, 111–120, 1985.

White Paper on Disaster Management, Ministry for Provincial Affairs and Constitutional Development, Pretoria, 1999.

World Food Programme, WFP, Hunger in Southern Africa: The Unfolding Crisis, 30-05-2002, http://www.wfp.org/index

Zucchini, W. Z., and P. T. Adamson, The occurrence and severity of droughts in South Africa, report to the Water Research Commission, WRC 91/1/84, prepared by the Department of Civil Engineering, University of Stellenbosch and the Department of Water Affairs, Pretoria, 1984.

CHAPTER 46

TRANSBOUNDARY FISHERIES: PACIFIC SALMON

KATHLEEN A. MILLER AND MARY W. DOWNTON

1 INTRODUCTION

Climate variations often affect the abundance, location, or migratory patterns of fish populations. Even when a fishery is entirely contained within a single jurisdiction, these climate impacts complicate the difficult task of maintaining economically efficient and biologically sound harvest management while balancing the interests of competing harvesters. When fish stocks are harvested by more than one nation, or when they cross internal jurisdictional boundaries, the management task is further complicated by the efforts of each nation or jurisdiction to promote the interests of its own harvesters.

The faltering attempts of the United States and Canada to cooperate on Pacific salmon management illustrate the fragile nature of such cooperation and the destabilizing role that climate variations can play. These two nations have a long and rocky history of alternating between cooperative salmon conservation efforts and predatory grabs at one another's returning adult salmon. The most recent breakdown in cooperation began in 1993. For 6 years, the United States and Canada were unable to agree on a full set of salmon "fishing regimes" under the terms of the Pacific Salmon Treaty. The conflict was sparked by strongly divergent trends in the abundance of northern and southern salmon stocks and a consequent change in the balance of each nation's interceptions of salmon spawned in the other nation's rivers. Alaska's salmon harvests (i.e., northern) have experienced a remarkable sustained increase over the past two decades, while harvests of some salmon stocks in British Columbia and chinook and coho harvests in Washington, Oregon and California (i.e., southern) have fared poorly. These trends appear to be influ-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

enced by the impacts of climatic variations on stock abundance, but climate is not the only source of harvest variability. Because it is not easy to disentangle natural and anthropogenic sources of variability, the negotiation process has been complicated by differences of opinion over the biological "facts."

A new agreement, signed on June 30, 1999, may end the conflict, but it is too early to judge its likelihood of success. The Canadians remain bitterly divided over the merits of the agreement, which has been labeled a "sellout" by Canadian fishing interests, and the arrangement is still contingent on U.S. Congressional approval of $140 million for two jointly managed endowment funds to be used for scientific cooperation, stock enhancement, and habitat restoration (Culbert and Beatty, 1999).

This case exemplifies many of the problems that arise in the management of transboundary fishery resources. Each nation has the power to significantly damage the other's interests and, in the short term, each could gain by interfering with the other's harvests. In the longer term, such opportunistic competition tends to squander the potential value of the fishery, as each nation commits excessive fishing capital and labor in its attempts to capture a larger share of the available fish. The ultimate result of competitive harvesting may be a "tragedy of the commons" (Hardin, 1968) in which overfished stocks are depleted and the fishery declines, sometimes abruptly.

Nations that exploit shared fishery resources often recognize the mutually destructive effects of unconstrained competitive harvesting, and they may attempt to improve the situation by negotiating harvest management agreements. The stability of such agreements hinges on the extent to which it is easy or difficult to monitor and enforce compliance and on the extent to which each party continues to expect to gain by cooperating. Many such agreements have proven to be unstable. The tendency for cooperation to degenerate into mutually destructive fish wars is a significant puzzle. The problem may have its roots in high costs of monitoring and enforcement as well as in the fact that the parties' incentives to cooperate change over time (Miller, 1996; McKelvey, 1997). Climatic variations contribute to these sources of instability by causing fluctuations in fish abundance and availability that are difficult to observe and predict. One can see the playing-out of this process in the collapse of cooperation in the Pacific Salmon Treaty case.

2 SALMON ABUNDANCE: CLIMATE AND OTHER INFLUENCES

Pacific salmon are anadromous fish that spawn in streams from California's Central Valley northward. In spring, juvenile salmon emerge from the freshwater environment and disperse into the coastal ocean. Some salmon stocks remain in coastal areas throughout their lives, but many others spend a year or more in a long-distance migration across the feeding grounds of the subarctic Pacific before returning to their natal streams to spawn and to die (Pearcy, 1992). There are five species of Pacific salmon, with a multitude of distinct breeding populations. All five species (chinook, coho, sockeye, pink, and chum) are present from Washington state northward, while in Oregon and California only chinook and coho spawn in significant numbers.

In the mid-1970s, ocean conditions in the North Pacific changed dramatically. Shortly thereafter, Alaskan salmon harvests entered a period of dramatic increase, rising nearly 10-fold from a low of 22 million salmon (of all species) in 1974 to three successive record highs in 1993, 1994 and 1995 (Fig. 1). At the 1995 peak, Alaska harvested a total of 217 million salmon. Harvests of most salmon species in northern British Columbia also fared well during this period, although British Columbia's commercial chinook harvests have declined steadily, and by the late 1990s it became apparent that many of British Columbia's southern and interior coho stocks are severely depleted. Southward, salmon harvests have been on a roller-coaster. Commercial chinook and coho catches in California, Oregon, and Washington dropped abruptly in the late 1970s, hitting El Niño–related lows in 1983 and 1984. A dramatic but brief recovery in 1986 and 1987 then gave way to a precipitous decline to record low harvests in recent years (Fig. 2). Production has declined to the point that some stocks are on the verge of extinction. Some observers attribute these changes in salmon productivity to human disruptions of the southern fisheries and good management of the northern ones (Royce, 1988). However, mounting evidence suggests that shifts in marine climate may have played a major role (Beamish and Bouillon 1993; Hare and Francis 1995; Mantua et al., 1997).

Winter climate over the North Pacific is dominated by a low-pressure system centered near the Aleutian Islands. From 1977 through 1988, the Aleutian Low was frequently deeper than normal, leading to severe storms, increased mixing, and cooler temperatures in the central North Pacific (Trenberth and Hurrell, 1994) (Fig. 3). Along the west coast of North America, sea surface temperatures (SST) were unusually warm during the 1977 to 1995 period (Fig. 4).

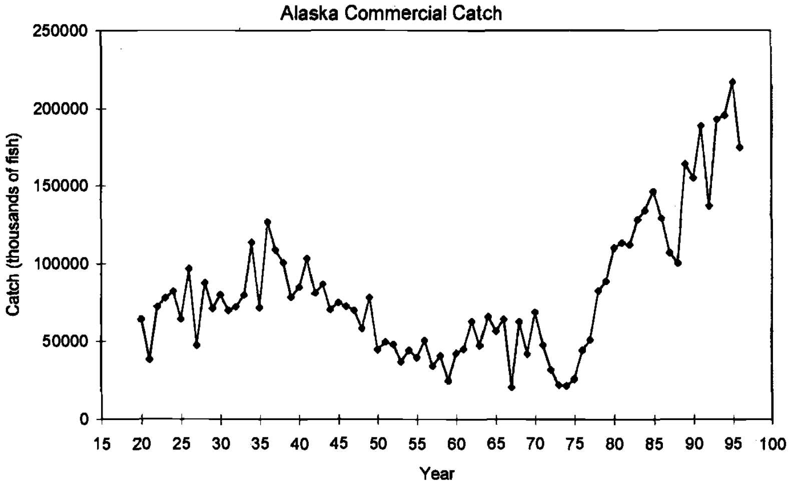


Figure 1 Alaskan commercial salmon harvest—all species.

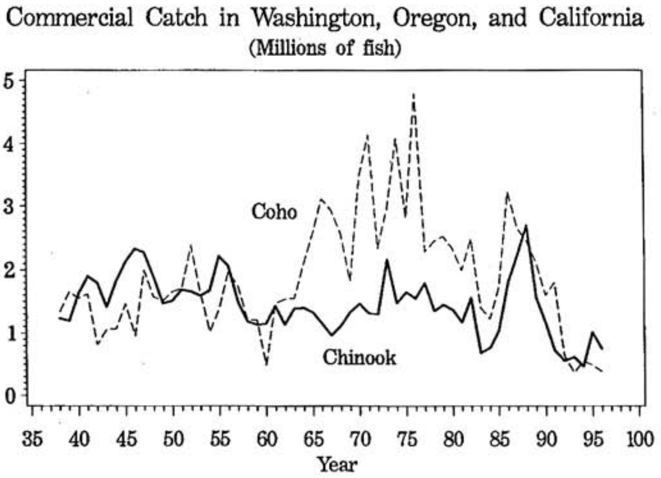


Figure 2 Commercial coho and chinook Salmon harvest in Washington, Oregon, and California, millions of fish.

Possible causes of the change are the subject of much research. In the last century, the North Pacific climate varied on an interdecadal scale, with shifts or trends in mean levels of sea-level pressure and SST that lasted for several decades (Zhang et al., 1996; Latif and Barnett, 1996). The pattern is characterized by alternate warm

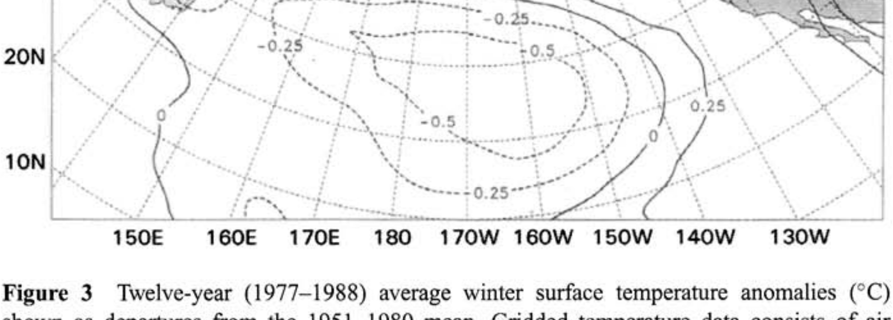


Figure 3 Twelve-year (1977–1988) average winter surface temperature anomalies (°C) shown as departures from the 1951–1980 mean. Gridded temperature data consists of air temperatures over land and sea surface temperatures over oceans (IPCC 1992; Trenberth et al., 1992). Figure courtesy of James Hurrell.

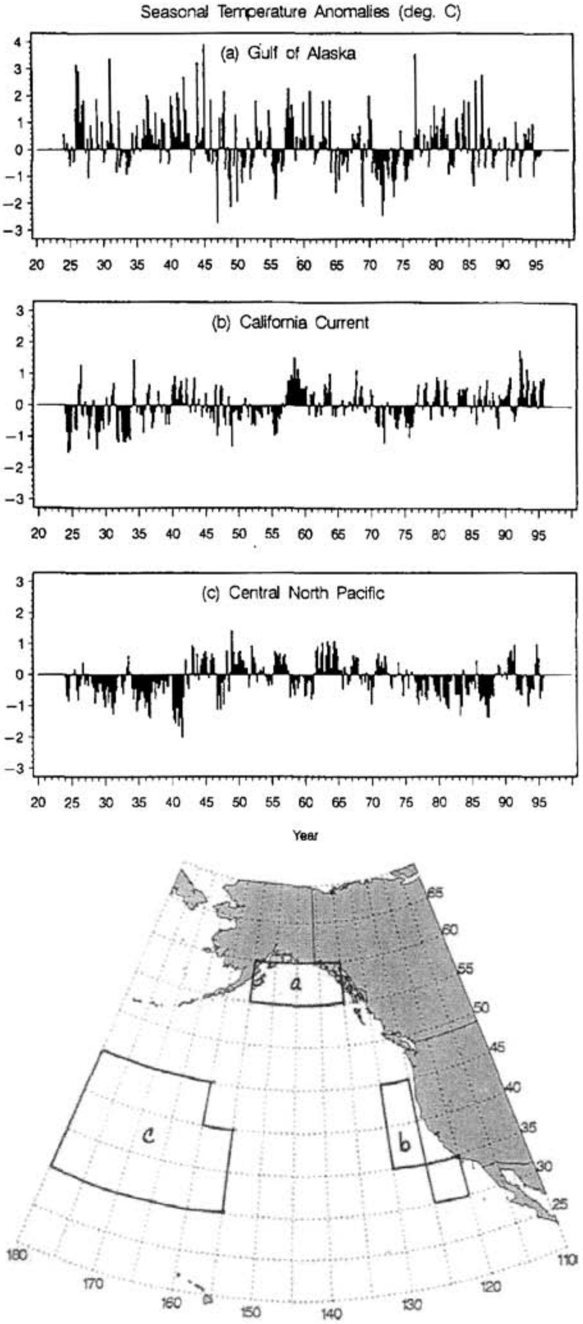


Figure 4 Seasonal surface temperature anomalies in three regions of the northeast Pacific: (*a*) Gulf of Alaska, (*b*) California current, and (*c*) central North Pacific. (Temperature data as described in Fig. 3.)

and cool periods in a large area of the western and central north Pacific, with shifts toward warmer temperatures in the mid-1940s and cooler temperatures in the mid-1970s (the eastern edge of this region is area C in Fig. 4). The cool periods in the central North Pacific are associated with an intensification of the Aleutian Low and a warming of coastal temperatures along the west coast of North America.

The North Pacific also is influenced by the El Niño–Southern Oscillation (ENSO) phenomenon (Kiladis and Diaz, 1989). ENSO-related warming of the equatorial Pacific occurs intermittently, at intervals of about 3 to 6 years, and frequently leads to intensification of the Aleutian Low. The effects of an ENSO warm event often propagate northward, warming the west coast of North America and cooling SSTs in the central north Pacific. An unususal sequence of closely spaced ENSO warm events have occurred from 1977 to 1995, possibly evidence of a change in large-scale climate (Trenberth and Hurrell, 1994; Trenberth and Hoar, 1996). These ENSO warm events (El Niño events) have tended to reinforce the decadal-scale shift to warmer coastal SSTs and cooler SSTs in the central north Pacific.

Intensification of the Aleutian low and warming of the coastal ocean appear to have positive effects on salmon abundance in the Gulf of Alaska, but negative effects on stocks that spend a portion of their lives in the California current (Pearcy, 1992; Hare et al., 1999). In the subarctic zone, the mixed layer has become shallower. This may have enhanced the survival of Alaskan and northern British Columbian salmon smolts by increasing the productivity of zooplankton, which is a frequent food source for juvenile salmon (Polovina et al., 1995; Brodeur and Ware, 1992). A general pattern of winter warming and increased winter precipitation in Alaska over the past two decades (Mantua et al., 1997) also may have contributed to favorable stream conditions for egg-to-smolt survival. From southern British Columbia southward, El Niño events have been associated with poor feeding conditions for maturing salmon and changes in species composition, including increased abundance of some species that prey on juvenile salmon (Pearcy, 1992). In addition, recent droughts in California and the Pacific Northwest have resulted in poor conditions for spawning and migration in the salmon's freshwater phase. Changes in ocean temperatures and circulation, and associated changes in stream conditions, thus appear to have contributed to the opposite trends in northern and southern salmon production.

3 HISTORY OF HARVEST MANAGEMENT

North America's commercial Pacific salmon fisheries were established and grew rapidly in the late nineteenth century. In many areas, returning adult salmon soon were running such a gauntlet of competing fishing gears and other hindrances, such as dams, that it was a lucky salmon that survived to spawn (Higgs, 1982). The resulting decline in salmon populations led to the creation of public agencies to establish fishing gear restrictions and fishing seasons. However, these authorities could never fully control harvests of the salmon stocks within their purview because many salmon could be caught as they passed through the waters of neighboring jurisdictions on their return migration. Such "interceptions" became increasingly important over time as fishing effort expanded in offshore areas.

The first significant international agreement on salmon harvests was a convention between the United States and Canada, signed in 1930 and ratified in 1937. That agreement divided the harvest of Fraser River sockeye salmon as well as management and restoration costs equally between the two nations (Munro and Stokes, 1989). The agreement was later extended to Fraser River pink salmon. Under that convention, the International Pacific Salmon Fishery Commission (IPSFC) regulated harvests of the Fraser River stocks. Although the Fraser River lies entirely in Canada, a large portion of the salmon spawning in that drainage typically approach the river through the Strait of Juan de Fuca where, historically, they had been harvested by Washington State fishing vessels. When a rock slide blocked access to part of their spawning habitat, and sent the Fraser's salmon resources into decline, the United States and Canada had a clear joint interest in removing the blockage and restoring the runs.

That agreement covered only a portion of the salmon runs jointly exploited by the United States and Canada. When negotiations for the Pacific Salmon Treaty began in 1971, Alaskan interceptions of salmon spawned in the rivers of Washington and Oregon were creating tensions among the states, while increasing Canadian troll harvests of those stocks precluded an effective internal solution. In addition, mutual interceptions of salmon of Canadian and Alaskan origin were seen as a barrier to effective management in the northern area (Yanagida, 1987). At the same time, the Canadians had become increasingly unhappy about their agreement to share one half of the Fraser River salmon with the United States because, by foregoing construction of hydropower dams on the Fraser, Canada was effectively bearing more than half of the cost of maintaining those runs. After 14 years of negotiations, the treaty went into effect in 1985.

The treaty created the Pacific Salmon Commission and gave it the authority to promulgate "fishing regimes" to govern harvests in six distinct fisheries. The commission is to periodically renegotiate these regimes as they expire. The body of the treaty lays out a set of general principles to guide the commission in this task. Of central importance are the equity and conservation objectives, which the treaty expresses as follows:

> "...cach Party shall conduct its fisheries and its salmon enhancement programs so as to:
>
> a) prevent overfishing and provide for optimum production; and
> b) provide for each Party to receive benefits equivalent to the production of salmon originating in its waters. (Pacific Salmon Treaty, Article III)"

The treaty then advises the parties to consider the following factors: the desirability of reducing interceptions, the desirability of avoiding disruption of existing fisheries, and annual variations in abundances of the stocks. These considerations are somewhat mutually inconsistent because many of the existing fisheries relied heavily on interceptions.

Until the June 1999 amendments to the treaty, the fishing regimes consisted primarily of harvest ceilings for specific locations and species. For example, in 1985 and 1986, the annual all-gear harvest of chinook in northern and central British Columbia and southeast Alaska was to be limited to 526,000 fish divided equally between the parties (*Treaty*, Annex IV, Chapter 3). Although the intent of the treaty was to control interceptions of fish produced in other jurisdictions, the difficulty of identifying the true origins of fish taken in an ongoing mixed-stock fishery led to the harvest ceiling approach as a proxy method for balancing interceptions.

In addition, the regimes were effective for only a few years. Negotiations for new regimes were to follow a consensus rule, but that allowed any of the parties to veto proposed fishing regimes seen as contrary to its constituent's interests (Yanagida, 1987; Miller, 1996; Schmidt, 1996; Munro et al., 1998). The relevant parties in this context are Canada and the 3 voting U.S. commissioners—representing Alaska, Washington/Oregon, and 24 treaty tribes located in Washington, Oregon, and Idaho. While the Canadian federal government has primary authority on the Canadian side, the British Columbia Provincial Government has often differed vociferously with federal policies, and those internal differences frequently colored the course of the negotiations. When the parties failed to agree on fishing regimes, regulatory authority reverted to the appropriate state or federal jurisdiction. In the United States, the states have authority within 3 nautical miles of the coast and federal jurisdiction extends from 3 to 200 miles offshore.

4 RECENT CONFLICT

The recent breakdown in efforts to renegotiate the expired fishing regimes revolved around two issues. The first was a long-standing dispute over the meaning and enforcement of the treaty's equity provisions. The second was disagreement regarding actions required to meet the treaty's stated goal of rebuilding chinook stocks from the Columbia River northward to southeastern Alaska. When the treaty went into effect, all parties recognized that interceptions could not be reduced to zero and that the interception balance would vary from year to year. They also recognized that the balance would tend to favor either the United States or Canada in each of the six covered fisheries. Canada hoped, however, that the treaty would lead to a rough balance in total interceptions. In particular, they expected that their own interceptions of U.S. coho and chinook would roughly offset the value of U.S. interceptions of Fraser River salmon (Munro and Stokes, 1989; Munro et al., 1998).

Nature and the actions of each party to the agreement have thwarted these expectations. The Canadian commissioners charged that the harvest ceilings failed to ensure an equitable division of the catch. They claimed that Alaska consistently intercepted an excessive number of Canadian salmon. Canada was unable to offset increasing Alaskan interceptions because declining southern coho and chinook abundance prevented Canadian harvesters from reaching the agreed-upon ceilings for harvests of those stocks along the west coast of Vancouver Island. At the same time, fishing interests along the U.S. West Coast claimed that Canada's efforts to reach the ceiling resulted in overharvesting of those fragile stocks. Alaskan officials countered the Canadian charges by arguing that increased interceptions were unavoidable given the increased abundance of their own intermingled stocks.

When the fishing regimes expired, the Canadians used the opportunity to reassert their demands for a quantitative approach to the equity issue. In previous rounds of regime setting, the Canadians had acquiesced to the quota approach. However in their view, the treaty principle that each party should receive "benefits equivalent to the production of salmon originating in its waters" (*Treaty*, Article III, para. 1) should be interpreted literally as a dollar-for-dollar balancing of the value of a nation's harvest with the value of the salmon spawned in its waters. According to Canadian calculations, the United States owed Canada a considerable debt. The U.S. delegation never favored a quantitative approach, arguing instead that: "[A]n effort to create an accounting scheme would invite costly, and perhaps divisive and inconclusive debate over biological and economic variables" (Yanagida, 1987, p. 591). U.S. officials have been quick to point out that slightly different biological assumptions and valuation rules can give vastly different results regarding amounts owed and even the direction of the equity imbalance (Personal communication, Tom Cooney, Washington State Department of Fish and Wildlife, June 22, 1995; Munro and Stokes, 1989). The U.S. side preferred to treat each of the covered fisheries separately, giving due recognition to the treaty provision disfavoring economic disruption of historic fisheries. Each party's refusal to accept the other's approach to the equity issue resulted in a protracted stalemate.

In addition to Canadian/U.S. differences over implementation of the treaty's equity provisions, a rift developed between Alaska and the other U.S. parties over chinook harvests. When the treaty was ratified, the parties agreed to a program of limiting harvests in order to rebuild naturally spawning chinook stocks by the year 1998. While Alaska's chinook harvests have not increased, declining runs in British Columbia and the southern U.S. jurisdictions pushed the rebuilding goal further out of reach. Tensions on the U.S. side reached a boiling point in 1995 when the Northwest treaty tribes and the states of Washington and Oregon sued Alaska and won an injunction that closed the southeastern Alaska chinook fishery for the remainder of the season (*Confederated Tribes and Bands v. Baldridge* [W.D. Wash. September 7, 1995].

As the treaty dispute escalated, the Canadians employed a variety of desperate tactics in an effort to force the United States back to the bargaining table. For example, in 1994, British Columbia tried to pressure the southern U.S. parties by pursuing an "aggressive fishing strategy." That strategy failed to win any concessions and resulted in dangerous overharvesting of part of the Fraser River sockeye run by the Canadian fleet (Fraser River Sockeye Public Review Board, 1995). By the summer of 1997, British Columbia's salmon harvesters had become so frustrated that approximately 150 fishing vessels participated in a blockade that held the Alaska Ferry hostage in the Canadian port of Prince Rupert for 3 days (Hogben et al., 1997; D'oro, 1997).

Canadian frustration was fueled by Alaska's unwillingness to take actions to reduce the interceptions imbalance. Such concessions made little sense from Alaska's standpoint because they would impose costs on Alaska without commensurate benefits. In fact, Alaska never had much to gain by participating in the Pacific Salmon Treaty, and apart from possible suits over interference with Native American treaty fishing rights, the other U.S. parties have very few bargaining chips to use in their negotiations with Alaska. Alaska's favorable position arises from the fact that many salmon stocks swim northward as juveniles to feed and mature in the Gulf of Alaska. As adults, they swim southward to return to their natal streams. This migratory pattern gives Alaska a natural advantage in exploiting chinook salmon from the southern U.S. jurisdictions and certain Canadian stocks while harvesters in the southern U.S. jurisdictions do not intercept Alaskan origin salmon.

In addition, as Alaska's own salmon became more numerous, Alaska's fishery managers found it increasingly difficult to limit interceptions. In southeastern Alaska, salmon harvesting occurs primarily in areas where Alaska's fish are intermingled with stocks originating elsewhere. Fish caught in those offshore areas are in top condition and at peak value. The Alaskans argue that interceptions cannot be kept constant when their own stocks increase, unless they allow a larger number of their own fish to escape harvesting in those prime areas. Those fish could contribute to spawning escapements or they might later be harvested in a river or estuary where their commercial value is often much lower. However, with spawning escapements already strong, and markets glutted with lower valued "canning quality" salmon, neither of those options is attractive. The Alaskan stance in the negotiations has consistently been that they should be allowed to reap the rewards of their own good salmon management and that they are not to blame for the declining southern stocks.

Canadian efforts to promote a fish-for-fish approach to the equity issue, Alaskan intransigence, and the helplessness of the other U.S. parties to halt the declines of their salmon resources collided in a dangerous muddle. This situation threatened the health of some highly valued salmon stocks and cast a rancorous cloud on relations between the two nations.

5 CURRENT AGREEMENT AND PROSPECTS FOR THE FUTURE

The 1999 agreement represents a dramatic break from the previous approach. Rather than relying on short-lived, ceiling-based regimes whose frequent renegotiation provided ample opportunity for disagreement and brinkmanship, the new agreement establishes a long-term commitment to define harvest shares as a function of the abundance of each salmon species in the areas covered by the treaty. For example, for the next 12 years, the U.S. share of Fraser River sockeye will be fixed at 16.5% of the annual harvest. This represents a decrease from the post-1985 average U.S. share of 20.5%, but an increase relative to the share actually attained by the U.S. fleet during the 1992 to 1997 salmon war period (DFO, 1999; O'Neil, 1999a). This percentage approach allows the number of Fraser River sockeye harvested by the U.S. fleet to increase in years of high sockeye abundance while requiring reduced harvests when abundance is depressed. In contrast, in the 1985 treaty, U.S. harvests of Fraser sockeye were to be held to a cap of 7 million fish over each of two successive 4-year periods (*Pacific Salmon Treaty,* Annex 4).

The new arrangements for chinook, which will be in effect for 10 years, take account of the fact that the various fisheries along the coast differ considerably in the extent to which they rely on healthy or depressed chinook stocks (U.S. Department of State, 1999). Accordingly, the agreement designates two types of fisheries: (1) aggregate abundance-based management (AABM) fisheries will be managed based on indices of the aggregate abundance of chinook present in the fishery and (2) individual stock-based management (ISBM) fisheries, which are primarily located in inside fishing areas, will be managed based on the status of individual stocks or groups of stocks (e.g., on the basis of the evolving status of currently endangered or threatened stocks). Abundance-based allocation rules for coho have not yet been developed, but the agreement instructs the parties to jointly develop such a management approach, and specifies various deadlines for the accomplishment of particular tasks.

It is not yet clear if the new approach will provide a path to sustainable cooperation. While the focus on conservation will tend to protect some of the weak stocks that were jeopardized by the recent turmoil, the new agreement does little to resolve long-standing differences over the division of benefits. In particular, many Canadians remain convinced that Canada will come out "short" under this agreement, and they have labeled it "profoundly disappointing" (Culbert and Beatty, 1999). They are also unhappy about the long-term implications of the agreement because they feel that it risks committing them to an unsatisfactory arrangement for many years. In fact, the Canadian delegation had unsuccessfully argued for a shorter term. Although the parties have formally stipulated that compliance with the terms of the new agreement shall be deemed to fulfill the requirements of Article III of the treaty, the stipulation applies only for the duration of the current agreement. If Canadians continue to feel that their interests have been compromised, there may be renewed turmoil when this agreement expires. Nevertheless, the agreement represents a step in the right direction and may perhaps lay the groundwork for more robust future cooperation.

One positive aspect is the agreement's provision for two endowment funds, the proceeds of which are to be used to support scientific research, habitat restoration, and enhancement of wild stock production in their respective (northern and southern) areas. Initially, the funding is to be provided entirely by the United States, and the entire agreement is contingent on U.S. Congressional approval of $75 million for the Northern Fund and $65 million for the Southern Fund. In the future, either party may make additional contributions, and even third parties may contribute, with the consent of the parties.

Some analysts have suggested that expanding the scope of the salmon negotiations to allow "side payments" might provide a path to sustained cooperation. These payments might be either in monetary form or in the form of concessions on other nonsalmon issues (Schmidt 1996; Munro et al., 1998). The endowment funds might be a vehicle for providing such side payments, although their initial yield will be far smaller than the debt claimed by Canada from the United States for the accumulated harvest imbalance. A portion of the money available from the Northern Fund will support Alaskan research and enhancement (O'Neil, 1999b), which, together with other payments, might elicit greater cooperation from Alaska. However, it remains to be seen if this tool will be put to good use, or if its promise will be squandered in quarreling over distribution of the proceeds.

In summary, the new agreement is a hopeful sign, but it does not ensure sustained cooperation. The division of benefits is still largely tied to the division of the catch. Depending on the vagaries of nature, the parties may or may not find that division to be satisfactory. If they do not, and if they fail to develop other methods to ensure a fair division of benefits, then another breakdown in cooperation will likely occur when this agreement expires. The Pacific salmon case demonstrates that the societal impacts of weather and climate are often a complex product of physical or biological impacts, institutional factors, and economic motivations operating in a context of incomplete information. We must attempt to understand all aspects of this complex interaction if we are to improve society's ability to cope with the effects of climatic variations.

REFERENCES

Beamish, R. J., and D. R. Bouillon, Pacific salmon production trends in relation to climate, *Can. J. Fish. Aquat. Sci.*, *50*, 1002–1016, 1993.

Brodeur, R. D., and D. M. Ware, Long-term variability in zooplankton biomass in the subarctic Pacific Ocean, *Fish. Oceanogr.*, *1*, 32–38, 1992.

Confederated Tribes and Bands v. Baldridge, Civil Case C80-342, Order and Judgment of U.S. District Judge Barbara J. Rothstein, W.D. Wash., September 7, 1995.

Culbert, L., and J. Beatty, Salmon pact "disappointing," *Vancouver Sun*, final ed., June 4, 1999, p. A1.

Department of Fisheries and Oceans (DFO), Government of Canada, *Canada and U.S. Reach a Comprehensive Agreement under the Pacific Salmon Treaty*, press Release, and Backgrounders, available on-line, http://www.ncr.dfo.ca, June 3, 1999.

D'oro, R., Judge orders end to blockade Alaska officials attempt to end ferry standoff, *Anchorage Daily News*, final ed., July 21, 1997, p. A1.

Fraser River Sockeye Public Review Board, *Fraser River Sockeye 1994; Problems and Discrepancies*, Canada Communication Group Publishing, Ottawa, 1995.

Hardin, G., The tragedy of the commons, *Science*, *162*, 1243–1248, 1968.

Hare, S. R., and R. C. Francis, Climate change and salmon production in the Northeast Pacific Ocean, in R. J. Beamish (Ed.), *Climate Change and Northern Fish Populations*, Can. Spec. Publ. Fish. Aquat. Sci. 121, 1995; pp. 357–372.

Hare, S. R., N. J. Mantua, and R. C. Francis, Inverse production regimes: Alaska and West Coast Pacific salmon, *Fisheries*, *24*(1), 6–14, 1999.

Higgs, R., Legally-induced technical regress in the Washington salmon fishery, *Res. Econ. History*, *7*(1), 55–86, 1982.

Hogben, D., D. Rinehart, and J. Beatty, Prince Rupert fish boats end blockade of Alaskan ferry: Ringleaders comply with injunction after meeting late Monday evening with federal fisheries minister David Anderson, and the ferry steams away, *Vancouver Sun*, final ed., July 22, 1997, p. A1.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums (Eds.), *Climate Change (1992)*, Cambridge University Press, Cambridge, 1992.

Kiladis, G. N., and H. F. Diaz. Global climatic anomalies associated with extremes in the Southern Oscillation, *J. Climate*, *2*, 1069–1090, 1989.

Latif, M., and T. P. Barnett, Decadal climate variability over the North Pacific and North America: Dynamics and predictability, *J. Climate*, *9*, 2407–2423, 1996.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, A Pacific interdecadal climate oscillation with impacts on salmon production, *Bull. Am. Meteorol. Soc.*, *78*(6), 1069–1079, 1997.

McKelvey, R., Game-theoretic insights into the international management of fisheries, *Natural Resour. Model.*, *10*(2), 129–171, 1997.

Miller K. A., Salmon stock variability and the political economy of the Pacific Salmon Treaty, *Contemp. Econ. Pol.*, *14*(3), 112–129, 1996.

Munro, G. R., T. McDorman, and R. McKelvey, Transboundary fishery resources and the Canada-United States Pacific Salmon Treaty, in *Occasional Papers: Canadian-American Public Policy*, Canadian-American Center, University of Maine, Orono, 1998.

Munro, G. R., and R. L. Stokes, The Canada-United States Pacific Salmon Treaty, in D. McRae and G. Munro, (Eds.), *Canadian Oceans Policy: National Strategies and the New Law of the Sea*, University of British Columbia Press, Vancouver, 1989, pp. 17–35.

O'Ncil, P., Taking stock of the Salmon Treaty, *Vancouver Sun*, final ed., July 5, 1999a, p. A1.

O'Neil, P. Conservation card played key role in salmon deal, *Vancouver Sun*, final ed., June 5, 1999b, p. A1.Pacific Salmon Treaty, March 18, 1985, U.S.-Can 99 Stat. 7 [codified at 16 U.S.C. 3631–3644 (1997)].

Pacific Salmon Treaty, March 18, 1985, U.S.-Can., 99 Stat. 7 [codified at 16 U.S.C. 3631–3644 (1997)].

Pearcy, W. G., *Ocean Ecology of North Pacific Salmonids*, University of Washington Press, Seattle, 1992.

Polovina, J. J., G. T. Mitchum, and G. T. Evans, Decadal and basin-scale variation in mixed layer depth and the impact on biological production in the Central and North Pacific, 1960–88, *Deep Sea Res.*, *42*, 1701–1716, 1995.

Royce, W. F., An interpretation of salmon production trends in W.J. McNeil (Ed.), *Salmon Production, Management and Allocation*, Oregon State University Press, Corvallis, 1988.

Schmidt, Jr., R. J., International negotiations paralyzed by domestic politics: Two-level game theory and the problem of the Pacific Salmon Commission, *Environ. Law*, *26*, 95–139, 1996.

Trenberth, K. E., J. R. Christy, and J. W. Hurrell, Monitoring global monthly mean surface temperatures, *J. Climate*, *5*, 1406–1423, 1992.

Trenberth, K. E., and T. J. Hoar, The 1990–1995 El Niño–Southern Oscillation event: Longest on record, *Geophys. Res. Lett.*, *23*(1), 57–60, 1996.

Trenberth, K. E., and J. W. Hurrell, Decadal atmosphere-ocean variations in the Pacific, *Climate Dynam.*, *9*, 303–319, 1994.

U.S. Department of State, Diplomatic Note No. 0225 from Canada to the United States; reply; attached Agreement, available on-line, http://www.state.gov, June 30, 1999.

Yanagida, J. A., The Pacific Salmon Treaty, *Am. J. Int. Law*, *81*, 577–592, 1987.

Zhang, Y., J. M. Wallace, and N. Iwasaka, Is climate variability over the North Pacific a linear response to ENSO? *J. Climate*, *9*, 1468–1478, 1996.

CHAPTER 47

TRANSBOUNDARY RIVER FLOW CHANGES

ROGER S. PULWARTY

1 INTRODUCTION

Water is a "fugitive" resource in the sense that it flows naturally from one place to another, from one reserve to another (e.g., groundwater to surface), and from one physical state (solid, liquid, and gas) to another. Thus *transboundary* can mean many things including transitions from wet to arid zones, from upstream to downstream, from one country or province to the next, etc. The Convention on the Protection and Use of Transboundary Watercourses and International Lakes (1992) defines "transboundary waters" to mean "any surface or ground waters which mark, cross, or are located on the boundaries between two or more states." Wherever transboundary waters flow directly to the sea, these transboundary waters end at a straight line across their respective mouths between points on the low water line of their banks. Groundwater resources are also frequently shared by two or more countries. Emerging issues in water resources emanate from three categories of problems: (1) transboundary water availability, (2) transboundary groundwater allocation, management, and conservation, and (3) transboundary water quality (Caldwell, 1993). This chapter is primarily concerned with surface water in large river basins crossing international boundaries.

Watersheds can be defined by crossing a physical line (or bench), dividing surfaces from which waters flow in different directions to different outlets. The term '*river basin*' is here used to denote channel length and catchment area. '*Catchment area*' refers to a drainage area in which surface water flows to a common outlet channel. '*Watersheds*' can also be crossed conceptually, such as in differentiating between upstream and downstream emphases in managing water and sediment

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

flows, water quantity and quality relationships, changes from soil type to landscape type as a basis for study and administration, trade-offs between centralized versus decentralized emphases in management, and concern for impacts on the environment and on basins of origin (White, 1997).

Shared watersheds constitute about 47% of the global land area and are inhabited by about 40% of the world's population. Worldwide there are more than 214 major transboundary river basins, of which 74% are shared between only 2 riparian states (Kaufman et al., 1997). Thirteen basins are shared by 5 or more countries, 4 of these (the Congo, Danube, Nile, and Niger) are shared by 9 or more countries; 3 (the Zambezi, Amazon, and Rhine) are shared by 7 riparians; 4 (Chad, Volta, Ganges–Brahmaputra, Mekong) are shared by 6 countries, and 2 (La Plata, Elbe) are shared by 5 countries. Nine basins have 4 riparians each, and 30 major basins are shared by 3 countries. Of the 9 international water bodies shared by 6 or more countries, 5 are in Africa. The largest river basin in Europe, the Danube, is one-seventh the size of the Amazon Basin, but its waters are shared among 17 countries. On the smaller scale, there are over 400 surface waters between Finland and Russia alone, 150 crossing the Netherlands including the Rhine, and over 150 in the Ganges–Brahmaputra–Meghna system.

Much recent attention has been focused on highly visible water-related problems in the Middle East, the Indian Subcontinent, and the Aral Sea Basin. The UN Food and Agriculture Organization (FAO) has identified more than 3600 treaties relating to non-navigational water use dating between the years 1805 and 1984. Since 1945, approximately 300 treaties have been negotiated dealing with water management allocations of international basins (Kaufman et al., 1997). But, one-third of the major international rivers have no international agreement, and fewer than 30 have any cooperative institutional arrangements.

Rivers have constituted boundaries long before the rise of the modern nation state. Interestingly, the notion of "rival," derived from the Latin "rivalis," was originally used to describe people living on opposite banks of a river. Among the most important factors conditioning transboundary conflicts are the historical patterns of use and the needs of new "arrivals" and changes in values over time. Several countries now rely on significant amounts of surface waters originating outside of their national borders (Fig. 1). Changing social factors in particular have made water-related resource conflicts within countries common and, in situations where water is shared between two or more countries, increasingly unavoidable (Kaufmann et al., 1997).

Frederick (1996), Dinar (1997), and others have highlighted several factors underlying most international disputes involving river flows, including: the variability and uncertainty of supply, interdependencies among users, increasing water scarcity, overallocation and rising costs, the increasing vulnerability of water quality and aquatic ecosystems to human activities, ways and means of supplying safe water facilities, and the mobilization of financial resources for water development and management. Many of these issues derive from concerns in water resources management in general. How these concerns are met is strongly shaped by the choice of the spatial unit within which studies and management actions are conducted, by the way

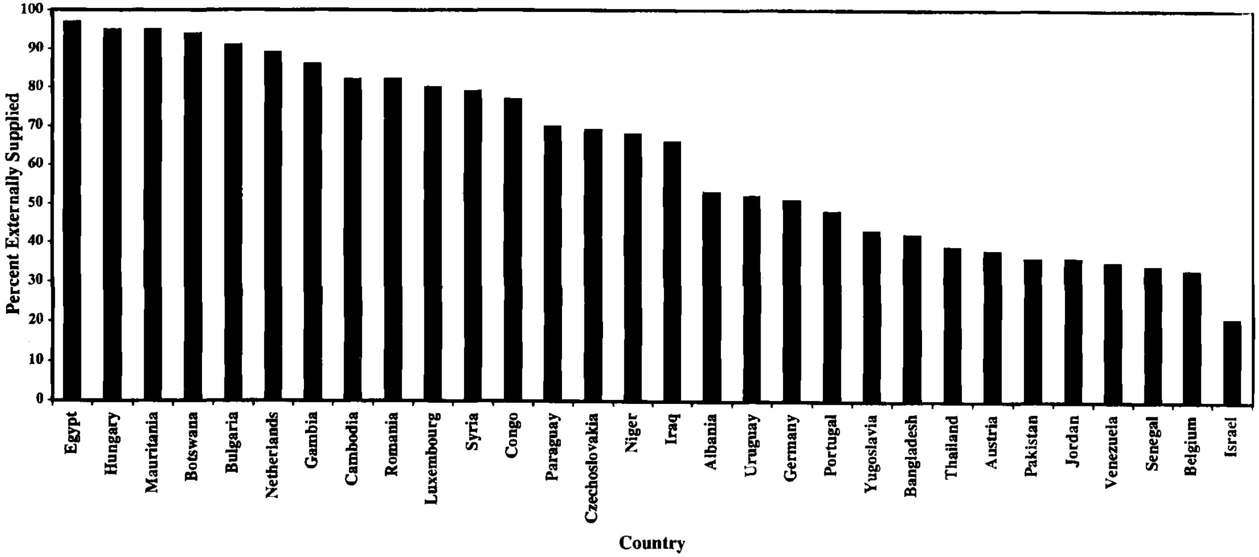


Figure 1 Percentage of water supply originating externally to selected countries (data from Gleick, 1993).

problems have been defined and changed over time, and by who benefits from defining problems in a particular way.

The primary concern in an international context is the need for international cooperation in the development of institutional and human resources for the efficient and equitable management of transboundary waters under variable and changing environments. In the following discussion the scales of human activities and interactions with large river basins are put in the context of streamflow changes on the time scales of century, decadal, seasonal, and extreme events.

2 IMPACTS

Transboundary fluctuations and changes in river flow can be attributed to (1) climate variations and change and (2) physical and biological transformations of basin hydrology including increased storage, diversions, and landscape changes. In this section, these conditioning factors on flow variability and change are discussed in general. Three cases are then selected for illustration in detail.

Climate Variability and Streamflow

At a given point along a river, streamflow is the product of the total catchment area above the gage and the average rate at which runoff is generated from snow and/or rain in that catchment. Runoff within a basin depends not only on rainfall but on its temporal distribution, vegetation cover (amounts and types), evapotranspiration, soil moisture storage capacity, rate of groundwater outflow, amount of paved area, etc. The seasonality of streamflow varies widely from river to river and is influenced mostly by the local seasonal cycle of precipitation, by timing of snowmelt (where appropriate), by the travel time of water from the runoff source areas, through surface and subsurface reservoirs and channels, and by large-scale human interventions.

Year-to-year variations in streamflow timing and magnitude play important roles in the development and management of water resources in most regions. Such interannual variations may be superimposed on longer (decadal to century-scale) fluctuations. Regime shifts from wetter to drier periods (and vice versa) on the decadal scale can be seen, for instance, in the annual streamflow data for the Colorado Basin in the southwestern United States (Fig. 2). Some but not all of these variations and their characteristics can be attributed to El Niño–Southern Oscillation events, climate forcings on the decadal time scale in the north Pacific, etc. There are many exceptions to large-scale patterns of streamflow seasonality in a region. Spatial differences in and between main streams and tributaries pose significant problems for flow estimation and planning. The shifts, and the surprises they introduce, provide a variable background against which allocations and water quality requirements are to be developed, agreed upon, and implemented.

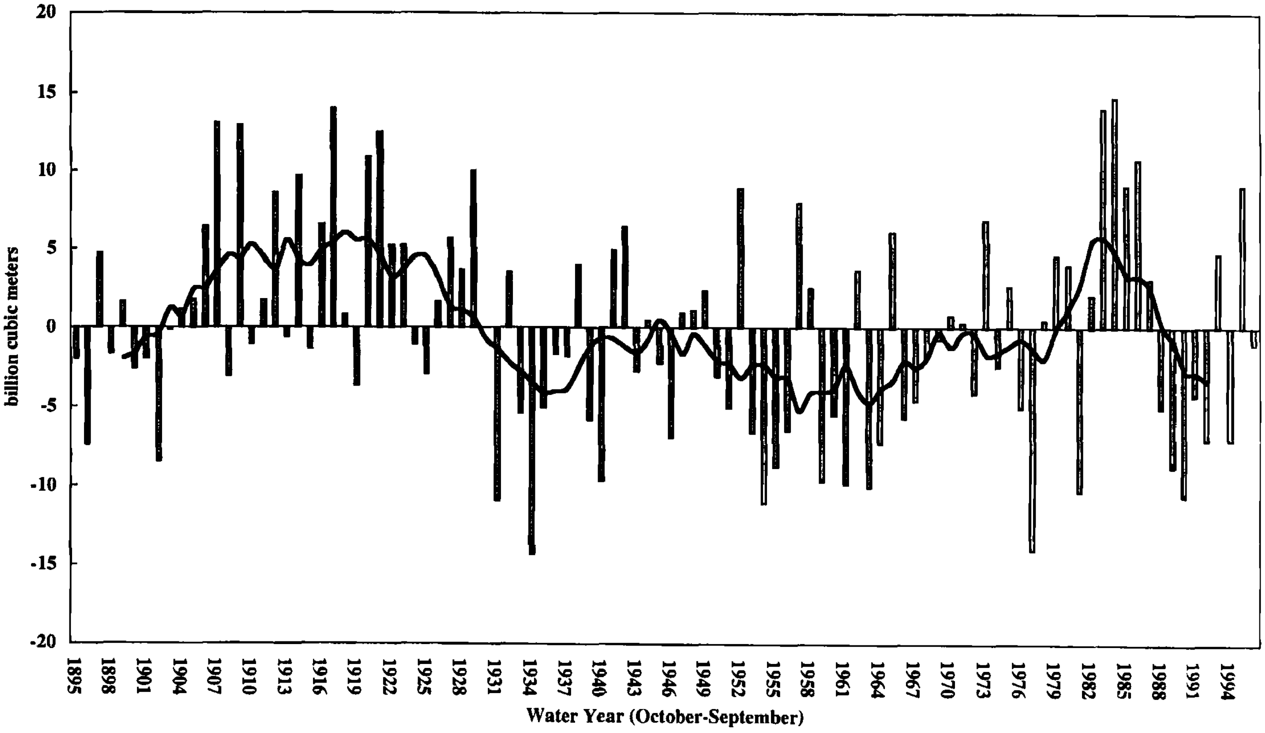


Figure 2 Colorado River streamflow: Annual deviations from the long-term mean (1896–1996) at Lee Ferry and 9-year moving average.

Physical and Biological Transformations: Storage, Diversions, and Landscape Changes

Human-built structures can either increase or decrease hydrologic connectivity of freshwater systems, rates of water movement, and transport and movement of organisms, materials, and heat. Construction of dams and diversions and modification of watersheds have greatly altered the natural flow regimes in streams and rivers (see, e.g., Fig. 3). Many existing flow regimes, particularly in large rivers, reflect human demands for water rather than natural cycles (Naiman et al., 1995). Two of the most dramatic changes to rivers in the twentieth century have taken place with regard to (1) water quality decline from return flows of agricultural and municipal waste water (L'vovich and White, 1990) and (2) large-scale diversions of water from one watershed to another. By 1990, the agricultural use of water worldwide was almost double that of all other uses combined, and canals in the former Soviet Union alone were estimated to be diverting more than 60 km^3 of water annually. In the United States diversions such as that from the Upper Colorado to Missouri and Rio Grande Basins and, more dramatically, from the Lower Basin to southern California have had significant impacts on the basins of origin and on streamflow into the Colorado River delta in Mexico.

At the global level, there has been a tripling of water use since 1950 (Postel, 1997). The number of large dams increased sevenfold to about 39,000, with reservoir capacity at about 9% of global annual river runoff. Until the 1930s, dam and reservoir designers were concerned primarily with single-purpose economic benefits (e.g., either transportation or irrigation). Since then, dams have been designed or altered to meet multiple-purpose criteria including flood control, hydropower, fishing, and recreation. Adverse social and environmental impacts include displacement, disease, siltation, scouring, reduced length of wild rivers, interference with migration and life cycles of aquatic species, introduction of exotic species, eutrophication and anoxia, and losses through evaporation and seepage.

Soil and vegetation act as intermediaries between precipitation and streamflow. Changes in landscapes brought about by urbanization, agriculture, forestry, industrialization, channelization, and construction of transportation corridors alter terrestrial and aquatic components of watersheds. Such alterations result in changes of flow (volume and timing) of water, sediments, nutrients and organisms in river channels, lake basins, wetlands, and groundwater. The rates, processes, and consequences of these changes are not well documented for most rivers. While the data on large-scale irrigated lands are for the most part reliable and comprehensive, no such detail exists for changes in drained wetlands and in low-lying grasslands.

In the following section, three cases (the Nile, the Colorado, and the Plata Basins) are chosen to illustrate problems, opportunities, and challenges in the sharing of transboundary streamflow under variable climate conditions and evolving management arrangements. By the late-1960s there were only 25 major rivers in the world with streamflow records extending back for at least 60 years (NAS, 1968). This discussion is limited to the reliable period of record for each basin considered.

3 THE NILE: CENTURIES OF CHANGE

The Middle East region, taken here to include the Tigris–Euphrates, the Jordan, and the Nile River Basins, has some of the highest population growth rates in the world and has a heavy reliance on irrigation for agricultural productivity. The region also has diverse historical religious, ideological, and ethnic disputes. Political friction and water scarcity have combined to produce perhaps the most volatile situation in the world (McCaffrey, 1993). At present, only the Tigris–Euphrates system has a significant amount of water left for allocation after present demands are met.

The Nile is the longest river in the world, flowing some 6500 km and draining an area of about 3.1 million km^2 spread over 10 countries. Most of its water is, however, generated over only 20% of this area. Its waters have been in use by Egypt for over 5500 years. The Nile differs from many transboundary rivers such as the Tigris–Euphrates and the Colorado River systems, in that the strongest economy, the strongest military force, and the best established water user in the basin, i.e., Egypt, is the downstream nation. Ninety-five percent of its runoff originates outside of Egypt (Table 1). The Nile experiences significant decadal to century-scale variability. Biblical stories of feasts and famines remind us of how the river dominated the climate experience and well-being of ancient Egypt (Riebsame et al., 1996).

The record of Nile flood levels at the Nilometer on Roda Island in Cairo has been called "the world's longest, best quantified climate proxy providing year-by-year values spanning an interval of 13 centuries" (see Diaz and Pulwarty, 1992). The Blue Nile flows from the Ethiopian Highlands to its confluence with the White Nile in Khartoum, Sudan. The White Nile loses half of its flow to the Sudan's Sudd (swamp lands region) as it leaves the Equatorial Plateau. The Atbara from Ethiopia is the last major tributary below Khartoum; here, the flow of the Nile averages 88 BCMY (billion cubic meters per year). From there the river flows 1200 km to Lake Nasser losing 4 BCMY to evaporation, with virtually no additional inflow for its remaining 2500 km run to the Mediterranean. Water levels are at their lowest from March to June and at their highest during the summer monsoon period from July to September. During August, the Ethiopian tributaries provide up to 95% of the flow. The Blue Nile as a watershed constitutes 16% of the entire basin, and at present contributes 62% of the annual flow at Cairo, while the White Nile contributes about 32%.

Changes in the contributions of the source regions to the aggregate streamflow have been noted during three main periods: (1) AD 650 to 1250 with about equal contributions from both sources, (2) AD 1250 to 1480 with greater contribution from the equatorial White Nile source, and (3) AD 1480 to 1870 with lower discharge from the equatorial source. Increases in flow from 1650 to 1750 were more widespread, roughly matching increases in Lake Chad levels. Variations also occur within these century-scale fluctuations on a 25- to 45-year time scale (Diaz and Pulwarty, 1992).

After the construction of the Aswan High Dam during the 1960s, the Nile maximum to minimum discharge ratio decreased from a natural condition of 12 : 1 to 2 : 1. In addition to reducing seasonal to interannual variability of the river and to providing hydropower and predictable flows for irrigation, the dam also reduced sediment transport, estimated at 110 million tonnes per year, by 98%. The end result of these modifications in the basin is that the Nile delta is slowly declining and Egypt, at this time, uses more fertilizer per hectare than any other nation. In addition, 30 of the 47 commercial fish species available before construction of the Aswan Dam have been reduced to below harvestable levels.

Allocations of Nile water are based on the 1959 Agreement for the Full Utilization of the Nile Waters, which allotted 55.5 BCMY to Egypt out of a "fixed" flow of 84 BCMY. Requested allocations for upper Nile Basin states, most of which were British colonies in 1959, were rebuffed by Egypt. The result has been that Sudan and Egypt have harnessed most of the Nile's water with negligible development by other basin states. Egypt pursues a status quo position of fixed or increased inflows, resisting the construction of dams by Ethiopia on Nile tributaries. At present, Nile water in an average year plays a role in producing food for about half the population of Egypt. Large food imports, thus represent the importation of "virtual water," that is, water that Egypt would have had to obtain itself in order to grow the same amount of food.

As discussed by Riebsame et al. (1996), any calculation of climate change impact in the Nile Basin is complicated by assumptions about intricate water allocation and institutional arrangements, chiefly between Sudan and Egypt. At present, the Sudan requires additional Nile water for irrigation, but additional withdrawals would be detrimental to Egypt's already fully used supplies. Countries dependent on hydropower also have to agree to release adequate amounts of water for irrigation. However, there is considerable reluctance to release water at the expense of hydroelectricity. In addition, several of the upstream countries have severe food security problems, with much less money to import food than does Egypt (Appelgren and Klohn, 1997). These countries increasingly view water resources development, including interbasin transfer, as their major hope and option for socioeconomic development. One such system, the Jonglei Canal, designed to increase available Nile flows, is over half completed. The canal was designed to divert water around the Sudd to enable larger quantities of White Nile flow to reach Aswan. Its development is halted, at present. While rules for regulation of the equatorial lakes to minimize downstream and upstream losses have been developed, the scale of ecological impacts of the canal, if completed, are anticipated to be large for the Sudd. Even with better technical information, it is still up to the states of the Nile Basin to decide which operational strategy represents the most desirable compromise (Georgakakos et al., 1997). The Quaternary record also shows evidence that for periods of time the equatorial sources formed a closed basin, i.e., they did not flow to the Sudd nor reach the confluence. Pressures on Nile resources have now reached the stage where countries are obliged to make the river both an object and instrument of domestic and international politics (Appelgren and Klohn, 1997), without the flexibility needed to adapt to natural changes in variability or social trends. Ethiopia, for instance, is expected to have 25% more people than Egypt by 2025.

4 THE COLORADO: DECADAL-SCALE VARIATIONS

The past and present alterations of hydrology in the southwestern United States and northwestern Mexico reflect complex histories of human settlement, large-scale water diversions, the development and evolution of water policy and law, and expanding frameworks of water resources management. Rapid population increases, economic growth (including agriculture), the rise of urban centers over the last century (and more so recently) have resulted in intense pressures being placed on western lands, water, and institutions. In addition, the focus on water supply and the resulting design of water management agencies in the U.S. West evolved under the presumption of water as an open resource. Recent emphases on water demand management, on meeting obligations to Native American tribes, and on environmental concerns have altered the traditional roles of federal, state, and local agencies. Even in the water-rich Pacific Northwest region, trade-offs, for example (between hydropower and endangered species requirements), have brought allocation systems to their limits, threatening the very sense of community and reducing the likelihood of water transfers to drier regions.

The Colorado River flows from the high mountain regions of Colorado through Utah and Arizona to the Sea of Cortez in Mexico. Its major tributaries, the Green, flowing from Wyoming, and the San Juan give the Colorado a drainage area of about 629, 370 km^2. The Upper Basin (above Lee Ferry, Arizona), just below the state border with Utah, provides 83% of the annual flow. The Colorado does not discharge a large volume of water with estimated annual flows being about one tenth that of the Columbia and one-eighteenth that of the St. Lawrence in the northeastern United States, both of which drain basins of comparable size. It is, however, an important source of water for seven semiarid states in the western United States and for northwestern Mexico.

Significant human alteration to seasonal streamflow began with the development of the Yuma Valley Irrigation Project, which in 1912 provided irrigation water in Arizona and California. More importantly, the completion of the largest concrete dam in the world in 1935, the Hoover Dam on the Colorado, initiated the golden Age of Dams in the United States (Reisner, 1986). The Glen Canyon Dam (hereafter referred to as the GCD) was completed in 1963, physically dividing the Colorado River into its Upper and Lower Basins. The full reservoir system now stores about four times the annual average streamflow. The Upper Basin provides 80 to 90% of the total flow in the Colorado. The primary role of the GCD is to enable the Upper Basin states of Utah, Colorado, Wyoming, and New Mexico to utilize their apportionment of Colorado River water, while meeting obligations for water delivery to the Lower Basin states of Arizona, California, and Nevada, and also Mexico, consistent with the laws, treaties, compacts, and court decisions regarding Colorado River operations, collectively known as the Law of the River. Decadal-scale climatic factors influencing present water allocations are discussed in greater detail by Stockton and Jacoby (1978). Briefly, the period 1905 to 1930 was the wettest such period in 400 years of record, with 19.8 BCM constructed annual average flow at Lees Ferry. The Colorado River Compact (1922) among basin states used this average as the base minimum for fixed allocation between Upper and Lower Basins. Since the signing of the Compact, the estimated annual virgin flow (1922 to 1997) has been 17.7 BCM, with an historic low flow of 6.9 BCM in 1934. During the 1931 to 1940 and 1954 to 1963 streamflow averaged about 12.6 BCM annually.

Under similar future conditions, if the Upper Colorado River Basin states consume the 9.3 BCM allocated to them by the Compact, they would default on the legal obligation to the Lower Basin. These Lower Basin states have the first right to the allotment. The engineering solution was to construct a dam (the GCD) near Lees Ferry that could store water in wet years and release water in dry ones.

Maintaining geopolitical equity between basins was therefore the major purpose served by the GCD (Ingram et al., 1990). Power generation itself was second to the need to generate revenue for other water projects primarily in the Upper Basin. Decisions in the Colorado River Basin now involve many temporal and spatial scales. A recent study by the consortium of western water resources institutes (i.e., the Powell Consortium) offers the counterintuitive result that while the Lower Colorado River Basin within the United States is indeed drier than the Upper Basin, it is the Upper Basin that is vulnerable to severe, long-term drought because of the 1922 agreement (Powell Consortium, 1995). In addition it was found that, while opportunities for "win–win" situations and rule changes exist, such changes are extremely difficult to implement. Minimum flow requirements are, at present, met with unused entitlements. As with many other river basins, in western North America, there is at present no single decision-making body that encompasses the entire basin.

Water managers have traditionally relied upon the historical record in order to plan for the future, inferring the probability that shortages and floods might occur given their frequency of occurrence in the past. As a result of the climatological droughts experienced during the 1930s, 1950s, and in 1977 (at 7.2 BCM the second driest year in the record), the system as a whole is operated to maximize the amount of water in storage for protection against dry years (see Fig. 3). Total storage within Lake Powell is over three times the annual Upper Basin allotment, or about 31.0 BCM. It should be noted that the region has also experienced sustained periods of high runoff as occurred from 1941 to 1950 and from 1983 to 1986.

The Colorado River, because of the scale of impoundments and withdrawals (including large-scale interbasin transfers), has been called the most legislated and managed river in the world. The river now virtually ends 16 to 30 km before reaching the Sea of Cortez. The impacts of these diversions and storage within Mexico are not well documented, but they have had the effect of disrupting fishing communities located along the river's delta and of decreasing water quality due to increased salinity. The 1944 Water Treaty between these two countries left important problems unresolved in the area of the quality of water delivered by the United States to Mexico. Indeed, the domestic realities at source regions, such as pollution and low flows, have forced the United States to assume costs for desalination of Colorado River water, before it enters Mexico. This has implications for possible long-term reductions in water availability during exceptionally dry periods. Since Mexico's entitlement to Colorado River water is less than 10% of the flow, it appears

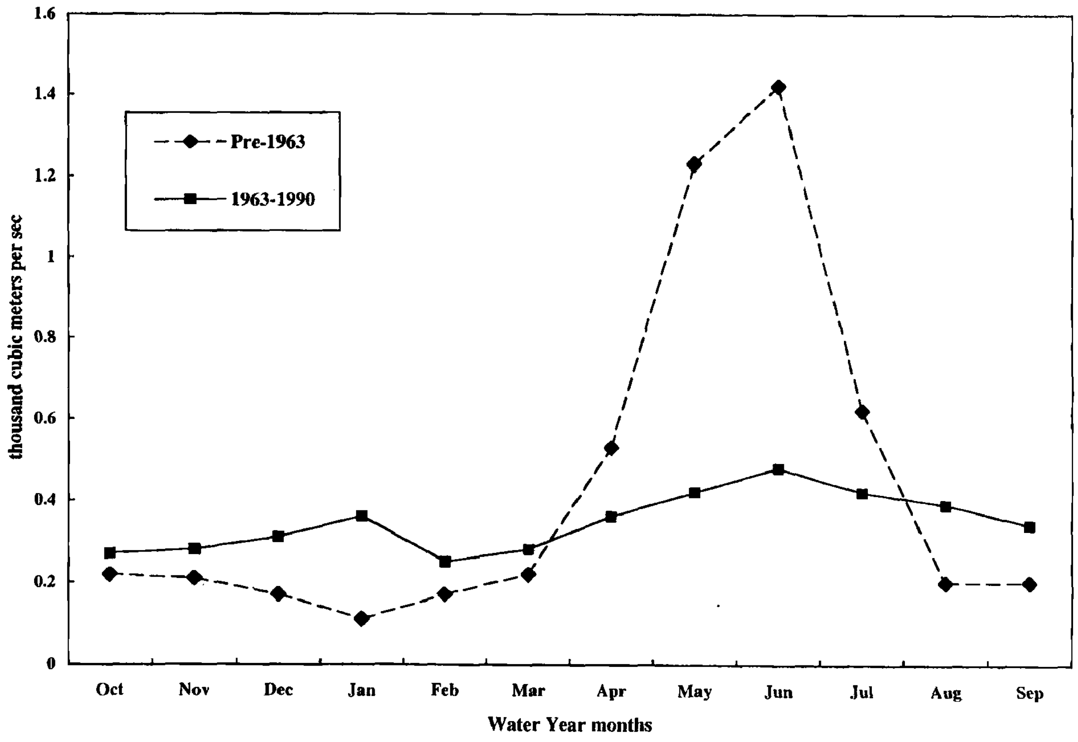


Figure 3 Colorado River discharge at Lee Ferry before and after completion of Glen Canyon Dam in 1963.

unreasonable to expect that Mexico will assume total responsibility for delta restoration. Such stresses are expected to produce serious conditions more often in the future because of the expansion of bilateral trade between the United States and Mexico, the projected rise of development and agriculture in western Sonora, and the full utilization of water allocations in the Upper Colorado Basin anticipated to occur by the year 2010.

5 THE PARANÁ–PARAGUAY RIVER BASIN: INTERANNUAL VARIABILITY AND EXTREME EVENTS

The Paraná–Paraguay River Basin in South America encompasses about 84% of the La Plata Basin with a population of about 100 million (1992 figures). It is the most developed agricultural and industrial zone in South America, accounting for 80% of the economic production in Argentina, Brazil, and Paraguay. The period of relatively small numbers of high floods between 1905 and 1960 coincided with the expansion of urban areas onto the valley bottoms. Three major changes have taken place in the basin since the 1960s (Penning-Rowsell, 1996). First, agricultural and industrial production has increased. The Upper Paraná Basin in Brazil has been converted from coffee plantations, developed in the 1940s, to fields of soybeans and sugar cane (for alcohol fuel production). Second, deforestation to establish cropland and pasture has been extensive in both Brazil and Paraguay. In Paraguay forested regions declined from about 45% in the eastern region in 1945 to about 15% in 1992. In northern Argentina and southern Brazil forest losses range from 60 to 90% in this century. Third, about 20 hydroelectric power plants have been built on the river, significantly changing system hydrology. For example, the width of the Paraná River varies dramatically from 4 km north of Guaira (Brazil) to about 60 m below the Itaipu Dam. As in other parts of the world, the increasing damages caused by extreme rainfall events result from urban and agricultural encroachment onto the floodplains, as much as from the events themselves. Most of these floodwaters (up to 85% on the Paraná) come from the Upper Basin in Brazil. Accompanying these changes are heavy sediment flows from agricultural lands bordering the Paraná.

The Upper Paraguay Basin has low river banks and is prone to flooding, creating a zone known as the 'Pantanal'. Flooding in the Paraná–Paraguay Basin has become both more frequent and more severe in recent years. In particular, the 1983 flooding cost an estimated US$1.8 billion, while 1992 flooding caused serious damage to infrastructure and capital stock resulting in estimated damages of up to US$1 billion and affecting 3.1 million people. The cost estimates of more recent and widespread flooding during the 1998 to 2000 period are as yet incomplete. Ten-year flood discharge rates are now over 15% greater than those in the early twentieth century (Anderson, 1993). In addition, the low-river flows in the watershed have been less frequent and less extreme (i.e., not as dry) in the latter part of the twentieth century. Five of the 10 largest floods recorded during this century (as measured by daily peak discharge) have occurred since 1982. At present El Niño (warm) events may provide the best explanation as a driver of increased precipitation during the rainy season

(March–May). The relative contribution of El Niño as opposed to La Niña–related teleconnections results in a difference of about 20% of annual streamflow. However, it is clear from the record that the general trend in streamflow is upward, and the hydrological regime of the rivers appears to have been changing primarily since 1940. It should be noted that there are also references to high flood events during the nineteenth century.

The human occupancy of these floodplains reflects their economic value for agriculture, communications, and transportation. Flood losses will increase because people are concentrated in flood-prone urban areas, there is increasing migration from rural areas, and land use is poorly regulated. Wet-season rainfall in the region has also been relatively high since the early 1980s. Good relations among Brazil, Paraguay, and Argentina are important for shared use of flood-forecasting data and mitigation strategies, including land-use changes and reforestation. Interagency rivalries within each of these countries restrict the capacity to act internationally, especially under crisis conditions. The prognosis for implementation of sustainable, preventive, nonstructural alternative approaches to reducing exposure (property at risk) and social vulnerability over the long-term does not appear promising.

6 PROBLEMS

Societies are always adapting incrementally and in diverse ways to a variety of integrated and cumulative changes. There is, however, little understanding of the long-term and widespread consequences of these adaptations (Dynesius and Nilsson, 1994). Questions remain as to how and when adaptation will occur and, in particular, how equity and environmental considerations will be addressed over time. A key component has been, and will continue to be, ensuring that the best available and most appropriate scientific information is employed in decision making (Pulwarty and Redmond, 1997). Unfortunately, detailed assessments of the direct human-induced changes of river hydrology of most large river systems are lacking, as are coherent assessments of discharge. Attempts to cope with the negative effects of technological interventions usually follow decades later (L'vovich and White, 1990).

Most major programs for complex utilization of rivers assume that average conditions for hydrologic records available at the time of project planning will continue indefinitely into the future (White, 1997). This has usually resulted in overestimations of supply or underestimations of demand over the long term. The complications of changes in the spatial and temporal distribution of rainfall, soil moisture, runoff, frequency, and magnitudes of droughts and floods have not been explicitly included in response planning. Systems design, operational inflexibility, and legal and institutional constraints reduce the adaptability of water systems to respond to climatic changes (Gleick, 1993). It is therefore difficult to plan for and justify expensive new projects on the basis of supply alone, when the magnitude, timing, and even direction of the changes in basins are unknown (Frederick, 1996).

The major stumbling blocks, exacerbated by shifts in climate or in the frequency of climate extremes, relate to the adversarial relationship that usually develops between upstream and downstream users of water. The determination as to when a particular use is equitable and reasonable involves definitions of broad concepts such as "no harm" and "optimal utilization." Problems are further compounded by lack of agreement on event definitions, such as what constitutes an "extraordinary" (i.e., severe and persistent) drought in different places. The spatial extent and persistence of drought may produce shortages not only in the locale considered but also in neighboring regions that otherwise are supposed to make surplus water available through interbasin tranfers. These concepts appear clear from the standpoint of water measurements, but difficulties emerge in (a) the practical and equitable sharing of quality water or (b) how an upstream country should share water with downstream countries, especially during periods of water stress.

As pointed out by Frederick (1996) and others, in the absence of clear and enforceable property rights, the strongest, most clever, and most advantageously positioned countries can claim and use scarce resources with little concern for the impacts on others. Scarcity does not arise on its own but is dependent on the quantity and quality of resources at particular times and on the degree of access to and capacity to use those resources. The call to the market as a first order of business does not address historical inequities, political differences, and environmental needs. The most powerful usually have the greatest resources to purchase property rights anyway.

All of the river basins described above exhibit characteristics of "closed or closing" water systems. In such systems the management of interdependence becomes a public function; development of mechanisms to get resource users to acknowledge interdependence and to engage in negotiations and binding agreements become necessary. The implementation of such mechanisms does not appear to be viable without focusing events (Keller et al., 1992). Focusing events, such as the La Plata floods, are usually associated with exceptional societal or environmental impacts, highlight critically vulnerable conditions, and elicit highly visible responses. Clusters of historic events may combine with physical events to precipitate or allow particular actions to be undertaken e.g. the cumulative roles of high flows prior to 1930, the 1930s drought, the Great Depression, World War II, and the 1950s drought in influencing dam building and management on the Colorado). Experience shows that decisions bringing rigidity to the management system ultimately generate more problems than they resolve (see in addition to cases cited here, Glantz, 1988; Gunderson et al., 1994). As is evident on the Lower Colorado River Basin and the Nile, early "winners" are unlikely to be willing to alter earlier terms of agreement even when changes in climate conditions are well documented. These problems are further complicated by the unique context imposed by transboundary resources at the borders themselves. According to Ingram et al. (1997), political boundaries, whether domestic or international (1) often separate the location where problems are felt from the location where the most effective and efficient solutions can be applied; (2) make restraint in using scarce water resources unlikely especially when the forces of global economic competition reinforce the focus on opportunities for immediate economic profit; (3) aggravate perceived inequalities; and (4) obstruct grassroots problem solving. In addition, they note that national and state policies are usually at odds with border needs and priorities.

As pointed out by one geographer, "it would be naive to envisage government-sponsored research on the cultural conflicts and politics of water management" (Wescoat, 1991 p. 392). The end result is that in the absence of explicit discussion of conflict the policy process is pushed toward the "technological fix," and time and resources are allocated to achieve near-term tangible results rather than long-term solutions (Caldwell, 1993). Practice thus becomes largely issue specific and incremental with the focus on winners and losers rather than on the development of a consensual vision of a preferred future. There is, in addition, limited experience for managing impacts of severe events, such as persistent drought, in the context of projected rates of development or in the context of closed international water systems. In meeting new challenges and trade-offs brought by a changing and variable environment and societal changes, we simply do not know how (precisely) we must plan.

7 LESSONS

As discussed above, climate and weather events form a variable background on which agreements, conflicts, and politics are constantly being played out. Demographic, political, and environmental changes can and do disrupt existing relationships and current wisdom about the interactions between society and the environment (Glantz, 1994). There have, however, been cases where regional cooperation has led to particular solutions. In 1996 the UN Economic Commission Convention on the Protection and Use of Transboundary Watercourse and International Lakes came into being (Wieriks and Leidig, 1997). Parties to the Convention are obligated to prevent, control, and reduce water pollution, primarily the influx of hazardous materials from point and nonpoint sources.

Lessons provided by the Convention and from earlier treaties (e.g., the 1960 Indus Treaty between Pakistan and India) lead to the following conclusions: (1) international water problems can only be effectively handled on the river basin scale with full acknowledgement of interdependence, (2) river basin management requires an overall integrated approach, including attention to ecological water quality and water quantity issues, (3) international strategies and policies should leave room for flexible implementation, (4) public and political support are also prerequisites for successful formulation, particularly regarding environmental policies, (5) major decisions cannot be taken without input from all stakeholders and ensuring adequate legal basis for participation, and (6) cooperation will not occur without mutual confidence among all parties involved. Most importantly, implementation must be explicitly provided for, and usually does not succeed without some shared vision for the future.

Recognition of variability and change in water resources is a first step in accepting that management occurs under changing conditions in which surprise and uncertainty will always exist. From the brief reviews provided above, several conclusions can be drawn about climate-weather and water relationships: (1) it is unwise to ignore the variability that is inherent in natural systems, since decisions that bring rigidity to a management system can ultimately generate more problems than they resolve; (2) it is important not to ignore changes that have and will occur in social systems; (3) major changes in streamflow can be regarded in retrospect as climate changes, and; (4) careful examination of past seasonal to decadal-scale variability and responses can provide useful organizational lessons for areas with increased or decreased water supply, as may be postulated under different climate change scenarios. Expectations about the future tend to be better understood by people within organizations if there is a clear parallel with the past. Incentives to conserve and opportunities to reallocate supplies as conditions change do not require long lead times, large financial commitments, or accurate information about the future (Frederick, 1996). There is, however, a clear need for exchange of experience and learning among different basins especially on how awareness of slow onset problems in the context of decadal-scale variability is developed and the ways in which societies have adjusted to them. A conspicuous aspect of water management has thus been the lack of careful postaudits (systematic and iterative evaluations) of the social, economic, and consequences of previous programs and ongoing projects (White, 1997).

Few assessments, intended to provide insight into future responses, show sensitivity to historical dimensions. For instance, it is impossible to understand the present context without acknowledging that for most of the twentieth century, following the disintegration of the Ottoman Empire and the post–World War I period, water disputes in the Middle East were closely associated with boundary drawing, state formation, nation building, domestic and international strife, and security issues. Attention to history shows that people have known about most problems for a long time but have not acted on better knowledge of these past changes, i.e., problems have been accumulating for many years not just when they are publicized (Glantz, 1999). Inattention to these changes or engaging in inadequate responses allows the incremental accumulation of problems to the point of system criticality or collapse. Mitigating future impacts requires greater emphases on social and ecological factors that prefigure these "surprises."

As a cautionary note, the idea that many solutions to reducing social and environmental vulnerabilities, cognizant of physical, social, and economic time frames, are available but remain unused is not new (Ascher and Healy, 1990; Pulwarty and Riebsame, 1997). There are no apparent quick fixes, technological, economic, or otherwise (Glantz, 1999). A better understanding of the links between domestic political concerns and foreign policy is needed in order to construct a more complete picture of issues underlying water disputes. One of the most important benefits that may be realized through a comparative study is an understanding of why some policies may be chosen over others, how these are related to particular climatic events, which ones rise to prominence, and which are allowed to persist. Identification of the barriers to implementation and evaluation in one setting may shed light on the likelihood of success of similar actions in another setting.

8 IMPORTANCE OF LINKING HUMAN AND PHYSICAL ASPECTS

Cultural, political, and economic conflicts are the most serious problems encountered in transboundary river issues. These may be exacerbated by environmental stresses and by the actions chosen in response. Gilbert White (1966) distinguished between the theoretical and practical ranges of choice in structuring the analysis of adjustment decisions. The physical environment at a given stage of technology sets the theoretical range of choice open to any resource manager or group. The practical range of choice is set by culture, institutions, types of analytical tools employed, etc., which permit, prohibit, or discourage a given choice. Water has thus been defined as a property of territorial units in the legal setting, as a natural resource transformable into products for human consumption in an engineering setting, and as a commodity that can be exchanged and traded between various places and various uses in an economic setting (Blatter and Ingram, 2001). The technological response is to regularize the flow and expand the total amount of available water. The recent experience on dispute resolution shows a contrast between an abiding belief in the rationality of public decision making on the part of some participants while others exhibit sharp suspicion toward the politics of "expert" managerial discourses (e.g., the commodification of water vs. communal values). Emerging knowledge of the complexity and varieties of meaning characterizing "fresh water" at the end of the twentieth century has led to greater appreciation of the limitations as well as the benefits of technological, legal, and economic approaches. In no other context are the divergent meanings of water likely to be more contested than in transboundary situations (Blatter and Ingram, 2001).

Attention has begun to focus on how impacts of chosen approaches exacerbate the root causes of social and ecological vulnerability. There is increasing appreciation for explicit consideration of peoples affected by decisions but who are usually excluded from participation or from the benefits of developed infrastructure (see Pulwarty and Riebsame, 1997; Milich and Varady, 1998). Reduction of vulnerability requires careful assessments of the range of alternative adjustments, among which societies may choose in arriving at a suitable plan for a given period (White, 1977). There is thus an ongoing need for guidance in integrating equity with efficiency considerations in transboundary water management through studies of place-based historical and cultural uses, understanding the role of public trust, the impact of new technologies, and of the flexibility provided by market-based approaches after basic human needs and environmental requirements are met. In particular more work needs to be done on the trade-offs involved between calls for increased participation and the formation of consensus. Promising partnerships are emerging but are in their early stages (see Milich and Varady, 1998).

From the perspective of the climate and water sciences, researchers, through ongoing dialog and joint studies, should engage practitioners as full partners to uncover issues of mutual significance, explicitly address uncertainties in both the scientific and decision domains, and to understand and overcome barriers to information use contingent in each situation (Pulwarty and Melis, 2001). The goals are to have better matches among what is needed, what is asked for, what is available, and what actions can be taken. These processes must be embedded within an understanding of the decision contexts (historical, policy, and operational) within which trade-offs take place.

Water by its very nature tends to introduce even hostile co-riparians to cooperate even as disputes continue over other issues (Wolf, 1999). At the international level the weight of historic evidence tends to favor water as a catalyst for cooperation for particular ends. This has not been the case on the subnational scales. While governing institutions that more closely correspond with the physical water system can help to assure appropriate consideration of efficiency and equity, domestic policy can pose major institutional barriers to international agreements and management across national borders. Ultimately, the main tasks in the foreseeable future will be uncovering how to share common but variable water resources in a catchment area between upstream and downstream users, between various sectors, between rural and urban areas, between preservation of functioning ecosystems, and more direct tangible needs (Falkenmark and Lundqvist, 1995). Engaging the many dimensions of transboundary river flow requires, more than ever, the need to understand these "regions" as integrators of social, cultural, climatic, economic, and ecological histories and networks, that help to shape shared community interests and values.

REFERENCES

Anderson, R., *An Analysis of Flooding in the Paraná/Paraguay River Basin World Bank*, LATEN #5, World Bank, Washington, DC, 1993.

Appelgren, B., and W. Klohn, Management of transboundary water resources for water security: Principle, approaches and state practices, *Nat. Resour. Forum*, *21*, 91–100, 1997.

Ascher, W., and R. Healy, *Natural Resource Policymaking in Developing Countries*, Duke University Press, Chapel Hill, NC, 1990.

Blatter, J., and H. Ingram (Eds.), *Reflections on Water: New Approaches to Transboundary Conflicts and Cooperation*, MIT Press, Cambridge, 2001.

Caldwell, L., Emerging boundary environmental challenges and institutional issues: Canada and the United States, *Nat. Resour. J.*, *33*, 10–31, 1993.

Correia, F., and J. da Silva, International framework for the management of transboundary water resources, *Water Int.*, *24*, 86–94, 1999.

Diaz, H., and R. Pulwarty, A comparison of Southern Oscillation and El Niño signals in the tropics, in H. Diaz and V. Markgraf (Eds.), *El Niño: Historical and Paleoclimate Aspects of the Southern Oscillation*, Cambridge University Press, 1992, pp. 175–192.

Dinar, A., Economic and social considerations of regional cooperation in River Basin comprehensive water resources development, keynote presented at Nile 2002 Conference, Addis Ababa, Ethiopia, February 24–28, 1997.

Dynesius, M., and C. Nilsson, Fragmentation and flow regulation of river systems in the northern third of the world, *Science*, *266*, 753–762, 1994.

Falkenmark, M., and J. Lundqvist, Looming water crisis: New approaches are inevitable, in L. Ohlsson (Ed.), *Hydropolitics: Conflicts over Water as a Development Constraint*, Zed Books, London, 1995.

Frederick, K., Water as a source of international conflict, *Resources*, *123*, 9–19, 1996.

Georgakakos, A., W. Klohn, and K. Georgakakos, A decision support system for the Nile River, in *Proc. 5Th Nile 2002 Conference*, Addis Ababa, Ethiopia, February 24–28, 1997.

Glantz, M., H., *Creeping Environmental Problems and Sustainable Development in the Aral Sea Basin*, Cambridge University Press, Cambridge, 1999.

Glantz, M. H. (Ed.), *The Role of Regional Organizations in the Context of Climate Change*, Springer-Verlag, Amsterdam, 1994.

Glantz, M., *Societal Responses to Regional Climatic Change: Forecasting by Analogy.* Westview Press Colorado, 1988.

Gleick, P. (Ed), *Water in Crisis*, Oxford University Press, New York, 1993.

Gunderson, L., C. Holling, and S. Light, *Barriers and Bridges to the Renewal of Ecosystems and Institutions.* Columbia University Press, 1994, pp. 3–34.

Ingram, H., L. Milich, and R. Varady, Managing transboundary resources: Lessons from Ambos Nogales, *Environment*, *36*, 6–38, 1994.

Ingram, H., D. Tarlock, and C. Oggins, The law and politics of the operation of Glen Canyon Dam, in *Colorado River Ecology and Dam Management*, National Research Council, National Academy Press, 1990, pp. 10–27.

Kaufman, E., J. Oppenheimer, A. Wolf, and A. Dinar, Transboundary fresh water disputes and conflict resolution: Planning an integrated approach, *Water Int.*, *22*, 37–48, 1997.

Keller, J., N. Peabody, D. Seckler, and D. Wichelns, *Water Policy Innovations in California.* Center for Economic Policy studies, Winrock International 1992.

L'vovich, M., and G. White, Use and transformation of water systems, in Turner et al. (Eds.), *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years*, Cambridge University Press, 1990; pp. 235–252.

McCaffrey, S., Water, politics and international law, in P. Gleick (Ed.), *Water in Crisis*, Oxford University Press, New York, 1993, pp. 92–104.

Milich, L., and R. Varady, Managing transboundary resources: Lessons from transboundary accords, *Environment*, *40*, 10–41, 1998.

Naiman, R., J. Magnuson, D. McKnight, and J. Stanford, *The Freshwater Imperative: A Research Agenda*, Island Press, Washington, DC, 1995.

National Academy of Science (NAS), *Alternatives in Water Management*, National Academy Press, Washington, DC, 1968.

Penning-Rowsell, E., Flood-hazard response in Argentina, *Geogr. Rev.*, *86*, 72–90, 1996.

Postel, S., *Last Oasis: Facing Water Scarcity*, W. W. Norton, New York, 1997.

Powell Consortium, Severe sustained drought: Managing the Colorado River in times of water shortages, *Water Resour. Bull. Spec. Issue*, *31*(5), 1995.

Pulwarty, R., and T. Melis, Climate extremes and adaptive management on the Colorado River. *J. Environ. Mgmt.*, *63*, 307–324, 2001.

Pulwarty, R., and K. Redmond, Climate and salmon restoration in the Columbia River basin: The role and usability of seasonal forecasts, *Bull. Am. Meteorol. Soc.*, *78*, 381–397, 1997.

Pulwarty, R., and W. Riebsame, The political ecology of natural hazards, in H. Diaz and R. Pulwarty (Eds.), *Hurricanes: Climate and Socio-Economic Impacts*, Springer-Verlag, 1997.

Reisner, M., *Cadillac Desert*, Penguin Books, 1986.

Riebsame, W. E., K. M. Strzepek, J. L. Wescoat, Jr., R. Perritt, G. L. Gaile, J. Jacobs, R. Leichenko, C. Magadza, H. Phein, B. J. Urbiztondo, P. Restrepo, W. R. Rose, M. Saleh, L. H. Ti, C. Tucci, and D. Yates, Complex river basins, in K. Strzepek and J. Smith (Eds.), *As Climate Changes: International Impacts and Applications*, Cambridge University Press, 1996, pp. 57–91.

Wescoat, J. L., Managing the Indus Basin in light of climate change, *Global Environ. Change*, *1*, 381–395, 1991.

White, G. F., Formation and role of public attitudes, in M. Jarrett (Ed.), *Environmental Quality in a Growing Environment*, Johns Hopkins Press, Baltimore, MD, 1966, pp. 105–127.

White, G. F., *Environmental Effects of Complex River Development*, Westview, Boulder, CO, 1977, pp. 1–22.

White, G. F., Watersheds and streams of thought, in H. Barakat and A. Hegazy (Eds.), *Reviews in Ecology: Desert Conservation and Development*, Cairo, Egypt, 1997, pp. 89–98.

Wieriks, K., and A. Schulte-Wulwer-Leidig, Integrated water management for the Rhine, *Nat. Resour. Forum*, *21*, 155–156, 1997.

Wolf, A., Conflict and Cooperation along International Waterways, *Water Policy*, 1&2, 251–265, 1998.

CHAPTER 48

LESSONS FROM THE RISING CASPIAN

IGOR S. ZONN

1 INTRODUCTION

The Caspian Sea is the biggest inland body of water in the world. Its surface area is roughly equivalent to the combined area of the Netherlands and Germany (about 400,000 km^2, or 144,000 mi^2). The surface water inflow into the sea is formed by the flow of the Volga, Ural, Terek, Sulak, Samur, Kura, small Caucasian rivers, and Iranian rivers. The watershed area of the Caspian Sea is 3.5 million square kilometers. The basin of the Volga River makes up nearly 40% of the territory of the catchment of the Caspian Sea, and it supplies about 80% of the total volume of annual water flow into the sea. All components of the Caspian ecosystem, directly or indirectly, to a greater or lesser extent, are influenced by river flow.

The Caspian Sea basin falls into three morphologically different parts: (I) the northern (25% of the sea area), a shallow area (less than 10 m deep; about 20% with depths less than 1 m) extending to a conventional line passing from the Terek river to the Mangyshlak Peninsula; (II) the medium (35%), with an average depth of 170 m (the maximum being 790 m); and (III) the southern (39%), the deepest area, with a maximum depth of 1025 m and an average depth of 325 m (see Fig. 1). Deep depressions in the northern and southern parts of the sea are divided by an underwater threshold running from the Apsheron Peninsula to Turkmenbashi (formerly Krasnovodsk) (Kosarev and Yablonskaya, 1994).

Before the breakup of the Soviet Union in December 1991, the USSR and Iran were the only two independent nations occupying the shores of the Caspian. With the breakup, three additional newly independent nations emerged along the coast: Azerbaijan, Kazakstan, and Turkmenistan. The Russian Federation's Caspian coastline is shared by three of its political units: Astrakhan Oblast, the Republic of Kalmykia, and the Republic of Dagestan.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

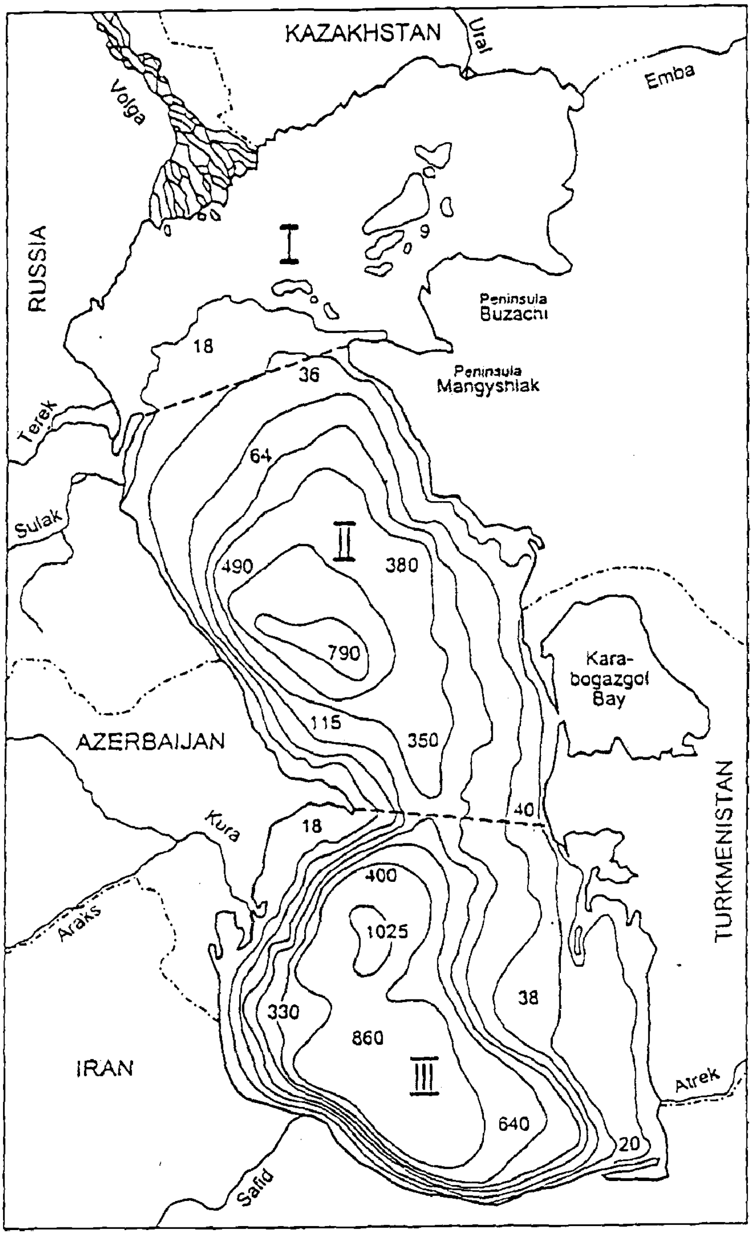


Figure 1 Depth isolines for the Caspian Sea, in meters.

2 NATURE OF SEA-LEVEL CHANGES IN CASPIAN SEA

The Caspian Sea is a closed basin in the inland part of Eurasia and this sea's water level is below that of the world ocean. The sea basin stretches almost 1200 km from north to south and its width varies between 200 and 450 km. The total length of the coastline is about 7000 km. Its water surface area is about 390,600 km^2 (as of January 1993). Water salinity in the northern part is 3 to 6‰, and reaches 12‰ in the middle and southern parts.

Fluctuations in sea level for various lengths of time can be found in the data of geomorphological and historical studies of the record of the Caspian Sea (Fig. 2). Within the last 10,000 years, the amplitude of fluctuations of Caspian Sea level has been 15 m (varying from −20 to −35 m). During the period of instrumental observations (from 1830 onward), this value was only about 4 m, varying from −25.3 m during the 1880s to −29 m in 1977. Annual increases in the level during this period met or exceeded 30 cm on three occasions (in 1867, 30 cm; in 1979, 32 cm; and in 1991, 39 cm). The mean annual increment in the level in the 1978 to 1991 period was 14.3 cm.

Natural factors are the primary cause of recent Caspian Sea level fluctuations (but not the only cause). Scientists have identified three distinct periods of level changes: 1830 to 1930, 1931 to 1977, and 1978 to the present. The first period of 100 years saw sea-level fluctuations not exceeding 1.5 m (5 ft). Researchers considered this period to have been relatively stable. The second period, from 1931 to 1977, is identified by a constant decline in level by 2.8 m (9.1 ft), and in 1977 the Caspian Sea reached its lowest level since the beginning of instrumental record-keeping in the 1830s.

As the sea level declined throughout the 1950s, 1960s, and early 1970s, Soviet scientists forecast that the decline would continue for at least a few decades into the future. Scientists have linked the reason for the decline to the regulation of Volga River flow. During these decades, major engineering activities were undertaken along the Volga, such as the construction of water diversion canals, reservoirs,

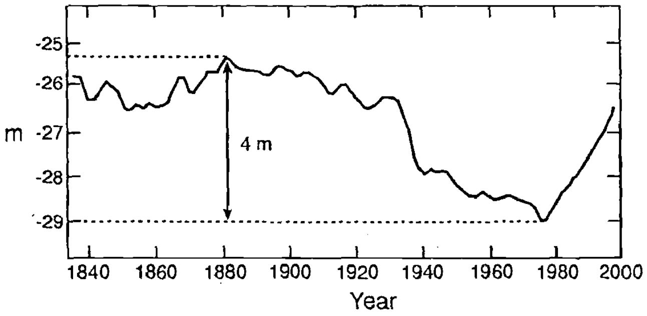


Figure 2 Caspian Sea level, 1835–1999, observed.

and dams. The construction of such engineering facilities diverted water away from the Caspian.

In response to this major drop in sea level, human settlements bordering the sea coast began to move toward the receding coastline. Fields and pasturelands were prepared for use, roads and rail lines were constructed, and housing and factories were built on the newly exposed seabed. During the Soviet era, many people emigrated from other parts of the region to settle along the border of the sea. Development of infrastructure along the coast took place to support the increasing population.

In an attempt to save the Caspian from drying out, Soviet scientists and engineers proposed the construction of a dam to block the flow of Caspian water to Kara-Bogaz-Gol Bay, a large desert depression in Turkmenistan adjacent to the sea's eastern shore. Political decisions made in the mid-1970s ordered the construction of the dam, but due primarily to bureaucratic inertia, the dam was not completed until the early 1980s. This was a few years after the Caspian's sea-level change had reversed direction. Before the dam was constructed, the bay took in 40 km^3 (8.6 mi^3) of Caspian water annually. It served as a huge evaporation pond, as well as a natural location for the accumulation of commercially useful mineral salts.

Another Soviet government response to the decline in the Caspian's sea level was a diversion of water into the Volga River from other Soviet rivers that flowed northward into the Arctic Ocean. River water flowing into the Arctic was viewed as wasted and without value to the Soviet Union because it was unused by human activity.

3 THE CASPIAN RISES

To the surprise of Soviet scientists, the level of the sea began to rise suddenly in 1978, the beginning of its third period of level changes. Since then, the Caspian has risen steadily by more than 2.5 m. One of the first actions the newly independent government of Turkmenistan took in 1992 was to tear down the dam in order to allow great amounts of water to flow into Kara-Bogaz-Gol Bay again and to replenish the supply of salts.

Scientists have proposed a variety of hypotheses about why the Caspian Sea level had increased so rapidly. These can be clustered into the following categories: tectonic plate movement on the seabed, climate fluctuations and change, and hydraulic construction along the Volga River, or some combination of these factors.

Tectonic Plate Movement Hypothesis

Tectonic movements over periods such as centuries and millennia have been the cause of many geologic changes in the Caspian basin. The region has been subjected to uplift, subsidence, overthrust of landforms, seabed mud-volcanic activity, and landslides, in addition to erosion processes and the accumulation on the Caspian seabed of river-transported sediments. However, it is difficult to see how tectonic movements could cause such sharp fluctuations in the Caspian's sea level over relatively short periods. Thus, it appears that such movements have had an insignificant impact on recent sea-level fluctuations.

Climate Change Hypothesis

Today, most Russian scientists believe that climatic factors are the real cause of the Caspian Sea level rise. Studies by Golitsyn (1989) and Golitsyn and McBean (1992) indicate that recent changes of the Caspian Sea level are 90% associated with corresponding changes in the water balance components of the sea, as opposed to possible tectonic activity. The volume of inflow from rivers to the sea increased sharply after 1978. During certain years (e.g., 1979, 1985, and 1990), more than 350 km^3 of river water entered the sea. From 1978 until 1990, Volga River flow exceeded 260 km^3/yr. At present, no arguments have challenged the view that the main contribution to seasonal and annual level fluctuations of the Caspian is accounted for by surface inflow and evaporation levels. Within recent decades, the sea's fluctuations have been subjected to anthropogenic impacts as well.

In this regard, climate has two dimensions: climate fluctuations and climate change. Climate fluctuations occur on various time scales, with those of interest to present-day society being on the order of decades and perhaps centuries. Climate-related fluctuation refers to the increase and decrease of sea level over the course of decades. During the past two centuries, the sea has undergone several fluctuations. Those of the twentieth century have adversely affected socioeconomic activities and infrastructure along the sea's coastline.

The view that climatic processes in the Volga basin are the dominant cause of sea-level fluctuations has been recently reinforced. Droughts in this basin and sharply reduced Volga flow into the Caspian from mid-1995 until early 1997 have been associated with a 25-cm (10-inch) drop in Caspian level. Nevertheless, Russian scientists still suggest that the sea level will continue to rise into the first decades of the twenty-first century.

Climate change associated with global warming induced by human activities has also been proposed as the forcing factor behind the Caspian's rise since 1978. Those who see global warming as the forcing factor suggest that the most recent sea-level rise can be associated with intensification of the hydrologic cycle (i.e., more active precipitation-producing processes), an intensification that some scientists have linked to the human-induced global warming of the atmosphere. An increase in precipitation within the Volga River basin would translate into increased sea level.

Hydraulic Construction Hypothesis

Some observers have argued that the recent fall and rise in sea level were the result of human activities. They suggest that the widespread development of hydraulic structures (e.g., dams, reservoirs, irrigation systems) in the Volga River basin, beginning in the 1950s, led to a sharp decline in Volga flow. The filling of many reservoirs built along the rivers flowing into the Caspian, the increase in industrial and municipal water use by several times, and changes in the water regime of the floodplains led to a decrease of streamflow into the sea. Such a hypothesis could be tested by constructing a water budget model for the Caspian. Such a model would need to identify all the inflows into the Caspian Sea (such as from rivers and groundwater) and all outflow from the sea (such as evaporation and water diversions). While it is a seemingly straightforward task, identifying all the sources and sinks of Caspian water is not easy.

There is also a hypothesis about an Aral Sea connection. Yet another suggestion that seems to be made at just about every Aral or Caspian Sea conference is that the decline in the level of the Aral Sea is linked to the rise in level of the Caspian. The reasoning is that water diverted from the Aral basin to the Caspian basin to irrigate the desert sands for cotton production in Turkmenistan ends up either being evaporated into the air or seeping into the groundwater, which eventually makes its way into the Caspian. However, it is important to point out that *both* the recent fall and rise in the Caspian Sea level occurred during three and a half decades of a constant decline in the Aral's level.

4 SOCIETAL IMPACTS OF SEA-LEVEL RISE

According to a UN Environment Programme estimate, the cost of the impact of the sea-level rise of the Caspian, as of 1994, was $30 to $50 billion (US). Coastal ecosystems have been destroyed, villages inundated and populations evacuated, sea banks eroded, and buildings destroyed. Coastal plains have been invaded by subsurface seawater or have become waterlogged. Fauna have changed, and pasturelands and sturgeon spawning grounds have been destroyed.

Each of the five countries sharing the coasts of the Caspian Sea has suffered losses, and those losses increased until the mid-1990s. They suffer from the different impacts of sea-level rise because the territory along its coastline is neither uniformly settled nor uniformly developed economically. Economic losses in the big cities and villages have been higher than in the rural areas. More specifically, in Astrakhan Oblast (equivalent to an American state), about 10% of its agricultural land was out of production by 1995 because of sea-level rise. The coastline of the Republic of Dagestan (also part of Russia) was affected by the flooding of at least 40 factories in its cities of Makhachkala, Kaspiysk, Derbent, and Sulak. Nearly 150,000 hectares (370,000 acres) of land have been inundated, with a loss of livestock production and breeding facilities. Much of the 650-km (390-mile) Caspian coastline of Turkmenistan is made up of low-lying sandy beaches and dunes that are vulnerable to coastal flooding and erosion. In fact, some Turkmen villages that were once several kilometers from the sea are now coastal communities. Similar adverse impacts of sea-level rise on human settlements and ecosystems are found in Kazakstan, Azerbaijan, and Iran.

The Caspian has been referred to as a "hard currency sea" because of its large oil and natural gas reserves and because of its highly valued caviar-producing sturgeon. Regional reserves contain upward of 18 billion metric tons of oil and 6 billion cubic meters (215 billion cubic feet) of natural gas. Experts suggest that the Caspian is second only to the Persian Gulf with respect to the size of its oil and gas reserves, and that Turkmenistan is a "second Kuwait." If the sea level were to continue to rise, a large part of the oil and gas mains along the Turkmen coast would become submerged and would also be subjected to corrosion by seawater. Coastal settlements, which include the greater part of Turkmenistan's oil, gas, and chemical enterprises, would also be threatened. Similar environmental problems would certainly affect other Caspian coastal countries as well (e.g., Ragozin, 1995).

The Caspian Sea is unique in yet another respect: It contains about 90% of the sturgeon that produce the lucrative prized black caviar for export to foreign markets. Sturgeon roe is often referred to as "black gold." Today, however, Caspian sturgeon is at risk of extinction from overexploitation by illegal poachers and by destitute fishermen desperately seeking funds to buy food for their families. The sea-level rise, with its destruction of sturgeon spawning grounds, adds yet another threat to the endangered Caspian sturgeon.

Poachers hunt sturgeon only for its caviar. Today, they catch sturgeon directly in the open sea. However, in the early 1960s, prohibition was introduced by the former USSR against catching sturgeon in the open sea. Since that time, catching sturgeon has been carried out in the river deltas. Sturgeon reproduce very slowly: The fish do not spawn for the first time until they reach the age of 20 to 25 years. In 1990, the permissible catch of sturgeon in the USSR was set at 13,500 metric tons. In 1996, permissible (legal) catch was only 1200 metric tons (Rosenberg, 1996).

5 SEA-LEVEL CHANGE AS A GLOBAL PROBLEM

Given the growing concern about, and possible evidence of, global warming, there has been considerable speculation about the potential impacts on coastal areas of a sea-level rise related to global warming. Scientists who participated in the 1995 Intergovernmental Panel on Climate Change (IPCC) Report (IPCC, 1996) suggested that global sea level may well increase by an additional 15 to 70 cm (6 to 27 inches) by the end of the 21st century. The exact amount of rise would depend on the actual increase in global temperatures. Clearly, any additional increase in sea level could have devastating consequences for coastal communities.

All states that border bodies of water, whether along the global oceans or inland seas, should pay attention to fluctuations in sea level as well as to the rise in sea level linked to global warming. Inland seas, for example, can be viewed as living bodies in the sense that they can expand and can shrink. These changes can occur on different time scales: from daily to seasonally, from a year to a decade, or a century, or a millennium. In fact, they fluctuate and change on all these scales. The same can be said of the open oceans, but they tend to fluctuate on much longer time scales than do the inland seas, over periods of many decades and centuries. Such time scales are difficult to factor into the thinking of economic development planners, whose time frames are on the order of years to a few decades at most.

In essence, one can consider the Caspian as a laboratory of sea-level change and its potential societal and environmental consequences. For the Caspian to serve as a true "laboratory," its environmental-monitoring network, which collapsed with the breakup of the Soviet Union, must be restored and maintained by regional cooperation among the Caspian states. Impacts on ecosystems that are managed (farms and pastures) and unmanaged (wetlands, forests, deserts) can be identified. Effective human responses to changes in the coastal zone (both land and sea) can also be identified and assessed; environmental engineering proposals to deal with sea-level changes (such as seawall construction, higher oil platforms in the sea, diversion of water from the Caspian to the drying Aral Sea) can be evaluated for effectiveness, taking into consideration the scientific uncertainties surrounding sea-level fluctuations.

Whether the global climate gets warmer, cooler, or stays as it has been for the last several decades, the level of inland seas will likely continue to fluctuate (the mean ocean level has already gone up by 5 to 6 inches in the twentieth century alone). Societies must learn to cope with both short- and long-term fluctuations. In the Middle Ages, people in the Caspian region were not allowed to settle too close to the sea's shore, under the threat of death. Apparently, leaders were then aware of the dangers that the Caspian's fluctuating levels posed to their citizens. Today's leaders would be well advised to pay attention to traditional wisdom.

REFERENCES

Golitsyn, G. S., Once more about the changes of the Caspian Sea level, *Vestnik AN SSR*, *9*, 59–63, 1989.

Golitsyn G. S., and G. A. McBean, Changes of the atmosphere and climate, *Proc. Russian Acad. Sci. Geogr. Ser.*, *2*, 33–43, 1992.

IPCC, *Climate Change 1995: The Science of Climate Change*, Cambridge University Press, Cambridge, UK, 1996.

Kosarev, A. N., and E. A. Yablonskaya, *The Caspian Sea*, SPB Academic Pub., The Hague, 1994.

Ragozin, A. L., Synergistic effects and consequences of the Caspian Sea level rise, in *Proceedings of the International Scientific Conference: Caspian Region: Economics, Ecology, Mineral Resources*, Geocenter- Moscow, 1995, pp. 120–121 (in Russian).

Rosenberg, I., Catching sturgeon, *Itogi Magazine*, 4 June 1996 (in Russian), 48–50.

CHAPTER 49

ACID RAIN AND SOCIETY

PAULETTE MIDDLETON

1 INTRODUCTION

Acid rain is one of many manifestations of how actions of society can have adverse effects on human health and welfare. Now more than ever before, the breadth of socioeconomic as well as environmental impacts associated with air pollutants, connections among pollutant contributions to these many impacts, and the implications of these connections are being recognized for policy making and development of management strategies.

It can no longer be argued that it is very costly to mediate acid rain and related air quality concerns. Assessments are beginning to suggest that multiple benefits associated with addressing acid rain in combination with other issues outweigh the costs of control of key responsible pollutants. In addition, when innovative strategies, which include market trading and incentives for conservation and use of clean fuels, are initiated, the costs of pollutant management become even lower. As factors that are not easily quantified monetarily are considered more directly in assessments, the benefits become even greater.

Recent analyses show that the implementation of the acid-rain-related part of the 1990 Clean Air Act Amendments has resulted in reductions in acidity in the northeastern United States. Improvements in acid-related impacts have also been suggested. However, projections of future conditions over the next 20 to 50 years suggest that, unless more dramatic steps are taken, the overall burden of harmful pollutants could continue to rise in general in different parts of the country. Dramatic reductions in sulfur oxides, the current main contributors to acidity, alone are probably not enough. Planned reductions in nitrogen oxides may or may not be adequate. Similarly, continued close monitoring, if not increased management, of volatile

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

organic compounds and fine particulates, and reassessment of their importance to acid rain and related concerns will be important over the next few years.

Of course, all of these issues are part of the bigger international picture of energy, environment, and economy. While the United States may be more aggressively and wisely addressing acid rain and related issues here at home, the projections for future fossil fuel use worldwide must be considered for the sake of regional air quality as well as the global climate condition. Capital investment in cleaner technologies such as renewable energy and the promotion of conservation strategies worldwide could bring long-term environmental and economic benefits that far surpass the initial costs. The alternative, continued growth in fossil fuel usage in developing countries, could exacerbate air quality problems that already exist in many of these areas and, in the long term, could cause the same acid rain damages experienced in many parts of North America and western Europe. In addition, it would contribute to adverse long-term carbon-dioxide-induced climate change.

As a contribution to our understanding of the atmospheric pollution problem and our role in the solution, this chapter summarizes:

- Acid rain and its relationship with other major issues
- U.S. response to the acid rain issue
- Current assessment of progress on reducing the effects of acid rain
- Projections and speculation on the future of acid rain

2 ACID RAIN: THE PHENOMENON

The relationships between chemical emissions into the atmosphere and the effects of the chemicals on various ecosystems, human health, and materials are highly complex. Many harmful chemicals (i.e., air pollutants) chemically interact to form other pollutants, which are perhaps even more harmful than the originally emitted chemicals. The most prominent examples of these dangerous chemical products are acid rain, ozone, and aerosols. In addition, many air pollutants are thought to interact in a synergistic fashion to cause even more harm as a group rather than individually. An example of the possible synergism is the hypothesized impact of acid rain and ozone on forest ecosystems. Understanding the causes and effects of acid rain and related air quality issues has become an important mission of atmospheric scientists around the globe.

3 DEFINITION OF ACID RAIN

Acid rain is the general term used to describe the removal, by rainfall, of acidic pollutants from the atmosphere. Acids also can be removed by other forms of precipitation, such as snow or fog. Acid pollutants may also fall as dry particles or gases that form acids when later combined with moisture. The term *acid deposi-*

tion is used to include all the possible forms of acid pollutant removal from the atmosphere, but *acid rain* remains the popular term.

The majority of the deposited acids are nitric acid and sulfuric acid. In some of the more rural regions of the world, organic acids also are important. In very remote areas where the level of acid is low, the "natural background" acid is carbonic acid, which is associated with carbon dioxide in the air. The overall acidity of precipitation also depends on the basic (or alkaline) constituents of the precipitation. Major bases include ammonia and geologic materials, such as dust and fly ash.

Acidity is measured in terms of a pH scale, which is a measure of the log of the hydrogen ion concentration in the precipitation. The scale runs from 0 to 14 with 0 being very acidic and 14 being very alkaline. A midscale value of 7 is considered neutral. A change in 1 pH unit indicates 10-fold increase or decrease in acidity. Unpolluted rain water is considered to have a pH of about 5.6. This acidity is assumed to contain only carbonic acid. In the highly polluted eastern states of the United States., the average acidity of water has a pH between 4 and 5. Even in some remote areas of the world, pH values of 5.2 have been found. These acidity levels suggest that there is a long-range transport of nitrogen and sulfur chemicals.

4 SOURCES OF ACIDITY

Atmospheric acids mainly are produced in the air as a result of complex chemical reactions of the acid precursor gases. Direct emissions of acids such as sulfuric acid, hydrogen chloride, and hydrogen fluoride have been estimated but are not thought to play a significant role in the acidic deposition processes (NAPAP, 1991). Sources of the harmful chemical precursors affecting the acidity of deposited materials can be natural or human caused. The three pollutants of most concern in the acidic deposition process are sulfur dioxide (SO_2), nitrogen oxides (NO_x) and volatile organic compounds (VOCs). Fossil-fuel-based power plants and motor vehicles are the major sources of all of these acid precursor pollutants in industrialized areas of the world (Graedel et al., 1993; Middleton, 1995).

Sulfur gases are primarily emitted from point sources, involving the combustion of coal in particular. Natural sources of sulfur gases include sea spray, volcanoes, and biologic activity. These sources, however, are at least a factor of 10 *less* than the human-caused emission for major industrial areas such as the United States. Nitrogen gases also result primarily from human activities involving fossil-fuel-derived energy use related to transportation and utilities. Major natural sources include soils and lightning and are thought to make up a significant portion of the overall emission totals, more than has been estimated for the sulfur gases. Estimates of these levels, however, are uncertain (NAPAP, 1991). The VOCs that produce the organic acids and influence the chemistry producing sulfuric and nitric acids also come mainly from automobile use. However, for the VOCs, natural production from vegetation can be quite significant. In highly vegetated, low industrialized regions natural sources become the dominant producers of VOCs.

Estimates of alkaline particulate and ammonia emissions indicate that there is a high potential for acid neutralization in some parts of the United States. The estimates, however, are subject to a high degree of uncertainty (NAPAP, 1991). On a global scale, emissions of these important acid neutralizers are among the least well-known chemical emissions (Graedel et al., 1993).

5 EFFECTS OF ACID RAIN

The effects of acid deposition are the subject of continuing controversy. The northeastern United States has experienced the worst reported impacts in the United States (NAPAP, 1991). Severe damages attributed to acid rain also have been documented for parts of western Europe (Graedel and Crutzen, 1989).

The most sensitive systems to acid deposition are poorly buffered lakes and streams. Buffering capacity refers to the availability of alkaline minerals from soil or rocks to neutralize the acids. When minerals are dissolved in a lake, buffering is able to diminish acid effects. However, this buffering ability or alkalinity can be used up with additions of acidic pollutants. Low alkalinity lakes have the greatest potential for damage, since their neutralizing minerals can be quickly depleted.

Vegetation is exposed to wet acidic deposition through rain, snow, and by direct contact with low, acid-laden clouds. There is currently no widespread forest or crop damage in the United States related to these possible pathways. However, cloud acidity, together with a complex combination of other factors (e.g., ozone, soil acidification, climate) contribute to reduced cold tolerance in high-elevation spruce in the eastern United States and in Europe. This can contribute to damage to trees above cloud level during winters with particularly low temperatures.

Adverse effects on forests in other regions of the world are associated with ozone, as is the case with high-elevation pines in California, or they are closely related to localized soil nutrient deficiencies, as is the case with sugar maples in eastern Canada. Acidic deposition may increase leaching rates of important base cations, principally magnesium and calcium, in forest solids and may be a contributing factor in sugar maple decline in some areas.

Generally, controlled experiments on trees and crops have indicated that ozone, at concentrations near ambient levels, adversely affects forests and crops primarily by growth reduction. Other controlled experiments have demonstrated that normal levels of atmospheric sulfur and nitrogen deposition cause no negative direct effects. Some areas actually may benefit through nutrient enrichment by nitrogen and sulfur deposition.

Computer models project that continued acidic deposition could result in long-term deficiencies of nutrients in some soils. However, currently, there is no evidence to indicate that forest health in general is currently affected by nutrient deficiency or will be affected in the next half century.

Air pollution and acidic deposition contribute to the corrosion of metals and deterioration of stone in buildings, statues, and other cultural resources. Although air pollution is an important concern for cultural objects, the magnitude of its effect on construction materials has been difficult to assess. Many construction materials have protective coatings such as paints; therefore, maintenance and service of protective coatings have an important role in determining the ultimate impact of air pollutants. Paints may also be affected by ambient levels of air pollution.

Another related side effect of acid rain is visibility degradation. Fine particles in the atmosphere containing sulfate, nitrate, and other chemical constituents, which when deposited are associated with acid deposition, cause visibility degradation while in the air. These fine particles have been the major factor in the reduction of visibility in rural and urban areas in the eastern United States since the beginning of the century. In the U.S. West, visibility degradation is being reported in major urban areas and in national parks and wilderness areas.

Direct adverse effects of these pollutants on humans occur largely through the respiratory system. Sensitive populations with existing respiratory or cardiovascular problems, such as those with asthma, are especially susceptible. These effects have been mainly associated with the acid precursor gases and ozone. Studies of the effects of acidic aerosols, composed primarily of nitric acid, ammonium bisulfate, and sulfuric acid, are still relatively new. It has been found that on rare occasions acid levels approach 10 times the long-term mean levels for sulfuric acid. Although substantial uncertainty exists, the body of data raises concern that acidic aerosols alone, or in concert with other pollutants, may be contributing to health effects in exposed populations at current concentration levels.

Human health also can be affected indirectly by pollutants related to acid deposition. People who eat large amounts of fish from acidic lakes or streams may experience exposure to methylmercury in some regions of the country. Drinking water from acidic sources may contain significantly elevated levels of lead. It is unlikely, however, that exposure to humans by this pathway is important, except in isolated cases.

6 SOCIAL RESPONSE TO ACID RAIN

Historically, toxic effects have been observed in populations acutely exposed to high concentrations of air pollutants. As early as the Middle Ages in London prohibitions on coal burning were instituted in response to perceived health effects of mixtures of dust, soot, and fog. The industrial revolution brought the air pollution issue to the United States, where air quality management continued to be considered local in nature into the twentieth century.

The severity of air pollution impacts became very obvious during the London "killer fog" of 1952, when a mixture of particulates, sulfur dioxide, and acidic fog was associated with severe respiratory effects and approximately 4000 deaths. Emergence of air pollution as a public health issue in the 1950s, as a result of this and other deadly episodes, led to the development of federally funded research programs, culminating in the Clean Air Act and in the establishment of the Environmental Protection Agency (EPA) in 1970. These were major stimuli for the establishment of the U.S. National Ambient Air Quality Standards (NAAQS) that today restrict the atmospheric concentration of pollutants such as sulfur dioxide, nitrogen oxides, and ozone.

Other countries around the world have been developing institutional responses to the threat to human health of air pollution. Air pollution effects on the environment, however, had been slower to be recognized as a serious issue. Acid rain and its ecological effects were first documented in England at the end of the nineteenth century and became regional issues in northwestern Europe and the northeastern United States and eastern Canada only recently—in the late 1960s. During this period and into the 1970s, the mounting anecdotal evidence of the effects of acid rain on aquatic and terrestrial ecosystems launched acid rain as perhaps the first pollution threat to the environment to receive international attention.

The origins of the pressures to regulate acid rain in the United States were primarily twofold. First, Canada protested, lobbied, and publicized its contention that major environmental damage was occurring in its eastern provinces because of acid deposition, and that the major sources of acid precursors were in the United States Second, elected officials and citizens in the eastern and New England states echoed the same concerns, elevating the acid rain controversy to the level of a growing interregional conflict between receptor states and polluting states (Rhodes and Middleton, 1983).

The U.S. responses to these concerns took the form of federal research and eventually control programs. The first step was the Acid Precipitation Act of 1980, which created the National Acid Precipitation Assessment Program (NAPAP). During its first 10 years, the research and periodic assessments conducted by NAPAP improved the understanding of the scientific processes and effects of acid deposition. The monitoring and research conducted in the 1980s and the subsequent integrated assessment completed in 1990 provided the scientific knowledge base for Title IV, the Acid Deposition Control Program, of the 1990 Clean Air Act Amendments.

Title IV is designed to reduce the adverse effects of acid deposition through the reductions in annual emissions of sulfur dioxide (SO_2) and nitrogen oxides (NO_x), the precursors to acid rain. Recognizing that the principal sources of acid rain precursors in the atmosphere are emissions from the combustion of fossil fuels, control measures were initiated to reduce emissions from electric utilities. However, rather than the traditional command-and-control approach to regulation, alternative methods of compliance were allowed. These methods included technological adaptation (e.g., scrubbers, higher-efficiency boilers), fuel-switching, and an innovative SO_2 emissions allowance trading program. This represented the first national effort to use market-based incentives to achieve environmental goals.

Due to the innovative nature of using market-based incentives for environmental regulation, Congress set up a mechanism for checking how well trading was working. As part of this activity, Congress asked NAPAP to assess the costs and economic impacts of the acid deposition control program as well as the effectiveness and benefits associated with the various human health and welfare effects. The effects included visibility, materials, and cultural resources damages and ecosystem effects. NAPAP was also asked to consider the deposition levels needed to protect sensitive ecosystems. The results of the assessment of Title IV are to be reported to Congress quadrennially, beginning with the 1996 Report to Congress (NAPAP, 1998).

7 CURRENT CONDITIONS

As of the completion of the first report to Congress, several observations have been made regarding the success of Title IV. It appears that the market-based approach has lowered compliance costs. Costs are lower than expected, probably due to a number of factors such as railroad deregulation, technological innovation, and lower operating costs for scrubbers. In addition, all affected utilities have fulfilled the compliance requirements of Title IV. In the first annual reconciliation of allowances and emissions, SO_2 allowances matched or exceeded SO_2 emissions. NO_x reductions have not been as dramatic. This is expected since mandates on NO_x reductions are not in place yet. However, NO_x emissions from all sources in 1995 were 1.5 million tons below 1980 levels. Utilities were responsible for 53% of that reduction.

Statistically significant reductions in acidity and sulfate in precipitation were reported at monitoring sites in the Midwest, mid-Atlantic, and northeastern United States. There is no real evidence of statistically significant decreases in nitrate concentration. Changes in aquatic ecosystems have not yet been detected. However, over the last 15 years, lakes and streams throughout many areas of the United States have experienced decreases in sulfate concentration in response to decreased emissions. While there is some evidence of recovery from acidification in New England, Adirondack lakes continue to acidify, suggesting that additional reductions may be needed in these areas (NAPAP, 1998).

Sulfur and nitrogen deposition has caused adverse impacts on certain sensitive forest ecosystems in the United States, with high-elevation spruce in the eastern United States being most sensitive. Other sensitive forests are apparently not experiencing the same effects in mortality and growth, at least for now, but some of the same processes appear to be slowly occurring.

The leaching of soil nutrients by continued acidic deposition is a gradual process that will eventually impact forest nutrition and growth in many areas. The recent reductions in sulfur should result in some small immediate improvements in sensitive forests, but large improvements will be slow to occur.

Reduced emissions of sulfur oxides are expected to reduce sulfate concentration and its contribution to haze. It is difficult to assess the extent to which recent reductions have contributed to changes in visibility over the past few years since meteorological and other factors determine the overall changes in visibility. Information is needed over the long term.

The recent reductions in SO_x and NO_x emissions are expected to reduce fine particulates and, as a result, lead to improved human health. It is suggested that reduced emissions will lead to a reduction in premature mortality from cardiovascular and respiratory causes and to a dramatic reduction in the number of asthma symptom days.

One difficulty in determining effects at this time is that many impacts have a response times that are longer than the few years since the passage of Title IV. Visibility and acute health effects can be detected on the order of hours to days. Episodic aquatic effects and soil and plant processes in the forest ecosystem respond on the order of days and weeks to months. Chronic human health, chronic aquatic effects, and forest health, on the other hand, indicate response times on the order of years to decades. Effects on forest solid nutrient reserves and effects on materials begin to show up on the order of decades to centuries. These latter effects are more on the order of climate change impacts response times.

The difference in response times, of course, makes an evaluation of actions taken in the early 1990s difficult to quantify. Improvements in health and visibility can serve as indicators of positive change. However, as already noted, even changes in visibility cannot be directly attributed to sulfate reductions alone, since other factors such as meteorological variability play a role in determining visibility changes especially in the humid part of the eastern United States.

8 KEEPING A BROAD BASIS OF ASSESSMENT AND ACTION

To review, acid deposition is an end product of a complex series of interactions among atmospheric chemical species emitted by both natural and human sources. For policy assessment purposes, the most important groups of chemical species are compounds containing sulfur and nitrogen compounds that are emitted from factories, power plants, and automobiles based on fossil-fuel combustion. In addition, volatile organic carbon compounds and fine particles play a role in modulating chemical processes and acidity. Some key compounds remain unchanged in the atmosphere and some are neutralized, but others are oxidized into more acidic forms through a complicated series of chemical, meteorological, physical, and biological interactions.

Decisions about the control of acid deposition must deal with the environmental impacts of estimated future emission levels as well as present levels. Projections depend on many complex and interacting socioeconomic factors. The predictability of how rapidly and to what extent fossil fuels will be replaced by clearer and safer fuels (society may change transportation and other energy use habits) and the interrelationships among countries of the world becoming driving influences for these changes is highly uncertain. Given the demonstrated value of examining multiple causes and effects together, it will continue to be important to keep the base of assessment broad in spite of the uncertainties. The elements of such assessments are illustrated in Figure 1.

On issues related to acid rain, other policy discussions going on throughout North America also illustrate the growing awareness of the interconnections, as well as the need to capitalize on the relationship in developing strategies for the future. For example, EPA, through the Federal Advisory Committee Act (FACA), is leading the development of combined ozone, particulate matter (aerosols), and regional haze implementation program rules and guidance. Other activities at the regional level,

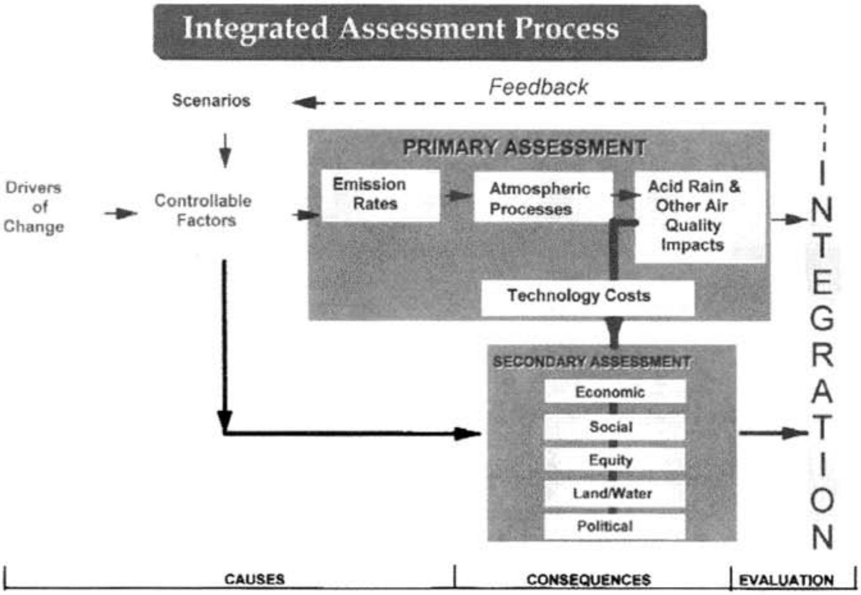


Figure 1 Important elements to consider in an assessment of acid rain in the context of other important environmental, energy, and economic concerns. See ftp site for color image.

such as the Western Governors' Association (WGA) Air Quality Initiative, the Ozone Transport and Analysis Group (OTAG), the Southern Appalachian Mountain Initiative (SAMI), the Southern Oxidant Study (SOS), and the North American Research Strategy for Tropospheric Ozone (NARSTO), are addressing various science and policy issues associated with ozone, particulate matter, and regional haze.

On a continental scale, the Commission on Environmental Cooperation (CEC) is working on North American strategies for addressing transboundary concerns, which include ozone, particulate matter, regional haze, and acid rain along with other hazardous pollutants. Finally, on a much broader scale, the U.S. global climate change program is addressing regional climate assessment for areas throughout the United States. The regional air quality concerns addressed by FACA and the climate concerns to be considered in these discussions are closely linked through the development of aerosols that can influence climate on regional scales as well as produce other problems. On a larger policy implementation level, the two are linked through the development of energy strategies aimed at reducing greenhouse gas emissions and the emission of other more traditionally harmful air pollutants.

All of these programs and approaches share the same fundamental concerns. The role of natural or background processes and the role of chemical interactions in determining the levels of impacts in different regions continue to be fundamental overarching scientific questions. The implementation issues of emissions trading versus pollution prevention versus technological controls are also part of each aspect of the various debates. Assessments of trade-offs for any decisions made with respect to any of these issues must consider the less quantifiable, and sometimes more uncertain, impacts associated with heath, social impacts and values, equity, and related environmental concerns about water and soil quality as well as air. Given these strong interconnections, it is important to make the best use of research and policy-making resources across organizations addressing acid rain and related issues, where energy and the environment are key factors.

REFERENCES

Graedel, T. E., T. S. Bates, A. F. Bouwman, D. Cunnold, J. Dignon, I. Fung, D.J. Jacob, B.K. Lamb, J.A. Logan, G. Marland, P. Middleton, J.M. Pacyna, M. Placet, and C. Veldt, A compilation of inventories of emissions to the atmosphere, *Global Biogeochem. Cycles*, *7*, 1–26, 1993.

Graedel, T. E., and P. J. Crutzen, The changing atmosphere, *Sci. Am, 261* (Sept.), *58*, 1989.

Middleton, P. Sources of air pollutants, in *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995.

Rhodes, S. L., and P. Middleton, The complex challenge of controlling acid rain, *Environment*, *25*, 6–9, 31–38, 1983.

U.S. National Acid Precipitation Assessment Program (NAPAP), *Acidic Deposition: State of Science and Technology, Summary Report of the U.S. National Acid Precipitation Assessment Program*. Washington, DC. 1991.

U.S. National Acid Precipitation Assessment Program (NAPAP), *NAPAP 1996 Report to Congress*, 1998, http://www.nnic.noaa.gov/CENR/NAPAP/NAPAP.96.htm

CHAPTER 50

IMPACTS OF CLIMATE CHANGE

STEWART J. COHEN

1 SCIENCE-POLICY CHALLENGE

The Intergovernmental Panel on Climate Change (IPCC) has concluded that if atmospheric concentrations of greenhouse gases continue to increase, largely as a result of fossil fuel combustion, agricultural practices, and deforestation, average temperatures will increase at a rate much faster than our world has experienced since the last Ice Age. This in itself represents a vision of the future that is substantially different from the past. It challenges long held notions of climate stability, slow rates of climate change (in human terms), and the dominance of natural forces over societal forces in influencing global climate.

Even as new evidence is presented about how civilizations over the last 5000 to 7000 years have been affected by changes in climate (Lamb, 1982), it is recognized that those historic shifts in annual temperatures were only around ± 1 to 2°C from current global averages. The medieval warm epoch of the tenth to thirteenth centuries was around 0.5 to 1°C above the current global average, while the Little Ice Age of the sixteenth to nineteenth centuries was around 1°C cooler. Although ancient and medieval civilizations may have been less technologically developed than twentieth-century society, modern individuals and nations still have to plan for and adapt to climate. Impacts associated with recent extreme events and El Niño episodes illustrate that despite technological advances, societies in developed and developing countries are still vulnerable to short-term variations in climate (Burton, 1997). There has been a clear upward trend in weather-related costs to insurance companies since the 1960s (Munich Re, 2000), leading to substantial losses by major reinsurance companies such as Lloyd's (IPCC, 1996a). This does not include uninsured losses that may be equal in magnitude to insured losses.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

A warming of up to 5.8°C during the twenty-first century, a rate of up to 0.5° per decade, would be unprecedented in human history (IPCC, 2001a). Such a change in mean temperatures, with accompanying changes in seasons and probabilities of extreme events, would have direct impacts on land, water, wildlife, and a myriad of indirect impacts on communities, businesses, and governments. If greenhouse gas emissions are not reduced, societies would be faced with the prospect of having to adapt to a new climate that current climate models are not yet able to precisely describe, especially at the regional scale. Adaptation in the twenty-first-century context would be a very different challenge than the one faced by our ancestors, and the costs of adaptation measures are not known.

Identifying societal impacts of a scenario of rapid warming is a complex interdisciplinary research activity that goes beyond assessments of changes in atmospheric processes alone. These changes would be superimposed on changing populations, landscapes, institutions, technologies, and perceptions of resources and environment.

What are the broad dimensions of the societal aspects of global climate change? An outline is presented in Figure 1. This chain of causality also represents a target for a research activity known as *integrated assessment* (IA). Several research groups have attempted to incorporate part or all of this chain into a series of linked models, which collectively have become known as *integrated assessment models* (IAM). The

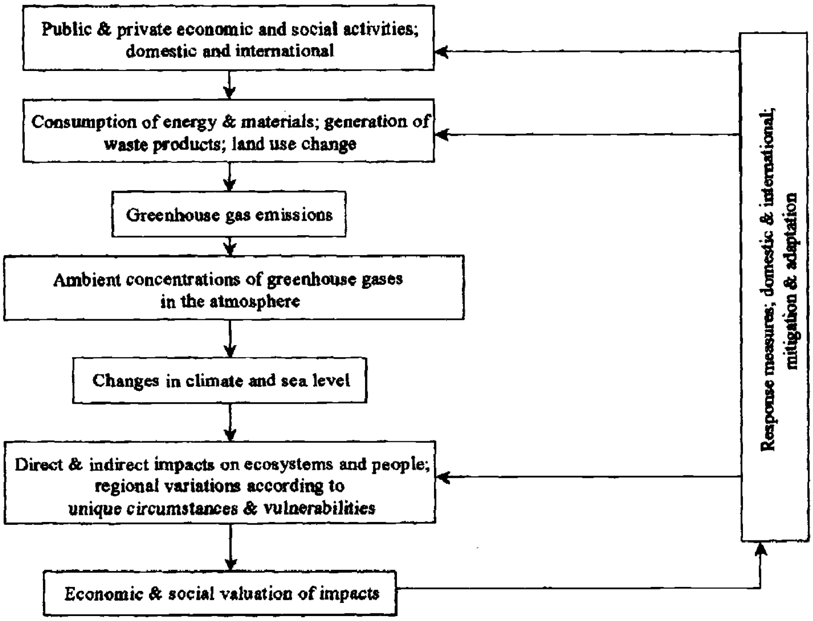

Public & private economic and social activities; domestic and international
Consumption of energy & materials; generation of waste products; land use change
Greenhouse gas emissions
Ambient concentrations of greenhouse gases in the atmosphere
Changes in climate and sea level
Direct & indirect impacts on ecosystems and people; regional variations according to unique circumstances & vulnerabilities
Economic & social valuation of impacts
Response measures; domestic & international; mitigation & adaptation

Figure 1 Science-policy dimensions of climate change. Mitigation responses focus on emission reduction or sink enhancement. Adaptation responses focus on reduction of vulnerabilities to climatic events or taking advantage of new climate-related opportunities.

IPCC lists 23 IAMs in its Second Assessment Report (IPCC, 1996b). Other IA techniques include economic models, decision support models, expert judgment exercises, policy exercises, and the use of themes or places as an interdisciplinary platform for collection and analysis of information (Cohen, 1997; Schneider, 1999; Kasemir et al., 1999).

2 NEED FOR INTEGRATED ASSESSMENT OF GLOBAL CLIMATE CHANGE

The outline sketched in Figure 1 suggests that climate change should not be treated in isolation from other environmental concerns, particularly the suite of challenges that are part of global environmental changes resulting from unsustainable development. Sustainability requires living within the carrying capacity of Earth, and this has become a highly political and emotional issue in many places as environment and development come into conflict (e.g., toxic wastes, overfishing, deforestation, desertification). Climate change is connected to many of these concerns, yet climate change has often been treated as a narrowly defined question of atmospheric change and greenhouse gas emissions. Climate change is not only about the meteorology and chemistry of the atmosphere. It is about the underlying human (i.e., economic, political, social) sources of this stress, and the potential victims. Understanding these other components requires a more holistic view of the climate change issue, well beyond consideration of atmospheric science alone.

The advantage of trying an integrated approach is that it represents an explicit attempt to incorporate both physical and human dimensions into research on climate change impacts and responses. For example, a study may indicate a projected change in the potential of a region or country to grow corn and wheat. However, just because there is a change in potential does not mean that farmers, other landowners, communities, businesses, governments, and other stakeholders would agree to a land-use change in response to this change in land capability. Another example is the current debate about measures to reduce greenhouse gas emissions, and the effectiveness of various alternatives such as a "carbon tax," technology transfer from developed to developing countries (through international or binational agreements), or an emissions trading system.

3 METHODOLOGY FOR IMPACT ASSESSMENT OF CLIMATE CHANGE SCENARIOS

Unlike assessments of current and historical events, impact assessment of projected climate change is based on *scenarios*, or plausible pictures of the future. This kind of research can include analysis of past observed events (e.g., the Dust Bowl in the United States during the 1930s), which could serve as societal analogs of a future warm climate (Glantz, 1988), or as climate change analogs superimposed on twenty-first-century society (Rosenberg, 1993), but there are also many studies at regional and global scales that use climate change scenarios as part of a forward-looking exercise to estimate future impacts.

Case studies of future climate scenarios have often been based on the outputs of climate models (IPCC, 1996a,b; Carter et al., 2000). These models (general circulation models or GCMs) produce estimates of climatic variables on a regular network of grid points for a base case (i.e., current climate), and as an "equilibrium" response to a doubling of carbon dioxide concentrations, or as a "transient" response over 50 years or more of incremental increases in carbon dioxide. Most assessments have taken the difference between the equilibrium or transient simulations and the base case simulation and combined this simulated "change" with baseline climate information from actual observations to produce a scenario (Carter et al., 1994, 2000).

These scenarios of changes in temperature, precipitation, and other elements are used as inputs to other analytical tools that would convert these climate changes to first-order changes in landscapes, ecosystems, renewable resources, and disease rates. Examples include (a) hydrologic models for streamflow and lake levels, (b) crop models for grain yields, (c) fire indicators for forest fire potential, and (d) pest indicators for seasonal ranges of insects. Outputs of first-order impact assessments have been applied to economic models (e.g., timber yields, food production, hydroelectric generation) to estimate impacts in monetary terms (IPCC, 1996a, 2001b).

There have also been attempts to combine these individual estimates into regional or national impact assessments. Scaling up from sectoral to national assessments adds complexity and uncertainty to an already difficult assignment. Economies and societies are composed of many stakeholders whose actions are not easily amenable to modeling. Governments, industries, and individuals may choose various response strategies depending on their knowledge and perceptions of the climate change issue and other forces of concern (e.g., population growth, changes in global trading patterns, and technology).

At the same time, concerns raised by atmospheric scientists about greenhouse gas emissions has led to international negotiations to establish a global strategy to reduce emissions. The main policy instrument, the United Nations Framework Convention on Climate Change (UNFCCC) has been ratified by more than 150 countries, and the emission targets for six greenhouse gases have been tentatively established for more than 30 industrialized countries by a recent agreement known as the Kyoto Protocol (negotiated in December, 1997, but not yet ratified). Critics of this agreement have argued that these targets will cause severe economic losses to most of these countries because, in their view, economic growth is directly linked to growth in energy consumption and hence increased greenhouse gas emissions. Two other concerns are raised as well: (a) global warming may not happen as predicted since there are uncertainties in climate models and their projections may be wrong, and (b) since societies in the past flourished during warm periods (e.g., medieval warm epoch), global warming (if it occurs) would actually be a good thing. Although the IPCC (1995, 1996a,b, 1998, 2001a,b) has published estimates of projected climate change, impacts of climate change scenarios, and impacts of greenhouse gas emission reductions, the rapid policy response has created a substantial new research challenge—determining the impacts of various policy options and comparing these with the costs of doing nothing about emissions. Given the magnitude of the climate change problem and proposed responses to this, the demand for answers will be high.

4 SUMMARY OF CASE STUDIES

Estimates of climate change impacts are summarized for several key sectors in Table 1. In fisheries and health, higher levels of impacts are estimated for developing countries, reflecting their vulnerabilities to climate change. Impacts in agriculture would result from damage due to heat stress, decreased soil moisture, increased incidence of pests and disease, and changes in plant growth cycles; but this could be offset in some circumstances by longer growing seasons and CO_2 fertilization. Some developing countries, however, could experience significant increases in population at risk from hunger. Coastal zone costs are high in many regions, reflecting the growth in built structures in areas vulnerable to sea-level rise and extreme events, as well as land loss itself (e.g., coastal wetlands). Estimates are also available for changes in water supply, wetlands, electricity demand, and some other sectors (IPCC, 1996a,b).

In all cases, scenario estimates are dependent on a variety of assumptions about regional changes in climate (downscaling from GCMs), indirect effects of climate change (e.g., CO_2 fertilization effects), technological change, population growth, changes in infrastructure (e.g., health services in developing countries), and responses of stakeholders to other issues besides climate change (e.g., changing international markets). Table 1 should be considered as a first attempt at determining

TABLE 1 Range of Sectoral Impacts for Different World Regions (2.5°C warming scenario)[a]

Region	Agriculture (% loss in GDP)	Forestry (area lost, km^2)	Coastal Zone (annual protection costs 10^6 U.S.$)	Health (number of deaths, 1000 s)	Fisheries (reduced catch, 1000 ton)
European Union	0.21	52	133	8.8	558
United States	0.16	282	176	6.6	452
FSU	0.24	908	51	7.7	814
China	2.10	121	24	29.4	464
Non-OECD	0.28	334	514	114.8	4,326
OECD	0.17	901	493	22.9	2,503
World	0.23	1,235	1,007	137.7	6,829

GDP, gross domestic product; FSU, former Soviet Union; OECD, Organization of Economic Cooperation and Development (developed countries); non-OECD, developing countries.

Source: Adapted from IPCC (1996b). Table 6.5

TABLE 2 Overall Annual Economic Impacts in Different World Regions for a 2.5°C Warming Scenario (% of current GDP)[a]

Region	IPCC, 1996b	IPCC, 2001
Developed countries	−2.8 to −1.3	−2.8 to 0.3
FSU	−0.7 to 0.3	−0.7 to 11.1
Developing countries	−8.7 to −4.1	−4.9 to 1.8
World	−1.9 to −1.4	−1.5 to 0.1

[a] See Table 1.

Source: Adapted from IPCC, 1996b (Table 6.6) and IPCC, 2001b (Table 19-4).

impacts, and changes in such estimates should be expected as new information becomes available.

Regional impact costs are summarized in Table 2. The range of uncertainty cannot be gauged from the existing literature, nor can the range of estimates provide a confidence interval. Some costs are hidden because of aggregation of communities and nations into large regions and because of lack of information on potential impacts of climate change scenarios on construction, insurance, nontropical extreme events (e.g., midlatitude river floods), transportation, and political institutions.

This work is still in its infancy, and economists have to make assumptions about markets, trading patterns, technological change, adaptation (direct, indirect), and the availability of information. Uncertainties associated with converting impacts into monetary units are due to many factors. One of the most controversial is the cost of health impacts. Is a premature death due to climate change (e.g., heat stress, tropical disease, etc.) worth the same monetary value if it were to occur in a developed or developing country? Since average incomes differ, initial estimates have used a higher "value of a statistical life" for a premature death in a developed country than in a developing country (IPCC, 1996b, Chapter 6). More recent assessments focus on global and regional development trends and their effects on vulnerability to climate change (IPCC, 2001b). These trends may increase adaptive capacity in some circumstances (e.g., community health), but may decrease it in others (e.g., protection of endangered species).

The dilemma of placing a value on life and death is an illustration of the political and social dimensions of the potential impacts of global climate change. Others include (a) the presence of different stakeholders with different visions and goals, (b) issues related to cultural preservation (especially in developing countries and the Arctic), (c) the influence of trade globalization on management of climate-sensitive resources (e.g., fish, water resources), (d) the market value of ecosystems (e.g., wetlands, rain forests, alpine tundra), and (e) intergenerational equity and the choice of discount rates (i.e., a percent change in monetary value due to depreciation, inflation, etc.) used in economic valuation of climate change damages and actions. All of these can influence development choices, including adaptation choices. Global climate change may have started as a theoretical problem of atmospheric science, but the transmission of information to the public has taken this issue outside of the laboratory and into the real world.

5 LESSONS AND A LOOK AHEAD

Climate impact research is a relatively young endeavor. The term *climate impact assessment* was coined only in the 1970s (Munn, 1979; Kates et al., 1985), and case studies of climate change scenarios have been undertaken for less than 20 years. Advances have been made in incorporating various natural and social science disciplines into the effort, and some important lessons have been learned in the process:

1. There is more than one scenario of future changes for any region, regardless of any scenarios of climate change.
2. There is more than one stakeholder, and one cannot assume that while temperatures change, stakeholders will continue historic patterns of activities. Historic and future decisions result from trade-offs and consensus reached by a broad array of decision makers with different visions.
3. Although there are some preferred options for assessing impacts on *sectors*, there is no single best method available to provide impact assessments for *places*. Parallel sectoral assessments have been the approach of choice in most countries, but some integration exercise needs to become part of this process. Integrated assessment models have become a highly visible option, but these tend to focus more on mitigation and are relatively weak on the impacts and adaptation dimensions, so alternative methods will continue to be important contributors.

Societal aspects of global climate change represent a significant interdisciplinary research challenge, in which atmospheric science and scientists will continue to play an important role. This collaboration will be beneficial for advancing the science as well as for providing better information for stakeholders as they grapple with the human dimensions of this issue.

As consumers of climate information, both from observations and models, researchers on impacts of and responses to climate change uncertainties represent a different type of client than those who work in the atmospheric sciences. This group needs value-added information that can be used as input to other analytical tools, which may or may not have been designed with climate as an explicit input element. Indeed, these tools (e.g., crop yield models, water management models) may be calibrated only to current climate conditions, and their response to climate change scenarios outside their calibration range represents one of many uncertainties in this process. Some of these tools require considerable amounts of data, often at spatial scales too fine to be visible in current GCMs. Impacts researchers are following the progress of downscaling activities with considerable interest (e.g., regional climate models, statistical techniques), and perhaps this can provide incentive to atmospheric scientists to continue research and development in this area. In the meantime, however, the urgent need for impacts information, as well as for testing and developing methodologies for impact and response assessments, means that currently available methods for scenario construction will continue to be used.

There will also be continued demand for assessments of historic events. Some of these events [e.g., effects of warm years, droughts (El Niño–Southern Oscillation (ENSO)] may serve as possible analogs of future climate change, but important questions arise. How can impacts and costs be attributed? Were these due to the climatic event (at what scale), or to changing vulnerabilities, or both? Did regional or global forces cause the climatic event, and was it consistent with modeled scenarios of future climate changes? The recent increase in insurance losses (Munich Re, 2000) is an example of an observed series of events that would benefit from such an analysis.

Finally, the Kyoto Protocol presents a challenge and an opportunity for new research into the potential impacts of measures to reduce emissions and improve adaptive capabilities. The decision to ratify or not ratify this agreement will be taken on the basis of technical information balanced against preset stakeholder interests. Atmospheric science, in collaboration with researchers from many other disciplines, will play an important role in determining the benefits and costs of various response scenarios.

Past and present impacts have occurred over landscapes and populations that are changing for many reasons. Such changes affect vulnerabilities and costs, as well as responses (e.g., changing land use, insurance programs). Future impacts will be influenced by global economic and institutional changes (e.g., globalization of trade) as well as policy initiatives at various scales. Stakeholders' responses will be determined by attitudes and beliefs about the importance of climate change in the context of other challenges. Atmospheric scientists were the first to call attention to global climate change. Now that this has become an international policy concern, there will be greater demands to make scientific views known not only in the traditional refereed literature, but in broader public forums as well.

REFERENCES

Burton, I., Vulnerability and adaptive response in the context of climate and climate change, *Climatic Change*, *36*, 185–196, 1997.

Carter, T. R., M. Hulme, J. E. Crossley, S. Malyshev, M. G. New, M. E. Schlesinger, and H. Tuomenvirta, *Climate Change in the 21st Century—Interim Characterizations based on the New IPCC Emissions Scenarios*, Finnish Environment Institute, Helsinki, 2000.

Carter, T. R., M. L. Parry, H. Harasawa, and S. Nishioka, *IPCC Technical Guidelines for Assessing Climate Change Impacts and Adaptations*, University College, London and Center for Global Environmental Research, Tsukuba, 1994.

Cohen, S. J., Scientist-stakeholder collaboration in integrated assessment of climate change: Lessons from a case study of Northwest Canada, *Environ. Model. Assess.*, *2*(4), 281–293, 1997.

Glantz, M. H. (Ed.), *Societal Responses to Regional Climate Change: Forecasting by Analogy*, Westview Boulder, CO, 1988.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group I to the second assessment report of the Intergovernmental Panel on Climate Change, in J.J. Houghton, L.G. Meiro Filho, B.A. Callandar, N. Harris, A. Kattenberg and K. Maskell (Eds.), *Climate Change 1995—The Science of Climate Change*, Cambridge University Press, Cambridge, 1995.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group II to the second assessment report of the Intergovernmental Panel on Climate Change, in R.T. Watson, M. C. Zinyowera, and R. H. Moss (Eds.), *Climate Change 1995—Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analysis*, Cambridge University Press, Cambridge, 1996a.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group III to the second assessment report of the Intergovernmental Panel on Climate Change, in J. P. Bruce, H. Lee, and E. F. Haites, (Eds.), *Climate Change 1995—Economic and Social Dimensions of Climate Change*, Cambridge University Press, Cambridge. 1996b.

Intergovernmental Panel on Climate Change (IPCC), A special report of working group II, in R. T. Watson, M. C. Zinyowera, and R. H. Moss, (Eds.), *The Regional Impacts of Climate Change: An Assessment of Vulnerability*, Cambridge University Press, Cambridge, 1998.

Intergovernmental Panel on Climate Change (IPCC), *Summary for Policymakers of the IPCC Working Group I Third Assessment Report*, approved in Shanghai, January 2001, available on-line, http://www.usgcrp.gov/ipcc/wglspm.pdf, 2001a.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group II to the Third assessment report of the Intergovernmental Panel on Climate Change, in J. McCarthy, O. Canziani, N. Leary, D. Dokken, and K. White (Eds.), *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, Cambridge University Press, Cambridge, 2001b.

Kasemir, B., M. B. A. van Asselt, G. Dürrenberger, and C. C. Jaeger, Integrated assessment of sustainable development: Multiple perspectives in interaction, *Int. J. Environ. Pollut.*, *11*(4), 407–425, 1999.

Kates, R. W., J. H. Ausubel, and M. Berberian (Eds.), *Climate Impact Assessment*, SCOPE 27, Wiley, New York, 1985.

Lamb, H. H., *Climate, History and the Modern World*, Methuen, London, 1982.

Munich Re, *Topics—Annual Review of Natural Disasters 1999*, Report 2946-M-e, Munich Reinsurance Group, Munich, Germany, 2000.

Munn, R. E., The framework for a climate impact assessment, in *Carbon Dioxide Issues and Impacts: Proceedings of the Workshop on Energy Carbon Dioxide Issues and Impacts*, Climate Planning Board (F.K. Hare, Chair), Canadian Climate Program (Eds.), 19–22, Atmospheric Environment Service, Downsview, Canada, 1979.

Rosenberg, N. J. (Ed.), Towards an integrated impact assessment of climate change: The MINK study, *Climatic Change*, *24*(1–2), 1–173, 1993.

Schneider, S. H. (Ed.), Topics related to integrated assessment, *Climatic Change*, *41*(3–4), 265–546, 1999.

CHAPTER 51

IMPACTS OF STRATOSPHERIC OZONE DEPLETION

MICHELE M. BETSILL

1 INTRODUCTION

In 1928, Charles Kettering and Thomas Midgley, Jr., scientists with General Motor's Research Corporation in Dayton, Ohio, invented chlorofluorocarbons (CFCs) as a safe alternative to toxic and flammable refrigerants. In 1974, Mario Molina and F. Sherwood Rowland published a paper in *Nature* that linked the use of CFCs to destruction of Earth's stratospheric ozone layer. Today, the international community has made substantial progress toward phasing out the production and consumption of CFCs and in setting up the mechanisms that should help to restore the damaged ozone layer.

Ozone is a molecule consisting of three oxygen atoms (O_3). Approximately 90% of Earth's ozone is found in the stratosphere, the region of the atmosphere that is between 10 and 15 km above Earth's surface. Stratospheric ozone helps regulate Earth's atmospheric temperature structure by absorbing damaging ultraviolet sunlight (UV-B).

Stratospheric ozone depletion is primarily caused by human-made chemicals containing various combinations of chlorine, fluorine, bromine, carbon, and hydrogen. Collectively, these compounds are called halocarbons. They can be divided into compounds containing carbon, chlorine, and fluorine (CFCs) and those containing carbon, bromine, and fluorine (halons). CFCs are used in refrigeration, air-conditioning systems, as foam blowing agents, for cleaning electronic components, and as solvents. Halons are primarily used in fire extinguishers. These compounds break down when they enter the atmosphere; then chlorine and bromine atoms react with ozone to catalyze its destruction (WMO, 1995).

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

The possibility that human-made substances may cause stratospheric ozone depletion was first raised in the late 1960s and early 1970s as part of debates about the development of a large fleet of high-flying supersonic transport (SST) aircraft. These aircraft were to be designed to fly at the speed of sound at an altitude of 45,000 ft (well into the stratosphere). Opponents of the SST program raised concern about the sonic booms the aircraft would cause. Several scientists also suggested that water vapor and nitrous oxide emitted by SSTs would catalyze a chemical process leading to the breakdown of stratospheric ozone molecules. The United States, fueled by public outcry over the sonic boom concern and with full realization of the adverse economic aspects, canceled its SST program before the link between nitrous oxide and ozone depletion could be further investigated (Cagin and Dray, 1993; NAS, 1975). It was Molina and Rowland's 1974 article that prompted the scientific community to explore the role of human-made substances in depletion of the stratospheric ozone layer. In 1995, Rowland and Molina were awarded the Nobel Prize in Chemistry for their work in explaining how the stratospheric ozone layer is destroyed. They shared the award with Paul Crutzen who was recognized for his research linking nitrogen oxides with ozone depletion. This marked the first time the Nobel Prize was awarded for environmental work (Lipkin, 1995).

2 IMPACTS OF STRATOSPHERIC OZONE DEPLETION

Stratospheric ozone depletion was recognized as an environmental problem in need of international attention because it impacts both humans and the natural environment. When stratospheric ozone levels decrease, the amount of UV-B reaching Earth's surface increases (WMO, 1995). The changes in UV-B radiation are highest at high and midlatitudes in both hemispheres while the increases are fairly small in the tropics (UNEP, 1994). Increased levels of UV-B affect human health, the productivity of plant and animal species, as well as the composition of ecosystems.

Impacts on Human Health

Ultraviolet exposure does have some benefits for humans. For example, it initiates the production of vitamin D_3, which is believed to inhibit the growth of tumor cells (UNEP, 1996). However, the balance of evidence indicates that the effects of stratospheric ozone depletion on human health are negative. The major risks include increased incidence of eye diseases, skin cancer, and infectious diseases. When UV-B levels increase, two main organ systems are exposed: the eyes and the skin. The impacts of ozone depletion are mediated through these two systems (Longstreth et al., 1995; UNEP, 1998).

Evidence suggests that increased UV-B radiation exposure may be associated with an increase in the incidence of cataracts, a clouding of the lens of the eye (Longstreth et al., 1995; UNEP, 1998). One review of research on this problem reported that a 1% increase in stratospheric ozone depletion would result in a 0.6 to 0.8% increase in the incidence of cataracts (UNEP, 1994; see also UNEP, 1998).

The most widely known impact of increased UV-B radiation on human health is skin cancer. UV-B radiation damages deoxyribonucleic acid (DNA), which may cause gene mutations and the formation of cancer cells. Some studies estimate that a sustained 10% decrease in average stratospheric ozone concentrations would result in 250,000 new cases of nonmelanoma skin cancer. This is in addition to the 1.2 million cases already reported each year (Longstreth et al., 1995; UNEP, 1996). Many animal species, such as cows, goats, sheep, cats, and dogs, are also at increased risk of developing skin cancer as a result of increased exposure to UV-B radiation (UNEP, 1998).

In an assessment of the effect of the Montreal Protocol and its amendments in protecting the ozone layer, Slaper and his colleagues (1996) concluded these efforts will substantially decrease the growth rate of the incidence of skin cancer over the next century. They found that under a scenario where there were no limits on the production and consumption of ozone-depleting substances, there would be a quadrupling in the incidence of skin cancer by the year 2100. Under the provisions of the Montreal Protocol (a 50% reduction in the production of CFCs by 1999), a doubling in the incidence of skin cancer could be expected in that same period. In contrast, they found the Copenhagen Amendments scenario (a complete phase-out in the production of 21 ozone-depleting substances by January 1, 1996) would result in a 10% increase in skin cancer incidence, peaking in the year 2060. This study lends support to the importance of international efforts to combat stratospheric ozone depletion.

Researchers believe that skin exposure to increased levels of UV-B radiation is also linked to modifications in the human immune system. As a result, the ability of the immune system to respond to certain infectious diseases, such as tuberculosis, leprosy, and Lyme disease, is impaired (UNEP, 1998). Longstreth and her colleagues (1995) predict that higher levels of UV-B will result in increased *severity* and *duration* of diseases such as lupus rather than an increase in their *incidence.*

Impacts on Aquatic Systems

The balance of evidence indicates that increased UV-B radiation can have harmful effects on many species of aquatic organisms and the aquatic systems in which they live (SCOPE, 1993; UNEP, 1998). For example, studies in the Antarctic have linked increased UV-B levels to reduced phytoplankton productivity. Phytoplankton are the basis for the oceanic food chain. UV-B radiation affects the DNA, photosynthesis, enzyme activity, and nitrogen incorporation of phytoplankton. Reduced phytoplankton productivity will likely lead to reduced productivity further up the food chain. It has been estimated that a 16% reduction in stratospheric ozone could lead to a 5% loss of phytoplankton causing a loss of 7 million tons of fish worldwide per year (Häder et al., 1995; UNEP, 1994, 1996). Figure 1 illustrates the effects of UV-B radiation on phytoplankton.

Researchers have also found that enhanced UV-B radiation disrupts the early development of several species of fish, shrimp, and crabs, ultimately affecting their motility (Häder et al., 1995). In damaging aquatic organisms, stratospheric

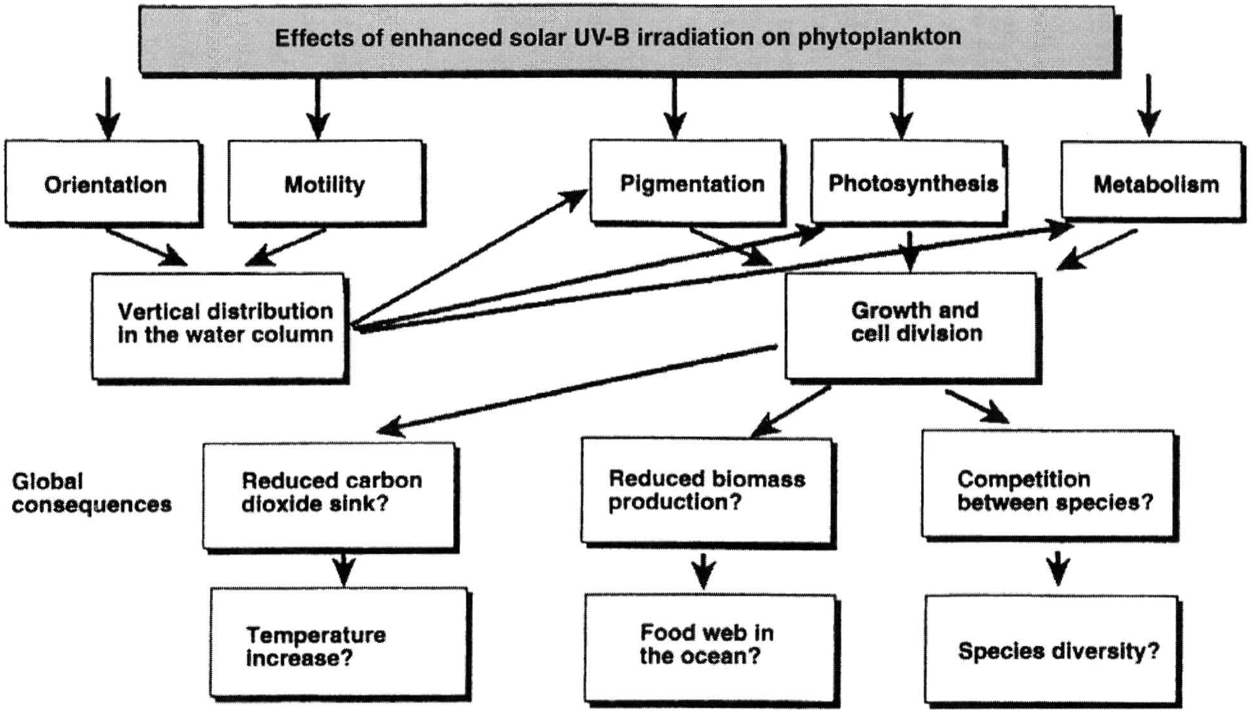

Effects of enhanced solar UV-B irradiation on phytoplankton
Orientation
Motility
Pigmentation
Photosynthesis
Metabolism
Vertical distribution in the water column
Growth and cell division
Global consequences
Reduced carbon dioxide sink?
Reduced biomass production?
Competition between species?
Temperature increase?
Food web in the ocean?
Species diversity?

Figure 1 Effects of UV-B radiation on phytoplankton (from Häder et al., 1995, p. 178).

ozone depletion has serious implications for the world food supply. Globally, 30% of the animal protein consumed by humans comes from the oceans. The percentage is much higher in developing countries (UNEP, 1998). These impacts are particularly worrisome in light of the growing world population.

Impacts on Terrestrial Plants and Ecosystems

Scientific understanding of the impact of enhanced UV-B on terrestrial plants and ecosystems is incomplete. The majority of studies have been conducted in growth chambers and greenhouses under controlled conditions, conditions that are often quite different from those experienced in the field. Thus, researchers contend it is necessary to use caution in making generalizations about the impacts of enhanced UV-B on terrestrial plants. The results of existing studies need to be verified under field conditions (Caldwell et al., 1995).

Keeping the limitations of existing research in mind, it is still possible to make some statements about the effect of enhanced UV-B on terrestrial plants. It appears that increased UV-B radiation may have both direct and indirect effects on plants. Some plant species exhibit a reduction in leaf area and/or stem growth when exposed to higher levels of UV-B. In addition, UV-B may also inhibit photosynthesis, damage plant DNA, and alter the time of flowering as well as the number of flowers in some species. The latter has implication for the availability of pollinators and thus the reproductive capacity of plants (Caldwell et al., 1995; UNEP, 1998). The effects of UV-B on plants are not always straightforward but rather depend on the species, the cultivar, and developmental stage of the plants as well as mineral nutrition in the soil, drought, and local air pollutants (Caldwell et al., 1995; UNEP, 1998).

In affecting plants, enhanced UV-B radiation may ultimately lead to changes in entire ecosystems. In nonagricultural ecosystems (e.g., forests and grasslands), the balance of plants may change as some species are less able to respond to increases in UV-B radiation and their productivity declines. At the same time, the productivity of more responsive species will likely increase. The overall species composition of ecosystems will change, as will species interactions and ecosystem dynamics (Caldwell et al., 1995; UNEP, 1998).

Link to Global Climate Change

Increased levels of UV-B radiation may also affect the balance of carbon dioxide (CO_2) into and out of the biosphere, ultimately contributing to the problem of global climate change. For example, phytoplankton absorb CO_2. As discussed above, stratospheric ozone depletion leading to enhanced UV-B is associated with decreased productivity in phytoplankton. This means the reduction of a major sink, resulting in increased levels of atmospheric CO_2 (Häder et al., 1995). In addition, higher levels of UV-B radiation may increase the decomposition rate of nonliving organic matter and reduce photosynthesis, thereby increasing the amount of CO_2 that is emitted into

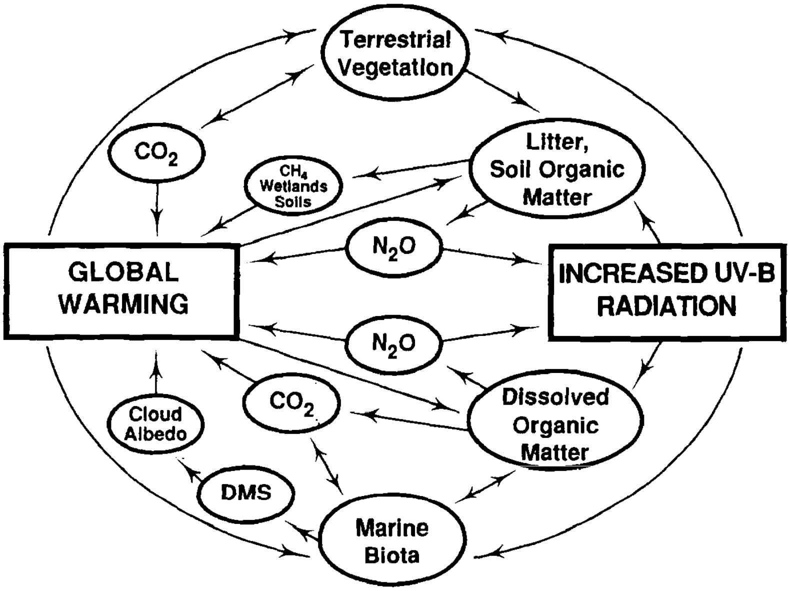


Figure 2 Ways in which increased levels of UV-B are linked to increased levels of CO_2 (from SCOPE, 1993, p. 18).

the atmosphere (SCOPE, 1993; UNEP, 1998). Figure 2 illustrates ways in which increased levels of UV-B are linked to increased levels of CO_2.

3 INTERNATIONAL RESPONSES TO STRATOSPHERIC OZONE DEPLETION

By the early 1980s, it became apparent that combating the problem of ozone depletion would require coordinated international action. Evidence began to mount that ozone depletion was linked to increased rates of skin cancer and that the ozone layer was being depleted at a faster rate than initially imagined (Benedick, 1991; Morrisette, 1989). In 1985, 20 countries signed the Vienna Convention for the Protection of the Ozone Layer, a very general agreement that called upon parties to cooperate in ozone research and to exchange information on the problem of ozone depletion (Roan, 1989; Caldwell, 1990). The Vienna Convention served to establish the recognition of a problem that needed to be dealt with through international cooperation.

The first reports of an "ozone hole" came almost immediately following the signing of the Vienna Convention. Joseph Farman and his colleagues at the British Antarctic Survey station in Halley Bay, Antarctica, who had been monitoring stratospheric ozone over the southern pole since 1957, reported that in September and

October (Austral spring) nearly 60% of the ozone layer over Antarctica was depleted (Farman et al., 1985). Researchers have also found that other regions of the globe experience seasonal decreases in the stratospheric ozone layer. In the midlatitudes, decreases are most severe during the winter/spring months (WMO, 1995).

The 1992 and 1993 Antarctic ozone "holes" were the most severe on record; in some areas, stratospheric ozone was depleted by more than 99%. Researchers believe, however, that these depletions were due at least in part to the eruption of the Mount Pinatubo volcano in 1991. The eruption emitted a substantial amount of sulfate aerosols into the stratosphere, which are believed to enhance the effectiveness of chlorine and bromine as catalysts for ozone destruction (WMO, 1995).

The Montreal Protocol on Substances that Deplete the Ozone Layer was signed by 28 countries in September 1987. The Montreal Protocol established specific measures for countries to control worldwide emissions of several ozone-depleting substances. Specifically, the protocol called for a freeze on the consumption and production of ozone-depleting substances at 1986 levels by 1990, a 20% reduction by 1994, followed by a further 30% reduction by 1999. The agreement entered into force in January 1989. While reports of the ozone hole may have prompted countries to move quickly on the issue of ozone depletion,* the fact that the major CFC producers promised alternatives could be developed in a relatively short time also facilitated international cooperation.

The Montreal Protocol has undergone four major revisions since 1987. The London Amendments to the Montreal Protocol were passed by the Conference of the Parties (COP) at their second meeting in June 1990. These amendments required industrialized countries to completely phase out CFCs by the year 2000 and expanded the number of ozone-depleting substances controlled by the agreement (Benedick, 1991; Parson and Greene, 1995). The protocol was further amended in 1992 at a COP meeting in Copenhagen. The most significant outcome of this meeting was that the phase-out date for CFCs was moved to January 1, 1996 (Parson and Greene, 1995). In September 1997, parties adopted the Montreal Amendments to the Protocol, which strengthened regulation of methyl bromide. The 1999 Beijing Amendment introduced a 2002 phase-out of bromochloromethane and placed controls on the production of and trade in hydrochlorofluorocarbons (HCFCs – a CFC substitute). As of April 2002, there are 183 parties to the Montreal Protocol.

According to a 1995 assessment by the World Meteorological Organization (WMO), the ozone layer is expected to recover by the middle of the next century thanks to these international efforts. CFCs have an atmospheric lifetime of several decades, thus it will take some time before concentrations of these ozone-depleting substances decrease. Researchers expect global UV levels to peak around the turn of the century after which they expect atmospheric concentrations of chlorine and bromine to begin to decline, thereby slowing ozone depletion (UNEP, 1998; WMO, 1995). Figure 3 illustrates the impact of the Montreal Protocol and its amendments on stratospheric ozone levels.

* There is ongoing debate on the role of the ozone hole in the international policy process. For more information, see Benedick (1991), Lambright (1995), Morrisette (1992), Ungar (1995).

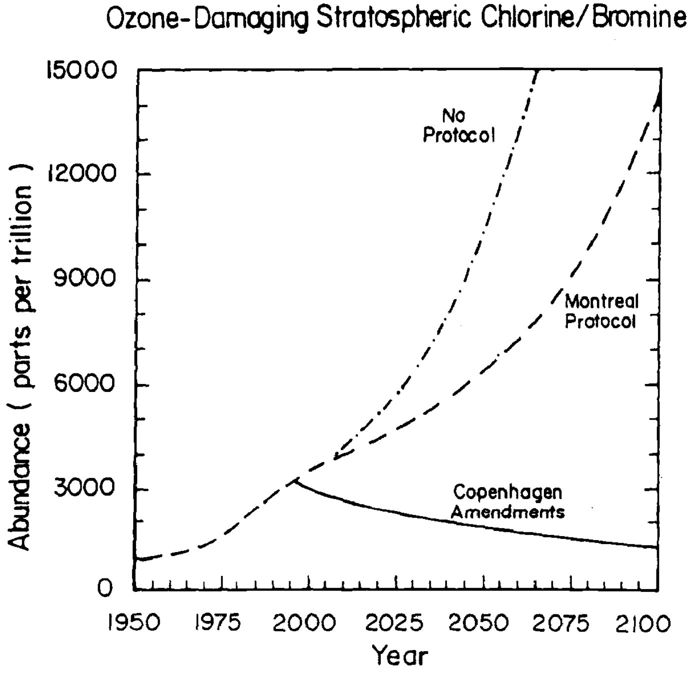


Figure 3 Impact of Montreal Protocol on stratospheric ozone levels (from WMO, 1994, p. 28).

4 REMAINING CHALLENGES IN ADDRESSING STRATOSPHERIC OZONE DEPLETION

Despite the tremendous progress made in addressing the problem of stratospheric ozone depletion, challenges remain. For example, the major replacements for CFCs, HCFCs and hydrofluorocarbons (HFCs), are not totally benign in their effects on stratospheric ozone and may pose other environmental risks. HCFCs do have the potential to deplete stratospheric ozone. However, their atmospheric lifetime is much shorter than for CFCs; thus their ozone-depleting potential is lower than for CFCs and halons (WMO, 1995). In addition, some HCFCs and HFCs react with agents in the atmosphere to produce trifluoroacetic acid (TFA). Once produced, TFA returns to Earth's surface via precipitation. TFA is mildly toxic to both marine and freshwater phytoplankton. There is concern that TFA levels in some areas, particularly those with restricted aquatic outflow, will become toxic (Häder et al., 1995; Schwarzbach, 1995).

Perhaps the greatest remaining challenge in addressing stratospheric ozone depletion is the need to regulate methyl bromide. Bromine, which is derived from methyl bromide, is estimated to be 50 times more efficient at destroying stratospheric ozone than chlorine (O'Meara, 1996; WMO, 1995). Methyl bromide, which has an average atmospheric lifetime of 1.3 years, accounts for between 5 and 10% of observed ozone depletion. This could increase to 17% if emissions continue to grow at current rates (O'Meara, 1996).

The majority of methyl bromide entering the atmosphere originates from the oceans. This source accounts for 60 to 160 ktons of methyl bromide a year. The primary anthropogenic sources derive from soil fumigation (20 to 60 ktons per year), biomass burning (10 to 50 ktons per year), and exhaust from cars using leaded gasoline (0.5 to 1.5 ktons per year) (WMO, 1995). In addition to its ozone-depleting potential, methyl bromide is also highly toxic to humans and animals. Exposure may result in a variety of symptoms including dizziness, headaches, nausea, respiratory irritation, persistent numbness in the extremities, convulsions, and, in extreme cases, coma or death (Longstreth et al., 1995).

In 1992, parties to the Montreal Convention agreed that developed countries should freeze production of methyl bromide at 1991 levels by 1995. They strengthened the regulation in 1995 and again in 1997, agreeing to a phase out by January 1, 20005.* Developing countries will be required to freeze production of methyl bromide in 2002 based on an average for the years 1995 to 1998. The Montreal Amendments, passed in 1997, established a schedule for phasing out methyl bromide production in developing countries by 2015. Unfortunately, these regulations pertain only to the production of methyl bromide; they do nothing to eliminate the use of methyl bromide.

5 CONCLUSION

The case of stratospheric ozone depletion and the international response to it offer several lessons for dealing with future similar environmental challenges. First, it illustrates that technology may be both a cause and a solution to global environmental problems. The development of CFCs and the use of methyl bromide as a pesticide beginning in the 1930s increased concentrations of ozone-depleting substances in the atmosphere and thus increased the rate of ozone depletion. At the same time, the development of alternative technologies such as HCFCs and HFCs is often seen as the key to motivating countries to take aggressive measures to address the problem of ozone depletion. Without these alternatives, the cost of phasing out the use of CFCs would have been prohibitive.

The ozone case also reminds us that there are no easy trade-offs. While CFCs have high ozone-depleting potential, their alternatives are not entirely environmentally sound. HCFCs and HFCs do have some ozone-depleting potential (though much lower than for CFCs) and may have other negative effects on the environment. Addressing environmental issues often requires making painful choices.

* The United States will phase out production of methyl bromide in 2001 as part of the Clean Air Act Amendments of 1990.

REFERENCES

Benedick, R., *Ozone Diplomacy*, Harvard Press, Cambridge, MA, 1991.

Cagin, S., and P. Dray, *Between Earth and Sky: How CFCs Changed Our World and Endangered the Ozone Layer*, Pantheon Books, New York, 1993.

Caldwell, L. K., *International Environmental Policy: Emergence and Dimensions*, Duke University Press, Durham, NC, 1990.

Caldwell, M. M., A. H. Teramura, M. Tevini, J. F. Bornman, L. O. Bjorn, and G. Kulandaivelu, Effects of increased solar ultraviolet radiation on terrestrial plants. *Ambio*, *24*, 166–173, 1995.

Farman J. C., B. G. Gardiner, and J. D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction, *Nature*, *315*, 207–210, 1985.

Häder, D. P., R. C. Worrest, H. D. Kumar, and R. C. Smith, Effects of increased solar ultraviolet radiation on aquatic ecosystems, *Ambio*, *24*, 174–180, 1995.

Lambright, W. H., NASA, ozone, and policy-relevant science, *Res. Policy*, *24*, 747–760, 1995.

Lipkin, R., Ozone depletion research wins Nobel, *Sci. News*, *148*, 262, 1995.

Longstreth, J. D., F. R. de Gruijl, M. L. Kripke, Y. Takizawa, and J. C. van der Leun, Effects of increased solar ultraviolet radiation on human health, *Ambio*, *24*, 153–165, 1995.

Molina, M. J., and F. S. Rowland, Stratospheric sink for chlorofluorocarbons: Chlorine atomic-atalysed destruction of ozone, *Nature*, *249*, 810, 1974.

Morrisette, P. M., The evolution of policy responses to stratospheric ozone depletion, *Nat. Resour. J.*, *29*, 793–820, 1989.

Morrisette, P. M., The Montreal protocol: Lessons for formulating policies for global warming, *Policy Stud. J.*, *19*, 152–161, 1992.

National Academy of Sciences (NAS), *Environmental Impact of Stratospheric Flight: Biological and Climatic Effects of Aircraft Emissions in the Stratosphere*, NAS, Washington, DC, 1975.

O'Meara, M., The next hurdle in ozone repair, *World Watch*, November/December 1996, P. 8.

Parson, E. A., and O. Greene, The complex chemistry of the international ozone agreements, *Environment*, March 1995; pp. 17–20, 35–43.

Roan, S., *Ozone Crisis: The 15-Year Evolution of a Sudden Global Emergency*, J Wiley, New York, 1989.

Schwarzbach, S. E., CFC alternatives under a cloud, *Nature*, *376*, 297–298, 1995.

Scientific Committee on Problems of the Environment (SCOPE), *Effects of Increased Ultraviolet Radiation on Global Ecosystems*, SCOPE Secretariat, Paris, 1993.

Slaper, H., G. J. M. Velders, J. S. Daniel, F. R. de Gruijl, and J. C. van der Leun, Estimates of ozone depletion and skin cancer incidence to examine the Vienna Convention achievements, *Nature*, *384*, 256–258, 1996.

United Nations Environment Programme, (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1994.

United Nations Environment Programme, (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1996.

United Nations Environment Programme (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1998.

Ungar, S., Social scares and global warming: Beyond the Rio Convention, *Soc Nat. Resour.*, *8*, 443–456, 1995.

World Meteorological Organization (WMO), *Scientific Assessment of Ozone Depletion: 1994, Executive Summary,* Global Ozone Research and Monitoring Project-Report No. 37, WMO, Geneva, 1995.

CHAPTER 52

TROPICAL DEFORESTATION AND CLIMATE

ROGER A. SEDJO

1 INTRODUCTION

Tropical forests cover a large portion of the globe's land surface running along the equator, roughly between the Tropic of Cancer and the Tropic of Capricorn. The largest expanse of tropical forest is found in the South American equatorial region, predominantly in the Amazon Basin, but extending up into Central America and down into northern Argentina. Large tropical forests are also found in the equatorial regions of Africa and West Africa and in Southeast Asia, running from India to Malaysia, north into China, and continuing to the islands of the East Indian Archipelago and extending into northeast Australia.

Tropical forests take many forms, largely controlled by variation in rainfall, temperature, and season, but also are affected by soil conditions. The climate of the region between the Tropics of Cancer and Capricorn is uniform in that there is a steady year-round temperature. However, annual rainfall may vary from less than 10 mm along the Peruvian coast to more than 10 m along the Colombian coast only a few hundred kilometers to the north (Terborgh, 1992).

Although tropical forests are often rainforest or wet tropical forest, large areas of dry tropical forests exist in almost all of the regions discussed, covering large areas in South and Central America as well as Africa and, to a lesser extent, Southeast Asia. Tropical forests range from open savannas where rainfall is limited to dense tropical rainforests where rainfall is most abundant. Obviously, the type of tropical forest that occurs in an area depends critically upon the availability of precipitation and moisture. The annual cycle of seasonal change is also an important feature of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

tropical climates, but the seasons are characterized by variation in rainfall rather than temperature. Evergreen forests occur where there is little or no dry season. Where dry and wet seasons are of approximately equal duration, deciduous forests are the norm.

A unique feature of tropical forests is the broad representation of biodiversity in a small area. Tropical forests contain much, if not most, of the world's biodiversity in the trees and plants that comprise the vegetative system and in the animals, especially anthropoids, that exist in the forest soils, floor, and canopy. Tropical forests, especially wet tropical forests, typically contain far more species of trees, plants, birds, butterflies, and so forth than their temperatc counterparts.

In general, the net primary productivity (NPP), a measure of plant growth given by the net amount of carbon fixed by plants in a given time period, increases the closer the forest is to the equator, although it is moderated by rainfall patterns.

As Table 1 shows, the NPP of tropical forests is substantially greater than that of other types of forest ecosystems, thus indicating higher levels of overall biological growth. This is the case even though tropical soils are often poor in nutrients and minerals as a result of exposure to torrential rainfalls over millennia, which have washed away most of the soluble materials. The high productivity found in these sterile soils is due to the ability of these forests to store the materials in the forest itself—either in living plants or in the litter of dead plants (Terborgh, 1992).

Without human intervention, tropical forest ecosystems are a mixture of living, growing, and dying systems that are periodically impacted by natural disturbances. Natural disturbances, as, for example, when an old tree falls, contribute to this biodiversity by creating numerous unique niches that create a host of varying environments thereby promoting the wide range of biodiversity. The disturbances vary from local disturbances that may open a small hole in the forest canopy that allows sunlight to reach the forest floor thereby stimulating small seedlings, to much broader disturbances. Examples of broader disturbances include those caused by land slides and floods, as well as those caused by forest fires, which occur not only in dry tropical forests but also in wet forests that experience distinct dry seasons, such as those found in some parts of Borneo and Southeast Asia.

TABLE 1 Predicted Net Primary Productivity (NPP) by Forest Ecosystem Types

Ecosystem Type	NPP (g/m^2/yr)
Boreal forest	224.20
Temperate conifcr	459.25
Temperate deciduous	361.17
Temperate broadleaf evergreen forest	590.29
Subtropical and tropical	892.74

Source: FAO (1992).

2 HUMAN INFLUENCES IN THE TROPICS

Humans introduce additional complexity to this system by introducing another possible source of disturbance through the process of utilizing forests for their own needs. Few tropical forests are free from the effects of humans. Most tropical forests have been inhabited by humans, many for thousands or tens of thousands of years. Human impacts typically involve the collection of various forest items for food and fiber. Very often humans provide management to increase the future supply of desired outputs. In a subsistence context such management may include modifying the forest to promote certain desired nontimber forest products, e.g., promoting the growth of certain fruit and nut-bearing trees or promoting the growth of young edible bamboo shoots.

In addition, shifting cultivation has been practiced in some tropical forests for millennium. Under this system a small area is cleared and usually burned, in part to provide a nutrient-rich ash. Crops are planted on the site for several years until it loses fertility or the weeds become untenable. The site is left fallow for a period of years, often more than 20, at which time the site is again cleared, burned, and replanted in crops. This cycle appears to be sustainable indefinitely. Thus, even in subsistence and relatively primitive societies, humans have influenced the forest ecosystem. When considering the sustainability of forests, human behavioral influences must be considered as well as natural adjustments.

Forests have also been utilized for millennia by people for building materials and timber. Although there has always been local use and some international trade in tropical timbers, the importance of trade has increased in the post–World War II global economy. Timbers from the West Coast of Africa have been supplemented by much larger volumes entering international trade from the East Indian countries of Malaysia, the Philippines, and Indonesia. This trade has provided these countries with substantial volumes of foreign currencies, which have provided some of the capital that financed their economic development.

3 VALUES OF FORESTS

Tropical forests, indeed all forests, generate a host of values to humans. Human benefits often involve the collection of various forest items for food and fiber such as the various timber and nontimber forest outputs discussed above. In addition, tropical forests provide local and regional ecological services in the form of watershed protection, mitigation of soil erosion, and reduction of downstream flooding. Additionally, tropical forests provide the residence for much of the world's biodiversity, with the majority of species believed to be found in the tropical forest habitat. Finally, tropical forest, together with the rest of the world's forests, provide the majority of the world's biomass that provides a "sink" for carbon, thus mitigating the buildup of carbon in the atmosphere, which is believed to contribute to global warming.

TABLE 2 Recent Estimates of Carbon Flux, Pg C yr^{-1}, from the Tropical Landscape for 1980 and 1990

Source	Year	Range	Average
Detwiler and Hall (1988)	1980	0.4–1.6	1.0
Hall and Uhlig (1991)	1980	0.52–0.64	0.58
Houghton et al. (1987)	1980	0.9–2.5	1.7
Houghton (1992)	1990	1.2–2.2	1.7

Source: Brown et al. (1993, p. 79).

Forests are well recognized as having the potential to affect the forestlands' microclimate. Additionally, forests have an impact on the global climate through their capacity to sequester carbon. Although other terrestrial systems also sequester carbon, forests by far constitute the largest terrestrial carbon pool; tropical forests account for about one-half of all forest area and perhaps a large portion of the total forest biomass (Brown et al., 1993). By holding carbon captive in pools of forest biomass and soils, the amount of carbon in the atmosphere is reduced, thereby mitigating the greenhouse warming provided by the atmosphere.

While it is widely agreed that the preponderance of human-generated releases of carbon into the atmosphere in recent decades has been due to the use of fossil fuels and that the global warming "problem" is largely a fossil fuel problem, it is also recognized that land-use changes play a role. It is generally believed that some of the buildup of carbon dioxide in the atmosphere experienced over the past century or so is due to land-use changes associated with land clearing and deforestation. In recent decades, probably all of the net carbon releases from forests have come from tropical deforestation (since temperate and boreal forests are in approximate carbon balance). An estimate of the carbon releases in recent years is provided in Table 2.

The range of carbon releases from tropical forests is from 0.4 to 2.5 Pg C yr^{-1}. This compares with an estimate of about 6.0 Pg C yr^{-1} total human-generated releases of carbon. Thus, although tropical forest carbon releases are significant, they are well below 50% of total annual releases and probably in the range of 10 to 25% of the total.

4 DEFORESTATION IN THE TROPICS

In 1990 the tropical forests of the world were estimated to cover an area of about 1,756.3 million ha (Table 3), or about 13.4% of the globe's land area, excluding Antarctica and Greenland. This is down from an estimated 1,910.4 million ha in 1981 (Table 3). The annual deforestation rate for this period averaged 0.1540 million km^2 or about 0.8% of the area of tropical forest (FAQ, 1993). The rate of deforestation, however, varied substantially throughout the tropics. Perhaps somewhat surprisingly, the tropical rainforest, the forest type over which the international community

TABLE 3 Estimates of Forest Area and Rate of Deforestation by Geographical Subregions[a]

Continent	Number of Countries	Total Land Area[a]	Forest Area 1980[a]	Forest Area 1990[a]	Annual Deforestation 1981–1990	Rate of Change 1981–1990 (percent per annum)
		million hectares				
Africa	40	2236.1	568.6	527.6	4.1	-0.7
West Sahelian Africa	9	528.0	43.7	40.8	0.3	-0.7
East Sahelian Africa	6	489.7	71.4	65.3	0.6	-0.8
West Africa	8	203.8	61.5	55.6	0.6	-1.0
Central Africa	6	398.3	215.5	204.1	1.1	-0.5
Tropical southern Africa	10	558.1	159.3	145.9	1.3	-0.8
Insular Africa	1	58.2	17.1	15.8	0.1	-0.8
Asia	17	892.1	349.6	310.6	3.9	-1.1
South Asia	6	412.2	69.4	63.9	0.6	-0.8
Continental South East Asia	5	190.2	88.4	75.2	1.3	-1.5
Insular South East Asia	5	244.4	154.7	135.4	1.9	-1.2
Pacific Islands	1	45.3	37.1	36.0	0.1	-0.3
Latin America	32	1650.1	992.2	918.1	7.4	-0.7
Central America	7	239.6	79.2	68.1	1.1	-1.4
Mexico						
Caribbean	19	69.0	48.3	47.1	0.1	-0.3
Tropical South America	7	1341.6	864.6	802.9	6.2	-0.7
Total	90	47,783	1910.4	1756.3	15.4	-0.8

[a] Totals may not tally due to rounding.

Source: FAO (1993).

seems to have the greatest concern, experienced the slowest rate of overall deforestation at 0.6% annually. The highest rates of deforestation were experienced in the upland forests. Both moist and dry upland forests experienced a 1.1% annual rate of deforestation (Table 4). By major region, Central America, including Mexico, experienced the highest rate of deforestation at about 1.4% annually while the Caribbean had the lowest rate at 0.3% annually. Of the large forest formations, continental Southeast Asia had the most rapid rate of deforestation at 1.5% annually while Central Africa had the lowest rate at 0.5% annually.

It is estimated that 90% of tropical deforestation has occurred since 1970 (Skole et al., 1994). If this estimate is correct, the tropical forest of the world at its apex would have covered about 22 million km^2 or about 16.8% of the globe's land surface. Although reduced in size the world's tropical forests still constitute an area equal to that of the whole of South America. Even at the current rate of tropical deforestation, the world's tropical forests would continue to exist through the entire twenty-first century and well into the twenty-second century. Of course, the rate of tropical deforestation will almost surely be changing over time.

TABLE 4 Estimates of Forest Cover Area and Rate of Deforestation by Main Forest[a]

Forest Formations	Land Area (million hectares)	Population Density 1990	Population Growth (1981–1990)	Forest Area 1990		Annually Deforested Area (1981–1990)	
		(inh./km)	(% per year)	(million hectares)	(% of land area)	(million hectares)	(%)
Forest Zone	4186.4	57	2.6	1748.2	42	15.3	0.8
Lowland formations	3485.6	57	2.5	1543.9	44	12.8	0.8
Tropical rainforest	947.2	41	2.5	718.3	76	4.6	0.6
Moist deciduous forests	1289.2	55	2.7	587.3	46	6.1	0.9
Dry deciduous forests	706.2	106	2.4	178.6	25	1.8	0.9
Very dry zone	543.0	24	3.2	59.7	11	0.3	0.5
Upland formations	700.9	56	2.9	204.3	29	2.5	1.1
Moist forests	528.0	52	2.7	178.1	34	2.2	1.1
Dry forests	172.8	70	3.2	26.2	15	0.3	1.1
Nonforest Zone (hot and cold deserts)	591.9	15	3.5	8.1	1	0.1	0.9
Total Tropics	4778.3	52	2.7	1756.3	37	15.4	0.8

[a] Totals may not tally due to rounding.

Source: FAO (1993).

The causes of tropical deforestation are complex and not well understood. The term *deforestation* refers to situations in which the land is more or less permanently converted from forest cover to other cover and/or uses. A common but simplistic view, now largely rejected by analysts familiar with tropical forests, is that tropical deforestation is due to commercial timber logging. Commercial logging in the tropics rarely results in significant direct land conversion; however, as discussed below, it does make indirect contributions to the process of deforestation.

Another common explanation of tropical deforestation is to attribute it to population growth. As populations rise, one would expect that the pressures on the forests might increase. It is also argued that population pressures force a reduction in the fallow period in slash-and-burn regimes increasing pressures to bring more forest land into agricultural use. However, it is difficult to directly link population and economics growth to tropical deforestation, for example, Skole et al. (1994) conclude that population growth alone does not explain tropical deforestation.

Most analysts now believe that most tropical deforestation is driven by human desire for land-use changes, primarily the replacement of forests by agricultural activities. In fact, many governments have currently, or have had in the past, explicit policies to promote the conversion of forests to agricultural uses. In Central and South America, for example, large areas of forest clearing reflect government policy to open forest areas to agricultural settlement. In Brazil, for example, one rationale for promoting migration into the Amazon region, with its attendant deforestation, was to solidity Brazil's sovereignty claims to the region. Other clearing for use as pasture and other agriculture results from spontaneous actions of individuals. In Southeast Asia, forest land was gradually converted to paddy lands in river bottomlands over many years, and more recently conversion has occurred where water development projects, which allow irrigation, have been implemented. Additionally, in Southeast Asia, substantial native forests have been replaced over the years by the introduction of tree crops including rubber, palm oil, coconut, and so forth.

As noted, although commercial logging rarely plays a direct role in deforestation, it often plays an indirect role by providing access to previously inaccessible forest areas. Logging penetrates the dense forest with roads, which then provide access for spontaneous migration. The forest, which is accessible due to the logging roads and is not less dense and impenetrable due to the logging, is now more vulnerable to alternative land-use activities and land-use changes. The improved access makes investments in land clearing, spontaneous and otherwise, more feasible and attractive.

Although governments are often involved in explicitly promoting forest conversion, they are also often involved by the absence of their management of the forests that fall under their jurisdiction. In much of the tropical world today the forests are under the control of the central government, even though in earlier periods they had tended to be under local control. Although the central government has responsibility for the forests, often it is unable or unwilling to exercise effective management control. Thus the forests often become degraded or destroyed by uncontrolled use. In effect, the forests become a type of open access resource over which the responsible authority does not exercise effective control and the users have only illegal or attenuated use rights that provide little or no incentive for long-term management. More generally, careless use of tropical forest land often signals the absence of clear well-defined property rights and/or effective management, either private or public. With unsecured rights, the incentives are for "cut-and-run" behavior.

5 SIMILARITIES WITH EARLIER DEFORESTATIONS

In many respects tropical deforestation today is not dramatically different from temperate deforestation that occurred one and two centuries earlier. During that period, pressures for land-use change, primarily the demand for new lands for agriculture, resulted in large-scale deforestation of areas of Europe and North America. In the United States much of the forestlands of the eastern seaboard, the south and the Great Lakes states, were converted to croplands and pastures. This same phenomenon had begun earlier in Europe but continued in places well into the early part of the twentieth century. The denuding of the forest landscape was often the result of spontaneous actions but also often reflected govemmental policies. In the United States, for example, the Homestead Act required land clearing as a prerequisite of obtaining land title. For much of North America and Europe the early land clearing has been offset by the renewal of the forest, largely through natural processes. Today, as the work of Kuusela (1994) of the European Forestry Institute has shown, the European forest has reclaimed large areas once deforested. Similarly, in America the forest has reclaimed much of the area deforested in New England (Barrett, 1988), the Great Lakes states, and the south as abandoned agricultural lands regenerated naturally into forest and, more recently, planted forest cover many former tobacco, cotton and other crop lands.

6 TIMBER HARVESTS IN THE TROPICS

Unlike much of the commercial logging in the temperate forest, even today tropical commercial logging almost never involves clear-cutting of the forest. Rather, the usual approach is to select only the trees that are suitable for commercial uses and log those trees, leaving large numbers of live trees in the forest. In past periods, relatively few trees were removed, and those that were felled by hand were commonly transported out of the forest by animals. The forest would be periodically relogged as trees reached desired sizes.

In recent decades logging has involved chainsaws, roads, and equipment. Additionally, larger areas have been logged, reflecting expanded demand. Nevertheless, selective cutting is still almost universally practiced, with only the larger trees of desired species harvested. This practice does not indicate benign motives by loggers but rather reflects the fact that, due to the high diversity of tree species, only a relatively few of the total trees of the tropical forest are suitable for commercial markets. Additionally, unlike many temperate forests that are even-aged (i.e., most all the mature trees are the same age and species), tropical forests typically are uneven-aged. The diverse mix of tree species and ages makes clear-cutting an unsuitable and uneconomic approach for commercial logging in most of the tropics.

This selective logging approach is also generally conducive to forest regeneration and regrowth. During the period immediately following the logging, the sunlight now reaching the forest floor stimulates the growth of seeds and seedlings, especially of the so-called pioneer species, which include many of the more important timber species such teak, mahogany, and many of the dipterocarp species. Typically the stock of seed and seedlings is adequate; however, this can be supplemented by human activities if required.

The idealized tropical forest management regime regarding logging varies with forest type. In the timber-rich forests of Southeast Asia it is common to follow the logging with a period of 30 to 70 years during which the forest recovers and has time to grow new trees of the desired size. Additionally, existing saplings and medium size trees will continue to grow and, in some cases, accelerate their growth now that the dominant trees are gone. When such an approach is followed, a viable and sustainable forest system can be achieved.

7 RENEWABILITY

Although it is sometimes claimed that tropical forests have difficulty renewing themselves, the evidence is to the contrary. For example, large areas of the American tropics had been in terraces, irrigated agriculture, and agroforestry in the pre-Columbian period but reverted to tropical forests as local populations were decimated by disruptions and disease. For example, Turner and Butzer (1992) argue that "the scale of deforestation, or forest modification, in the American tropics has only recently begun to rival that undertaken prior to the Columbian encounter." Similarly, the great temples of Angkor Watt in Cambodia, Borobodor in Java, and other similar large structures in Southeast Asia, once located in the mist of a high level of human activity, were lost for centuries due to the incursion of tropical forest when human activity declined. Also, the banks of the Panama Canal, which were almost wholly defoliated during the canal's construction in the early 1900, are now covered with lush tropical forests.

8 CONCLUSIONS

A major difference between the prior and current view of tropical deforestation is that the global community is now aware of and concerned for both the preservation of biodiversity and the sequestration of carbon, two global ecological functions that were largely unknown and generated little concern one and one half centuries ago. Another consideration being faced by the global community is increasing awareness that, although deforestation is reversible, species loss is not. Thus, the tropical forests that could be regenerated in the future may not have all of the constituent parts of the original forest. Although there is no guarantee that large amounts of additional deforestation will be prevented nor that reforestation will follow the pattern of the temperate industrial countries, recent events including the Earth Summit in Rio in 1992 have demonstrated the growing concern over deforestation by the global community and the emerging of a commitment to moderate, if not preclude, its continuance.

REFERENCES

Barrett, J. W., ed; The northeastern region, in *Regional Siliviculture of the United States*, Wiley, New York, pp. 25–65, 1988.

Brown, S., C. Hall, W. Knabe, J. Raich, M. Trexler, and P. Woomer, Tropical forests: Their past, present, and potential future role in the terrestrial carbon budget, in terrestrial biospheric carbon fluxes, in J. Wisniewski and R. N. Sampson (Eds.), Quantification of Sinks and Sources of CO_2, Kluwer Academic, Dordreceht, 1993, pp. 71–94.

Detwiler, R. P., and C. A. S. Hall, Tropical forests and the global carbon cycle, *Science*, *239*, 42–47, 1988.

Food and Agricultural Organization of the United Nations (FAO) *FAO Yearbook: Forest Products, 1981–1992*, FAO Forestry Series No. 27, FAO, Rome, 1992.

Food and Agricultural Organization of the United Nations, (FAO), *Forest Resources Assessment 1990: Tropical Countries*, FAO Forestry Paper 112, FAO, Rome, 1993.

Hall, C. A. S., and J. Uhlig, Refining estimates of carbon released from tropical land-use change, *Canadian Journal of Forest Research*, *21*, 118–131, 1991.

Houghton, R. A., R. D. Boone, J. R. Fruci, J. E. Hobbie, J. M. Melillo, C. A. Palm, B. J. Peterson, G. R. Shaver, G. M. Woodwell, B. Moore, D. L. Skole, and N. Meyers, The flux of carbon from terrestrial ecosystems to the atmosphere in 1980 due to changes in land use: geographic distribution of the global flux, *Tellus*, *39B*, 122–139, 1987.

Houghton, R. A., Tropical forests and climate, paper presented at the International Workshop Ecology, Conservation, and Management of Southeast Asian Rainforests, October 12–14, Kuching, Sarawak, 1992.

Kuusela, K., *Forest Resources in Europe*, Cambridge University Press, Cambridge, 1994.

Skole, D. L., W. H. Chomentowski, W. A. Salas, and A. D. Nobre, Physical and human dimensions of deforestation in Amazonia, *BioScience*, *44*(5), 314–322, 1994.

Terborgh, J., *Diversity and the Tropical Rain Forest*, Scientific American Library, New York 1992.

Turner II, B. L., and K. I. Butzer, The Columbian encounter and land-use change, *Environment*, *34*(8), 16–20, 37–44, 1992.

CHAPTER 53

DESERTIFICATION

R. L. HEATHCOTE

1 INTRODUCTION: ORIGINS OF CONCERNS

Desertification is commonly understood to mean the spread of desertlike conditions. The term came to international attention with the environmental devastation and loss of human and animal life accompanying extensive droughts in West Africa's Sahel region over the period 1968 to 1973, which were captured on televised documentary programs and in the print media. This media coverage likely played an important part in the subsequent extensive United Nations–sponsored international and national aid programs and associated scientific investigations into what appeared to be evidence of long-term environmental deterioration. These activities led to the United Nations Conference on Desertification (UNCOD) in Nairobi, Kenya, in 1977.

In fact, however, concern for evidence of adverse environmental change can be found in the accounts of European explorers in the deserts of Africa, Arabia, and Asia from the late eighteenth century onward. A Danish expedition to Arabia of 1761 to 1767 (Hansen, 1965) and the activities of explorers such as Charles Doughty (1843–1926), Sven Hedin (1865–1952), Baron von Richthofen (1833–1905), and Elsworth Huntington (1876–1947) uncovered evidence of past civilizations amid the desert sands. For example, Huntington's 1907 *The Pulse of Asia* drew upon his own and the earlier explorations to suggest that the evidence of climatic variation and associated environmental deterioration and reduced capacity to support the population in central Asia might have caused the folk migrations that led to the Mongol invasions of southwest Asia and Europe in the thirteenth century.

In the areas of nineteenth-century European colonization, droughts on the Great Plains of North America in the 1880s to 1990s (e.g., Brown, 1948), eastern Australia

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

1895 to 1902 (Heathcote, 1965), and southern Africa in 1918 and the early 1920s (Kokot, 1955) raised concerns for the long-term viability of human occupation. The 1930s saw further concerns for what appeared to be a combination of climatic variability and human mismanagement of the land's resources, with the evidence of the Dust Bowl on the southern Great Plains in 1935 and the first global study of soil erosion (Jacks and Whyte, 1938). In West Africa, Stebbing (1935, 1937) anticipated an advance of the southern edge of the Sahara as a result of overzealous clearance of the forest and savannah vegetation, and Ratcliffe (1938) studied the link between overgrazing and the expansion of the Australian deserts.

Post–World War II reconstruction efforts reflected continuing popular concerns about human use of the land in such writings as Sear's *Deserts on the March* (1949) for the Great Plains, Calder's *Men against the Desert* (1951), and Aubreville's *Climats, Forêts et Desertification de l'Afrique Tropicale* (1949) for African deserts. Pick and Aldis (1944) published concerns for *Australia's Dying Heart*. Lowdermilk (1953) undertook a global review of soil resource mismanagement and repeated earlier calls to reverse the loss of resources.

That such reversals were possible was the claim of the Israeli author, Reifenberg, whose book, *The Struggle between the Desert and the Sown* (1955), documented the efforts of the new nation of Israel to reclaim the deserts as the culmination of a long history of desert advances and retreats in this part of southwest Asia. In 1966 a retired British forester (Baker, 1966) with experience in Kenya and Nigeria and founder of the Men of the Trees Society, published his *Sahara Conquest*, yet another scenario to reclaim the desert, in this case by massive tree-planting schemes.

International scientific interest culminated in the United Nations Arid Zone Research Program, which ran from 1951 to 1971 (UNESCO, 1953). This program brought together scientists from around the world to study the globe's arid areas in an effort to understand both the physical characteristics and the past and present land uses in order to plan better for future management and possible desert reclamation. In effect this program laid the foundations for future United Nations interest in the desertification phenomenon.

This concern over the increasing evidence of environmental deterioration seems to have reflected in part not only the increased scientific evidence of that deterioration, but also a growing moral concern for the global environment and human relations with it. The "environmental movements" of the 1970s and their subsequent "green movements," by focusing upon evidences of resource mismanagement through human ignorance or greed, saw desertification as but one example of the adverse impact of human resource use upon the environment—an impact that had been of concern earlier (Glacken, 1967) and that gained an added ethical dimension (White, 1967; Passmore, 1980; Nash, 1990). Thus, by the 1990s desertification had been consolidated as an international "catch-cry," a political force and a global scientific problem for some of the reasons noted above, but also possibly as a result of the end of the Cold War removing political issues (Driver and Chapman, 1996) that had been competing with environmental ones.

2 DEFINING THE PHENOMENON

Specific definitions of desertification have been many and varied, partly reflecting the fact that it has been seen on the one hand as a *process of environmental deterioration* and on the other as the *product of environmental deterioration*—the devastated environment itself (Glantz and Orlovsky, 1986).

All definitions agree, however, upon certain basic criteria:

- Explicit or implicit evidence of changes in the characteristics of the present environment by comparison with previous characteristics.
- Those changes have resulted in a reduced capacity of the environment to support human life.
- The resultant degraded environment has a desertlike appearance.
- Unless rectified, those changes may continue to further reduce the capacity of the environment to support human life in the future.

To adequately and convincingly identify desertification in those terms, scientists need compatible objective data for a considerable time period for specific areas of the world. Identifying either the processes at work or the end product—the desertified landscape—has proved to be extremely difficult.

The first problem has been the lack of the requisite historical data sets and the assumption of linear trends between those data sets that do exist. The degradation of the soil resource capacity to sustain vegetation or crops (by removal from wind or water erosion, by modification of chemical content through salinization, alkalization, or acidification, or by modification of the soil texture and thus moisture retention capacity by compaction) is recognized to be a significant indicator of desertification. Global descriptions of soil characteristics are still sparse and of varying quality, although formulas for estimates of soil erosion rates are available and have been partly used in the Global Assessment of Human-Induced Soil Degradation (GLASOD). GLASOD was commissioned by the United Nations Environment Programme and was incorporated in the *World Atlas of Desertification* as the most scientifically acceptable measure of desertification (UNEP, 1992).

The other main indicator of desertification is the change in vegetation cover, whether of natural or domesticated plants. Change here may be indicated by reduced vegetation quantity and/or quality (in terms of reduced biodiversity, and/or reduced biomass, and/or productivity), but these indicators are much more difficult to measure. The problem here is the interseasonal and interannual fluctuations in that cover and the extent to which the cover at any given time reflects a linear trend toward increasing or decreasing density or whether it reflects merely cyclical (natural) changes or fluctuations.

In West Africa, where the contemporary concern for desertification originated, there have been serious disagreements among scientists as to the extent of recent vegetation change (Table 1). The differences reflect contrasting assumptions on the

TABLE 1 Conflicting Views of West African Desertification: Forest Loss during the Twentieth Century

	Orthodox View[a] Forest Cover[b]			Revisionist View[a] Forest Cover[b]		
Country	1900	1990s	Loss/Gain (±)	1900	1990s	Loss/Gain (±)
Ghana	9.9	1.6	−8.3	2.5	1.6	−0.9
Cote d'Ivoire	14.5	2.7	−11.8	6.0	2.7	−3.3
Benin	1.1	0.4	−0.7	0.5	0.4	−0.1
Sierra Leone	5.0	0.5	−4.5	0.1	0.5	+0.4
Liberia	6.5	2.0	−4.5	5.5	4.8	−0.7

[a] Sources noted in Fairhead and Leach (1996, p. 189).

[b] Forest cover areas in millions of hectares.

Source: Fairhead and Leach (1996).

historical extent of natural vegetation and the relationships between climate and natural vegetation, differences in the classification of "natural" vegetation, and a failure to recognize historical human-induced deliberate or accidental revegetation of previously sparsely vegetated areas.

The global satellite coverage of the early 1970s onward has improved upon the earlier estimates of vegetation conditions, which had been based upon subjective assessments by individual explorers and other observers. Satellite coverage has provided the base for a Global Vegetation Index—the global vegetation cover averaged over the 1983 to 1990 period as a baseline for subsequent documentation of trends (UNEP, 1992). This of course does not help the debates on changes prior to the 1970s.

A second problem has been the interpretation of the data itself. Faced with what appears to be a desertified landscape in terms of apparently degraded vegetation and/or soils, the scientist must assess whether the landscape is in decline from a more productive past or in process of rehabilitation to a more productive future.

A third problem is the time scale chosen for the analysis of the significance of the environmental changes. While significance on a human scale may be set on scales ranging from interseasonal variations to trends over decades or even a generation (usually seen as 20 to 30 years), significance in ecological terms may range from decades to centuries or millennia, and while these latter scales may be seen as irrelevant for human planning purposes, they may be relevant to any attempt to explain what appears to us to be desertification processes. The basic difficulty remains, however, for the vegetation cover, as for the soil degradation observations, that the implied trend will depend upon the length of time between observations. Oscillations in conditions from whatever cause, which may occur between those observations in time, may not be noticed.

A fourth problem facing definitions of desertification is the fact that there are several interested parties concerned with the phenomenon, some of whom may have an interest in stressing the extent of and dangers from desertification, while others may be more interested in playing down its significance. For scientists seeking research funding for projects of personal interest, for self-identified victims of the process and their political leaders, for political groups who have identified particular organizations or ideologies as contributing to the process, there may be an incentive to exaggerate the significance of the phenomenon, or at least its natural as opposed to the human causes (Glantz, 1977; Heathcote, 1986). For scientists seeking funding for other research areas, for resource managers accused of exacerbating the desertification processes or seeking to gain time to complete exploitive management strategies, and for political groups anxious to defend specific resource management orthodoxies, there may be an incentive to play down the dangers. There seems to be evidence of both approaches to the phenomenon in the literature (Heathcote, 1980; Garcia, 1981; Watts, 1983; Beinart, 1996; Chapman and Driver, 1996).

A fifth and final problem in defining and specifically in accounting for desertification is that the phenomena are of interest to both natural and social scientists. As a result, the investigations of one group may underestimate the significance of factors of interest to the other. This division has been most evident in the debates about whether desertification stems from natural fluctuations; for example, from climatic variations such as drought periods and the associated vegetation death and soil desiccation and erosion, or from the effects of human activities producing environmental stresses that cannot be sustained in particular locations (Rhodes, 1991; Thomas, 1993).

Bearing in mind such difficulties, the variability of definitions is not surprising. However, the consensus view as adopted by the United Nations has evolved from the definition adopted by the UNCOD in 1977:

> Desertification: The intensification or extension of desert conditions; it is a process leading to reduced biological productivity, with consequent reduction in plant biomass, in the land's carrying capacity for livestock, in crop yields and human wellbeing. (UNCOD, 1977, p. 3)

to that used in the *World Atlas of Desertification*:

> desertification/land degradation is defined as: land degradation in arid, semiarid and dry subhumid areas resulting mainly from adverse human impact. (UNEP, 1992, p. vii)

and further modified as:

> Desertification is land degradation in arid, semi-arid and dry sub-humid areas resulting from various factors including climatic variations and human activities. (UN Convention to Combat Desertification in Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa, June 1994)

Thus, consensus has defined the phenomenon as limited to the drier areas of the world, where highly variable climatic conditions, particularly drought, combined with the adverse impacts of human activities appear to constitute a major threat to the long-term sustainability of support for human occupation in those areas.

3 DOCUMENTING DESERTIFICATION

Since UNCOD in 1977, there have been various estimates of the area desertified. The *World Map of Desertification*, showing areas at risk and prepared for the conference, identified 4.56 billion hectares as "affected or likely to be affected" (UNCOD, 1977, Annex 1). If the extreme deserts were excluded, on the grounds that their condition could not worsen, the area was reduced to 3762 million hectares. In 1984, when the progress of the UN Plan of Action to Combat Desertification was reviewed, the area was reduced to 3.48 billion hectares (Tolba, 1984). Dregne and colleagues estimated 3.56 billion hectares in 1991 (Dregne et al., 1991), while the latest UN estimate is 1.04 billion hectares (UNEP, 1992).

In fact, these varying figures are not strictly compatible, since they refer to drylands defined in different ways at the different times and to the inclusion of areas suffering vegetation degradation but that may not have been suffering soil degradation. At least the most recent UN estimate (UNEP, 1992) does include estimates of soil degradation based upon GLASOD.

The 1992 estimates (Table 2) provide basic information on soil rather than vegetation degradation and while, therefore, marginally more acceptable, must still be viewed with caution. The message is that over 15% of the global soils appear to be degraded, over half of that area (53% of the 1.04 billion hectares) is located within the global drylands. The largest degraded areas are in Asia and Africa, with 38 and 25%, respectively, of the global total, and the proportions of the global degraded drylands are similar—36 and 31%, respectively. Interestingly, the largest proportion of regional drylands desertified is in Europe (33%), but all regions have at least 10% of their drylands showing desertification.

TABLE 2 Global Soil Degradation, c. 1992

	Soil Degradation (million hectares)					
Region	In Susceptible Drylands (%)[a]		In Other Areas (%)[b]		Total Degraded Area (%)[c]	
Africa	319.4	24.8	174.8	10.4	494.2	25.2
Asia	370.3	22.1	376.6	14.6	746.9	38.0
Australasia	87.5	13.2	15.4	7.0	102.9	5.2
Europe	99.4	33.1	119.4	18.3	218.8	11.1
North America	79.5	10.9	78.7	5.4	158.2	8.1
South America	79.1	15.3	164.3	13.1	243.4	12.4
Total global	1035.2	20.0	929.2	11.9	1964.4	100.0

[a] Percentage of total area of regional susceptible drylands.

[b] Percentage of total of regional other areas.

[c] Percentage of global total degraded area.

Total degraded area of 1964.4 million hectares is 15.1% of global land area.

Source: UNEP (1992, p. 25).

TABLE 3 The Most Serious Soil Degradation Areas in Drylands, c. 1992

Region	Soil Degradation Areas (strong and extreme) (million hectares)	Percentage of Global Total Areas (strong and extreme) (%)
Africa	74.0	53.8
Asia	43.7	31.8
Australasia	1.6	1.1
Europe	4.9	3.5
North America	7.1	5.2
South America	6.3	4.6
Total	137.6	100.0

Source: UNEP (1992, p. 13).

The most serious dryland soil degradation and implied desertification is identified in Table 3. Using the worst two categories of soil degradation identified (strong and extreme degradation), Asia and Africa again dominate the scene, although it is Africa's turn to lead with almost 54% of the world's seriously degraded dryland areas compared to Asia's 32%. By comparison, the other regions each contain less than 6% of the global areas.

Maps of the desertified areas of the world, regardless of the definitions used, all show the most intensive areas of desertification to lie on the humid edge of the drylands (UNCOD, 1997; Dregne et al., 1991; UNEP, 1992). This is usually explained as the result of the encroachment of agriculture, with its implied soil disturbance, into areas traditionally given over to purely livestock grazing, together with increasing pressure on vegetation resources for fuel and building materials from increased human populations. This somewhat simplistic view, however, has been criticised as noted below.

4 EXPLAINING DESERTIFICATION

There is an extensive literature devoted to alternative explanations for the desertification phenomenon. Essentially, the arguments fall into three camps. First are those that claim natural processes, specifically climate change or variability—mainly through drought, and feedback linkages between the characteristics of the ground surface and air temperatures, are the cause (Bryson and Murray, 1977); second are those that place the blame squarely on human mismanagement of the environment, mainly through human resource demands that exceed the capacity of the environment to supply in the long term (Sinclair and Fryxell, 1985); and third are those that see a mix of both natural processes and human activities combining and interacting to create an unstable environment (Hare et al., 1977).

In the *World Atlas of Desertification* (UNEP, 1992) the causes of desertification were listed as entirely human derived (Table 4) and no specific listing of natural causes was provided. Implicit, however, were links between human activities and the natural ecosystems. As identified, *deforestation* and *overgrazing* imply the reduction of vegetative cover resulting in less protection for the soils from solar insolation and increased wind speeds, leading to increased evaporation and potential wind and water erosion, along with possible increases in groundwater levels and soil waterlogging from reduced vegetation soil moisture requirements. *Agricultural* activities include the plowing up of fragile soils, thus baring them for wind erosion; attempts to grow crops in areas with insufficient soil moisture leading to crop failure and soil exposure to further erosion; and the excessive application of irrigation water leading to the buildup of salts in the soil (salinization).

Overexploitation implies excessive use of vegetation for fuel or building materials that reduces the capacity of the vegetation to reproduce or recycle essential nutrients to maintain soil fertility. *Bioindustrial* impacts imply contamination from pollution sources and are usually associated with intensive land uses outside the drylands, hence the relatively small area shown.

There is no doubt that climate variability, particularly in terms of precipitation trends over decades, has had measurable impacts upon the success of human occupation of the drylands of the globe. Some of the most telling evidence comes from the Sahel (Kates, 1981; Hare, 1983; Nicholson, 1994). In such locations seasonal climates resemble desert climates (low precipitation, high evaporation, and high solar insolation) so that any human resource management involving reduction of protective vegetation cover at such time would run the risk of permanently damaging the environment (Aubreville, 1949; Bullock and Le Houerou, 1996).

Such damage might be interpreted as the result of longer-term climate change, such as a trend toward increasing aridity, and labeled desertification. However, dryland ecosystems demonstrate considerable ability to recover from periods of desiccation so that whether or not desertification is identified could depend upon the time period chosen for the study, as suggested earlier. Indeed, the edge of the Sahara Desert defined in terms of vegetation cover has been identified from satellite imagery to have shifted 240 km south between 1980 and 1984, but after several annual retreats and advances was by 1990 only 130 km south of the 1980 location (Tucker et al., 1991). Such variation suggests the importance of short-term fluctuations rather than long-term climatic changes.

Certainly, the clearance of vegetation for cropping or through excessive livestock grazing may change the albedo (reflectivity) of the ground surface and, thus, increase the air temperatures, thereby increasing evaporation and creating droughtlike conditions that encourage soil erosion as a result of increased wind speeds. Significantly, most desert reclamation techniques involve attempts to establish new vegetation cover to directly protect the soil and to act in part as wind breaks (Baker, 1966; Zhu and Liu, 1981).

Table 4, however, omits mention of some of the less obvious but possibly equally important factors, such as the history of colonialism and recent economic development both of which have been identified as relevant to the spread of desertification. Most colonial powers in Africa, for example, introduced new tax systems that forced a monetary economy upon a previously subsistence pastoral or agricultural commu-

TABLE 4 Causes of Desertification

Attributed Cause	Area Affected (million hectares)	Percentage of Total Desertified
Deforestation	578.6	29.4
Overgrazing	678.7	34.5
Agricultural	551.6	28.1
Overexploitation	132.8	6.8
Bioindustrial	22.7	1.2
Total	1964.4 m.ha.	100.0%

Source: UNEP (1992, p. 25).

nity, requiring extra numbers of livestock to be grazed for cash sale to pay taxes, or food crops on the better soils to be replaced by cash crops and necessary food production to be pushed into areas climatically more vulnerable or possessing poorer soils more prone to erosion (Franke and Chasin, 1980; Garcia, 1981; Watts, 1983; Baker, 1984; Macdonald, 1986; Morgan and Solarz, 1994). More recently, improved veterinary services and permanent water supplies developed with foreign aid, which replaced traditionally limited seasonal supplies, allowed larger herds to be carried all year, and this often led to overgrazing of the ranges (Glantz, 1977).

In addition, the increase of human populations in the drylands threatened by desertification, from 57 millions in 1977 to 135 millions by 1984 (UNEP, 1992, p. iv), and particularly the rising populations in Africa (Caldwell and Caldwell, 1990) have brought increased pressure on the environment and, in association with the periodic devastation from civil strife and warfare, have no doubt contributed to the desertification process (Glantz, 1987).

5 FUTURE OF DESERTIFICATION

Despite over 20 years of international efforts, the complexity of the factors involved in desertification has meant that those efforts to reverse the trends have had limited success. Reviews of the results of the 1977 UNCOD initiative were not particularly impressive (Mabbutt, 1987; Odingo, 1992; Rapp, 1987; Spooner, 1987). And despite ongoing research and publications by the United Nations Environment Programme through its *Desertification Control Bulletin*, which documents attempts to halt desertification, the debate about definitions cannot hide the fact that the phenomenon continues to be extensive and locally increasing in the area it affects.

A new UN "Convention to Combat Desertification in those Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa," was agreed to in 1994 and came into force in December 1996. The new emphasis is on support for local schemes to rehabilitate desertified areas or to reduce the potential for their expansion. This shift in scale holds better promise for the future since the specific causes of desertification are more often related to local economic, social, and political events in the context of the seasonal weather rather than to broad changes in climatic patterns.

Having said that, however, global warming climate scenarios (in essence forecasts) could result in increased aridity in the drylands, with associated increases in natural soil erosion, even without human interference (Bullock and Le Houerou, 1996). Such a scenario does not offer much hope for reversing desertification in the drylands in the future, unless the pressure imposed by human resource uses can, itself, be reduced.

REFERENCES

Aubreville, A., *Climats, Forêts et Dèsertification de l'Afrique Tropicale*, Societès d'Èditions Geographiques, Maritimes et Coloniales, Paris, 1949.

Baker, R., *Sahara Conquest*, Butterworth, London, 1966.

Baker, R., Protecting the environment against the poor: The historical roots of the soil erosion orthodoxy in the Third World, *Ecologist*, *14*, 53–60, 1984.

Beinert, W., Environmental destruction in Southern Africa, in T. S. Driver, and G. P. Chapman (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996, pp. 256–268.

Brown, R. H., *Historical Geography of the United States*, Harcourt Brace, New York, 1948.

Bryson, R. A., and R. J. Murray, *Climates of Hunger*, University of Wisconsin, Madison, 1977.

Bullock, P., and H. Le Houerou, Land degradation and desertification, in R. T. Watson, M. C. Zinyowera, and R. H. Moss (Eds.), *Climate Change 1995, Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses*, Cambridge University Press, Cambridge, 1996, pp. 177–189.

Calder, R., *Men Against the Desert*, Allen and Unwin, London, 1951.

Caldwell, J. C., and P. Caldwell, High fertility in sub-Saharan Africa, *Sci. Am.*, *262*(5); 82–88, 1990.

Chapman, G. P., and T. S. Driver, Time, mankind, and the Earth, in T. S. Driver and G. P. Chapman (Eds.), *Time-scales and Environmental Change*, Routledge, London, 1996, pp. 1–24.

Driver, T. S., and G. P. Chapman, (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996.

Dregne, H., M. Kassas, and B. Rosanov, A new assessment of the world status of desertification, *Desertif. Contr. Bull.*, *20*, 6–18, 1991.

Fairhead, J., and M. Leach, Reframing forest history, in T. S. Driver, and G. P. Chapman (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996, pp. 169–195.

Franke, R. W., and B. H. Chasin, *Seeds of Famine: Ecological Destruction and the Development Dilemma in the West African Sahel*, Landmark, Montclair, NJ, 1980.

Garcia, R. V., *Drought and Man: The 1972 Case History*, Vol. 1: *Nature Pleads Not Guilty*, Pergamon, New York, 1981.

Glantz, M. H. (Ed.), *Desertification: Environmental Degradation in and around Arid Lands*, Westview, Boulder, CO, 1977.

Glantz, M. H. (Ed.), *Drought and Hunger in Africa: Denying Famine a Future*, Cambridge University Press, London, 1987.

Glantz, M. H., and N. S. Orlovsky, Desertification: Anatomy of a complex environmental process, in K. A. Dahlberg, and J. W. Bennett (Eds.), *Natural Resources and People: Conceptual Issues in Interdisciplinary Research*, Westview, Boulder, CO, 1986, pp. 213–229.

Hansen, T., Arabia Felix, in *The Danish Expedition of 1761–1767*, J. McFarlane, and K. McFarlane (Trans.), Readers Union, Collins, London, 1965.

Hare, K., *Climate and Desertification: A Revised Analysis*, World Meteorogical Organization (WMO) World Climate Program Publication 44, WMO Geneva, 1983.

Hare, F. K., R. W. Kates, and A. Warren, The making of deserts: Climate, ecology and society, *Econ. Geogr.*, *53*(4), 332–346, 1977.

Heathcote, R. L. (Ed.), *Perception of Desertification*, United Nations University, Tokyo, 1980.

Heathcote, R. L., *Back of Bourke: A Study of Land Appraisal and Settlement in Semi-Arid Australia*, Melbourne University Press, London, 1965.

Heathcote, R. L., Climate and famine: Differing interpretations of the linkages, *Austral. Overseas Disaster Response Org. Newslett.*, *4*(4), 6–8, 1986.

Huntington, E., *The Pulse of Asia*, Yale University Press, New Haven, CT, 1907.

Jacks, G. V., and R. O. Whyte, *The Rape of the Earth: A World Survey of Soil Erosion*, Faber and Faber, London, 1938.

Kates, R. W., Drought in the Sahel: Competing views on what really happened in 1910–14 and 1968–74, *Mazingira*, *5*(2), 72–80, 1981.

Kokot, D. F., Desert encroachment in South Africa, *Afr. Soils*, *3*(3), 404–409, 1955.

Lowdermilk, W. C., *Conquest Land Through 7,000 Years*, United States Department of Agriculture Soil Conservation Service Agricultural Information Bulletin No. 99, U.S. Government Printing Office, Washington, DC, 1953.

Mabbutt, J. A., A review of progress since the UN conference on desertification, *Desertif. Control Bull.*, *15*, 12–23, 1987.

Macdonald, L. H., *Natural Resources Development in the Sahel: The Role of the United Nations System*, United Nations University, Tokyo, 1986.

Morgan, W. B., and J. A. Solarz, Agricultural crisis in Sub-Saharan Africa: Development constraints and policy problems, *Geogr. J.*, *160*(1), 57–73, 1994.

Nash, R., *The Rights of Nature: A History of Environmental Ethics*, University of Wisconsin Press, Madison, 1990.

Nicholson, S. E., Variability of African rainfall on interannual and decadal time scales, in D. Martinson, K. Bryan, M. Ghil, T. Karl, E. Sarachik, S. Sorooshian, and L. Talley, (Eds)., *Natural Climatic Variability on Decade-to-Century Time Scales*, National Academy of Sciences, Washington, DC, 1994.

Odingo, R. S., Implementation of the Plan of Action to Combat Desertification (PACD) 1978–1991, *Desertif. Control Bull. 21*, 6–14, 1992.

Passmore, J. Man's responsibility for nature, in *Ecological Problems and Western Traditions*, 2nd ed., Duckworth, London, 1980.

Pick, J. H., and V. R. Alldis, *Australia's Dying Heart: Soil Erosion and Station Management in the Inland*, 2nd ed., Melbourne University Press, Melbourne, 1944.

Rapp, A. Reflections on desertification 1977–1987: Problems and prospects, *Desertif. Control Bull.*, *15*, 27–33, 1987.

Ratcliffe, F. N., *Flying Fox and Drifting Sand*, 2nd ed., Angus and Robertson, Sydney, 1938.

Reifenberg, A., *The Struggle between the Desert and the Sown*, Jewish Agency, Jerusalem, 1955.

Rhodes, S. L., Rethinking desertification: What do we know and what have we learned? *World Devel.*, *19*(9), 1137–1143, 1991.

Sears, P. B., *Deserts on the March*, Routledge and Kegan Paul, London, 1949.

Sinclair, A. R. E., and J. M. Fryxell, The Sahel of Africa: Ecology of a disaster, *Can. Zool. J.*, *63*, 987–994, 1985.

Spooner, B., The (apparent) paradoxes of desertification, *Desertif. Control Bull.*, *15*, 40–45, 1987.

Stebbing, E. P., The encroaching Sahara: The threat to the West African colonies, *Geogr. J.*, *85*, 506–524, 1935.

Stebbing E.P., The forests of West Africa and the Sahara: A Study of modern conditions, *Geogr. J.*, *90*, 550–552, 1937.

Thomas, D. S. G., Sandstorm in a teacup? Understanding desertification, *Geogr. J.*, *159*(3), 318–331, 1993.

Tolba, M. K., A harvest of dust? *Environ. Conserv.*, *11*(2), 1–2, 1984.

Tucker, C. J., H. E. Dregne, and W. W. Newcomb, Expansion and contraction of the Saharan Desert from 1980 to 1990, *Science*, *253*, 299–301, 1991.

United Nations Conference on Desertification (UNCOD), *World Map of Desertification at a Scale of 1:25000000,* FAO and UNESCO, New York, 1977.

UNCOD, *Convention to Combat Desertification in Those Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa*, Secretariat of the UN Convention to Combat Desertification, United Nations, New York, 1994.

United Nations Conference on Desertification (UNCOD), *Round-Up, Plan of Action and Resolutions*, United Nations, New York, 1978.

United Nations Environmental Programme (UNEP), *World Atlas of Desertification*, Edward Arnold, London, 1992.

United Nations Economic Scientific and Cultural Organisation (UNESCO), *Arid Zone Research*, Series No. 1 to No. 29, UNESCO, Paris, France, 1968.

Watts, M., *Silent Violence*, University of California Press, Berkeley, CA, 1983.

White, L., The historical roots of our ecologic crisis, *Science*, *155*(3767), 1203–1207, 1967.

Zhu, Z., and S. Liu, Desertification and desertification control in northern China, *Desertif. Control Bull.*, *5*, 13–19, 1981.

CHAPTER 54

IMAGINABLE SURPRISE

STEPHEN H. SCHNEIDER

1 INTRODUCTION

Although unexpected events are often considered "surprises," it is often the case that many events are anticipated by at least some observers. Thus, an unexpected event may be labeled more appropriately as an "imaginable surprise," which is defined as an event or process that departs from the expectations of some definable community. Imaginable surprise is a concept related to, but distinct from, risk and uncertainty. Risk is typically defined as the condition in which the event and the probability that it will occur is known. However, risk almost always is accompanied by a certain degree of uncertainty. Uncertainty remains a difficult concept to define or codify but is usually defined as the condition in which the event, process or outcome is known, but the probability that it will occur is not known. Two basic options are appropriate in the face of uncertainty: (1) reduce the uncertainties through data collection, research, modeling, simulation techniques, etc. and (2) manage or integrate uncertainty directly into the decision-making or policy-making process. When uncertainties are large, a strategic approach that considers a wide range of possible outcomes, including low-probability events, may be a more appropriate way to manage uncertainty. It may be possible to identify "imaginable conditions for surprise" when the conditions that might induce surprises are known even though the actual surprise events are not.

Decision makers at all levels (individuals, firms, and local, national, and international governmental organizations) are concerned about reducing their vulnerability to (or the likelihood of) unexpected events or surprises. After briefly and selectively reviewing the literature on uncertainty and surprise, I introduce a definition of surprise that does not include the strict requirement that it apply to a wholly unexpected outcome, but rather recognizes that many events are often anticipated by

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts,
Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

some, even if not most observers. Thus, an imaginable surprise is defined as an event or process that departs from the expectations of some definable community, yet is a concept related to, but distinct from, risk and uncertainty. Therefore, what gets labeled as "surprise" depends on the extent to which what happens departs from community expectations and on the salience of the problem. Impediments to overcoming ignorance range from the need for more "normal science" to phenomenological impediments (e.g., inherent unpredictability in some chaotic systems) to epistemological ignorance (e.g., ideological blocks to reducing ignorance). The substantive focus in this chapter will concentrate on the theme of global change. Examples of imaginable surprises in the global change context are presented, as well as their potential salience for creating unexpectedly high or low carbon dioxide emissions. Improving the anticipation of surprises is an interdisciplinary enterprise that should offer a skeptical welcoming of outlier ideas and methods.

Strictly speaking, a surprise cannot be anticipated; by definition it is an unexpected event. Potential climate change, and more broadly global environmental change, is replete with the truly unexpected because of the enormous complexities of processes and interrelationships involved and our insufficient understanding of them (such as coupling ocean, atmosphere, and terrestrial systems). However, risk, hazard, and related research demonstrate repeatedly that the event, process, or outcome registered as a surprise by the community in question was frequently known or forecast by others or the same event was knowable within the competing frameworks of understanding (Darmstadter and Toman, 1993).

2 UNCERTAINTY

Much of the current work on surprise has grown out of an extensive body of research on uncertainty. Yet, although widely acknowledged and studied, uncertainty remains a difficult concept to define or codify. Different conceptualizations and approaches to uncertainty abound in the literature, cross numerous fields of study, and touch a wide range of problem types. Two basic options are invariably followed in the face of uncertainty. The first is to reduce the uncertainties through data collection, research, modeling, simulation techniques, and so forth. Following this option, the objective is to overcome uncertainty, to make the unknown known. But the daunting nature of uncertainties surrounding global environmental change, as well as the need to make decisions before the normal science option can provide resolution, force a second option—that of managing or integrating uncertainty directly into the decision-making or policy-making process. Before uncertainty can be so integrated, however, the nature and extent of the uncertainty must be clarified.

The fields of mathematics, statistics, and, more recently, physics provide the "science of uncertainty" with many powerful means and techniques to conceptualize, quantify, and manage uncertainty, ranging from the frequency distributions of probability theory to the possibility and belief statements of Bayesian statistics. Addressing other aspects of uncertainty, fuzzy set logic offers an alternative to classical set theory for situations where the definitions of set membership are vague, ambiguous, or nonexclusive. More recently, researchers have proposed chaos theory and complexification theory to focus on expecting the unexpected in models and theory (Casti, 1994).

The practical application of many of these techniques was originally pioneered by researchers in decision analysis (see Raiffa, 1968). In the fields of economics and decision theory, researchers continue to study rational decision making under uncertainty and how to assess the value of collecting additional information (Clemen, 1991). Methods for modeling risk attitudes, leading to the terms *risk-prone* and *risk-averse*, attempt to capture how different people faced with making a decision react to the uncertainty surrounding the expected outcomes. Uncertainty and its related context, surprise, are treated largely as the realization that events, currently unknown, will occur affecting the final outcome of a process or decision.

This acknowledgment of uncertainty has found a prominent place in many other fields of study, each one speaking its own language of uncertainty. Research on uncertainty cross-cuts a number of different disciplines. For example, researchers making risk assessments and setting safety standards find it most useful to distinguish between risk (the probability of a certain negative effect resulting from a hazard occurrence, given the specified level of exposure), variability (interindividual differences in vulnerability and susceptibility), and uncertainty (model parameter variability and any unexplained residual). In work related to hazards and risk, sociologists, anthropologists, psychologists, and geographers have made important contributions to the discussions on risk perception, risk communication, and the social amplification of risk (Kahneman et al., 1982; Kasperson et al., 1988; see Gigerenzer, 1996, for a criticism). Similarly, work on visualizing or graphically conveying uncertainty also crosses a diverse set of disciplines including psychology, computer science, and geographic information systems (GIS) (MacEachren, 1992).

Wynne (1992) emphasizes that the modeling of environmental risk systems requires examination of not only the scientific evidence and competing interpretations, but also investigation of the nature, assumptions, and inherent limitations of the scientific knowledge behind the data and the model. He specifies four types of uncertainty—risk, uncertainty, ignorance, and indeterminacy—each overlaying dimensions of uncertainty. *Risk* refers to a situation when the system behavior is well known and the chances of different outcomes can be quantified by probability distributions. If, however, the important system parameters are known but not the associated probabilities, then in Wynne's definitions, *uncertainty* exists. *Ignorance* is that which is not known (or even that we are aware that we do not know it) and, for Wynne, is endemic because scientific knowledge must set the bounds of uncertainty in order to function. *Indeterminacy* captures the unbounded complexity of causal chains and open networks. Uncertainty, in part, stems not only from an incomplete understanding of determinate relationships, but also from the interaction of these relationships with contingent and unpredictable actors and processes. However, the extent to which situations are truly "indeterminate," as opposed to simply containing a very broad distribution of subjective probability estimates, is not a straightforward classification, for often very ill understood phenomena can still be bounded to some extent by existing knowledge, and thus are not truly indeterminate.

3 OVERCOME OR JUST MANAGE UNCERTAINTY

In the areas of environmental policy and resource management, policymakers struggle with the need to make decisions utilizing vague and ambiguous concepts (such as sustainability), with sparse and imprecise information, in decisions that have far-reaching, and often irreversible, impacts on both environment and society. Not surprisingly, efforts to incorporate uncertainty into the decision-making process quickly move to the forefront with the advent of decision-making paradigms, such as the precautionary principle, adaptive environmental management, the preventative paradigm, or stewardship. Ravetz (1986) takes the concept of "usable knowledge in the context of incomplete science" one step further by introducing the idea of usable ignorance. To Ravetz, acknowledging the "ignorance factor" means becoming aware of the limits of our knowledge. Ravetz argues that ignorance cannot be overcome with any amount of sophisticated calculations. Rather, coping with ignorance demands a better articulation of the policy process and a greater awareness of how that process operates. He recognizes that one can only replace ignorance by gaining more knowledge, but stresses that by "being aware of our ignorance we do not encounter disastrous pitfalls in our supposedly secure knowledge or supposedly effective technique" (p. 429).

The emphasis on managing uncertainty rather than mastering it can be traced to work on resilience in ecology (Holling, 1986). Whereas resistance implies an ability to withstand change or impact within some measure of performance, resilience captures the ability to give with the forcing function, without disrupting the overall health of the system. In this framework, adaptation is an ecological mechanism whose aim is not to overcome or control environmental uncertainty but to live with and, in some cases, thrive upon it.

Risk is typically defined as the condition in which the event, process, or outcome, and the probability that each will occur, is known. In reality, of course, complete or perfect knowledge of complex systems, which would permit the credible calculation of objective or frequentist probabilities, rarely exists. Likewise, the full range of potential outcomes is usually not known. Thus, risk almost always is accompanied by varying degrees of uncertainty. Uncertainty is usually defined as the condition in which the event, process, or outcome is known (factually or hypothetically), but the probabilities that it will occur are not known or are highly subjective estimates (see, e.g., Moss and Schneider, 2000).

4 SURPRISE

Strictly speaking, surprise is the condition in which the event, process, or outcome is not known or expected. In this "strict" meaning, the attribution of surprise shifts toward the event, process, or outcome itself. We may expect surprises to occur, but we are surprised by the specific event, process, or outcome involved. This meaning, as noted, begs the issue of anticipation because the very act of anticipation implies some level of knowledge or foresight. However, it may be possible to identify "imaginable conditions for surprise" where the conditions that might induce surprises are known even though the actual surprise events are not—e.g., rapid forcing of nonlinear systems (Moss and Schneider, 2000).

Because of the impracticality of the strict definition of surprise for policy making, various studies advocate the use of another meaning for surprise, one in which the attribution of surprise shifts more toward the expectations of the observer. Holling (1986; p. 294) recognized this meaning of surprise as a condition in which perceived reality departs qualitatively from expectations. It is this more interpretive or relational meaning of surprise—which has been labeled imaginable surprise—that portends to be most useful for global change studies [e.g., see Schneider et al. (1998) from which much of this material has been adopted].

Almost every event may constitute an imaginable surprise to someone. But since global change phenomena and their environmental and societal impacts are a community-scale set of issues, little can be gained for our purposes by focusing on whether someone, somewhere, may or may not have once predicted or hinted at some surprise event. More fruitful is the recognition that groups, communities, and cultures may share expectations such that a particular event is likely to qualify as a surprise for most within them. In these cases, what gets labeled as a surprise depends upon the extent to which reality departs from community expectations, and on the salience of the problems imposed.

Imaginable surprise applies to communities of experts, policymakers, managers, and educators who share common ranges of expectation that are generated by group dynamics, leaders, and signal processors, including the dominant educational and research paradigms (Kasperson et al., 1988). For these communities, shared expectations follow from dominant interpretations among the expert community (e.g., global warming is likely), from their fit with broader policy agendas (e.g., environmentally benign economic development is possible), and from vested interest, conscious or unconscious, of an agency or group to maintain a particular view (e.g., global population growth is environmentally damaging, or, alternatively, good for the economy). Since policy making often reflects a blend of public and interest group perceptions of reality, the imaginable surprise formulation is much more relevant to global change policy issues than a strict definition of surprise as an unimaginable outcome.

5 APPLICATION TO GLOBAL CHANGE

Since natural and social global change science remains in a range of developmental stages, the unknowns are sufficiently large to warrant attention to divergent themes about similar processes and outcomes. To facilitate this range of research, (a) measures should be taken to ensure a more open discourse and evaluation of alternatives, such as by a more open airing and professional evaluation as opposed to uncritical, "equal time," and equal credibility often afforded to polarized viewpoints in the popular media of less dominant or unconventional views, including those by advocacy science and scientists; and (b) by reducing the redundancy of research focused on the dominant views and theses while still preserving a diversity of approaches within dominant paradigms—i.e., create research "overlap without cloning" (Schneider et al., 1998).

The assumptions associated with the standard paradigm of global climate change impact assessment, for example, although recognizing the wide range of uncertainty, are essentially surprise free. One approach is to postulate low, or uncertain, probability cases in which little climate change, on the one hand, or catastrophic surprises, on the other hand, might occur and multiply the lower probability times the much larger potential costs or benefits. Analysts, however, customarily use a few standard general circulation model CO_2-doubling scenarios to "bracket the uncertainty" rather than to postulate extremely serious or relatively negligible climatic change outcomes (e.g., see Schneider, 2001). A strategic approach, that is, one that considers a wide range of probabilities and outcomes, may be more appropriate for global climate change impact assessments given the high plausibility of surprises, even if we have but limited capacity to anticipate specific details right now (Moss and Schneider, 2000).

An assumption in cost–benefit calculations within the standard assessment paradigm is that "nature" is either constant or irrelevant. For example, ecological services such as pest control or waste recycling are assumed as constants or of no economic value in most assessment calculations. Yet should climatic change occur in the middle to upper range of that typically projected, it is highly likely that communities of species will be disassembled, and the probability of significant alterations to existing patterns of pests and weeds seem virtually certain (Root, 2000). Some argue that pests, should their patterns be altered, can simply be controlled by pesticides and herbicides. The side effects of many such controls are well known. What is not considered in the standard paradigm is the consideration of a "surprise" scenario such as a change in public consciousness regarding the value of nature that would reject pesticide or herbicide application as a "tech-fix" response to global changes.

Finally, global change portends alterations to the basic processes that govern the state of the biosphere. Global change research, therefore, might do well to anticipate these alterations, an effort that will require more than the study of extant processes and conditions alone. Various modes of analysis and approaches appropriate for such explorations, but typically underutilized in the research community, should be encouraged. Among these are (a) backcasting scenarios from posited future states and/or reconstructing past scenarios in alternative ways to identify events or processes that might happen (recognizing, of course, that diffusion processes usually are not reversible and diffusion-dominated systems cannot be uniquely backcast); (b) increasing attention to and support for the study of "outlier" outcomes, searching for the reasons they appear deviant and the lessons that might be drawn from them (Hassol and Katzenberger, 1997); and, (c) exploring the "resilience" paradigm (e.g., precautionary principle) alongside the "optimization" paradigm (e.g., aggregated cost–benefit analyses) to inform policy making and diagnose alternative outcomes and risk management strategies. Other means of improving the anticipation of surprise in global change science would emerge from convening additional expert groups and asking them for more exhaustive assessments of the issues. Balanced assessments will likely lead to recommendations that "research as usual" be tempered with more alternative or even unusual research.

In summary, global change science and policy making will have to deal with uncertainty and surprise for the foreseeable future. Thus, more systematic analysis of surprise issues and more formal and consistent methods of incorporation of uncertainty into global change assessments will become increasingly necessary. Improvements in dealing with scientific surprise in climate change in particular and global change in general, therefore, require the research and funding communities to seek a better balance among traditional and experimental research alternatives (see also Kates and Clark, 1996, p. 31). This aim, in turn, requires strategies that will facilitate this balance, including the difficult problem of assessing "quality" in an interdisciplinary context.

ACKNOWLEDGMENTS

Modified from S. H. Schneider, B. L. Turner, and H. Morehouse Garriga (1998), "Imaginable Surprise in Global Change Science," *Journal of Risk Research* 1(2): 165–185. I thank Kristin Kuntz-Duriseti for editorial help and Billie Lee Turner for the co-chairing the effort that produced much of the perspective here.

REFERENCES

Casti, J. L., *Complexification: Explaining a Paradoxical World through the Science of Surprise*, Harper Collins, New York, 1994.

Clemen, R. T., *Making Hard Decisions: An Introduction to Decision Analysis*, PWS-Kent, Boston, 1991.

Darmstadter, J., and M. A. Toman (Eds.), Nonlinearities and surprises in climate change: An introduction and overview, in *Assessing Surprise and Nonlinearities in Greenhouse Warming. Proceedings of an Interdisciplinary Workshop*, Resources for the Future, Washington, DC, 1993, pp. 1–10.

Gigerenzer, G., On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky, *Psychol. Rev.*, *103*, 592–596, 1996.

Hassol, S. J., and J. Katzenberger (Eds.), *Elements of Change 1996*, Aspen Global Change Institute, Aspen, CO, 1997.

Holling, C. S., The resilience of terrestrial ecosystems: Local surprise and global change, in W. C. Clark, and R. E. Munn (Eds.), *Sustainable Development of the Biosphere*, Cambridge University Press for the International Institute for Applied Systems Analysis, Cambridge, 1986, pp. 292–317.

Kahneman, D., P. Slovic, and A. Tversky, *Judgment under Uncertainty*, Cambridge University Press, New York, 1982.

Kasperson, R. E., O. Renn, P. Slovic, H. Brown, J. Emel, R. Goble, J. X. Kasperson, and S. J. Ratick, The social amplification of risk: A conceptual framework, *Risk Anal. 8*, 177–187, 1988.

Kates, R. W., and W. C. Clark, Environmental surprise: Expecting the unexpected? *Environment, 38*, 6–11, 28–34, 1996.

MacEachren, A., Visualizing uncertain information, *Cartogr. Perspect.*, (13), 10–19, 1992.

Moss, R. H., and S. H. Schneider, Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting, in R. Pachauri, T. Taniguchi, and K. Tanaka (Eds.), *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, Intergovernmental Panel on Climate Change, Geneva, 2000; available on-line, http://www.gispri.or.jp.

Raiffa, H., *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison-Wesley, Reading, MA, 1968.

Ravetz, J. R., Usable knowledge, usable ignorance: Incomplete science with policy implications, in W. C. Clark and R. E. Munn (Eds.), *Sustainable Development of the Biosphere*, Cambridge University Press, New York, 1986, pp. 415–432.

Root, T. L., Ecology: Possible consequences of rapid global change, in G. Ernst (Ed.), *Earth Systems: Processes and Issues*, Cambridge University Press, Cambridge, MA, 2000, pp. 315–324.

Schneider, S. H., What constitutes "dangerous" climate change? *Nature*, *411*, 17–19, 2001.

Schneider, S. H., B. L. Turner II, and H. Morehouse Garriga, Imaginable surprise in global change science, *J. Risk Res.*, *1*(2), 1998, 165–185.

Wynne, B., Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm, *Global Environ. Change*, *20*, 111–127, 1992.

HANDBOOK OF WEATHER, CLIMATE, AND WATER

DYNAMICS, CLIMATE, PHYSICAL METEOROLOGY, WEATHER SYSTEMS, AND MEASUREMENTS

THOMAS D. POTTER
BRADLEY R. COLMAN

HANDBOOK OF WEATHER, CLIMATE, AND WATER

HANDBOOK OF WEATHER, CLIMATE, AND WATER

Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements

Edited by

THOMAS D. POTTER

BRADLEY R. COLMAN

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

The handbook of weather, climate, and water : dynamics, climate
physical meteorology, weather systems, and measurements / co-chief
editors, Thomas D. Potter and Bradley R. Colman.
p. cm.
Includes bibliographical references.
ISBN 0-471-21490-6 (acid-free paper)
1. Dynamic meteorology. 2. Atmospheric physics. 3. Weather.
I. Potter, Thomas D., 1955- II. Colman, Bradley Roy.
QC880.H34 2003
551.51′5--dc21
2003003584

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

The *Handbook of Weather*, *Climate*, *and Water* provides an authoritative report at the start of the 21st Century on the state of scientific knowledge in these exciting and important earth sciences. Weather, climate, and water affect every person on earth every day in some way. These effects range from disasters like killer storms and floods, to large economic effects on energy or agriculture, to health effects such as asthma or heat stress, to daily weather changes that affect air travel, construction, fishing fleets, farmers, and mothers selecting the clothes their children will wear that day, to countless other subjects.

During the past two decades a series of environmental events involving weather, climate, and water around the globe have been highly publicized in the press: the Ozone Hole, Acid Rain, Global Climate Change, El Niños, major floods in Bangladesh, droughts in the Sahara, and severe storms such as hurricane Andrew in Florida and the F5 tornado in Oklahoma. These events have generated much public interest and controversy regarding the appropriate public policies to deal with them. Such decisions depend critically upon scientific knowledge in the fields of weather, climate, and water.

One of two major purposes of the Handbook is to provide an up-to-date accounting of the sciences that underlie these important societal issues, so that both citizens and decision makers can understand the scientific foundation critical to the process of making informed decisions. To achieve this goal, we commissioned overview chapters on the eight major topics that comprise the Handbook: Dynamics, Climate, Physical Meteorology, Weather Systems, Measurements, Atmospheric Chemistry, Hydrology, and Societal Impacts. These overview chapters present, in terms understandable to everyone, the basic scientific information needed to appreciate the major environmental issues listed above.

The second major purpose of the Handbook is to provide a comprehensive reference volume for scientists who are specialists in the atmospheric and hydrologic areas. In addition, scientists from closely related disciplines and others who wish to get an authoritative scientific accounting of these fields should find this work to be of great value. The 95 professional-level chapters are the first comprehensive and integrated survey of these sciences in over 50 years, the last being completed in 1951 when the American Meteorological Society published the Compendium of Meteorology.

The *Handbook of Weather, Climate, and Water* is organized into two volumes containing eight major sections that encompass the fundamentals and critical topic areas across the atmospheric and hydrologic sciences. This volume contains sections on the highly important topics of Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements. Dynamics describes the nature and causes of atmospheric motions mostly through the laws of classical Newtonian physics supplemented by chaos theory. Climate consists of the atmosphere, oceans, cryosphere and land surfaces interacting through physical, chemical, and biological processes. Physical Meteorology presents the laws and processes governing physical changes in the atmosphere through chapters on atmospheric thermodynamics, atmospheric radiation, cloud physics, atmospheric electricity and optics, and other physical topics. The section on Weather Systems describes remarkable advances in weather forecasting gained by an increased understanding of weather systems at both large and small scales of motion and at all latitudes. The section on Measurements describes the many advances in the sensing of atmospheric conditions through constantly improving instrumentation and data processing.

To better protect against weather, climate, and water hazards, as well as to promote the positive benefits of utilizing more accurate information about these natural events, society needs improved predictions of them. To achieve this, scientists must have a better understanding of the entire atmospheric and hydrologic system. Major advances have been made during the past 50 years to better understand the complex sciences involved. These scientific advances, together with vastly improved technologies such as Doppler radar, new satellite capabilities, numerical methods, and computing, have resulted in greatly improved prediction capabilities over the past decade. Major storms are rarely missed nowadays because of the capability of numerical weather-prediction models to more effectively use the data from satellites, radars, and surface observations, and weather forecasters improved understanding of threatening weather systems. Improvements in predictions are ongoing. The public can now rely on the accuracy of forecasts out to about five days, when only a decade or so ago forecasts were accurate to only a day or two. Similarly, large advances have been made in understanding the climate system during the past 20 years. Climate forecasts out beyond a year are now made routinely and users in many fields find economic advantages in these climate outlooks even with the current marginal accuracies, which no doubt will improve as advances in our understanding of the climate system occur in future years.

Tom Potter
Brad Colman

Color images from this volume are available at ftp://ftp.wiley.com/public/sci_tech_med/weather/.

DEDICATION AND ACKNOWLEDGMENTS

Many people have assisted in the production of this Handbook—the Contributing Editors, the Authors, our editors at Wiley, friends too numerous to mention, and our families who supported us during the long process of completing this work. Professor Peter Shaffer, University of Washington, is owed deep appreciation for his untiring generosity in sharing his experience and talent to solve many problems associated with this large project. They all deserve much credit for their contributions and we want to express our deep thanks to all of them.

Finally, we want to dedicate this work to Tom Lockhart, the Contributing Editor of the Measurements part of the Handbook. Tom passed away in early 2001 and we regret that he will not be able to see the results of his efforts and those of his colleagues in final form.

Tom Potter
Brad Colman

CONTRIBUTORS

STEVEN A. ACKERMANN, NOAA-CIRES Climate Diagnostic Center, 1225 West Dayton Street, Madison, WI 53706-1695

AMANDA ADAMS, University of Wisconsin-Madison, Department of Atmospheric and Ocean Sciences, 1225 West Dayton Street, Madison, WI, 53706-1695

W. D. BACH, Jr., Army Research Program, AMXRL-RO-EN, P.O. Box 12211, Research Triangle Park, NC 27709-2211

R. A. BAXTER, 26106 Alejandro Road, Valencia, CA 91355

CRAIG F. BOHREN, Pennsylvania State University, Department of Meteorology, College of Earth and Mineral Sciences, 503 Walker Building, University Park, PA 16802-5013

GORDON BONAN, NCAR, Box 3000, Boulder, CO 80307-3000

FRED V. BROCK, Oklahoma Climate Survey, 100 East Zboyd Street, Room 1210, Norman, OK 73019

H. BROOKS, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

JOHN R. CHRISTY, University of Alabama in Huntsville, Earth Sciences Center, Huntsville, AL 35899

ROBERT E. DICKINSON, Georgia Tech, School of Earth and Atmospheric Sciences, Atlanta, GA 30332-0340

Nolan J. Doesken, Colorado State University, Department of Atmospheric Sciences, Colorado Climate Center, Fort Collins, CO 80523

C. Doswell III, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

D. Dowell, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

Lawrence B. Dunn, National Weather Service, 2242 West North Temple, Salt Lake City, UT 84116-2919

David R. Easterling, NCDC, 151 Patton Avenue, Room 120, Asheville, NC 28801-5001

Paul M. Fransoli, 1112 Pagosa Way, Las Vegas, NV 8912

Grant W. Goodge, P.O. Box 2178, Fairview, NC 28730

John Gyakum, McGill University, Department of Atmospheric and Oceanic Sciences, 805 Sherbrooke Street West, Montreal, Quebec H3A 2K6

W. H. Haggard, 150 Shope Creek Road, Asheville, NC 28805

John Hallett, Desert Research Center, Atmospheric Sciences Center, 2215 Raggio Parkway, Reno, NV 89512-1095

Jackson R. Herring, NCAR, Box 3000, Boulder, CO 80307-3000

Alan L. Hinckley, Campbell Scientific Inc., 815 West 1800 North, Logan, UT 84321-1784

R. Holle, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

John Horel, University of Utah, Meteorology Department, 135 South 1460 East, Salt Lake City, UT 84112-0110

B. Johns, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

D. Jorgenson, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

Eugenia Kalnay, University of Maryland, Department of Meteorology, 3433 Computer and Space Sciences Building, College Park, MD 20742-2425

Thomas R. Karl, NCDC, 151 Patton Avenue, Room 120, Asheville, NC 28801-5001

Kristina Katsaros, NOAA/AOML, 4301 Rickenbacker Causeway, Miami, FL 33149

Paul Kushner, GFDL, Princeton Forrestal Route, P.O. Box 308, Princeton, NJ 08542-0308

Thomas J. Lockhart, deceased

Walter A. Lyons, FMA Research Inc., 46050 Weld County Road 13, Fort Collins, CO 80524

FRANK D. MARKS, Jr., NOAA Hurricane Research Division, 43001 Rickenbacker CSWY, Miami, FL 33149

THOMAS B. MCKEE, Colorado State University, CIRA, Foothills Campus, Fort Collins, CO 80523-1375

GERALD MEEHL, NCAR, Box 3000, Boulder, CO 80307-3000

JOHN MITCHELL, Meteorology Office, Hadley Center, London Road, Bracknell, Berkshire RG12 2SY

JOHN W. NIELSEN-GAMMON, Texas A&M University, Department of Atmospheric Sciences, 3150 TAMU, College Station, TX 77843-3150

HAROLD D. ORVILLE, South Dakota School of Mines and Technology, 501 East St. Joseph Street, Rapid City, SD 57701-3995

ROBERT PINCUS, NOAA-CIRES Climate Diagnostic Center, 1225 West Dayton Street, Madison, WI 53706-1695

ARTHUR L. RANGNO, Department of Atmospheric Sciences, University of Washington, Box 351640, Seattle, WA 98195

ROBERT M. RAUBER, University of Illinois at Urbana-Champaign, Department of Atmospheric Sciences, 105 South Gregory Street, Urbana, IL 61801-3070

SCOTT J. RICHARDSON, Oklahoma Climate Survey, 100 East Zboyd Street, Room 1210, Norman, OK 73019

MURRY SALBY, University of Colorado, Box 311, Boulder, CO 80309-0311

EDWARD S. SARACHIK, University of Washington, JISAO, 4909 25th Avenue NW, Box 354235, Seattle, WA 98195-4235

JOSEPH W. SCHIESL, 9428 Winterberry Lane, Fairfax, VA 22032

D. SCHULTZ, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

D. STENSRUD, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

KYLE SWANSON, Geosciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201

ROBERT N. SWANSON, 1216 Babel Lane, Concord, CA 94518-1650

KEVIN TRENBERTH, NCAR, Climate Analysis Section, Box 3000, Boulder, CO 80307-3000

JOSEPH TRIBBIA, NCAR, P.O. Box 3000, Boulder, CO 80307-3000

GREG TRIPOLI, University of Wisconsin-Madison, Department of Atmospheric and Oceanic Sciences, 1225 W. Dayton Street, Madison, WI, 53706-1695

JEFFREY B. WEISS, NCAR, Box 3000, Boulder, CO 80307-3000

S. WEISS, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

L. WICKER, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

EARLE R. WILLIAMS, FMA Research Inc., 46050 Weld County Road 13, Fort Collins, CO 80524

JOSH WURMAN, 1945 Vassar Circle, Boulder, CO 80305

ROBERT YOUNG, RM Young Co., 2801 Aeropark Drive, Traverse City, MI 49686

D. ZARAS, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069

EDWARD J. ZIPSER, University of Utah, Department of Meteorology, 135 South 1460 East, Salt Lake City, UT 84112-0110

SECTION 1

DYNAMIC METEOROLOGY

Contributing Editor: Joseph Tribbia

CHAPTER 1

OVERVIEW—ATMOSPHERIC DYNAMICS

JOSEPH TRIBBIA

The scientific study of the dynamics of the atmosphere can broadly be defined as the attempt to elucidate the nature and causes of atmospheric motions through the laws of classical physics. The relevant principles are Newton's second law of motion applied to a fluid (the atmosphere), the equation of mass continuity, the ideal gas law, and the first law of thermodynamics. These principles are developed in detail in the contribution by Murry Salby. Since the empirical discovery and mathematical statement of these laws was not completed until the middle of the nineteenth century, as defined above, atmospheric dynamics was nonexistent before 1875. Nonetheless, attempts at applying dynamical reasoning using principles of dynamics appeared as early as 1735, in a work discussing the cause of the trade winds. Hadley's contribution and a complete history of theories of the atmospheric general circulation can be found in the monograph by Lorenz (1967).

The recognition that the laws enumerated above were sufficient to describe and even predict atmospheric motions is generally attributed to Vilhelm Bjerknes. He noted this fact in a study (1904) detailing both the statement of the central problem of meteorology (as seen by Bjerknes), weather prediction, and the system of equations necessary and sufficient to carry out the solution of the central problem. The chapter by Eugenia Kalnay describes the progress toward the solution of the central problem made since 1904 and the current state-of-the-art methods that marry the dynamical principles spelled out by Bjerknes and the computational technology brought to applicability by John von Neumann, who recognized in weather prediction a problem ideally suited for the electronic computer.

If Bjerknes' central problem and its solution were the sole goal of dynamical meteorology, then the chapters by Salby and Kalnay would be sufficient to describe

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

both the scientific content of the field and its progress to date. However, as noted above, atmospheric dynamics also includes the search for dynamical explanations of meteorological phenomena and a more satisfying explanation of why weather patterns exist as they do, rather than simply Force = (mass)(acceleration). The remaining chapters in the part demonstrate the expansion of thought required for this in three ways. The first method, exemplified by Paul Kushner's chapter, is to expand the quantities studied so that important aspects of atmospheric circulation systems may be more fully elucidated. The second method, exemplified by the chapters of Gerry Meehl and Kyle Swanson, develops dynamical depth by focusing on particular regions of Earth and the understanding that can be gained through the constraints imposed by Earth's geometry. The third method of expanding the reach of understanding in atmospheric dynamics is through the incorporation of techniques and ideas from other related scientific disciplines such as fluid turbulence and dynamical systems. These perspectives are brought to bear in the chapters of Jackson Herring and Jeffrey Weiss, respectively.

The focus of the chapter by Kushner is vorticity and potential vorticity. Anyone familiar with the nature of storms, e.g., both tropical and extratropical cyclones, will note the strong rotation commonly associated with these circulations. As Kushner shows, the local measure of rotation in a fluid can be quantified by either the vorticity or the related circulation. The recognition of the importance of circulation and vorticity in atmospheric systems can be traced at least as far back as von Helmholtz (1888). However, the most celebrated accomplishment in the first half of the twentieth century within atmospheric dynamics was the recognition by Carl G. Rossby (1939) that the most ubiquitous aspects of large-scale atmospheric circulations in middle latitudes could be succinctly described through a straightforward analysis of the equation governing vorticity. Rossby was also one of the first to see the value of the dynamical quantity, denoted by Ertel as potential vorticity, which, in the absence of heating and friction, is conserved by fluid parcels as they move through the atmosphere. The diagnostic analysis and tracking of this quantity forms the basis of many current studies in atmospheric dynamics, of both a theoretical and observational nature, and Kushner's chapter gives a concise introduction to these notions.

The chapters by Meehl and Swanson review the nature of motions in the tropics and extratropics, respectively. These geographic areas, distinguished from each other by their climatic regimes, have distinctive circulation systems and weather patterns that necessitate a separate treatment of the dynamics of each region. The dominant balance of forces in the tropics, as discussed by Meehl, is a thermodynamic balance between the net heating/cooling of the atmosphere by small-scale convection and radiation and the forced ascent/descent of air parcels that leads to adiabatic cooling/heating in response. This thermodynamic balance is sufficient to explain the mean circulations in the equatorial region, the north–south Hadley circulation and east–west Walker cell, the transient circulations associated with the El Niño–Southern Oscillation (ENSO) phenomenon, the monsoon circulations of Australia and Asia, and the intraseasonal Madden–Julian Oscillation. Meehl also explains the interactions among these circulations.

In contrast to the tropics, the main focus in the extra-tropics are the traveling cyclones and anticyclones, which are the dominant cause of the weather fluctuations seen at midlatitudes in all seasons except summer. These variations, which are symbolized on weather maps with the familiar high- and low pressure centers and delimiting warm and cold fronts, are dynamically dissected by Swanson and explained in terms of inherent instabilities of the stationary features that arise due to the uneven distribution of net radiative heating, topography, and land mass over Earth's surface. In the process of dynamically explaining these systems, Swanson makes use of the quasi-geostrophic equations, which are a simplification of the governing equations derived by Salby. This quasi-geostrophic system is a staple of dynamical meteorology and can be formally derived as an approximation of the full system using scale analysis (cf. Charney, 1948, or Phillips, 1963). The advantage of such reduced equations is twofold: the reduction frequently leads to analytically tractable equations as shown by Swanson's examples and, with fewer variables and degrees of freedom in the system, it is almost always easier to directly follow the causal dynamical mechanisms.

The chapters by Herring and Weiss bring in paradigms and tools from the physics of fluid turbulence and the mathematics of dynamical systems theory. The entire field of atmospheric dynamics is but a subtopic within the physics of fluid dynamics. The study of fluid motions in the atmosphere, ocean, and within the fluid earth is frequently referred to as geophysical fluid dynamics (GFD), so it is not surprising that ideas from fluid turbulence would be used in atmospheric dynamics, as well as in the rest of GFD. What is different in the application in the large-scale dynamics of the atmosphere is the notion of viewing the atmosphere as a turbulent (nearly) two-dimensional flow. The perspective given by Herring was conceived in the late 1960s by George Batchelor and Robert Kraichnan, and further developed by C. Leith, Douglas Lilly, and Herring. Prior to this time it was thought that two-dimensional turbulence was an oxymoron since turbulence studied in the laboratory and observed in nature is inherently three dimensional. As Herring shows, the two-dimensional turbulence picture of the atmosphere has enabled a dynamical explanation of the spectrum of atmospheric motions and elucidated the growth in time of forecast errors, which initiate in small scales and propagate up the spectrum to contaminate even planetary scales of motion. This notion of forecast errors contaminating the accuracy of forecasts was first investigated by Philip Thompson (1957) using the methodology of Batchelor's statistical theory of homogeneous turbulence. Herring's chapter is a summary of subsequent developments using this methodology.

A seminal study by Edward Lorenz (1963) is the predecessor of the review given by Weiss, detailing the use of a dynamical system's perspective and deterministic chaos in quantifying the predictability of the atmosphere. Lorenz' study was the starting point for two research fields: the application of dynamical systems theory to atmospheric predictions and the mathematical topic of deterministic chaos. Weiss' chapter summarizes the scientific developments relevant to atmospheric dynamics and climate and weather prediction since 1963.

In any brief summarization of an active and growing field of research as much, or more, will be left out as will be reviewed. The chapters presented in this part are to be viewed more as a sampler than an exhaustive treatise on the dynamics of atmospheric motions. For those intrigued by works presented here and wishing to further learn about the area, the following texts are recommended in addition to those texts and publications cited by the individual authors: *An Introduction to Dynamical Meteorology* (1992) by J. R. Holton, Academic Press; *Atmosphere-Ocean Dynamics* (1982) by A. E. Gill, Academic Press; and *Geophysical Fluid Dynamics* (1979) by J. Pedlosky, Springer.

REFERENCES

Bjerknes, V. (1904). Das problem von der wettervorhersage, betrachtet vom standpunkt der mechanik un der physik, *Meteor. Z.* **21**, 1–7.

Charney, J. G. (1948). On the Scale of Atmospheric Motions, *Geofys. Publik.* **17**, 1–17.

Hadley, G. (1735). Concerning the cause of the general trade-winds, *Phil. Trans.* **29**, 58–62.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow, *J. Atmos. Sci.* **20**, 130–141.

Lorenz, E. N. (1967). The Nature and Theory of the General Circulation of the Atmosphere, World Meteorological Organization/Geneva.

Phillips, N. A. (1963). Geostrophic motion, *Rev. Geophys.* **1**, 123–176.

Thompson, P. D. (1957). Uncertainty of the initial state as a factor in the predictability of large scale atmospheric flow patterns, *Tellus* **9**, 257–272.

von Helmholtz, H. (1888). On Atmospheric motions, *Sitz.-Ber. Akad. Wiss. Berlin* 647–663.

CHAPTER 2

FUNDAMENTAL FORCES AND GOVERNING EQUATIONS

MURRY SALBY

1 DESCRIPTION OF ATMOSPHERIC BEHAVIOR

The atmosphere is a fluid and therefore capable of redistributing mass and constituents into a variety of complex configurations. Like any fluid system, the atmosphere is governed by the laws of continuum mechanics. These can be derived from the laws of mechanics and thermodynamics governing a discrete fluid body by generalizing those laws to a continuum of such systems. The discrete system to which these laws apply is an infinitesimal fluid element or *air parcel*, defined by a fixed collection of matter.

Two frameworks are used to describe atmospheric behavior. The *Eulerian description* represents atmospheric behavior in terms of field properties, such as the instantaneous distributions of temperature and motion. It lends itself to numerical solution, forming the basis for most numerical models. The *Lagrangian description* represents atmospheric behavior in terms of the properties of individual air parcels, the positions of which must then be tracked. Despite this complication, the laws governing atmospheric behavior follow naturally from the Lagrangian description because it is related directly to transformations of properties within an air parcel and to interactions with its environment.

The Eulerian and Lagrangian descriptions are related through a kinematic constraint: The field property at a given position and time must equal the property possessed by the air parcel occupying that position at that instant. Consider the property ψ of an individual air parcel, which has instantaneous position $(x, y, z, t) =$

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

(**x**, *t*). The incremental change of $\psi(\mathbf{x}, t)$ during the parcel's displacement $(dx, dy, dz) = d\mathbf{x}$ follows from the total differential:

$$
\begin{aligned}
d\psi &= \frac{\partial \psi}{\partial t} dt + \frac{\partial \psi}{\partial x} dx + \frac{\partial \psi}{\partial y} dy + \frac{\partial \psi}{\partial z} dz \\
&= \frac{\partial \psi}{\partial t} dt + \nabla \psi \cdot d\mathbf{x}
\end{aligned}
\tag{1}
$$

The property's rate of change following the parcel is then

$$
\begin{aligned}
\frac{d\psi}{dt} &= \frac{\partial \psi}{\partial t} + u \frac{\partial \psi}{\partial x} + v \frac{\partial \psi}{\partial y} + w \frac{\partial \psi}{\partial z} \\
&= \frac{\partial \psi}{\partial t} + \mathbf{v} \cdot \nabla \psi
\end{aligned}
\tag{2}
$$

where $\mathbf{v} = dx/dt$ is the three-dimensional velocity. Equation (2) defines the *Lagrangian derivative* of the field variable $\psi(\mathbf{x}, t)$, which corresponds to its rate of change following an air parcel. The Lagrangian derivative contains two contributions: $\partial \psi / \partial t$ represents the rate of change introduced by transience of the field property ψ, and $\mathbf{v} \cdot \nabla \psi$ represents the rate of change introduced by the parcel's motion to positions of different field values.

Consider now a finite body of air having instantaneous volume $V(t)$. The integral property

$$
\int_{V(t)} \psi(x, y, z, t)\, dV'
$$

changes through two mechanisms (Fig. 1), analogous to the two contributions to $d\psi/dt$: (1) Values of $\psi(\mathbf{x}, t)$ within the material volume change due to unsteadiness of the field. (2) The material volume moves to regions of different field values. Relative to a frame moving with $V(t)$, this motion introduces a flux of property ψ across the material volume's surface $S(t)$. Collecting these contributions and applying Gauss' theorem [see, e.g., Salby (1996)] leads to the identity

$$
\begin{aligned}
\frac{d}{dt} \int_{V(t)} \psi\, dV' &= \int_{V(t)} \left\{ \frac{\partial \psi}{\partial t} + \nabla \cdot (\mathbf{v}\psi) \right\} dV' \\
&= \int_{V(t)} \left\{ \frac{d\psi}{dt} + \psi \nabla \cdot \mathbf{v} \right\} dV'
\end{aligned}
\tag{3}
$$

Known as *Reynolds' transport theorem*, (3) constitutes a transformation between the Lagrangian and Eulerian descriptions of fluid motion, relating the time rate of change of some property of a finite body of air to the corresponding field variable and the motion field $\mathbf{v}(\mathbf{x}, t)$. Applying Reynolds' theorem to particular properties

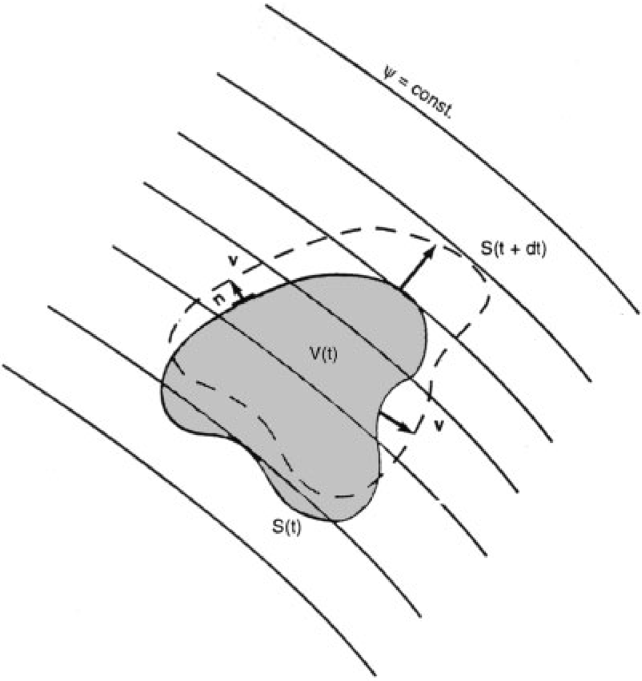


Figure 1 Finite material volume $V(t)$, containing a fixed collection of matter, that is displaced across a field property ψ. (Reproduced from Salby, 1996.)

then yields the field equations governing atmospheric behavior, which are known collectively as the *equations of motion*.

2 MASS CONTINUITY

A material volume $V(t)$ has mass

$$\int_{V(t)} \rho(\mathbf{x}, t)\, dV'$$

where ρ is density. Since $V(t)$ is comprised of a fixed collection of matter, the time rate of change of its mass must vanish

$$\frac{d}{dt}\int_{V(t)} \rho(\mathbf{x}, t)\, dV' = 0 \tag{4}$$

Applying Reynolds' theorem transforms (4) into

$$\int_{V(t)} \left\{ \frac{d\rho}{dt} + \rho \nabla \cdot \mathbf{v} \right\} dV' = 0 \tag{5}$$

where the Lagrangian derivative is given by (2). This relation must hold for "arbitrary" material volume $V(t)$. Consequently, the quantity in brackets must vanish identically. Conservation of mass for individual bodies of air therefore requires

$$\frac{d\rho}{dt} + \rho\nabla\cdot\mathbf{v} = 0 \tag{6}$$

which is known as the *continuity equation.*

Budget of Constituents

For a specific property f (i.e., one referenced to a unit mass), Reynolds' transport theorem simplifies. With (6), (3) reduces to

$$\frac{d}{dt}\int_{V(t)} \rho f\, dV' = \int_{V(t)} \rho \frac{df}{dt} dV' \tag{7}$$

where ρf represents the concentration of property f (i.e., referenced to a unit volume).

Consider now a constituent of air, such as water vapor in the troposphere or ozone in the stratosphere. The specific abundance of this species is described by the mixing ratio r, which measures its local mass referenced to a unit mass of dry air (i.e., of fixed composition). The corresponding concentration is then ρr.

Conservation of the species is then expressed by

$$\frac{d}{dt}\int_{V(t)} \rho r\, dV' = \int_{V(t)} \rho\dot{P}\, dV' \tag{8}$$

which equates the rate that the species changes within the material volume to its net rate of production per unit mass $\dot{P}$, collected over the material volume. Applying (7) and requiring the result to be satisfied for arbitrary $V(t)$ then reduces the constituent's budget to

$$\frac{dr}{dt} = \dot{P} \tag{9}$$

Like the equation of mass continuity, (9) is a partial differential equation that must be satisfied continuously throughout the atmosphere. For a long-lived chemical species or for water vapor away from regions of condensation, $\dot{P} \cong 0$, so r is approximately conserved for individual air parcels. Otherwise, production and destruction of the species are balanced by changes in a parcel's mixing ratio.

3 MOMENTUM BUDGET

In an inertial reference frame, Newton's second law of motion applied to the material volume $V(t)$ may be expressed

$$\frac{d}{dt}\int_{V(t)} \rho\mathbf{v}\, dV' = \int_{V(t)} \rho\mathbf{f}\, dV' + \int_{S(t)} \boldsymbol{\tau}\cdot\mathbf{n}\, dS' \tag{10}$$

where $\rho\mathbf{v}$ is the concentration of momentum, $\mathbf{f}$ is the body force per unit mass acting internal to the material volume, and $\boldsymbol{\tau}$ is the *stress tensor* acting on its surface (Fig. 2). The stress tensor $\boldsymbol{\tau}$ represents the vector force per unit area exerted on surfaces normal to the three coordinate directions. Then $\boldsymbol{\tau}\cdot\mathbf{n}$ is the vector force per unit area exerted on the section of material surface with unit normal $\mathbf{n}$, representing the flux of vector momentum across that surface.

Incorporating Reynolds' theorem for a concentration (7), along with Gauss' theorem, casts the material volume's momentum budget into

$$\int_{V(t)} \left\{\rho\frac{d\mathbf{v}}{dt} - \rho\mathbf{f} - \nabla\cdot\boldsymbol{\tau}\right\} dV' = 0 \tag{11}$$

As before, (11) must hold for arbitrary material volume, so the quantity in brackets must vanish identically. Newton's second law for individual bodies of air thus requires

$$\rho\frac{d\mathbf{v}}{dt} = \rho\mathbf{f} + \nabla\cdot\boldsymbol{\tau} \tag{12}$$

to be satisfied continuously.

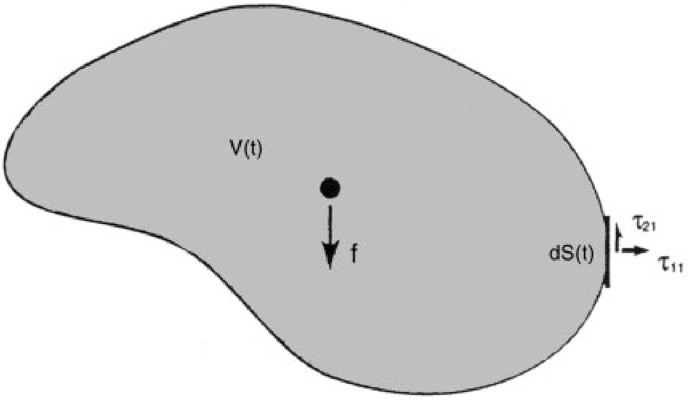


Figure 2 Finite material volume $V(t)$ that experiences a specific body force $\mathbf{f}$ internally and a stress tensor τ on its surface. (Reproduced from Salby, 1996.)

The body force relevant to the atmosphere is gravity

$$\mathbf{f} = \mathbf{g} \tag{13}$$

which is specified. The stress term, on the other hand, is determined autonomously by the motion. It represents the divergence of a momentum flux, or force per unit volume, and has two components: (i) a normal force associated with the pressure gradient ∇p and (ii) a tangential force or drag $\mathbf{D}$ introduced by friction. For a *Newtonian fluid* such as air, tangential components of the stress tensor depend linearly on the shear. Frictional drag then reduces to (e.g., Aris, 1962)

$$\begin{aligned}\mathbf{D} &= -\frac{1}{\rho}\nabla \cdot \boldsymbol{\tau} \\ &= -\frac{1}{\rho}\nabla \cdot (\mu \nabla \mathbf{v})\end{aligned} \tag{14}$$

where μ is the coefficient of viscosity. The right-hand side of (14) accounts for diffusion of momentum, which is dominated in the atmosphere by turbulent diffusion (e.g., turbulent mixing of an air parcel's momentum with that of its surroundings). For many applications, frictional drag is expressed in terms of a turbulent diffusivity ν as

$$\mathbf{D} = -\nu \frac{\partial^2 \mathbf{v}}{\partial z^2} \tag{15}$$

in which horizontal components of $\mathbf{v}$ and their vertical shear prevail.

With these restrictions, the momentum budget reduces to

$$\frac{d\mathbf{v}}{dt} = \mathbf{g} - \frac{1}{\rho}\nabla p - \mathbf{D} \tag{16}$$

which comprise the so-called *Navier–Stokes equations*. Also called the *momentum equations*, (16) assert that an air parcel's momentum changes according to the resultant force exerted on it by gravity, pressure gradient, and frictional drag.

Momentum Budget in a Rotating Reference Frame

The momentum equations are a statement of Newton's second law, so they are valid in an inertial reference frame. The reference frame of Earth, on the other hand, is rotating and consequently noninertial. The momentum equations must therefore be modified to account for acceleration of the frame in which atmospheric motion is observed.

Consider a reference frame that rotates with angular velocity $\boldsymbol{\Omega}$. A vector $\mathbf{A}$ that appears constant in the rotating frame rotates when viewed from an inertial frame.

During an interval dt, $\mathbf{A}$ changes by a vector increment $d\mathbf{A}$ that is perpendicular to the plane of $\mathbf{A}$ and $\boldsymbol{\Omega}$ (Fig. 3) and has magnitude

$$|d\mathbf{A}| = A\ \sin\theta \cdot \Omega\, dt$$

where θ is the angle between $\mathbf{A}$ and $\boldsymbol{\Omega}$. The time rate of change of $\mathbf{A}$ apparent in an inertial frame is then described by

$$\left(\frac{d\mathbf{A}}{dt}\right)_i = \boldsymbol{\Omega} \times \mathbf{A} \tag{17}$$

More generally, a vector $\mathbf{A}$ that has the time rate of change $d\mathbf{A}/dt$ in the rotating reference frame has the time rate of change

$$\left(\frac{d\mathbf{A}}{dt}\right)_i = \frac{d\mathbf{A}}{dt} + \boldsymbol{\Omega} \times \mathbf{A} \tag{18}$$

in an inertial reference frame.

Consider now the position $\mathbf{x}$ of an air parcel. The parcel's velocity $\mathbf{v} = d\mathbf{x}/dt$ apparent in an inertial reference frame is then

$$\mathbf{v}_i = \mathbf{v} + \boldsymbol{\Omega} \times \mathbf{x} \tag{19}$$

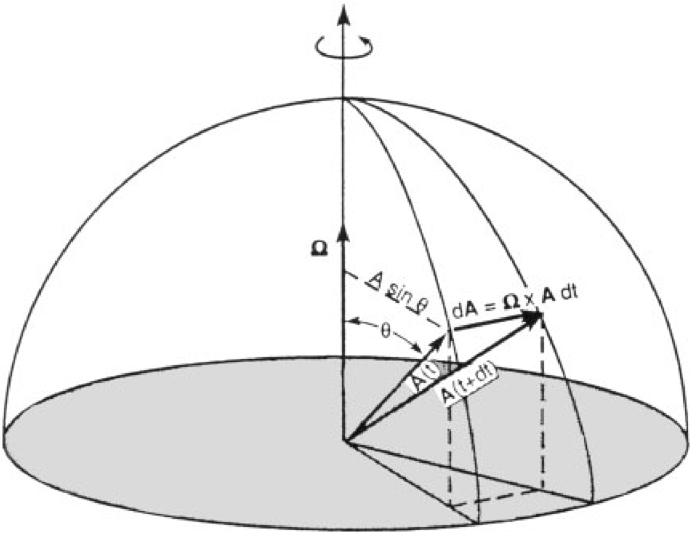


Figure 3 A vector $\mathbf{A}$ fixed in a rotating reference frame changes in an inertial reference frame during an interval dt by the increment $|d\mathbf{A}| = A\ \sin\theta \cdot \Omega\, dt$ in a direction orthogonal to the plane of $\mathbf{A}$ and Ω, or by the vector increment $d\mathbf{A} = \boldsymbol{\Omega} \times \mathbf{A}\, dt$. (Reproduced from Salby, 1996.)

Likewise, the acceleration apparent in the inertial frame is given by

$$\left(\frac{d\mathbf{v}_i}{dt}\right)_i = \frac{d\mathbf{v}_i}{dt} + \mathbf{\Omega} \times \mathbf{v}_i$$

which upon consolidation yields for the acceleration apparent in the inertial frame:

$$\left(\frac{d\mathbf{v}_i}{dt}\right)_i = \frac{d\mathbf{v}}{dt} + 2\mathbf{\Omega} \times \mathbf{v} + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{x}) \tag{20}$$

Earth's rotation introduces two corrections to the parcel's acceleration: (i) $2\mathbf{\Omega} \times \mathbf{v}$ is the *Coriolis acceleration*. It acts perpendicular to the parcel's motion and the planetary vorticity $2\mathbf{\Omega}$. (ii) $\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{x})$ is the *centrifugal acceleration* of the air parcel associated with Earth's rotation. When geopotential coordinates are used (e.g., Iribarne and Godson, 1981), this correction is absorbed into the *effective gravity*: $\mathbf{g} = -\nabla\Phi$, which is defined from the geopotential Φ.

Incorporating (20) transforms the momentum equations into a form valid in the rotating frame of Earth:

$$\frac{d\mathbf{v}}{dt} + 2\mathbf{\Omega} \times \mathbf{v} = -\frac{1}{\rho}\nabla p - g\mathbf{k} - \mathbf{D} \tag{21}$$

where g is understood to denote effective gravity and $\mathbf{k}$ the upward normal in the direction of increasing geopotential. The correction $2\mathbf{\Omega} \times \mathbf{v}$ is important for motions with time scales comparable to that of Earth's rotation. When moved to the right-hand side, it enters as the Coriolis force: $-2\mathbf{\Omega} \times \mathbf{v}$, a fictitious force that acts on an air parcel in the rotating frame of Earth. Because it acts orthogonal to the parcel's displacement, the Coriolis force performs no work.

Component Equations in Spherical Coordinates

In vector form, the momentum equations are valid in any coordinate system. However, those equations do not lend themselves to standard methods of solution. To be useful, they must be cast into component form, which then depend on the coordinate system in which they are expressed.

Consider the rectangular Cartesian coordinates $\mathbf{x} = (x_1, x_2, x_3)$ having origin at the center of Earth (Fig. 4) and the spherical coordinates

$$\mathbf{x} = (\lambda, \phi, r)$$

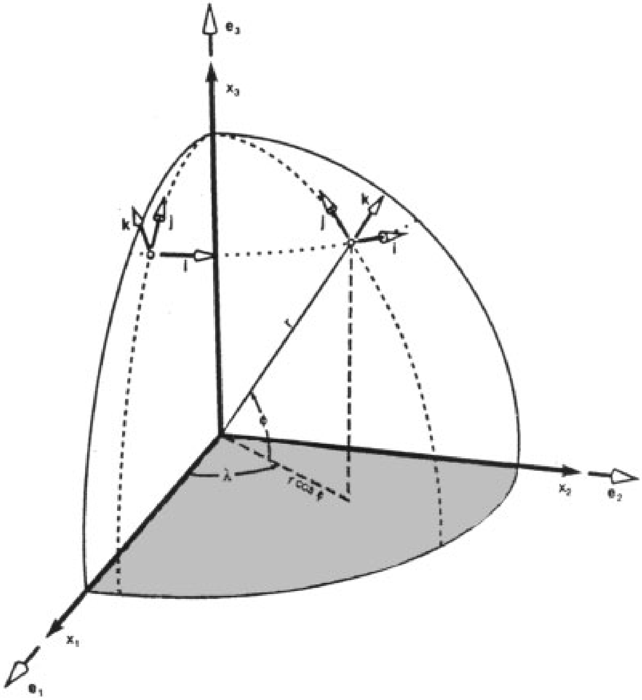


Figure 4 Spherical coordinates: longitude λ, latitude ϕ, and radial distance r. Coordinate vectors $\mathbf{e}_\lambda = \mathbf{i}$, $\mathbf{e}_\phi = \mathbf{j}$, and $\mathbf{e}_\gamma = \mathbf{k}$ change with position (e.g., relative to fixed coordinate vectors $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ of rectangular Cartesian coordinates). (Reproduced from Salby, 1996.)

both fixed with respect to Earth. The corresponding unit vectors

$$\begin{aligned} \mathbf{e}_\lambda &= \mathbf{i} \\ \mathbf{e}_\phi &= \mathbf{j} \\ \mathbf{e}_r &= \mathbf{k} \end{aligned} \tag{22}$$

point in the directions of increasing λ, ϕ, and r and are mutually perpendicular.

The rectangular Cartesian coordinates can be expressed in terms of the spherical coordinates as

$$\begin{aligned} x_1 &= r \cos\phi \cos\lambda \\ x_2 &= r \cos\phi \sin\lambda \\ x_3 &= r \sin\phi \end{aligned} \tag{23a}$$

which may be inverted for the spherical coordinates:

$$\begin{aligned}
\lambda &= \tan^{-1}\left(\frac{x_2}{x_1}\right) \\
\phi &= \tan^{-1}\left(\frac{x_3}{x_1^2 + x_2^2}\right) \\
r &= \sqrt{x_1^2 + x_2^2 + x_3^2}
\end{aligned} \tag{23b}$$

Parcel displacements in the directions of increasing longitude, latitude, and radial distance are then described by

$$\begin{aligned}
dx &= r \cos\phi \, d\lambda \\
dy &= r \, d\phi \\
dz &= dr
\end{aligned} \tag{24a}$$

in which vertical distance is measured by height

$$z = r - a \tag{24b}$$

where a denotes the mean radius of Earth. Velocity components in the spherical coordinate system are then expressed by

$$\begin{aligned}
u &= \frac{dx}{dt} = r \cos\phi \frac{d\lambda}{dt} \\
v &= \frac{dy}{dt} = r \frac{d\phi}{dt} \\
w &= \frac{dz}{dt} = \frac{dr}{dt}
\end{aligned} \tag{25}$$

The vector momentum equations may now be expressed in terms of the spherical coordinates. Lagrangian derivatives of vector quantities are complicated by the dependence on position of the coordinate vectors **i**, **j**, and **k**. Each rotates in physical space under a displacement of longitude or latitude. For example, an air parcel moving along a latitude circle at a constant speed u has a velocity $\mathbf{v} = u\mathbf{i}$, which appears constant in the spherical coordinate representation, but which actually rotates in physical space (Fig. 4). Consequently, the parcel experiences an acceleration that must be accounted for in the equations of motion.

Consider the velocity

$$\mathbf{v} = u\mathbf{i} + v\mathbf{j} + w\mathbf{k}$$

Since the spherical coordinate vectors **i**, **j**, and **k** are functions of position $\hat{\mathbf{x}}$, a parcel's acceleration is actually

$$\begin{aligned}\frac{d\mathbf{v}}{dt} &= \frac{du}{dt}\mathbf{i} + \frac{dv}{dt}\mathbf{j} + \frac{dw}{dt}\mathbf{k} + u\frac{d\mathbf{i}}{dt} + v\frac{d\mathbf{j}}{dt} + w\frac{d\mathbf{k}}{dt} \\ &= \left(\frac{d\mathbf{v}}{dt}\right)_C + u\frac{d\mathbf{i}}{dt} + v\frac{d\mathbf{j}}{dt} + w\frac{d\mathbf{k}}{dt}\end{aligned} \tag{26a}$$

where the subscript refers to the basic form of the Lagrangian derivative in Cartesian geometry

$$\left(\frac{d}{dt}\right)_C = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla \tag{26b}$$

Corrections appearing on the right-hand side of (26) can be evaluated by expressing the spherical coordinate vectors in terms of the fixed rectangular Cartesian coordinate vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Lagrangian derivatives of the spherical coordinate vectors then follow as

$$\begin{aligned}\frac{d\mathbf{i}}{dt} &= u\left(\frac{\tan\phi}{r}\mathbf{j} - \frac{1}{r}\mathbf{k}\right) \\ \frac{d\mathbf{j}}{dt} &= -u\frac{\tan\phi}{r}\mathbf{i} - \frac{v}{r}\mathbf{k} \\ \frac{d\mathbf{k}}{dt} &= \frac{u}{r}\mathbf{i} + \frac{v}{r}\mathbf{j}\end{aligned} \tag{27}$$

With these, material accelerations in the spherical coordinate directions become

$$\frac{du}{dt} = \left(\frac{du}{dt}\right)_C - \frac{uv\tan\phi}{r} + \frac{uw}{r} \tag{28a}$$

$$\frac{dv}{dt} = \left(\frac{dv}{dt}\right)_C + \frac{u^2\tan\phi}{r} + \frac{vw}{r} \tag{28b}$$

$$\frac{dw}{dt} = \left(\frac{dw}{dt}\right)_C - \left(\frac{u^2+v^2}{r}\right) \tag{28c}$$

Introducing (28), making use of the atmosphere's shallowness ($z \ll a$), and formally adopting height as the vertical coordinate then casts the momentum equations into component form:

$$\frac{du}{dt} - 2\Omega(v \sin\phi - w \cos\phi) = -\frac{1}{\rho a \cos\phi}\frac{\partial p}{\partial \lambda} + uv\frac{\tan\phi}{a} - \frac{uw}{a} - D_\lambda \tag{29a}$$

$$\frac{dv}{dt} + 2\Omega u \sin\phi = -\frac{1}{\rho a}\frac{\partial p}{\partial \phi} - \frac{u^2 \tan\phi}{a} - \frac{uw}{a} - D_\phi \tag{29b}$$

$$\frac{dw}{dt} - 2\Omega u \cos\phi = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g + \frac{u^2 + v^2}{a} - D_z \tag{29c}$$

4 FIRST LAW OF THERMODYNAMICS

For the material volume $V(t)$, the first law of thermodynamics is expressed by

$$\frac{d}{dt}\int_{V(t)} \rho c_v T \, dV' = -\int_{S(t)} \mathbf{q} \cdot \mathbf{n} \, dS' - \int_{V(t)} \rho p \frac{d\alpha}{dt} dV' + \int_{V(t)} \rho \dot{q} \, dV' \tag{30}$$

where c_v is the specific heat at constant volume, so $c_v T$ represents the internal energy per unit mass, $\mathbf{q}$ is the local heat flux so $-\mathbf{q} \cdot \mathbf{n}$ represents the heat flux "into" the material volume, $\alpha = 1/\rho$ is the specific volume so $p d\alpha/dt$ represents the specific work rate, and $\dot{q}$ denotes the specific rate of internal heating (e.g., associated with the latent heat release and frictional dissipation of motion).

In terms of specific volume, the continuity equation (6) becomes

$$\frac{1}{\alpha}\frac{d\alpha}{dt} = \nabla \cdot \mathbf{v} \tag{31}$$

Incorporating (31), along with Reynolds' transport theorem (3) and Gauss' theorem, transforms (30) into

$$\int_{V(t)} \left\{ \rho c_v \frac{dT}{dt} + \nabla \cdot \mathbf{q} + p \nabla \cdot v - \rho \dot{q} \right\} dV' = 0$$

Since $V(t)$ is arbitrary, the quantity in brackets must again vanish identically. Therefore, the first law applied to individual bodies of air requires

$$\rho c_v \frac{dT}{dt} = -\nabla \cdot \mathbf{q} - p \nabla \cdot \mathbf{v} + \rho \dot{q} \tag{32}$$

to be satisfied continuously.

The heat flux can be separated into radiative and diffusive components:

$$\begin{aligned} \mathbf{q} &= \mathbf{q}_R + \mathbf{q}_T \\ &= \mathbf{F} - k\nabla T \end{aligned} \tag{33}$$

where **F** is the net radiative flux and k is the thermal conductivity in Fourier's law of heat conduction. The first law then becomes

$$\rho c_v \frac{dT}{dt} + p\nabla \cdot \mathbf{v} = -\nabla \cdot \mathbf{F} + \nabla \cdot (k\nabla T) + \rho \dot{q} \tag{34}$$

Known as the *thermodynamic equation*, (34) expresses the rate that a material element's internal energy changes in terms of the rate that it performs work and the rate that it absorbs heat through convergence of radiative and diffusive energy fluxes.

The thermodynamic equation is expressed more compactly in terms of another thermodynamic property, one that accounts collectively for a change of internal energy and expansion work. The *potential temperature* θ is defined as

$$\frac{\theta}{T} = \left(\frac{p_0}{p}\right)^{\kappa} \tag{35}$$

where $\kappa = R/c_p$, R is the specific gas constant for air, and temperature and pressure are related through the ideal gas law

$$p = \rho RT \tag{36}$$

and θ is related to a parcel's entropy. It is conserved during an adiabatic process, as characterizes a parcel's motion away from Earth's surface and cloud. Incorporating θ into the second law of thermodynamics (e.g., Salby, 1996) then leads to the fundamental relation

$$c_p T \frac{d \ln \theta}{dt} = \frac{du}{dt} + p \frac{d\alpha}{dt} \tag{37}$$

where d/dt represents the Lagrangian derivative. Making use of the continuity equation (31) transforms this into

$$\frac{\rho c_p T}{\theta} \frac{d\theta}{dt} = \rho c_v \frac{dT}{dt} + p\nabla \cdot \mathbf{v} \tag{38}$$

Then incorporating (38) into (34) absorbs the expansion work into the time rate of change of θ to yield the thermodynamic equation

$$\rho \frac{c_p T}{\theta} \frac{d\theta}{dt} = -\nabla \cdot \mathbf{F} + \nabla \cdot (k \nabla T) + \rho \dot{q} \tag{39}$$

Equation (39) relates the change in a parcel's potential temperature to net heat transfer with its surroundings. In the absence of such heat transfer, θ is conserved for individual air parcels.

REFERENCES

Aris, R. (1962). *Vectors*, *Tensors*, *and the Basic Equations of Fluid Mechanics*, Englewood Cliffs, NJ, Prentice Hall.

Iribarne, J., and W. Godson (1981). *Atmospheric Thermodynamics*, Reidel, Dordrecht.

Salby, M. (1996). *Fundamentals of Atmospheric Physics*, San Diego, Academic.

CHAPTER 3

CIRCULATION, VORTICITY, AND POTENTIAL VORTICITY

PAUL J. KUSHNER

1 INTRODUCTION

Vorticity and circulation are closely related quantities that describe rotational motion in fluids. Vorticity describes the rotation at each point; circulation describes rotation over a region. Both quantities are of fundamental importance to the field of fluid dynamics. The distribution and statistical properties of the vorticity field provide a succinct characterization of fluid flow, particularly for weakly compressible or incompressible fluids. In addition, stresses acting on the fluid are often interpreted in terms of the generation, transport, and dissipation of vorticity and the resulting impact on the circulation.

Potential vorticity (PV) is a generalized vorticity that combines information about both rotational motion and density stratification. PV is of central importance to the field of geophysical fluid dynamics (GFD) and its subfields of dynamical meteorology and physical oceanography. Work in GFD often focuses on flows that are strongly influenced by planetary rotation and stratification. Such flows can often be fully described by their distribution of PV. Similarly to vorticity, the generation, transport, and dissipation of PV is closely associated with stresses on the fluid.

Vorticity, circulation, and PV are described extensively in several textbooks (e.g., Holton, 1992; Gill, 1982; Kundu, 1990; Salmon, 1998). This review is a tutorial, with illustrative examples, that is meant to acquaint the lay reader with these concepts and the scope of their application. An appendix provides a mathematical summary.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

2 CIRCULATION AND VORTICITY: DEFINITIONS AND EXAMPLES

Circulation is a physical quantity that describes the net movement of fluid along a chosen *circuit*, that is, a path, taken in a specified direction, that starts at some point and returns to that point.

Statement 1. The *circulation* at a given time is the average, over a circuit, of the component of the flow velocity tangential to the circuit, multiplied by the length of the circuit. Circulation therefore has dimensions of length squared per unit time.

Consider the circuit drawn as a heavy curve in Figure 1 for a fluid whose velocity $\mathbf{u}$ is indicated by the arrows. At each point we may split the velocity into components tangential to and perpendicular to the local direction of the circuit. The tangential component is defined as the dot product $\mathbf{u} \cdot \hat{\mathbf{l}} = |\mathbf{u}| \cos \theta$, where $\hat{\mathbf{l}}$ is a unit vector that points in the direction tangential to the circuit and θ is the angle between $\mathbf{u}$ and $\hat{\mathbf{l}}$. Note that $\hat{\mathbf{l}}$ points in the specified direction chosen for the circuit. Where θ is acute, the tangential component is positive, where oblique, negative, and where a right angle, zero. To calculate the circulation, we average over "all" points along the circuit. (Although we cannot actually average over all points along the circuit, we can approximate such an average by summing over the tangential component $\mathbf{u} \cdot \hat{\mathbf{l}} = |\mathbf{u}| \cos \theta$ at evenly and closely spaced points around the circuit and by dividing by the number of points.) The circulation would then be this average, multiplied by the length of the circuit. (In the Appendix, the circulation is defined in terms of a line integral.)

The circulation is not defined with reference to a particular point in space but to a circuit that must be chosen. For example, to measure the strength of the primary near-surface flow around a hurricane in the Northern Hemisphere, a natural circuit is a circle, oriented horizontally, centered about the hurricane's eye, of radius somewhat larger than the eye, and directed counterclockwise when viewed from above. We thus have specified that the circuit run in the "cyclonic" direction, which is the direction of the primary flow around tropical storms and other circulations with low-pressure centers in the Northern Hemisphere (Fig. 2).

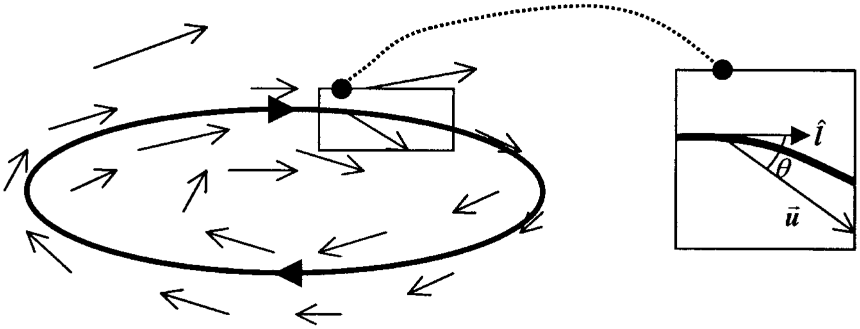


Figure 1 Heavy curve, dark arrows: circuit. Light arrows: flow velocity.

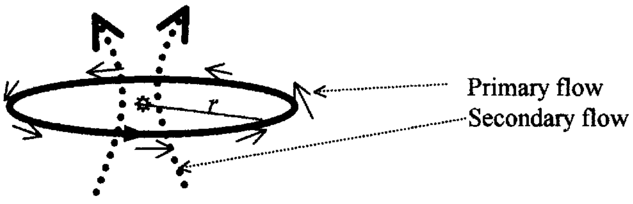


Figure 2 Circuit shown with bold curve. ✿ symbol represents a paddle wheel with its axis in the vertical and located at the center of the circle.

Suppose that the average flow tangential to the circuit is approximately $u = 40\,\text{m/s}$ and that the radius of the circle is $r = 100\,\text{km}$. The circulation, Γ, using the definition above, is then the average tangential flow times the circumference of the circle: $\Gamma = 2\pi r u \approx 2.5 \times 10^7\,\text{m}^2/\text{s}$. If the circuit were chosen to run clockwise, i.e., opposite to the primary flow, instead of counterclockwise, the circulation would be of the same strength but of opposite sign.

Circulation is a selective measure of the strength of the flow: It tells us nothing about components of motion perpendicular to the circuit or components of the flow that average to zero around the circuit. For example, suppose the hurricane is blown to the west with a constant easterly wind. This constant wind will not contribute to the circulation: Since $\cos(\theta + \pi) = -\cos\theta$, every point on the north side of the circle that contributes some amount to the circulation has a counterpart on the south side that contributes an equal and opposite amount. As another example, the circuit in Figure 2 would not measure the strength of the hurricane's secondary flow, which runs toward the eye near the surface and away from the eye aloft (dotted arrows in Fig. 2). The reader might try to imagine a circuit that would measure this secondary flow.

The hurricane example might suggest that the fluid must flow around the chosen circuit for the circulation to be nonzero; however, this need not be true. For example, consider the circuit shown in Figure 3, which shows a vertically oriented 50- × 200-m rectangular circuit in a west-to-east flow whose strength increases linearly with height, according to the formula $u(z) = az$, where z is the height in meters and $a = 0.01/\text{s} = (10\,\text{m/s})\,\text{km}$ is the "vertical shear," that is, the rate of change of the wind with respect to height. This is a value of vertical shear that might be encountered in the atmospheric boundary layer. If the direction of the circuit is chosen, as in the figure, to be clockwise, the circulation in this example is the average of the along-circuit component of the flow weighted by the length of each side,

$$\frac{50 \times u(\text{top}) - 50 \times u(\text{bottom}) + 200 \times 0 + 200 \times 0}{50 + 50 + 200 + 200} = \frac{50 \times 200a}{500} = 0.2\,\text{m/s}$$

times the perimeter of the rectangle, giving $\Gamma = 100\,\text{m}^2/\text{s}$. Note that the vertical sides make no contribution to the circulation since the flow is perpendicular to them.

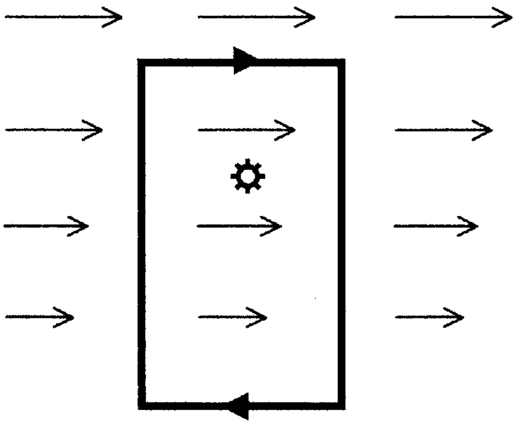


Figure 3 Circuit shown with bold line. ⚙ symbol represents paddle wheel with rotation axis coming out of page.

Thus, shear alone can make the circulation nonzero even though the flow is not actually moving around the circuit.

What, then, does the circulation represent in these examples, if not the strength of the flow around the circuit? In general, circulation represents the ability of a given flow to rotate mass elements in the fluid and objects placed in the flow. Imagine a paddle wheel placed within the circuits in Figures 2 and 3 [see, e.g., Holton (1992), Fig. 4.6], with axis perpendicular to the plane of the circuit. In the hurricane illustration, the paddle wheel rotates counterclockwise; in the vertical shear-flow, clockwise. In both examples, when the paddle wheel turns in the same direction as the circuit in the illustration, the sign of the circulation is positive.

Circulation depends on many details about the circuit: its size, shape, orientation, direction, and location. For example, in Figure 3, we could make the circulation arbitrarily large by increasing the length of the sides of the circuit. It is useful to define a quantity that measures rotation but that does not refer to the details of any particular circuit. It is also useful to define a "microscopic" quantity that measures rotation at individual points instead of a "macroscopic" quantity that measures rotation over a region. This leads us to vorticity.

> **Statement 2.** The *vorticity* in the right-hand-oriented direction perpendicular to the plane of a circuit is equal to the circulation per unit area of the circuit in the small-area limit. Vorticity is, therefore, a quantity with dimensions of inverse time. In order to define the x, y, and z components of the vorticity, we consider circuits whose perpendicular lies in each of these directions.

Statement 2 requires some explanation. Consider the circulation for a circuit which, for simplicity, we assume to be flat. This defines a unique direction that is perpendicular to the plane of the circuit if we use the "right-hand rule." This rule works as follows: Curl the fingers of the right hand in the direction of the circuit, e.g., clockwise in Figure 3. Then point the right-hand thumb in a direction perpendicular to the circuit: into the page in Figure 3. In statement 2, we refer to this direction as the "right-hand-oriented direction normal to the plane of the circuit." Now, imagine reducing the size of the circuit until it becomes vanishingly small. In Figure 3, for example, we could imagine reducing the loop to $5 \times 20\,\mathrm{m}$, then $5 \times 20\,\mathrm{mm}$, and so on. In statement 2, we refer to this as the "small-area limit." Therefore, vorticity describes rotation as a microscopic quantity.

We can calculate vorticity in the previous examples. For the shear-flow example (Fig. 3), it is easy to show that the circulation around the loop is aA, where A is the area of the rectangular loop. With more effort, it can be shown that the circulation is aA for any loop if A is taken to be the cross-sectional area of the loop in the plane of the flow. Therefore, the vorticity in the plane of the flow is into the page and equal to $a = 0.01/\mathrm{s}$ everywhere. The reader should try to verify that the other two components of the vorticity are zero because the associated circulation is zero. In hurricanes, as is typical of other vortices, the vorticity varies strongly as a function of distance from the eye. The average vertical component of vorticity in our example is upward and equal to the circulation divided by the total area of the circle: $\Gamma/\mathrm{A} = 2\pi ru/(\pi r^2) = 2u/r \approx 10^{-3}/\mathrm{s}$.

In GFD, there are three types of vorticity: absolute, planetary, and relative. The *absolute vorticity* is the vorticity measured from an inertial frame of reference, typically thought of as outer space. The *planetary vorticity* is the vorticity associated with planetary rotation. Earth rotates with a period of $T \approx 24\,\mathrm{h}$ from west to east around the north pole–south pole rotation axis (Fig. 4). Consider a fluid with zero ground speed, meaning that it is at rest when measured from the frame of reference of Earth. This fluid will also be rotating with period T from west to east. The fluid is said to be "in solid-body rotation" because all the fluid elements maintain the same distance from one another over time, just as the mass elements do in a rotating solid. The distance from the axis of rotation is $r\cos\theta$, the component of the velocity tangential to this rotation is $2\pi r\cos\theta/T$, the circumference of the latitude circle is $2\pi r\cos\theta$, and the circulation, by statement 1, is the product of the two last

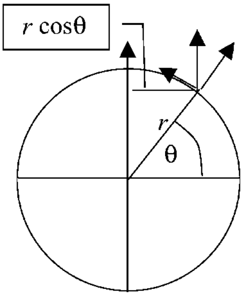


Figure 4 Geometry of Earth. Latitude is θ The vertical vector is the vorticity of solid-body rotation. The distance from the surface to the axis of rotation is $r\cos\theta$. The component of the planetary vorticity perpendicular to the surface is $4\pi\sin\theta/T$. The component tangential to the surface is $4\pi\cos\theta/T$.

quantities $(2\pi r \cos\theta)^2/T$. The vorticity points from the south pole to the north pole and has magnitude $4\pi/T = 1.4 \times 10^{-4}$ per second, where we have divided the circulation by the area of the latitude circle. This is the planetary vorticity. Finally, the *relative vorticity* is the vorticity of the velocity measured with respect to the ground, i.e. in the rotating frame of reference of Earth. The absolute vorticity is the sum of the relative vorticity and the planetary vorticity (since the velocities are additive).

The comparative strengths of planetary vorticity and relative vorticity determine, in large part, the different dynamical regimes of GFD. The planetary vorticity has a component perpendicular to the planet's surface with value $4\pi \sin\theta/T$. This component points radially away from the center of the planet in the Northern Hemisphere where $\sin\theta > 0$, and toward the center of the planet in the Southern Hemisphere where $\sin\theta < 0$. From the viewpoint of an observer on the surface in the Northern Hemisphere, this component points vertically up toward outer space and, in the Southern Hemisphere, vertically down toward the ground. For motions that are characterized by scales of 1000 km or greater in the atmospheric midlatitudes, or 100 km or larger in the oceanic midlatitudes, the radial, "up/down" component of the planetary vorticity is an order of magnitude larger than the relative vorticity in this direction. This is the dynamical regime for midlatitude cyclones and oceanic mesoscale eddies, for which geostrophic balance and quasi-geostrophy hold (see chapter by Salby). At scales smaller than this, the relative vorticity can be comparable to or larger than the planetary vorticity. For instance, in the hurricane example, the vertical component of the vorticity was determined to be 10^{-3}/s in a region 100 km across. Often even larger are the values of the horizontal components of vorticity associated with vertical shear—recall the vertical shear example, with vorticity of magnitude 10^{-2}/s. Although it is large values of the vertical component of vorticity that are associated with strong horizontal surface winds, the availability of large values of horizontal vorticity associated with vertical shear can have a potentially devastating impact. For example, the tilting into the vertical of horizontal shear vorticity characterizes the development of thunderstorms and tornadoes (e.g., Cotton and Anthes, 1989).

Having defined vorticity in terms of the circulation, it is also useful to go from the microscopic to the macroscopic and define circulation in terms of the vorticity. We first introduce the idea of a vector *flux*: the flux of any vector field through a surface is the average of the component of the vector perpendicular to the surface, multiplied by the area of the surface. For example, the flux of a 10-m/s flow passing through a pipe of cross-sectional area 0.5 m^2 is a volume flux of 5 m^3/s.

Statement 3. The circulation for a given circuit is equal to the flux of the vorticity through any surface bounded by the circuit.

To illustrate statement 3, we consider Figure 5, which shows adjacent rectangular circuits with circulation values C_1, C_2 and areas A_1, A_2 that are small enough to have unique values of the vorticity $Z_1 = C_1/A_1$, $Z_2 = C_2/A_2$ pointing into the page. The total circulation for the larger rectangular region formed by joining the two smaller

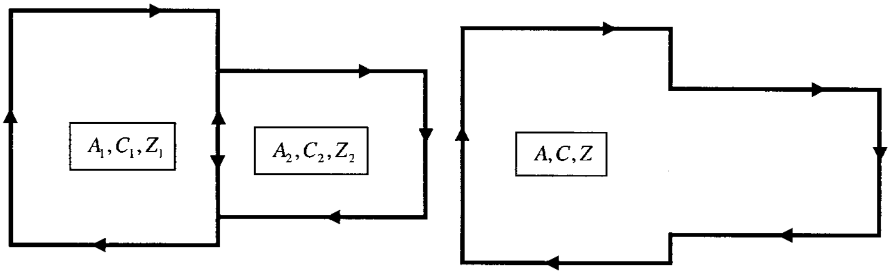


Figure 5 Calculation of net circulation.

rectangles is $C = C_1 + C_2$. This is because, along the shared side (marked in the figure with circulation arrows pointing in both directions), the tangential flow component for circuit 1 is equal and opposite to that for circuit 2. The average component of the vorticity for this region is $(Z_1A_1 + Z_2A_2)/(A_1 + A_2) = (C_1 + C_2)/(A_1 + A_2) = C/A$, where A is the total area. The flux of vorticity is then $(C/A) \times A = C$, which is the total circulation, consistent with statement 3. We can repeat this calculation to include more rectangles in order to determine the average component of the vorticity over a large region. Although we have illustrated the simple case in which the surface bounded by the circuit is flat, statement 3 holds for any surface, flat or curved, bounded by the circuit. This is because statement 3 is a statement of the Stokes's theorem of vector calculus (see Appendix).

Statement 3 shows that vorticity distributed over a small region can be associated with a circulation well away from that region. Typical vortices in geophysical flows tend to have a core of strong vorticity surrounded by a region of relatively weak or zero vorticity. Suppose the vortex covers an area A and has a perpendicular component of vorticity of average value Z. The circulation induced by this flux, for any circuit enclosing the vortex, is AZ. Consider a circuit that spans an area larger than A. The perimeter of such a circuit is proportional to its average distance l from the center of the vortex. (Recall the hurricane example for which the radius is l and the perimeter is $2\pi l$.) Then the induced tangential flow speed associated with the vortex, from statement 1, is proportional to AZ/l. That is, for typical localized vortex distributions, the flow around the vortex varies as the reciprocal of distance from the center. The ability of vorticity to induce circulation "at a distance" and the variation with the reciprocal of distance of the flow strength are key to understanding many problems in GFD.

Statement 3 also implies that mechanisms that change the vorticity in some region can change the circulation of any circuit that encloses that region. For instance, in the hurricane example of Figure 2, suppose that drag effects near the surface reduce the vertical component of the vorticity in some location (see next section). By statement 3, this would reduce the average circulation around the circuit. In other words, the reduction in vorticity would decelerate the flow around the circuit. In this way, we see that vorticity transport, generation, and dissipation are associated with stresses that accelerate or decelerate the flow.

In the final example of this section, we will illustrate, with an idealized model, how vorticity transport gives rise to flow accelerations on a planetary scale (I. Held, personal communication). A thin layer of fluid of constant density and depth surrounds a solid, featureless, "billiard ball" planet of radius *r*. By "thin," we mean that the depth of the fluid is much less than *r*. Both planet and fluid layer are rotating, with period *T*, from west to east. As we have seen, the vorticity of the solid-body rotation is the planetary vorticity, and the component of this vorticity perpendicular to the surface is $4\pi \sin\theta/T$. We will learn, shortly, that this normal component, in the absence of applied forcing or friction, acts like a label on fluid parcels and stays with them as they move around. In other words, the component of the absolute vorticity normal to the surface of the fluid is a tracer.

Suppose, now, that a wave maker near the equator generates a wave disturbance, and that this wave disturbance propagates away from the region of the wave maker. Away from the wave maker, the parcels will be displaced by the propagating disturbance. Given that the component of the absolute vorticity normal to the surface is a tracer, fluid elements so displaced will carry their initial value of this quantity with them. For example, a fluid element at 45N latitude will preserve its value of vorticity, $4\pi \sin(\pi/4)/T = 1.0 \times 10^{-4}$ per second as it moves north and south. Now, $4\pi \sin\theta/T$ is an increasing function of latitude: there is a gradient in this quantity, from south to north. Therefore, particles from low-vorticity regions to the south will move to regions of high vorticity to the north, and vice versa. There will be, therefore, a net southward transport of vorticity, and a reduction in the total vorticity poleward of that latitude. Notice that this transport is down-gradient. Thus, the circulation around that latitude, for a west-to-east circuit, will be reduced and the fluid at that latitude will begin to stream westward as a result of the disturbance. The transport of vorticity by the propagating disturbance gives rise to a stress that induces acceleration on the flow. This acceleration is in the direction perpendicular to the vorticity transport.

We can estimate the size of the disturbance-induced acceleration. Suppose, after a few days, that particles are displaced, on average, by 10 degrees latitude, which corresponds to a distance of about 1000 km. Since $4\pi \sin\theta/T$ has values between -1.4×10^{-4} and 1.4×10^{-4} per second for $T = 24\,\text{h}$, a reasonable estimate for the difference between a displaced particle's vorticity and the background vorticity is 10^{-5} per second for a 10 degrees latitude displacement. The average estimated southward transport of the vorticity is then the displacement times the perturbation vorticity: $1000\,\text{km} \times 10^{-5}$ per second $= 10\,\text{m/s}$. This is an estimate of the westward flow velocity induced by the displacement over a few days and corresponds to reasonable values of the observed eddy-induced stresses on the large-scale flow.

3 POTENTIAL VORTICITY: DEFINITION AND EXAMPLES

The previous example describes the effect on the horizontal circulation of redistributing the vorticity of solid-body rotation. Although the example seems highly idealized, the wave-induced stress mechanism it illustrates is fundamental to Earth's large-scale atmospheric circulation. The physical model of a thin, fixed-density, and constant-depth fluid, which is known as the barotropic vorticity model, is deceptively simple but has been used as the starting point for an extensive body of work in GFD. This work ranges from studies of the large-scale ocean circulation (Pedlosky 1996), to the analysis of the impact of tropical disturbances such as El Niño on the midlatitude circulation, and to early efforts in numerical weather forecasting. Such applications are possible because the idealizations in the example are not as drastic as they initially appear. For example, Earth's atmosphere and ocean are quite thin compared to the radius of Earth: Most of the atmosphere and world ocean are contained within a layer about 20 km thick, which is small compared to Earth's radius (about 6300 km). In addition, large-scale atmospheric speeds are characteristically 10 to 20 m/s, which is small compared to the speed of Earth's rotation ($2\pi r \cos\theta/T = 460\cos\theta\,\mathrm{m/s} \approx 300\,\mathrm{m/s}$ in the midlatitudes)—the atmosphere is not so far from solid-body rotation. This is consistent with the idea that at large scales, atmospheric and oceanic flows have vertical relative vorticity components that are much smaller than the planetary vorticity. Perhaps the most drastic simplifications in the example are that the layer of fluid has constant depth and constant density. In this section, we consider variations of fluid layer depth and of fluid density; this will lead us to potential vorticity (PV).

Since large-scale atmospheric circulations typically occur in thin layers, these flows are typically *hydrostatic*. This means that there is a balance between the vertical pressure force and the force of gravity on each parcel, that vertical accelerations are weak, and that the pressure at any point is equal to the weight per unit area of the fluid column above that point. Consider a thin hydrostatic fluid of constant density but of variable depth. It can be shown (e.g., Salmon 1998) that the state of the fluid may be completely specified by three variables: the two horizontal components of the velocity and the depth of the fluid. The system formed by these three variables and the equations that govern them is known as the *shallow-water model*. These three variables are independent of depth, which implies that there is only a single component of vorticity: the vertical component of vorticity associated with the north–south and east–west components of motion.

The shallow-water model provides a relatively simple context to begin to think about PV:

Statement 4. The *shallow-water PV* is equal to the absolute vorticity divided by the depth of the fluid; it has dimensions of inverse time–length.

By this definition, the PV can be changed by either changing the vorticity or by changing the depth of the fluid column. The depth of the fluid column can be changed, in turn, by the fluid column encountering variations in depth of the topography below the fluid or in the height of the fluid's surface. For example, if the PV of a fluid column is held constant and the depth of the fluid decreased, by, for example, shoaling (i.e., moving the column up a topographic slope), the result is a reduction in the column's vorticity.

The generalization of PV to fluids in which the density is not constant can be approached in stages. To start with, consider, instead of a single layer of fluid, a system consisting of two or more thin, constant-density, hydrostatic layers in which each layer lies under another layer of lighter fluid. For this system, the PV in each layer is the ratio of the vertical vorticity to the layer depth. If density is instead taken to vary continuously, extra complications are added. First, at least six variables are required to specify the state of the fluid: the three components of the velocity, the pressure, the density, and another thermodynamic variable such as the temperature or the specific entropy. The specific entropy is the entropy per unit mass (see Section 4). Additional variables are needed to account for salinity in the ocean and moisture in the atmosphere; we neglect these additional factors. The second complication is that, since the velocity is three dimensional, so now is the vorticity vector. The generalization of the shallow-water PV to fluids with variable density is known as Ertel's PV.

> **Statement 5.** *Ertel's PV* is equal to the projection of the absolute vorticity vector onto the spatial gradient of the specific entropy, divided by density. Its dimensions depend on the physical dimensions of the entropy measure used.

To explain statement 5 in more detail: Recall that the projection of vector **a** onto vector **b** is **a** · **b**. The gradient of the specific entropy is a vector, pointing perpendicular to a surface of constant specific entropy, in the direction of increasing specific entropy. Its value is equal to the rate of change of specific entropy per unit along-gradient distance (Fig. 6). Despite the additional complexity, there are analogies between the shallow-water PV (statement 4) and Ertel's PV (statement 5). For example, for fixed Ertel's PV, we can decrease a column's vorticity by decreasing the thickness between two surfaces of specific entropy, since this would increase the gradient. This reduction in thickness is analogous to compressing the fluid column, as in the shallow-water example. The connection between statements 4 and 5 will be discussed in more detail below.

We have defined circulation, vorticity, and PV but have made little explicit reference to fluid dynamics, that is, to the interactions and balances within a fluid that allow us to predict the behavior of these quantities. Dynamical considerations justify our interest in defining PV as we have here, and link the PV back to our original discussion of circulation. Another important loose end is to look at the impact of planetary rotation in more detail, since rotation dominates large-scale motions on Earth. These topics will be taken up in the next section.

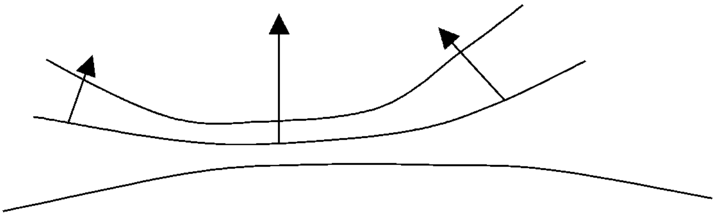


Figure 6 Contours: surfaces of constant specific entropy. Arrows: specific-entropy gradient, pointing in direction of increasing specific entropy.

4 DYNAMICAL CONSIDERATIONS: KELVIN'S CIRCULATION THEOREM, ROTATION, AND PV DYNAMICS

We begin our discussion of the dynamical aspects of circulation, vorticity, and PV by returning to the shallow-water system, which consists of a thin hydrostatic layer of fluid of constant density. In this section, we will need to introduce several new concepts. First, we introduce the concept of a *material circuit*. Points on a material circuit do not stay fixed in space, as in the circuits in Figures 2 and 3, but instead follow the motion of the fluid. In the presence of shear, mixing, and turbulence, material circuits become considerably distorted over time.

> **Statement 6.** In the absence of friction and applied stresses, the absolute circulation in the shallow-water model is constant for a material circuit.

Statement 6 asserts that the circulation measured along a material circuit, as it follows the flow, will not change. This is a special case of Kelvin's circulation theorem, which applies to a fluid with variable density, and which will be discussed shortly.

Statement 6 follows from applying, to a material circuit, the fluid momentum equations that express Newton's second law of motion. Newton's second law expresses the balance between the acceleration of the flow and the sum of the forces per unit mass acting on the fluid. For the shallow-water fluid, these forces include the pressure force, friction near solid boundaries, and applied stresses (such as the stress exerted by the wind on the ocean). The pressure force is not mentioned in statement 6 because, when averaged around a circuit, it cannot generate circulation on any circuit in the shallow-water system. To understand this, we need to remind the reader of the concept of *torque*. Torque is a force that acts on a body through a point other than the body's center of mass. Forces that act through a body's center of mass cause acceleration in the direction of the force. Torque, on the other hand, causes rotation. For example, it is a torque imparted by the vorticity in the fluid that sets spinning the paddle wheels in Figures 2 and 3. Torque is a vector, which, by convention, points in the direction of the imparted rotation using the right-hand rule.

The pressure force in the shallow-water system acts through the center of fluid *elements*, that is, fluid columns in the small-area limit. Therefore, the pressure force cannot directly impart a rotational torque to fluid elements. This can be used to show that the pressure force cannot generate circulation on a circuit of finite size. Nevertheless, pressure forces can compress or expand the circuit horizontally. Variations in the area of the circuit then induce rotation indirectly because of statement 6. To see this, suppose a material circuit that surrounds a fluid column is compressed horizontally. We note that the mass of the column is another quantity that is constant following the motion—in the absence of mass sources and sinks, no mass will leak into or out of the vertical column that the circuit encloses. Since the density is a constant throughout the fluid, and since the mass is constant following the motion, the volume of the column must be conserved following the motion of the fluid. Therefore, horizontal compression increases proportionally the height of the column.

At the same time, if the area of the material circuit is reduced, statement 6 shows that the vorticity within the circuit must increase to maintain a constant circulation. The increase in vorticity associated with horizontal compression and vertical stretching reflects local angular momentum conservation [see, e.g., Salmon (1998)].

As pointed out in the previous section, the depth of the fluid column may also change if the column passes over topographic features of the solid underlying surface. Columns passing over ridges obtain negative vorticity, and over valleys, positive vorticity, relative to their initial vorticity.

The interpretation of statement 6 in the presence of planetary rotation is fundamental to the study of large-scale GFD. Statement 6 holds for the absolute circulation, that is, the circulation measured from an inertial, nonrotating frame. An important mechanism for generating relative vorticity involves the constraint of having to conserve the absolute circulation as fluid moves north and south. Consider a ring of fluid, at rest in the rotating frame, of small surface area A, and located at latitude θ. From our earlier discussion of the rotating fluid on the sphere, the vorticity for this ring is the planetary vorticity $4\pi \sin\theta/T$ and the circulation for this ring is the planetary circulation $4\pi A \sin\theta/T$ for the right-hand oriented path around the circuit. If the fluid is free of friction and external applied stresses, then statement 6 tells us that the sum of the planetary and relative circulations will be constant following the motion of the ring. Thus, if the ring maintains its original area and is displaced northward, it will acquire a positive (cyclonic, counterclockwise in the Northern Hemisphere) relative circulation.

The final application of statement 6 we will consider concerns the effect of applied stresses and friction on the circulation. The proof of statement 6, which we have not detailed, shows how circulation may be generated under conditions that depart from the statement's assumptions. In particular, frictional and applied stresses, such as those found in atmospheric and oceanic boundary layers, can impart vorticity to a fluid by a mechanism known as *Ekman pumping*. Ekman pumping describes the way a frictional fluid boundary layer responds to interior circulation. For example, a cyclonic-relative circulation, i.e. one with positive relative vorticity, can cause upwelling in the boundary layer. This tends to decrease the depth of the fluid column, to reduce the relative vorticity, and therefore to counteract the positive circulation. On the other hand, applied stresses (such as atmospheric wind stress on the ocean) play a major role in generating large-scale oceanic circulation. These effects can be modeled in a simple way within the shallow-water system.

Just as the macroscopic circulation has a microscopic counterpart, namely the vorticity, the macroscopic statement of conservation of circulation (statement 6) has a microscopic counterpart.

Statement 7. In the absence of friction and applied stresses, the shallow-water PV (statement 4) is constant following the motion of the fluid.

To demonstrate statement 7, we apply the following argument, using statement 6 in the small-area limit (see Fig. 7):

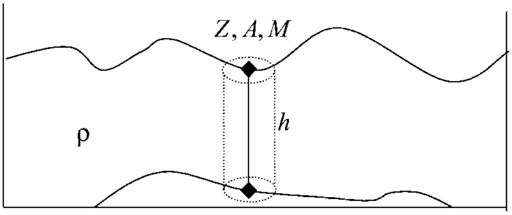


Figure 7 Demonstration of statement 7.

1. In the small-area limit, from statement 6, the circulation ZA is a constant of the motion, where Z is the vorticity and A is the area of the material circuit.
2. $ZA = (ZM)/(\rho h)$, where M is the mass of the fluid enclosed by the circuit, ρ the density, and h the depth.
3. Recall that M is constant following the motion and that ρ is a simple constant.
4. Thus, Z/h, which is the shallow-water PV (statement 4), is also a constant of the motion.

That PV is a constant following the motion is consistent with the example in which the vorticity and the height increased proportionally to the horizontal compression of the column. All the important mechanisms discussed in the context of definition 6—the roles of fluid column stretching and compression by the pressure force and topography, and the ability of planetary rotation to induce relative circulation as systems move north and south—can be interpreted in terms of the conservation of shallow-water PV (statement 7). The PV can be thought of as the planetary vorticity a fluid column would have if it were brought to a reference latitude and depth. This is the origin of the word *potential* in the name *potential vorticity.*

At the end of Section 2, we discussed an example that used the barotropic vorticity model, that is, a thin fluid of constant density and depth. For fixed depth, statement 6 is the same, but the PV is now, simply, the absolute vorticity. That is, in the absence of column stretching, the absolute vorticity is constant following the motion. This property was used to argue that particles disturbed by the propagating wave in the example would retain their initial values of absolute vorticity.

The final topic of this section concerns the dynamics of fluids with variable density. If density is not constant, it is possible for the pressure force to induce a net rotational torque. This will result in an important modification to the circulation theorem. To start with, we define a *barotropic* fluid to be one for which surfaces of constant density are aligned with surfaces of constant pressure. Put in another way, in a barotropic fluid, the density is a function only of the pressure, and not of the temperature or other thermodynamic variables.

Statement 8. In a barotropic fluid without friction or applied stresses, the circulation is constant following the motion of the fluid (Kelvin's circulation theorem).

Similarly to statement 6, statement 8 is proved by applying the momentum equations to a circuit. In a barotropic fluid, the net pressure force around a fluid element points through its center of mass, as in the shallow-water system, and the pressure force cannot exert a net rotational torque. A fluid that is not barotropic is said to be *baroclinic*. In baroclinic fluids, the net pressure force on a fluid element does not pass through the center of mass of the element. This results in a torque on the element that generates rotation, that is, vorticity, and, therefore, circulation. The impact of baroclinicity can be felt at small and large scales in the atmosphere and ocean. On scales of a few kilometers, the baroclinicity associated with the thermal contrast between land and sea generates sea breeze circulations. On planetary scales, it is the baroclinicity associated with the large-scale equator-to-pole temperature gradient that provides the source of energy for midlatitude atmospheric and oceanic eddies.

As for the shallow-water case, Kelvin's circulation theorem (statement 8) has a microscopic counterpart.

Statement 9. In the absence of friction, applied stresses, and applied heating, Ertel's PV, (statement 5), is constant following the motion of the fluid.

In order to justify statement 9, we need to discuss the concept of specific entropy in more detail. In thermodynamics, entropy is a thermodynamic variable that can be expressed as a function of other thermodynamic variables, such as temperature, pressure, and density. When heat is applied to a thermodynamic system, the change in entropy is related to the amount of heat transferred to the system. In the absence of heating, the entropy of the system does not change: the system is said to be *adiabatic*. In an adiabatic fluid, this implies that the entropy per unit mass, i.e., the specific entropy, is a constant of the motion. Therefore, surfaces of constant entropy (isentropic surfaces) are material surfaces, that is, surfaces that move with the fluid. In other words, the motion in an adiabatic fluid is along isentropic surfaces. Adiabatic fluids are relevant to GFD because large-scale atmospheric and oceanic flows are often approximately adiabatic.

Now, let us consider a simple fluid such as an ideal gas, for which entropy is a function of pressure and density. Suppose the fluid is adiabatic, so that the flow is along isentropic surfaces. On these surfaces, the pressure is a function of density. That is, the flow along isentropic surfaces in an adiabatic fluid is barotropic. This implies that statement 8 holds: The circulation is constant following the flow on an isentropic surface. Now, consider statement 8 applied in the small-area limit to two closely spaced isentropic surfaces (Fig. 8). The demonstration of statement 9 proceeds as follows:

1. In the small-area limit, from statement 8, the circulation ZA is a constant of the motion, where Z is the component of the vorticity normal to the isentropic surface and A is the area of each circuit in the small-area limit.

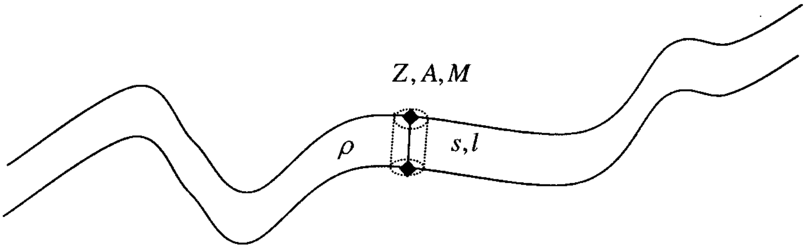


Figure 8 Demonstration of statement 8.

2. The entropy difference between the two isentropic surfaces, s, is a constant of the motion, since the entropy value of each surface is a constant of the motion.
3. Therefore, the product AZs is a constant of the motion.
4. The area $A = V/l$, where V is the volume enclosed by the material column and l is the perpendicular distance between the isentropic surfaces.
5. The volume $V = M/\rho$, where ρ is the density, and M is the mass of the column enclosed by the circuit and the two isentropic surfaces.
6. The product $MZs/(l\rho)$ is a constant of the motion, as is M. Thus, $(Z/\rho)(s/l)$ is a constant of the motion.
7. The quantity s/l represents the change of entropy per unit distance between the isentropic surfaces, and therefore represents the specific entropy gradient.
8. Therefore, the quantity $(Z/\rho)(s/l)$ is Ertel's PV (statement 5) and is a constant of the motion. This demonstrates statement 9.

The dynamics of Ertel's PV incorporate similar mechanisms to the one discussed for shallow-water PV dynamics, with the fluid depth replaced by the distance between isentropic surfaces. It is very useful to have simpler model systems, such as the shallow-water model, with which to build up our intuition concerning PV dynamics. For example, there is a strong analogy between changes in thickness of the fluid column brought about by mass sources and sinks in the shallow-water model and thermally induced changes in the distribution of isentropic surfaces.

5 CONCLUSION

Potential vorticity is a powerful unifying tool in the atmospheric and oceanic sciences because it combines apparently distinct factors, such as topography, stratification, relative vorticity, and planetary vorticity, into a single dynamical quantity that is conserved, following the motion, like a chemical tracer. Indeed, much current work in the field is characterized by "PV thinking:" how PV is generated, maintained, converted from one form to another, transported, and dissipated. To conclude, there follows a list of a few important topics of current interest in which PV dynamics is central. Hoskins et al. (1985), Salmon (1998), and Andrews et al. (1987) are excellent starting points for further investigation.

- *Rossby-Wave Dynamics* The primary large-scale oscillations in the extratropical atmosphere and oceans are Rossby waves, which are supported by the PV gradients largely associated with planetary rotation. In the "billiard-ball planet" example discussed in Section 2, it is Rossby waves that carry the signal of the wave maker away from the equator. Rossby-wave dynamics underlies our understanding of the midlatitude remote response to El Niño events in the tropical Pacific, of the propagation of upper tropospheric disturbances, and of the dynamics of large-scale waves in the stratosphere (see Chapter 4).
- *Baroclinic Instability* The development of atmospheric cyclones and of oceanic eddies can be understood in terms of the interactions of regions of opposite-signed PV gradients (see Chapter 4).
- *Geostrophic Turbulence* Our understanding of atmosphere/ocean large-scale turbulence is framed by PV, which behaves as an "active" tracer that can be mixed and diffused downgradient (see Chapter 6).
- *Balance Models* Balance models, such as the quasi-geostrophic model (Chapter 4, Swanson), are filtered equations that model the slow, large-scale motions associated with Rossby waves, baroclinic instability, and geostrophic turbulence. These models are based on some of the dynamical regimes mentioned above; for example, the assumption that the vertical component of relative vorticity is small compared to the vertical component of planetary vorticity. In these models, the PV is the sole dynamical variable, and all other variables can be obtained from the PV by an inversion procedure. This inversion procedure is analogous to the one involved in obtaining the circulation from the vorticity distribution (see Section 2).
- *Rossby-Wave Activity Propagation and Wave-Induced Circulations* Recall that the southward flux of vorticity in the barotropic vorticity model example at the end of Section 2 gives rise to a westward acceleration. More generally, the flux of PV in both the shallow-water model and fluids with variable density can be associated with a wave-induced torque, generally in the direction transverse to the flux. In this way, the existence of persistent extratropical jets, such as the atmospheric jet stream and the jets in the Antarctic Circumpolar Current, can be understood in terms of the flux, often downgradient, of PV. These fluxes are also the starting point for a theory of "wave activity" propagation that is the foundation of models of the overturning circulation in the stratosphere.
- *Symmetries, Conservation Laws, and Hamiltonian Structure* The existence of PV as a constant of the motion has a profound connection to the mathematical structure of the equations of motion. All conserved quantities for a physical system—spatially distributed quantities such as energy and momentum, and local quantities such as PV and wave activity, are connected to symmetries in the system through Noether's theorem. This idea is the starting point for Hamiltonian fluid dynamics, which is reviewed in Salmon (1998). The symmetry connected to the PV is the so-called particle relabeling

symmetry, which reflects the fact that the dynamics of an adiabatic fluid is unaffected by changing the particle labels on isentropic surfaces. Such considerations provide a theoretical basis for wave activity invariants and generalized nonlinear stability theorems for GFD, as well as for the systematic derivation of balance models with desirable symmetry.

6 APPENDIX: MATHEMATICAL SUMMARY

In this section, we provide a summary, using the notation of vector calculus, of the main results. First, the circulation, *C*, in statement 1, may be represented by a line integral

$$C = \oint_{\text{circuit}} |\mathbf{u}| \cos\theta \, ds$$

where **u** is the velocity measured from the inertial frame. In this notation, the length of the circuit is $L = \oint_{\text{circuit}} ds$ and the average of the tangential flow component is C/L. This is consistent with statement 1. By the Stokes theorem, the circulation also satisfies

$$C = \iint_{\text{region}} \text{curl } \mathbf{u} \cdot \hat{\mathbf{n}} \, dA$$

where the double-integral sign and the word *region* indicate that a flux integral is being taken over the region bounded by the circuit, and $\hat{\mathbf{n}}$ is a unit vector perpendicular to the surface. The vector curl of the velocity is, in Cartesian coordinates,

$$\text{curl } \mathbf{u} = \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}, \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}, \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right)$$

Statement 2 in the small-area limit implies that the vorticity, $\boldsymbol{\zeta}$, is equal to the curl of the velocity:

$$\boldsymbol{\zeta} = \text{curl } \mathbf{u}$$

Statement 3 is therefore equivalent to the Stokes's theorem for finite-area surfaces.

For the shallow-water system, the vorticity is vertical:

$$\boldsymbol{\zeta} = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) \mathbf{z}$$

and the potential vorticity, statement 4, is

$$q = \frac{(\partial v/\partial x) - (\partial u/\partial y)}{h - h_b}$$

where h is the height of the fluid surface and h_b is the height of the solid surface beneath the fluid. Ertel's PV, statement 5, is

$$q = \boldsymbol{\zeta} \cdot \frac{\nabla s}{\rho}$$

where s is the specific entropy and ρ is the density.

The circulation theorem for the shallow-water model, statement 6, and Kelvin's circulation theorem, statement 8, are written

$$\frac{DC}{Dt} = 0$$

where D/Dt is the material derivative following the motion, defined by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla$$

and C is the circulation of a material circuit. The conservation of PV following the motion for shallow-water (statement 7) or three-dimensional stratified flow (statement 9) is

$$\frac{Dq}{Dt} = 0$$

REFERENCES

Andrews, D. G., J. R. Holton, and C. B. Leovy (1987). *Middle Atmosphere Dynamics*, Orlando, Academic.

Cotton, W. R., and R. A. Anthes (1989). *Storm and cloud dynamics*, San Diego, Academic.

Gill, A. E., (1982). *Atmosphere-Ocean Dynamics*, San Diego, Academic.

Holton, J. R. (1992). *An Introduction to Dynamic Meteorology*, 3rd Edition, San Diego, Academic.

Hoskins, B. J., M. E. McIntyre, and A. W. Robertson (1985). On the use and significance of isentropic potential vorticity maps, *Q. J. Roy. Met. Soc.* **111**, 877–946.

Kundu, P. K. (1990). *Fluid Mechanics*, San Diego, Academic.

Pedlosky, J. (1996). *Ocean Circulation Theory*, Berlin, Springer.

Salmon, (1998) *Lectures on Geophysical Fluid Dynamics*, New York, Oxford University Press.

CHAPTER 4

EXTRATROPICAL ATMOSPHERIC CIRCULATIONS

KYLE SWANSON

1 INTRODUCTION

As one moves poleward in Earth's atmosphere, the character of the circulation changes markedly. At large scales and low frequencies, tropical circulations are fundamentally ageostrophic, epitomized by the thermally direct Hadley, Walker, and monsoon circulations. In contrast, at large scales and low frequencies in the extratropics, the effect of Coriolis accelerations due to Earth's rotation become important, and circulations are approximately in geostrophic balance. As such, away from the surface, the extratropical large-scale wind satisfies

$$(u, v, w) \approx (u_g, v_g, 0) \equiv \lfloor -a^{-1}\partial_\phi \psi, (a\cos\phi)^{-1}\partial_\lambda \psi, 0 \rfloor \tag{1}$$

where the geostrophic stream function is given by:

$$\psi \equiv f_0^{-1}(\Phi - \Phi_0) \tag{2}$$

Here $\Phi_0(z)$ is a suitable reference geopotential profile and $f_0 = 2\Omega \sin\phi_0$ is the Coriolis parameter evaluated at some extratropical latitude ϕ_0. The extent to which geostrophic balance adequately describes the extratropical circulation is measured by the smallness of the Rossby number, $\mathrm{Ro} = U/(f_0 L)$, which is approximately 0.1 for velocity and length scales $U \sim 10\,\mathrm{m/s}$ and $L \sim 1000\,\mathrm{km}$ that, respectively, characterize large-scale extratropical dynamical motions.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

In addition to geostrophic balance, the extratropical atmospheric circulation is also in hydrostatic balance. Hydrostatic balance provides a relation between the potential temperature and mass fields of the form

$$\theta_e \equiv \theta - \theta_0(z) = HR^{-1} f_0 e^{\kappa z/H} \partial_z \psi \tag{3}$$

where $\theta_0(z) = HR^{-1} f_0 e^{\kappa z/H} \partial_z \Phi_0$ is a reference potential temperature. The combination of hydrostatic and geostrophic balance, through cross differentiation of (1) and (3), yields the thermal wind relations that link the vertical shear of the geostrophic wind to horizontal potential temperature gradients. Therefore, the jets that dominate the zonally averaged tropospheric flow have their origin in the density contrasts of the equator–pole temperature gradient and are referred to as baroclinic.

Approximate geostrophic balance, along with the existence of a robust zonally averaged flow provide the basis for theoretical inquiries into the dynamics of observed large-scale motions in the extratropical atmosphere. The former allows for a simpler description of the underlying dynamics, while the latter allows for meaningful insights into dynamical phenomena to be gained through the study of linear fluctuations about a zonal basic state flow. In this chapter, we apply these simplifications to examine the dynamics of several important classes of large-scale motions in the extratropical atmospheric circulation.

2 QUASI-GEOSTROPHIC THEORY

Geostrophy by itself does not provide an adequate description of large-scale dynamical motions, as it is a time-independent diagnostic relation between the mass and velocity fields. To understand how flows in approximate geostrophic balance evolve in time, it is necessary to include the important effects of $O(\mathrm{Ro})$ ageostrophic winds, which includes the vertical wind. The quasi-geostrophic (QG) approximation includes the effects of these ageostrophic winds on the $O(1)$ geostrophic fields in a simplified but consistent manner and provides a concise, idealized system with which to study large-scale extratropical dynamical motions.

The QG approximation has its roots in the geometrical simplification of replacing the horizontal spherical coordinates (λ, ϕ) by the eastward and northward Cartesian coordinates (x, y) in the full primitive equations and restricting the flow domain to some neighborhood of the latitude ϕ_0. Here x is eastward distance and y northward distance from some origin (λ_0, ϕ_0). This geometric simplification allows the QG dynamical approximations to be made with full rigor (Pedlosky 1987, Section 6.2-5) and once made provides the important conceptual simplification that the *only* effect of Earth's sphericity is the variation of the Coriolis parameter f with latitude,

$$f = f_0 + \beta y \tag{4}$$

where $\beta = 2\Omega a^{-1} \cos \phi_0$. This so-called *β-plane approximation* captures the most important dynamical effect of the variation of *f* with latitude. Once this simplification is made, the QG dynamical equations naturally emerge from an expansion of the equations of motion in the small parameter Ro.

In the absence of forcing and dissipation,* QG dynamics are described by the material conservation of quasi-geostrophic potential vorticity (QGPV) by the horizontal geostrophic flow

$$(\partial_t + u_g \partial_x + v_g \partial_y) q = 0 \tag{5}$$

where

$$\begin{aligned} q &= \zeta_g + f_0 \rho_0^{-1} \partial_z \lfloor \rho_0 \theta_e (\partial \theta_0 / \partial z)^{-1} \rfloor + f_0 + \beta y \\ &= \partial_{xx} \psi + \partial_{yy} \psi + \rho_0^{-1} \partial_z (\rho_0 \varepsilon \partial_z \psi) + f_0 + \beta y \end{aligned} \tag{6}$$

is the QGPV, ζ_g is the vertical component of the relative geostrophic vorticity, and $\varepsilon(z) \equiv f_0^2 / N^2(z)$. QG dynamics are layer-wise equivalent to a shallow water system insofar as variations in the vertical only appear implicitly in the elliptic nature of the q-ψ relation. Although ageostrophic circulations do not explicitly occur in the QGPV evolution equation, these circulations are vital to QG dynamics and may be obtained diagnostically from the geostrophic mass field and its time evolution (Holton, 1992, Section 6.4).

This system is completed by the specification of upper and lower boundary conditions. The lower boundary condition over a rigid surface is

$$(\partial_t + u_g \partial_x + v_g \partial_y)(\partial_z \psi + N^2 h_T / f_0) = 0 \quad \text{at} \quad z = 0 \tag{7}$$

where $h_T(x, y)$ is the bottom topography. Recalling the connection between the vertical derivative of the geostrophic stream function and the perturbation potential temperature (3), condition (7) simply describes the material conservation of potential temperature in the geostrophic wind at the lower boundary. If there is a rigid upper boundary, another condition of the form (7) is applied there. In the atmosphere, this condition is replaced by the requirement that either the perturbation energy density decay as $z \to \infty$, or alternatively, that the energy flux be directed upward as $z \to \infty$.

The robust extratropical zonally averaged flow suggests studying small-amplitude fluctuations about that flow. Dividing the stream function into a zonally symmetric background and a perturbation,

$$\psi(x, y, z, t) = \psi_0(y, z) + \psi'(x, y, z, t) \tag{8}$$

*Forcing and dissipation are neglected not because they are unimportant to large-scale atmospheric dynamical motions, but because their inclusion is not necessary to understand the dynamical features examined herein. However, in general they will have important modifying effects in any given circumstances and will be crucial in the study of any steady-state motions.

the evolution of the perturbation QGPV is described by

$$(\partial_t + u_{g0}\partial_x)q' + v'_g q_{0y} + \mathbf{u}'_g \cdot \nabla q' = 0 \tag{9}$$

Here $q' = \partial_{xx}\psi' + \partial_{yy}\psi' + \rho_0^{-1}\partial_z(\rho_0 \varepsilon \partial_z \psi')$; $\mathbf{u}'_g = (u'_g, v'_g) = (-\partial_y \psi', \partial_x \psi')$; $u_{g0} = -\partial_y \psi_0$; and

$$q_{0y} = \beta - \partial_{yy}u_{g0} - \rho^{-1}\partial_z\big(\rho\varepsilon\partial_z u_{g0}\big) \tag{10}$$

is the background QGPV gradient. The perturbation (rigid surface) boundary condition is

$$(\partial_t + u_{g0}\partial_x)(\partial_z\psi' + N^2 h'/f_0) + v'_g \eta_{0y} + \mathbf{u}'_g \cdot \nabla(\partial_z \psi') = 0 \quad \text{at} \quad z = 0 \tag{11}$$

where

$$N^2 f_0^{-1}\partial_y[h_T] - \partial_z u_{g0} \tag{12}$$

is proportional to the basic state potential temperature gradient measured along the sloping lower boundary, and the brackets indicate a zonal average.

We are free to seek wavelike solutions to Eqs. (8) and (9) of the form,

$$\psi'(x, y, z, t) = \Psi(y, z)e^{ik(x-ct)} \tag{13}$$

where k is the (non-negative) wavenumber in the x (zonal) direction and c is the complex phase speed, related to the frequency $\omega = kc$. Substituting in solutions of this form and neglecting terms that are quadratic in perturbation quantities yields the equation

$$(u_{g0} - c)\big(\rho^{-1}\partial_z\rho\varepsilon\partial_z + \partial_{yy} - k^2\big)\Psi + q_{0y}\Psi = 0 \tag{14}$$

with the lower boundary condition

$$(u_{g0} - c)\partial_z\Psi + \eta_{0y}\Psi = 0 \quad \text{at} \quad z = 0 \tag{15}$$

along with the appropriate upper boundary condition for the problem at hand. Jumps in N or the shear, common at the tropopause, may be included in (14) by applying analytically derived jump conditions.

The linear, homogeneous system [Eqs. (14) and (15)] defines an eigenvalue problem for c. A particularly illuminating case is where u_{g0} and N are constant, $\rho = \rho_0 \exp(-z/H)$, and the topography h_T vanishes. In this situation, solutions to (14) may be sought of the form

$$\Psi = A\exp(imz)\exp\Big(\frac{z}{2H}\Big)\exp(ily) \tag{16}$$

where the complex constant A may be adjusted to meet the boundary condition (15). For this particular solution, the zonal phase speed is given by

$$c = \frac{\omega}{k} = u_{g0} - \frac{\beta}{k^2 + l^2 + \varepsilon(m^2 + \frac{1}{4}H^2)} \tag{17}$$

Waves of this type are called *Rossby waves* and provide an archetype for low-frequency ($\omega < f_0$) motions in the extratropics. Rossby waves differ from other, more familiar wave types such as gravity waves, since it is the planetary vorticity gradient β that acts as a restoring mechanism for the wave. An illuminating discussion of the vorticity gradient as a restoring mechanism is found in Pedlosky (1987, Section 3.16).

Rossby waves have a number of remarkable properties. First, the zonal phase speed is everywhere westerly relative to the basic state zonal flow u_{g0}. However, as the total wavenumber $K^2 = k^2 + l^2 + \varepsilon(m^2 + 1/4H^2)$ gets larger, Rossby waves move more and more nearly with the basic flow u_{g0}. These properties are entirely consistent with observed large-scale atmospheric waves. Rossby waves are also dispersive, i.e., the group velocity $\mathbf{c}_g \equiv (\partial\omega/\partial k, \partial\omega/\partial l, \partial\omega/\partial m)$ that describes how a packet of Rossby waves moves with time differs from the phase velocity of the waves comprising that packet in both magnitude and direction. Since information travels with the group velocity, rather than the phase velocity, the group velocity also describes the propagation of the energy of the packet. The difference between the Rossby wave phase and group velocities is most apparent for vertical propagation, since the vertical group velocity and phase velocity are oppositely directed. If the lines of constant phase propagate upwards (downwards), the energy in the wave propagates downwards (upwards).

While Rossby waves provide a nontrivial example of nearly geostrophic motions in the extratropics, the actual extratropical flow situation is significantly more complicated than that considered above. Specifically, vertical shear and associated meridional temperature gradients allow for the existence of instabilities that can amplify with time as well as propagate. The characteristics of such instabilities, along with their importance in the extratropical atmospheric circulation form the topic of the next section.

3 BAROCLINIC INSTABILITY AND FRONTOGENESIS

The most striking observational feature of the extratropical atmosphere is the spectrum of vigorous synoptic-scale transient eddies that provide much of the day-to-day variability in the middle latitude weather. These eddies arise from the release of potential energy stored in the equator–pole temperature gradient. The release of this potential temperature is marked by a down gradient flux of heat from equator to pole, leading to an equator–pole temperature gradient that is significantly smaller than radiative processes would provide acting alone. Curiously, this heat flux can be linked to the growth of transient eddies through the budget for the perturbation eddy energy

$$E' = \int\int\int \frac{1}{2}\rho[u_g'^2 + v_g'^2 + \varepsilon(\partial_z\psi')^2]\,dx\,dy\,dz \tag{18}$$

The equation governing E' is obtained by multiplying (9) by ψ', integrating over all space, and employing the boundary condition (11). This equation can be written

$$\frac{dE'}{dt} = \int\int \mathbf{Q}\cdot\nabla u_{g0}\,dy\,dz \tag{19}$$

where the **Q** vector is given by

$$\mathbf{Q} = (Q_y, Q_z) \equiv (-\rho\overline{u_g'v_g'}, \rho\varepsilon\overline{v_g'\partial_z\psi'}), \tag{20}$$

and the overbar denotes an integral over x. It has been assumed that the vertical flux terms vanish by a radiation condition as $z \rightarrow \infty$, and the analogous lateral flux terms vanish by virtue of meridional confinement in a channel, or equivalently, by decay of perturbations at large y.

The integral in (19) represents the conversion of energy between the background flow and perturbations. In the integrand, $Q_z\partial_z u_{g0}$ is the baroclinic conversion term, and represents energy exchange with the potential energy of the basic state. By thermal wind balance, positive vertical shear implies colder air toward the pole; and, in this situation, perturbations with a poleward heat flux tap the baroclinic energy and grow. This term releases potential energy by reducing the tilt of the mean isentropes.

The term $Q_y\partial_y u_{h0}$ is the barotropic conversion term and represents exchanges with the basic state kinetic energy. In the extratropics, it typically acts as an energy sink. Synoptic eddies thus act as an intermediary, transforming potential energy due to differential heating into kinetic energy of barotropic jets.

For any basic state flow with vertical or horizontal shear, an initial perturbation can be crafted that will extract energy and begin to grow. However, as time passes, the structure of the perturbation will generally change, and growth may cease or even turn to decay. Stability theory seeks to identify energy-extracting structures that can persist long enough to amplify by a significant amount and to understand what basic state flow circumstances allow for such growth. This is elegantly discussed with reference to the wave action budget of the perturbation. If we linearize (9), multiply by q', and average over x, we obtain

$$\partial_t A + \nabla\cdot\mathbf{Q} = 0 \tag{21}$$

where the wave action (or more properly the pseudomomentum) A is

$$A = \frac{1}{2}\rho\overline{q'^2}q_{0y}^{-1} \tag{22}$$

An analogous budget for the variance of temperature at the ground can be obtained by multiplying (11) by $\rho\partial_z\psi'$ and averaging in x. Doing so yields

$$\partial_t B + Q_z(0) = 0 \quad \text{at} \quad z = 0 \tag{22}$$

where

$$B = \frac{1}{2}\rho\varepsilon\overline{(\partial_z\psi')^2}\eta_{0y}^{-1} \tag{23}$$

Integrating (21) over space, assuming boundary conditions that make the horizontal flux terms vanish, and using (22) to reduce the bottom flux contribution yields the following relation for the total pseudomomentum $\prod$:

$$\frac{d}{dt}\prod = 0 \qquad \prod = \int\int A\,dy\,dz + \int B\,dy \tag{24}$$

This conservation relation yields the desired stability criterion. Specifically, if q_{0y} and η_{0y} each have uniform sign throughout the domain, have the same sign as each other, and the magnitudes of q_{0y} and η_{0y} are finite and bounded away from zero, then the system is stable in the sense that the perturbation enstrophy

$$Z' = \int\int \frac{1}{2}\rho\overline{q'^2}\,dy\,dz \tag{25}$$

remains small for all times if it is initially small. If this criterion is not met, sustained exponential disturbance growth is possible. However, such growth is not inevitable, as (24) provides a necessary rather than a sufficient condition for disturbance growth.

The archetypal scenario permitting baroclinic instability in the extratropics is q_{0y} greater than zero everywhere, positive vertical shear at the ground, and a sufficiently flat lower boundary to keep η_{0y} negative. This condition is almost ubiquitously satisfied throughout the extratropical troposphere in both hemispheres. In this situation, a growing disturbance can be regarded as a coupled QGPV/surface temperature motion. The QGPV anomaly aloft induces a motion that stirs up a surface temperature anomaly, and the surface temperature anomaly in turn further stirs up the QGPV aloft. The influence depth of a QGPV or surface temperature perturbation of length L is the "deformation depth" f_0L/N provided this is not much in excess of the density scale height H. If a QGPV anomaly is at altitude D, motions with

$L \ll ND/f_0$ will have QGPV dynamics uncoupled from the surface dynamics, and cannot be expected to cohere and grow. Thus, short-wave instabilities must also be shallow, and deep modes filling out a density scale height must have horizontal scales comparable to the radius of deformation $L_d = NH/f_0$, if not longer. These arguments are quite general and carry over to strongly nonlinear motions.

The simplest possible model that satisfies the necessary conditions for instability in a continuous atmosphere is the Eady model. The Eady model has $\rho = \text{constant}$, $\beta = 0$, $\partial_z u_{g0} = \Lambda = \text{constant}$, and rigid lids at $z = 0$ and H_T. While this scenario provides only a crude model of the atmosphere, it allows for insights into the dependence of a growing disturbance on horizontal scale and stability. The primary simplification is the vanishing potential vorticity gradient in the interior of the fluid. Since the QGPV is zero initially in the interior, it will remain so for all time. Therefore, all the dynamics involve the time-dependent boundary condition (11) applied at each boundary. Despite the vanishing of the QGPV gradient, the Eady model satisfies the necessary conditions for instability because vertical shear of the basic state flow at the upper boundary provides an additional term in (24) that is equal and opposite to the lower boundary integral.

For this situation, (11) and the condition that q' vanish in the interior lead to wave solutions on each boundary (considered in isolation), whose effective restoring force is the meridional temperature gradient. Such waves are called Rossby edge waves, where edge waves are boundary-trapped disturbances. The wave at the lower boundary propagates eastward (relative to the surface flow), and the wave at the lid propagates westward. Localized warm surface air is associated with low pressure and cyclonic flow, and is effectively a positive QGPV perturbation. Cold air at the surface features high pressure and anticyclonic flow, analogous to a negative QGPV perturbation.

Baroclinic instability occurs in this model when the waves at each boundary amplify the temperature at the opposite boundary. For this to happen, the "deformation depth" defined above must be at least comparable to the distance between the boundaries. If the disturbance horizontal length scale is too short, or if the boundaries are too far apart, unstable modes do not exist.

Figure 1 shows the disturbance structure for a wave with equal zonal and meridional wavenumber ($k = l$). For this situation, the wavelength of the perturbation that grows fastest is

$$L_m = \frac{2\sqrt{2}\pi L_d}{(H_T \alpha_m)} \approx 4000 \text{ km} \tag{26}$$

similar to the length scale of observed synoptic-scale transients. The growth rate of the instability is approximately given by

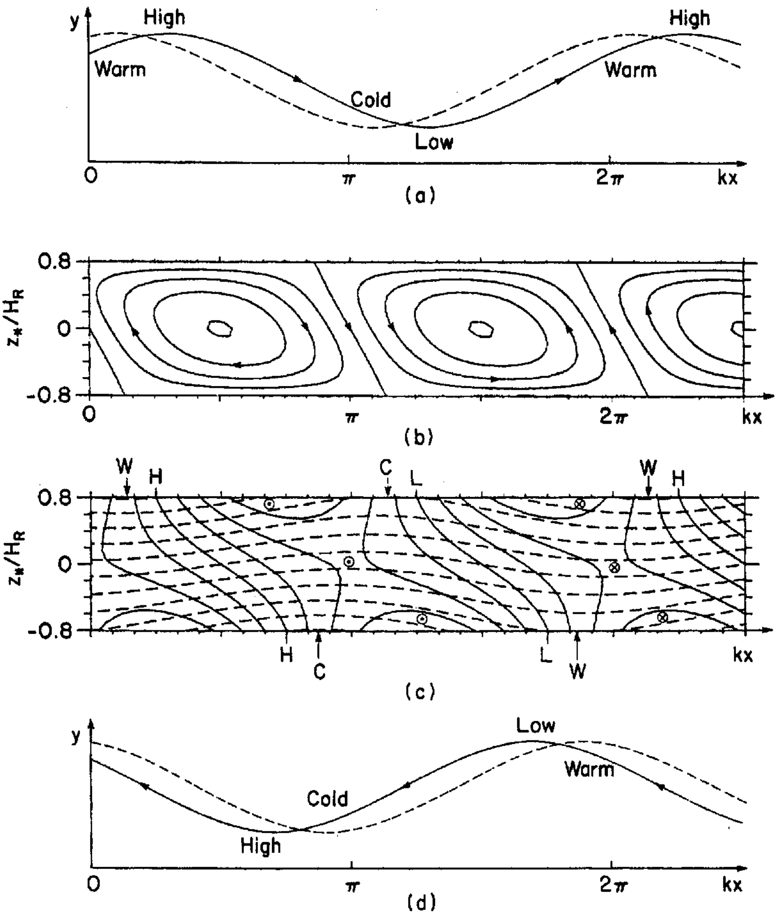


Figure 1 Properties of the most unstable Eady wave. (*a* and *d*) Geopotential height and temperature patterns on the upper and lower boundaries, respectively. (*b*) The stream function for the ageostrophic flow, showing the overturning in the *x*-*z* plane. Ascent is just to the east of the surface trough and descent is just to the west of the surface ridge. (*c*) Contours of meridional geostrophic wind v'_g (solid) and isentropes (dashed) in the *x*-*z* plane. *H* and *L* indicate the ridge and trough axes; *W* and *C* denote the phase of the warmest and coldest air. Circles with enclosed crosses indicate the axis of maximum poleward flow, and circles with enclosed dots indicate the axis of maximum equatorward flow (from Gill, 1982).

$$\mathrm{Im}(\omega) \approx \frac{0.25 f_0 \partial_z u_{g0}}{N} \tag{27}$$

This leads to an *e*-folding time of 2 or 3 days for typical extratropical values of the shear and *N*, again consistent with observed synoptic eddy growth rates. Figure 1*b* shows that for this particular wavelength, as for all unstable baroclinic waves, the trough and ridge axes slope westward with height. This phase displacement allows the velocity and temperature perturbations to be partly in phase at each boundary, thus leading to a poleward heat flux. The axes of the warmest and coldest air, however, tilt eastward with height. East of the trough, where the perturbation meridional velocity is poleward, Figure 1*c* shows that the ageostrophic vertical motion is upward. Thus, parcel motion is poleward and upward in the region where temperature perturbations are positive, and vice versa in regions where the temperature perturbations are negative. Through condensation in ascending, adiabatically cooling trajectories, this vertical velocity pattern provides a link to the hydrological cycle in the extratropical troposphere and, in particular, provides the "comma cloud" signature of surface troughs.

The scaling of the growth rate with N in (27) suggests that baroclinic instability would produce a statically stable extratropical atmosphere on Earth even in the absence of moisture. Local radiative convective equilibrium would drive N to zero but would have a strong equator–pole temperature gradient. Such a flow would be violently unstable to baroclinic instability. It is presumed that baroclinic instability increases N so as to adjust toward a more stable state. Eddies could do this by taking cold air manufactured in polar regions and sliding it under warmer air, and vice versa for warm air produced near the tropics. Either by vertical heat fluxes that increase N or by horizontal heat fluxes that reduce the equator–pole temperature gradient (and hence the vertical shear), baroclinic instability acts to adjust the extratropical flow toward a state that is neutrally stable to baroclinic instability.

In addition to their role in the tropospheric general circulation, synoptic-scale eddies also drive weather-related events at smaller scales. It has long been understood that extratropical cyclones have cold frontal regions in which the temperature typically changes rapidly, often accompanied by shifts in the wind and by precipitation. In general, these fronts are considered to form in a growing nonlinear baroclinic wave, though this secondary role in no way diminishes their importance in practical meteorology. Rather, the theoretical emphasis shifts toward the mechanism for generating fronts, namely frontogenesis.

The basis of frontogenesis lies in the advection of surface temperature (7). The kinematics of frontogenesis can be understood by considering a certain class of horizontal velocity fields called *deformation fields*. Such deformation fields, which are *local* features of large-scale horizontal wave motions, contain confluent regions that tend to concentrate a large-scale preexisting temperature gradient, squeezing isotherms together. A prototypical deformation field is given by $u = -\gamma x$, $v = \gamma y$, where γ is a constant with dimension (time)$^{-1}$. Suppose the potential temperature field is initially oriented so the isotherms are parallel to the y axis. Then at the ground, where w vanishes, the time evolution of the potential temperature must satisfy

$$\partial_t \theta = -u \partial_x \theta = \gamma x \partial_x \theta \tag{28}$$

since θ is independent of y. The solution to (28) is easily verified to be

$$\theta(t) = \theta_0(x e^{\gamma t}) \tag{29}$$

where $\theta_0(x)$ is the surface distribution of potential temperature at the initial instant. The temperature gradient at the surface is

$$\partial_x \theta = e^{\gamma t} \dot{\theta}(x e^{\gamma t}) \tag{30}$$

where the dot indicates derivative with respect to the argument. Therefore, the surface temperature gradient increases exponentially with time in regions of confluence in the deformation field.

Naturally, as time goes by and the temperature field changes as a result of this confluence, the changing thermal wind in the y direction will produce, by Coriolis accelerations, flow in the x direction that will alter the initial deformation velocity field. Thus, the above solution will be valid only initially and even then describes the structure of the temperature gradient at the ground. However, this solution does show that large-scale flow features can generate strong density gradients.

A more complete theoretical explanation of frontogenesis lies outside of the class of motions described by quasigeostrophic theory. This is because fronts are inherently *anisotropic*, with a very narrow frontal zone compared to the spatial extent along the front. A different scaling of the primitive equations, namely the semigeostrophic theory originally developed by Hoskins and Bretherton, explicitly recognizes this anisotropy and provides a more complete dynamical theory for the formation of fronts [see Hoskins (1982) for a review]. While such a theory lies beyond the scope of our treatment, it does present a number of features absent in the deformation model of frontogenesis examined above. Foremost among these features is that the frontogenetic mechanism is so powerful in semigeostrophic theory that infinite temperature gradients can be generated in a finite amount of time. Of course, in reality instabilities can be expected to occur in regions of such gradients, and accompanying mixing of surface temperature will tend to limit the sharpness of the frontal zone. However, the prediction of a discontinuity indicates the vigor of the frontogenetic process.

4 STATIONARY PLANETARY WAVES

Another striking component of the observed extratropical atmospheric circulation are the large-scale stationary planetary (Rossby) waves that dominate the zonally asymmetric flow during the Northern Hemisphere winter. Flow over larger-scale surface features (e.g., the Tibetan plateau), along with longitudinally varying diabatic heating force these waves, which subsequently propagate zonally and meridionally throughout the troposphere, and vertically into the upper atmosphere. In the troposphere, these waves influence the regional climates of Earth, as they affect the position and strength of the midlatitude storm tracks. In the stratosphere, breaking vertically propagating Rossby waves instigate strong changes in the stratospheric flow, called sudden warmings (Andrews et al., 1987, Section 6). More generally, the breaking of these vertically propagating waves plays an important role in the overall exchange of air between the stratosphere and the troposphere, as this breaking forces planetary-scale vertical ascent and descent of air (Holton et al., 1995).

First, let us consider the zonal and meridional propagation of stationary planetary waves. This problem is most easily treated using a barotropic model on the sphere, linearized about a zonally symmetric but latitudinally varying mean flow. In this model, stationary waves are solutions of the equation

$$u_{g0}(a\cos\phi)^{-1}\partial_\lambda \zeta'_g + v'_g(a^{-1}\partial_\phi \zeta_g + \beta) = -u_{gs} f(h_0 a\cos\phi)^{-1}\partial_\lambda h_T - r\zeta' \qquad (31)$$

where $\zeta_g \equiv -a^{-1}\partial_\phi u_g + (a\cos\phi)^{-1}\partial_\lambda v_g$ is the relative perturbation vorticity, primes indicate deviation from zonal mean background states, and h_0 is the resting depth of the barotropic fluid. The damping coefficient r is taken to be independent of latitude.

The response of this system with h_T equal to the Northern Hemisphere topography is shown in Figure 2, where u_{g0} and u_{gs} equal the zonally averaged Northern Hemisphere wintertime 300 mb and surface winds, respectively. The response field appears to be comprised of two wavetrains propagating equatorwards, one produced by the Rockies and the other by the Tibetan plateau.

These structures can be explained by Rossby wave ray-tracing theory. If we transform the linearized, unforced inviscid vorticity equation into Mercator coordinates $x = a\lambda$, $a^{-1}dy = (\cos\phi)^{-1}d\phi$, then (30) takes the form

$$\hat{u}\partial_x(\partial_{xx} + \partial_{yy})\psi' = -\hat{\beta}\partial_x\psi' \qquad (32)$$

where $\hat{u} \equiv u_{g0}(\cos\phi)^{-1}$ and $\hat{\beta} \equiv \cos\phi(\beta + a^{-1}\partial_\phi\zeta_{g0})$. For a wave of the form $\psi' = \mathrm{Re}\,\tilde{\psi}\exp(ikx)$, one has

$$\partial_{yy}\tilde{\psi} = (k^2 - \hat{\beta}/\hat{u})\tilde{\psi} \qquad (33)$$

Given a source localized in latitude, the structure of (32) requires that zonal wavenumbers $k > k_s \equiv (\hat{\beta}/\hat{u})^{1/2}$ remain meridionally trapped in the vicinity of the source, while wavenumbers $k < k_s$ are free to propagate away from the source.

A source localized in longitude as well as latitude can be regarded as producing two rays for each $n < n_s$, where $n = ak$ is the zonal wavenumber on the sphere, and $n_s = ak_s$. These rays correspond to the two possible meridional wavenumbers, $al = \pm(n_s^2 - n^2)^{1/2}$. Since the meridional group velocity of nondivergent Rossby waves $c_{gy} = \partial\omega/\partial l = 2\hat{\beta}kl(k^2 + l^2)^{-2}$ has the same sign as l, the positive (negative) sign corresponds to poleward (equatorward) propagation. The zonal group velocity $c_{gx} = \partial\omega/\partial k = u_{g0} + \hat{\beta}(k^2 - l^2)(k^2 + l^2)^{-2}$ is eastward for stationary planetary waves with $k \sim l$. Therefore, the rays can be traced downstream from their source, noting that the zonal wavenumber remains unchanged because the mean flow is independent of x, whereas the meridional wavenumber adjusts to satisfy the local dispersion relation. The ray path is tangent to the vector group velocity (c_{gx}, c_{gy}) for a particular k and l, and is always refracted toward the larger "refraction index" $(n_s^2 - n^2)^{1/2}$, i.e., rays turn toward larger n_s in accordance with Snell's law for

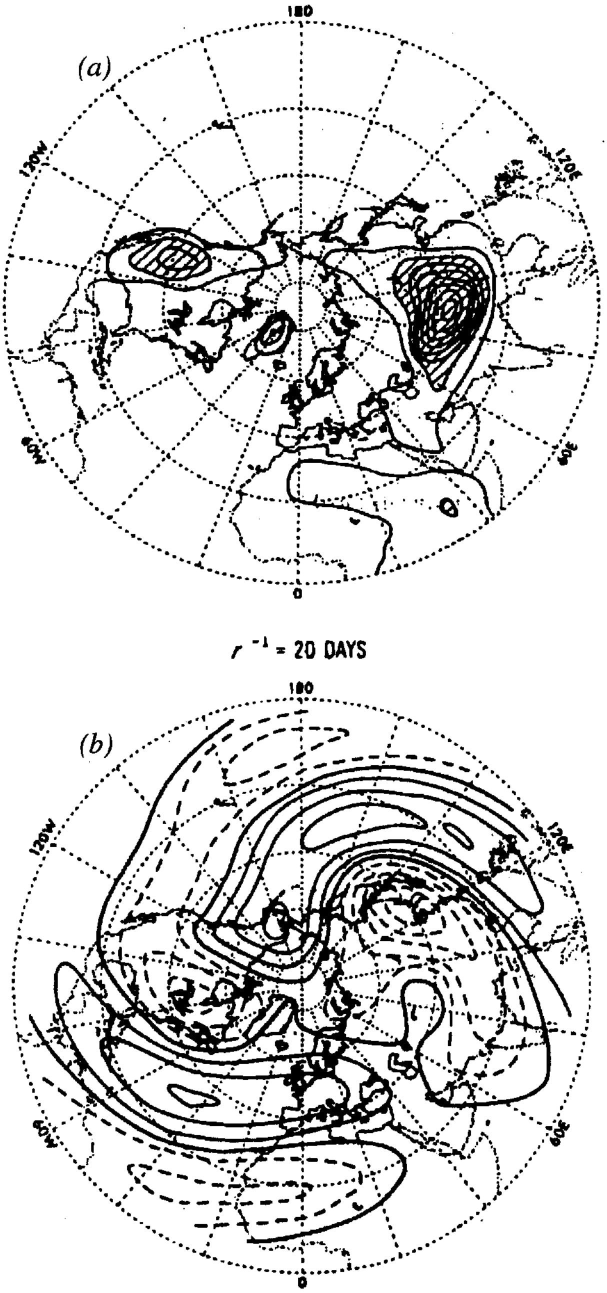


Figure 2 (*a*) Topography of the Northern Hemisphere. (*b*) The stream function response of the Northern Hemisphere of a barotropic model, forced using the zonal mean winds and Northern Hemisphere topography. Solid contours are positive or zero; dashed contours are negative (from Held, 1982).

optics. The atmosphere tends on average to exhibit low values of the index of refraction toward the polar regions, with high values toward the subtropical regions. Therefore, a wave excited somewhere in extratropics will refract away from the polar latitudes and propagate in the direction of the tropics.

When the zonal mean flow varies with latitude, two important complications occur. The first is the existence of a turning latitude where $n = n_s$. At this latitude, both the meridional wavenumber and meridional group velocity go to zero; thus, for a wave propagating poleward and eastward out of the tropics, it would continue to be refracted until at some point its troughs and ridges are aligned entirely north/south. After this point, the wave continues to turn and begins to propagate back toward the tropics.

The second complication is posed by the existence of a critical latitude, where $u_{g0} = 0$. Critical latitudes are generally marked by complicated dynamics because the implicit linear assumption that particle position deviations from rest are small does not hold. For the case of a Rossby wave train, as it approaches its critical latitude, linear theory predicts that the meridional wavenumber $l \rightarrow \infty$, while the meridional group velocity vanishes. The development of small scales while propagation slows leads linear theory to predict the absorption of the Rossby wave train at critical latitudes. However, the inclusion of nonlinearities opens a whole set of possibilities, including the reflection of the equatorward propagating wave train back into the extratropics. More generally, down gradient fluxes of potential vorticity associated with stationary wave breaking (or for that matter, with synoptic transient wave breaking) at critical latitudes located on the flanks of the upper tropospheric or stratospheric jets plays an important role in the momentum budget of the extratropical atmospheric general circulation. Specifically, this wave breaking results in significant wave-induced forcing of the zonal basic state flow.

Stationary Rossby waves not only propagate zonally and meridionally but also vertically. To understand this vertical propagation, let us return to the continuously stratified QG equations of Section 2. We are free to seek stationary solutions to Eqs. (14) and (15) with $c = 0$. For the case where u_{g0} and N are constant, but including topography $h_T = \exp(ily)\exp(ikx)$, the dispersion relation (17) leads to a condition on the vertical wavenumber m. Provided that $0 < u_{g0} < \beta\ (k^2 + l^2 + \varepsilon(2H)^{-2})^{-1}$, m will be real and given by

$$m = \pm N f_0^{-1} (\beta {u_{g0}}^{-1} - (k^2 + l^2) - \varepsilon(2H)^{-2})^{1/2} \tag{34}$$

Examination of the vertical group velocity $(\partial\omega/\partial m)$ reveals that it has the same sign as m, so choosing the positive branch of (34) corresponds to a wave propagating upward from a source at the ground and is an acceptable solution. For winds outside this range, m will be imaginary, and the waves will be evanescent rather than propagate with height. The most important fact highlighted by this simple system is that only sufficiently weak westerlies permit vertical stationary wave propagation. For parameters relevant to Earth's atmosphere, stationary wavenumbers three and greater will not readily propagate into the stratosphere, thus accounting for the predominance of wavenumbers one and two in the winter stratosphere. Additionally, the summer stratospheric easterlies effectively block the propagation of all stationary waves, accounting for the observed zonal character of the summer stratospheric circulation.

Of particular importance is the external Rossby wave, defined as the evanescent mode with the gravest vertical structure, which tends to dominate the response downstream of a localized forcing. For realistic flows such as the one shown in Figure 3*a*, the external mode is equivalent barotropic, with geopotential amplifying with height in the troposphere and decaying with height in the stratosphere. The stationary Green's function response to a "spike" imposed topographic forcing $h_T = \sin(ly)\delta(x)$, where $\delta(x)$ is the Dirac delta function, is shown in Figure 3*c*. The emergence of the external mode in the far-field is apparent, due to the rapid vertical propagation of longer wavelength stationary Rossby waves and destructive interference among shorter wavelength stationary Rossby waves as they propagate upward and downward, confined by their midtropospheric turning levels.

5 ISENTROPIC POTENTIAL VORTICITY

In any fluid dynamical problem, the existence of conserved quantities provides a foundation on which to build firm theoretical descriptions of dynamical phenomena. The usefulness of having such a conserved quantity has been apparent throughout the above discussion, as, for example, the development of the necessary conditions for stability in Section 3 explicitly used the conservation of QGPV to develop the global pseudomomentum conservation relation (24). In the atmosphere, the Rossby–Ertel isentropic potential vorticity (IPV) goes beyond QGPV in that it is conserved by the full dynamics of the atmosphere in the absence of forcing and dissipation, rather than by just the simplified QG subset of those dynamics.

The conservation of IPV in the full dynamical equations of motion can be shown as follows. In a frictionless, adiabatic atmosphere a close material contour on an isentropic surface remains on that surface and its absolute circulation

$$C_a = \oint \mathbf{u} \cdot d\mathbf{l} = \int \zeta \cdot \mathbf{n}\, dS \tag{35}$$

is conserved. For an elemental cylinder of cross-sectional area S between the isentropic surfaces θ and $\theta + \delta\theta$, $C_a = \zeta \cdot \mathbf{n}S$, and the mass $m = \rho S\,\delta h$ and $\delta\theta$ are also conserved. Here $\mathbf{n}$ is the unit normal to the isentropic surface and h is the distance in this direction. Therefore

$$\frac{C_a}{m}\delta\theta = \rho^{-1}\zeta \cdot \mathbf{n}\frac{\delta\theta}{\delta h} \tag{36}$$

is also conserved. In the infinitesimal limit this gives material conservation of the IPV

$$P = \rho^{-1}\zeta \cdot \nabla\theta \tag{37}$$

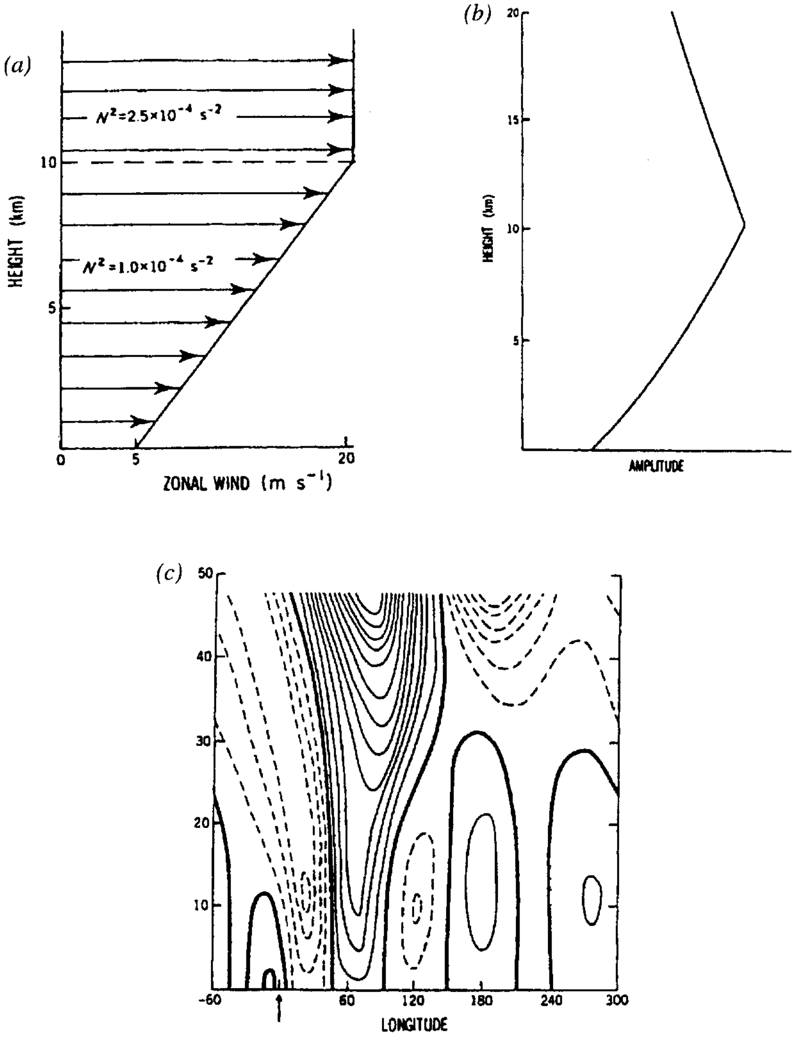


Figure 3 (*a*) Idealized zonal wind and static stability profile. (*b*) Vertical structure of the external Rossby wave for this flow profile. (*c*) Stationary response to "spike" topography for this flow profile, where the arrow marks the location of the source. Solid lines are positive contours and dashed lines are negative (from Held, 1982).

At the level of the hydrostatic approximation, $\zeta_a \cdot \mathbf{n} = \zeta_{a\theta}$, a quantity resembling the vertical component of absolute vorticity but with horizontal derivatives evaluated on isentropic surfaces. Then (37) is similarily approximated by

$$P = \sigma^{-1}\zeta_{a\theta} \tag{38}$$

where $\sigma = -g^{-1}\partial p/\partial\theta$ plays the role of density in isentropic coordinates.

In a frictionless, adiabatic atmosphere IPV is conserved by the two-dimensional (2D) motion on an isentropic surface; and similarly θ is conserved by the 2D motion on an iso-IPV surface. Most importantly, if the interior IPV and boundary θ distributions are known, then *balanced* motions may be determined by inversion of an elliptic operator. The simplest balance assumption is geostrophy; under this assumption, given the IPV/surface θ distributions one could deduce the structure of the waves and instabilities that comprise the large-scale extratropical atmospheric circulation. However, much more general balance assumptions can be made; for example, the assumption of some form of semigeostrophic balance would allow the study of IPV-dynamical interactions in the vicinity of fronts.

Provided the elliptic operator used for the IPV inversion is linear, a straightforward examination of the relative roles of various IPV disturbances in a given dynamical situation is possible. For example, the use of a QG inversion operator to study observed cyclogenesis leads to a similar dynamical picture to that outlined for the Eady model above, namely the intensification of the surface temperature anomalies by winds that stem from the upper tropospheric IPV anomalies, and vice versa. However, IPV allows more general investigations within this framework than does QGPV; for example, the importance of diabatic production of IPV by latent heat release and surface friction during cyclogenesis can be using such a decomposition. Hoskins et al. (1985) review a number of applications of IPV inversion, along with an in-depth review of the associated concepts underlying IPV conservation and atmospheric dynamics.

A cross section of the time-averaged IPV (Fig. 4) shows that the strongest gradients of IPV along isentropic surfaces occur at the tropopause. The other region of

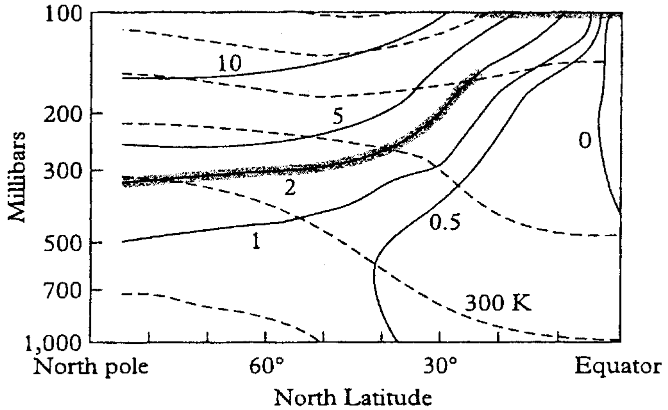


Figure 4 Climatology of isentropic potential vorticity (IPV) and potential temperature θ for the Northern Hemisphere winter. Isentropes are dashed and drawn in intervals of 30 K; IPV has units of $10^{-6}\,\text{m}^2\,\text{K}\,(\text{kg}\,\text{s})^{-1} = 1$ PV unit. Contours of IPV are 0, 0.5, 1, 2, 5, and 10 PV units, with the contour at 2 PV units stippled to indicate the approximate position of the dynamical tropopause.

importance is the large horizontal temperature gradient at the surface in midlatitudes. The natural decomposition of the general circulation into those isentropic surfaces that are everywhere above the tropopause (the "overworld"), those that intersect the tropopause ("middleworld"), and those that intersect the surface ("underworld") provides an alternative framework in which to view global-scale dynamical processes in the extratropical atmosphere. The qualitative link between IPV and the mass field in (38) makes IPV an ideal tool with which to study planetary-scale dynamical motions; examples of such motions include the structure of the mean meridional mass transport in the troposphere, the interaction between thermal and mechanical sources of IPV at Earth's surface and their role in the overall general circulation, and the exchange of air between the stratosphere and troposphere. Coupled with the invertibility principle, this link makes IPV a powerful diagnostic tool for studying the extratropical atmospheric circulation.

REFERENCES

Andrews, D. G., J. R. Holton, and C. B. Leovvy (1987). *Middle Atmosphere Dynamics*, International Geophysics Series, Vol. 40, Academic, New York.

Gill, A. E. (1982). *Atmosphere-Ocean Dynamics*, International Geophysics Series, Vol. 30, Academic, New York.

Held, I. M. (1982). Stationary and quasi-stationary eddies in the extratropical tropopshere: Theory, In *Large-Scale Dynamical Processes in the Atmosphere*, B. J. Hoskins and R. P. Pearce (Eds.), Academic, New York, pp. 127–168.

Held, I. M., and B. J. Hoskins (1985). Large-scale eddies and the general circulation of the troposphere, *Adv. Geophys*. **28A**, 3–31.

Holton, J. R. (1992). *An Introduction to Dynamical Meteorology*, 3rd ed., International Geophysics Series, Vol. 48, Academic, New York.

Holton, J. R., P. H. Haynes, M. E. McIntyre, A. R. Douglass, R. B. Rood, and L. Pfister (1995). Stratosphere–troposphere exchange, *Rev. Geophys*. **33**, 403–439.

Hoskins, B. J. (1982). The mathematical theory of frontogenesis, *Ann. Rev. Fluid Mech.* **14**, 131–151.

Hoskins, B. J., M. E. McIntyre, and A. W. Robertson (1985). On the use and significance of isentropic potential vorticity maps, *Quart. J. Roy. Met. Soc.* **111**, 877–946.

Pedlosky, J. P. (1987). *Geophysical Fluid Dynamics*, 2nd ed. Springer, New York.

Pierrehumbert, R. T., and K. L. Swanson (1995). Baroclinic instability, *Ann. Rev. Fluid Mech.* **27**, 419–167.

CHAPTER 5

DYNAMICS OF THE TROPICAL ATMOSPHERE

GERALD MEEHL

1 THERMALLY DIRECT CIRCULATIONS AND MONSOONS

The heating of Earth's surface that follows the annual cycle of solar forcing produces fundamentally ageostrophic thermally direct circulations in the tropical atmosphere. Hadley postulated that two thermally driven north–south cells, roughly straddling the equator, must be necessary to transport heat and energy surpluses from the tropics toward the poles. These meridional cells now bear his name. Subsequently, Sir Gilbert Walker recognized that, in addition to the north–south Hadley cells, important components of large-scale tropical circulations are associated with tremendous east–west exchanges of mass. The term *Walker circulation* was first used by Bjerknes (1969) to describe these exchanges of mass in the equatorial Pacific. Subsequently, Krishnamurti (1971) showed that the Walker circulation in the equatorial Pacific is just a component of large-scale east–west circulations emanating from centers of organized tropical convective activity associated with regional monsoon regimes.

For example, applications of the Hadley and Walker circulation concepts have linked the monsoon regimes of Asia and northern Australia to weather and climate in other parts of the tropics and subtropics. Figure 1 shows one such interpretation of the east–west Walker circulation connecting the Asia–Australia region to the eastern Pacific (Webster et al., 1998). Another branch of east–west circulation, termed the *transverse monsoon* here, connects those monsoon regions to eastern Africa. The north–south Hadley components in Figure 1 are termed *lateral monsoon* and involve large-scale mass overturning from the respective summer monsoon regimes to

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

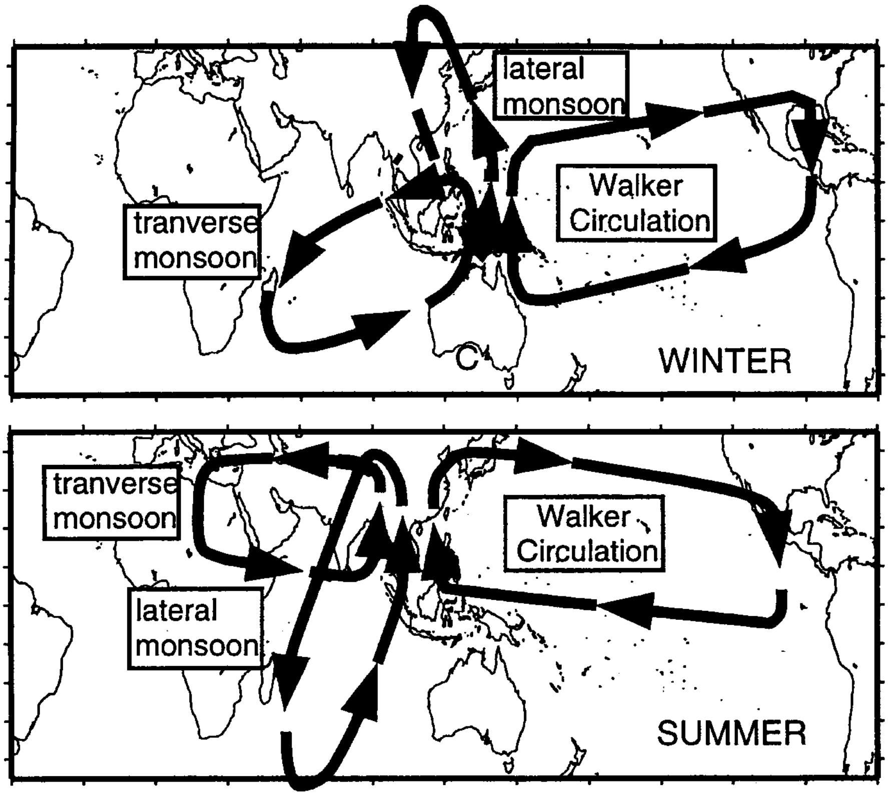


Figure 1 Schematic of the (*a*) south Asian and (*b*) Australian monsoon divergent wind circulations. Three major components identified are the transverse monsoon (Walker-type east–west circulation west of the monsoons), lateral monsoon (Hadley-type north–south circulations), and Walker circulation (east–west circulation east of the monsoons) (Webster et al., 1998).

the winter hemisphere subtropics, thus depicting the classic Hadley cell type of circulation.

The latitude where the sun is most directly overhead in its seasonal march is where the surface is most strongly heated in the tropics. The air directly above that warm surface rises and draws nearby air in to take its place. The warm air rises, cools and condenses, and forms convective clouds. The low-level convergence at that latitude forms a zonally oriented line of convection called the intertropical convergence zone (ITCZ). The arrangement of land and ocean in the tropics dictates that the ITCZ does not simply follow the sun's motion from northern to southern hemisphere with the seasonal cycle. Off-equatorial land masses in the tropics tend to heat more than the adjacent oceans, thus giving rise to seasonal regional convective systems called monsoons. Though the word *monsoon* has a number of interpretations, monsoon regimes exist wherever there is a regional organization of convection associated with an off-equatorial land mass and an equatorward ocean moisture source. This particular distribution of land and ocean disrupts the normal zonal orientation of the ITCZ and contributes to the east–west Walker-type circulations (e.g., Fig. 1). These regional monsoon regimes are recognized to exist in tropical regions of south Asia, northern Australia, southwestern North America, west Africa, and, in some definitions, South America.

All these monsoon regimes share certain characteristics. With the approach of local summer, intensifying solar radiation heats the land surface forming a heat low. Since the land surface has heated faster than the adjacent ocean, the heat low begins to draw in moist low-level air from the ocean. This moist warm air rises and forms convective clouds and rainfall. The latent heating from the convective systems powers a positive feedback, with even stronger low-level moist inflow from the ocean intensifying convection and precipitation over land. An upper level high forms overhead with outflow in all directions from the regional convective center associated with the monsoon circulation. This feeds into the thermally direct mass circulations mentioned above and connects the monsoon regimes with other areas in the tropics and subtropics. The massive amounts of latent heating associated with the monsoon regimes provide a major energy source for the general circulation of the global atmosphere.

2 EL NIÑO–SOUTHERN OSCILLATION

It was the efforts of Walker and others to try to understand the factors that influence the Indian monsoon in particular that led to the discovery of the east–west circulations evidenced by large-scale exchanges of mass between the Indian and Pacific Oceans described above. This phenomenon became known as the Southern Oscillation. These investigators associated the fluctuations of mass between the Indian and Pacific regions with temperature and precipitation variations over much of the globe. Bjerknes (1969) later linked sea surface temperature (SST) anomalies in the equatorial Pacific (which came to be known as El Niño and La Niña), and changes in the east–west atmospheric circulation near the equator, to the Southern Oscillation. The entire coupled ocean–atmosphere phenomenon is called *El Niño–Southern Oscillation* (ENSO).

ENSO and its connections to the large-scale east–west circulation can best be illustrated by correlations of sea level pressure between one node, representing the Indian–western Pacific region (in this case Darwin on the north coast of Australia), with all other points (Fig. 2). Positive values cover most of the south Asian and Australian regions, with negative values east of the dateline in the Pacific. Also note positive values over northeastern South America and the Atlantic. Thus, when convection and precipitation in the south Asian or Australian monsoons tends to be below normal, there is high pressure in those regions with a corresponding decrease of pressure and enhanced rainfall over the tropical Pacific, with higher pressure and lower rainfall over northeastern South America and tropical Atlantic.

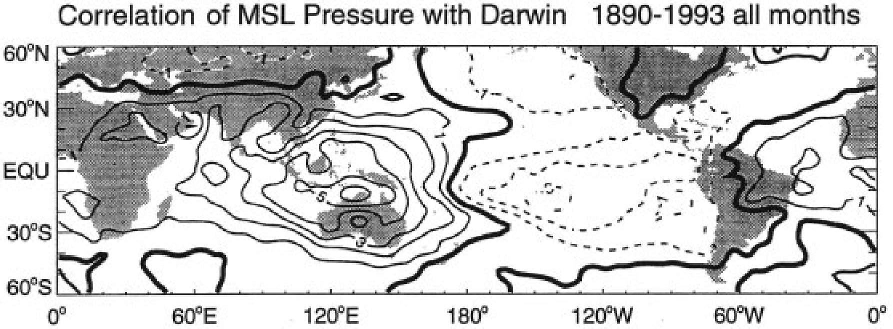

Correlation of MSL Pressure with Darwin 1890-1993 all months
60°N
30°N
EQU
30°S
60°S
0°
60°E
120°E
180°
120°W
60°W
0°
.5
.7
.3

Figure 2 Distribution of the zero-lag correlation of surface pressure variations with the Darwin surface pressure and all other points for monthly data for the period 1890–1993 (Webster et al., 1998).

Since the atmosphere and ocean are dynamically coupled in the tropics, ENSO and its associated fluctuations of atmospheric mass is associated with SST anomalies as a consequence of the mostly ageostrophic surface winds involved with sea level pressure anomalies such as those implied in Figure 2. This dynamic coupling is reflected with the association between El Niño and La Niña events and the Southern Oscillation noted above. For example, anomalously low pressure and enhanced precipitation over the tropical Pacific are usually associated with anomalously warm SSTs there. In the extreme, this is known as an El Niño event. In the opposite extreme, anomalously cold SSTs east of the dateline have come to be known as a La Niña event. These are often associated with low pressure and strong Asian–Australian monsoons west of the dateline, with high pressure and suppressed rainfall east of the dateline. The conditions that produce El Niño and La Niña events, with a frequency of roughly 3 to 7 years, involve dynamically coupled interactions of ocean and atmosphere over a large area of the tropical Indian and Pacific Oceans (Rasmusson and Carpenter, 1982). Figure 3 shows low-level wind, precipitation, and SST anomalies for the northern summer season during the 1997–1998 El Niño event. Westerly surface wind anomalies occur to the west of the largest positive SST anomalies in the central and eastern equatorial Pacific. These wind anomalies reduce upwelling of cool water and contribute to the positive SST anomalies. The warmer water is associated with greater evaporation and increased precipitation. As a result of the anomalous Walker circulation, there is suppressed precipitation over much of Southeast Asia.

El Niño and La Niña events are extremes in the system that tends to oscillate every other year to produce the tropospheric biennial oscillation, or TBO (Meehl, 1997). Thus there are other years, in addition to the Pacific SST extremes, that have similar coupled processes but lower amplitude anomalies also involving the Asian–Australian monsoons and tropical Indian and Pacific SST anomalies (Fig. 4). The TBO encompasses all the ENSO processes of dynamically coupled air–sea interaction in the tropical Indian and Pacific Oceans, large-scale east–west cir-

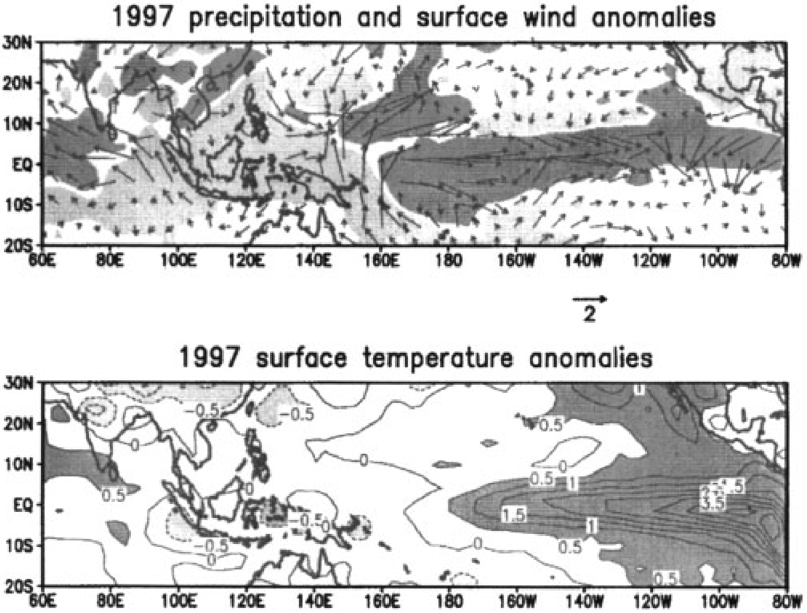
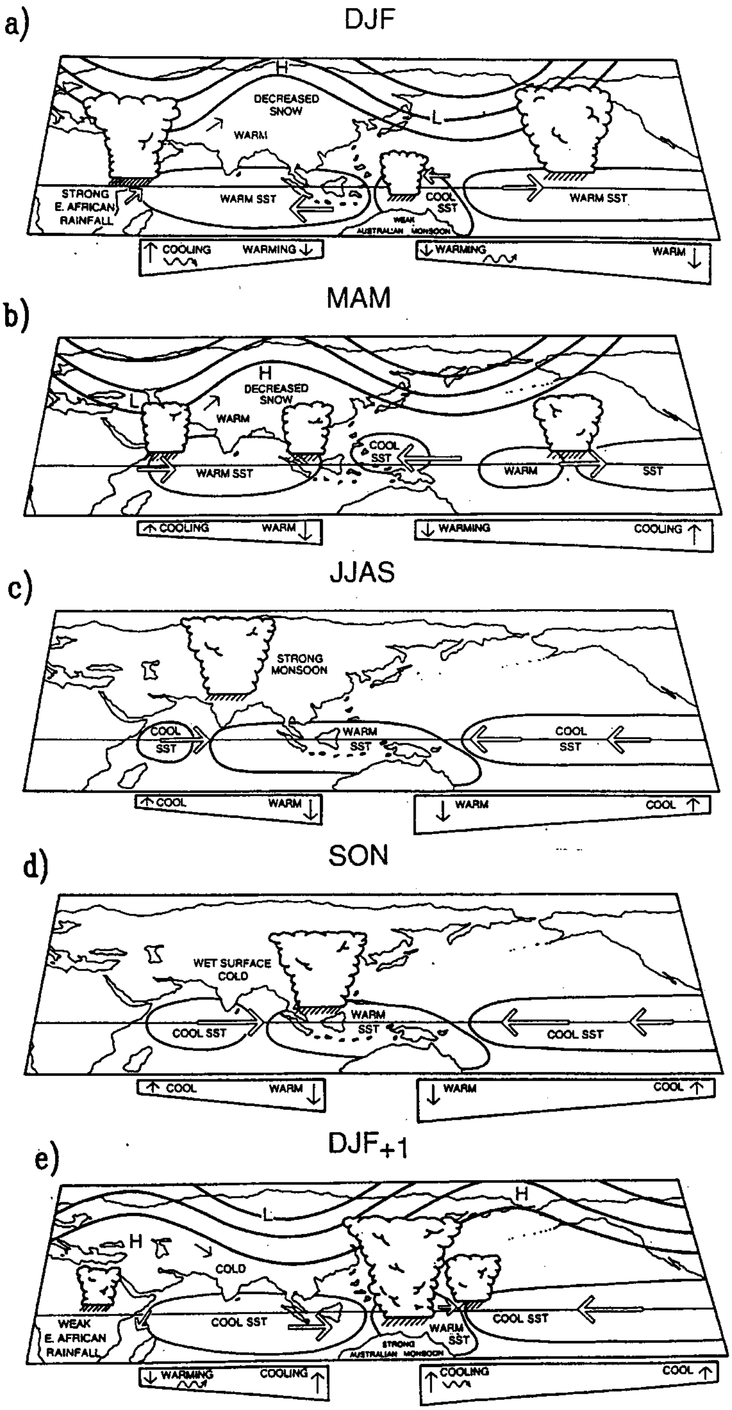


Figure 3 (*a*) Precipitation and surface wind anomalies for the 1997 June–July–August–September seasonal average during the 1997–1998 El Niño event; dark gray shading indicates precipitation anomalies greater than 0.5 mm/day, light gray less than −0.5 mm/day, scaling arrow below part (*a*) indicates surface wind anomalies of 2 m/s; (*b*) as in (*a*) except for SST anomalies with dark gray greater than 0.5°C, light gray less than −0.5°C.

culations in the atmosphere, the Asian–Australian monsoons, and coupled SST anomalies in Indian and Pacific involving internal ocean wave dynamics.

Figure 4 illustrates these coupled interactions for a sequence from a weak Australian monsoon during anomalously warm conditions in the tropical Pacific that, in the extreme, is an El Niño event, through to the onset of anomalously cold SSTs in the Pacific (in the extreme, a La Niña event) associated with a strong Indian monsoon, and on to a strong Australian monsoon the following southern summer. In the December–January–February (DJF) season prior to a strong Indian monsoon, there is a relatively weak Australian monsoon (Fig. 4*a*), with warm SST anomalies to the west in the Indian Ocean, to the east in the central and eastern Pacific, and relatively cool SSTs north of Australia. Relatively strong convection over the western Indian Ocean and eastern African areas, along with strong convection over the central equatorial Pacific, are associated with an atmospheric Rossby wave response over Asia such that there is an anomalous ridge of positive 500 hPa heights and warm land temperatures. Anomalous easterly winds in the western Pacific induce upwelling Kelvin waves in the ocean that propagate eastward. These begin to raise the thermocline in the central and eastern Pacific over the course of the next few months, as indicated in Figure 4*b*. Westerly anom-


a)
DJF
H
DECREASED SNOW
WARM
L
STRONG E. AFRICAN RAINFALL
WARM SST
COOL SST
WEAK AUSTRALIAN MONSOON
WARM SST
COOLING
WARMING
WARMING
WARM
b)
MAM
H
DECREASED SNOW
L
WARM
WARM SST
COOL SST
WARM
SST
COOLING
WARM
WARMING
COOLING
c)
JJAS
STRONG MONSOON
COOL SST
WARM SST
COOL SST
COOL
WARM
WARM
COOL
d)
SON
WET SURFACE COLD
COOL SST
WARM SST
COOL SST
COOL
WARM
WARM
COOL
e)
DJF+1
L
H
H
COLD
WEAK E. AFRICAN RAINFALL
COOL SST
WARM SST
STRONG AUSTRALIAN MONSOON
COOL SST
WARMING
COOLING
COOLING
COOL

alous surface winds in the far western Indian Ocean set off downwelling equatorial Kelvin waves in the ocean and contribute to a deepening of the thermocline in the eastern equatorial Indian Ocean in March–April–May (MAM, Fig. 4*b*).

In the MAM season prior to a strong Indian monsoon (Fig. 4*b*), the upwelling Kelvin waves from the easterly anomalous winds in the western Pacific during DJF continuing into MAM act to raise the thermocline in the central and eastern Pacific.

During June–July–August–September (JJAS) (Fig. 4*c*), convection and precipitation over the Indian monsoon region is strong. Anomalous winds over the equatorial Indian ocean remain westerly, and the shallow thermocline in the west is evidenced by anomalously cool SSTs appearing in the western equatorial Indian Ocean. There are strong trades, and a shallow thermocline and cool SSTs in the central and eastern equatorial Pacific, thus completing the SST transition there.

The September–October–November (SON) season after the strong Indian monsoon (Fig. 4*d*) has anomalously strong convection traversing with the seasonal cycle to the southeast, encountering warm SSTs set up by the dynamical ocean response to the westerly winds in the equatorial Indian Ocean in JJAS with a dipole of SST anomalies in the Indian Ocean.

Finally, in Figure 4*e* in the DJF following a strong Indian monsoon, there is a strong Australian monsoon, and strong easterlies in the equatorial Pacific as part of the strengthened Walker circulation associated with cool SSTs there. Anomalous westerlies in the far western equatorial Pacific associated with the strong Australian monsoon convection start to set off downwelling equatorial oceanic Kelvin waves, which begin to deepen the thermocline to the east and set up the next transition to warm SSTs in the central and eastern equatorial Pacific the following MAM–JJAS.

3 SUBSEASONAL TROPICAL ATMOSPHERIC WAVES

The westerly winds in the far western equatorial Pacific involved with transitions of SST anomalies in the Pacific have been associated with a subseasonal phenomenon called the *Madden–Julian Oscillation* or MJO (Madden and Julian, 1972). The MJO is a dominant mode of subseasonal tropical convection. It is associated with an eastward-moving envelope of convection that progresses around the global tropics with a period of about 30 to 70 days and is most evident in clouds and convection in

Figure 4 Schematic representation of coupled processes for (*a*) the DJF season after a weak Indian monsoon with a weak Australian monsoon and anomalously warm SSTs in the tropical Pacific, (*b*) MAM, (*c*) a strong Indian monsoon associated with anomalously cold SSTs in the Pacific, (*d*) SON and (*e*) DJF with a strong Australian monsoon. Note that conditions in DJF preceding the strong monsoon in (*a*) are opposite to those in the DJF following the strong monsoon in (*e*) indicative of the TBO in the coupled system. The TBO encompasses El Niño and La Niña events along with other years with similar but lower amplitude anomalies. Thus the coupled processes depicted here apply to TBO and ENSO (Meehl and Arblaster, 2002).

the Indian and Pacific sectors. Particularly active MJO events, as they move across the western Pacific, are associated with bursts of westerly anomalous winds that force eastward propagating Kelvin waves in the ocean. These oceanic Kelvin waves affect the depth of the thermocline to the east and can contribute to a transition of SST anomalies in the eastern Pacific associated with the onset of some El Niño events (McPhaden, 1999). For example, Kessler et al. (1995) show how successive MJO westerly wind events in the western Pacific can lead to the forcing of oceanic Kelvin waves and successive or stepwise advection of warm SSTs from the western Pacific toward the east to contribute to El Niño onset.

The MJO is just one of a number of convectively coupled tropical waves (i.e., atmospheric wave dynamics coupled to convection such that their signatures are seen in analysis of tropical clouds) that exist in the equatorial atmosphere. Subseasonal tropical convection is organized on larger scales by certain modes that are unique to the equatorial region. An entire class of equatorially trapped modes organizes atmospheric motions and convection at low latitudes. The theory for these "shallow water" modes was developed by Matsuno (1966).

Formally, equatorial wave theory begins with a separation of the primitive equations, linearized about a basic state with no vertical shear, governing small motions in a three-dimensional stratified atmosphere on an equatorial Beta plane, into the "vertical structure" equation and "shallow-water" equations. Four parameters characterize the equatorial wave modes that are the zonally (and vertically) propagating equatorially trapped solutions of the shallow-water equations. These include meridional mode number n, frequency v, "equivalent depth" h of the "shallow" layer of fluid, and zonal wavenumber s. The internal gravity wave speed c is related to the equivalent depth as

$$c = \sqrt{gh} \tag{1}$$

and links the vertical structure equation and shallow-water equations as a separation constant. The equatorial Rossby radius R, is related to the vertical wavelength of free waves and to the meridional scaling by

$$R_e = \left(\frac{\sqrt{gh}}{\beta}\right)^{1/2} \tag{2}$$

where β is the latitudinal gradient of the Coriolis parameter.

The theoretical dispersion relationship will fully characterize the wave, given the meridional mode number and wave type, provided two out of h, v, and s are specified. It is assumed that tropical waves associated with convective heating are internal modes with wavelike vertical structures, and the resulting solution of the shallow-water equations are either antisymmetric or symmetric about the equator. Convection is usually related to the divergence or temperature field, and modes of even meridional mode number n are antisymmetric, and odd n are symmetric. Figure 5 shows examples of circulation and divergence associated with some of these symmetric

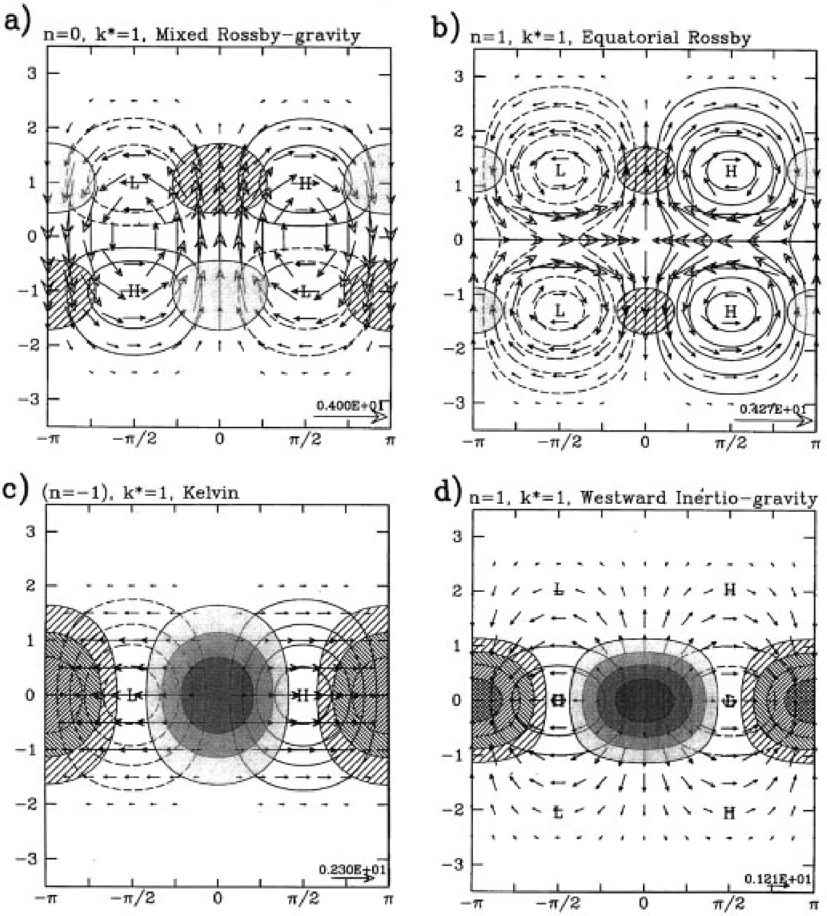

a)
n=0, k*=1, Mixed Rossby-gravity
b)
n=1, k*=1, Equatorial Rossby
c)
(n=-1), k*=1, Kelvin
d)
n=1, k*=1, Westward Inertio-gravity
3
2
1
0
-1
-2
-3
-π
-π/2
0
π/2
π
L
H
0.400E+01
0.327E+01
0.230E+01
0.121E+01

Figure 5 Depiction of various atmospheric tropical waves from shallow water theory: (*a*) $n = 0$ mixed Rossby gravity, (*b*) $n = 1$ equatorial Rossby, (*c*) $n = -1$ Kelvin wave, and (*d*) $n = 1$ westward inertiogravity wave. Hatching indicates convergence, shading is divergence.

modes, the $n = 0$ mixed Rossby gravity, $n = 1$ equatorial Rossby, $n = -1$ K, and $n = 1$ westward inertiogravity waves.

A space–time analysis of satellite-observed outgoing longwave radiation (OLR), a proxy for tropical convection, is shown in Figure 6. By separating the OLR into its antisymmetric and symmetric components about the equator, as well as removing a red background spectrum, the spectral peaks of these convectively coupled equatorial waves are shown in Figure 6*a* for the antisymmetric components and Figure 6*b* for the symmetric (Wheeler and Kiladis, 1999). The modes occur for a wide range of space and time scales, from synoptic to planetary, and submonthly to intraseasonal. Waves depicted in Figure 5 appear in Figure 6 as Kelvin waves, equatorial Rossby

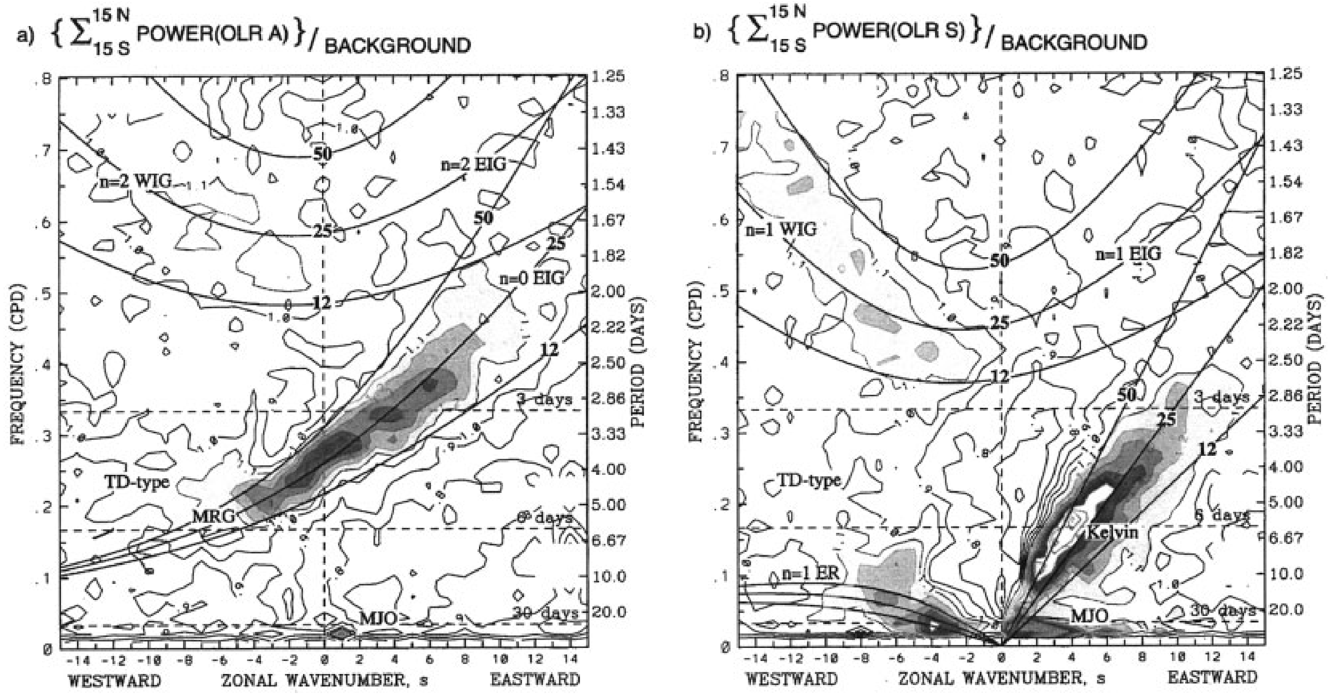


Figure 6 (*a*) Antisymmetric (with respect to equator) observed OLR power divided by the background power. The background power is calculated as a smoothed average of the power of the antisymmetric and symmetric components. The power is calculated for the 1979 to 1996 period and summed between 15 S and 15 N. Contour interval is 0.1, and shading begins at a value of 1.1 for which the spectral signatures are statistically significant above the background at the 95% level (based on 500 degrees of freedom). Contours less than 1.0 are dashed. Superimposed are the dispersion curves of the even meridional mode-numbered equatorial waves for the three equivalent depths of $h = 12$, 25, and 50 m. (*b*) Same as (*a*) except for the symmetric component of OLR, and the corresponding odd meridional mode-numbered equatorial waves. Frequency spectral bandwidth is 1/96 cpd (Wheeler and Kiladis, 1999).

(ER) waves, mixed Rossby gravity (MRG) waves, and westward inertiogravity (WIG) waves. Eastward inertiogravity (EIG) waves, along with tropical depression (TD) type and the MJO are also depicted. The mode with the most power is the MJO, which is unique in these spectra since it does not correspond to any particular shallow-water mode. It has most of its power in the eastward-propagating symmetric zonal wavenumbers 1 through 6 (Fig. 6*b*), with a substantial contribution to the zonal mean as well. There is also some MJO power in the same region of the antisymmetric spectrum in Figure 6*a*.

The periodic and apparently linear appearance of these waves would imply that they are predictable for a lead time of about half their period. Such relationships can be applied to predict tropical rainfall variability on the submonthly time scale using knowledge of the base state provided by the MJO (Lo and Hendon, 2000). The MJO itself can provide forecast information concerning the onset of ENSO or swings in the TBO, while the interannual time scale base states affect MJO and submonthly convection in the tropics in a continuum of upscale and downscale interactions involving tropical dynamics (Meehl et al., 2001).

REFERENCES

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial Pacific, *Mon. Wea. Rev.* **97**, 163–172.

Kessler, W. S., M. J. McPhaden, and K. M. Weickmann (1995). Forcing of intraseasonal Kelvin waves in the equatorial Pacific, *J. Geophys. Res.* **100**, 10,613–10,631.

Krishnamurti, T. N. (1971). Tropical east-west circulations during the northern summer, *J. Atmos. Sci.* **28**, 1342–1347.

Lo, F., and H. H. Hendon (2000). Empirical extended-range prediction of the Madden-Julian Oscillation, *Mon. Wea. Rev.* **128**, 2528–2543.

Madden, R. A., and P. R. Julian (1972). Description of global-scale circulation cells in the tropics with a 40–50 day period, *J. Atmos. Sci.* **29**, 1109–1123.

Matsuno, T. (1966). Quasi-geostrophic motions in the equatorial area, *J. Meteorol. Soc. Jpn.* **44**, 25–43.

McPhaden, M. (1999). Genesis and evolution of the 1997–98 El Niño, *Science* **283**, 950–954.

Meehl, G. A. (1997). The South Asian monsoon and the tropospheric biennial oscillation, *J. Climate* **10**, 1921–1943.

Meehl, G. A., and J. M. Arblaster (2002). The tropospheric biennial oscillation and Asian-Australian monsoon rainfall, *J. Climate* **15**, 722–744.

Meehl, G. A., R. Lukas, G. N. Kiladis, M. Wheeler, A. Matthews, and K. M. Weickmann (2001). A conceptual framework for time and space scale interactions in the climate system, *Clim. Dyn.* **17**, 753–775.

Rasmusson, E. M., and T. H. Carpenter (1982). Variations in tropical sea surface temperature and surface wind fields associated with the southern oscillation/EL Niño, *Mon. Wea. Rev.* **110**, 354–384.

Webster, P. J., V. O. Magana, T. N. Palmer, J. Shukla, R. A. Tomas, M. Yanai, and T. Yasunari (1998). Monsoons: Processes, predictability, and the prospects for prediction, *J. Geophys. Res.* **103**, 14,451–14,510.

Wheeler M., and G. N. Kiladis (1999). Convectively-coupled equatorial waves: Analysis of clouds and temperature in the wavenumber-frequency domain, *J. Atmos. Sci.* **56**, 374–399.

CHAPTER 6

TURBULENCE

JACKSON R. HERRING

1 TWO-DIMENSIONAL AND QUASI-GEOSTROPHIC TURBULENCE

Large-scale motions of Earth's atmosphere are perforce nearly two dimensional since the atmosphere's thickness is thin compared to Earth's radius. However, more important dynamical considerations also dictate near two dimensionality: Flows that are both stable against convection, and rapidly rotating, exhibit strong two-dimensional motion regardless of their vertical dimensions. The atmosphere, on average, meets these conditions and it is this latter constraint that determines the dynamic nature of its approximate two dimensionality. To illustrate this point, we consider a simple model consisting of a thin layer of incompressible fluid, subjected to the gravitational force **g** and rotation **Ω**. The equations of motion for such a fluid are the Boussinesq approximation to the Navier–Stokes equations:

$$\{\partial_t + \mathbf{u}\cdot\nabla\}\mathbf{u} = \hat{\mathbf{g}}\alpha T - 2\boldsymbol{\Omega}\times\mathbf{u} - \nabla p + \nu\nabla^2\mathbf{u} \tag{1}$$

$$\{\partial_t + \mathbf{u}\cdot\nabla\}\theta = \hat{\beta}w + \nabla\cdot\overline{\mathbf{u}\theta} + \kappa\nabla^2\theta \tag{2}$$

$$\nabla\cdot\mathbf{u} = 0 \tag{3}$$

In Eq. (1), $\mathbf{u} = (u, v, w)$ denotes the velocity, with w the velocity component parallel to Earth's radius, u the eastward component, and v the northern. The temperature T is split into its horizontal average $\bar{T}$, and a fluctuation about that average θ, with $\hat{\beta} \equiv -d\bar{T}/dz$; α is the coefficient of thermal expansion of air $(= 1/T)$. We consider these equations in a Cartesian coordinate frame tangential to Earth's sphere at a latitude $\pi/2 - \vartheta$. Here, δp is the deviation of the pressure field

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

from that necessary to provide a balance with gravitational and centrifugal forces. Its role in Eq. (1) is to preserve Eq. (3).

Consider now Eqs. (1) to (3) in the limit of large **g** and $\mathbf{\Omega}$. Then under suitable conditions [so as to neglect breaking gravity waves, and hence $\nabla \cdot \overline{\mathbf{u}\theta}$ in (2)] the flow separates into two nearly noninteracting components. The first is comprised of rapidly fluctuating gravity and inertial waves, and the second is a slowly evolving component in which terms in (1) $\sim\hat{\mathbf{g}}$ and $\sim\Omega$ nearly cancel. This separation is much analogous to the separation of acoustics and incompressible flow for low Mach numbers. It is the second "balanced" part that constitutes the focus of this chapter. For it, Eqs. (1) to (3) may be replaced by a simpler set, the *quasi-geostrophic* system (hereafter indicated by QG), whose derivation we now sketch. The QG equations do not contain gravity waves or other rapid fluctuations, as do Eqs. (1) to (3), hence their simplicity. Since the right-hand side of (1) contains individual terms (each of which is large), we may expect, in the balanced state, a cancellation among the largest of these. The equation for vorticity, $\omega = \nabla \times \mathbf{u}$ follows from (1) as

$$\partial_t\omega = -\mathbf{u}\cdot\nabla\omega + \omega\cdot\nabla\mathbf{u} + 2(\mathbf{\Omega}\cdot\nabla)\mathbf{u} - \mathbf{g}\times\nabla\theta - \beta v + \nu\nabla^2\omega \tag{4}$$

An independent constraint balancing the effects of rotating and gravity may be formed from the evolution equation for $\mathbf{g}\cdot(\nabla \times \omega)$:

$$+2(\mathbf{\Omega}\cdot\nabla)\omega - \alpha\mathbf{g}\nabla_\perp^2\theta = 0 + \cdots \tag{5}$$

Here $\beta = 2\mathbf{\Omega}\sin\theta/R$, with R the earth's radius, and θ the polar angle.

In Eq. (5), $\nabla_\perp^2 \equiv \partial_x^2 + \partial_y^2$, and $+\cdots$ indicates that only leading-order terms are recorded. Anticipating that **u** is nearly two dimensional, but with a possible z dependency, we represented it by a stream function, $\psi(x, y, z)$: $(u, v) = (-\partial_y, \partial_x)\psi(x, y, z)$. In terms of ψ, Eq. (5) simplifies to

$$\alpha g\theta = 2\Omega\partial_z\psi + \cdots \tag{6}$$

The evolution equations for QG may now be obtained by eliminating $(\partial_z w)$ between (2) and the z component of (4):

$$(\partial_t + \mathbf{u}\cdot\nabla)\left\{\omega - \frac{4\Omega^2}{\alpha\hat{\beta}g}\frac{\partial^2}{\partial z^2}\psi\right\} = -\beta v \tag{7}$$

In Eq. (7), we omit the viscous contribution ($\sim v$) from (4). In deriving (7), we neglect $(\Omega_x O_x + \Omega_y\partial_y)\omega$ as compared to $\Omega_3\partial_3\omega$ since the vertical extent of the atmosphere is much smaller than its horizontal extent. The divergence of Eq. (1) and Eq. (6) may be used to relate the stream function to the pressure δp,

$$-\delta p = 2\Omega\psi + \cdots \tag{8}$$

For Eq. (7) to be a reasonable approximation, it is clear [from Eq. (4)] that the term proportional to Ω on the right-hand-side of (4) should dominate the term: $\omega \cdot \nabla \mathbf{u}$. Such dominance is indicated by the *smallness* of the Rossby number:

$$R_o = |\omega|/\Omega \sim |u|/(\Omega\ell) \tag{9}$$

where ℓ estimates the length scale of typical horizontal gradients of the flow. For a more comprehensive statement of conditions under which the quasi-geostrophic equations are valid; see Pedlosky (1987).

We next make a few comments concerning the physics of the quasi-geostrophic approximation, confining our remarks to the case of inviscid flows, with $\beta = 0$.* First note that (7) simply states that the quantity

$$Q \equiv \left\{\omega - \frac{4\Omega^2}{\alpha\beta g}\frac{\partial^2}{\partial z^2}\psi + \beta y\right\}, \qquad \omega = -\nabla_\perp^2\psi \tag{10}$$

is conserved along streamlines. Q is the potential vorticity, and for flows confined to a given spatial volume, both it and an associated kinetic energy,

$$E = \int_{\mathcal{U}} d\mathbf{r}\left\{(\nabla_\perp\psi)^2 + \frac{4\Omega^2}{\alpha\hat{\beta} g}\left\{\frac{\partial}{\partial z}\psi\right\}^2\right\}, \quad \text{and} \quad \mathcal{E}(z) \equiv \int dxdy\, Q^2 \tag{11a,b}$$

are conserved. $\mathcal{U}$ denotes the volume to which the flow is confined, which, for simplicity, we take as a rectangular box, L_x, L_y, L_z. $\mathcal{E}$ is called the total enstrophy, and $Q(x, y, z)^2$ the enstrophy density. In Eq. (11a,b), $\nabla_\perp = \partial_x + \partial_y$.

Flows described by (7) are known to be subject to a variety of instabilities, which lead to intrinsic temporal fluctuations in Q. We shall not discuss these instabilities here, except to remark that they lead, through the effects of nonlinearities, to the existence in the flow of eddies with a broad distribution of sizes. A convenient description of this distribution stems from a Fourier analysis of $Q(\mathbf{r})$:

$$Q(r) = \sum_{\mathbf{k}} \exp(i\mathbf{k}\cdot\mathbf{r})Q(\mathbf{k}) \tag{12}$$

where $\mathbf{k} = 2\pi\ (\pm n/L_x, \pm m/L_y, \pm l/L_z)$, and (n, m, l) range over all positive integers.

A basic question in the study of turbulent flows is the distribution of Q among available modes $\mathbf{k}$. The simplest indicators of this distribution are moments

$$\langle E(\mathbf{k})\rangle \equiv (k_\perp^2 + (4\Omega^2/(\alpha\hat{\beta} g))k_z^2)\langle|\psi(\mathbf{k})|^2\rangle$$
$$\mathcal{E}(\mathbf{k}) \equiv \langle|Q(\mathbf{k})|^2\rangle$$

*Two-dimensional turbulence with $\beta \neq 0$ (Rossby wave turbulence) is important in the study on large-scale planetary flows. Statistical studies of the sort described here were initiated by Holloway and Hendershott (1977). For recent interesting advances, see Galperin et al. (2001).

where the angular brackets denote ensemble averages.

We now explore briefly the implications of the inviscid constraints E and $\mathcal{E}$ constants of motion for QG flow, confining our attention to the case $\beta = 0$ in (10). For the vertical variability, we take a simple form

$$\psi(\rho, z, t) = \frac{1}{2}(\psi_2 + \psi_1) + \frac{1}{2}(\psi_1 - \psi_2)\cos(\pi z/L) \tag{13}$$

Here $\rho = (x,y)$. Evaluating (7) at $z = (0,L)$ with (13) gives

$$(\partial_t + \mathbf{u_i} \cdot \nabla)\{-\nabla_\perp^2 \psi_i - \varepsilon(-1)^i(\psi_1 - \psi_2)\} = 0, \qquad i = (1, 2) \tag{14a,b}$$

Here, $\mathbf{u_i} = (-\partial_y, \partial_x)\psi_i$, etc., and

$$\varepsilon = 2\Omega^2\pi^2/(\alpha\hat{\beta}gL^2) \equiv 1/L_R^2 \tag{14c}$$

In (14c), L_R is the Rossby radius of deformation, which for the atmosphere is ~400 km. Note that $\psi_1 \equiv \psi_2$ is a solution to (14). Such flows are barotropic, and strictly two dimensional. Flows for which $\psi_1 - \psi_2 \neq 0$ are designated as baroclinic since in that case the pressure gradient is inclined to $\mathbf{g}$, according to (8).[†] Our discussion now focuses on the qualitative behavior of the flow at horizontal scales above and below L_R. For Eqs. (14a) and (14b), conservation laws (11a) and (11b) specialize to

$$\frac{d}{dt}\mathcal{J}_i = 0, \qquad \mathcal{J}_i = \frac{1}{2}\sum_{\mathbf{k}} |Q_i(\mathbf{k})|^2 = 0, \qquad (i = 1, 2) \tag{15a}$$

$$\frac{d}{dt}\mathcal{I} = 0, \qquad \mathcal{I} = \frac{1}{2}\sum_{\mathbf{k}}\{u_1^2 + u_2^2 + \varepsilon(\psi_1 - \psi_2)^2\} \tag{15b}$$

where

$$Q_i(\mathbf{k}) = k^2\psi_i - \varepsilon(-1)^i(\psi_1 - \psi_2) \quad i = 1, 2, \quad u_i^2 = k^2|\psi_i|^2 \tag{15c}$$

It is of interest to ask what distribution in $\mathbf{k}$ emerges as the most probable having given values of total energy, $\mathcal{I}$, and total enstrophy, $\mathcal{J}_1 + \mathcal{J}_2$. Such a distribution fixes $\langle|\psi_i(\mathbf{k})|^2\rangle$, as well as $\langle\psi_1(\mathbf{k})\psi_2(\mathbf{k})\rangle$. If $\mathcal{I}(\psi_1, \psi_2)$ and $\mathcal{J}_1(\psi_1, \psi_2) + \mathcal{J}_2(\psi_1, \psi_2)$ are, separately, constants of motion, then so is their linear combination,

[†] Equations (14a) and (14b) have more respectability than just a simple vertical collocation of (7). They describe exactly the behavior of two immiscible fluids (the lighter above) under QG conditions.

$\bar{\alpha}\mathcal{I} + \bar{\beta}(\mathcal{J}_1 + \mathcal{J}_2)$, where $\bar{\alpha}$, and $\bar{\beta}$ are arbitrary constants. We may diagonalize the quadratic form $\bar{\alpha}\mathcal{I} + \bar{\beta}(\mathcal{J}_1 + \mathcal{J}_2)$ by a change of variables:

$$\psi_1 = \frac{1}{\sqrt{2}}(\phi_1 + \phi_2), \quad \psi_2 = \frac{1}{\sqrt{2}}(\phi_1 - \phi_2) \tag{16}$$

In terms of ϕ_1, ϕ_2,

$$2(\bar{\alpha}\mathcal{I} + \bar{\beta}(\mathcal{J}_1 + \mathcal{J}_2)) = \left\langle \sum_{\mathbf{k}} a(k)|\phi_1|^2 + (a(k) + \delta(k))|\phi_2|^2 \right\rangle \tag{17}$$

where

$$\alpha(k) = \bar{\alpha}k^2 + \bar{\beta}k^4, \quad \delta(k) = 2\varepsilon(\bar{\alpha} + 2\varepsilon(k^2 + \varepsilon)) \tag{18}$$

Here, ϕ_1 is the barotropic part of the flow, and ϕ_2, the baroclinic. According to standard thermodynamics, the most probable distribution of $\psi_1(\mathbf{k})$, $\psi_2(\mathbf{k})$ should be a negative exponential of $\bar{\alpha}\mathcal{I} + \bar{\beta}(\mathcal{J}_1 + \mathcal{J}_2)$, with $\bar{\alpha}$, $\bar{\beta}$ adjusted so that the invariants $\mathcal{I}$ and $\mathcal{J}_1 + \mathcal{J}_2$ have prescribed values. This yields;

$$\langle|\phi_1|^2\rangle(\mathbf{k}) \sim \frac{1}{a(k)}, \quad \langle|\phi_2|^2\rangle(\mathbf{k}) \sim \frac{1}{a(k) + \delta(k)} \tag{19}$$

$$\langle|\psi_1(\mathbf{k})|^2\rangle = \langle|\psi_2(\mathbf{k})|^2\rangle \sim \frac{1}{a(k)} + \frac{1}{a(k) + \delta(k)} \tag{20}$$

and,

$$R_{12} \equiv \frac{\langle\psi_1\psi_2\rangle}{\sqrt{\langle|\psi_1|^2\rangle\langle|\psi_2|^2\rangle}} = \frac{\delta(k)}{2a(k) + \delta(k)} \tag{21}$$

Thus, the flow approaches maximum two dimensionality at large scales $[R_{12}(k) \to 1,\ k \to 0]$. As k increases beyond $\sqrt{\varepsilon}$, levels 1 and 2 becoming increasingly decorrelated so that $R_{12}(k) \to 4\varepsilon^2/k^2$, as $k/\sqrt{\varepsilon} \to \infty$.[‡]

Distributions (19) to (21) hold for inviscid flows confined to a finite wavenumber domain and have no energy or enstrophy transfer among scales of differing sizes. They represent an end state toward which nonlinear interactions impel the system, but which molecular dissipation prevents.

[‡] The analysis sketched here has been given more substantial form by Salmon et al. (1978). For an application of equipartitioning ideas to ocean basin circulation, see Holloway (1992) and Griffa and Salmon (1989).

2 ENERGY AND ENSTROPHY CASCADE

We turn next to dissipative flows and the scale-size distribution of two-dimensional flows. For simplicity, we focus on barotropic flows, however much of the discussion may be made for baroclinic flows as well. We again ignore the variation of Coriolis force with latitude.

Although equipartitioning arguments of the last section are qualitatively instructive, they are unable to incorporate dissipative effects, which are essential in determining the scale-size distribution of the flow. In atmospheric flows molecular dissipation (viscous and diffusive effects) are centered at very small scales, well below the range of most meteorological interest. However, even if dissipation is negligible in a given spectral range of **k**, its presence at very small scales induces a flux of energy (or enstrophy, for QG flow) which is vital in determining scale-size distributions, such as that for $E(\mathbf{k})$, or $Q(\mathbf{k})$. Thus, for the viscous form of (4), the principal range of interest is that for which inertial forces $(\mathbf{u} \cdot \nabla\omega)$ far exceed dissipation $(\nu\nabla^2\omega)$, so that

$$R_e = |\mathbf{u} \cdot \nabla\omega|/(\nu|\nabla^2\omega|) \gg 1 \tag{21}$$

This equation defines a Reynolds number. It is clear that its value depends on specifying a length scale to estimate the gradient operator.

In three-dimensional flows, it is well known that as $R_e \to \infty$ dissipation acts to significantly reduce the total energy:

$$E(t) = \int_0^\infty dkE(k, t)$$

on a time scale $\sim(\sqrt{E}/\mathcal{L})^{-1}$, where $\mathcal{L}$ pertains to the energy containing range (that region in wavenumber that contributes most significantly to E). Thus, the decrease of total energy is independent of ν. The underlying physics is that of a cascade. If we discretize wavenumber space in bins Δk_i, in which Δk_1 contains most of the total energy, and $(\Delta k_{i+1}/\Delta k_i) = 2$, then energy is passed from a given bin to its right-hand neighbor at a rate independent of i, until a value of k is reached for which $R_e(1/k) \sim 1$. The range over which this constant flux is maintained is called the inertial range, and the range for which $R_e(k) \leq 1$, the dissipation range. The wavenumber distribution of energy in the inertial range is wholly calculable from the constant flux of energy, F, and the decay time scale noted above. Thus, the energy within $\Delta k_i \sim k_iE(k_i)$, and its residency time is $\sim 1/k_i\sqrt{k_iE(k_i)}$, so that $F \sim [k_iE(k_i)(k_i\sqrt{E(k_i)k_i}]$, or

$$E(k) = CF^{2/3}k^{-5/3} \tag{22}$$

A central question in turbulence theory is to determine the equation of motion of $E(k,t)$. We write this evolution equation in the form

$$\dot{E}(k,t) = T(k,t) - 2\nu k^2 E(k,t) + \mathcal{F}(k) \tag{23}$$

where, for simplicity, we assume isotropy. In (23) $\mathcal{F}(k)$ is a possible external forcing function. The energy transfer function, $T(k,t)$ represents the effects of the nonlinearity and is a functional of the energy spectrum, during the entire history of the flow. The idea of a constant of energy flux noted above for the inertial range may be more formally put in terms of (22) as

$$\mathcal{F} = -\int_0^k T(p,t)dp \tag{24}$$

where k is within the inertial range. Energy conservation for inviscid flows is expressed by

$$\int_0^\infty dk\, T(k,t) = 0 \tag{25}$$

We now discuss how ideas of cascade apply to two-dimensional and QG flows, for which both the energy and enstrophy are inviscid constants of motion. In this case, we must have in addition to (25),

$$\int dp\, p^2 T(p,t) = 0 \tag{26}$$

We illustrate the physics of energy and enstrophy transfer by means of a simple model of $T(k,t)$, originally proposed by Leith (1967) and modified by Pouquet et al. (1975). Leith proposed that T be modeled as a diffusion in wavenumber space, with a diffusion rate fixed by the large-scale strain field. It seems a plausible suggestion, and if we assume the straining field acts rapidly, such approximation follows in a nearly exact manner, if we assume the large-scale strain is independent of the vorticity upon which it acts (Kraichnan, 1968). The form for $T(k)$ is

$$T(k) = \frac{\partial}{\partial k}\left\{\frac{1}{k}\frac{\partial}{\partial k}\left\{\sqrt{\int_0^k p^2 dp\, E(p)}k^3 E(k)\right\}\right\} \tag{27}$$

We note that both $\int_0^\infty dk\, T(k)$, and $\int_0^\infty k^2 dk\, T(k)$ vanish, so that for $\nu = 0$ Eq. (27) conserves total energy and total enstrophy. In our discussion of three-dimensional turbulence, we saw that at large R_e, a cascade, together with a constant flux, was associated with a constant of motion (i.e., energy). The question we now address is how does the existence of two constants of motion translate into possible inertial ranges (with associated constant fluxes) for energy and enstrophy? The use of

Eq. (27) in an initial value problem shows that T spreads energy to both small and large k, but primarily to small k. The converse is true for $k^2E(k)$. However, for decaying turbulence, the only consistent constant flux inertial range is a forward fluxed enstrophy inertial range. Then defining

$$\eta = -d\mathcal{E}/dt \tag{28}$$

Eq. (28) gives

$$E(k) \sim \eta^{2/3}k^{-3}/O(\ln^{1/3}(k)) \tag{29}$$

On the other hand, we may verify by direct substitution that for (27), $T(k^{-5/3}) = 0$. These remarks apply to freely decaying turbulence, and the situation changes if we consider turbulence forced at some intermediate wavenumber $k_f[\mathcal{F}(\mathbf{k}) \sim \delta(k - k_f)$, for example]. Then we may solve Eqs. (23) to (27) outside the dissipation range to find:

$$E(k) \sim \eta^{2/3}k_f^{-3}(k_f/k)^{-5/3}, k \leq k_f, \sim \eta^{2/3}k^{-3}/(1 + 2\ln(k/k_f))^{1/3}, k > k_f \tag{30}$$

Thus, for stationary two-dimensional turbulence, energy is fluxed to small k, and enstrophy to large k. However, for a steady state to exist, the energy fluxed toward $k \rightarrow 0$ must somehow be dissipated by friction or some other large-scale dissipation.

The ideas of a dual cascade carries over to QG flow in which the continuous degrees of freedom are permitted in the vertical direction. First, we may note that the equipartitioning formalism implies in this case an isotropic form for $\psi(\mathbf{k}) = \psi(|\mathbf{k}|)$. This, combined with the form of the enstrophy inertial range sketched above comprises the basis for Charney's (1971) discussion of QG turbulence in the atmosphere. The problem has subsequently been examined *via* the statistical theory of turbulence (Herring, 1980) and high-resolution direct numerical simulation (McWilliams et al., 1994). Both studies examined decaying QG flows. In general, scales larger than the energy peak developed into strong two-dimensional flow with scales smaller than the energy peak mixed barotropic and baroclinic. Above k_{peak}, the statistical theory's prediction for $Q(\mathbf{k})$ displayed weak baroclinic anisotropy just beyond the energy peak, which weakens into isotropy as $k \rightarrow \infty$. The simulation, on the other hand, had a much more pronounced baroclinicity beyond the energy peak that persisted over the entire range $k\phi > k_{\text{peak}}$.

3 COHERENT STRUCTURES

Our considerations of spectra have said nothing of the structures imbedded in the turbulence, and in fact two-point covariances cannot distinguish between strain and vorticity, except to say that their root-mean-square (rms) values are equal. Thus, flows with intense, isolated vortices surrounded by expansive strain fields would have the same spectra as flows in which the vortex regions are spatially merged, inseparable from the strain regions. The latter situation could be obtained by picking the amplitudes, $Q(\mathbf{k})$ according to a Gaussian distribution. Such a field would have $T(k) = 0$, thereby excluding energy transfer, which gives rise to inertial ranges. This suggests that some simple measure of non-Gaussianity would be an indicator of the presence of structures, although the converse is not the case. The simplest such measure is the kurtosis,

$$\mathcal{K} \equiv \langle Q^4(\rho)\rangle / \langle Q^2(\rho)\rangle^2 \tag{31}$$

Here, the angular brackets indicate an average over the spatial domain of the flow. For the case in which $Q(\rho)$ consists of isolated pulses of width w separated by empty areas, of area A, $\mathcal{K} = A/w$, while for Gaussian fields, $\mathcal{K} = 3$.

Numerical simulations of large R_e two-dimensional decaying flows (McWilliams, 1984) indicate that flows that start with Gaussian initial conditions develop intense, well-separated vortices, and that $\mathcal{K}(t)$ systematically increases during the flows development with values of $\mathcal{K} = 60$ typical at late stages. The same comment may be made of homogeneous quasi-geostrophic flows. For comparison, similar studies of three-dimensional flows have values of $\mathcal{K} \sim 6$. As the flow develops, the vortices become more isolated and more circular, and the vortex cores probably do not participate in the cascade. Thus regions contributing to enstrophy cascade are confined to thin regions surrounding vortex cores, so that the turbulence lives on a subspace available to it, with transfer to small scales that is progressively reduced during the decay.

One simple consequence derived from Eqs. (23) to (27) is that the decay of enstrophy is $\mathcal{E}(t) \sim t^{-p}, p = 2$, first proposed by Batchelor (1969). Such conclusions are also reached by more complete statistical theories, which are reviewed by Lesieur (1990, p. 269 et seq.). But high-resolution numerical simulations show a much slower decay. Thus Carnavale et al. (1991) find $p = 0.5$, while Chasnov (1997) finds from his 4096 resolution simulation that $p = 0.8$. However, these results for decaying turbulence appear not to bring into question the nature of the inverse cascade discussed elsewhere [see, e.g., Gotoh (1998) or Foias and Chae (1993)].

4 STRATIFIED THREE-DIMENSIONAL TURBULENCE AND WAVES

Stably stratified turbulence shares certain features of two-dimensional turbulence; indeed it has been proposed that motion in the larger scales of the mesoscale are explained as inverse cascading quasi two-dimensional turbulence (Gage, 1978; Lilly, 1983). To explore this issue, we write the Boussinesq equations in a convenient nondimensional form:

$$(\partial_t - \nabla^2)\mathbf{u} = -\nabla p - \mathbf{u} \cdot \nabla \mathbf{u} + \hat{g} N \theta - 2\boldsymbol{\Omega} \times \mathbf{u} \tag{32}$$

$$(\partial_t - \sigma\nabla^2)\theta = -Nw - \mathbf{u} \cdot \nabla\theta \tag{33}$$

$$\nabla \cdot u = 0 \tag{34}$$

Here, $N = \sqrt{g\alpha\hat{\beta}T_0}$. We take the stratification to be constant and stable $N > 0$. Now introduce, in Fourier space, the representation

$$\mathbf{u} = \mathbf{e}_1\phi_1 + \mathbf{e}_2\phi_2, \mathbf{e}_1(\mathbf{k}) = \mathbf{k} \times \hat{\mathbf{g}}/|\mathbf{k} \times \hat{\mathbf{g}}|, \mathbf{e}_2(\mathbf{k}) = \mathbf{k} \times \mathbf{e}_1/|\mathbf{k} \times \mathbf{e}_1| \qquad (35)$$

where ϕ_1 describes horizontal motion (and hence its name "vortical mode"), while ϕ_2 describes gravity waves (for the case $\Omega = 0$). In terms of $(\phi_1\phi_2, \theta)$ (33) may be written:

$$\partial_t \begin{pmatrix} \phi_1 \\ \phi_2 \\ \theta \end{pmatrix} = M \begin{pmatrix} \phi_1 \\ \phi_2 \\ \theta \end{pmatrix} + \mathrm{NL}\{\phi, \theta\} \qquad (36)$$

Nonlinearities are indicated symbolically here by 'NL', and

$$M = \begin{pmatrix} 0 & 2\Omega\cos\varphi & 0 \\ -2\Omega\cos\varphi & 0 & -N\sin\varphi \\ 0 & N\sin\varphi & 0 \end{pmatrix} \qquad (37)$$

Here, $\hat{\mathbf{g}} \cdot \hat{\mathbf{k}} = \cos\varphi$. For simplicity, we take $\hat{\mathbf{g}}||\boldsymbol{\Omega}$. The eigenvalues of M are

$$\lambda = (0, \pm i\mathcal{R}), \mathcal{R} = \sqrt{4\Omega^2\cos^2\varphi + N^2\sin^2\varphi} \qquad (38)$$

Some insight may be gained from supposing that the nonlinear terms are weak and act as random agitators of (ϕ_1, ϕ_2, θ). Such a linearization of (36) is called rapid-distortion theory. Then for a given N, Ω the dominant structures activated would be those that minimize the frequency, λ. For $N/\Omega \gg 1$, gravity waves having $\vartheta \sim 0$ dominate, and for $N/\Omega \ll 1$, inertial waves with $\vartheta \sim \pi/2$ dominating. The former condition is satisfied by shear layers ("pancakes"), while the latter by vertically oriented columns. Of course, simple linear reasoning cannot resolve structural issues, for example, whether the shear layers are organized into circular pancakes or the vertical spacing between "pancakes."

Turbulence and waves coexist in stably stratified flows, but their time scales may be disparate: The time scale for waves is $\sim 1/N$, while the eddy turn over time scale is $1/\sqrt{k^3E(k)}$. Thus, if the scale-size distribution is less steep than k^{-3}, waves will dominate the large scales, and turbulence the small. If small scales are isotropic [with spectra (22)], the scale at which these two time scales are equal is $L_O = \sqrt{\varepsilon/N^3}$, known as the Ozmidov scale (Ozmidov 1965). Eddies whose $k > k_O$ are unstable to vertical overturning; and vice versa: at scales smaller than k_O, an rms Richardson number $R_i \sim N^2/|\partial_z u|^2 \leq 1$. We should mention here that this picture may be a gross oversimplification of the physics: There has yet to emerge a direct verification of it from either numerical simulations or experiments.

Stable stratification severely suppresses vertical eddy diffusion of scalar fields. Neglecting molecular processes, a vertically moving parcel of fluid exhausts its kinetic energy and stops. In reality the parcel's kinetic energy is refreshed through thermal exchanges with neighboring parcels and can continue its vertical migration. In terms of turbulence concepts, we may argue that the gravity wave part of the spectrum ($k \leq k_O$) contributes little to eddy diffusion, with only scales with $k > k_O$ active in vertical diffusion. Eddy diffusion may be estimated by treating each Fourier mode of the vertical velocity as a random component contributing to the parcel's velocity,

$$dZ_k(t)/dt = w(k,t),\ \kappa_{\text{eddy}} = \sum_{\mathbf{k}} \langle Z_k dZ_k/dt \rangle \sim \int dk E(k)\tau(k) \tag{39}$$

where $\tau(k)$ is the correlation time associated with $\langle w(k,t)w(-k,t')\rangle \sim 1/\sqrt{k^3E(k)}$. Here, $w(k,t)$ is the kth component of the Lagrangian vertical velocity field. The equation for κ_{eddy} may be found in Lesieur (1990, p. 285). If in the sum in (39) we suppress scales larger than $1/k_O$,

$$\kappa_{\text{eddy}} \sim \varepsilon/N^2 \tag{40}$$

This represents a reduction by a factor $(L_O/L)^{4/3}$ of the unstratified value of eddy diffusivity, $u_{\text{rms}}\ L \sim \varepsilon^{1/3}\ L^{4/3}$. Although the above estimate has been given a more quantitative form by the theory of Weinstock (1978), numerical simulations (Kimura and Herring, 1996) as well as observational studies (Britter et al., 1985) suggest much smaller eddy diffusivity than (40). It is interesting to note that rapid distortion theory is able to account qualitatively for the suppression of eddy diffusivity (Kaneda and Ishida, 2000).

The search for the "balanced" dynamics of (32) and (34) in the limit $\boldsymbol{\Omega} \to 0$, $N \to \infty$ is of vital interest. As noted above, it was thought early on that the "reduced system" would sufficiently resemble two-dimensional turbulence so that the latter paradigm would be useful in understanding the large horizontal scales of mesoscale variability [for which $E(k) \sim k^{-5/3}$ is observed]. Early analysis suggested two-dimensional turbulence in horizontal, independent layers (Riley et al., 1982; Lilly, 1983). Current mathematical reformulation of the reduced system (Majda and Grote, 1997) now seems able to explain how "pancakes" form, as well as their vertical spacing. Of course, with rotation present, quasi-geostrophy may be invoked as an explanation, but the difficulty is that as the smaller scales of mesoscale variability are approached, R_o [see (9)] becomes ~ 1 with the approximate $-\frac{5}{3}$ range still present. The issue has been examined numerically by Herring and Métais (1989), who concluded that forced flow did indeed become two dimensionally layered with horizontal motion that varies from layer to layer. But layer edges are rough, and frictional effects associated with roughness tends to seriously suppress inverse cascade. A recent study of the mesoscale variability by Lindborg (1999) indicates that the dynamics of the $\sim k^{-5/3}$ range (thought by Gage and Lilly to be inverse cascading two-dimensional turbulence) does not conform to that of inverse cascading two-dimensional turbulence. His analysis compared theoretical estimates of the third-order structure function [which gives a measure of $T(k, t)$ as in (23)] to that gathered from an ensemble of aircraft observations. It remains to be seen if the theory of Majda and Grote will be able to give insight into the issue of inverse cascade.

REFERENCES

Batchelor, G. K. (1969). Computation of the energy spectrum in homogeneous two-dimensional turbulence, *Phys. Fluids Suppl.* **12**(II), 233–239.

Britter, R. E., J. C. R. Hunt, G. L. Marsh, and W. H. Snyder (1983). The effects of stable stratification on the turbulent diffusion and the decay of grid turbulence, *J. Fluid Mech.* **127**, 27–44.

Carnavale, G. F., J. C. McWilliams, Y. Pomeau, J. B. Weiss, and W. R. Young (1991). Evolution of vortex statistics in two-dimensional turbulence, *Phys. Rev. Lett.* **66**, 2735–2737.

Charney, J. G. (1971). Quasigeostrophic turbulence, *J. Atmos. Sci.* **28**, 1087–1095.

Chasnov, J. R. (1997). On the decay of two-dimensional homogeneous turbulence, *Phys. Fluids* **9**, 171–180.

Foias, C., and D. Chae (1993). A probability measure representing 2-D homogeneous turbulence, in S. Kida (Ed.), *Unstable and Turbulent Motion of Fluid*, World Scientific, Singapore, pp. 131–140.

Gage, K. S. (1979). Evidence for a $k^{-5/3}$ law inertial range in mesoscale two-dimensional turbulence, *J. Atmos. Sci.* **36**, 1950–1954.

Galperin, B., S. Sukoriansky, and H.-P. Huang (2001). Universal n^{-5} spectrum of zonal flows on giant planets, *Phys. Fluids* **13**, 1545–1548.

Gotoh, T. (1998). Energy spectrum in the inertial and dissipation ranges of the two-dimensional steady turbulence, *Phys. Rev. E* **57**, 2984–2991.

Griffa, A., and R. Salmon (1989). Wind-driven ocean circulation and equilibrium statistical mechanics, *J. Marine Res.* **47**, 457–492.

Herring, J. R. (1980). Statistical theory of quasi-geostrophic turbulence, *J. Atmos. Sci.* **37**, 969–977.

Herring, J. R. and O. Métais (1989). Numerical experiments in forced stably stratified turbulence. *J. Fluid Mech.* **202**, 97–115.

Holloway, G. (1992). Representing topographic stress for large-scale ocean models, *J. Phys. Ocean* **22**, 1033–1046.

Holloway, G., and M. C. Hendershott (1977). Stochastic modeling for non-linear Rossby waves, *J. Fluid Mech.* **82**, 747–765.

Kaneda, Y., and T. Ishida (2000). Suppression of vertical diffusion in strongly stratified turbulence, *J. Fluid Mech.* **402**, 311–327.

Kimura, Y., and J. R. Herring (1996). Diffusion in stably stratified turbulence, *J. Fluid Mech.* **328**, 253–269.

Kraichnan, R. H. (1968). Small-scale structure convected by turbulence, *Phys. Fluids* **11**, 945–953.

Kraichnan, R. H. (1975). Statistical dynamics of two dimensional turbulence, *J. Fluid Mech.* **67**, 155–175.

Leith, C. E. (1971). Atmospheric predictability and two-dimensional turbulence, *J. Atmos. Sci.* **28**, 145–161.

Lesieur, M. (1990). *Turbulence in Fluids*, 2nd ed., Dordrecht, Kluwer Academic.

Lilly, D. G. (1983). Stratified turbulence and the mesoscale variability of the atmosphere, *J. Atmos. Sci.* **40**, 749–761.

Lindborg, E. (1999). Can the atmospheric energy spectrum be explained by two-dimensional turbulence? *J. Fluid Mech.* **388**, 259–288.

Majda, A. J., and M. J. Grote (1997). Model dynamics and vertical collapse in decaying strongly stratified flows, *Phys. Fluids* **9**(10), 2932–2940.

McWilliams, J. C. (1984). The emergence of isolated vortices in turbulent flows, *J. Fluid Mech.* **146**, 21–43.

McWilliams, J. C., J. B. Weiss, and I. Yavneh (1994). Anisotropy and coherent vortex structures in planetary turbulence, *Science* **264**, 410–413.

Ozmidov, R. V. (1965). On the turbulent exchange in a stably stratified ocean. *Izu. Accad. Sci. USSR Atmos. Oceanic Phys.* **1**, 493–497.

Pedlosky, J. (1987). *Geophysical Fluid Dynamics*, 2nd ed., New York, Springer.

Pouquet, A., M. Lesieur, J. C. Andre, and C. Basedevant (1975). Evolution of high Reynolds number two-dimensional turbulence, *J. Fluid Mech.* **72**, 305–319.

Riley, J. J., R. W. Metcalfe, and M. A. Weissman (1982). Direct numerical simulations of homogeneous turbulence in density stratified fluids, Proc. AIP Conf. On Nonlinear Properties of Internal Waves, (AIP, Woodbury, 1981) pp. 79–112.

Salmon, R., G. Holloway, and M. C. Hendershott (1978). The equilibrium statistical mechanics of simple quasi-geostrophic models, *J. Fluid Mech.* **75**, 691–703.

Weinstock, J. (1978). Vertical turbulent diffusion in a stably stratified fluid, *J. Atmos. Sci.* **35**, 1022–1027.

CHAPTER 7

PREDICTABILITY AND CHAOS

JEFFREY B. WEISS

Attempts to predict the weather are probably as old as humanity itself. One can imagine our earliest ancestors anxiously examining a stormy sky trying to decide whether to stay in their caves or venture out to hunt and gather. By the early nineteenth century, science viewed the universe as a deterministic and predictable clockworks, a view concisely expressed by Laplace (1825):

> An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up, if moreover it were vast enough to submit these data to analysis, would encompass in the same formula the movements of the greatest bodies of the universe and those of the lightest atoms. For such an intelligence nothing would be uncertain, and the future, like the past, would be open to its eyes.

Laplace's viewpoint reduces the problem of weather prediction to that of finding the right equations of motion, Laplace's forces, and the exact initial state of the appropriate system, Laplace's respective positions of the beings. Apart from the inherent randomness arising from the quantum nature of the universe, Laplace's view is still correct, but of limited utility. The problem is one of approximations. We will never know the exact equations for the entire universe and the exact positions and velocities of all its particles. For linear systems, such as an archetypical clock, approximate knowledge of only a small portion of the universe, the current location of the clock's hands to within some uncertainty, allows prediction of its entire future to within a similar uncertainty. However, for nonlinear dynamical systems, such as the atmosphere, the question is: How accurately do we need to know the equations and the initial state to make a prediction within a desired error?

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

In what follows, the focus will be on the atmosphere, but the issues are similar for any other component of the climate system and for its entirety.

Consider the one-dimensional linear dynamical system

$$\frac{dx}{dt} = ax \tag{1}$$

which has a solution

$$x(t) = e^{at}x(0) \tag{2}$$

If $a > 0$, the system is unstable and $x(t) \to \infty$, while if $a < 0$ the system is stable, $x(t) \to 0$. When $a < 0$, the point $x = 0$ is said to be the attractor of the dynamical system in that all initial conditions are attracted to $x = 0$ as $t \to \infty$.

Suppose we know the initial condition $x(0)$ and the parameter a to within (positive) uncertainties $\delta x(0)$ and δa, respectively. Then, if those uncertainties are small, the uncertainty, or error, $\delta x(t)$ in the prediction $x(t)$ for $t > 0$ is

$$\delta x(t) = e^{at}|x(0)|\delta a\ t + e^{at}\delta x(0) \tag{3}$$

When the system is unstable, the error, as well as the system itself, grows exponentially to infinity; but, when the system is stable, the error eventually decays to zero. Assume now that we know the dynamics exactly, $\delta a = 0$. Since the error growth is exponential, it is convenient to define the mean rate of exponential error growth λ as

$$\lambda = \frac{1}{t}\ln\frac{\delta x(t)}{\delta x(0)} = a \tag{4}$$

So stability corresponds with errors shrinking and negative λ, while instability causes error growth and positive λ.

Most natural systems are, however, not linear. The atmosphere, ocean, and climate system are presumed to be extremely high-dimensional deterministic dissipative dynamical systems. They are dynamical systems in that their state at a future time is a function of their state at a previous time. They are dissipative in that energy is not conserved but is input through forcing, e.g., shortwave solar radiation, and is drained away through damping, e.g., viscosity and outgoing longwave radiation. The dimensionality of a dynamical system refers to the number of variables, or degrees of freedom, needed to fully describe its state. The atmosphere is a continuous fluid described by partial differential equations and its dimensionality is infinite. Discretization of the continuum, necessary for numerical modeling, renders the dimensionality finite, but any reasonable discretization results in a high dimensionality. Current numerical weather prediction models have $O(10^6)$ degrees of freedom.

Lorenz (1975) classified predictability into the first and second kinds. Predictions of the first kind are deterministic predictions, predictions of the specific state of a system evolving from an initial condition. This is the kind of predictability addressed by Laplace: Given an initial state with some uncertainty how well can we predict the future? Predictability of the second kind takes over after predictability of the first kind is no longer possible. Once the errors in an initial value problem have grown so large that a prediction of the first kind is useless, all that can be predicted are the statistics of the system. Those statistics will depend on the properties of the external forces. Some of the external forces for the atmosphere are incoming solar radiation, sea surface temperature (SST), and CO_2 concentration. Prediction of the second kind, statistical prediction, attempts to predict how the structure of the attractor and its resulting statistics respond to changes in the forcing.

1 NONLINEAR DYNAMICS AND CHAOS

Dynamical systems theory discusses the time evolution of a physical system in terms of its phase space, which is defined by the collection of all the variables needed to determine the state of the system at a given time (Lichtenberg and Lieberman, 1992). A single state is then a point in phase space. For a simple pendulum, the state is defined by its position and velocity, and its phase space is two dimensional. The state of the atmosphere is defined by the temperature, wind velocity, humidity, etc. at every point in space. Determining the appropriate state variables of a complex system is often extremely difficult. As a continuous dynamical system evolves in time, the state changes continuously, and the point representing the system's state moves through phase space tracing out a one-dimensional curve called the phase space trajectory.

If one starts a dissipative dynamical system at an arbitrary initial condition, it will typically display some transient behavior and then eventually settle down to its long-term behavior. This long-term behavior is the system's attractor and can take several forms. If the system sits at an equilibrium state, such as the pendulum hanging straight down, then the attractor is a single point in phase space, called a stable fixed point. If the system undergoes regular oscillations, the attractor is a closed curve. If the system undergoes quasiperiodic oscillations, i.e., the superposition of oscillations with irrationally related frequencies, the attractor covers a torus. Motion on all of these attractors is regular and predictable.

Most nonlinear dynamical systems, however, do not display any of these behaviors. First, the attractors are typically fractals, i.e., complex sets with fractional dimension. Attractors with fractal dimension are said to be strange attractors. Second, even though the dynamics is restricted to evolve on an attractor, two nearby states will rapidly separate. Dynamics with this sensitive dependence on initial conditions is said to be chaotic. While there are examples of chaotic nonstrange attractors, as well as nonchaotic strange attractors, it seems that dissipative nonlinear systems typically have chaotic strange attractors and the two words are often used interchangeably. Sensitive dependence on initial conditions refers to predictions of the first kind, while the properties of the attractor as a whole are predictions of the second kind.

Sensitive dependence was first discovered by Lorenz using the so-called Lorenz equations:

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y-x)\\ \frac{dy}{dt} &= -xz+rx-y\\ \frac{dz}{dt} &= xy-bz\end{aligned} \tag{5}$$

Despite being only three dimensional, the Lorenz system and its variants continue to be used in predictability studies. The Lorenz equations display sensitive dependence on initial conditions (Fig. 1) and have a strange attractor (Fig. 2).

Sensitive dependence obviously has a significant impact on predictability and is quantified in terms of Lyapunov exponents, which are a generalization of the error growth rate defined in Eq. (4). Consider a nonlinear dynamical system

$$\frac{d\boldsymbol{y}}{dt} = \boldsymbol{F}(\boldsymbol{y}) \tag{6}$$

with a solution $\boldsymbol{X}(t)$. Consider an initially small perturbation to this trajectory, $\boldsymbol{x}(0)$, which gives a new trajectory $\boldsymbol{X}'(t) = \boldsymbol{X}(t) + \boldsymbol{x}(t)$. The difference, or error, $\boldsymbol{x}(t)$ evolves via the linearization of the dynamical equations:

$$\begin{aligned}\frac{d\boldsymbol{x}}{dt} &= \frac{d\boldsymbol{X}'}{dt} - \frac{d\boldsymbol{X}}{dt}\\ &= \boldsymbol{F}(\boldsymbol{X}+\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{X})\\ &= \mathbf{M}(\boldsymbol{X})\boldsymbol{x}\end{aligned} \tag{7}$$

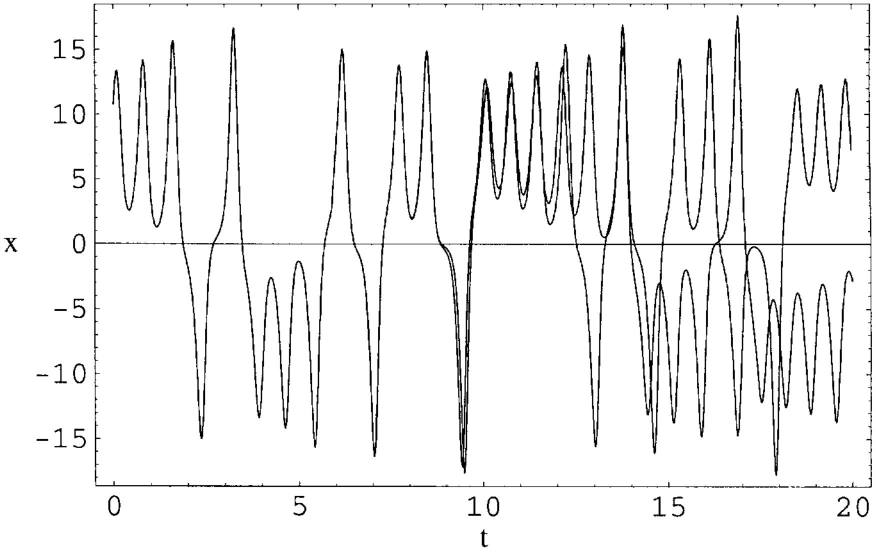


Figure 1 Two trajectories of the Lorenz equations with initial separation 0.00001.

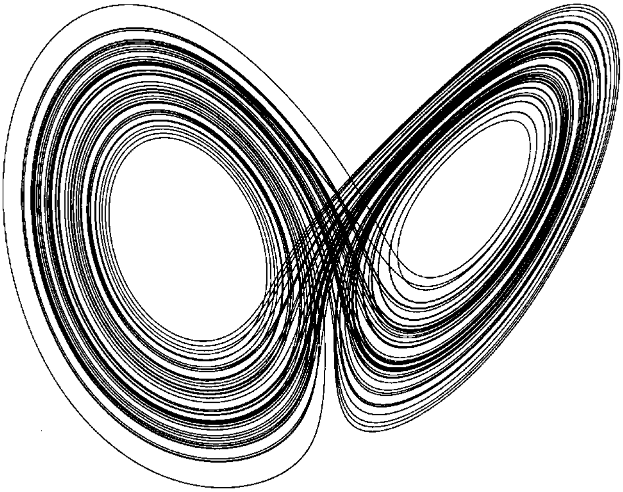


Figure 2 Lorenz attractor at $b=\frac{8}{3}$, $\sigma=10$, and $r=28$.

where the linearization of the dynamics at $\boldsymbol{X}$ is given by the matrix

$$M_{ij}(\boldsymbol{X}) = \left.\frac{\partial F_i(\boldsymbol{y})}{\partial y_j}\right|_{\boldsymbol{y}=\boldsymbol{X}} \tag{8}$$

and terms of $O(x^2)$ are ignored because the perturbation is assumed to be small.

The distance, or error, between the two trajectories is given by $d = \|\boldsymbol{x}\|$ where the norm $\|\cdot\|$ must be specified. The choice of norm is highly subjective and the appropriate choice depends on the situation. Some simple common norms are quadratic functions of the state variables, such as the total energy or total enstrophy. However, if your goal is to plan a picnic, then a local precipitation norm would be more appropriate.

Since Eq. (7) shows that the error growth is linear in the error we expect the error to grow exponentially as in (3). However, due to the nonlinear nature of the dynamics, when d becomes large enough that $O(d^2) \sim O(d)$, the linearized equation (7) breaks down. To follow error growth for long times requires either including nonlinear terms in the error growth equation or shrinking the initial d as the prediction time grows so that the final error remains small. Including nonlinear effects causes difficulty over long times since eventually the error grows to the size of the attractor and then remains roughly constant. Averaging error growth over these long times results in zero averaged error growth. A measure of the error growth over the whole attractor thus requires making the initial error smaller as the time gets longer. The result is the Lyapunov exponent λ:

$$\lambda = \lim_{\substack{t\to\infty \\ d(0)\to 0}} \frac{1}{t} \ln \frac{d(t)}{d(0)} \tag{9}$$

The Lyapunov exponent is thus the mean exponential error growth rate linearized about the true trajectory averaged over the entire attractor. If the Lyapunov exponent is positive, errors grow, while if it is negative, errors decay.

For an n-dimensional dynamical system there are actually n Lyapunov exponents, typically ordered from largest to smallest. Over long times, the largest Lyapunov exponent will dominate error growth and is given by Eq. (9). The spectrum of Lyapunov exponents can be understood in terms of the propagator matrix $\mathbf{G}$ of Eq. (7), defined by:

$$\boldsymbol{x}(t) = \mathbf{G}(t, 0)\boldsymbol{x}(0) \tag{10}$$

For a quadratic norm, the error is given by:

$$\begin{aligned} \|\boldsymbol{x}\|^2 &= \langle \mathbf{G}(t, 0)\boldsymbol{x}(0), \mathbf{G}(t, 0)\boldsymbol{x}(0)\rangle \\ &= \langle \boldsymbol{x}(0), \mathbf{G}(t, 0)^\dagger \mathbf{G}(t, 0)\boldsymbol{x}\rangle \end{aligned} \tag{11}$$

where $\mathbf{G}^\dagger$ is the adjoint of $\mathbf{G}$ with respect to the norm. The matrix $\mathbf{G}^\dagger\mathbf{G}$ is sometimes called the Oseledec matrix. The n eigenvalues σ_i and the n eigenvectors of $\mathbf{G}^\dagger\mathbf{G}$ are called the singular values and singular vectors, respectively, of $\mathbf{G}$. The Lyapunov exponents λ_i are related to the singular values by:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \ln \sigma_i \tag{12}$$

Since chaotic systems have sensitive dependence on initial conditions, their largest Lyapunov exponent must be positive. Positive largest Lyapunov exponents impose a practical limit on predictability. While the Laplacian view is still valid in that knowledge of the exact initial condition, $\boldsymbol{x}(0) = 0$, allows perfect prediction, any small errors in the initial condition grow rapidly. Increasing the length of time a forecast error is below some threshold requires decreasing the errors in the initial condition. Since errors grow exponentially, the gain in prediction time is much smaller than the associated decrease in initial error. In particular, decreasing the initial error by a factor of e results in an increase in prediction time of only $1/\lambda$. Current estimates for the atmosphere indicate that the maximum practical prediction time is about 2 weeks.

2 STOCHASTIC DYNAMICS

Although the atmosphere and other components of the climate system are fundamentally deterministic chaotic dynamical systems, they can often be considered as random systems. Due to the high dimensionality of the atmosphere, there is a broad range of space and time scales. If one is interested in the slower motion of the large scales, then the faster variation of the small scales can be parameterized as random noise. Loss of predictability in the resulting stochastic system is not due to the

Lyapunov exponent of the deterministic part of the system, but rather to the lack of predictability in the noise itself. Physically, however, since the noise is merely a parameterization of fast processes with large Lyapunov exponents, it is the deterministic chaos of the parameterized processes that ultimately is the cause of the loss of predictability.

The simplest such system is a deterministic stable fixed point perturbed by white noise,

$$\frac{d\boldsymbol{x}(t)}{dt} = \mathbf{A}\boldsymbol{x}(t) + \vec{\xi}(t), \tag{13}$$

where $\boldsymbol{x}$ represent the state of the system in terms of its deviation from a stable fixed point, $\mathbf{A}$ is the linear operator describing the deterministic dynamics, and $\vec{\xi}$ is multivariate Gaussian white noise. A stable fixed point requires that all of the eigenvalues of $\mathbf{A}$ be negative. In climatic applications, the fixed point is usually taken to be the mean state of the system and $\boldsymbol{x}$ represents the anomaly. This kind of stochastic system has been used to study a number of climatic systems including the El Niño–Souther Oscillation (ENSO) and extratropical atmospheric dynamics.

One of the most interesting features of these systems is that they can amplify the noise. As for the calculation of Lyapunov exponents, the growth of anomalies $d = \|\boldsymbol{x}\|$, is governed by Eq. (11). Now, however, the propagator $\mathbf{G}$ has both deterministic and stochastic components. In most climatic applications the deterministic dynamics is non-normal, i.e., $\mathbf{A}\mathbf{A}^{\dagger} - \mathbf{A}^{\dagger}\mathbf{A} \neq 0$. In this case the eigenvectors of $\mathbf{A}$ are not orthogonal. For normal systems, the stability of the fixed point ensures that all perturbations decay to zero monotonically. Non-normality, however, allows the nonorthogonal eigenvectors to interfere and transient perturbation growth is possible. For a given time interval t, the maximum growth ratio is the largest eigenvalue of $\mathbf{G}^{\dagger}(t, 0)\mathbf{G}(t, 0)$, which is the largest singular value of $\mathbf{G}$. The perturbation that grows the most is the eigenvector of $\mathbf{G}^{\dagger}(t, 0)\mathbf{G}(t, 0)$ corresponding to this eigenvalue, which is the corresponding singular vector of $\mathbf{G}$.

One can treat a time-evolving state perturbed by noise in a similar fashion. The deterministic dynamics is described by its linearization about the time-evolving state, Eq. (7), which is perturbed by adding noise. The propagator now depends on the time-evolving state, but the error growth is still determined by its singular values and singular vectors.

3 ENSEMBLE FORECASTING

When producing a forecast, it is becoming more common to provide an estimate of how correct the forecast is likely to be (Ehrendorfer, 1997). For a perfect model, forecast errors are due only to initial condition errors. The Lyapunov exponents are measures of error growth averaged over the entire attractor. However, the short-time error growth can vary significantly from one location on the attractor to another. This variation is described by so-called local Lyapunov exponents. The distribution of local Lyapunov exponents on the Lorenz attractor shows that the dynamics is least predictable near the unstable fixed point that divides the two lobes (Fig. 3). The variation in local Lyapunov exponents indicates that some atmospheric states are inherently more predictable than others.

An estimate of the believability of a forecast can be made by producing a large number, or ensemble, of forecasts, each starting from a different initial condition, where the ensemble of initial conditions is chosen to reflect the estimated distribution of initial condition errors. A forecast is more likely to be correct when the resulting distribution of forecasts is narrow than when it is large. Furthermore, any detailed structure in the distribution of forecasts indicates the type of forecast errors likely to occur.

Due to the cost of making forecasts, and the high dimensionality of atmospheric dynamics, it is impractical to use an ensemble large enough to adequately sample all possible initial condition errors. Since many directions in phase space correspond to stable directions, errors in these directions will decay and not affect the forecast quality. Thus, a good estimate of forecast error can be obtained with a small ensemble, provided the ensemble of initial conditions is distributed among the directions responsible for the largest error growth.

Currently, there are two main techniques for producing such an ensemble in an operational forecast: breeding methods and singular vectors. The European Center for Medium-Range Weather Forecasting (ECMWF) ensemble prediction system uses singular vectors, discussed above. The breeding method, used by National Center for Environmental Prediction (NCEP), starts two initial conditions some small distance apart and integrates them both using the fully nonlinear model resulting in a control solution and a perturbed solution. As the model evolves the two

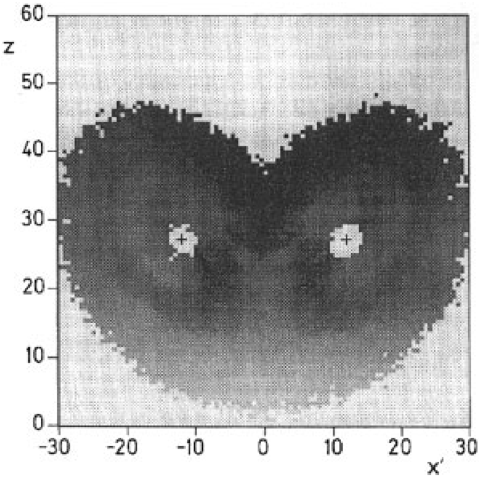


Figure 3 Distribution of local Lyapunov exponents on the Lorenz attractor, with darker (lighter) shadings indicating more (less) predictable regions (Eckhardt and Yao, 1993).

solutions separate. After some time interval the difference between the control and perturbation is rescaled back to its original amplitude and a new perturbed solution is begun. After this cycle has been repeated several times, components of the perturbation in stable and less unstable directions have been removed due to the rescalings. The resulting bred modes are then used to initialize the ensemble.

Ensemble forecasts accurately represent forecast error only if the model itself is accurate enough to approximately follow the true trajectory of the system. If the model error is large, then the entire ensemble will evolve into a completely different region in phase space than the true system. In this case, the ensemble spread is irrelevant to the forecast error.

4 SEASONAL TO CLIMATE PREDICTABILITY

Predicting the atmosphere beyond the roughly 2-week limit of deterministic predictability requires predicting its statistics rather than its actual state. Statistical prediction is possible if there is some slow external forcing that significantly affects the structure of the attractor. Note that the division between internal and external variability is somewhat subjective. For example, in predicting the effect of ENSO on the extratropics, the variation in tropical SST is considered an external forcing, while when considering the climate response to increasing CO_2, the behavior of tropical SST is internal variability.

The land and ocean surface provide boundary conditions for the atmosphere and, to the extent that they evolve slowly compared to the rapid dynamics of the weather, they allow longer term prediction. For example, the variation of tropical SSTs associated with ENSO evolves on a seasonal time scale. In addition, the relatively weak nonlinearity in the tropical atmosphere–ocean system allows successful predictions of ENSO well beyond the 2-week limit of extratropical weather. Thus, ENSO prediction provides some degree of climate prediction. The degree to which the state of the ENSO system affects the structure of the seasonal attractor can be seen in Figure 4, which shows the effects of an El Niño episode on December–February climate.

One factor that can significantly impact predictability is regime behavior, which occurs when a system spends significant amounts of time in localized regions of the attractor, followed by relatively rapid transitions to other localized regions of the attractor. For example, the two lobes of the Lorenz attractor in Figure 2 are regimes. There is very little understanding of the generic regime behavior of high dimensional chaotic attractors, and the role of regimes in intraseasonal to climate variability is still a subject of debate despite decades of research.

Intraseasonal regimes have been described in a number of ways including variations in the Rossby wave amplitude, blocking events, and teleconnection patterns. It has been suggested that the warming of the equatorial Pacific in 1976, which affected subsequent ENSO cycles, is an example of an interdecadal regime shift. Climate events such as the Little Ice Age may indicate centennial-scale regimes and glacial–interglacial transitions indicate regimes on even longer time scales.

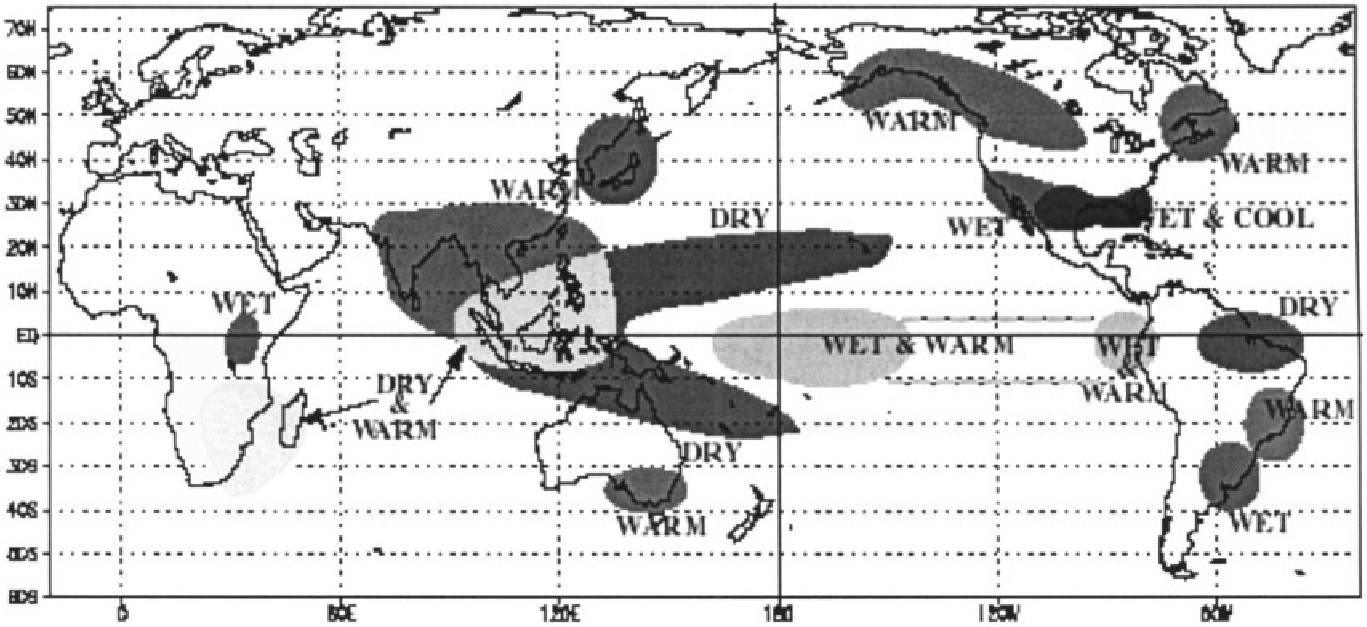


Figure 4 Effect of an El Niño episode on December–February climate (courtesy Climate Prediction Center, NOAA).

Since weather statistics depend on the regime of the atmosphere, one avenue for longer-term prediction is to focus on regime dynamics. Rather than predict the specific evolution of the system trajectory, the goal is to predict the sequence of regimes and their duration. Unfortunately, attempts at predicting the timing and outcome of regime transitions have not been particularly successful.

Recently, Palmer (1999) has analyzed the predictability of climate under the assumption that the climate attractor is composed of a number of distinct quasi-stationary regimes. Regimes are regions of the attractor that are relatively stable. For example, the regimes in the Lorenz attractor are localized around unstable fixed points, which, despite being unstable, have stable subspaces that attract trajectories. In the context of stochastic systems, the regimes may be modeled as truly stable attractors and the transitions are then solely due to the random noise. The areas of phase space between the regimes, the "saddles," are more unstable, and the system passes through these areas relatively quickly. An important feature of such saddles is that they show sensitive dependence: The relationship between the specific location where the system crosses the saddle and the regime it next visits can have fractal structure. Palmer noted that small-amplitude forcing will affect this fine-scale structure of the saddle much more than the relatively stable regime structure. The stability of the behavior of the attractor to external forcing is called structural stability. Thus, small-amplitude forcing will not change the structure of the regimes, i.e., they are structurally relatively stable, but will affect the relative probability of the system visiting a regime, which is structurally unstable. In the context of anthropogenic CO_2 increase, the resulting climate change will not show new climate regimes but, rather, a shift in the occurrence of existing regimes. Thus, the fact that recent climate changes are similar to naturally occurring variations does not disprove the possibility that the changes have an anthropogenic cause.

REFERENCES

Eckhardt, B., and D. Yao (1993). Local Lyapunov exponents in chaotic systems, *Physica D* **65**, 100–108.

ECMWF (1996). *Predictability*, Proceedings of a seminar held at ECMWF on predictability: 4–8 September 1995 Reading, European Centre for Medium-Range Weather Forecasts.

Ehrendorfer, M. (1997). Predicting the uncertainty of numerical weather forecasts: a review, *Meteorol. Z.* **6**, 147–183.

Laplace, P. S. (1825). *Philosophical Essay on Probabilities*, translated from the fifth French edition of 1825 by A. I. Dale, New York, Springer, 1995.

Lichtenberg, A. J., and M. A. Lieberman (1992). *Regular and Chaotic Dynamics*, New York, Springer.

Lorenz, E. N. (1975). Climatic predictability, in *The Physical Basis of Climate and Climate Modelling*, GARP Publication Series, (ICSU/WMO, Geneva), Vol. **16**, pp. 132–136.

NATO (1996). *Decadal Climate Variability: Dynamics and Predictability*, Proceedings of the NATO Advanced Study Institute held at Les Houches, France, February 13–24, 1995, L. T. Anderson and J. Willebrand (eds.), Berlin, Springer.

Palmer, T. N. (1999). A nonlinear dynamical perspective on climate prediction, *J. Climate* **12**, 575–591.

CHAPTER 8

HISTORICAL OVERVIEW OF NUMERICAL WEATHER PREDICTION

EUGENIA KALNAY

1 INTRODUCTION

In general, the public is not aware that our daily weather forecasts start out as initial-value problems on the major National Weather Service supercomputers. Numerical weather prediction provides the basic guidance for weather forecasting beyond the first few hours. For example, in the United States, computer weather forecasts issued by the National Centers for Environmental Prediction (NCEP) in Washington, DC, guide forecasts from the U.S. National Weather Service (NWS). NCEP forecasts are performed running (integrating in time) computer models of the atmosphere that can simulate, given today's weather observations, the evolution of the atmosphere in the next few days. Because the time integration of an atmospheric model *is an initial-value problem*, the ability to make a skillful forecast requires both that *the computer model be a realistic representation of the atmosphere and that the initial conditions be accurate*. In what follows we will give examples of the evolution of numerical weather prediction at NCEP, but they are representative of what has taken place in all major international operational weather centers such as the European Centre for Medium Range Weather Forecasts (ECMWF), the United Kingdom Meteorological Office (UKMO), and the weather services of Japan, Canada, and Australia.

Formerly the National Meteorological Center (NMC), the NCEP has performed operational computer weather forecasts since the 1950s. From 1955 to 1973, the forecasts included only the Northern Hemisphere; they have been global since 1973. Over the years, the quality of the models and methods for using atmospheric observations has improved continuously, resulting in major forecast improvements.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

Figure 1 shows the longest available record of the skill of numerical weather prediction. The S_1 score (Teweles and Wobus, 1954) measures the relative error in the horizontal gradient of the height of the constant-pressure surface of 500 hPa (in the middle of the atmosphere, since the surface pressure is about 1000 hPa) for 36-h forecasts over North America. Empirical experience at NMC indicated that a value of this score of 70% or more corresponds to a useless forecast, and a score of 20% or less corresponds to an essentially perfect forecast. Twenty percent was the average S_1 score obtained when comparing analyses hand-made by several experienced forecasters fitting the same observations over the data-rich North America region.

Figure 1 shows that current 36-h 500-hPa forecasts over North America are close to what was considered "perfect" 30 years ago: The forecasts are able to locate generally very well the position and intensity of the large-scale atmospheric waves, major centers of high and low pressure that determine the general evolution of the weather in the 36-h forecast. Smaller-scale atmospheric structures, such as fronts, mesoscale convective systems that dominate summer precipitation, etc., are still difficult to forecast in detail, although their prediction has also improved very significantly over the years. Figure 1 also shows that the 72-h forecasts of today are as accurate as the 36-h forecasts were 10 to 20 years ago. Similarly, 5-day forecasts, which had no useful skill 15 years ago, are now moderately skillful,

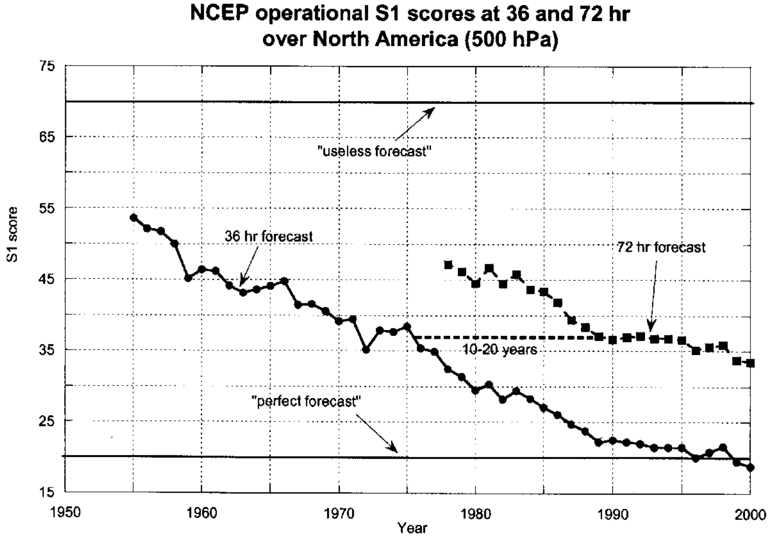


Figure 1 Historic evolution of the operational forecast skill of the NCEP (formerly NMC) models over North America. The S_1 score measures the relative error in the horizontal pressure gradient, averaged over the region of interest. The values $S_1 = 70\%$ and $S_1 = 20\%$ were empirically determined to correspond, respectively, to a "useless" and a "perfect" forecast when the score was designed. Note that the 72-h forecasts are currently as skillful as the 36-h forecasts were 10 to 20 years ago. (Data courtesy C. Vlcek, NCEP.)

and during the winter of 1997–1998, ensemble forecasts for the second week average showed useful skill (defined as anomaly correlation close to 60% or higher).

The improvement in skill over the last 40 years of numerical weather prediction apparent in Figure 1 is due to four factors:

- Increased power of supercomputers, allowing much finer numerical resolution and fewer approximations in the operational atmospheric models.
- Improved representation of small-scale physical processes (clouds, precipitation, turbulent transfers of heat, moisture, momentum, and radiation) within the models.
- Use of more accurate methods of data assimilation, which result in improved initial conditions for the models.
- Increased availability of data, especially satellite and aircraft data over the oceans and the Southern Hemisphere.

In the United States, research on numerical weather prediction takes place in national laboratories of the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), the National Center for Atmospheric Research (NCAR), and in universities and centers such as the Center for Prediction of Storms (CAPS). Internationally, major research takes place in large operational national and international centers (such as the European Center for Medium Range Weather Forecasts, NCEP, and the weather services of the United Kingdom, France, Germany, Scandinavian, and other European countries, Canada, Japan, Australia, and others). In meteorology there has been a long tradition of sharing both data and research improvements, with the result that progress in the science of forecasting has taken place in many fronts, and all countries have benefited from this progress.

2 PRIMITIVE EQUATIONS, GLOBAL AND REGIONAL MODELS, AND NONHYDROSTATIC MODELS

As envisioned by Charney (1951, 1960), the filtered (quasi-geostrophic) equations, introduced by Charney et al. (1950), were found not accurate enough to allow continued progress in Numerical Weather Prediction (NWP), and were eventually replaced by primitive equations models. These equations are conservation laws applied to individual parcels of air: conservation of the three-dimensional momentum (equations of motion), conservation of energy (first law of thermodynamics), conservation of dry air mass (continuity equation), and equations for the conservation of moisture in all its phases, as well as the equation of state for perfect gases. They include in their solution fast gravity and sound waves, and therefore in their space and time discretization they require the use of a smaller time step. For models with a horizontal grid size larger than 10 km, it is customary to replace the vertical component of the equation of motion with its hydrostatic approximation, in which the vertical acceleration is neglected compared with gravitational acceleration (buoyancy). With this approximation, it is convenient to use atmospheric pressure, instead of height, as a vertical coordinate.

The continuous equations of motions are solved by discretization in space and in time using, for example, finite differences. It has been found that the accuracy of a model is very strongly influenced by the spatial resolution: In general, the higher the resolution, the more accurate the model. Increasing resolution, however, is extremely costly. For example, doubling the resolution in the 3-space dimensions also requires halving the time step in order to satisfy conditions for computational stability. Therefore, the computational cost of a doubling of the resolution is a factor of 2^4 (3-space and one time dimension). Modern methods of discretization attempt to make the increase in accuracy less onerous by the use of implicit and semi-Lagrangian time schemes (Robert, 1981), which have less stringent stability conditions on the time step, and by the use of more accurate space discretization. Nevertheless, there is a constant need for higher resolution in order to improve forecasts, and as a result running atmospheric models has always been a major application of the fastest supercomputers available.

When the "conservation" equations are discretized over a given grid size (typically a few kilometers to several hundred kilometers) it is necessary to add "sources and sinks," terms due to small-scale physical processes that occur at scales that cannot be explicitly resolved by the models. As an example, the equation for water vapor conservation on pressure coordinates is typically written as:

$$\frac{\partial \bar{q}}{\partial t} + \bar{u}\frac{\partial \bar{q}}{\partial x} + \bar{v}\frac{\partial \bar{q}}{\partial y} + \bar{\omega}\frac{\partial \bar{q}}{\partial p} = \bar{E} - \bar{C} + \frac{\partial \overline{\omega' q'}}{\partial x}$$

where q is the ratio between water vapor and dry air, x and y are horizontal coordinates with appropriate map projections, p pressure, t time, u and v are the horizontal air velocity (wind) components, $\omega = dp/dt$ is the vertical velocity in pressure coordinates, and the primed product represents turbulent transports of moisture on scales unresolved by the grid used in the discretization, with the overbar indicating a spatial average over the grid of the model. It is customary to call the left-hand side of the equation, the "dynamics" of the model, which are computed explicitly.

The right-hand side represents the so-called physics of the model, i.e., for this equation, the effects of physical processes such as evaporation, condensation, and turbulent transfers of moisture, which take place at small scales that cannot be explicitly resolved by the dynamics. These subgrid-scale processes, which are sources and sinks for the equations, are then "parameterized" in terms of the variables explicitly represented in the atmospheric dynamics.

Two types of models are in use for numerical weather prediction: global and regional models. Global models are generally used for guidance in medium-range forecasts (more than 2 days), and for climate simulations. At NCEP, for example, the global models are run through 16 days every day. Because the horizontal domain of global models is the whole Earth, they usually cannot be run at high resolution. For more detailed forecasts it is necessary to increase the resolution, and this can only be done over limited regions of interest, because of computer limitations.

Regional models are used for shorter-range forecasts (typically 1 to 3 days) and are run with resolutions several times higher than global models. In 1997, the NCEP global model was run with 28 vertical levels, and a horizontal resolution of 100 km for the first week, and 200 km for the second week. The regional (Eta) model was run with a horizontal resolution of 48 km and 38 levels, and later in the day with 29 km and 50 levels. Because of their higher resolution, regional models have the advantage of higher accuracy and ability to reproduce smaller scale phenomena such as fronts, squall lines, and orographic forcing much better than global models. On the other hand, regional models have the disadvantage that, unlike global models, they are not "self-contained" because they require lateral boundary conditions at the borders of the horizontal domain. These boundary conditions must be as accurate as possible because otherwise the interior solution of the regional models quickly deteriorates. Therefore, it is customary to "nest" the regional models within another model with coarser resolution, whose forecast provides the boundary conditions. For this reason, regional models are used only for short-range forecasts. After a certain period, proportional to the size of the model, the information contained in the high-resolution initial conditions is "swept away" by the influence of the boundary conditions, and the regional model becomes merely a "magnifying glass" for the coarser model forecast in the regional domain. This can still be useful, for example, in climate simulations performed for long periods (seasons to multiyears), and which therefore tend to be run at coarser resolution. A "regional climate model" can provide a more detailed version of the coarse climate simulation in a region of interest.

More recently the resolution of regional models has been increased to just a few kilometers in order to resolve better mesoscale phenomena. Such storm-resolving models as the Advanced Regional Prediction System (ARPS) cannot be hydrostatic since the hydrostatic approximation ceases to be accurate for horizontal scales of the order of 10 km or smaller. Several major nonhydrostatic models have been developed and are routinely used for mesoscale forecasting. In the United States the most widely used are the ARPS, the MM5, the RSM, and the U.S. Navy model. There is a tendency toward the use of nonhydrostatic models that can be used globally as well.

3 DATA ASSIMILATION: DETERMINATION OF INITIAL CONDITIONS FOR NWP PROBLEM

As indicated previously, NWP is an initial value problem: Given an estimate of the present state of the atmosphere, the model simulates (forecasts) its evolution. The problem of determination of the initial conditions for a forecast model is very important and complex, and has become a science in itself (Daley, 1991). In this brief section we introduce the main methods used for this purpose [successive corrections method (SCM), optimal interpolation (OI), variational methods in three and four dimensions, 3D-Var and 4D-Var, and Kalman filtering (KF)]. More detail is available in Chapter 5 of Kalnay (2001) and Daley (1991).

In the early experiments, Richardson (1922) and Charney et al. (1950) performed hand interpolations of the available observations, and these fields of initial conditions were manually digitized, which was a very time-consuming procedure. The need for an automatic "objective analysis" became quickly apparent (Charney, 1951), and interpolation methods fitting data were developed (e.g., Panofsky, 1949; Gilchrist and Cressman, 1954; Barnes, 1964).

There is an even more important problem than spatial interpolation of observations: There is not enough data to initialize current models. Modern primitive equations models have a number of degrees of freedom on the order of 10^7. For example, a latitude–longitude model with typical resolution of one degree and 20 vertical levels would have $360 \times 180 \times 20 = 1.3 \times 10^6$ grid points. At each grid point we have to carry the values of at least 4 prognostic variables (two horizontal wind components, temperature, moisture) and surface pressure for each column, giving over 5 million variables that need to be given an initial value. For any given time window of ± 3 h, there are typically 10,000 to 100,000 observations of the atmosphere, two orders of magnitude fewer than the number of degrees of freedom of the model. Moreover, their distribution in space and time is very nonuniform (Fig. 2), with regions like North America and Eurasia, which are relatively data rich, and others much more poorly observed.

For this reason, it became obvious rather early that it was necessary to use additional information (denoted background, first guess, or prior information) to prepare initial conditions for the forecasts (Bergthorsson and Döös, 1955). Initially climatology was used as a first guess (e.g., Gandin, 1963), but as forecasts became

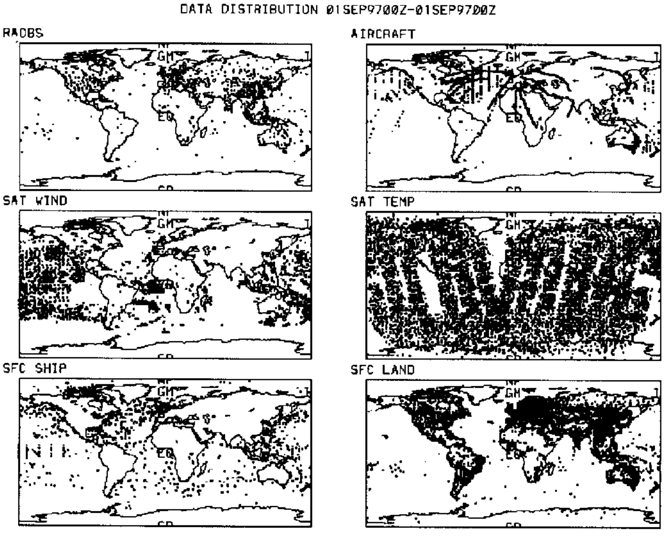


Figure 2 Typical distribution observations in a ±3-h window.

better, a short-range forecast was chosen as first guess in the operational data assimilation systems or "analysis cycles." The intermittent data assimilation cycle shown schematically in Figure 3 is continued in present-day operational systems, which typically use a 6-h cycle performed 4 times a day.

In the 6-h data assimilation cycle for a global model, the model 6-h forecast x^b (a three-dimensional array) is interpolated to the observation location and, if needed, converted from model variables to observed variables y^o (such as satellite radiances or radar reflectivities). The first guess of the observations is therefore $H(x^b)$, where H is the observation operator that performs the necessary interpolation and transformation. The difference between the observations and the model first guess $y^o - H(x^b)$ is denoted "observational increments" or "innovations." The analysis x^a is obtained by adding the innovations to the model forecast (first guess) with weights that are determined based on the estimated statistical error covariances of the forecast and the observations:

$$x^a = x^b + W[y^o - H(x^b)] \tag{1}$$

Different analysis schemes (SCM, OI, variational methods, and Kalman filtering) differ by the approach taken to combine the background and the observations to

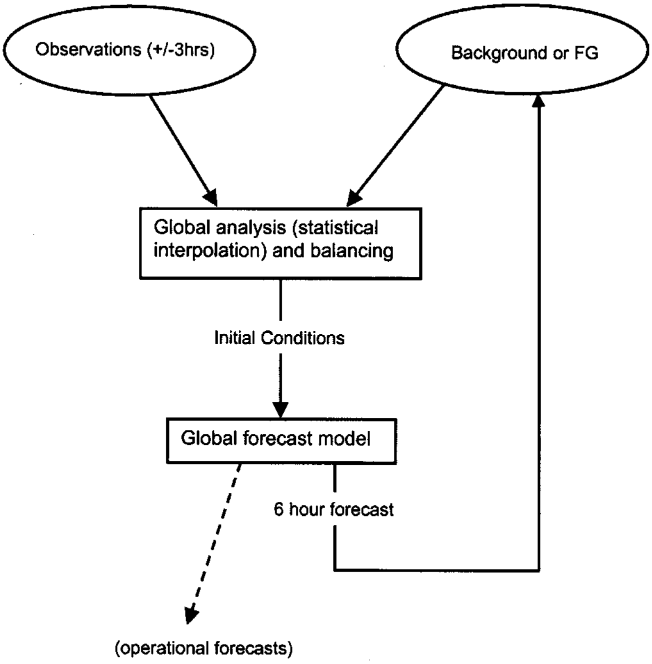


Figure 3 Flow diagram of a typical intermittent (6 h) data assimilation cycle. The observations gathered within a window of about ± 3 h are statistically combined with a 6-h forecast denoted background or first guess. This combination is called "analysis" and constitutes the initial condition for the global model in the next 6-h cycle.

produce the analysis. In optimal interpolation (OI; Gandin, 1963) the matrix of weights W is determined from the minimization of the analysis errors. Lorenc (1986) showed that there is an equivalency of OI to the variational approach (Sasaki, 1970) in which the analysis is obtained by minimizing a cost function:

$$J = \frac{1}{2}\{[y^o - H(x)]^T R^{-1}[y^o - H(x)] + (x - x^b)^T B^{-1}(x - x^b)\} \tag{2}$$

This cost function J in (2) measures the distance of a field x to the observations (first term in the cost function) and the distance to the first guess or background x^b (second term in the cost function). The distances are scaled by the observation error covariance R and by the background error covariance B, respectively. The minimum of the cost function is obtained for $x = x^a$, which is the analysis. The analysis obtained in (1) and (2) is the same if the weight matrix in (1) is given by

$$W = BH^T(HBH^T + R^{-1})^{-1}[y^o - H(x)^b] \tag{3}$$

In (3) **H** is the linearization of the transformation H.

The difference between optimal interpolation (1) and the three-dimensional variational (3D-Var) approach (2) is in the method of solution: In OI, the weights W are obtained using suitable simplifications. In 3D-Var, the minimization of (2) is performed directly, and therefore allows for additional flexibility.

Earlier methods such as the successive corrections method, (SCM; Bergthorsson and Döös, 1955; Cressman, 1959; Barnes, 1964) were of a form similar to (1), but the weights were determined empirically, and the analysis corrections were computed iteratively. Bratseth (1986) showed that with a suitable choice of weights, the simpler SCM solution will converge to the OI solution. More recently, the variational approach has been extended to four dimensions, by including in the cost function the distance to observations over a time interval (assimilation window). A first version of this expensive method was implemented at ECMWF at the end of 1997 (Bouttier and Rabier, 1998; Andersson et al., 1998). Research on the even more advanced and computationally expensive Kalman filtering [e.g., Ghil et al. (1981) and ensemble Kalman filtering, Evensen and Van Leeuwen (1996), Houtekamer and Mitchell (1998)] is included in Chapter 5 of Kalnay (2001). That chapter also discusses in some detail the problem of enforcing a balance in the analysis in such a way that the presence of gravity waves does not mask the meteorological signal, as it happened to Richardson (1922). This "initialization" problem was approached for many years through "nonlinear normal mode initialization" (Machenauer, 1976; Baer and Tribbia, 1976), but more recently balance has been included as a constraint in the cost function (Parrish and Derber, 1992), or a digital filter has been used (Lynch and Huang, 1994).

In the analysis cycle, no matter which analysis scheme is used, the use of the model forecast is essential in achieving "four-dimensional data assimilation" (4DDA). This means that the data assimilation cycle is like a long model integration in which the model is "nudged" by the data increments in such a way that it remains

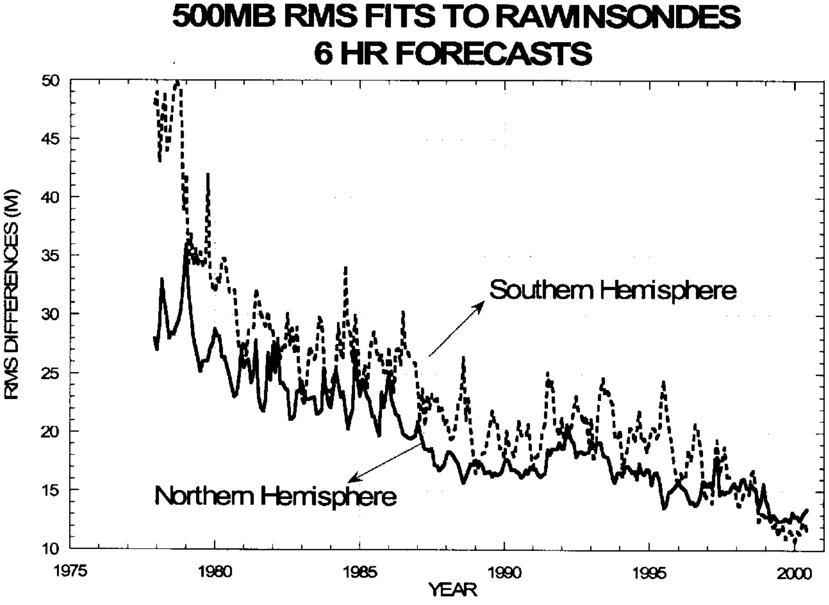


Figure 4 RMS observational increments (differences between 6-h forecast and rawinsonde observations) for 500-hPa heights. (Data courtesy of Steve Lilly, NCEP.)

close to the real atmosphere. The importance of the model cannot be overemphasized: It transports information from data-rich to data-poor regions, and it provides a complete estimation of the four-dimensional state of the atmosphere. Figure 4 presents the root-mean-square (rms) difference between the 6-h forecast (used as a first guess) and the rawinsonde observations from 1978 to the present (in other words, the rms of the observational increments for 500-hPa heights). It should be noted that the rms differences are not necessarily forecast errors since the observations also contain errors. In the Northern Hemisphere the rms differences have been halved from about 30 m in the late 1970s to about 15 m in the present, equivalent to a mean temperature error of about 0.75 K, not much larger than rawinsonde observational errors. In the Southern Hemisphere the improvements are even larger, with the differences decreasing from about 47 m to about 17 m, close to the present forecast error in the Northern Hemisphere. The improvements in these short-range forecasts are a reflection of improvements in the model, the analysis scheme used in the data assimilation, and, to a lesser extent, the quality and quality control of the data.

4 OPERATIONAL NUMERICAL PREDICTION

Here we focus on the history of operational numerical weather prediction at the NMC (now NCEP), as reviewed by Shuman (1989) and Kalnay et al. (1998), as an example of how major operational centers have evolved. In the United States operational numerical weather prediction started with the organization of the Joint Numerical Weather Prediction Unit (JNWPU) on July 1, 1954, staffed by members of the U.S. Weather Bureau (later National Weather Service, NWS), the Air Weather Service of the U.S. Air Force, and the Naval Weather Service.* Shuman (1989) pointed out that in the first few years, numerical predictions could *not* compete with those produced manually. They had several serious flaws, among them overprediction of cyclone development. Far too many cyclones were predicted to deepen into storms. With time, and with the joint work of modelers and practicing synopticians, major sources of model errors were identified, and operational NWP became the central guidance for operational weather forecasts.

Shuman (1989) included a chart with the evolution of the S_1 score (Teweles and Wobus, 1954), the first measure of error in a forecast weather chart that, according to Shuman (1989), was designed, tested, and modified to correlate well with expert forecasters' opinions on the quality of a forecast. The S_1 score measures the average relative error in the pressure gradient (compared to a verifying analysis chart). Experiments comparing two independent subjective analyses of the same data-rich North American region made by two experienced analysts suggested that a "perfect" forecast would have an S_1 score of about 20%. It was also found empirically that forecasts with an S_1 score of 70% or more were useless as synoptic guidance.

Shuman (1989) pointed out some of the major system improvements that enabled NWP forecasts to overtake and surpass subjective forecasts. The first major improvement took place in 1958 with the implementation of a barotropic (one-level) model, which was actually a reduction from the three-level model first tried, but which included better finite differences and initial conditions derived from an objective analysis scheme (Bergthorsson and Döös, 1954; Cressman, 1959). It also extended the domain of the model to an octagonal grid covering the Northern Hemisphere down to 9 to 15°N. These changes resulted in numerical forecasts that for the first time were competitive with subjective forecasts, but in order to implement them JNWPU (later NMC) had to wait for the acquisition of a more powerful supercomputer, an IBM 704, replacing the previous IBM 701. This pattern of forecast improvements, which depend on a combination of better use of the data and better models, and would require more powerful supercomputers in order to be executed in a timely manner has been repeated throughout the history of operational NWP. Table 1 (adapted from Shuman, 1989) summarizes the major improvements in the first 30 years of operational numerical forecasts at the NWS. The first primitive equations model (Shuman and Hovermale, 1968) was implemented in 1966. The first regional system (Limited Fine Mesh or LFM model; Howcroft, 1971) was implemented in 1971. It was remarkable because it remained in use for over 20 years, and it was the basis for Model Output Statistics (MOS). Its development was frozen in 1986. A more advanced model and data assimilation system, the Regional Analysis and Forecasting System (RAFS) was implemented as the main guidance for North

*In 1960 the JNWPU divided into three organizations: the National Meteorological Center (National Weather Service), the Global Weather Central (U.S. Air Force), and the Fleet Numerical Oceanography Center (U.S. Navy).

TABLE 1 Major Operational Implementations and Computer Acquisitions at NMC between 1955 and 1985

Year	Operational Model	Computer
1955	Princeton three-level quasi-geostrophic model (Charney, 1954). Not used by the forecasters	IBM 701
1958	Barotropic model with improved numerics, objective analysis initial conditions, and octagonal domain	IBM 704
1962	Three-level quasi-geostrophic model with improved numerics	IBM 7090 (1960) IBM 7094 (1963)
1966	Six-layer primitive equations model (Shuman and Hovermale, 1968)	CDC 6600
1971	Limited-area fine mesh (LFM) model (Howcroft, 1971) (first regional model at NMC)	
1974	Hough functions analysis (Flattery, 1971)	IBM 360/195
1978	Seven-layer primitive equation model (hemispheric)	
1978	Optimal interpolation (Bergman, 1979)	Cyber 205
Aug 1980	Global spectral model, R30/12 layers (Sela, 1982)	
March 1985	Regional Analysis and Forecast System based on the Nested Grid Model (NGM; Phillips, 1979) and optimal interpolation (DiMego, 1988)	

Adapted from Shuman (1989).

America in 1982. The RAFS was based on the multiple Nested Grid Model (NGM; Phillips, 1979) and on a regional optimal interpolation (OI) scheme (DiMego, 1988). The global spectral model (Sela, 1982) was implemented in 1980.

Table 2 (from Kalnay et al., 1998) summarizes the major improvements implemented in the global system starting in 1985 with the implementation of the first comprehensive package of physical parameterizations from GFDL. Other major improvements in the physical parameterizations were made in 1991, 1993, and 1995. The most important changes in the data assimilation were an improved OI formulation in 1986, the first operational three-dimensional variational data assimilation (3D-VAR) in 1991, the replacement of the satellite retrievals of temperature with the direct assimilation of cloud-cleared radiances in 1995, and the use of "raw" (not cloud-cleared) radiances in 1998. The model resolution was increased in 1987, 1991, and 1998. The first operational ensemble system was implemented in 1992 and enlarged in 1994.

Table 3 contains a summary for the regional systems used for short-range forecasts (to 48 h). The RAFS (triple-nested NGM and OI) were implemented in 1985. The Eta model, designed with advanced finite differences, step-mountain coordi-

TABLE 2 Major Operational Implementations and Computer Acquisitions at NMC between 1955 and 1985

Year	Operational Model	Computer Acquisition
April 1985	GFDL physics implemented on the global spectral model with silhouette orography, R40/18 layers	
Dec. 1986	New optimal interpolation code with new statistics	
1987		2nd Cyber 205
Aug. 1987	Increased resolution to T80/18 layers, Penman-Montieth evapotranspiration and other improved physics (Caplan and White, 1989; Pan, 1989)	
Dec. 1988	Implementation of hydrostatic complex quality control (Gandin, 1988)	
1990		Cray YMP/8cpu/ 32 megawords
Mar. 1991	Increased resolution to T126 L18 and improved physics, mean orography (Kanamitsu et al., 1991)	
June 1991	New 3D variational data assimilation (Parrish and Derber, 1992; Derber et al., 1991)	
Nov. 1991	Addition of increments, horizontal, and vertical OI checks to the CQC (Collins and Gandin, 1990)	
Dec. 7, 1992	First ensemble system: one pair of bred forecasts at 00Z to 10 days, extension of AVN to 10 days (Toth and Kalnay, 1993; Tracton and Kalnay, 1993)	
Aug. 1993	Simplified Arakawa-Schubert cumulus convection (Pan and Wu, 1995). Resolution T126/28 layers	
Jan. 1994		Cray C90/16cpu/ 128 megawords
March 1994	Second ensemble system: 5 pairs of bred forecasts at 00Z, 2 pairs at 12Z, extension of AVN, a total of 17 global forecasts every day to 16 days	
Jan. 10, 1995	New soil hydrology (Pan and Mahrt, 1987), radiation, clouds, improved data assimilation; reanalysis model	

(*continued*)

TABLE 2 (*Continued*)

Year	Operational Model	Computer Acquisition
Oct. 25, 1995	Direct assimilation of TOVS cloud-cleared radiances (Derber and Wu, 1997). New PBL based on nonlocal diffusion (Hong and Pan, 1996); improved CQC	Cray C90/16cpu/ 256 megawords
Nov. 5, 1997	New observational error statistics; changes to assimilation of TOVS radiances and addition of other data sources	
Jan. 13, 1998	Assimilation of non cloud-cleared radiances (Derber et al., pers. comm.); improved physics	
June 1998	Resolution increased to T170/40 layers (to 3.5 days); improved physics; 3D ozone data assimilation and forecast; nonlinear increments in 3D; VAR	

Adapted from Shuman (1989).

TABLE 3 Major Changes in the NMC/NCEP Regional Modeling and Data Assimilation Since 1985

Year	Operational Model	Computer
March 1985	Regional Analysis and Forecast System (RAFS) based on the triply Nested Grid Model (NGM; Phillips, 1979) and optimal interpolation (OI; DiMego, 1988); resolution: 80 km/16 layers	Cyber 205
August 1991	RAFS upgraded for the last time: NGM run with only two grids with inner grid domain doubled in size; Implemented Regional Data Assimilation System (RDAS) with 3 hourly updates using an improved OI analysis using all off-time data including Profiler and ACARS wind reports (DiMego et al., 1992) and complex quality control procedures (Gandin et al., 1993)	Cray YMP 8 processors 32 megawords
June 1993	First operational implementation of the Eta model in the 00Z & 12Z early run for North America at 80 km and 38 layer resolution (Mesinger et al., 1988; Janjic, 1994; Black et al., 1994)	

(*continued*)

TABLE 3 (*Continued*)

Year	Operational Model	Computer
September 1994	The Rapid Update Cycle (RUC; Benjamin et al., 1994) was implemented for CONUS domain with 3 hourly OI updates at 60-km resolution on 25 hybrid (sigma-theta) vertical levels	Cray C-90 16 processors 128 megawords
September 1994	Early Eta analysis upgrades (Rogers et al., 1995)	
August 1995	Mesoscale version of the Eta model (Black, 1994) was implemented at 03Z and 15Z for an extended CONUS domain, with 29-km and 50-layer resolution and with NMC's first predictive cloud scheme (Zhao and Black, 1994) and new coupled land–surface–atmosphere package (2 layer soil)	Cray C-90 16 processors 256 megawords
October 1995	Major upgrade of early Eta runs: 48-km resolution, cloud scheme and Eta Data Assimilation System (EDAS) using 3 hourly OI updates (Rogers et al., 1996)	
January 1996	New coupled land–surface–atmosphere scheme put into early Eta runs (Chen et al., 1997; Mesinger, 1997)	
July–August 1996	Nested capability demonstrated with twice daily support runs for Atlanta Olympic Games with 10-km 60-layer version of Meso Eta	
February 1997	Upgrade package implemented in the early and Meso Eta runs	
February 1998	Early Eta runs upgraded to 32 km and 45 levels with 4 soil layers; OI analysis replaced by 3-dimensional variational (3D-VAR) with new data sources; EDAS now partially cycled (soil moisture, soil temperature, cloud water/ice, and turbulent kinetic energy)	
April 1998	RUC (3 hourly) replaced by hourly RUC II system with extended CONUS domain, 40-km and 40-level resolution, additional data sources and extensive physics upgrades	

TABLE 3 (*Continued*)

Year	Operational Model	Computer
June 1998	Meso runs connected to early runs as single 4 day system for North American domain at 32-km and 45-level resolution, 15z run moved to 18z, added new snow analysis; all runs connected with EDAS, which is fully cycled for all variables	

From compilations by Fedor Mesinger and Geoffrey DiMego, personal communication, 1998.

nates, and physical parameterizations, was implemented in 1993, with the same 80-km horizontal resolution as the NGM. It was denoted "early" because of a short data cutoff. The resolution was increased to 48 km, and a first "mesoscale" version with 29 km and reduced coverage was implemented in 1995. A cloud prognostic scheme was implemented in 1995, and a new land-surface parameterization in 1996. The OI data assimilation was replaced by a 3D-VAR analysis in 1998, and at this time the early and meso-Eta models were unified into a 32-km/ 45-level version. Many other less significant changes were also introduced into the global and regional operational systems and are not listed here for the sake of brevity. The Rapid Update Cycle (RUC), which provides frequent updates of the analysis and very short-range forecasts over the continental United States (CONUS), developed at NOAA's Forecast System Laboratory, was implemented in 1994 and upgraded in 1998 (Benjamin et al., 1994).

The 36-h S_1 forecast verification scores constitute the longest record of forecast verification available anywhere. They were started in the late 1940s for subjective surface forecasts, before operational computer forecast guidance, and for 500 hPa in 1954, with the first numerical forecasts. Figure 1 includes the forecast scores for 500 hPa from 1954 until the present, as well as the scores for the 72-h forecasts. It is clear that the forecast skill has improved substantially over the years, and that the current 36-h 500-hPa forecasts are close to a level that in the 1950s would have been considered "perfect" (Shuman, 1989). The 72-h forecasts have also improved and are now as accurate as the 36-h forecasts were about 15 years ago.

Figure 5 shows threat scores for precipitation predictions made by expert forecasters from the NCEP Hydrometeorological Prediction Center (HPC, the Meteorological Operations Division of the former NMC). The threat score (TS) is defined as the intersection of the predicted area of precipitation exceeding a particular threshold (P), in this case 0.5 inches in 24 h, and the observed area (O), divided by the union of the two areas: $\mathrm{TS} = (P \cap O)/(P \cup O)$. The bias (not shown) is defined by P/O. The TS, also known as critical success index (CSI) is a particularly useful score for quantities that are relatively rare. Figure 4 indicates that the forecasters' skill in predicting accumulated precipitation has been increasing with time, and that the current average skill in the 2-day forecast is as good as the 1-day forecasts were in the 1970s. Beyond the first 6 to 12 h, the forecasts are based mostly on numerical

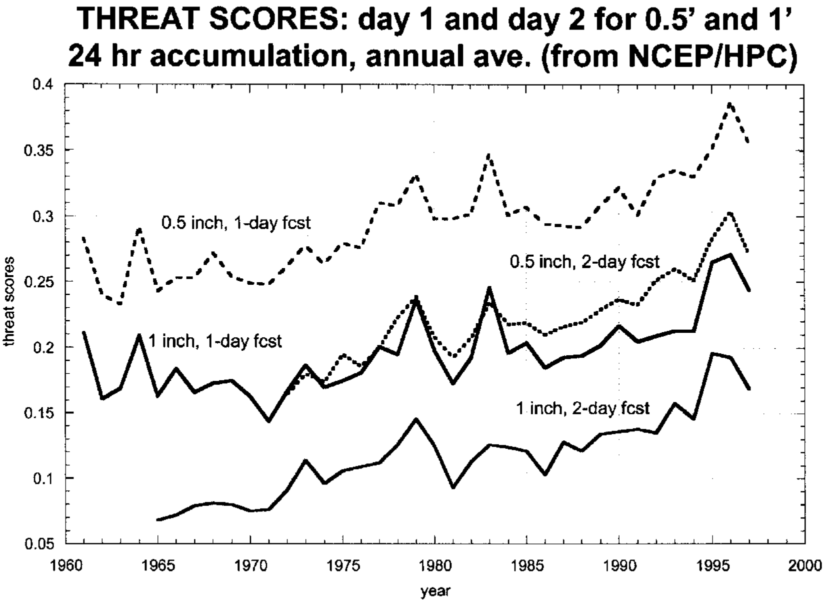


Figure 5 Threat scores of human forecasters at NCEP. (Data courtesy of J. Hoke.)

guidance, so that the improvement reflects, to a large extent, improvements of the numerical forecasts, which the human forecasters in turn improve upon based on their knowledge and expertise. The forecasters also have access to several model forecasts, and they use their judgment in assessing which one is more accurate in each case. This constitutes a major source of the "value-added" by the human forecasters.

The relationship between the evolution of human and numerical forecasts is clearly shown in a record compiled by the late F. Hughes (1987), reproduced in Figure 6. It is the first operational score maintained for the "medium-range" (beyond the first 2 days of the forecasts). The score used by Hughes was a standardized anomaly correlation (SAC), which accounted for the larger variability of sea level pressure at higher latitudes compared to lower latitudes. The fact that until 1976 the 3-day forecast scores from the model were essentially constant is an indication that their rather low skill was more based on synoptic experience than on model guidance. The forecast skill started to improve after 1977 for the 3-day forecast, and after 1980 for the 5-day forecast. Note that the human forecasts are on the average significantly more skillful than the numerical guidance, but it is the improvement in NWP forecasts that drives the improvements in the subjective forecasts.

5 THE FUTURE

The last decades have seen the expectations of Charney (1951) fulfilled and an amazing improvement in the quality of the forecasts based on NWP guidance.

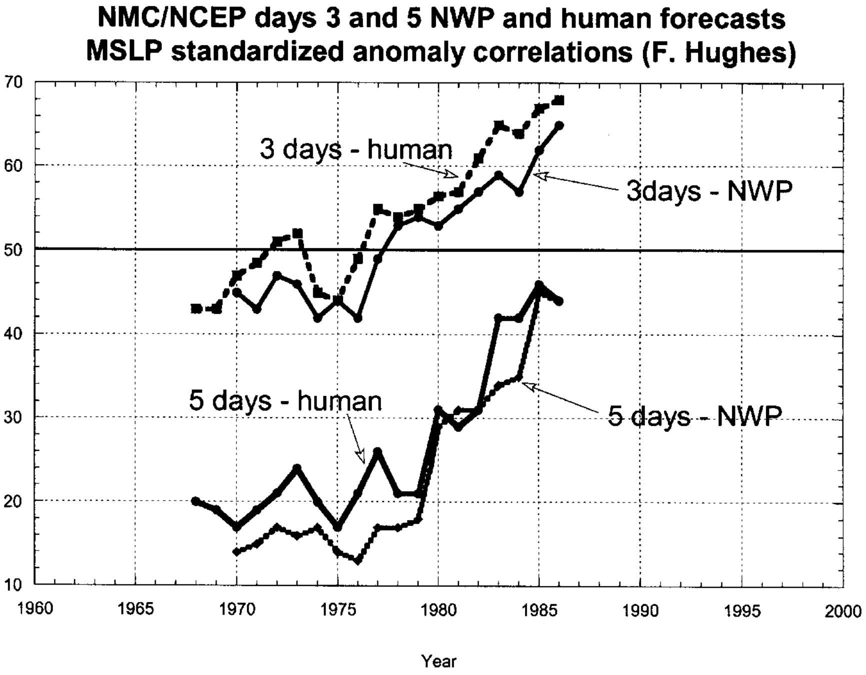


Figure 6 Hughes data: comparison of the forecast skill in the medium-range from NWP guidance and from human forecasters.

The next decade will continue seeing improvements, especially in the following areas:

- Detailed short-range forecasts, using storm-scale models able to provide skillful predictions of severe weather
- More sophisticated methods of data assimilation able to extract the maximum possible information from observing systems, especially remote sensors such as satellites and radars
- Development of adaptive observing systems, where additional observations are placed where ensembles indicate that there is rapid error growth (low predictability)
- Improvement in the usefulness of medium-range forecasts, especially through the use of ensemble forecasting
- Fully coupled atmospheric-hydrological systems, where the atmospheric model precipitation is appropriately scaled down and used to extend the length of river flow prediction
- More use of detailed atmosphere–ocean–land coupled models, where the effect of long-lasting coupled anomalies such as sea surface temperatures (SST) and

soil moisture anomalies leads to more skillful predictions of anomalies in weather patterns beyond the limit of weather predictability (about 2 weeks)

- More guidance to government and the public on areas such as air pollution, ultraviolet (UV) radiation and transport of contaminants, which affect health
- An explosive growth of systems with emphasis on commercial applications of NWP, from guidance on the state of highways to air pollution, flood prediction, guidance to agriculture, construction, etc.

REFERENCES

Andersson, E., J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljevic, C. Brankovic, C. Cardinali, C. Gaffard, A. Hollingsworth, C. Jakob, P. Janssen, E. Klinker, A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J.-N. Thepaut, and P. Viterbo (1998). The ECMWF implementation of three-dimensional variational assimilation (3D-Var). III. Experimental results, *Quart. J. Roy. Meteor. Soc.* **124**, 1831–1860.

Baer, F., and J. Tribbia (1977). On complete filtering of gravity modes through non-linear initialization, *Mon. Wea. Rev.* **105**, 1536–1539.

Barnes, S. (1964). A technique for maximizing details in numerical map analysis, *J. Appl. Meteor.* **3**, 395–409.

Bengtsson, L. (1999). From short-range barotropic modelling to extended-range global weather prediction: A 40-year perspective, *Tellus* **51 (A-B)**, 13–32.

Benjamin, S. G., G. A. Grell, J. M. Brown, R. Bleck, K. J. Brundage, T. L. Smith, and P. A. Miller (1994). An operational isentropic/sigma hyrid forecast model and data assimilation system, in *Proceedings, The Life Cycles of Extratropical Cyclones*, Vol. III, Bergen, Norway, June 27–July 1, 1994, S. Gronas and M. A. Shapiro (Eds.), Geophysical Institute, Bergen, Norway, University of Bergen, pp. 268–273.

Bergthorsson, P., and B. Döös (1955). Numerical weather map analysis, *Tellus* **7**, 329–340.

Bergthorsson, P., B. Döös, S. Frykland, O Hang, and R. Linquist (1955). Routine forecasting with the barotropic model, *Tellus* **7**, 329–340.

Black, T. L. (1994). The new NMC mesoscale Eta Model: Description and forecast examples, *Wea. Forecasting* **9**, 265–278.

Bratseth, A. (1986). Statistical interpolation by means of successive corrections, *Tellus* **38A**, 439–447.

Caplan, P., and G. White (1989). Performance of the National Meteorological Center's Medium-Range Model, *Wea. Forecasting* **4**, 391–400.

Charney, J. G. (1951). *Dynamical Forecasting by Numerical Process. Compendium of Meteorology*, Boston, American Meteorological Society.

Charney, J. G. (1962). Integration of the primitive and balance equations, in Proc. of the International Symposium on Numerical Weather prediction, Nov. 1960, Tokyo, Meteorological Society of Japan.

Charney, J. G., R. Fjørtoft, and J. von Neuman (1950). Numerical integration of the barotropic vorticity equation, *Tellus* **2**, 237–254.

Collins, W. G. (1998). Complex quality control of significant level radiosonde temperatures, *J. Atmos. Oceanogra. Tech.* **15**, 69–79.

Collins, W. G., and L. S. Gandin (1990). Comprehensive hydrostatic quality control at the National Meteorological Center, *Mon. Wea. Rev.* **118**, 2752–2767.

Courtier, P., J.-N. Thepaut, and A. Hollingsworth (1994). A strategy for operational implementation of 4d-Var using an incremental approach, *Quart. J. Roy. Meteor. Soc.* **120**, 1367–1387.

Daley, R. (1991). *Atmospheric Data Analysis*, Cambridge University Press.

Derber, J. C., and W.-S. Wu (1998). The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system, *Mon. Wea. Rev.* **126**, 2287–2302.

Derber, J. C., D. F. Parrish, and S. J. Lord (1991). The new global operational analysis system at the National Meteorological Center, *Wea. Forecasting* **6**, 538–547.

DiMego, G. J. (1988). The National Meteorological Center regional analysis system, *Mon. Wea. Rev.* **116**, 977–1000.

DiMego, G. J., K. E. Mitchell, R. A. Petersen, J. E. Hoke, J. P. Gerrity, J. J. Tuccillo, R. L. Wobus, and H. M. H. Juang (1992). Changes to NMC's regional analysis and forecast system, *Wea. Forecasting* **7**, 185–198.

Evensen, G., and P. J. Van Leeuwen (1996). Assimilation of GEOSAT altimeter data for the Aghulas current using the ensemble Kalman filter with a quasigeostrophic model, *Mon. Wea. Rev.* **124**, 85–96.

Gandin, L. S. (1963). Objective analysis of meterological fields, *Gidrometerologicheskoe Izdatelstvo*, Leningrad; English translation by Israeli Program for Scientific Translations, Jerusalem, 1965.

Gandin, L. S. (1988). Complex quality control of meteorological observations, *Mon. Wea. Rev.* **116**, 1137–1156.

Gandin, L. S., L. L. Morone, and W. G. Collins (1993). Two years of operational comprehensive hydrostatic quality control at the NMC, *Wea. Forecasting*, **8**(1), 57–72.

Gilchrist, B., and G. Cressman (1954). An experiment in objective analysis, *Tellus* **6**, 309–318.

Haltiner, G. J., and R. T. Williams (1980). *Numerical Prediction and Dynamic Meteorology*, New York, Wiley.

Hong, S. Y., and H. L. Pan (1996). Nonlocal boundary layer vertical diffusion in a medium-range forecast model, *Mon. Wea. Rev.* **124**, 2322–2339.

Howcroft, J. G. (1971). Local forecast model: Present status and preliminary verification. NMC Office note 50, National Weather Service, NOAA, US Dept of Commerce. Available from the National Centers for Environmental Prediction, 5200 Auth Rd., Rm. 100, Camp Springs, MD 20746.

Hughes, F. D. (1987). Skill of Medium-range Forecast Group, Office Note #326, National Meteorological Center, NWS, NOAA, US Dept of Commerce.

Janjic, Z. I. (1994). The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes, *Mon. Wea. Rev.* **122**, 927–945.

Kalnay, E. (2001). *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press.

Kalnay, E. S., S. J. Lord and R. McPherson (1998). Maturity of operational numerical weather prediction: The medium range, *Bull. Am. Meteor. Soc.* **79**, 2753–2769.

Kanamitsu, M., J. C. Alpert, K. A. Campana, P. M. Caplan, D. G. Deaven, M. Iredell, B. Katz, H.-L. Pan, J. Sela, and G. H. White (1991). Recent changes implemented into the global forecast system at NMC, *Wea. Forecasting* **6**, 425–436.

Lorenc, A. (1981). A global three-dimensional multivariate statistical interpolation scheme, *Mon. Wea. Rev.* **109**, 701–721.

Lorenc, A. (1986). Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteor. Soc.* **112**, 1177–1194.

Lorenz, E. N. (1963). Deterministic nonperiodic flow, *J. Atmos. Sci.* **20**, 130–141.

Lorenz, E. N. (1965). A study of the predictability of a 28-variable atmospheric model, *Tellus* **17**, 321–333.

Lorenz, E. N. (1982). Atmospheric predictability experiments with a large numerical model, *Tellus* **34**, 505–513.

Lorenz, E. N. (1993). *The Essence of Chaos*, University of Washington Press.

Lynch, P., and X.-Y. Huang (1994). Diabatic initialization using recursive filters, *Tellus* **46A**, 583–597.

McPherson, R. D., K. H. Bergman, R. E. Kistler, G. E. Rasch, and D. S. Gordon (1979). The NMC operational global data assimilation system, *Mon. Wea. Rev.* **107**, 1445–1461.

Mesinger, F. (1996). Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade, *Am. Meteor. Soc.* **77**, 2637–2649.

Mesinger, F., Z. I. Janjic, S. Nickovic, D. Gavrilov, and D. G. Deaven (1988). The step-mountain coordinate: Model description and performance for cases of Alpine lee cyclogenesis and for a case of an Appalachian redevelopment, *Mon. Wea. Rev.* **116**, 1493–1518.

Pan, H.-L. (1990). A simple parameterization scheme of evapotranspiration over land for the NMC Medium-Range Forecast Model, *Mon. Wea. Rev.* **118**, 2500–2512.

Pan, H.-L., and L. Mahrt (1987). Interaction between soil hydrology and boundary-layer development, *Boundary-Layer Meteor.* **38**, 185–202.

Pan, H.-L., and W.-S. Wu (1995). Implementing a mass flux convection parameterization package for the NMC Medium-Range Forecast Model, Office note 409, National Meteorological Center.

Parrish, D. F., and J. D. Derber (1992). The National Meteorological Center spectral statistical interpolation analysis system, *Mon. Wea. Rev.* **120**, 1747–1763.

Phillips, N. A. (1956). The general circulation of the atmosphere, a numerical experiment, *Q. J. Roy Met. Soc.* **82**, 123–164.

Phillips, N. A. (1979). The nested grid model, NOAA, Tech. Report NWS 22, US Dept of Commerce, Washington DC. Available from the National Centers for Environmental Prediction, 5200 Auth Rd., Rm. 100, Camp Springs, MD 20746.

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*, Cambridge University Press; reprinted by Dover (1965) with a new introduction by Sydney Chapman.

Robert, A. J. (1981). A stable numerical integration scheme for the primitive meteorological equations, *Atmosphere-Ocean* **19**, 35–46.

Rogers, E., T. L. Black, D. G. Deaven, G. J. DiMego, Q. Zhao, M. Baldwin, N. W. Junker, and Y. Lin (1996). Changes to the Operational Early Eta Analysis/Forecast System at the National Centers for Environmental Prediction, *Wea. Forecasting* **11**, 319–413.

Sasaki, Y. (1970). Some basic formalisms in numerical variational analysis, *Mon. Wea. Rev.* **98**, 875–883.

Sela, J. G. (1980). Spectral modeling at the National Meteorological Center, *Mon. Wea. Rev.* **108**, 1279–1292.

Shuman, F. G. (1989). History of numerical weather prediction at the National Meteorological Center, *Wea. Forecasting* **4**, 286–296.

Shuman, F. G., and J. B. Hovermale (1968). An operational six-layer primitive equation model, *J. Appl. Meteor.* **7**, 525–547.

Toth, Z., and E. Kalnay (1993). Ensemble forecasting at NMC: The generation of perturbations, *Bull. Am. Meteor. Soc.* **74**, 2317–2330.

Toth, Z., and E. Kalnay (1997). Ensemble forecasting at NCEP: The breeding method, *Mon. Wea. Rev.* **125**, 3297–3318.

Tracton, M. S., and E. Kalnay (1993). Ensemble forecasting at NMC: Practical aspects, *Wea. Forecasting* **8**, 379–398.

Zhao, Q., T. L. Black, and M. E. Baldwin (1997). Implementation of the cloud prediction scheme in the eta model at NCEP, *Wea. Forecasting* **12**, 697–711.

SECTION 2

THE CLIMATE SYSTEM

Contributing Editor: Robert E. Dickinson

CHAPTER 9

OVERVIEW: THE CLIMATE SYSTEM

ROBERT E. DICKINSON

The climate system consists of the atmosphere, cryosphere, oceans, and land interacting through physical, chemical, and biological processes. Key ingredients are the hydrological and energy exchanges between subsystems through radiative, convective, and fluid dynamical mechanisms. Climate involves changes on seasonal, year-to-year, and decadal or longer periods in contrast to day-to-day weather changes. However, extreme events and other statistical measures are as, or more, important than simple averages. Climate is seen to impact human activities most directly through the occurrence of extremes. The frequency of particular threshold extremes, as, for example, the number of days with maximum temperatures above 100°F, can change substantially with shifts in climate averages.

1 THE ATMOSPHERE

The atmosphere is described by winds, pressures, temperatures, and the distribution of various substances in gaseous, liquid, and solid forms. Water is the most important of these substances. Also important are the various other radiatively active ("greenhouse") gases, including carbon dioxide and liquid or solid aerosol particulates. Most of the mass of the atmosphere is in the troposphere, which is comprised of the layers from the surface to about 12 km (8 km in high latitudes to 16 km at the equator) where the temperature decreases with altitude. The top of the troposphere is called the tropopause. Overlying this is the stratosphere, where temperatures increase with altitude to about 50 km or so (Fig. 1). The tropospheric temperature decreases with altitude are maintained by vertical mixing driven by moist and dry convection.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

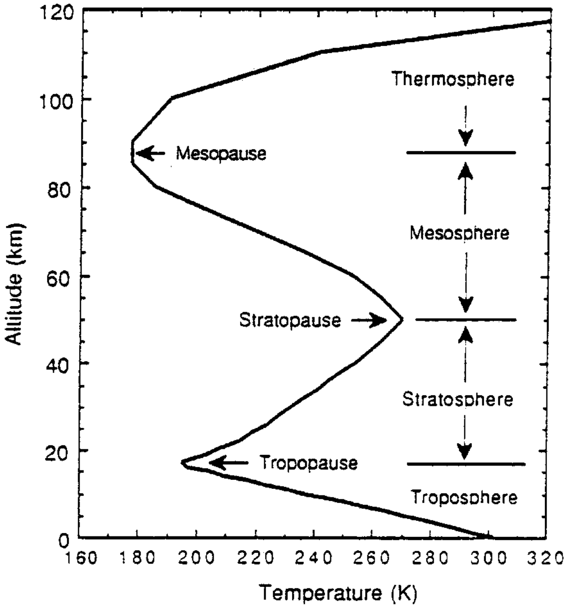


Figure 1 Main zones of the atmosphere defined according to the temperature profile of the standard atmosphere profile at 15°N for annual-mean conditions (Hartmann, 1994).

The temperature increases with altitude in the stratosphere in response to increasing heating per the unit mass by ozone absorption of ultraviolet radiation. The variation of temperature structure with latitude is indicated in Figure 2. The troposphere is deepest in the tropics because most thunderstorms occur there. Because of this depth and stirring by thunderstorms, the coldest part of the atmosphere is the tropical tropopause. In the lower troposphere temperatures generally decrease from the equator to pole, but warmest temperatures shift toward the summer hemisphere, especially in July. Longitudinally averaged winds are shown in Figure 3. Because of the geostrophic balance between wind and pressures, winds increase with altitude where temperature decreases with latitude. Conversely, above about 8 km, where temperatures decrease toward the tropical tropopause, the zonal winds decrease with altitude. The core of maximum winds is referred to as the jet stream. The jet stream undergoes large wavelike oscillations in longitude and so is usually stronger at a given latitude than in its longitudinal average. These waves are especially noticeable in the winter hemisphere as illustrated in Figure 4.

2 GLOBAL AVERAGE ENERGY BALANCE

Solar radiation of about 342 W/m^{-2} entering Earth's atmosphere is absorbed and scattered by molecules. The major gaseous absorbers of solar radiation are water vapor in the troposphere and ozone in the stratosphere. Clouds and aerosols likewise

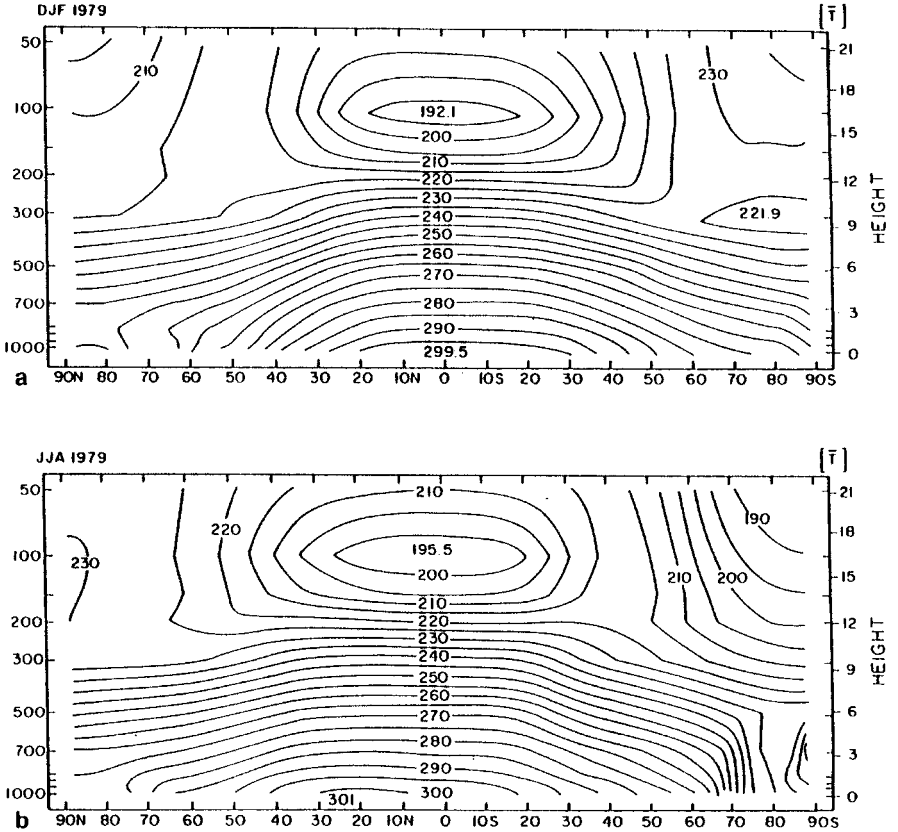


Figure 2 Zonal mean meridional cross sections of temperature during two seasons. (*a*) December 1978–February 1979 and (*b*) June–August 1979. Height scale is approximate and the contour interval is 5 K (Grotjahn, 1993).

scatter and absorb. Clouds are the dominant scatterer and so substantially enhance the overall planetary reflected radiation, whose ratio to incident solar radiation, about 0.31, is referred to as *albedo*. Thermal infrared radiation, referred to as *longwave*, is controlled by clouds, water vapor, and other greenhouse gases. Figure 5 (Kiehl and Trenberth, 1997) illustrates a recent estimate of the various terms contributing to the global energy balance. The latent heat from global average precipitation of about 1.0 m per year is the dominant nonradiative heating term in the atmosphere.

Because of the seasonally varying geometry of Earth relative to the sun, and the differences in cloudiness and surface albedos, there are substantial variations in the distribution of absorbed solar radiation at the surface and in the atmosphere, as likewise in the transfer of latent heat from the surface to the atmosphere. This heterogeneous distribution of atmospheric heating drives atmospheric wind systems, either directly or through the creation of available potential energy, which is utilized

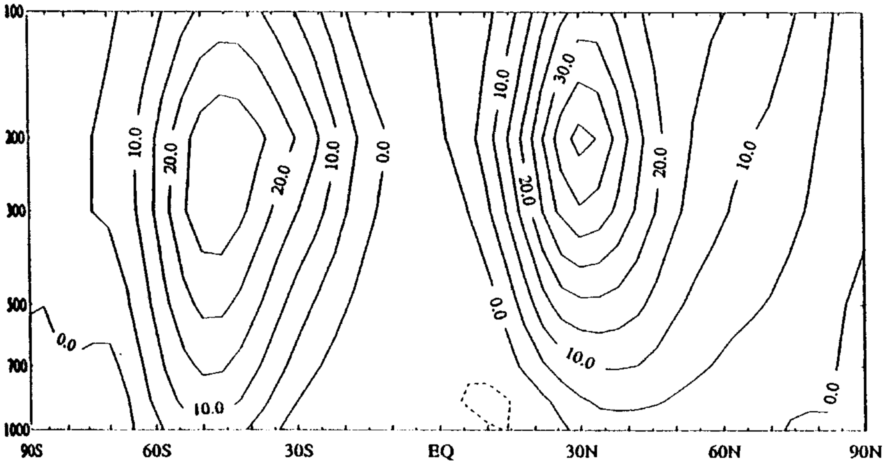


Figure 3 Meridional cross sections of longitudinally averaged zonal wind (top panels, m/s) for DJF (Holton, 1992).

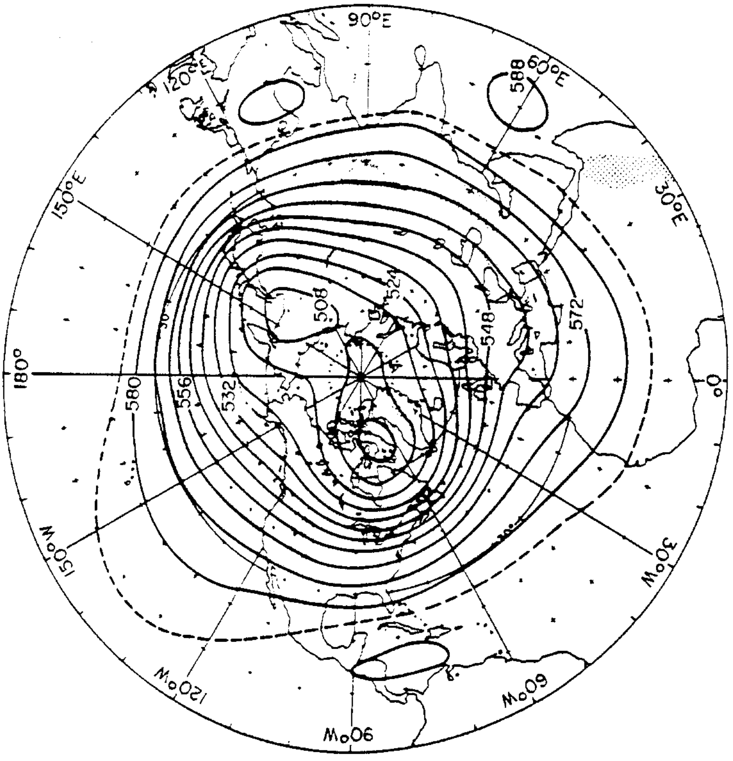


Figure 4 Mean 500-mb contours in January, Northern Hemisphere. Heights shown in tens of meters (Holton, 1992).

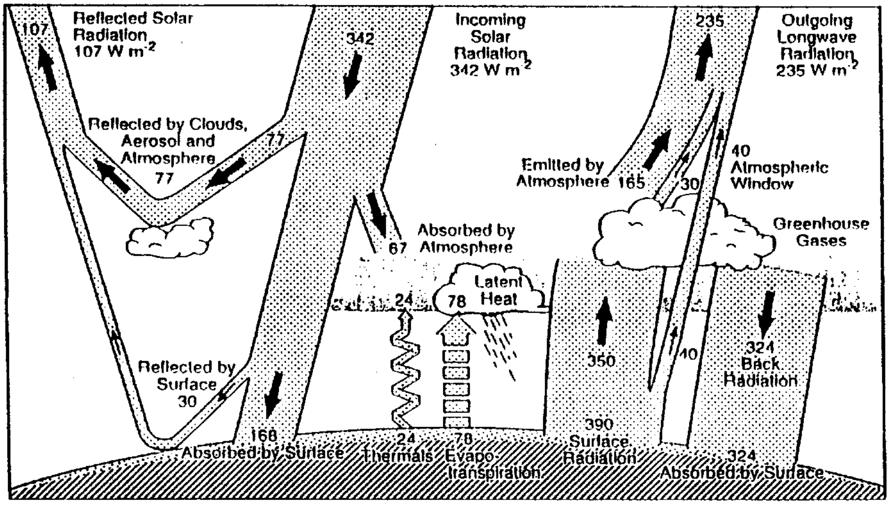


Figure 5 Earth's annual global mean energy budget based on the present study. Units are W/m^2 (Kiehl and Trenberth, 1997).

to maintain random occurrences of various kinds of instabilities, such as thunderstorms and wintertime cyclonic storm systems. These dynamical systems hence act to redistribute energy within the atmosphere and so determine the distributions of temperature and water vapor. Likewise, the balances at the surface between fluxes of radiative, latent, and thermal energies determine surface temperatures and soil moistures. The properties of the near-surface air we live in are determined by a combination of surface and atmospheric properties, according to processes of the atmospheric boundary layer. Thus climatic anomalies in surface air may occur either because of some shift in atmospheric circulation patterns or through some modification of surface properties such as those accompanying deforestation or the development of an urban area.

3 THE ATMOSPHERIC BOUNDARY LAYER

The term *boundary layer* is applied in fluid dynamics to layers of fluid or gas, usually relatively thin, determining the transition between some boundary and the rest of the fluid. The atmospheric boundary layer is the extent of atmosphere that is mixed by convective and mechanical stirring originating at Earth's surface. Such stirring is commonly experienced by airplane travelers as the bumps that occur during takeoff or landing, especially in the afternoon, or as waves at higher levels in flying over mountainous regions. The daytime continental boundary layer, extending up to several kilometers in height, is most developed and vigorously mixed, being the extent to which the daytime heating of the surface drives convective overturning of the atmosphere. The land cools radiatively at night, strongly stabilizing the atmosphere against convection, but a residual boundary layer extends up to about 100 m stirred by the flow of air over the underlying rough surface. This diurnal variation of fluxes over the ocean is much weaker and the boundary layer is of intermediate height. The temperature of the atmosphere, when stirred by dry mixing, decreases at a rate of 9.8 K/km. Above the boundary layer, temperatures decrease less rapidly with height, so that the atmosphere is stable to dry convection. A layer of clouds commonly forms at the top of the daytime and oceanic boundary layers and contributes to the convection creating the boundary layer through its radiative cooling (convection results from either heating at the bottom of a fluid or cooling at its top). Also, at times the clouds forming near the top of the boundary layer can be unstable to moist convection, and so convect upward through a deep column such as in a thunderstorm.

4 ATMOSPHERIC HYDROLOGICAL CYCLE

The storage, transport, and phase changes of water at the surface and in the atmosphere are referred to as the hydrological cycle. As already alluded to, the hydrological cycle is closely linked to and driven by various energy exchange processes at the surface and in the atmosphere. On the scale of continents, water is moved from oceans to land by atmospheric winds, to be carried back to the oceans by streams and rivers as elements of the land hydrological cycle. Most of the water in the atmosphere is in its vapor phase, but water that is near saturation vapor pressure (relative humidity of 100%) converts to droplets or ice crystals depending on temperature and details of cloud physics. These droplets and crystals fall out of the atmosphere as precipitation. The water lost is replenished by evaporation of water at the surface and by vertical and horizontal transport within the atmosphere. Consequently, much of the troposphere has humidities not much below saturation. Saturation vapor pressure increases rapidly with temperature (about 10% per kelvin of change). Hence, as illustrated in Figure 6, the climatological concentrations of water vapor vary from several percent or more when going from near-surface air to a few parts per million near the tropical tropopause. Water vapor concentrations in the stratosphere are close to that of the tropical tropopause, probably because much of the air in the lower stratosphere has been pumped through the tropical tropopause by moist convection.

5 CLIMATE OF THE STRATOSPHERE

The dominant radiative processes in the stratosphere are the heating by absorption of solar ultra violet (UV) radiation and cooling by thermal infrared emission from carbon dioxide and, to a lesser extent, ozone molecules. The stratospheric absorption of UV largely determines how much harmful UV reaches the surface. Ozone in the upper troposphere and lower stratosphere additionally adds heat by absorption of thermal emission from the warmer surface and lower layers. The stratosphere,

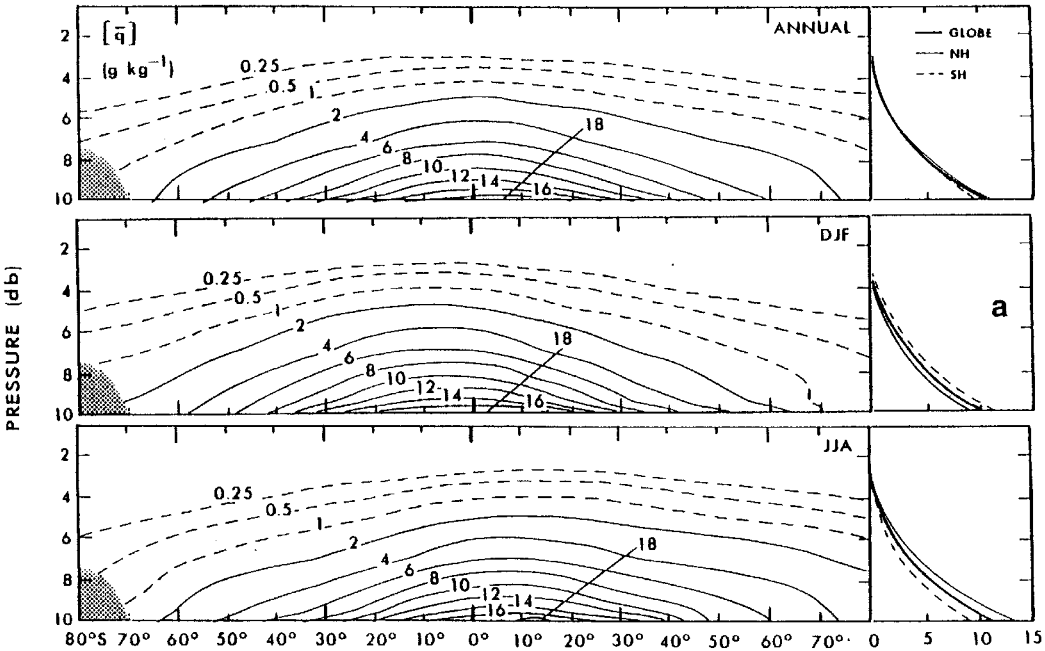


Figure 6 Zonal mean cross sections of the specific humidity in g/kg for annual, DJF, and JJA mean conditions. Vertical profiles of hemispheric and global mean values are shown on the right (Peixoto and Oort, 1992).

furthermore, enhances the greenhouse warming of CO_2 in the troposphere through substantial downward thermal emissions to the troposphere.

How changes of ozone change stratospheric and tropospheric radiative heating depends on the amounts of overlying ozone and, for thermal effect, on pressure and radiative upwelling depending on underlying temperatures.

Besides radiative processes, stratospheric climate is characterized by its temperature and wind patterns and by the chemical composition of its trace gases. At midstratosphere, temperature increases from winter pole to summer pole with an accompanying eastward jet stream in the winter hemisphere extending upward from the tropospheric jet steam. This wind configuration allows planetary wave disturbances to propagate into the stratosphere, contributing significant temporal and longitudinal variabilities. Conversely, the westward jet, found in the summer stratosphere attenuates wave disturbances from below, and so is largely zonally symmetric, changing only with the seasonal heating patterns.

6 THE CRYOSPHERE

The term *cryosphere* refers to the components of the climate system dominated by water in its frozen phase, that is, in high latitudes and extratropical winter conditions. Elements include snow, its distribution and depths, sea ice, its distribution and properties, high-latitude ice caps, and temperate glaciers. The largest volume of frozen water is stored in ice caps, and glaciers. This storage acts to remove water from the oceans. How it changes with climate change is, hence, of interest for determining changing sea levels.

Ice is highly reflective of sunlight, especially in crystal form. The loss of solar heating because of this high albedo acts to substantially reduce high-latitude temperatures especially in spring and early summer where near-maximum solar radiation sees white snow-covered surfaces. This high albedo can be substantially masked by cloud cover and, over land, tall vegetation such as conifer forests.

7 THE OCEAN

Oceans are a major factor in determining surface temperatures and fluxes of water into the atmosphere. They store, release, and transport thermal energy, in particular, warming the atmosphere under wintertime and high-latitude conditions, and cooling it under summer and tropical conditions.

How the oceans carry out these services depends on processes coupling them to the atmosphere. Atmospheric winds push the oceans into wind-driven circulation systems. Net surface heating or cooling, evaporation, and precipitation determine oceanic densities through controlling temperature and salinity, hence oceanic buoyancy. This net distribution of buoyancy forcing drives “thermohaline” overturning of the ocean, which acts to transport heat. Climate of the surface layers of the ocean includes the depth to which waters are stirred by waves and net heating or cooling.

Heating acts to generate shallow warm stable layers, while cooling deepens the surface mixed layers. Under some conditions, convective overturning of cold and/or high-salinity water can penetrate to near the ocean bottom.

REFERENCES

Grotjahn, R. (1993). Zonal average observations, Chapter 3, in *Global Atmospheric Circulations: Observations and Theories*, New York, Oxford University Press.

Hartmann, D. L. (1994). Atmospheric temperature, Chapter 1.2, in *Global Physical Climatology*, San Diego, Academic Press.

Holton, J. R. (1992). The observed structure of extratropical circulations, Chapter 6.1, in *An Introduction to Dynamic Meteorology*, San Diego, Academic.

Kiehl, J. T. and K. E. Trenberth (1997). Earth's annual global mean energy budget, *J. Clim.* **78**, 197–208.

Peixoto, J. P., and A. H. Oort (1992). Observed atmospheric branch of the hydrological cycle, Chapter 12.3, *Physics of Climate*, New York, American Institute of Physics.

CHAPTER 10

THE OCEAN IN CLIMATE

EDWARD S. SARACHIK

1 INTRODUCTION

Earth's present climate is intrinsically affected by the ocean—the climate without the ocean would be different in many essential ways: Without the evaporation of water from the sea surface, the hydrological cycle would be different; without ocean heat transport and uptake, the temperature distribution of the globe would be different; and without the biota in the ocean, the total amount of carbon in the atmosphere would be many time its current value. Yet, while we may appreciate the role of the ocean in climate, the difficulty and expense of making measurements below the ocean's surface has rendered the vast volume of the ocean a sort of *mare incognita.* Why is the ocean so important in Earth's climate, and which of its properties are of special significance for climate? How have we learned about the ocean and its role in climate, and what more do we need to know?

2 PROPERTIES OF THE OCEAN AND PROCESSES IN THE OCEAN

The ocean covers 70% of Earth's surface to an average depth of about 4 km. The mass of the ocean is 30 times and its heat capacity 120 times that of the atmosphere, and the ocean contains 80 times the carbon dioxide stored in the atmosphere.

The density of the ocean is controlled both by its temperature and by its salt content. Ocean density increases with salinity and decreases with temperature. Unlike fresh water, which has a maximum density at 4°C (so that colder water and ice float on 4°C water and the temperature at the bottom of a nonfrozen lake is 4°C), water saltier than 26 parts per thousand of water is continuously denser as

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

the temperature is lowered, and the temperature at the bottom of the ocean is closer to 1°C.

Since heat always diffuses from warm to cool temperatures, why does not the temperature of the deep ocean eventually adjust and become the same as its surface temperature? Cold water constantly sinks at high latitudes (both in the Northern and Southern Hemispheres) and fills the deeper parts of the oceans with cold water so that the water at depth is always cold, even when the surface temperature is very warm. This circulation is called the thermohaline circulation.

About 7% of the ocean surface is covered by sea ice. Growth of sea ice radically changes the nature of the ocean surface: Sea ice reflects solar radiation, thereby preventing it from being absorbed by the surface and blocks the transfer of heat and moisture from the surface of the ocean to the atmosphere.

The average salinity in the global oceans is 34.7 parts per thousand salt to water by weight. As the total amount of salt in the ocean is constant, changes in salinity only occur because of additions and subtractions of fresh water. Salinity decreases as rain falls on the ocean or river water enters the ocean, and it increases as water evaporates from the surface of the ocean. As sea ice grows, it rejects salt into the ocean thereby increasing its salinity. Similarly, as ice melts, it dilutes the surrounding ocean and lowers its salinity. A specific parcel of water can either increase or decrease its salinity by mixing with parcels with different salinities.

3 HOW THE OCEAN INTERACTS WITH THE ATMOSPHERE TO AFFECT THE CLIMATE

The ocean interacts with the atmosphere at (or very near) the sea surface where the two media meet. Visible light can penetrate into the ocean several tens of meters, but heat, moisture, and momentum, carbon dioxide, and other gases exchange directly at the surface. Sea ice forms at the surface and helps to determine the local exchanges. The basic problem of the ocean in climate is to explain these interchanges and to determine those characteristics of the ocean that affect these exchanges.

The ocean may be considered to interact with the atmosphere in two distinct ways: passively and actively. It interacts passively when the ocean affects the atmosphere but does not change the essential manner in which the atmosphere is operating. An example of a passive interaction is the oceanic response to adding CO_2 to the atmosphere where the ocean simply delays the greenhouse warming of the atmosphere as heat and CO_2 enters the ocean.

Active interaction with the atmosphere produces results that would not otherwise be there—an example is where the warming of the atmosphere reduces the thermohaline circulation and produces a climate reaction that could not have been obtained without the essential interaction of the ocean. In particular, since the northern branch, say, of the thermohaline circulation brings cold water from high latitudes toward the equator, and since the water must be replaced by warm water that is cooled as it moves northward, the net effect of the thermohaline circulation is to transport heat northward and thereby warm the higher latitudes. As the atmosphere warms, the water becomes less dense both by the effect of temperature and by increased rainfall, a necessary concomitant of global warming. Since the atmosphere sees a reduced north–south temperature gradient at the sea surface, it reacts fundamentally differently than if the thermohaline circulation were at full strength.

Our present models of greenhouse warming have the ocean acting in both the active and passive modes—active when warming leads to a slowed thermohaline circulation and passive when heat and CO_2 simply enter the ocean surface and is therefore lost to the atmosphere. Another example of active interaction is El Niño, a phenomenon that would not exist were it not for the active interaction of the atmosphere and the ocean (see the chapter by Trenberth). The ocean also has been inferred (by examining the composition of ancient ice stored in the Greenland and Antarctic ice sheets) to have, and probably actively take part in causing, climatic variability on time scales anywhere from decades to a few thousand years, a type of variability not seen on Earth since it emerged from the last glacial maximum some 18,000 years ago.

4 MEASURING THE OCEAN

The ocean is remarkably poorly measured. While the global atmosphere is constantly probed and analyzed for the purposes of weather prediction, until recently no such imperative existed for the ocean. Our ability to measure the ocean is severely limited basically by the inability of radiation to penetrate very far into the ocean—this requires direct in situ measurements of the interior of the ocean. The world's oceanographic research fleet is small and incapable of monitoring the breadth and depth of the world's ocean, although valuable research measurements are constantly being taken at selected places. As a result of ocean observations, we know the basic pathways of water in the ocean, we have a good idea of the transports by the ocean, we have some idea of the basic mechanisms for much of the major ocean currents, and we have a good idea of how the exchanges at the surface are accomplished. Yet we cannot measure vertical velocity (it is far too small), and so we remain completely ignorant of the processes by which the interior of the ocean affects the surface. Similarly, we are ignorant of the basic processes of mixing and friction in the ocean, both basic to being able to model the ocean for climate purposes.

Major oceanographic programs have been conducted in the last decade (the World Ocean Circulation Experiment, WOCE) and (the Tropical Ocean–Global Atmosphere, TOGA), and while they have taught us much about the ocean circulation and El Niño, respectively, the basic lesson is that, unless we can make continuous long-term measurements beneath the surface of the ocean, it will forever remain unknown territory.

5 MODELING THE OCEAN

Because the ocean is poorly measured, and because much of what we need to know about the past and predict about the future cannot be directly known, only inferred, models have played a particularly important role in the development of oceanography and, in particular, the role of the ocean in climate.

The basic tool of climate studies is the coupled model, where the various components of the climate system—the atmosphere, ocean, cryosphere, biosphere, and chemosphere—are simultaneously and consistently coupled. The ocean component of such a coupled model exchanges its heat, fresh water, and momentum with the atmosphere at the sea surface. The test of the successful coupling of the atmosphere and ocean is the correct simulation of the time-varying sea surface temperature and surface winds, both of which are relatively easy to measure: Directly by ship or mooring, remotely by satellite, or by a combination of the two.

The development of off-line ocean-only models requires the heat, momentum, and freshwater forcing from the atmosphere to be known. Since precipitation and evaporation, in particular, are so poorly measured over the ocean, it is a continual struggle to know whether errors in the ocean model are due to errors in the model itself or errors in the forcing of the ocean by the atmosphere.

Ocean models themselves are relatively simple in concept: The known equations of water and salt are discretized and time stepped into the future. The discretization process requires a trade-off between fine resolution for accuracy and the need to simulate over long periods of time, which, because of limited computer resources, requires coarser resolution. While the equation of state of seawater relating density to salt, temperature, and pressure cannot be written down simply, it has, over the course of time, become known to high accuracy.

What makes ocean modeling difficult is the specification of those mixing processes that unavoidably cannot be resolved by whatever resolution is chosen. We are beginning to understand that enhanced small-scale mixing occurs near bottom topography and near boundaries: Purposeful release experiments, where a dye is released and then followed in time to see how the dye cloud evolves, has revealed this to us. Larger scale mixing, where parcels are interchanged because of the large-scale circulation (but still unresolved by the ocean models) itself is more problematic, but recent advances in parameterizing these unresolved mixing effects have shown promise.

6 THE FUTURE OF THE OCEAN IN CLIMATE

It is clear that the ocean is a crucial component of the climate system. Since so much of what is not known about the past and future of the climate system depends on active interactions with the ocean, it is clear that we have to learn more about its essential processes. How to go about learning about the ocean is the difficult question.

Direct measurements are very expensive, and satellites, while giving a global look at the entire ocean, see only its surface. Designs are currently underway for a Global Ocean Observing System (GOOS), but the cost of implementing such a system in toto is prohibitive, even if shared among the wealthier countries of the world.

It is likely that a combination of studies, perhaps conducted for entirely different purposes, will advance the field most rapidly. In particular, the advent of the El Niño–Southern Oscillation (ENSO) prediction, which requires subsurface ocean data as initial conditions, has made almost permanent the Tropical Atmosphere-Ocean (TAO) array in the tropical Pacific, giving an unprecedented and continuous view of a significant part of the tropical ocean. We may extend the reasoning to say that, where predictability is indicated and shows societal or economic value, the measurement systems to produce the initial data will almost certainly be implemented. The promise of predicting climate from seasons to a few years will expand the ocean-observing system considerably. Additional expansions will come from resource monitoring, pollution monitoring, and various types of monitoring for national security purposes. While monitoring for security has traditionally meant the data is classified, once taken, data can eventually reach the research arena—the vast amount of Soviet and U.S. data that was declassified after the end of the cold war has shown this.

Observations can be also combined with models to give "value-added" observations. Data at individual points in the ocean exist without reference to neighboring points unless models are used to dynamically interpolate the data to neighboring points using the equation of motion of a fluid. This so-called four-dimensional data assimilation is in the process of development and shows promise as a powerful way of optimally using the ocean data that can be taken.

Models can also be compared with other models. While this might seem sterile, fine-resolution models can be used to develop parameterizations of large-scale mixing for use in coarse-resolution ocean models that can be run the long times needed to participate in coupled model simulations of climate. Advances in computer power will ultimately allow successive refinements in resolution so that finer scale resolution models can be run directly.

We close by reemphasizing the crucial role that the ocean plays in climate and climate variability and the necessity to know more about the ocean for all aspects of the climate problem.

CHAPTER 11

PROCESSES DETERMINING LAND SURFACE CLIMATE

GORDON BONAN

1 INTRODUCTION

Energy is continually flowing through the land–atmosphere system. As the sun's radiation passes through the atmosphere, some of it is absorbed, primarily by water vapor and clouds, and some is reflected back to space by clouds and particles suspended in the air. The remainder reaches Earth's surface, where it is either absorbed or reflected upwards. The solar radiation absorbed by the surface provides the warmth needed to maintain life and is used in biological activities such as photosynthesis. In turn, it provides energy to warm the atmosphere. Its surface emits radiation in the infrared waveband in proportion to its temperature raised to the fourth power. Most of this longwave radiation is absorbed by water vapor, clouds, carbon dioxide, and other gases in the atmosphere, heating the atmosphere. Without this heating, Earth's effective temperature would be 33°C cooler than it is now.

The solar and longwave radiation absorbed by Earth's surface constitute the net radiation at the surface. This energy is either stored or returned to the atmosphere as sensible or latent heat. Objects that absorb radiation become warmer than their surroundings and lose some of that energy by convection. For example, heat will normally be lost from a warm surface to the cooler air surrounding it. The transfer of this heat determines the temperature of air and is called sensible heat because it can be felt. Heat is also lost from the surface by evapotranspiration. Evapotranspiration determines the amount of water in the atmosphere, but it also cools the surface because the change of water from liquid to gas requires a large amount of heat,

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

which is transferred from the surface to the atmosphere as latent heat. The net radiation at the surface that is not returned to the atmosphere as sensible or latent heat is stored at the surface. Heat storage is very important for the diurnal cycle of temperature over land. Soils have a much smaller heat capacity than water. Consequently, land heats and cools faster than water.

Surface properties determine these energy fluxes and the resulting surface climate. Forests are "darker", hence absorbing more solar radiation, than grasslands. Forests are also taller—that is "rougher"—than shrubs or grasses and exert more stress on the fluid motion of the atmosphere. Deserts and shrublands, with their dry soils, have less evapotranspiration than well-watered forests or crops. Observational studies and the advent of sophisticated mathematical models of atmospheric physics, atmospheric dynamics, and surface ecological and hydrological processes have allowed scientists to examine how these different surfaces affect climate. Numerous studies all point to the same conclusion: The distribution of vegetation on Earth's surface is an important determinant of regional and global climates; consequently, natural and human-induced changes in Earth's surface can alter the climate.

Tropical deforestation is one example of the way in which human alterations of the natural landscape are changing climate. Climate model experiments show that the replacement of tropical forests with pastures causes a warmer, drier climate. Desertification, in which deserts encroach into forest landscapes, also results in a warmer, drier climate. Conversely, vegetation expansion into deserts, as happened 6000 years ago in North Africa, causes a cooler, wetter climate. By masking the high albedo of snow, the boreal forest creates a warmer climate compared to simulations in which the boreal forest is replaced with tundra vegetation. In Asia, monsoon circulations are created by land–sea temperature contrasts. A high land albedo, such as occurs when there is increased snow cover in Tibet, cools the surface, decreasing the land–sea temperature contrast and causing less rainfall.

2 SURFACE ENERGY FLUXES AND TEMPERATURE

The radiation that impinges on a surface or object must be balanced by the energy re-radiated back to the atmosphere, energy lost or gained as sensible and latent heat, and heat storage. More formally, the energy balance at the surface is

$$(1 - r) \cdot S{\downarrow} + a \cdot L{\downarrow} = L{\uparrow} + H + \lambda \cdot E + G$$

The first term in this equation, $(1 - r) \cdot S{\downarrow}$, is the solar radiation absorbed by the surface. $S{\downarrow}$ is the radiation onto the surface and r is the albedo, which is defined as the fraction of $S{\downarrow}$ that is reflected by the surface. The remainder, $(1 - r)$, is absorbed by the surface. The second term, $a \cdot L{\downarrow}$, is the atmospheric longwave radiation absorbed by the surface, where a is the fraction of the incoming radiation $L{\downarrow}$ that is absorbed. Together, the absorbed solar radiation and longwave radiation comprise the radiative forcing, Q_a. This must be balanced by:

1. Longwave radiation emitted by the surface ($L\uparrow$) in proportion to its absolute temperature, in kelvins, raised to the fourth power.
2. Energy transferred to or from the surface by convection (H). This sensible heat flux is directly proportional to the temperature difference between the surface and air and inversely proportional to a transfer resistance.
3. Energy used to evaporate water ($\lambda \cdot E$). The latent heat flux is directly proportional to the vapor pressure difference between the surface and air and is inversely proportional to a transfer resistance.
4. Energy stored in the soil via conduction (G). When the ground beneath the surface is colder than the surface, heat is conducted into the ground. At night, when the surface is colder than the ground, heat is transferred from the ground to warm the surface.

These energy flows at the surface must be balanced. This balance is maintained by changing the surface temperature. For example, the temperature of a surface will rise as more radiation is received on the surface. As a result, more energy is returned to the atmosphere as longwave radiation and as sensible heat. More energy is stored in the ground via conduction. As the latent heat flux increases, the surface temperature cools. During the day, when the surface is warmer than the air, this decreases the sensible heat flux. If the underlying soil is a good conductor of heat, and there is a large soil heat flux, the surface will not be as hot and the sensible heat flux will decrease to compensate for the increased heat storage. Conversely, if the soil is a poor conductor of heat, little heat will be transferred from the surface to the soil and the surface will be hot.

The importance of these energy fluxes in determining surface temperature can be illustrated with a simple example. Suppose a leaf has a radiative forcing of 1000, 700, and 400 W/m^2, which are representative of values for a clear sky at mid-day, a cloudy sky at mid-day, and at night, when solar radiation is zero and the surface receives only longwave radiation. If the only means to dissipate this energy is through re-radiation (i.e., $H = 0$ and $\lambda \cdot E = 0$) and there is no heat storage ($G = 0$), the leaf surface would attain temperatures of 91, 60 and 17°C with the high, moderate, and low radiative forcings (Table 1). When heat loss by convection

TABLE 1 Temperatures of Well-Watered Leaf for Radiative Forcings[a]

	Temperature (°C)						
		$L\uparrow + H$			$L\uparrow + H + \lambda E$		
Q_a (W/m^2)	$L\uparrow$	0.1 m/s	0.9 m/s	4.5 m/s	0.1 m/s	0.9 m/s	4.5 m/s
1000	91	53	39	34	39	33	31
700	60	40	34	32	32	29	29
400	17	26	28	29	23	26	27

[a] Air temperature is 29°C, relative humidity is 50%, and wind speeds are 0.1, 0.9, and 4.5 m/s.

is included, leaf temperature depends on wind speed because the transfer resistance decreases with high wind speeds. Under calm conditions, with a wind speed of 0.1 m/s, leaf temperature decreases by 38°C with the high radiative forcing and by 20°C with the moderate forcing (Table 1). Higher wind speeds lead to even cooler temperatures. At the low radiative forcing, convection warms the leaf because it is colder than the surrounding air and heat is transferred from the air to the leaf. Latent heat exchange is also a powerful means to cool a surface because of the large amount of energy required to evaporate water. For a well-watered leaf, under calm conditions and high radiative forcing, evapotranspiration cools the leaf an additional 14°C to a temperature of 39°C (Table 1). Higher winds result in even lower temperatures.

3 HYDROLOGIC CYCLE

As the preceding example shows, evapotranspiration is an effective means to cool a surface, particularly at high radiative forcing and low wind speed. The rate of latent heat loss depends on the amount of water present. A well-watered site has more water to evaporate than a dry site. Because more energy goes into latent heat rather than sensible heat, the lower atmosphere is likely to be cool and moist. A dry surface, on the other hand, has low latent heat flux, high sensible heat flux, and the air is likely to be warm and dry. Typical values of the Bowen ratio (the ratio of sensible to latent heat) are: 0.1 to 0.3 for tropical rain forests, where high annual rainfall keeps the soil wet; 0.4 to 0.8 for temperate forests and grasslands, where less rainfall causes drier soils; 2.0 to 6.0 for semiarid regions; and greater than 10.0 for deserts.

The water stored on land (ΔW) is the difference between water input as precipitation (P) and water loss via evapotranspiration (E) and runoff (R):

$$\Delta W = P - E - R$$

In many regions, water is stored as snow in winter and not released to the soil until the following spring. Snow is climatologically important because its high albedo reflects a majority of the solar radiation. Snow has a low thermal conductivity, and a thick snow pack insulates the soil. In spring, a large portion of the net radiation at the surface is used for snow melt, preventing the surface from warming. Precipitation is greatly modified by vegetation. Some of the rainfall is intercepted by foliage, branches, and stems. The remainder reaches the ground as throughfall, stemflow, and snowmelt.

Only a portion of the liquid water reaching the ground surface infiltrates into the soil. The actual amount depends on soil wetness, soil type, and the intensity of the water flux. Table 2 shows representative hydraulic conductivity when the soil is saturated. Sandy soils, because of their large pore sizes, can absorb water at fast rates. Loamy soils absorb water at slower rates. Clay soils, because of their small pores, have the lowest hydraulic conductivity. The water that does not infiltrate into

TABLE 2 Hydraulic Conductivity at Saturation and Water Contents

	Volumetric Water Content (mm^3/mm^3)			
	Wilting Point	Field Capacity	Saturation	Hydraulic Conductivity (mm/s)
Sand	0.07	0.23	0.40	0.176
Sandy loam	0.11	0.32	0.44	0.035
Loam	0.15	0.39	0.45	0.007
Silty clay loam	0.22	0.42	0.48	0.002
Clay	0.29	0.45	0.48	0.001

the soil accumulates in small depressions or is lost as surface runoff, which flows overland to streams, rivers, and lakes.

The water balance of the soil is the difference between water input via infiltration and water loss from evapotranspiration and subsurface drainage. Since the latent heat flux decreases as soil becomes drier, the amount of water in the soil is a crucial determinant of the surface climate. Two hydraulic parameters determine the amount of water a soil can hold. Field capacity is the amount of water after gravitational drainage. Sandy soils, because of their large pores, hold less water at field capacity than clay soils, with their small pores (Table 2). Wilting point is the amount of water in the soil when evapotranspiration ceases. Because water is tightly bound to the soil matrix, clay soils have very high wilting points. The difference between field capacity and wilting point is the available water holding capacity. Loamy soils hold the most water; sands and clays hold the least amount of water.

4 VEGETATION

The ecological characteristics of the surface vary greatly among vegetation types. Some plants absorb more solar radiation than others; some plants are taller than others; some have more leaves; some have deeper roots. For example, the albedo of coniferous forests generally ranges from 0.10 to 0.15; deciduous forests have albedos of 0.15 to 0.20; grasslands have albedos of 0.20 to 0.25. As albedo increases, the amount of solar radiation absorbed at the surface decreases and if all other factors were equal the surface temperature would decrease. However, plants also vary in height. Trees are taller than shrubs, which are taller than grasses. Taller surfaces are “rougher” than shorter surfaces and exert more stress on atmospheric motions. This creates more turbulence, increasing the transfer of sensible and latent heat away from the surface. Plants also increase the surface area from which sensible and latent heat can be transferred to the atmosphere. The leaf area of plant communities is several times that of the underlying ground. A typical ratio of leaf area to ground area (called the leaf area index) is 5 to 8. Plant communities differ greatly in leaf area index. A dry, unproductive site supports much less foliage than a moist, nutrient-rich grassland or forest. The rooting depth of plants is important because this determines how deep in the soil water can be obtained for transpiration. Shallow rooted plants have a much smaller volume of water for evapotranspiration than deep rooted plants.

Leaves have microscopic pores, called stomata, through which they absorb carbon dioxide from the atmosphere during photosynthesis. These pores open and close in response to environmental factors such as light, temperature, atmospheric CO_2 concentration, and soil water. When they are open, the plant absorbs CO_2; but water also diffuses out of the leaf to the surrounding air—a process known as transpiration. Plants differ greatly in their stomatal physiology, especially responses to environmental factors. Some plants photosynthesize, and hence have open stomata, at lower light levels than others. Plants differ in the water content at which stomata close. They differ in optimum temperatures for photosynthesis, and they differ in their responses to increasing CO_2 concentration. These different stomatal physiologies contribute to variations in latent heat flux, and hence surface temperature, among vegetation types.

5 COUPLING TO ATMOSPHERIC MODELS

Different surfaces, with different soil and vegetation types, can create vastly different climates. These differences can be examined by coupling models of surface energy fluxes, which depend on the hydrological and ecological state of the land, with atmospheric numerical models. The surface model provides to the atmospheric model albedo and emitted longwave radiation, which determine the net radiative heating of the atmosphere; sensible and latent heat fluxes, which determine atmospheric temperature and humidity; and surface stresses, which determine atmospheric winds. In turn, the atmospheric model provides the atmospheric conditions required to calculate these fluxes: temperature, humidity, winds, precipitation, and incident solar and longwave radiation.

Land surface models account for the ecological effects of different vegetation types and the hydrological and thermal effects of different soil types. Although the equations needed to model these processes at a single point are well understood, the scaling over large, heterogeneous areas comprised of many soils and many plant communities is much less exact. Moreover, when coupled to a global climate model that may be used to simulate the climate of thousand of points on the surface for tens to hundreds of years, there is a need to balance model complexity with computational efficiency.

CHAPTER 12

OBSERVATIONS OF CLIMATE AND GLOBAL CHANGE FROM REAL-TIME MEASUREMENTS

DAVID R. EASTERLING AND THOMAS R. KARL

1 INTRODUCTION

Is the planet getting warmer?
Is the hydrologic cycle changing?
Is the atmospheric/oceanic circulation changing?
Is the weather and climate becoming more extreme or variable?
Is the radiative forcing of the climate changing?

These are the fundamental questions that must be answered to determine if climate change is occurring. However, providing answers is difficult due to an inadequate or nonexistent worldwide climate observing system. Each of these apparently simple questions are quite complex because of the multivariate aspects of each question and because the spatial and temporally sampling required to address adequately each question must be considered on a global scale. A brief review of our ability to answer these questions reveals many successes, but points to some glaring inadequacies that must be addressed in any attempt to understand, predict, or assess issues related to climate and global change.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

2 IS THE PLANET GETTING WARMER?

There is no doubt that measurements show that near-surface air temperatures are increasing. Best estimates suggest that the warming is around 0.6°C (+ − 0.2°C) since the late nineteenth century (IPCC, 2001). Furthermore, it appears that the decade of the 1990s was the warmest decade since the 1860s, and possibly for the last 1000 years. Although there remain questions regarding the adequacy of this estimate, confidence in the robustness of this warming trend is increasing (IPCC, 2001). Some of the problems that must be accounted for include changes in the method of measuring land and marine surface air temperatures from ships, buoys, land surface stations as well as changes in instrumentation, instrument exposures and sampling times, and urbanization effects. However, recent work evaluating the effectiveness of corrections of sea surface temperatures for time-dependent biases, and further evaluation of urban warming effects on the global temperature record have increased confidence in these results. Furthermore, by consideration of other temperature-sensitive variables, e.g., snow cover, glaciers, sea level and even some proxy non-real-time measurements such as ground temperatures from boreholes, increases our confidence in the conclusion that the planet has indeed warmed. However, one problem that must be addressed is that the measurements we rely upon to calculate global changes of temperature have never been collected for that purpose, but rather to aid in navigation, agriculture, commerce, and in recent decades for weather forecasting. For this reason there remain uncertainties about important details of the past temperature increase and our capabilities for future monitoring of the climate. The IPCC (2001) has summarized latest known changes in the temperature record, which are summarized in Figure 1.

Global-scale measurements of layer averaged atmospheric temperatures and sea surface temperatures from instruments aboard satellites have greatly aided our ability to monitor global temperature change (Spencer and Christy, 1992a,b; Reynolds, 1988), but the situation is far from satisfactory (Hurrell and Trenberth, 1996). Changes in satellite temporal sampling (e.g., orbital drift), changes in atmospheric composition (e.g., volcanic emissions), and technical difficulties related to overcoming surface emissivity variability are some of the problems that must be accounted for, and reduce the confidence that can be placed on these measurements (NRC, 2000). Nonetheless, the space-based measurements have shown, with high confidence, that stratospheric temperatures have decreased over the past two decades. Although perhaps not as much as suggested by the measurements from weather balloons, since it is now known that the data from these balloons high in the atmosphere have an inadvertent temporal bias due to improvements in shielding from direct and reflected solar radiation (Luers and Eskridge, 1995).

3 IS THE HYDROLOGIC CYCLE CHANGING?

The source term for the hydrologic water balance, precipitation, has been measured for over two centuries in some locations, but even today it is acknowledged that in

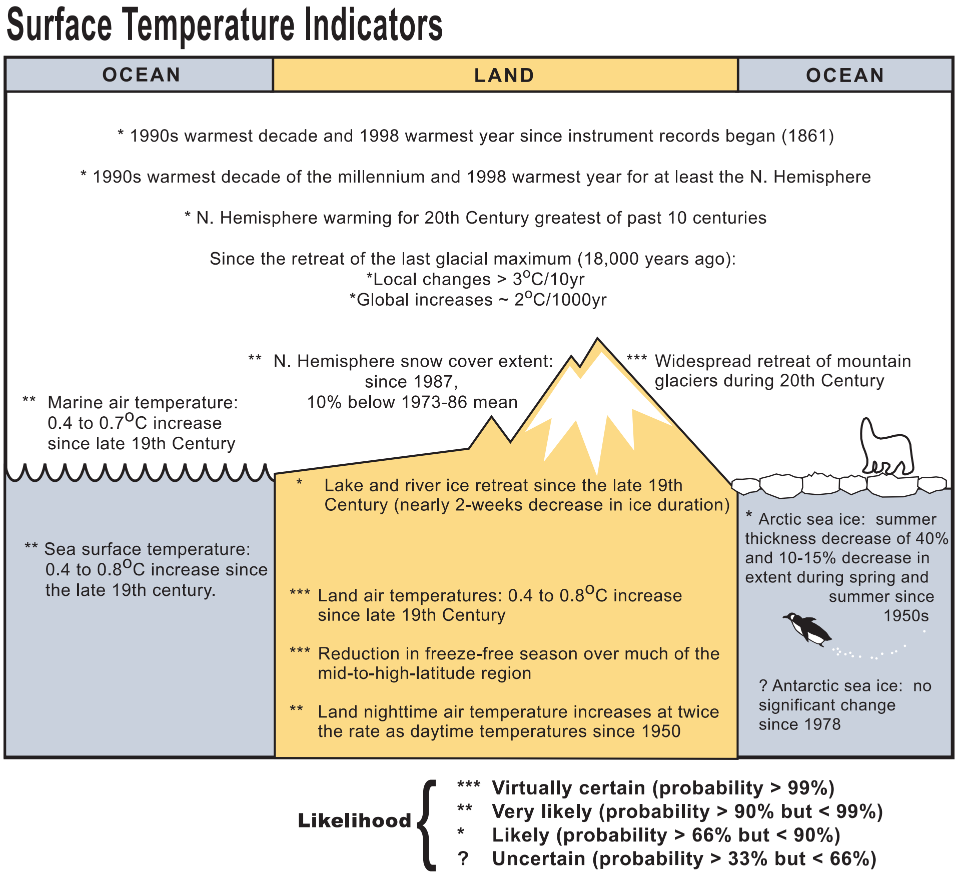

Surface Temperature Indicators
OCEAN
LAND
OCEAN
* 1990s warmest decade and 1998 warmest year since instrument records began (1861)
* 1990s warmest decade of the millennium and 1998 warmest year for at least the N. Hemisphere
* N. Hemisphere warming for 20th Century greatest of past 10 centuries
Since the retreat of the last glacial maximum (18,000 years ago):
*Local changes > 3°C/10yr
*Global increases ~ 2°C/1000yr
** Marine air temperature: 0.4 to 0.7°C increase since late 19th Century
** N. Hemisphere snow cover extent: since 1987, 10% below 1973-86 mean
*** Widespread retreat of mountain glaciers during 20th Century
** Sea surface temperature: 0.4 to 0.8°C increase since the late 19th century.
* Lake and river ice retreat since the late 19th Century (nearly 2-weeks decrease in ice duration)
*** Land air temperatures: 0.4 to 0.8°C increase since late 19th Century
*** Reduction in freeze-free season over much of the mid-to-high-latitude region
** Land nighttime air temperature increases at twice the rate as daytime temperatures since 1950
* Arctic sea ice: summer thickness decrease of 40% and 10-15% decrease in extent during spring and summer since 1950s
? Antarctic sea ice: no significant change since 1978
Likelihood
*** Virtually certain (probability > 99%)
** Very likely (probability > 90% but < 99%)
* Likely (probability > 66% but < 90%)
? Uncertain (probability > 33% but < 66%)

Figure 1 Schematic of observed variations of selected indictors regarding (*a*) temperature and (*b*) the hydrologic cycle (based on IPCC, 2001). See ftp site for color image.

many parts of the world we still cannot reliably measure true precipitation (Sevruk, 1982). For example, annual biases of more than 50% due to rain gauge undercatch are not uncommon in cold climates (Karl et al., 1995), and, even for more moderate climates, precipitation is believed to be underestimated by 10 to 15% (IPCC, 1992). Progressive improvements in instrumentation, such as the introduction of wind shields on rain gauges, have also introduced time-varying biases (Karl et al., 1995). Satellite-derived measurements of precipitation have provided the only large-scale ocean coverage of precipitation. Although they are comprehensive estimates of large-scale spatial precipitation variability over the oceans where few measurements exist, problems inherent in developing precipitation estimates hinder our ability to have much confidence in global-scale decadal changes. For example, even the landmark work of Spencer (1993) in estimating worldwide ocean precipitation using the microwave sounding unit aboard the National Oceanic and Atmospheric Administration (NOAA) polar orbiting satellites has several limitations. The observations are limited to ocean coverage and hindered by the requirement of an unfrozen ocean. They do not adequately measure solid precipitation, have low spatial resolution, and are affected by the diurnal sampling inadequacies associated with polar orbiters, e.g., limited overflight capability. Blended satellite/in situ estimates also show promise (Huffman et al., 1997); however, there are still limitations, including a lack of long-term measurements necessary for climate change studies.

Information about past changes in land surface precipitation, similar to temperature, has been compared with other hydrologic data, such as changes in stream flow, to ascertain the robustness of the documented changes of precipitation. Figure 1 summarizes some of the more important changes of precipitation, such as the increase in the mid to high latitude precipitation and the decrease in subtropical precipitation. Evidence also suggests that much of the increase of precipitation in mid to high latitudes arises from increased autumn and early winter precipitation in much of North America and Europe. Figure 2 depicts the spatial aspects of this change, reflecting rather large-scale coherent patterns of change during the twentieth century.

Other changes related to the hydrologic cycle are summarized in Figure 1. The confidence is low for many of the changes, and it is particularly disconcerting relative to the role of clouds and water vapor in climate feedback effects.* Observations of cloud amount long have been made by surface-based human observations and more recently by satellite. In the United States, however, human observers have been replaced by automated measurements, and neither surface-based or spaced-based data sets have proven to be entirely satisfactory for detecting changes in clouds. Polar orbiting satellites have an enormous difficulty to overcome related to sampling aliasing and satellite drift (Rossow and Cairns, 1995). For human observers changes in observer schedules, observing biases, and incomplete sampling have created major problems in data interpretations, now compounded by a change to new automated measurements at many stations. Nonetheless, there is still some confi-

*An enhancement or diminution of global temperature increases or decreases due to other causes.

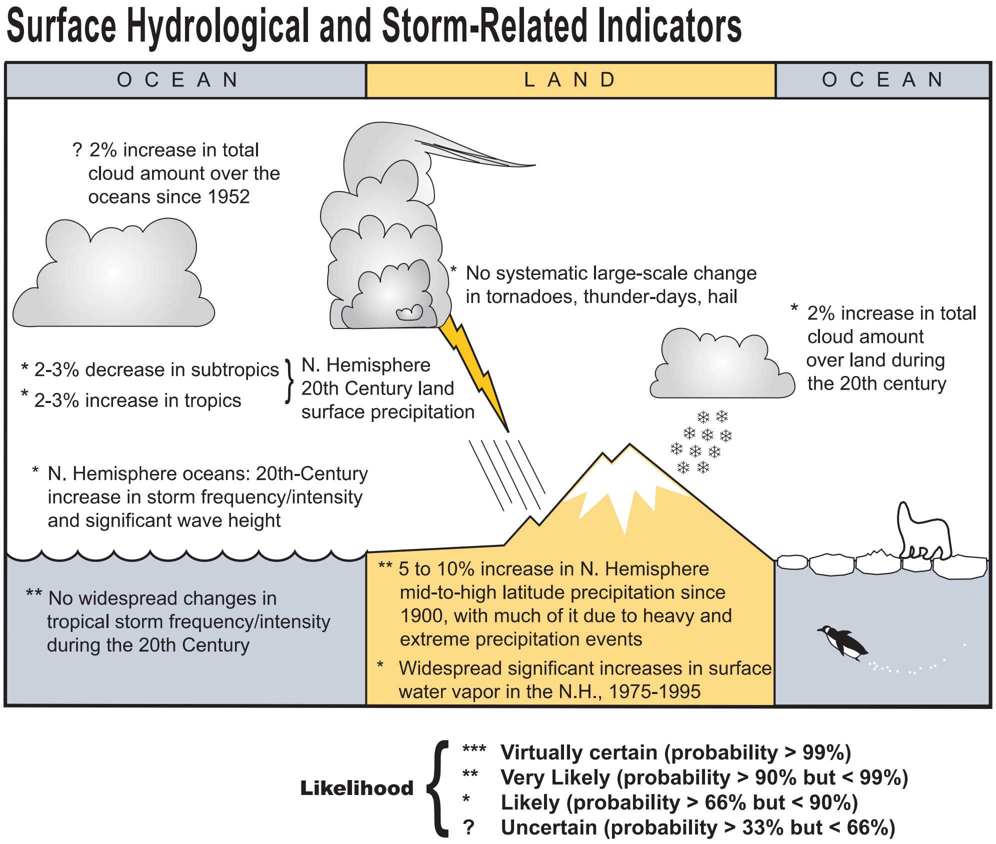

Surface Hydrological and Storm-Related Indicators
O C E A N
L A N D
O C E A N
? 2% increase in total cloud amount over the oceans since 1952
* No systematic large-scale change in tornadoes, thunder-days, hail
* 2% increase in total cloud amount over land during the 20th century
* 2-3% decrease in subtropics
* 2-3% increase in tropics
N. Hemisphere 20th Century land surface precipitation
* N. Hemisphere oceans: 20th-Century increase in storm frequency/intensity and significant wave height
** No widespread changes in tropical storm frequency/intensity during the 20th Century
** 5 to 10% increase in N. Hemisphere mid-to-high latitude precipitation since 1900, with much of it due to heavy and extreme precipitation events
* Widespread significant increases in surface water vapor in the N.H., 1975-1995
Likelihood
*** Virtually certain (probability > 99%)
** Very Likely (probability > 90% but < 99%)
* Likely (probability > 66% but < 90%)
? Uncertain (probability > 33% but < 66%)

Figure 2 Precipitation trends over land 1900–1999. Trend is expressed in percent per century (relative to the mean precipitation from 1961–1990) and magnitude of trend is represented by area of circle with green reflecting increases and brown decreases of precipitation. See ftp site for color image.

dence (but low) that global cloud amounts have tended to increase. On a regional basis this is supported by reductions in evaporation as measured by pan evaporimeters over the past several decades in Russia and the United States, and a worldwide reduction in the land surface diurnal temperature range. Moreover, an increase in water vapor has been documented over much of North America and in the tropics (IPCC, 2001).

Changes in water vapor are very important for understanding climate change since water vapor is the most important greenhouse gas in the atmosphere. The measurement of changes in atmospheric water vapor is hampered by both data processing and instrumental difficulties for both weather balloon and satellite retrievals. The latter also suffers from discontinuities among successive satellites and errors introduced by changes in orbits and calibrations. Upper tropospheric water vapor is particularly important for climate feedbacks, but, as yet, little can be said about how it has varied over the course of the past few decades.

4 IS THE ATMOSPHERIC/OCEANIC CIRCULATION CHANGING?

Surprisingly, there is a considerable lack of reliable information about changes in atmospheric circulation, even though it is of daily concern to much of the world since it relates to day-to-day weather changes. Analyses of circulation are performed every day on a routine basis, but the analysis schemes have changed over time, making them of limited use for monitoring climate change. Moreover, even the recent reanalysis efforts by the world's major numerical weather prediction centers, whereby the analysis scheme is fixed over the historical record, contains time-varying biases because of the introduction of data with time-dependent biases and a changing mix of data (e.g., introducing satellite data) over the course of the reanalysis (Trenberth and Guillemot, 1997). Even less information is available on measured changes and variations in ocean circulation.

A few major atmospheric circulation features have been reasonably well measured because they can be represented by rather simple indices. This includes the El Niño–Southern Oscillation (ENSO) index, the North Atlantic Oscillation (NAO) index, and the Pacific–North American (PNA) circulation pattern index. There are interesting decadal and multidecadal variation, but it is too early to detect any long-term trends. Evidence exists that ENSO has varied in period, recurrence interval, and strength of impact. A rather abrupt change in ENSO and other aspects of atmospheric circulation seems to have occurred around 1976–1977. More frequent ENSOs with rare excursions into its other extreme (La Niña) became much more prevalent. Anomalous circulation regimes associated with ENSO and large-amplitude PNA patterns persisted in the North Pacific from the late 1970s into the late 1980s, affecting temperature anomalies. Moreover, the NAO has been persistent in its association with strong westerlies into the European continent from the late 1980s until very recently when it abruptly shifted. As a result, temperature anomalies and storminess in Europe have abruptly changed over the past 2 years compared to the past 7 or 8 years.

Increases in the strength of the Southern Hemisphere circumpolar vortex during the 1980s have been documented (van Loon et al., 1993; Hurrell and van Loon, 1994) using station sea level pressure data. This increase was associated with a delayed breakdown in the stratospheric polar vortex and ozone deficit in the Antarctic spring. A near-global sea level pressure data set has been used to identify changes in circulation patterns in the Indian Ocean. Allan et al. (1995) and Salinger et al. (1995) find that circulation patterns in the periods 1870–1900 and 1950–1990 were more meridional than those in the 1900–1950 period, indicating intensified circulation around anticyclones. These changes may be related to changes in the amplitude of longwave troughs to the south and west of Australia and the Tasman Sea/New Zealand area and a subsequent decrease in precipitation in Southwest Australia (Nicholls and Lavery, 1992; Allan and Haylock, 1993).

5 IS THE WEATHER AND CLIMATE BECOMING MORE EXTREME OR VARIABLE?

Climate and weather extremes are of great interest. Due to inadequate monitoring as well as prohibitively expensive access to weather and climate data held by the world's national weather and environmental agencies, only limited reliable information is available about large-scale changes in extreme weather or climate variability. The time-dependent biases that affect climate means are even more difficult to effectively eliminate from the extremes of the distributions of various weather and climate elements. There are a few areas, however, where regional and global changes in weather and climate extremes have been reasonably well documented (Easterling et al., 2000a).

Interannual temperature variability has not changed significantly over the past century. On shorter time scales and higher frequencies, e.g., days to a week, temperature variability may have decreased across much of the Northern Hemisphere (Karl and Knight, 1995). Related to the decrease in high-frequency temperature variability there has been a tendency for fewer low-temperature extremes, but widespread changes in extreme high temperatures have not been noted.

Trends in intense rainfall have been examined for a variety of countries. Some evidence suggests an increase in intense rainfalls (United States, tropical Australia, Japan, and Mexico), but analyses are far from complete and subject to many discontinuities in the record. The strongest increases in extreme precipitation are documented in the United States and Australia (Easterling et al., 2000b)

Intense tropical cyclone activity may have decreased in the North Atlantic, the one basin with reasonably consistent tropical cyclone data over the twentieth century, but even here data prior to World War II is difficult to assess regarding tropical cyclone strength. Elsewhere, tropical cyclone data do not reveal any long-term trends, or if they do they are most likely a result of inconsistent analyses. Changes in meteorological assimilation schemes have complicated the interpretations of changes in extratropical cyclone frequency. In some regions, such as the North Atlantic, a clear trend in activity has been noted, as also in significant wave heights in the northern half of the North Atlantic. In contrast, decreases in storm frequency and wave heights have been noted in the south half of the North Atlantic over the past few decades. These changes are also reflected in the prolonged positive excursions of the NAO since the 1970s.

6 IS THE RADIATIVE FORCING OF THE PLANET CHANGING?

Understanding requires a time history of forcing global change. The atmospheric concentration of CO_2, an important greenhouse gas because of its long atmospheric residence time and relatively high atmospheric concentration, has increased substantively over the past few decades. This is quite certain as revealed by precise measurements made at the South Pole and at Mauna Loa Observatory since the late 1950s, and from a number of stations distributed globally that began operating in subsequent decades. Since atmospheric carbon dioxide is a long-lived atmospheric constituent and it is well mixed in the atmosphere, a moderate number of well-placed stations operating for the primary purpose of monitoring seasonal to decadal changes provides a very robust estimate of global changes in carbon dioxide.

To understand the causes of the increase of atmospheric carbon dioxide, the carbon cycle and the anthropogenic carbon budget must be balanced. Balancing the carbon budget requires estimates of the sources of carbon from anthropogenic emissions from fossil fuel and cement production, as well as the net emission from changes in land use (e.g., deforestation). These estimates are derived from a combination of modeling, sample measurements, and high-resolution satellite imagery. It also requires measurements for the storage in the atmosphere, the ocean uptake, and uptake by forest regrowth, the CO_2 and nitrogen fertilization effect on vegetation, as well as any operating climate feedback effects (e.g., the increase in vegetation due to increased temperatures). Many of these factors are still uncertain because of a paucity of ecosystem measurements over a sustained period of time. Anthropogenic emissions from the burning of fossil fuel and cement production are the primary cause of the atmospheric increase.

Several other radiatively important anthropogenic atmospheric trace constituents have been measured for the past few decades. These measurements have confirmed significant increases in atmospheric concentrations of CH_4, N_2O, and the halocarbons including the stratospheric ozone destructive agent of the chloroflourocarbons and the bromocarbons. Because of their long lifetimes, a few well-placed high-quality in situ stations have provided a good estimate of global change. Stratospheric ozone depletion has been monitored both by satellite and ozonesondes. Both observing systems have been crucial in ascertaining changes of stratospheric ozone that was originally of interest, not because of its role as a radiative forcing agent, but its ability to absorb ultraviolet (UV) radiation prior to reaching Earth's surface. The combination of the surface- and space-based observing systems has enabled much more precise measurements than either system could provide by itself. Over the past few years much of the ozonesonde data and satellite data has been improved using information about past calibration methods, in part because of differences in trends between the two observing systems.

Figure 3 depicts the IPCC (1995) best estimate of the radiative forcing associated with various atmospheric constituents. Unfortunately, measurements of most of the forcings other than those already discussed have low or very low confidence, not only because of our uncertainty about their role in the physical climate system, but because we have not adequately monitored their change. For example, estimates of changes in sulfate aerosol concentrations are derived from model estimates of source emissions, not actual atmospheric concentrations. The problem is complicated because of the spatially varying concentrations of sulfate due to its short atmospheric lifetime. Another example is measurements of solar irradiance, which have been taken by balloons and rockets for several decades, but continuous measurements of top-of-the-atmosphere solar irradiance did not begin until the late 1970s with the *Nimbus 7* and the Solar Maximum Mission satellites. There are significant absolute differences in total irradiance between satellites, emphasizing the critical need for overlaps between satellites and absolute calibration of the irradiance measurements to determine decadal changes. Spectrally resolved measurements will be a key element in our ability to model the effects of solar variability, but at the present time no long-term commitment has been made to take these measurements. Another important forcing that is estimated through measured, modeled, and estimated changes in optical depth relates to the aerosols injected high into the atmosphere by major volcanic eruptions. The aerosols from these volcanoes are sporadic and usually persist in the atmosphere for at most a few years. Improved measurements of

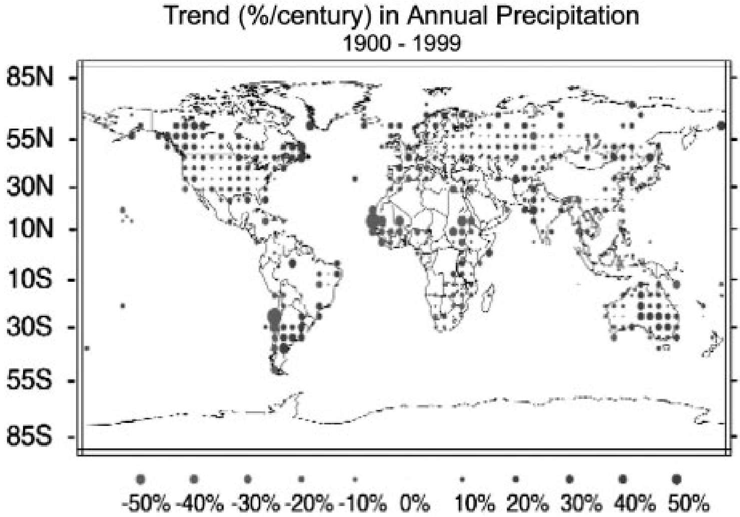


Figure 3 (see color insert) Estimates of globally and annually averaged radiative forcing (in W/m^{-2}) for a number of agents due to changes in concentrations of greenhouse gases and aerosols and natural changes in solar output from 1750 to the present day. Error bars are depicted for all forcings (from IPCC, 2001). See ftp site for color image.

aerosol size distribution and composition will help better understand this agent of climate change.

7 WHAT CAN WE DO TO IMPROVE OUR ABILITY TO DETECT CLIMATE AND GLOBAL CHANGE?

Even after extensive reworking of past data, in many instances we are incapable of resolving important aspects concerning climate and global change. Virtually every monitoring system and data set requires better data quality, continuity, and fewer time-varying biases if we expect to conclusively answer questions about how the planet has changed, because of the need to rely on observations that were never intended to be used to monitor the physical characteristics of the planet of the course of decades. Long-term monitoring, capable of resolving decade-to-century-scale changes, requires different strategies of operation.

In situ measurements are currently in a state of decay, decline, or rapid poorly documented change due to the introduction of automated measurements without adequate precaution to understand the difference between old and new observing

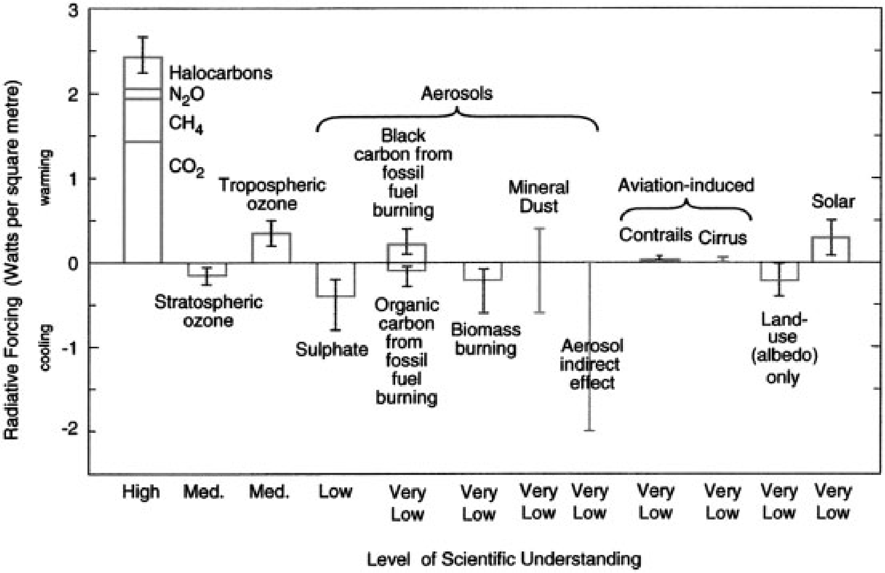


Figure 4 Global, annual-mean radiative forcings (Wm^{-2}) due to a number of agents for the period from pre-industrial (1750) to the present. The height of the vertical bar denotes the central or "best" estimate, no bar indicates that it is not possible to provide a "best" estimate. The vertical line indicates an estimate of the uncertainty range and the level of scientific understanding is a subjective judgement about the reliability of the forcing estimate based on such factors as assumptions, degree of knowledge of the physical/chemical mechanisms, etc. (From IPCC 2001). See ftp site for color image.

systems. Satellite-based systems alone will not and cannot provide all the necessary measurements. Much wiser implementation and monitoring practices must be adopted for both space-based and surface-based observing systems in order to adequately understand global change. The establishment of the Global Climate Observing System (GCOS) is a high priority (Spence and Townsend, 1995), and continued encouragement by the World Meteorological Organization (WMO) of a full implementation of this system in all countries is critical. Furthermore, in the context of the GCOS, a number of steps can be taken to improve our ability to monitor climate and global change.

These include:

1. Prior to implementing changes to existing environmental monitoring systems or introducing new observing systems, standard practice should include an assessment of the impact of these changes on our ability to monitor environmental variations and changes.
2. Overlapping measurements in time and space of both the old and new observing systems should be standard practice for critical environmental variables.
3. Calibration, validation, and knowledge of instrument, station, and/or platform history are essential for data interpretation and use. Changes in instrument sampling time, local environmental conditions, and any other factors pertinent to the interpretation of the observations and measurements should be recorded as a mandatory part of the observing routine and be archived with the original data. The algorithms used to process observations must be well documented and available to the scientific community. Documentation of changes and improvements in the algorithms should be carried along with the data throughout the data archiving process.
4. The capability must be established to routinely assess the quality and homogeneity of the historical database for monitoring environmental variations and change, including long-term high-resolution data capable of resolving important extreme environmental events.
5. Environmental assessments that require knowledge of environmental variations and change should be well integrated into a global observing system strategy.
6. Observations with a long uninterrupted record should be maintained. Every effort should be made to protect the data sets that document long-term homogeneous observations. Long term may be a century or more. A list of prioritized sites or observations based on their contribution to long-term environmental monitoring should be developed for each element.
7. Data-poor regions, variables, regions sensitive to change, and key measurements with inadequate temporal resolution should be given the highest priority in the design and implementation of new environmental observing systems.

8. Network designers, operators, and instrument engineers must be provided environmental monitoring requirements at the outset of network design. This is particularly important because most observing systems have been designed for purposes other than long-term monitoring. Instruments must have adequate accuracy with biases small enough to resolve environmental variations and changes of primary interest.

9. Much of the development of new observation capabilities and much of the evidence supporting the value of these observations stem from research-oriented needs or programs. Stable, long-term commitments to these observations, and a clear transition plan from research to operations, are two requirements in the development of adequate environmental monitoring capabilities.

10. Data management systems that facilitate access, use, and interpretation are essential. Freedom of access, low cost, mechanisms that facilitate use (directories, catalogs, browse capabilities, availability of metadata on station histories, algorithm accessibility and documentation, on-line accessibility to data, etc.), and quality control should guide data management. International cooperation is critical for successful management of data used to monitor long-term environmental change and variability.

REFERENCES

Allan, R. J., and M. R. Haylock (1993). Circulation features associated with the winter rainfall decrease in southwestern Australia, *J. Climate* **6**, 1356–1367.

Allan, R. J., J. A. Lindesay, and C. J. C. Reason (1995). Multidecadal variability in the climate system over the Indian Ocean region during the austral summer, *J. Climate* **8**, 1853–1873.

Easterling, D. R., G. Meehl, C. Parmesan, S. Changnon, T. Karl, and L. Mearns (2000a). Climate extremes: Observations, modeling, and impacts, *Science* **289**, 2068–2074.

Easterling, D. R., J. L. Evans, P. Ya. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje (2000b). Observed variability and trends in extreme climate events: A brief review, *Bull. Am. Meteor. Soc.* **81**, 417–426.

Elliott, W. P. (1995). On detecting long-term changes in atmospheric moisture, *Climatic Change* **31**, 219–237.

Groisman, P. Y., and D. R. Legates (1995). Documenting and detecting long-term precipitation trends: Where we are and what should be done, *Climatic Change* **31**, 471–492.

Huffman, G. J., R. F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider (1997). The Global Precipitation Climatology Project (GPCP) Combined Precipitation Dataset, *Bull. Am. Meteor. Soc.* **78**, 5–20.

Hurrell, J. W., and K. E. Trenberth (1996). Satellite versus surface estimates of air temperature since 1979, *J. Climate* **9**, 2222–2232.

Hurrell, J. W., and H. van Loon (1994). A modulation of the atmospheric annual cycle in the Southern Hemisphere, *Tellus* **46A**, 325–338.

IPCC (1992). *Climate Change, 1992, Supplementary Report*, WMO/UNEP, J. T. Houghton, B. A. Callander, and S. K. Varney (Eds.), New York, Cambridge University Press, pp. 62–64.

IPCC (2001). *Climate Change, 2001: The Scientific Basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson (Eds.), New York, Cambridge University Press.

Karl, T. R., R. W. Knight, and N. Plummer (1995). Trends in high-frequency climate variability in the twentieth century, *Nature* **377**, 217–220.

Luers, J. K., and R. E. Eskridge (1995). Temperature corrections for the VIZ and Vaisala radiosondes, *Appl. Meteor.* **34**, 1241–1253.

National Research Council (NRC) (2000). *Reconciling Observations of Global Temperature Change, Report of the Panel on Reconciling Temperature Observations*, Washington, DC, National Academy Press.

Nicholls, N., and B. Lavery (1992). Australian rainfall trends during the twentieth century, *Int. J. Climatology* **12**, 153–163.

Reynolds, R. W. (1988). A real-time global sea surface temperature analysis, *J. Climate* **1**, 75–86.

Rossow, W. B., and B. Cairns (1995). Monitoring changes in clouds, *Climatic Change* **31**, 175–217.

Salinger, M. J., R. Allan, N. Bindoff, J. Hannah, B. Lavery, L. Leleu, Z. Lin, J. Lindesay, J. P. MacVeigh, N. Nicholls, N. Plummer, and S. Torok (1995). Observed variability and change in climate and sea level in Oceania, in *Greenhouse: Coping with Climate Change*, W. J. Bouma, G. I. Pearman, and M. R. Manning (eds.), CSIRO, Melbourne, Australia 100–126.

Sevruk, B. (1982). Methods of correcting for systematic error in point precipitation measurements for operational use, *Hydrology Rep.* **21**, World Meteorological Organization, Geneva 589.

Spence, T., and J. Townsend (1995). The Global Climate Observing System (GCOS), *Climatic Change* **31**, 131–134.

Spencer, R. W. (1993). Global oceanic precipitation from the MSU during 1979–1992 and comparison to other climatologies, *J. Climate* **6**, 1301–1326.

Spencer, R. W., and J. R. Christy (1992a). Precision and radiosonde validation of satellite gridpoint temperature anomalies, Part I: MSU channel 2. *J. Climate* **5**, 847–857.

Spencer, R. W., and J. R. Christy (1992b). Precision and radiosonde validation of satellite gridpoint temperature anomalies, Part II: A tropospheric retrieval and trends during 1979–90, *J. Climate* **5**, 858–866.

Trenberth, K. E., and C. J. Guillemot (1997). Evaluation of the atmospheric moisture and hydrologic cycle in the NCEP reanalysis, *Clim. Dyn.* **14**, 213–231.

van Loon, H., J. W. Kidson, and A. B. Mullan (1993). Decadal variation of the annual cycle in the Australian dataset, *J. Climate* **6**, 1227–1231.

CHAPTER 13

WHY SHOULD WE BELIEVE PREDICTIONS OF FUTURE CLIMATE?

JOHN MITCHELL

1 INTRODUCTION

Three-dimensional models of the atmosphere are based on laws of classical physics, including the conservation of momentum (the Navier–Stokes equations), heat (the first law of thermodynamics, the perfect gas law), mass and water vapor, allowing for sources and sinks. The state variables are temperature. the northerly and easterly wind components, and water vapor. For the ocean, salt is included rather than water vapor, winds are replaced by currents, and the equation of state for seawater is used (Fig. 1).

The state variables are held on a three-dimensional grid, which for current atmospheric models is typically 200 to 300 km in the horizontal and on about 20 levels in the vertical. This gives over 60,000 basic variables. The equations are solved to produce the rates of change of these variables, and hence new values of the state variables a small time interval ahead. This process is repeated over and over again to produce the evolution of the system. In practice, a variety of numerical techniques (explicit, implicit, semi-implicit, semi-Lagrangian) and spatial representations (spectral and finite difference) are used. The time step varies typically from 10 min to an hour.

Other variables are diagnosed each time step from the state variables (e.g., cloudiness, precipitation and latent heat release, radiative heating rates, surface evaporation, ground wetness, and snow cover) and, where appropriate, are used for the source and sink terms in the basic equations. Processes that occur on a scale too small to be represented explicitly by the model grid have to be included approxi-

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

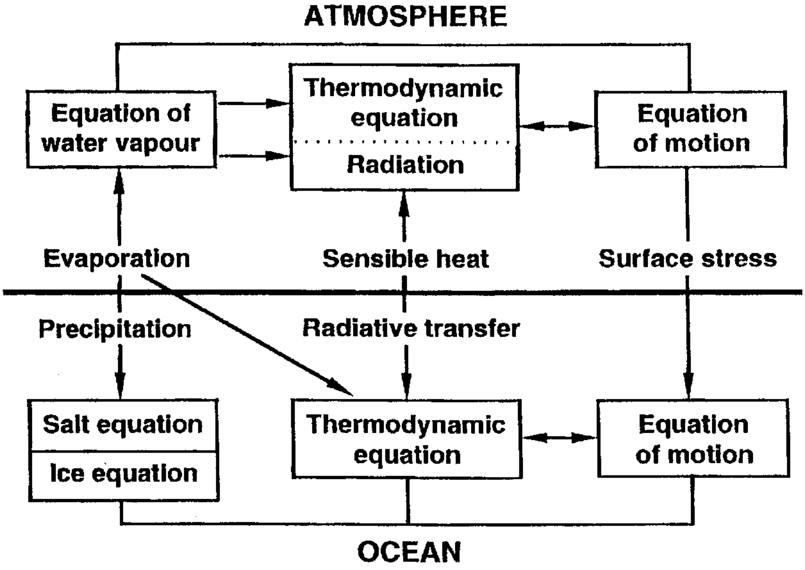


Figure 1 Some factors that affect climate.

mately, by representing them in an idealized way in terms of the grid-scale variables (*parameterization*). The parameterization may be based on one or a combination of the following: well-established theory (e.g., for radiative transfer), field or laboratory observations (e.g., for turbulent mixing of heat, moisture and momentum in the boundary layer), finer scale models (e.g., clouds) and experimentation with the general circulation models (GCM).

The oceanic component is similar to the atmosphere in resolution, although, ideally, higher resolution is required to represent mesoscale eddies, which are the oceanic equivalent of weather, and the parameterizations are generally simpler than in the atmosphere.

Climate models are numerical models of the atmosphere, ocean, and other components of the climate system that are used to understand climate and past climate change and to predict future climate. They range in complexity from simple globally averaged models that attempt to model changes in the energy balance of the climate system to complex three-dimensional GCMs. GCMs were originally developed in parallel with numerical weather prediction models in the 1960s. Indeed, several current GCMs are low-resolution versions of weather prediction models.

Simple climate models are useful in illustrating some aspects of climate change and can be tuned to mimic some of the global mean changes found in GCMs. However, only GCMs can provide information on the detailed geographical distribution of climate change. State-of-the-art GCMs include a full representation of the atmosphere, ocean, sea ice, and land surface (Fig. 2). The exact formulation of a climate model will depend on the use to which it is put. For example, accurate

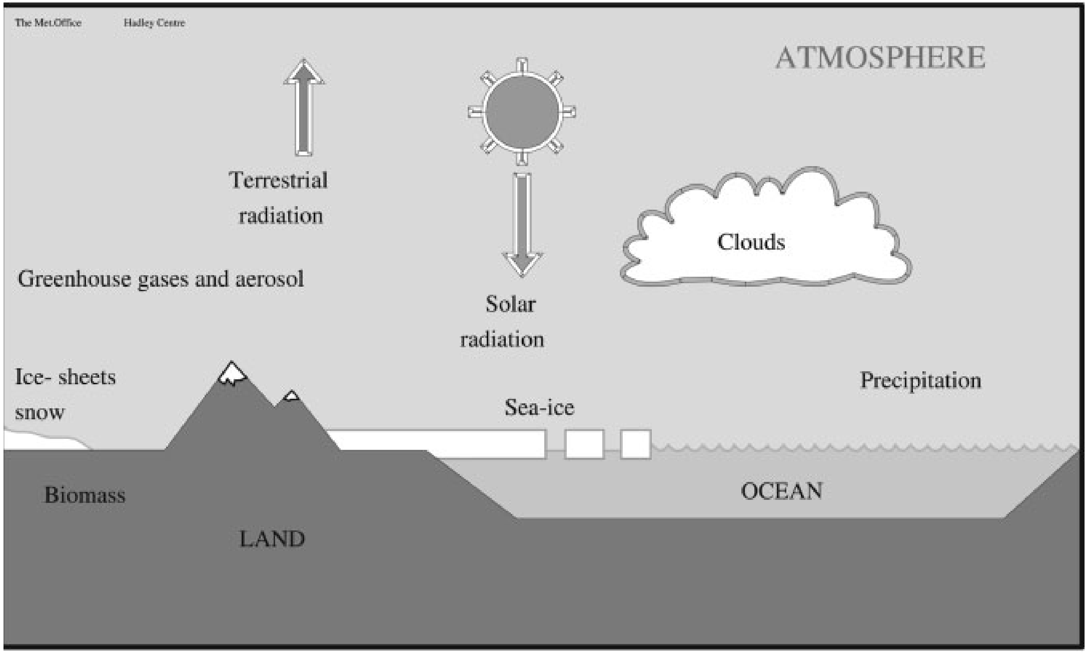


Figure 2 Basic structure of coupled ocean atmosphere GCM. See ftp site for color image.

representation of the radiative effects of greenhouse gases is required if the model is to be used for studies of anthropogenic climate change. Some new GCMs include carbon and sulfur models (to understand and predict changes in atmospheric CO_2 and sulfate aerosol concentrations and their effect on climate) and detailed atmospheric chemistry (to study ozone depletion and changes in tropospheric trace gases).

Climate models are used differently from weather prediction models. Weather prediction models prescribe an initial state of the basic variables from observations, and the equations are stepped forward in time to give the evolution of the atmosphere. For a few days, it is possible to relate individual features in the forecast to features that evolve at the corresponding time in the real world. As the forecast proceeds further, this correspondence disappears as errors due to inexact initial data and model inadequacies grow because of the nonlinear nature of the equations of motion. In other words, the chaotic nature of the equations limits the length of time of deterministic forecasts to about a week or so (depending upon processes involved and the scale of the system being forecast).

A climate model is usually run for long enough that its statistics are independent of the initial conditions. The experiment is then continued over a number of years, decades, or even centuries (depending on the availability of computer time and the application), and the statistics of the simulation over the final period are analyzed. This will include not only the time means, but variability on daily to annual or longer time scales, storm tracks, extreme events, and so on. If the effect of a particular change is being investigated (e.g., doubling atmospheric carbon dioxide concentrations, or changing Earth's orbital parameters), then the experiment is repeated with this change in the model, and the statistics of the two experiments are compared. Both the GCM and observed climate display variations on all time scales due to

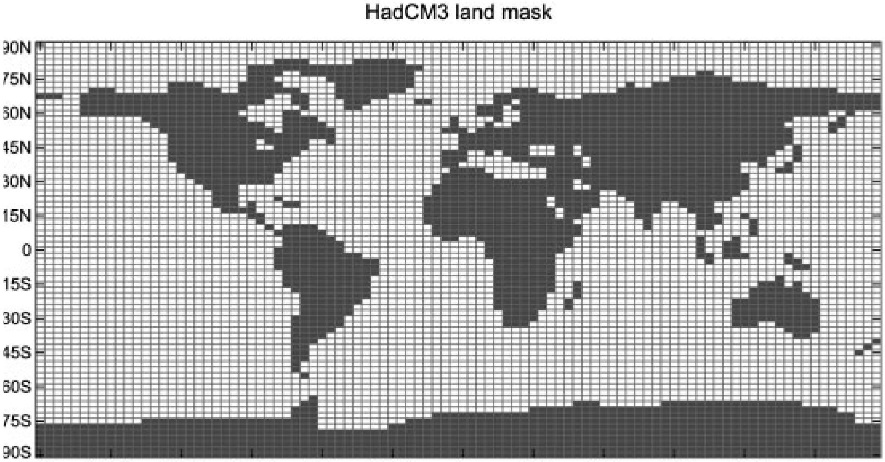


Figure 3 Typical climate model grid.

internal variability; statistical tests may be needed to demonstrate that the differences between the control and anomaly simulations are due to the change in the model, and not to internal variations.

What has been the evolution of climate change in the recent past, and likely changes in the near future? These questions are addressed using coupled ocean atmosphere GCMs. Models have been integrated from preindustrial times to the present using the estimated changes in climate forcing (factors governing climate due to human and natural causes), assuming that the preindustrial climate was in quasi-equilibrium, and then extended assuming some future scenario of greenhouse gas concentrations. As in weather forecasting, errors in initial data and the model will contaminate these "hindcasts" and forecasts. Hence, the most recent studies use an ensemble of simulations started from different initial conditions to give a range of possible future projections taking into account the uncertainty in initial conditions.

How reliable are GCMs for predicting future climate change? With climate, this is argued in several ways. First, they are based on well-established physical laws and principles based on a wealth of observations. Second, they reproduce many of the main features of current climate and its interannual variability, including the seasonal cycle of temperature, the formation and decay of the major monsoons, the seasonal shifts of the major rain belts, the average daily temperature cycle, and the variations in outgoing radiation at high elevations in the atmosphere, as measured by satellites. Similarly many features of the large-scale features in the ocean are reproduced by current climate models.

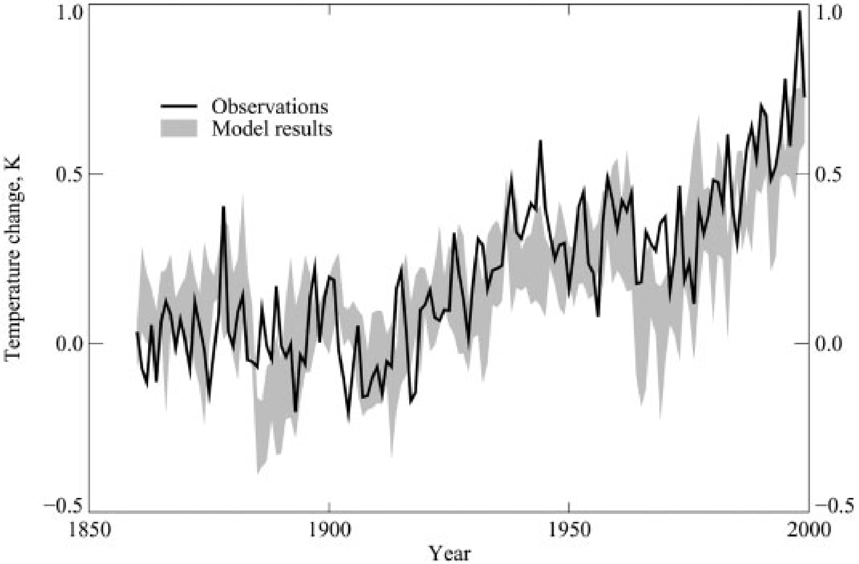


Figure 4 Simulations of recent changes in global mean surface temperature (greyband) compared to observations (black line).

A model may produce a faithful representation of current climate, yet give unreliable predictions of climate change. Hence models have been tested to reproduce both past climates and recent climatic events. Simulations of temperature and precipitation changes 6000 years ago (due to changes in Earth's orbital parameters) and the last glacial maximum (prescribing reconstructed changes in ice sheets, orbital parameters, and greenhouse gases) compare tolerably well with reconstructions from palaeodata. For example, models reproduce the strengthening of the North African monsoon 6000 years ago and the approximate level of cooling estimated for the last ice age. The value of these comparisons is limited by the possibility that other factors not included in the model may have contributed to change in these periods, and the uncertainty in the reconstructions of temperature and precipitation from the palaeodata.

Simulations of the last 130 years have been made driven with the observed increase in greenhouse gases and an estimate of the increase in sulfate particles. The simulated global mean warming agrees tolerably well with observations. Forecasts of the global mean cooling due to the eruption of Mount Pinatubo 1991 were also very successful. However, it is possible in both these instances that errors in the estimate of the factors governing climate in these cases (e.g., changes in ozone and aerosol concentrations) were fortuitously canceled by errors in the model's sensitivity to those changes. Recent patterns of change including cooling of the upper atmosphere, a reduction in diurnal temperature range, and a tentative increase in precipitation in high northern latitudes are in broad qualitative agreement with available observations.

Various factors limit our confidence in models. The sensitivity of global mean temperature change with increasing atmospheric carbon dioxide had different models which varied by a factor of over 2, largely as a result of uncertainty in the treatment of cloud and related processes. Pronounced differences exist in the regional changes in model responses to increasing greenhouse gases, although the broad scale features are quite robust, including enhanced warming in high northern latitudes in all seasons but summer, generally greater warming over the land than ocean, limited warming over the Southern Ocean and northern North Atlantic, and increased annual mean runoff in high latitudes. The validation against recent observed climate change is hampered not only by uncertainty in the factors affecting climate but also in how much of the recent observed changes are due to natural variability and naturally forced events (e.g., due to changes in solar intensity).

In summary, climate models are numerical models used to study climate and climate change. They reproduce many features of observed climate and climate variability, some of the broad features of past climates. Predictions of future global mean temperature change vary by a factor of 2 or so for a given scenario of anthropogenic greenhouse gas emissions and are reasonably consistent on global scales, but not on regional scales. The main areas of current research are reducing model errors by improving the representation of physical process (particularly clouds), reducing uncertainties in the factors affecting climate, particularly in the recent past, and estimating the magnitude of natural climate variability.

BIBLIOGRAPHY

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. S. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, (Eds.) (2001). *Climate Change 2001: The Scientific Basis*. The Third Assessment Report of the IPCC: Contribution of Working Group I, Cambridge University Press, Cambridge, MA.

Trenberth, K. E. (1995). *Numerical Modeling of the Atmosphere—Numerical Weather Prediction Haltiner and Martin Climate Modeling—Climate Systems Modeling*, Cambridge University Press, Cambridge, MA.

CHAPTER 14

THE EL NIÑO–SOUTHERN OSCILLATION (ENSO) SYSTEM

KEVIN TRENBERTH

1 ENSO EVENTS

Every 3 to 7 years or so, a pronounced warming occurs of the surface waters of the tropical Pacific Ocean. The warmings take place from the international dateline to the west coast of South America and result in changes in the local and regional ecology and are clearly linked with anomalous global climate patterns. These warmings have come to be known as *El Niño events*. Historically, El Niño referred to the appearance of unusually warm water off the coast of Peru as an enhancement of the normal warming about Christmastime (hence Niño, Spanish for "the boy Christ-child") and only more recently has the term come to be regarded as synonymous with the basinwide phenomenon. The atmospheric component tied to El Niño is termed the *Southern Oscillation* (SO) whose existence was first noted late in the 1800s. Scientists call the interannual variations where the atmosphere and ocean collaborate together in this way El Niño–Southern Oscillation (ENSO).

The ocean and atmospheric conditions in the tropical Pacific are seldom close to average, but instead fluctuate somewhat irregularly between the warm phase of ENSO, the El Niño events, and the cold phase of ENSO consisting of cooling of the central and eastern tropical Pacific, referred to as *La Niña events* (La Niña is "the girl" in Spanish). The most intense phase of each event lasts about a year.

This chapter briefly outlines the current understanding of ENSO and the physical connections between the tropical Pacific and the rest of the world and the issues involved in exploiting skillful but uncertain predictions.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

2 THE TROPICAL PACIFIC OCEAN–ATMOSPHERE SYSTEM

The distinctive pattern of average sea surface temperatures in the Pacific Ocean sets the stage for ENSO events. Key features are the "warm pool" in the tropical western Pacific, where the warmest ocean waters in the world reside and extend to depths of over 150 m, warm waters north of the equator from about 5 to 15° N, much colder waters in the eastern Pacific, and a cold tongue along the equator that is most pronounced about October and weakest in March. The warm pool migrates with the sun back and forth across the equator but the distinctive patterns of sea surface temperature are brought about mainly by the winds.

The existence of the ENSO phenomenon is dependent on the east–west variations in sea surface temperatures in the tropical Pacific and the close links with sea level pressures, and thus surface winds in the tropics, which in turn determine the major areas of rainfall. The temperature of the surface waters is readily conveyed to the overlying atmosphere, and, because warm air is less dense, it tends to rise while cooler air sinks. As air rises into regions where the air is thinner, the air expands, causing cooling and therefore condensing moisture in the air, which produces rain. Low sea level pressures are set up over the warmer waters while higher pressures occur over the cooler regions in the tropics and subtropics, and the moisture-laden winds tend to blow toward low pressure so that the air converges, resulting in organized patterns of heavy rainfall. The rain comes from convective cloud systems, often as thunderstorms, and perhaps as tropical storms or even hurricanes, which preferentially occur in the *convergence zones*. Because the wind is often light or calm right in these zones, they have previously been referred to as the *doldrums*. Of particular note are the Inter-Tropical Convergence Zone (ITCZ) and the South Pacific Convergence Zone (SPCZ), which are separated by the equatorial dry zone. These atmospheric climatological features play a key role in ENSO as they change in character and move when sea surface temperatures change.

There is a strong coincidence between the patterns of sea surface temperatures and tropical convection throughout the year, although there is interference from effects of nearby land and monsoonal circulations. The strongest seasonal migration of rainfall occurs over the tropical continents, Africa, South America and the Australian–Southeast Asian–Indonesian maritime region. Over the Pacific and Atlantic, the ITCZ remains in the Northern Hemisphere year round, with convergence of the trade winds favored by the presence of warmer water. In the subtropical Pacific, the SPCZ also lies over water warmer than about 27°C. The ITCZ is weakest in January in the Northern Hemisphere when the SPCZ is strongest in the Southern Hemisphere.

The surface winds drive surface ocean currents, which determine where the surface waters flow and diverge, and thus where cooler nutrient-rich waters upwell from below. Because of Earth's rotation, easterly winds along the equator deflect currents to the right in the Northern Hemisphere and to the left in the Southern Hemisphere and thus away from the equator, creating upwelling along the equator. The presence of nutrients and sunlight in the cool surface waters along the equator and western coasts of the Americas favors development of tiny plant species called phytoplankton, which are grazed on by microscopic sea animals called zooplankton, which in turn provide food for fish.

The winds largely determine the sea surface temperature distribution along with differential sea levels and the heat content of the upper ocean. The latter is related to the configuration of the thermocline, which denotes a region of sharp temperature gradient within the ocean separating the well-mixed surface layers from the cooler abyssal ocean waters. Normally the thermocline is deep in the western tropical Pacific (on the order of 150 m) and sea level is high as waters driven by the easterly tradewinds pile up. In the eastern Pacific on the equator, the thermocline is very shallow (on the order of 50 m) and sea level is relatively low. The Pacific sea surface slopes up by about 60 cm from east to west along the equator.

The tropical Pacific, therefore, is a region where the atmospheric winds are largely responsible for the tropical sea surface temperature distribution which, in turn, is very much involved in determining the precipitation distribution and the tropical atmospheric circulation. This sets the stage for ENSO to occur.

3 INTERANNUAL VARIATIONS IN CLIMATE

Most of the interannual variability in the tropics and a substantial part of the variability over the Southern Hemisphere and Northern Hemisphere extratropics is related and tied together through ENSO. ENSO is a natural phenomenon arising from coupled interactions between the atmosphere and the ocean in the tropical Pacific Ocean, and there is good evidence from cores of coral and glacial ice in the Andes that it has been going on for millennia. The region of the Pacific Ocean most involved in ENSO is the central equatorial Pacific (Fig. 1), especially the area 5° N to 5° S, 170° E to 120° W, not the traditional El Niño region along the coast of South America. The evolution of sea surface temperature covering ENSO events after 1950 is shown in Figure 2.

Inverse variations in pressure anomalies (departures from average) at Darwin (12.4° S 130.9° E) in northern Australia and Tahiti (17.5° S 149.6° W) in the South Pacific Ocean characterize the SO. Consequently, the difference in pressure anomalies, Tahiti minus Darwin, is often used as a Southern Oscillation Index (SOI), also given in Figure 2. The warm ENSO events clearly identifiable since 1950 occurred in 1951, 1953, 1957–1958, 1963, 1965, 1969, 1972–1973, 1976–1977, 1982–1983, 1986–1987, 1990–1995, and 1997–1998. The 1990–1995 event is sometimes considered as three events as it faltered briefly in late 1992 and early 1994 but reemerged strongly in each case, and the duration is unprecedented in the past 130 years. Worldwide climate anomalies lasting several seasons have been identified with all of these events.

The SO is principally a global-scale seesaw in atmospheric sea level pressure involving exchanges of air between eastern and western hemispheres (Fig. 3) centered in tropical and subtropical latitudes with centers of action located over Indonesia (near Darwin) and the tropical South Pacific Ocean (near Tahiti). Higher than normal pressures are characteristic of more settled and fine weather,

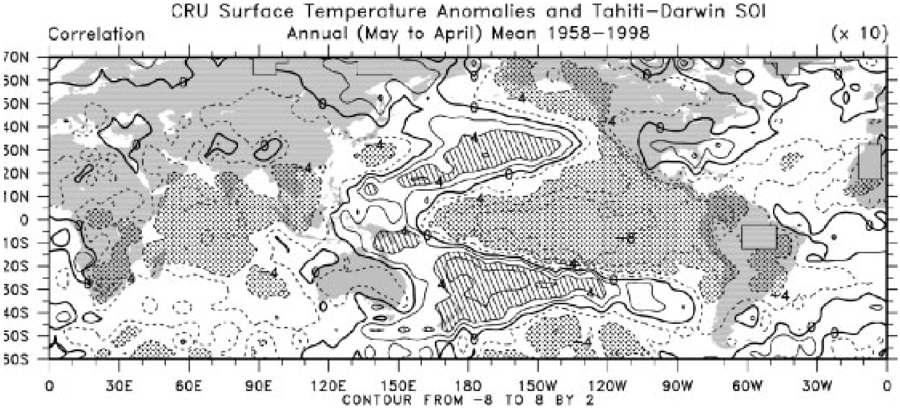

CRU Surface Temperature Anomalies and Tahiti–Darwin SOI
Annual (May to April) Mean 1958–1998
Correlation
(x 10)
70N
60N
50N
40N
30N
20N
10N
0
10S
20S
30S
40S
50S
60S
0
30E
60E
90E
120E
150E
180
150W
120W
90W
60W
30W
0
CONTOUR FROM –8 TO 8 BY 2

Figure 1 (see color insert) Correlation coefficients of the SOI with sea surface temperature seasonal anomalies for January 1958 to December 1998. It can be interpreted as the sea surface temperature patterns that accompany a La Niña event, or as an El Niño event with signs reversed. Values in the central tropical Pacific correspond to anomalies of about 1.5°C [From Trenberth and Caron, 2000]. See ftp site for color image.

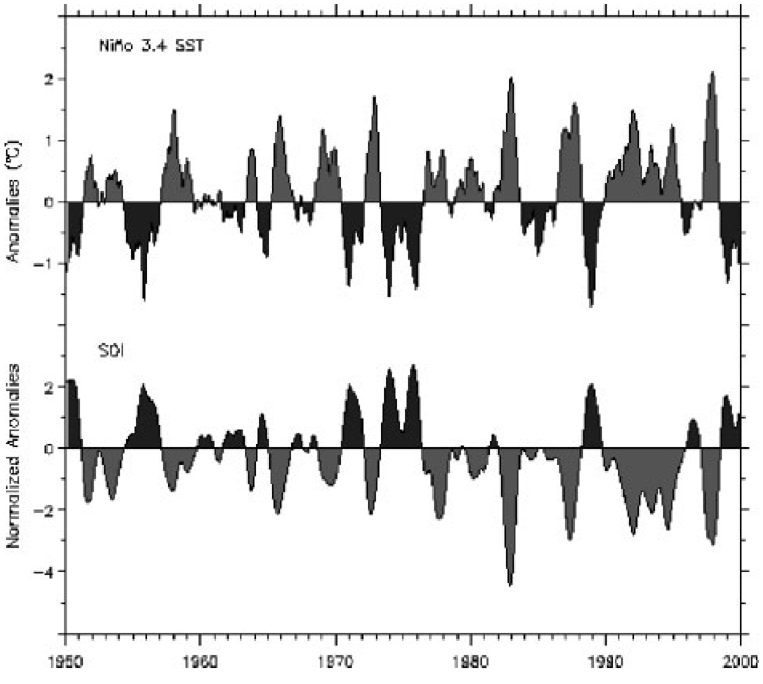


Figure 2 (see color insert) Time series of areas of sea surface temperature anomalies from 1950 through 2000 relative to the means of 1950–79 for the region most involved in ENSO 5° N–5° S, 170° E –120° W (top) and for the Southern Oscillation Index. El Niño events are in grey and La Niña events in black. See ftp site for color image.

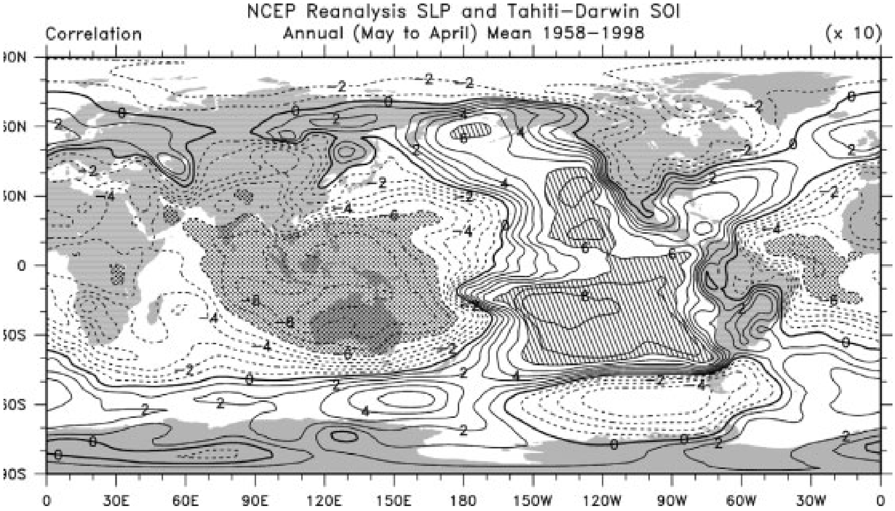

NCEP Reanalysis SLP and Tahiti–Darwin SOI
Correlation
Annual (May to April) Mean 1958–1998
(x 10)
90N
60N
30N
0
30S
60S
90S
0
30E
60E
90E
120E
150E
180
150W
120W
90W
60W
30W
0

Figure 3 (see color insert) Map of correlation coefficients (×10) of the annual mean sea level pressures (based on the year May to April) with the SOI showing the Southern Oscillation pattern in the phase corresponding to La Niña events. During El Niño events the sign is reversed [From Trenberth and Caron, 2000]. See ftp site for color image.

with less rainfall, while lower than normal pressures are identified with "bad" weather, more storminess, and rainfall. So it is with the SO. Thus for El Niño conditions, higher than normal pressures over Australia, Indonesia, southeast Asia, and the Philippines signal drier conditions or even droughts. Dry conditions also prevail at Hawaii, parts of Africa, and extend to the northeast part of Brazil and Colombia. On the other end of the seesaw, excessive rains prevail over the central and eastern Pacific, along the west coast of South America, parts of South America near Uruguay, and southern parts of the United States in winter (Fig. 4). In the winters of 1992–1993, 1994–1995, and 1997–1998 this included excessive rains in southern California, but in other years (e.g., 1986–1987 and 1987–1988 winters) California is more at risk for droughts. Because of the enhanced activity in the central Pacific and the changes in atmospheric circulation throughout the tropics, there is a decrease in the number of tropical storms and hurricanes in the tropical Atlantic during El Niño.

The SO has global impacts; however, the connections to higher latitudes (known as teleconnections) tend to be strongest in the winter of each hemisphere and feature alternating sequences of high and low pressures accompanied by distinctive wave patterns in the jet stream and storm tracks in midlatitudes. For California, all of the winter seasons noted above were influenced by El Niños, but their character differed and the teleconnections to higher latitudes were not identical. Although warming is

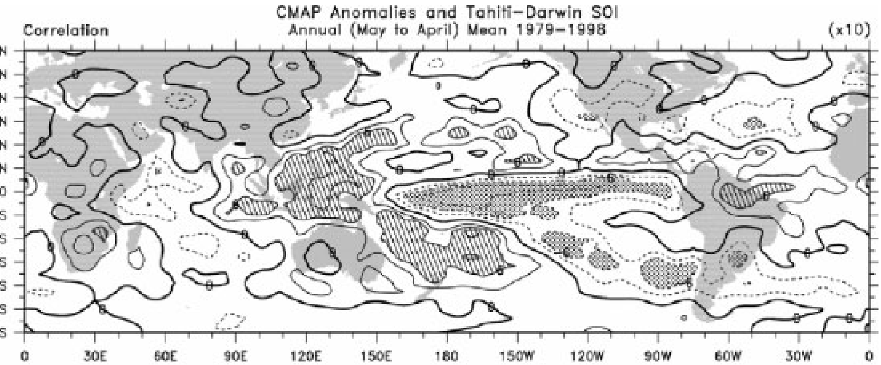


Figure 4 Correlations of annual mean precipitation with the SOI for 1979–1998. The pattern has the phase corresponding to La Niña events, and during El Niño events the sign is reversed, i.e., the stippled areas are wetter and hatched areas drier in El Niño events [From Trenberth and Caron, 2000]. See ftp site for color image.

generally associated with El Niño events in the Pacific and extends, for instance, into western Canada, cool conditions typically prevail over the North and South Pacific Oceans. To a first approximation, reverse patterns occur during the opposite La Niña phase of the phenomenon. However, the latter is more like an extreme case of the normal pattern with a cold tongue of water along the equator.

The prominence of the SO has varied throughout this century. Very slow long-term (decadal) variations are present; for instance, SOI values are mostly below the long-term mean after 1976. This accompanies the generally above normal sea surface temperatures in the western Pacific along the equator (Fig. 2). The prolonged warm ENSO event from 1990–1995 is very unusual and the 1997–1998 event is the biggest on record in terms of sea surface temperature anomalies. The decadal atmospheric and oceanic variations are even more pronounced in the North Pacific and across North America than in the tropics and also present in the South Pacific, with evidence suggesting they are at least in part forced from the tropics. It is possible that climate change associated with increasing greenhouse gases in the atmosphere, which contribute to global warming, may be changing ENSO, perhaps by expanding the west Pacific warm pool.

Changes associated with ENSO produce large variations in weather and climate around the world from year to year, and often these have a profound impact on humanity and society because of droughts, floods, heat waves, and other changes that can severely disrupt agriculture, fisheries, the environment, health, energy demand, and air quality and also change the risks of fire. Changes in oceanic conditions can have disastrous consequences for fish and sea birds and thus for the fishing and guano industries along the South American coast. Other marine creatures may benefit so that unexpected harvests of shrimp or scallops occur in some places. Rainfalls over Peru and Ecuador can transform barren desert into lush growth and benefit some crops but can also be accompanied by swarms of grasshoppers and increases in the populations of toads and insects. Human health is affected by mosquito-borne diseases such as malaria, dengue, and viral encephalitis and by water-borne diseases such as cholera. Economic impacts can be large, with losses typically overshadowing gains. This is also true in La Niña events and arises because of the extremes in flooding, drought, fires, and so on that are outside the normal range.

4 MECHANISMS OF ENSO

During El Niño, the trade winds weaken, causing the thermocline to become shallower in the west and deeper in the eastern tropical Pacific (Fig. 5), while sea level falls in the west and rises in the east by as much as 25 cm as warm waters surge eastward along the equator. Equatorial upwelling decreases or ceases and so the cold tongue weakens or disappears and the nutrients for the food chain are substantially reduced. The resulting increase in sea temperatures warms and moistens the overlying air so that convection breaks out and the convergence zones and associated rainfall move to a new location with a resulting change in the atmospheric circulation. A further weakening of the surface trade winds completes the positive feedback cycle leading to an ENSO event. The shift in the location of the organized rainfall in the tropics and the latent heat released alters the heating patterns of the atmosphere. Somewhat like a rock in a stream of water, the anomalous heating sets up waves in the atmosphere that extend into midlatitudes, altering the winds and changing the jet stream and storm tracks.

Although the El Niños and La Niñas are often referred to as "events" that last a year or so, they are somewhat oscillatory in nature. The ocean is a source of moisture, and its enormous heat capacity acts as the flywheel that drives the system through its memory of the past, resulting in an essentially self-sustained sequence in which the ocean is never in equilibrium with the atmosphere. The amount of warm water in the tropics builds up prior to and is then depleted during ENSO. During the cold phase with relatively clear skies, solar radiation heats up the tropical Pacific Ocean, the heat is redistributed by currents, with most of it being stored in the deep warm pool in the west or off the equator. During El Niño heat is transported out of the tropics within the ocean toward higher latitudes in response to the changing currents, and increased heat is released into the atmosphere mainly in the form of increased evaporation, thereby cooling the ocean. Added rainfall contributes to a general warming of the global atmosphere that peaks a few months after a strong El Niño event. It has therefore been suggested that the time scale of ENSO is determined by the time required for an accumulation of warm water in the tropics to essentially recharge the system, plus the time for the El Niño itself to evolve. Thus a major part of the onset and evolution of the events is determined by the history of what has occurred 1 to 2 years previously. This also means that the future evolution is predictable for several seasons in advance.

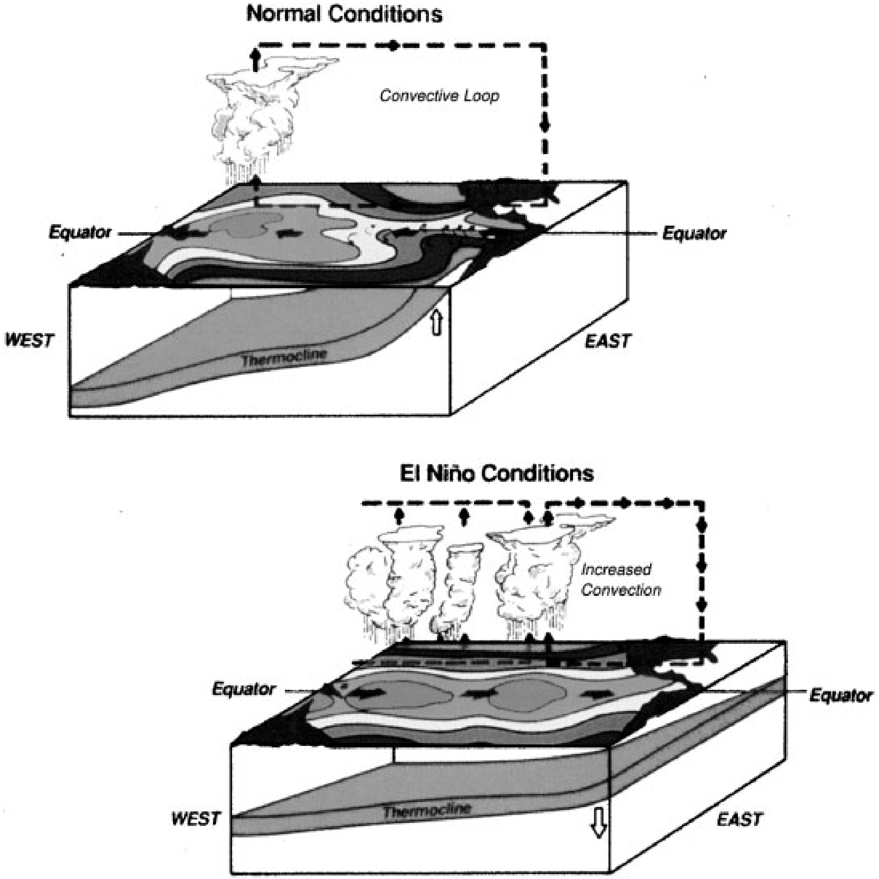


Figure 5 (see color insert) Schematic cross section of the Pacific Basin with Australia at lower left and the Americas at right depicting normal and El Niño conditions. Total sea surface temperatures exceeding 29°C are in gold and the colors change every 1°C. Regions of convection and overturning in the atmosphere are indicated. The thermocline in the ocean is shown in blue. Changes in ocean currents are shown by the black arrows. (Copyright University Corporation for Atmospheric Research. Reprinted by permission). See ftp site for color image.

5 OBSERVING ENSO

In the early 1980s, the observing system for weather and climate revolved around requirements for weather forecasting, and there was little routine monitoring of ocean conditions. Oceanographic observations were mostly made on research cruises, and often these were not available to the community at large until months or years later. Satellite observations of the ocean surface and atmosphere provided some information on sea surface temperatures and clouds, but processing of all the relevant information for ENSO was not routinely available. This meant that in 1982–1983, during the largest El Niño seen to that point, the tropical Pacific was so poorly observed that the El Niño was not detected until it was well underway. A research program called Tropical Oceans–Global Atmosphere (TOGA) was begun in 1985 until 1994 to explore ENSO and to build a system to observe and perhaps predict it. The understanding of ENSO outlined above has developed largely as a result of the TOGA program and an observing system has been put in place.

The ENSO observing system has developed gradually and was fully in place at the end of 1994, so the benefits from it and experience with it are somewhat limited. A centerpiece of this observing system is an array of buoys in the tropical Pacific moored to the ocean bottom known as the TAO (Tropical Atmosphere–Ocean) array. The latter is maintained by a multinational group spearheaded in the United States by the National Oceanic and Atmospheric Administration's (NOAA's) Pacific Marine Environmental Laboratory (PMEL). This array measures the quantities believed to be most important for understanding and predicting ENSO. Each buoy has a series of temperature measurements on a sensor cable on the upper 500 m of the mooring, and on the buoy itself are sensors for surface wind, sea surface temperature (SST), surface air temperature, humidity, and a transmitter to a satellite. Some buoys also measure ocean currents down to 250 m depth. Observations are continually made, averaged into hourly values, and transmitted via satellite to centers around the world for prompt processing. Tapes of data are recovered when the buoys are serviced about every 6 months by ships.

Other key components of the ENSO observing system include surface drifting buoys, which have drogues attached so that the buoy drifts with the currents in the upper ocean and not the surface wind. In this way, displacements of the buoy provide measurements of the upper ocean currents. These buoys are also instrumented to measure sea surface temperatures and, outside of the tropics, surface pressure. These observations are also telemetered by satellite links for immediate use.

Observations are also taken from all kinds of ships, referred to as "volunteer observing ships." As well as making regular observations of all surface meteorological observations, some of these ships are recruited to do expendable bathythermograph (XBT) observations of the temperatures of the upper 400 m of the ocean along regular ship lines. Another valuable part of the observing system is the network of sea level stations. Changes in heat content in the ocean are reflected in changes in sea level. New measurements from satellite-borne altimeters are providing much more comprehensive views of how sea level changes during ENSO.

Considerable prospects exist for satellite-based remote sensing, including observations of SSTs, atmospheric winds, water vapor, precipitation, aerosol and cloud properties, ocean color, sea level, sea ice, snow cover, and land vegetation. Continuity of consistent calibrated observations from space continues to be an issue for climate monitoring because of the limited lifetimes of satellites (a few years); replacement satellites usually have somewhat different orbits and orbits decay with time. All the ship, buoy, and satellite observations are used together to provide analyses, for instance, of the surface and subsurface temperature structure.

6 ENSO AND SEASONAL PREDICTIONS

The main features of ENSO have been captured in models that predict the anomalies in sea surface temperatures. Lead times for predictions in the tropics of up to about a year have been shown to be practicable. It is already apparent that reliable prediction of tropical Pacific SST anomalies can lead to useful skill in forecasting rainfall anomalies in parts of the tropics, notably those areas featured in Figure 4. While there are certain common aspects to all ENSO events in the tropics, the effects at higher latitudes are more variable. One difficulty is the vigor of weather systems in the extratropics, which can override relatively modest ENSO influences. Nevertheless, systematic changes in the jet stream and storm tracks do occur on average, thereby allowing useful predictions to be made in some regions, although with some uncertainty inherent.

Skillful seasonal predictions of temperatures and rainfalls have the potential for huge benefits for society. However, because the predictability is somewhat limited, a major challenge is to utilize the uncertain forecast information in the best way possible throughout different sectors of society (e.g., crop production, forestry resources, fisheries, ecosystems, water resources, transportation, energy use). Considerations in using a forecast include how decisions are made and whether they can be affected by new information. The implication of an incorrect forecast if acted upon must be carefully considered and factored into decisions along with estimates of the value of taking actions. A database of historical analogs (of past societal behavior under similar circumstances) can be used to support decision making. While skillful predictions should positively impact the quality of life in many areas, conflicts among different users of information (such as hydroelectric power versus fish resources in water management) can make possibly useful actions the subject of considerable debate. The utility of a forecast may vary considerably according to whether the user is an individual versus a group or country. An individual may find great benefits if the rest of the community ignores the information, but if the whole community adjusts (e.g., by growing a different crop), then supply and market prices will change, and the strategy for the best crop yield may differ substantially from the strategy for the best monetary return. On the other hand, the individual may be more prone to small-scale vagaries in weather that are not predictable. Vulnerability of individuals will also vary according to the diversity of the operation, whether there is irrigation available, whether the farmer has insurance, and whether he or she can work off the farm to help out in times of adversity.

ACKNOWLEDGMENTS

This research was partially sponsored by NOAA Office of Global Programs grant NA56GP0247 and the joint NOAA/NASA grant NA87GP0105.

BIBLIOGRAPHY

Chen, D., S. E. Zebiak, A. J. Busalacchi, and M. A. Cane (1995). An improved procedure for El Nino forecasting: Implications for predictability, *Science* **269**, 1699–1702.

Glantz, M. H. (1996). *Currents of Change. El Nino's Impact on Climate and Society*, Cambridge University Press, Cambridge, New York.

Glantz, M. H., R. W. Katz, and N. Nicholls (Ed.) (1991). *Teleconnections Linking Worldwide Climate Anomalies*, Cambridge University Press, Cambridge, New York.

Kumar, A., A. Leetmaa, and M. Ji (1994). Simulations of atmospheric variability induced by sea surface temperatures and implications for global warming, *Science* **266**, 632–634.

McPhaden, M. J. (1995). The Tropical Atmosphere–Ocean Array is completed, *Bull. Am. Meteor. Soc.* **76**, 739–741.

Monastersky, R. (1993). The long view of the weather, *Science News* **144**, 328–330.

National Research Council (1996). *Learning to Predict Climate Variations Associated with El Nino and the Southern Oscillation: Accomplishments and Legacies of the TOGA Program*, National Academy Press, Washington D.C.

Patz, J. A., P. R. Epstein, T. A. Burke, and J. M. Balbus (1996). Global climate change and emerging infectious diseases, *J. Am. Med. Assoc.*, **275**, 217–223.

Philander, S. G. H. (1990). *El Nino, La Nina, and the Southern Oscillation*, Academic Press, San Diego, CA.

Suplee, C. (1999). El Nino/La Nina, *National Geographic*, March, pp. 72–95.

Time (1983). Australia's "Great Dry," Cover story, March 28, pp. 6–13.

Trenberth, K. E. (1997a). Short-term climate variations: Recent accomplishments and issues for future progress, *Bull. Am. Met. Soc.* **78**, 1081–1096.

Trenberth, K. E. (1997b). The definition of El Nino, *Bull. Am. Met. Soc.* **78**, 2771–2777.

Trenberth, K. E. (1999). The extreme weather events of 1997 and 1998, *Consequences* **5**(1), 2–15, (http://www.gcrio.org/CONSEQUENCES/vol5no1/extreme.html).

Trenberth, K. E., and J. M. Caron (2000). The Southern Oscillation revisited: Sea level pressures, surface temperatures and precipitation, *J. Climate*, **13**, 4358–4365.

SECTION 3

PHYSICAL METEOROLOGY

Contributing Editor: Gregory Tripoli

CHAPTER 15

PHYSICAL ATMOSPHERIC SCIENCE*

GREGORY TRIPOLI

1 OVERVIEW

Perhaps the oldest of the atmospheric sciences, physical meteorology is the study of laws and processes governing the physical changes of the atmospheric state. The underlying basis of physical meteorology is the study of *atmospheric thermodynamics*, which describes relationships between observed physical properties of air including pressure, temperature, humidity, and water phase. Generally, thermodynamics as well as most of physical meteorology does not seek to develop explicit models of the processes it studies, such as the evolution of individual air molecules or individual water droplets. Instead, it formulates relationships governing the statistics of the microscale state and processes. Temperature and the air density are two examples of such statistics representing the mean kinetic energy of a population of molecules and the mass of molecules of air per unit volume, respectively.

A second major branch of physical meteorology is the study of the growth of water and ice precipitation called *cloud microphysics*. Here theories of how liquid and ice hydrometeors first form and subsequently evolve into rain, snow, and hail are sought. The interaction of water with atmospheric aerosols is an important part of atmospheric microphysics, which leads to the initial formation of hydrometeors. The quantification of these theories leads to governing equations that can be used to simulate and predict the evolution of precipitation as a function of changing atmospheric conditions. By studying the processes that lead to the formation of clouds,

*The derivations, concepts, and examples herein are based partially on course notes from the late Professor Myron Corrin of Colorado State University. Other derivations are taken from the Isbane and Godson text, the Emanuel (1994: *Atmospheric Convection*, Oxford University Press) text, and the Dutton (*The Ceaseless Wind*) book.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

clouds themselves must be categorized by the processes that form them, giving rise to the study of *cloud types*. From the discipline of microphysics emerges the study of *weather modification*, which studies the possibility of purposefully modifying natural precipitation formation processes to alter the precipitation falling to the surface. Another discipline associated with microphysics is *atmospheric electricity*, which studies the transfer of electrical energy by microphysical processes.

A third major branch of physical meteorology is the study of *atmospheric radiation*. This area of study develops theories and laws governing the transfer of energy through the atmosphere by radiative processes. These include the adsorbtion, transmission, and scattering of both solar radiation and terrestrial radiation. A primary goal of atmospheric radiation studies is to determine the net radiative loss or gain at a particular atmospheric location that leads to thermodynamic change described by thermodynamics theory. Since radiative transfer is affected by details of the atmospheric thermal, humidity, and chemical structure, it is possible to recover some details of those structures by observing the radiation being transferred. This has given rise to a major branch of atmospheric radiation called *remote sensing*, the goal of which is to translate observations of radiation to atmospheric structure. The study of *atmospheric optics* is another branch of atmospheric radiation that applies concepts of radiative transfer to study the effect of atmospheric structure on visible radiation passing through the atmosphere. Phenomena such as blue sky, rainbows, sun-dogs, and so on are subjects of this area.

Finally, *boundary layer* meteorology studies the transfer of heat, radiation, moisture, and momentum between the surface and the atmosphere. The transfer of energy occurs on a wide range of scales from the molecular scale all the way to the scales of thermal circulations. The challenge is to quantify these transfers on the basis of observable or modelable atmospheric quantities. This also gives rise to the need to represent the evolution of the soil, vegetation, and water surfaces. This has given rise to branches of atmospheric and other earth sciences aimed at developing models of the soil, water, and vegetation.

In this chapter we will concentrate our efforts on laying down the foundations of basic thermodynamics, microphysics, and radiative transfer and surface layer theory.

2 ATMOSPHERIC THERMODYNAMICS

Atmospheric thermodynamics is the study of the macroscopic physical properties of the atmosphere for which temperature is an important variable. The thermodynamic behavior of a gas, such as air, results from the collective effects and interactions of the many molecules of which it is composed. However, because these explicit microscale processes are too small, rapid, and numerous to be observable and too chaotic to be predictable, their effects are not described by the Newtonian laws governing the molecular processes. Instead, classical thermodynamics consists of laws governing the observed macroscopic statistics of the behavior of the microscopic system of air and the suspended liquids and solids within. As a consequence, thermodynamic laws are only valid on the macroscale where a sufficient population of molecular entities is present to justify the statistical approach to their behavior.

Classical atmospheric thermodynamics seeks to:

- Develop an understanding for a statistically significant mass of air called a *parcel.*
- Classify the interacting forms of energy contained within a parcel.
- Never ask "why" in terms of the first principles of atomic and molecular structure and the energetics of interactions between atoms and molecules.
- Derive all principles from *observables*. For example, the first law of thermodynamics begins with a classification of energy into thermal and nonthermal forms as is observed.
- Create an efficient *book-keeping* system in which the interacting forms of energy and mass are defined in a convenient and nonredundant form.
- Quantify the amount and rate of energy transfer from one form to another.
- Provide a means of calculating a true equilibrium state among these energies as well as defining the conditions for constrained equilibrium.

The atmospheric application of classical thermodynamics differs in that particular attention is paid to the properties of air and forms of the thermodynamic laws and relationships that are most useful for application to the atmosphere. We define *air* to be the gas present in Earth's lower atmosphere where air is sufficiently dense for the empirical laws to apply. This only excludes regions on the order of 50 km and higher above the surface.

This section will begin by defining the basic concepts needed, then move into the thermodynamics of gasses and then specifically into the thermodynamics of air. Finally the thermodynamics of the mixed-phase system, including liquid and ice processes, will be described in Chap. 16.

Basic Concepts, Definitions, and Systems of Units

We begin by defining a *system* to represent a portion of the universe selected for study. It may or may not contain matter or energy. Systems are classified as *open*, *isolated*, or *closed*. In contrast, *surroundings* refer to the remainder of the universe outside the system. An *isolated system* can exchange neither matter nor energy with its surroundings. The universe is defined by the first law as an isolated system! A *closed system* can exchange energy but not matter with its surroundings. An *open system*, on the other hand, can exchange both energy and matter with its surroundings.

Next we define a *property* of the system to be an observable. It results from a physical or chemical measurement. A system is therefore characterized by a set of properties. Typical examples are mass, volume, temperature, pressure, composition, and energy. Properties may be classified as intensive or extensive. An *intensive property* is a characteristic of the system as a whole and not given as the sum of these properties for portions of the system. Pressure, temperature, density, and specific heat are intensive properties. We will adopt a naming convention that uses capital letters for variables describing an "intensive" property. An *extensive property* is dependent on the dimensions of the system and is given by the sum of the extensive properties of each portion of the system. Examples are mass, volume, and energy. *Extensive variables will be named with lowercase letters.*

A homogeneous system is one in which each intensive variable has the same value at every point in the system. Then any extensive property z can be represented by the mass-weighted intensive properties:

$$z = mz \tag{1}$$

where z is the specific property.

We define *phases* to be subsets of a system that are homogeneous in themselves but are physically different from other portions. A *heterogeneous system* is defined to be a system composed of two or more phases. In this case, any extensive property will be the sum of the mass-weighted phases of the system, i.e.,

$$Z = \sum_{\alpha} m_{\alpha} z_{\alpha} \tag{2}$$

where α refers to the particular phase.

An *inhomogeneous system* is one in which intensive properties change in a continuous way, such as how pressure or temperature change from place to place in the atmosphere. Our thermodynamic theory will be applied only locally where the system can be considered homogeneous or heterogeneous as an approximation.

The study of the atmosphere requires the definition of an *air parcel.* This is a small quantity of air whose mass is constant but whose volume may change. The parcel should be considered to be small enough to be considered infinitesimal and homogeneous but large enough for the definitions of thermodynamic and microphysical statistics to be valid. In an atmosphere, where air movements are not confined to a rigid container and therefore expand and contract against local forces of pressure gradient, the parcel quantification is a natural choice for defining an air sample.

Energy is the *ability to perform work* and is a *property* defined in the first law of thermodynamics. Energy is an extensive variable that is normally defined as a product of an intensive and extensive variable. The intensive term defines, by differencing, the direction of the energy transfer. The extensive term is also called the capacity term. An example is mechanical energy in the air, which is written as a product of a pressure (intensive) and volume change (extensive). When other intensive variables are substituted, other forms of energy are represented. For instance, replace pressure in the example above with temperature and the energy is thermal energy; replace it with chemical potential and one defines chemical energy and so on.

The *state* of a system is defined by a set of system properties. A certain minimum number of such properties are required to specify the state. For instance, in an ideal gas, any three of the four properties—pressure, volume, number of moles, and temperature—will define the state of the system. The so-called *equation of state* states an empirical relationship between the state properties that define the system.

If the state of a system remains constant with time if not forced to change by an outside force, the system is said to be in a state of *equilibrium*; otherwise we say the system is in a state of *nonequilibrium*. Independence of the state to time is a necessary but not sufficient condition for equilibrium. For instance, the air within an evaporating convective downdraft may have a constant cool temperature due to the evaporation rate equaling a rate of warming by convective transport. The system is not in thermodynamic equilibrium but is still not varying in time. In that case we say the system is *steady state*.

When there is an equilibrium that if slightly perturbed in any way will accelerate away from equilibrium, it is referred to an *unstable equilibrium*. *Metastable equilibrium* is similar to unstable equilibrium except with respect to perhaps only a single process. For instance, when the atmosphere becomes supersaturated over a plain surface of ice, no ice will grow unless there is a crystal on which to grow. Introduce that crystal and growth will begin, but without it the system remains time independent even to other perturbations such as a temperature or pressure perturbation.

A *process* is the description of the manner by which a change of state occurs. Here are some process definitions:

- *Change to the System* Any difference produced in a system's state. Therefore any change is defined only by the initial and final states.
- *Isothermal Process* A change in state occurring at constant temperature.
- *Adiabatic Process* A change of state occurring without the transfer of thermal energy or mass between the system and its surroundings.
- *Diabatic Process* A process resulting from the exchange of mass or energy with the surroundings. An example is radiative cooling.
- *Cyclic Process* A change occurring when the system (although not necessarily its surroundings) is returned to its initial state.
- *Reversible or Quasi-Static Process* A special, idealized thermodynamic process that can be reversed without change to the universe. It will be shown that this is possible only when the conditions differ infinitesimally from equilibrium. Hence at any given point during a reversible process the system is nearly in equilibrium. One can think of a reversible process as being composed of small irreversible steps, each of which have only a small departure from equilibrium (see Fig. 1).

In association with these concepts of processes, we also define the *internal derivative* (d_i) to be the change in a system due to a purely adiabatic process. Conversely, the *external derivative* (d_e) is defined to be the change in a system due to a purely diabatic process.

We can say that the total change of a system is the sum of its internal and external derivatives, i.e.,

$$d = d_i + d_e \tag{3}$$

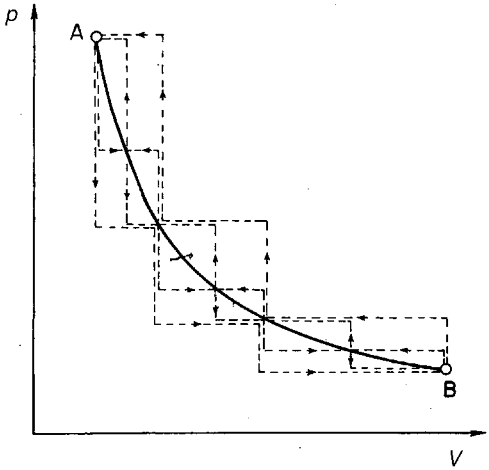


Figure 1 Reversible process as a limit of irreversible processes (from Iribarne and Godson, 1973).

where d is the total derivative. Now, it follows from the definition of adiabatic that

$$\oint d = \oint d_e$$

$$\oint d_i = 0$$

Zeroth Principle of Thermodynamics: Definition of Temperature (*T*)

We define a *diathermic wall* as one that allows thermal interaction of a system with surroundings but not mass exchange. The *zeroth principle of thermodynamics* states: *If some system A is in thermal equilibrium with another system B separated by a diathermic wall, and if A is in thermal equilibrium with a third system C, then systems B and C are also in thermal equilibrium.* It follows that all bodies in equilibrium with some reference body will have a common property that describes this thermal equilibrium and that property is defined to be *temperature*. The so-called *reference body* defines this property and is called a *thermometer*. A number is assigned to the property by creating a temperature scale. This is accomplished by finding a substance that has a property that changes in correspondence with different thermal states. For instance, Table 1 defines five thermometric substances and corresponding properties.

A thermometer is calibrated to changes in its observed thermometric property, obeying the zeroth principle. The thermometer must be much smaller than the substance measured so that the changes brought to the substance by exchanges from the thermometer can be neglected.

A linear temperature scale may be defined generally as:

$$T_X = aX + b \tag{4}$$

where T_X is an arbitrary temperature scale, a represents the slope, b is a constant representing the intercept of the linear scale, and X is the thermometric property of the substance. Use of this scale requires two well-defined thermal states and an approximate linear relationship of T to changes in the substance's thermometric property in order to determine a and b. The Celsius, or Centigrade, temperature scale is defined in this way by the freezing point of pure water at one atmosphere to be 0°C and the boiling point of pure water at one atmosphere to be 100°C. The Fahrenheit temperature scale, used primarily in the United States for nonscientific measurements and unofficially in the United Kingdom, is calibrated by assigning the boiling point of pure water at one atmosphere to be 212°F and the freezing point of pure water at one atmosphere to be 32°F.

A more general scale can be derived using the thermometric properties of gasses listed as the first two examples in Table 1. Using gas as the thermometric substance and pressure as the thermometric property, it is observed that pressure decreases as the thermal state of the gas falls. We then can define the lowest possible thermodynamic state to be that where pressure decreases to zero (at constant volume). Letting the temperature be zero at that point, defines the intercept to also be zero. Now only a single well-defined thermodynamic state is necessary to define the scale. This is chosen to be the triple point for pure water where all three phases of water can coexist. The actual value of this "absolute" temperature scale at the triple point is arbitrary and its choice defines the value of the slope a. We can define this scale to be such that the thermal increments for one degree of temperature equal that of the Celsius scale. We do this by solving for the intercept of the Celsius scale, using pressure as the thermometric property. This yields $b = -273.15$°C. Then defining the freezing point at 1 atm to be 273.15° and the triple point of water to be 273.16° of the absolute temperature, the scale is set to have equal increments to Celsius temperature but have an absolute scale beginning at absolute zero. This scale is called the *Kelvin* scale. The Kelvin temperature is detonated by K. Similarly, an absolute temperature scale is also defined for Fahrenheit units and is called the

TABLE 1 Empirical Scales of Temperature

Thermometric Substance	Thermometric Property X
Gas, at constant volume	Pressure
Gas, at constant pressure	Specific Volume
Thermocouple, at constant pressure and tension	Electromotive force
Pt wire, at constant pressure and tension	Electrical resistance
Hg, at constant pressure	Specific Volume

Source: Irabarne and Godson (1973).

Rankine scale having units of degrees Rankine (R). Generally, atmospheric scientists work with the Celsius and Kelvin scales.

The relationships between the four major scales are

$$\begin{aligned} T &= T_C + 273.15 \\ T_R &= T_F + 459.67 \\ T_F &= \frac{5}{9} T_C + 32 \end{aligned}$$

where T, T_C, T_F, and T_R are the Kelvin, Celsius and Fahrenheit and Rankine temperatures, respectively.

Ideal Gas and Equation of State for the Atmosphere

A gas is composed of molecules moving randomly about, spinning, vibrating and occasionally colliding with each other. If held within a container, the gas molecules will also collide with the container wall, exerting a force on the walls. The nature of a molecule's movement and the movement of the entire population of molecules is chaotic and unpredictable on a time scale of only a fraction of a second. We make no attempt to quantify a physical description of this process with Newtonian mechanics.

The statistical properties of a gas that are measured include its volume, its temperature, and its pressure. The temperature is a measure of the average kinetic energy of the molecules, which is proportional to the average of the square of the speed of the molecules. The pressure is proportional to force per unit area exerted on a plane surface resulting from the collisions of gas molecules with that surface. Newtonian mechanics tells us that force will be proportional to the average change in momentum of the molecules as they reflect off the surface, which is then proportional to the mass and normal velocity component of each molecule and the total number of molecules hitting the surface.

The mass of the gas is defined to be

$$m \equiv nM \tag{5}$$

where n is the number of moles and M is the molecular weight (in grams per mole). It is often advantageous to express an extensive thermodynamic quantity as the ratio of that quantity to its total mass. That ratio is referred to as the *specific* value of that quantity. For instance, the *specific volume* is given by:

$$\alpha \equiv \frac{V}{m} \equiv \frac{1}{\rho} \tag{6}$$

where V is the volume and ρ is the *density* K/g^3.

We can now define the *equation of state* for the atmosphere. Observations suggest that under normal tropospheric pressures and temperatures, air gases exhibit nearly identical behavior to an ideal gas. To quantify this observation, we define an *ideal* *gas* to be one that obeys the empirical relationship; also called the *ideal gas law* or the *equation of state*.

$$pV = nR^*/T = mR^*T/M = mRT \tag{7}$$

where n is the number of moles, M is the molecular weight of the gas, R^* is the universal gas constant, and $R = R^*/M$ is the *specific gas constant* valid only for a particular gas of molecular weight M. Observations show that, $R^* = 8.3143$ J/mol/K and is the same for all gases having pressures and temperatures typical of tropospheric conditions.

Employing these definitions, the equation of state for an ideal gas can be written as:

$$p = \frac{RT}{\alpha} = \rho RT \tag{8}$$

It may be shown, by comparison with measurements, that for essentially all tropospheric conditions, the ideal gas law is valid within 0.1%. Under stratospheric conditions the departure from ideal behavior may be somewhat higher.

We now consider multiple constituent gases such as air. In such a case, the partial pressure (p_j) is defined to be the pressure that the jth individual gas constituent would have if the same gas at the same temperature occupied an identical volume by itself. We can also define the *partial Volume* (V_j) to be the volume that a gas of the same mass and pressure as the air parcel would occupy by itself.

Dalton's law of partial pressure states that the total pressure is equal to the sum of the partial pressures:

$$p = \sum_j p_j \tag{9}$$

where p is the total pressure and the sum is over all components (j) of the mixture. Since each partial pressure independently obeys the ideal gas law, volume can be written

$$V = \sum V_j \tag{10}$$

where V is the total volume. Also, it follows that

$$P_j = n_j R^*T/V \tag{11}$$

and

$$p = (R^*T/V)\sum_j n_j \tag{12}$$

If we divide through by the total mass of the gas (m_j) and note that $n_j = m_j/M_j$, where M_j is the molecular weight of the species i, and total mass is

$$m = \sum m_j \tag{13}$$

The statement of the ideal gas law (equation of state) for a mixture of gases is then

$$p = \rho \frac{R^*}{\bar{M}} T = \rho \bar{R} T \tag{14}$$

where $\bar{M}$ is the mean molecular weight of the mixture, which is given by:

$$\bar{M} \equiv \frac{\sum m_j}{\sum \frac{m_j}{M_j}} = \sum M_j N_j \tag{15}$$

where N_j is the molar fraction of a particular gas given by $N_j = p_j/p = V_j/V$.

Composition of Air The current atmosphere of Earth is composed of "air," which we find to contain:

1. *Dry air*, which is a mixture of gases described below
2. *Water*, which can be in any of the three states of liquid, solid, or vapor
3. *Aerosols*, which are solid or liquid particles of small sizes

The chemical composition of dry air is given in Table 2.

Water vapor and the liquid and solid forms of water vary in its volume fraction from 0% in the upper atmosphere to as high as 3 to 4% near the surface under humid conditions. Because of this variation, dry air is treated separately from vapor in thermodynamic theory.

Among the trace constituents are carbon dioxide, ozone, chlorofluorocarbons (CFCs), and methane, which, despite their small amounts, have a very large impact on the atmosphere because of their interaction with terrestrial radiation passing through the atmosphere, or in the case of CFCs because of their impact on the ozone formation process.

Note from Table 2 that the molecular weighted average gas constant for dry air is $R_d = \bar{R} = 287.05\,\mathrm{J/kg/K}$, which we also call the *dry air gas constant.*

Work by Expansion

If a system is not in mechanical equilibrium with its surroundings, it will expand or contract. Assume that λ is the surface to the system that expands infinitesimally to λ'

TABLE 2 Composition of Dry Air near the Earth's Surface

Gas	Symbol	Molecular Weight	Molar (or Volume) Fraction	Mass Fraction	Specific Gas Constant (J/kg K)	m_jR_j/m ($\bar{R}$) (J/kg K)
Nitrogen	N_2	28.013	0.7809	0.7552	296.80	224.15
Oxygen	O_2	31.999	0.2095	0.2315	259.83	60.15
Argon	Ar	39.948	0.0093	0.0128	208.13	2.66
Carbon Dioxide	CO_2	44.010	0.0003	0.0005	188.92	0.09
Helium	He		0.0005			
Methane	CH_4		0.00017			
Hydrogen	H		0.00006			
Nitrous Oxide	N_2O		0.00003			
Carbon Monoxide	CO		0.00002			
Neon	Ne		0.000018			
Xenon	X		0.000009			
Ozone	O_3		0.000004			
Krypton	Kr		0.000001			
Sulfur Dioxide	SO_2		0.000001			
Nitrogen Dioxide	NO_2		0.000001			
Chlorofluorocarbons	CFC		0.00000001			
Total			1.0000	1.0000		$\bar{R} = 287.05$

in the direction ds (see Fig. 2). Then the surface element $d\sigma$ has performed work against the external pressure P. The work performed is

$$(dW)_{d\sigma} = P\, d\sigma\, d\lambda \cos\phi = P_{\text{surr}}\, dV \tag{16}$$

where dV is the change in volume.

For a finite expansion, then

$$W = \int_i^f p\, dV \tag{17}$$

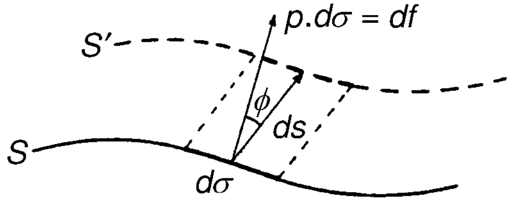


Figure 2 Work of expansion (from Iribarne and Godson, 1973).

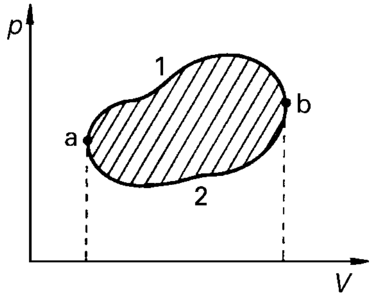


Figure 3 Work of expansion in a cycle (from Iribarne and Godson, 1973).

where *i*, *f* stand for the initial and final states. Since the integrand is not a total derivative, total work over a cyclic process can be nonzero. Therefore

$$W = \oint p\,dV = \left(\int_a^b P\,dV\right)_1 - \left(\int_a^b P\,dV\right)_2 \tag{18}$$

which is defined to be positive if integrated in the clockwise sense. This is the area enclosed by the trajectory in the graph given in Figure 3.

The work by expansion is the only kind of work that we shall consider in our atmospheric systems. Assuming the pressure of the system is homogeneous and in equilibrium with the surroundings, we can adopt the notation

$$dW = P\,dV \tag{19}$$

where W is the work performed on the surroundings by the system.

First Law (or Principle) of Thermodynamics

The *first law of thermodynamics* can be simply stated: *The energy of the universe is constant.* Let us consider the general concept of just what is meant by *energy*. We know that if we apply force to an object over some distance and as a result we accelerate the object, we will have performed *work* on the object, measured in the energy units of *force* × *distance* = *joules*. That work will precisely equal the increase in kinetic energy of the object, the total of which is measured by half its mass multiplied by the square of its new velocity. Moreover, we understand that our effort expends energy from the *working substance*. Depending on the method used to accelerate the object, the energy was previously stored in some other form; perhaps the "potential energy" of a compressed spring, "chemical energy" stored in the muscle cells of one's arm, or perhaps the "kinetic energy" of another object that strikes the object that was accelerated. There are obviously many possibilities for working substances featuring different types of stored energy. Energy, in a general sense, quantifies a potential to apply a force and so perform work.

The first law requires that for any thermodynamic system an energy budget must exist that requires any net energy flowing into a system to be accounted for by changes in the *external energy* of the system or the *internal energy* stored within the system. The external energy of the system includes its kinetic energy of motion and its energy of position, relative to forces outside the system such as gravitational, electrical, magnetic, chemical, and many other forms of potential energy between the system and its surroundings. The same energies can also exist within the system between the molecules, atoms, and subatomic particles composing the system and are termed internal energies. Kinetic energy of molecular motions relative to the movement of the center of mass of the system is measured by the system's temperature and represent the *thermal internal energy.* The other internal energies are *potential internal energies* and may include a potential energy against the intermolecular forces of attraction (*latent heat*), chemical potential energy (e.g., a gas composed of oxygen and hydrogen can potentially react), and so on.

Because we probably are not even aware of all of the forms of internal forces that exist in a system, we make no attempt to evaluate the total internal energy. Instead, for the purposes of classical thermodynamics, we need only consider changes in internal energy occurring during *allowable processes*. For instance, for atmospheric studies we choose to ignore nuclear reactions and even most chemical reactions. But we cannot neglect the changes in internal potential energy due to intermolecular forces of liquid and ice water phases in the atmosphere. The intermolecular forces are substantial, and the energy needed to overcome them is the latent heat of vaporization and melting. Thermodynamic energy transfers between thermal and these internal potential energies drive the general circulation of Earth's atmosphere! Our limited thermodynamic discussion will ignore the treatment of some internal energies such as surface energy of droplets, chemical energy in photochemical processes, electrical energy in thunderstorms and the upper atmosphere, chemical processes involving CFCs, and other processes known to have important secondary impacts on the evolution of the atmospheric thermodynamic system. These processes can be added to the system when needed, following the methods described in this chapter.

In most texts of classical thermodynamics, *closed* thermodynamic systems are assumed. A *closed* system allows no mass exchange between the parcel (system) and the surroundings. This approximation greatly simplifies the thermodynamic formulation. Later, we can still account for molecular transfer by representing it as a noninteracting and externally specified source of mass or energy rather than as an explicitly interacting component of the formulation.

One such open mass flux that must be accounted for with an open system is the mass flux of precipitation. We affix our parcel coordinate relative to the center of mass of the dry air, which is assumed to move identically to vapor. The liquid and ice components of the system, however, may attain a terminal velocity that allows it to flow into and out of the parcel diabatically. As a result, we will derive thermodynamic relationships for an *open system*, or one where external fluxes of mass into and out of the system are allowed. To retain some simplicity, however, we will make the assumption that external mass fluxes into and out of the system will be of constituents having the same state as those internal to the system. Hence we will allow rain to fall into our parcel, but it will be assumed to have the same temperature as the system. Although the effect of this assumption can be scaled to be small, there are instances where it can be important. One such instance is the case of frontal fog, where warm droplets falling into a cool air mass result in the formation of fog. These moist processes will be described in detail in the next chapter.

As demanded by the first law, we form an energy budget, first for a simple one-component system of an ideal gas:

$$\begin{aligned} \text{Energy exchanged with surroundings} &= \text{ change in energy stored} \\ Q - W &= \Delta U - \sum_k \Delta A_k \end{aligned} \tag{20}$$

where Q represents the flow of thermal energy into the system including radiative energy and energy transferred by conduction, W is the work performed on the surroundings by the system, U is the thermal internal energy, and A_k is the kth component of potential internal energy, summed over all of the many possible sources of internal energy. The work, as described in the previous section, is the work of expansion performed by the parcel on the surroundings. It applies only to our gas system and does not exist in the same form for our liquid and ice systems. In those systems, it should be reflected as a gravitational potential term; however, it is typically neglected.

In general, we treat the air parcel as a *closed* system, whereby there are no mass exchanges with the surroundings. This is generally a good approximation as molecular transfers across the boundaries can usually be neglected or included in other ways. One exception occurs and that is when we are considering liquid and ice hydrometeors. If we fix our parcel to the center of mass of the dry air, then there can be a substantial movement of liquid or ice mass into and out of the parcel. Clearly this process must be considered with an *open* thermodynamic system, where at least fluxes of liquid and ice are allowed with respect to the center of mass of the dry air parcel. Hence we will consider the possibility that A_k may change in part because of exchanges of mass between the parcel and the environment, particularly present with falling precipitation. Combining Eq. (20) with (19) for an infinitesimal process, we obtain the form of the first law:

$$\delta Q = dU + P\,dV - \sum_k dA_k \tag{21}$$

where it should be noted that P is the pressure of the surroundings, as $P\,dV$ defines work performed on the surroundings and $-dV$ is the change in volume of the surroundings. Note that for the adiabatic case, $\delta Q = 0$ by definition.

Because we require conservation of energy, it is evident that the change in both internal thermal energy (dU) and internal potential energy (dA_k) must be exact differentials, only dependent on the initial and final states of the process and not the path that the process takes. Hence *U and A_k must also be state variables.*

Heat Capacities *Heat capacity* is a property of a substance and is defined to be the rate at which the substance absorbs (loses) thermal energy (Q) compared to the rate at which its temperature (T) rises (falls) as a result. It is a property usually defined in units of dQ/dT (energy per temperature change) or in units of dq/dT (energy per mass per temperature change) in which case it is called specific heat capacity. If heat is passed into a system, it may be used to either increase the internal thermal or potential energy of the system or can be used to perform work. Hence there are an infinite number of possible values for heat capacity depending on the processes allowable. It is useful to determine the heat capacity for special cases where the allowable processes are restricted to only one. Hence we neglect the storage of potential internal energy and restrict the system to constant volume or constant pressure processes. Hence,

$$\begin{aligned} C_v &= \left(\frac{\delta Q}{dT}\right)_V & c_v &= \left(\frac{\delta q}{dT}\right)_\alpha \\ C_p &= \left(\frac{\delta Q}{dT}\right)_p & c_p &= \left(\frac{\delta q}{dT}\right)_p \end{aligned} \tag{22}$$

where C_v and C_p are the heat capacities at constant volume and pressure, respectively, and where c_v and c_p are the specific heat capacities at constant volume and pressure, respectively. It can be expected that $C_p > C_v$, since in the case of constant pressure, the parcel can use a portion of the energy to expand, and hence perform work on the surroundings, reducing the rise in temperature for a given addition of heat.

Since the first law requires that the change in internal energy, U, be an exact differential, U must also be a state variable; i.e., its value is not dependent upon path. Hence $U = u(P, \alpha, T)$. If we accept that air can be treated as an ideal gas, then the equation of state eliminates one of the variables, since the third becomes a function of the other two. Employing Euler's rule,

$$dU = \left(\frac{\partial U}{\partial V}\right)_T dV + \left(\frac{\partial U}{\partial T}\right)_V dT \tag{23}$$

Substituting Eq. (23) into the first law [Eq. (20)], we get

$$\delta Q = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V} + P\right)_T dV - \sum_k dA_k. \tag{24}$$

For a *constant volume process*,

$$C_v = \left(\frac{\delta Q}{dT}\right)_V = \left(\frac{\partial U}{\partial T}\right)_V \tag{25}$$

Hence, returning to Eq. (23),

$$dU = C_v\, dT + \left(\frac{\partial U}{\partial V}\right)_T dV. \tag{26}$$

It has been shown experimentally that $(\partial U/\partial V)_T = 0$, indicating that air behaves much as an ideal gas. Hence the internal energy of a gas is a function of temperature only provided that C_v is a function of temperature only. If we assume that possibility, $\partial C_v/\partial V = 0$. Experiments show that for an ideal gas the variation of C_v with temperature is small, and since we hold air to be approximately ideal, it follows that C_v is a constant for air. The first law [Eq. (24)] is therefore written as:

$$\delta q = C_v\, dT + P\, dV - \sum_k dA_k. \tag{27}$$

One can also obtain an alternate form of the first law by replacing dV with the ideal gas law and then combining q^{ith} Eq. (27) to yield

$$\delta Q = (C_v + R^*)\, dT - V\, dP - \sum_k dA_k. \tag{28}$$

For an isobaric process, ($dP = 0$), one finds

$$C_p = \left(\frac{\delta q}{\partial T}\right)_p = C_v + R^* \tag{29}$$

or

$$C_p - C_v = R^*. \tag{30}$$

Similarly for dry air, the specific heat capacities are

$$c_{pd} - c_{vd} = R_d \tag{31}$$

Equation (30) demonstrates that $C_p > C_v$. Statistical mechanics show that specific relationships between specific heats at constant volume and at constant pressure can be found for a gas depending on how many atoms form the gas molecule. Table 3

TABLE 3 Relationship between Specific Heats and Molecular Structure of Gas

Monotonic gas	$C_v = \frac{3}{2}R$	$C_p = C_v + R = \frac{5}{2}R^*$
Diatomic gas	$C_v = \frac{5}{2}R$	$C_p = \frac{7}{2}R$

shows these relationships for diatomic and monoatomic gases. Since dry air, composed mostly of molecular oxygen and nitrogen is nearly a diatomic gas,

$$c_{pd} = 1004 \text{ J/kg/K}, \qquad c_{vd} = 717 \text{ J/kg/K}. \tag{32}$$

Substituting Eq. (30) into Eq. (28) we obtain

$$\delta Q = C_p\, dT - V\, dP - \sum_k dA_k \tag{33}$$

This form of the first law is especially useful because changes of pressure and temperature are most commonly measured in atmospheric science applications. Note that $C_p\, dT$ is not purely a change in energy, nor is $V\, dp$ purely the work.

Enthalpy For convenience, we introduce a new state variable called *enthalpy*, defined as:

$$H \equiv U + PV. \tag{34}$$

Differentiating across we obtain

$$dH = dU + p\, dV + V\, dP, \tag{35}$$

and substituting Eq. (21), to eliminate dU:

$$\delta Q = dH - V\, dP - \sum_k dA_k. \tag{36}$$

Using Euler's rule with Eq. (33), we obtain

$$\left(\frac{\partial H}{\partial T}\right)_p = C_p. \tag{37}$$

Enthalpy differs only slightly from energy but arises as the preferred energy variable when the work term is expressed with a pressure change rather than a volume change. The enthalpy variable exists purely for convenience (since pressure is more easily measured than volume) and is not demanded or defined by any thermodynamic law.

Reversible Adiabatic Processes in the Atmosphere: Definition of Potential Temperature (θ) These types of processes are of great importance to the atmospheric scientist. Ignoring radiative effects, or heating effects of condensation, the dry adiabatic process represents the thermal tendencies that air will experience as it rises and the pressure lowers or as it sinks and the pressure rises. In the absence of condensation, this represents the bulk of the temperature change a parcel will experience if its rise rate is fast enough to ignore radiative effects. We defined an adiabatic process as one for which $Q = 0$. Then for an adiabatic process, combining Eq. (33), the equation of state [Eq. (7)], and then integrating we obtain

$$T = kP^{\kappa} \tag{38}$$

where $\kappa = R^*/C_p = 0.29$ and k is a constant of integration that can be determined from a solution point. Equation (38) is known to atmospheric scientists as the *Poisson equation*, not to be confused with the second-order partial differential equation also known as "Poisson's equation." The Poisson equation states that, given a relationship between temperature and pressure at some reference state, the value of temperature can be calculated for all other pressures the system might obtain through reversible dry adiabatic processes.

It is often convenient to define *potential temperature (θ)* to be the temperature that a parcel would have at 1000 kPa pressure. It follows directly from Eq. (38) that

$$\theta \equiv T\left(\frac{p_{00}}{P}\right)^{R/c_p}, \tag{39}$$

where $p_{00} = 1000\,\text{kPa}$ is the reference pressure at which potential temperature (θ) equals temperature (T). The potential temperature proves extremely useful to meteorologists. For all reversible dry adiabatic atmospheric thermodynamic processes, θ remains constant and represents a conserved property of the flow. Later we will show θ also represents the dry entropy of the flow and that it is one in a hierarchy of entropy variables that are conserved in flow under various permitted dry and moist processes.

Second Law (or Principle) of Thermodynamics

The zeroth principle of thermodynamics defined temperature to be a quantity that is determined from two bodies in thermal equilibrium. The first law stated the principle of energy conservation during any thermodynamic process. The second law deals with the direction of energy transfer during a thermodynamic process occurring when two bodies are not in thermodynamic equilibrium. It is again an empirical statement, not derived in any way from first principles but rather from observations of our perceived universe. The second law can be stated in two ways that can be shown to be equivalent:

1. *Thermal energy will not spontaneously* flow from a colder to a warmer object.
2. A thermodynamic process always acts down gradient in the universe to reduce differentiation overall and hence mix things up and reduce order overall. If we define entropy to be a measure of the degree of disorder, then the second law states: *The entropy of the universe increases or remains the same as a result of any process. Hence, the entropy of the universe is always increasing.*

Just as the first law demanded that we define the thermal variable temperature, the second law requires a new thermodynamic variable *entropy.* To form a mathematical

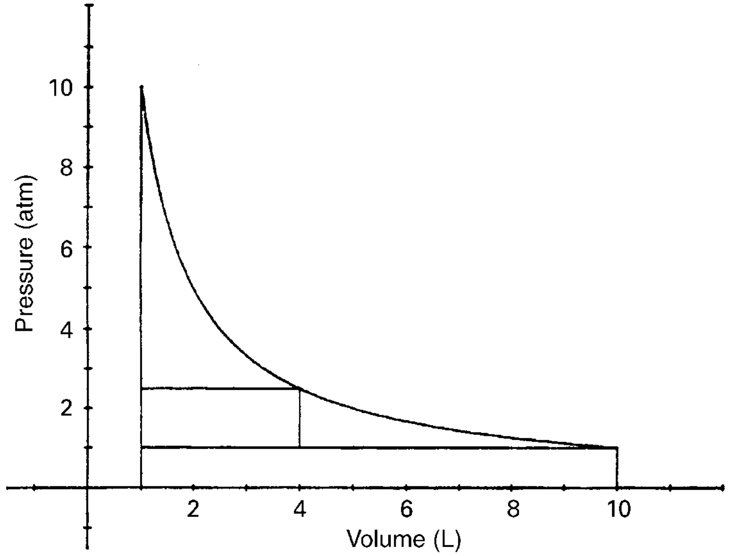


Figure 4 Illustration of three isothermal processes.

treatment for entropy, we consider several special thermodynamic processes that reveal the nature of entropy behavior and provide guidance to its form.

Mathematical Statement of the Second Law It is illustrative to examine the amount of work performed by a parcel on its surroundings under *isothermal* conditions, implying that the internal energy of the system is held constant. This will demonstrate an interesting behavior addressed by the second law.

Consider a fixed mass of an ideal gas confined in a cylinder fitted with a movable piston of variable weight. The weight of the piston and its cross-sectional area determine the pressure on the gas. Let the entire assembly be placed in a constant temperature bath to maintain isothermal conditions. Let the initial pressure of the gas be 10 atm and the initial volume be 1 liter. Now consider three isothermal processes (Fig. 4) (Sears, 1953):

Process I The weight on the piston is changed so as to produce an effective pressure of 1 atm and since $PV = nR^*T$, its volume becomes 10 liters. The work of expansion is given by:

$$W_{\text{exp}} = P\Delta Va$$

$$(1 \text{ atm} = 1013 \times 10^2 \text{ Pa}) \qquad (40)$$

Since P_{surr} is constant at 1 atm, the work is $1(10 - 1) = 9$ Latm. This work is performed on the surroundings. With respect to the system, the energy transfer (δQ) is therefore -9 Latm.

Process II This will occur in two stages: (1) Decrease the pressure exerted by the piston to 2.5 atm so that the specific volume will be 4 liters and then (2) reduce the pressure to 1 atm with a specific volume of 10 liters. The work of expansion becomes the sum:

$$\begin{aligned} w_{\text{exp}} &= P_1\,\Delta V_1 + P_2\,\Delta V_2 \\ &= 7.5 + 6.0 \\ &= 13.5\ \text{Latm} \end{aligned}$$

The mechanical energy transfer from the system (δQ) is thus -13.5 Latm.

Process III Let the pressure exerted by the piston be continuously reduced such that the pressure of the gas is infinitesimally greater than that exerted by the piston (otherwise no expansion would occur). Then the pressure of the gas is essentially equal to that of the surroundings and since $dw = P\,dV$,

$$\begin{aligned} W_{\text{exp}} &= \int_{V_1}^{V_2} P\,d\alpha = R^*T\ln\left(\frac{V_2}{V_1}\right) \\ &- -23.03\ \text{Latm} \end{aligned}$$

In each of the above three cases, the system responded to the removal of external weight to the piston by changing state and performing a net work on the surroundings while maintaining isothermal conditions within the system by absorbing heat from the attached heat reservoir, the reservoir being part of the surroundings. In each case, the total change in the system state was identical, although the energy absorbed by the system from the surroundings, equal also to the work performed on the surroundings by the process varied greatly.

Note that process III differed from the first two processes in that it was an *equilibrium* process, or one where the difference between the intensive driving variable (pressure in this case) of the system differed only infinitesimally from the surroundings. The maximum work possible for the given change in system state occurs in such an equilibrium process. All other processes, not in equilibrium, produce less net work on the surroundings and thus also absorb less energy from the heat reservoir!

It is important to note that, unlike the nonequilibrium processes, *the net work for the equilibrium process is dependent only on the initial and final states of the system*. Hence if the equilibrium process is run in reverse, the same amount of work is performed on the system as the system performed on the surroundings in the forward direction. In the case of the nonequilibrium process, this is not the case and so the amount of energy released to the surroundings for the reverse process is unequal to that absorbed in the forward direction. Hence the system is *irreversible*. Hence we can equate an irreversible process to one which is not in equilibrium.

For an isothermal reversible process involving an ideal gas, $dU = 0$ and hence $\delta Q = p\ dV = dW_{\max}$. Thus δQ is dependent only on the initial and final states of the system. *Hence, for a reversible process, Q and W behave like state functions.* As a consequence, since for either the irreversible or reversible processes, $\delta Q = \delta W$, it follows that:

$$\delta Q_{\text{rev}} = \delta W_{\max} \tag{42}$$

Since the δQ_{rev} is related to a change in state, then the isothermal reversible result described by process III must apply for the general isothermal case, i.e.,

$$\frac{\delta Q_{\text{rev}}}{T} = R^* d\ \ln V \tag{43}$$

where the right-hand side of Eq. (43) is an exact differential. By this reasoning it is concluded that the left-hand side of Eq. (43) must also be an exact differential of some variable that we will define to be the *entropy* (*S*) of the system (joules per kelvin). Since the change in entropy of the system is always an exact differential, its value after any thermodynamic process is independent of the path of that process and so *entropy, itself, is also a state function*. We can write:

$$dS = \frac{\delta Q_{\text{rev}}}{T} \tag{44}$$

From Eq. (42) we can also show that

$$\delta W_{\max} = T\ dS \tag{45}$$

which states that the maximum work that can be performed by a change in state is equal to the temperature multiplied by the total change in entropy from the initial to the final state of the system. Since we showed that the net work performed on the surroundings as a result of the isothermal process must be less than or equal to $dW_{\max}$, then it is implied that

$$dS > \frac{\delta Q}{T} \tag{46}$$

for an *irreversible* process. It is also reasoned that by virtue of the second law, and obvious from the example described above, that a process where $dS < \delta Q/T$ would be a *forbidden* process. Since the entropy of the surroundings must also satisfy these constraints, the entropy of the universe can either remain constant for a reversible process or increase overall for an irreversible process. This is the mathematical statement of the second law for the case of isothermal processes.

We must now expand this concept to a system not constrained to isothermal conditions. To build a theory applicable to a general process, we consider the following additional particular processes:

1. *Adiabatic Reversible Expansion of an Ideal Gas* For a reversible adiabatic expansion, $\delta Q = 0$ and so $dS = 0$. This is an *isentropic process*. Since a

reversible dry adiabatic process is one where θ is conserved, a process with constant θ is also called an isentropic process.

2. *Heating of a Gas at Constant Volume:* For a reversible process at constant volume, the work term vanishes:

$$dS = \frac{dQ_{\text{rev}}}{T} = C_v \frac{dT}{T} = C_v d \ \ln T \tag{47}$$

3. *Heating of an Ideal Gas at Constant Pressure:* For a reversible process:

$$dS = \frac{dQ_{\text{rev}}}{T} = C_p \frac{dT}{T} = C_p d \ \ln T \tag{48}$$

The Carnot Cycle We demonstrated earlier that a net work can also be performed on the surroundings as a system goes through a cyclic process. That work, we showed, was equal to the area traced out on a *P–V* diagram by that cyclic process. The maximum amount of work possible by any process was that performed by a *reversible process*. We can call our *system* that performs work on the surroundings as the *working substance*.

Now we ask, *Where does the energy for the work performed originate?* For a cyclic process the initial and final states of the working substance are the same so any work performed over that process must come from the *net* heat absorbed by the system from the surroundings during that process. We say *net* because the system typically *absorbs* heat and also *rejects* heat to the surroundings. If the system converted all of the heat absorbed to work without rejecting any we would say that the system is 100% efficient. That cannot happen typically. For instance, the engine in a car burns gasoline to heat air within the cylinder that expands to do work. But much of the heat given off from the gas actually ends up warming the engine and ultimately the surroundings rather than making the engine turn. The degree to which the energy warms the surroundings compared to how much is used to turn the engine and ultimately move the car (more energy is lost to friction in the engine and transmission and so on) is a measure of how efficient the car is.

For any engine, we can quantify its efficiency as:

$$\eta \equiv \frac{\text{Mechanical work performed by engine}}{\text{Heat absorbed by engine}} \equiv \frac{Q_1 - Q_2}{Q_1} = \frac{q_1 - q_2}{q_1} \tag{49}$$

where Q_1 is the heat absorbed by the engine and Q_2 is the heat rejected by the engine to the surroundings. In terms of specific heat:

$$\eta = \frac{q_1 - q_2}{q_1} \tag{50}$$

The maximum efficiency that any engine can ever achieve would be that of an engine powered by a reversible process. Processes such as friction in the engine or trans-

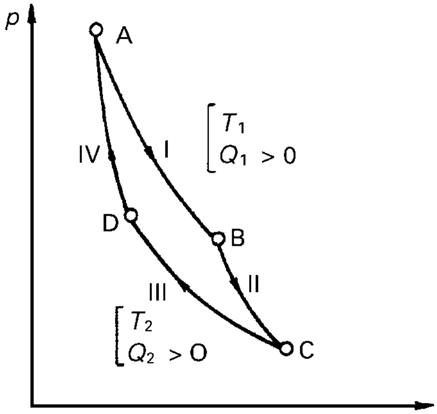


Figure 5 Carnot cycle (from Iribarne and Godson, 1973).

mission, not perfectly insulated walls of the pistons, and so on would all be irreversible parts of the cyclic process in a gasoline engine. Now let's imagine an engine where the thermodynamic cyclic process is carried out perfectly reversibly. One simple reversible thermodynamic process that fits this bill is the *Carnot cycle*, which powers an imaginary engine called a *Carnot engine*. No one can build such a perfect frictionless engine, but we study it because it tells us what the *maximum efficiency possible* is for cyclic process. Yes, even the Carnot engine is not 100% efficient and below we will see why.

First let's describe the Carnot cycle. Simply put, the Carnot cycle is formed by two adiabatic legs and two isothermal legs so that the net energy transfers can be easily evaluated with simple formulations for adiabatic or isothermal processes. Figure 5 shows such a process. The four legs of the process are labeled I to IV beginning and ending at point A. To visualize this physically, we can imagine some working substance contained within a cylinder having insulated walls and a conducting base, and a frictionless insulated beginning piston. The cycle is explained as:

I The cylinder is placed on a heat reservoir held at $T = T_1$. Beginning at $T = T_1$, and $V = V_A$, the working substance is allowed to slowly expand maintaining isothermal conditions.

II The cylinder is placed on an insulated stand and allowed to expand further to volume V_C and temperature T_2. Since the working substance is totally insulated for stage II, the process is *adiabatic*.

III The cylinder is taken from the stand and placed on a heat reservoir where the working substance is held at $T = T_2$ and slowly compressed (isothermally) to specific volume V_D.

IV The cylinder is replaced on the insulated stand and the working substance is slowly compressed (perhaps adding weight to piston slowly) causing the substance to warm adiabatically back to temperature T_1 and specific volume V_A

This process is illustrated graphically in Figure 5. The cycle is reversible and consists of two isotherms at temperatures T_1 and T_2, where $T_2 < T_1$ and two adiabats, i.e., θ_1 and θ_2.

The net work performed by the system is the area formed by the intersection of the four curves. We can compute the work A and the heat Q for the four steps of the process (direction indicated by arrows) in the following table:

$$\begin{array}{lll}
\text{(I)} & \Delta U_{\mathrm{I}} = 0 & \delta Q_{\mathrm{I}} = -A_{\mathrm{I}} = \int_A^B P\,dV = R^* T_1 \ln V_B/V_A \\
\text{(II)} & \delta Q_{\mathrm{II}} = 0 & -A_{\mathrm{II}} = -\Delta U_{\mathrm{II}} = C_V(T_1 - T_2) \\
\text{(III)} & \Delta U_{\mathrm{III}} = 0 & \delta Q_2 = -A_{\mathrm{III}} = -R^* T_2 \ln V_C/V_D \\
\text{(IV)} & \delta Q_{\mathrm{IV}} = 0 & -A_{\mathrm{IV}} = -\Delta U_{\mathrm{IV}} = -C_V(T_1 - T_2)
\end{array}$$

Since this is a cyclic process, the sum of the four terms Δu is zero. The work terms on the two adiabats cancel each other. The gas effectively absorbs the quantity of heat $Q_1 > 0$ from the warmer reservoir and rejects it $Q_2 < 0$ in the colder reservoir. The total work is then $W = A_{\mathrm{I}} + A_{\mathrm{III}} = -(Q_1 + Q_2)$.

We can now relate q_1 and q_2 to the temperatures of the two heat reservoirs. From Poisson's equation,

$$\frac{T_1}{T_2} = \left(\frac{V_C}{V_B}\right)^{R^*/C_p} = \left(\frac{V_D}{V_A}\right)^{R^*/C_p}, \tag{51}$$

and so:

$$\frac{V_B}{V_A} = \frac{V_C}{V_D}, \tag{52}$$

which, substituting into the expressions for Q_1 and Q_2 gives

$$\frac{Q_1}{T_1} + \frac{Q_2}{T_2} = 0, \tag{53}$$

or, alternatively

$$\left|\frac{|Q_2|}{Q_1}\right| = \frac{T_2}{T_1}. \tag{54}$$

Now we can calculate the thermodynamic efficiency of the Carnot engine:

$$\eta = \frac{T_1 - T_2}{T_1}. \tag{55}$$

Hence, *the Carnot cycle achieves its highest efficiencies when the temperature difference between the two heat reservoirs is the greatest!*

It is useful to generalize this result to the case of irreversible processes. To do so, we consider the second law, and its implication that heat can only flow from a warmer toward a cooler temperature. As a first step, it is noted that Eq. (55) is valid for any reversible cycle performed between two heat sources T_1 and T_2, independent of the nature of the cycle and of the systems. This is a statement of *Carnot's theorem*.

Just as Eq. (55) holds for a reversible cycle, it can be shown that for any *irreversible* cycle, the second law requires that $(Q_1/T_1) + (Q_2/T_2) < 0$, i.e., heat *must* flow from warm to cold and not vice versa. It can also be shown that any reversible cycle can be decomposed into a number of Carnot cycles (reversible cycles between two adiabats and two isotherms). Therefore, for any irreversible process it can be shown that

$$\oint \frac{\delta Q}{T} \leq 0 \tag{56}$$

which implies Eq. (46).

The Carnot cycle can also be viewed in reverse, in which case it is a refrigerating machine. In that case a quantity Q_2 of heat is taken from a cold body (the cold reservoir) and heat Q_1 is given to a hot reservoir. For this to happen, mechanical work must be performed on the system by the surroundings. In the case of a refrigerator, an electric motor supplies the work.

In recent years, atmospheric scientists have discovered that certain types of weather systems derive their energy from a Carnot cycle. In particular, the tropical cyclone* is powered by a Carnot-like cycle created by the radial circulation inward to the storm center at the surface, up within the eye-wall clouds, outward at the storm top, and then back downward far away from the storm center. State I, or the Carnot cycle, is equivalent to the air moving inward toward the storm center along the ocean surface at relatively constant temperature and toward low pressure in the storm center. The heat reservoir is the ocean. In this case it heats not only through the transfer of sensible heat but also through the transfer of latent heat. Stage II is found in the eye wall, where the heated air rises moist adiabatically under falling temperatures. Stage III occurs in the outflow at high levels away from the storm center. The outflowing air maintains a constant temperature while slowly subsiding and pressure

*Tropical cyclones are known as "hurricanes" in the Atlantic and eastern Pacific oceans, while in the western Pacific they are known as "typhoons."

rising. The heat is lost to space by transfer through the emission of long-wave radiation. Hence the cold plate is the radiation sink in space. Finally, stage IV occurs as the air finally sinks back to the ground approximately moist adiabatically. This occurs mainly within the downdrafts of convective clouds in the periphery of the storm. The result is a net gain in energy of the entire system, which is manifested in the cyclone winds. The storm loses energy primarily to surface friction, which is proportional to the speed of the winds. Therefore, the more energy released by the Carnot cycle, the stronger the surface winds that come into equilibrium with the Carnot cycle release. As with the discussion above, the efficiency of the cycle is proportional to the temperature difference between storm top and the sea surface. As a result, the strength attainable by a tropical cyclone is closely related to the sea surface temperature.

Energy harnessed by the Carnot cycle is likely the basis for other types of weather for which thermally driven convection is important. These include a host of convective systems and possibly some aspects of Earth's general circulation that derive their energy from thermally direct circulations.

Restatement of First Law with Entropy The second law is a statement of inequality regarding the limits of entropy behavior. When combined with the first law, the resulting relationships become inequalities:

$$T\,ds \geq dU + P\,dV - \sum_k dA_k \tag{57}$$

$$T\,dS \geq dH - V\,dP - \sum_k dA_k \tag{58}$$

for the internal energy and enthalpy forms, respectively. Alternatively, we may write for the special case of a reversible process:

$$dU = T\,dS - P\,dV + \sum_k dA_k \quad \text{and} \tag{59}$$

$$dH = T\,dS + V\,dP + \sum_k dA_k. \tag{60}$$

Note this differs from previous forms of the first law in that we have an equation for the parcel (or system) in terms of state function variables only. The difference between the general Eqs. (57) and (58) and the special case Eqs. (59) and (60) is a positive source term for entropy resulting from a nonequilibrium reaction that has not yet been determined. In the world of dry atmospheric thermodynamics, we can live with assuming the dry system is in equilibrium and so the second form is sufficient. However, when we begin to consider moist processes, the system is not always in equilibrium, and so the sources for entropy from nonequilibrium processes will have to be evaluated. The framework for the evaluation of these effects involves the creation of free energy relations described in the next chapter.

Free Energy Functions In their present form, Eqs. (57) and (58) relate thermodynamic functions that involve dependent variables that include the extensive variable S. It is again convenient to make a variable transformation to convert the dependence to variations in T instead of S. The transformation is made by defining the so-called *free energy functions* called the *Helmholtz free energy* (F) and the *Gibbs free energy* (G) are defined as:

$$F \equiv U - TS$$

$$G \equiv H - TS$$

In derivative form, they are written:

$$dF = dU - T\,dS - S\,dT$$

$$dG = dH - T\,dS - S\,dT$$

For our application to atmospheric problems, we will work primarily with the Gibbs free energy because of its reference to a constant pressure process.

Combining Eq. (61) with Eq. (60) we obtain

$$dG = -S\,dT + V\,dP + \sum_k dA_k + \sum_k dG_k \tag{61}$$

where G_k refers to Gibbs energies for each constituent of the system (k) necessary to form an equality of Eq. (58). The restrictions imposed by the second law and reflected by those inequalities result in the requirement that $\sum_k dG_k \leq 0$. Because $G = G(T, V, n_j)$ is an exact differential, then

$$dG = \left(\frac{\partial G}{\partial T}\right)_{P,n_k} dT + \left(\frac{\partial G}{\partial P}\right)_{T,n_k} dP + \left(\frac{\partial G}{\partial n_k}\right)_{P,T,n_j} dn_k \tag{62}$$

We can evaluate the potential internal energy and Gibbs free energy terms of Eq. (61) as:

$$\sum_k dA_k = \left(\frac{\partial U}{\partial n_k}\right)_{T,P} dn_k$$

$$\sum_k dG_k = \left(\frac{\partial G_\mu}{\partial n_k}\right)_{T,P} dn_k$$

where G_μ is the Gibbs free energy resulting from chemical potential. The terms of Eq. (62) are then evaluated:

$$\left(\frac{\partial G}{\partial n_k}\right)_{T,P} = \left(\frac{\partial U}{\partial n_k}\right)_{T,P} dn_k + \left(\frac{\partial G_\mu}{\partial n_k}\right)_{T,P} dn_k,$$

$$\left(\frac{\partial G}{\partial T}\right)_{V,n_k} = -S, \text{ and}$$

$$\left(\frac{\partial G}{\partial V}\right)_{T,n_k} = -P.$$

Concept of Equilibrium Equilibrium is not an absolute but is defined in terms of permitted processes. In statics, equilibrium is described as the state when the sum of all forces is zero. This may be stated as a principle of work: *If a system is displaced minutely from an equilibrium state by a change in one of the system properties, the sum of all the energy changes is zero*. Neglecting internal potential energies (A_k), this leads to the statement that for a single component system of ideal gas, undergoing a constant temperature, constant volume process, $dF = 0$ while for a constant temperature process and constant pressure process $dG = 0$ at equilibrium. Hence if one were to plot G (or F) as a function of system properties, equilibrium will appear as a minimum along the G (or F) curve.

In general, we define a spontaneous process as one in which the system begins out of equilibrium and moves toward equilibrium, resulting in an increase in entropy. The rate of the process is undetermined thermodynamically. Consider the Gibbs free energy for a constant temperature and constant pressure process. Since equilibrium occurs at a minimum in Gibbs free energy, we can define:

A *spontaneous* process	$\Delta G < 0$
An *equilibrium* process	$\Delta G = 0$
A *forbidden* process	$\Delta G > 0$

The *Molar chemical potential* of species k (μ_k) is defined as:

$$\mu_j \equiv \left(\frac{\partial G_\mu}{\partial n_k}\right)_{T,p,n_j}. \tag{63}$$

For a system of only one component of a set on noninteracting components:

$$\mu_k \, dn_k = G_k(m) = H_k - TS_k \tag{64}$$

in which $G_k(m)$ is the molar free energy of species k. We now write:

$$dG = -S\,dT + V\,dP + \sum_k dA_k + \sum_k \mu_k\,dn_k. \tag{65}$$

Hence the third term on the right-hand side (RHS) would represent free energies resulting from the potential for interaction between all components in the system, while the last term on the RHS represents the noninteracting free energy of each component, for example, surface free energy.

Consider a system composed of a single ideal gas and held in a state of equilibrium where $dG = 0$. Ignoring internal potential energies, Eq. (65) becomes

$$d\mu = S\,dT - V\,dP. \tag{66}$$

Using the equation of state and integrating, we obtain

$$\mu = \mu_0(T) + nR^*T\ln P, \tag{67}$$

where $\mu_0(T)$ is the *standard chemical potential* at unit pressure (usually taken to be 1 atm). Hence, for an ideal gas with two constituents:

$$\Delta G = -W_{\max} = (\mu_2 - \mu_1)(n_2 - n_1) = nRT\ \ln\frac{p_2}{p_1}. \tag{68}$$

We can also look at the chemical potential of a condensed phase, using the concept of equilibrium. Consider liquid water in equilibrium with water vapor at a temperature T and a partial pressure of e_s. Applying the principle of virtual work, we transfer dn moles of water from one phase to the other under the condition:

$$\mu_l\,dn = \mu_v\,dn,$$
$$\mu_l = \mu_v = \mu_l^0 + R_vT\ \ln e_s,$$

where μ_l and μ_v are the channel potentials of liquid and vapor respectively.

This is a general statement and demonstrates that the chemical potential of any species is constant throughout the system (if the equilibrium constraints permit transfer of the species). We can also see that for a nonequilibrium phase change:

$$\Delta G = \mu_v - \mu_l = R_vT\ \ln\frac{e_v}{e_s}, \tag{69}$$

where e_v is the partial pressure of vapour not at equilibrium.

Hence if $e_v < e_s$, then $\Delta G < 0$ for a process of condensation and hence that is a forbidden process. By the same token, evaporation becomes a spontaneous process. When $\mu_v = \mu_l$, the system is in equilibrium and at *saturation*. These concepts can be extended to include other free energies affecting condensation or sublimation such as solution effects, curvature effects, to process of chemical equilibrium and so on.

REFERENCES

Dutton, J. (1976). *The Ceaseless Wind*, McGraw-Hill.

Emanuel, K. A. (1994). *Atmospheric Convection*, Oxford University Press.

Hess, S. L. (1959). *Introduction to Theoretical Meteorology*, Holt, Reinhardt, and Winston.

Iribarne, J. V., and W. L. Godson (1973). *Atmospheric Thermodynamics*, D. Reidel Publishing Co.

Sears, F. W. (1953). *Thermodynamics*, Addison-Wesley.

Wallace, J. M., and P. V. Hobbs, (1977). *Atmospheric Science: An Introductory Survey*, Academic Press.

CHAPTER 16

MOIST THERMODYNAMICS*

AMANDA S. ADAMS

1 LATENT HEATS—KIRCHOFF'S EQUATION

Consider the effect resulting from mass fluxes between constituents of the parcel within the system. In particular, mass fluxes between the three possible phases of water are considered. We can express these phase transformations by:

$$dH = \delta Q + V\,dP + \sum_k \left(\frac{\partial H_k}{\partial n_k}\right) dn_k \tag{1}$$

where $\partial H_k/\partial n_k$ represents the energy per mole of the n_k constituent of the system. We let the system of moist air be composed of n_d moles of dry air, n_v moles of vapor gas, n_l moles of liquid water and n_i moles of ice. Then

$$dH = \delta Q + V\,dP + \left(\frac{\partial H_v}{\partial n_v} - \frac{\partial H_l}{\partial n_l}\right) dn_v + \left(\frac{\partial H_i}{\partial n_i} - \frac{\partial H_l}{\partial n_l}\right) dn_i \tag{2}$$

*The derivations, concepts, and examples herein are based partially on course notes from the late Professor Myron Corrin of Colorado State University. Other derivations are taken from the Isbane and Godson text, the Emanuel (1994: *Atmospheric Convection*, Oxford University Press) text, and the Dutton (*The Ceaseless Wind*) book.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

We define latent heats to be the differences in enthalpy per mole between two phases at the same temperature and pressure and is expressed as:

$$\begin{aligned} L_{lv} &\equiv \left(\frac{\partial H_l}{\partial n_l} - \frac{\partial H_v}{\partial n_v}\right) \\ L_{il} &\equiv +\left(\frac{\partial H_i}{\partial n_i} - \frac{\partial H_l}{\partial n_l}\right) \\ L_{iv} &\equiv L_{il} + L_{lv} \end{aligned} \tag{3}$$

L_{lv}, L_{il}, and L_{iv} are the latent heats of condensation, melting, and sublimation defined as positive quantities with units of joules per mole. Employing these definitions, Eq. (2) can be written:

$$dH = \delta Q + V\,dP - L_{iv}\,dn_v + L_{il}\,dn_i \tag{4}$$

In order to find Kirchoff's equation, we write Eq. (4) for three adiabatic homogenous systems consisting of vapor, liquid, and ice held at constant pressure:

$$\begin{aligned} C_{pv}T &= \frac{\partial H_v}{\partial n_v}, \\ C_l T &= \frac{\partial H_l}{\partial n_l}, \text{ and} \\ C_i T &= \frac{\partial H_i}{\partial n_i}. \end{aligned} \tag{5}$$

where C_{pv}, C_l, and C_i are the heat capacity of vapor at constant pressure, the heat capacity of liquid, and the heat capacity of ice. Applying (5) to (3) and differencing with respect to temperature we obtain Kirchoff's equation:

$$\begin{aligned} \frac{\partial L_{lv}}{\partial T} &= C_{pv} - C_l, \\ \frac{\partial L_{il}}{\partial T} &= C_l - C_i, \text{ and} \\ \frac{\partial L_{iv}}{\partial T} &= C_{pv} - C_i. \end{aligned} \tag{6}$$

In the "enthalpy per mass" form, Kirchoff's equation is:

$$\begin{aligned}\frac{\partial l_{lv}}{\partial T} &= c_{pv} - c_l,\\ \frac{\partial l_{il}}{\partial T} &= c_l - c_i, \text{ and}\\ \frac{\partial l_{iv}}{\partial T} &= c_{pv} - c_i.\end{aligned} \tag{7}$$

Given heat capacities determined observationally as a function of temperature, we can determine the variation of latent heat with temperature. Kirchoff's law can also be used to study reaction heats of chemical changes.

2 GIBBS PHASE RULE

For the case of a simple homogenous gas, we found that there were three independent variables, being the pressure, temperature, and volume assuming the number of moles was specified. Once we require that the gas behave as an ideal gas, the imposition of the equation of state (and the implicit assumption of equilibrium) reduces the number of independent variables by one to two.

If we now look at a system consisting entirely of water, but allow for both liquid and gas forms, and again assume equilibrium, then not only must the vapor obey the ideal gas law, but the chemical potential of the liquid must equal that of the vapor. This means that if we know the pressure of the vapor, only one water temperature can be in true equilibrium with that vapor, i.e., the temperature that makes the liquid exert a vapor pressure equal to the vapor pressure in the air. Hence by requiring equilibrium with two phases, the number of degrees of freedom is reduced to one.

If we allow for all three phases, i.e., vapor, liquid, and ice, then there is only one temperature and pressure where all three states can exist simultaneously, called the *triple point*.

Now consider a two-component system such as one where we have water and air mixed. Consider a system allowing only liquid but not the ice phase of water. Now we have to consider the equilibrium as is applied to the dry air by the ideal gas law in addition to the equilibrium between the liquid and ice water. For this case, the water alone had one independent variable and the dry air had two, for instance, its partial pressure and temperature. If the vapor pressure is specified, then the temperature of the water is specified and, for equilibrium, so is the air temperature. Hence adding the additional component of air to the two-phase water system increased the number of independent variables to two.

A general statement concerning the number of independent variables of a heterogeneous system is made by *Gibbs phase rule*, which states:

$$v = c - \phi + 2, \tag{8}$$

where v is the number of independent variables (or degrees of freedom), c is the number of independent species, and ϕ is the number of phases total. Hence for a water–air mixture, allowing two phases of water and one phase of dry air, there is 1 degree of freedom. So if we specify vapor pressure, then liquid temperature is known, air temperature is known, and so air pressure is fixed assuming the volume of the system is specified.

3 PHASE EQUILIBRIUM FOR WATER

Figure 1 shows where equilibrium exists between phases for water. Note that equilibrium between any two phases is depicted by a line, suggesting one degree of freedom, while all three phases are only possible at the triple point. Note that the curve for liquid–ice equilibrium is nearly of constant temperature but weakly slopes to cooler temperatures at higher pressures. This is attributable to the fact that water increases in volume slightly as it freezes so that freezing can be inhibited some by applying pressure against the expansion.

Note also that at temperatures below the triple point there are multiple equilibria, i.e., one for ice–vapor and another (dashed line) for liquid–vapor. This occurs because the free energy necessary to initiate a freezing process is high, and local equilibrium between the vapor and liquid phase may exist.

Note that at high temperatures, the vapor pressure curve for water abruptly ends. This is the critical point (p_c) where there is no longer a discontinuity between the liquid and gaseous phase. As we showed earlier, the latent heat of evaporation l_{lv} decreases with increasing temperature. It becomes zero at the critical point.

Since the triple point is a well-defined singularity, we define thermodynamic constants for water at that point, and these values are given in Table 1. At the critical point, liquid and vapor become indistinguishable. Critical point statistics are given in

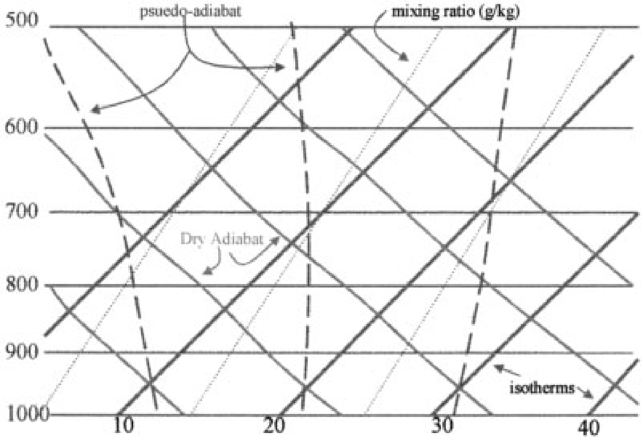


Figure 1 Phase-transition equilibria (from Iribarne and Godson, 1973).

TABLE 1 Triple Point Values for Water

Variable	Symbol	Value
Temperature	T_t	273.16 K
Pressure	p_t	610.7 Pa, 6.107 mb
Ice density	$\rho_{i,t}$	$917\,\mathrm{kg/m^{-3}}$
Liquid density	$\rho_{l,t}$	$1000\,\mathrm{kg/m^{-3}}$
Vapor density	$\rho_{v,t}$	$0.005\,\mathrm{kg/m^{-3}}$
Specific volume ice	$\alpha_{i,t}$	$1.091 \times 10^{-3}\ \mathrm{m^3/kg}$
Specific volume liquid	$\alpha_{l,t}$	$1.000 \times 10^{-3}\ \mathrm{m^3/kg}$
Specific volume vapor	$\alpha_{v,t}$	$2.060 \times 10^{2}\ \mathrm{m^3/kg}$
Latent heat of condensation	$l_{vl,t}$	$2.5008 \times 10^{6}\ \mathrm{J/kg}$
Latent heat of sublimation	$l_{vi,t}$	$2.8345 \times 10^{6}\ \mathrm{J/kg}$
Latent heat of melting	$l_{il,t}$	$0.3337 \times 10^{6}\ \mathrm{J/kg}$

Table 2. These phase relationships can also be seen schematically with an Amagat–Andrews diagram (Fig. 2). Note how the isotherms follow a hyperbola-like pattern at very high temperatures as would be expected by the equation of state for an ideal gas. However, once the temperature decreases to values less than T_C, an isothermal process goes through a transition to liquid as the system is compressed isothermally. The latent heat release and the loss of volume of the system keep the pressure constant while the phase change occurs. After the zone of phase transition is crossed, only liquid (or ice at temperatures below T_t) remains, which has very low compressibility and so a dramatic increase in pressure with further compression.

Phase equilibrium surfaces can be displayed also in three dimensions as a *p*-*V*-*T* surface and seen in Figure 3. Here we see that at very high temperatures, the region where phases coexist is lost. Note how the fact that water expands upon freezing results in the kink backward of the liquid surface. Note that we can attain the saturation curves shown in Figure 1 by cutting cross sections through the *p*-*V*-*T* diagram and constant volume or temperature (Fig. 3).

We can use these figures to view how the phase of a substance will vary under differing conditions, since the phases of the substance must lie on these surfaces. Figure 4 shows one such system evolution beginning at point *a*. Note the liquid system exists at point *a* under pressure p_1. The pressure presumably is exerted by an atmosphere of total pressure p_1 above the surface of the liquid. Holding this pressure constant and heating the liquid, we see that the system warms with only a small volume increase to point *b* where the system begins to evolve to a vapor phase at

TABLE 2 Critical Point Values for Water

Variable	Symbol	Value
Temperature	T_c	647 K
Pressure	p_c	2.22×10^{7} Pa (218.8 atm)
Specific volume vapor	$\alpha_{c,t}$	$3.07 \times 10^{-3}\ \mathrm{m^3/kg}$

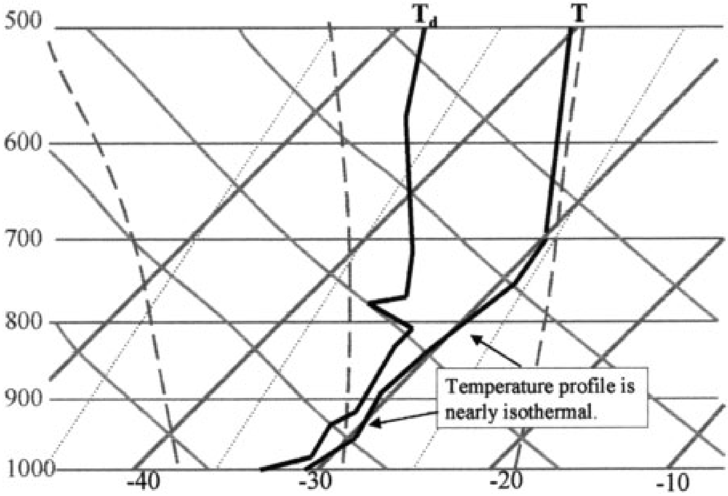


Figure 2 Amagat–Andrews diagram (from Iribarne and Godson, 1973).

much higher volume as the temperature holds constant. The evolution from b to c is commonly called *boiling* and occurs when the vapor pressure along the vapor–liquid interface becomes equal to the atmospheric pressure. Hence we can also view the saturation curves as boiling point curves for various atmospheric pressures.

Alternatively, if the system cools from point a, the volume very slowly decreases until freezing commences at point d. There we see the volume more rapidly decrease as freezing occurs. Of course, for water the volume change is reversed to expansion and is not depicted for the substance displayed on this diagram.

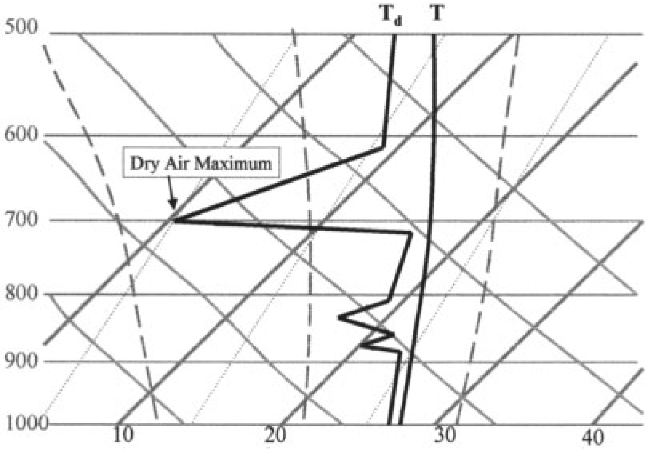


Figure 3 Projection of the p-V-T surface on the p-T and p-V planes. (Sears, 1959).

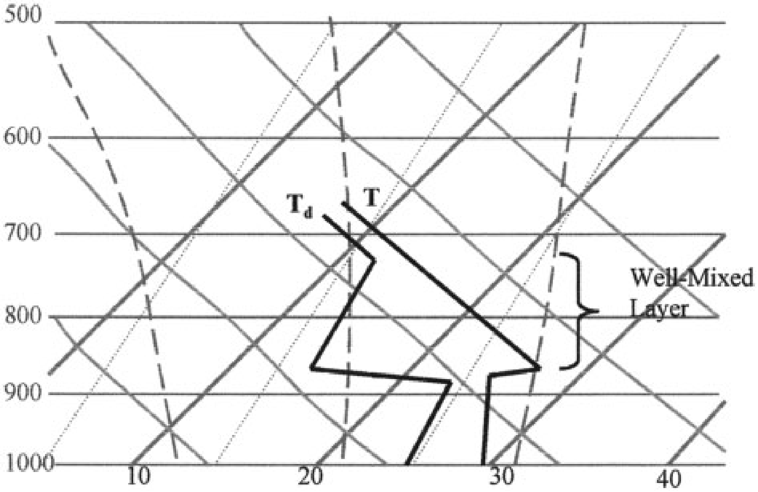


Figure 4 Projection of the p-V-T surface on the p-T and p-V planes. (Sears, 1959).

4 CLAUSIUS–CLAPEYRON EQUATION

We will now express mathematically the relation between the changes in pressure and temperature along an equilibrium curve separating two phases. The assumption of equilibrium between two systems *a* and *b*, requires that at the interface between the systems:

$$\begin{aligned} g_a &= g_b, \\ \mu_a &= \mu_b, \\ T_a &= T_b, \text{ and} \\ u_a &= u_b. \end{aligned}$$

where we will assume that *a* and *b* are of different phases. For equilibrium, we also require infinitesimal changes in conditions along the interface to preserve equilibrium. Hence,

$$\begin{aligned} dg_a &= dg_b, \\ d\mu_a &= d\mu_b, \\ dT_a &= dT_b, \text{ and} \\ du_a &= du_b. \end{aligned} \tag{9}$$

The Gibbs free energy was defined to be

$$g = u + p\alpha - Ts, \tag{10}$$

$$= h - Ts. \tag{11}$$

By virtue of Eq. (11),

$$\begin{aligned} dg_a &= (du_a) - s_a dT + (-T ds_a + p d\alpha_a) + \alpha_a dp \\ &= (du_b) - s_b dT + (-T ds_b + p d\alpha_b) + \alpha_b dp = dg_b \end{aligned}$$

where the terms in parenthesis drop out [because of Eq. (9) and the first law] to give:

$$\begin{aligned} (s_b - s_a)\, dT &= (\alpha_b - \alpha_a)\, dp, \text{ and} \\ \frac{s_b - s_a}{\alpha_b - \alpha_a} &= \frac{dp}{dT} \end{aligned} \tag{12}$$

From Eq. (11) it follows that along the interface,

$$g_b - g_a = h_b - h_a - T(s_b - s_a) = 0, \tag{13}$$

and since, $l_{ab} = h_b - h_a$, we can write

$$l_{ab} = T(s_b - s_a). \tag{14}$$

Hence, we can rewrite Eq. (12) as:

$$\frac{dp}{dT} = \frac{l_{ab}}{T(\alpha_b - \alpha_a)} \tag{15}$$

which is the general form of the Clapeyron equation.

The physical meaning of this equation can be illustrated by the process depicted in Figure 5. Consider the four-step process shown. Beginning at T, P in the lower left point of the cycle we can move to the point $T + dT$ and $p + dp$ in either of the two paths shown. Since g is a state function, the path does not matter for the change in g and hence the change in g along both paths can be equated, yielding the equation. Hence the equation defines how the equilibrium vapor pressure must vary with temperature based on the value of latent heat.

The Clapeyron equation can be applied between any two systems having differing phases to produce expressions for the variations of equilibrium vapor pressure with temperatures. Its integrated solution leads directly to expressions for the value of

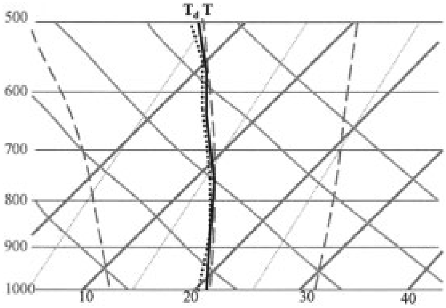


Figure 5 Cycle related to Clapeyron equation (Hess, 1959).

equilibrium vapor pressures with respect to a particular phase change for any temperature.

Equilibrium Between Liquid and Solid Clapeyron Equation

Applying the phase equilibrium to a phase transition between liquid and ice yields

$$\frac{de}{dT} = -\frac{l_{il}}{T(\alpha_i - \alpha_l)} \tag{16}$$

where we are using the symbol e for pressure of the liquid and ice, and we will always take the latent heat l_{il} as the positive definite difference in enthalpy between a liquid and frozen phase where the liquid phase has the greater enthalpy. For a freezing process considered here, the enthalpy change passing from liquid to ice will be negative. Hence, the negative sign appears on the right-hand side (RHS) of Eq. (16). Since the volume of the ice is greater than that of the liquid, the change in vapor pressure with temperature is negative. This is evidenced in Figure 1 as a negative slope to the liquid–ice phase equilibrium curve. Note also that the slope of the ice–liquid curve is very large resulting from the relatively small volume change between liquid and ice phases [denominator of Eq. (16)].

Equilibrium Between Liquid and Vapor: Clausius–Clapeyron Equation

Applying the phase equilibrium to a phase transition between liquid and vapor under equilibrium conditions yields

$$\frac{de_s}{dT} = -\frac{l_{lv}}{T(\alpha_l - \alpha_v)} \tag{17}$$

where e_s is the vapor pressure of the equilibrium state, and the latent heat of vaporization l_{lv} is defined as the absolute value of the difference in enthalpies between the liquid and ice phase. Since the enthalpy of the vapor state is higher than that of the liquid, the negative sign appears in front of Eq. (17). Since $\alpha_v \gg \alpha_l$, we can drop the specific volume of liquid water compared to that of vapor. Hence, applying the equation of state to the vapor specific volume:

$$\frac{d \ln e_s}{dT} = \frac{l_{lv}}{R_v T^2}. \tag{18}$$

This is the Clausius–Clapeyron equation and we will use it for a number of applications. The equation tells us how the saturation vapor pressure varies with temperature. Hence given an observation point, such as the triple point, we can determine e_s at all other points on the phase equilibrium diagram.

Equilibrium Between Ice and Vapor

Similarly, for equilibrium between ice and vapor, we can write

$$\frac{d \ln e_{si}}{dT} = \frac{l_{iv}}{R_v T^2}, \tag{19}$$

where we again take l_{il} to be the absolute value of the difference in enthalpy between the vapor and solid phase of water.

Computation of Saturation Vapor Pressure (e_s and e_{si})

For a precise integration of the Clausius–Clapeyron equation, one must consider the latent heat variation with temperature. To a good first approximation, let temperature be independent of latent heat and then integrate, yielding

$$\ln e_s = -\frac{l_{lv}}{R_v T} + \text{const.} \tag{20}$$

This formulation has assumed latent heat to be independent of temperature. This is a good approximation for sublimation, but not as good for freezing or condensation. We can improve on this by using a more precise formulation for latent heat variation and specific heat variation with temperature with a series expansion:

$$\ln e_s = \frac{1}{R_v}\left[-\frac{l_0}{T} + \Delta\alpha \ln T + \frac{\Delta\beta}{2} T + \frac{\Delta\gamma}{6} T^2 + \cdots\right] + \text{const.} \tag{21}$$

where the integration constant is determined empirically.

We need not achieve this much accuracy for most meteorological considerations, since we cannot measure vapor pressure that precisely anyway. Hence we can make approximations that allow us to calculate e_s or e_{si} with sufficient accuracy. Here are three approximations:

1. Solving the constant at the triple point temperature (T_t), and assuming latent heat constant, we obtain

$$e_s = 10^{(9.4051 - 2354/T)}$$
$$e_{si} = 10^{(10.5553 - 2667/T)}$$

2. Assuming heat capacity constant and retaining two terms in a series, we get *Magnus's formula (liquid–vapor only)*:

$$e_s = 10^{((-2937.4/T) - 4.9283 \log T + 23.5518)} \tag{22}$$

TABLE 3 Constants for Teten's Formula

	Water	Ice
a	17.2693882	21.8745584
b	35.86	7.66

This has been simplified by Tetens (1930) and later by Murray (1966) to be easily computed from the relationship (for pressure in millibars):

$$e_s = 6.1078 \ \exp\left[\frac{a(T - 273.16)}{T - b}\right] \tag{23}$$

where the constants are defined in Table 3.

3. Many atmospheric scientists use the Goff–Gratch (1946) formulation given by:

$$\begin{aligned} e_s = 7.95357242x10^{10} \ \exp\Bigg\{ & -18.1972839\left(\frac{T_s}{T}\right) + 5.02808 \ \ln\left(\frac{T_s}{T}\right) \\ & - 70242.1852 \ \exp\left(\frac{-26.1205253}{T_s/T}\right) \\ & + 58.0691913 \ \exp\left[-8.03945282\left(\frac{T_s}{T}\right)\right]\Bigg\} \end{aligned} \tag{24}$$

for saturation over liquid where $T_s = 373.16\,\mathrm{K}$. For saturation over ice:

$$\begin{aligned} e_s = 5.57185606x10^{10} \ \exp\Bigg\{ & -20.947031\left(\frac{T_0}{T}\right) - 3.56654 \ \ln\left(\frac{T_0}{T}\right) \\ & - \frac{2.01889049}{T_0/T}\Bigg\} \end{aligned} \tag{25}$$

where $T_0 = 273.16\,\mathrm{K}$.

Tables 4 and 5 show how the accuracy of Teten's formula compares to Goff–Gratch as reported by Murray (1966). Both of these forms of saturation vapor pressure are commonly used as a basis to compute saturation vapor pressure.

5 GENERAL THEORY FOR MIXED-PHASE PROCESSES WITHIN OPEN SYSTEMS

Often thermodynamic theory is applied to precipitating cloud systems containing liquid and ice that often are not in equilibrium. In many thermodynamic texts, the simplifying assumption of equilibrium is made that can invalidate the result.

TABLE 4 Saturation Vapor Pressure over Liquid

t (°C)	Goff–Gratch (mb)	Tetens (mb)	Difference (%)
−50	0.06356	0.06078	-1.6×10^{-2}
−45	0.1111	0.1074	-1.6×10^{-2}
−40	0.1891	0.1842	-1.6×10^{-2}
−35	0.3139	0.3078	-1.7×10^{-2}
−30	0.5088	0.5018	-2.1×10^{-2}
−25	0.8070	0.7993	-4.5×10^{-2}
−20	1.2540	1.2462	2.8×10^{-2}
−15	1.9118	1.9046	5.8×10^{-2}
−10	2.8627	2.8571	1.9×10^{-3}
−5	4.2149	4.2117	5.2×10^{-4}
0	6.1078	6.1078	-1.8×10^{-7}
5	8.7192	8.7227	-1.9×10^{-4}
10	12.272	12.2789	-2.2×10^{-4}
15	17.044	17.0523	-1.8×10^{-4}
20	23.373	23.3809	-1.1×10^{-4}
25	31.671	31.6749	-3.7×10^{-5}
30	42.430	42.426	2.5×10^{-5}
35	56.237	56.221	7.1×10^{-5}
40	73.777	73.747	9.5×10^{-5}
45	95.855	95.812	9.7×10^{-5}
50	123.40	123.35	7.5×10^{-5}

Moreover, ice processes are often neglected, despite the existence of ice processes within the vast majority of precipitating clouds. We will adopt the generalized approach to the moist thermodynamics problem, developed by Dutton (1973) and solve for the governing equations for a nonequilibrium three-phase system first and then find the equilibrium solution as a special case.

TABLE 5 Saturation Vapor Pressure over Ice

t (°C)	Goff–Gratch (mb)	Tetens (mb)	Difference (%)
−50	0.03935	0.03817	-9.4×10^{-3}
−45	0.07198	0.07032	-8.8×10^{-3}
−40	0.1283	0.1261	-8.5×10^{-3}
−35	0.2233	0.2205	-8.5×10^{-3}
−30	0.3798	0.3764	-9.2×10^{-3}
−25	0.6323	0.6286	-1.3×10^{-2}
−20	1.032	1.028	1.2×10^{-1}
−15	1.652	1.648	4.1×10^{-3}
−10	2.597	2.595	9.9×10^{-4}
−5	4.015	4.014	1.7×10^{-4}
0	6.107	6.108	-6.3×10^{-5}

Besides the nonequilibrium phase changes, the formation of liquid and ice hydrometeors create a heterogeneous system having components of gas, solids, and liquids, each of which may be moving at different velocities. We must therefore forego our traditional assumption of an adiabatic system and instead consider an *open system*. In doing so, we fix our coordinate system relative to the center of the *dry air parcel*. Although we can assume that vapor will remain stationary relative to the dry air parcel, we must consider movements of the liquid and solid components of the system relative to the parcel, hence implying diabatic effects. We therefore generalize the derivation of Eq. (4) to include these effects. It will be convenient to work in the system of specific energies (energy per mass).

We will assume that our system is composed of several constituents, each of mass m_j. Then the total mass is

$$m = \sum_j m_j. \tag{26}$$

We account for both internal changes (d_i), which result from exchanges between phases internal to the parcel, and external changes (d_e), which result from fluxes of mass into and out of the parcel. The total change in mass is thus

$$dm_j = d_i m_j + d_e m_j. \tag{27}$$

By definition and for mass conservation, we require

$$d_i m = \sum_j d_i m_j = 0, \text{ and}$$

$$d_e m = \sum_j d_e m_j = d_e m_l + d_e m_i.$$

As discussed above, external mass changes are restricted to liquid or ice constituents, while gas constituents are assumed not to move relative to the system. An exception can be made, although not considered here, for small-scale turbulent fluxes relative to the parcel. From our original discussion of the first law, we can write the generalized form of Eq. (4) in a relative mass form as:

$$dH = \delta Q + V\, dP + \sum_j h_j\, dm_j \tag{28}$$

where $h_j = (\partial H_j / \partial m_j)$. We found earlier that the terms involving h_j eventually result in the latent heat term.

We can split the heating function, $\delta Q = Q_i + Q_e$. Here Q_i is the diabatic change due to energy flowing into or out of the system that does not involve a mass flow. Q_e accounts for diabatic heat fluxes resulting from mass fluxes into or out of the system. The diabatic heat source of conduction between the parcel and an outside source would be accounted for in Q_i if no mass back and forth flow into and out of the parcel is not explicitly represented. Even the case of turbulent heat transfer into and out of the parcel would not affect Q_i unless the explicit fluxes of mass in and out were represented. The heat flux associated with the movement of precipitation featuring a different temperature than the parcel, into and out of the parcel, would also be represented by Q_e, as it would appear as a different enthalpy for the precipitation constituent, and be accounted for by the $d_e m_j$ term. As is conventional, we will assume $Q_e = 0$. Now;

$$Q_i = d_i H - V\,dp - \sum_j h_j\,d_i m_j, \text{ and} \tag{29}$$

$$Q_e = d_e H - \sum_j (h_j - h)\,d_e m_j. \tag{30}$$

We can similarly split the enthalpy tendency into its internal and external parts:

$$dH_i = Q_i + V\,dp - \sum_j h_j\,d_i m_j, \text{ and} \tag{31}$$

$$dH_e = -\sum_j (h_j - h)\,d_e m_j. \tag{32}$$

Ignoring the noninteracting free energies, Gibbs' relation for the mixed-phase system is written in "per mass" form as:

$$T\,dS = dH - V\,dp - \sum_j h_j\,dm_j - \sum_j \mu_j\,dm_j, \tag{33}$$

where μ_j is the *chemical potential per mass* rather than per mole as we first defined it. We can now divide the entropy change between internal and external changes, and employing Eq. (29)

$$\begin{aligned} T\,d_i S &= Q_i - \sum_j h_j\,d_i m_j - \sum_j \mu_j\,d_i m_j, \text{ and} \\ T\,d_e S &= -\sum_j \mu_j\,d_e m_j, \end{aligned} \tag{34}$$

for the external change. Since, $\mu_j = g_j = h_j - Ts_j$, and Eq. (30), then

$$T\,d_e S = -\sum_j h_j\,d_e m_j - \sum_j Ts_j\,d_e m_j. \tag{35}$$

Now we apply our results to the particular water–air system. Consider a mixture composed of m_d grams of dry air, m_v grams of vapor, m_l grams of liquid water, and m_i grams of ice. We require

$$\begin{gathered} d_i m_d = d_e m_d = 0, \\ d_i m_v + d_i m_l + d_i m_i = 0, \text{ and} \\ d_e m_v = 0. \end{gathered} \tag{36}$$

The internal enthalpy equation [Eq. (31) and internal entropy Eq. (36)] can then be written for this mixed-phase system:

$$d_iH = Q_i + V\,dp + \sum_j h_j\,d_im_j, \text{ and} \tag{37}$$

$$d_iS = \frac{Q_i}{T} - \frac{\sum_j h_j\,d_im_j}{T} - \frac{\sum_j \mu_j\,d_im_j}{T}. \tag{38}$$

For the air–water system Eq. (38) becomes

$$d_iS = -\frac{l_{iv}}{d_i}m_v + \frac{l_{il}}{T}d_im_i - \frac{a_{iv}}{T}d_im_v + \frac{a_{il}}{T}d_im_i + \frac{Q_i}{T}, \tag{39}$$

where the a_{iv} and a_{il} are the specific affinities of sublimation and melting. They are defined as:

$$a_{iv} \equiv \mu_l - \mu_i,$$
$$a_{il} \equiv \mu_l - \mu_i.$$

The affinity terms take into account the entropy change resulting if the two interacting phases are out of equilibrium. They would be expected to be nonzero, for instance, if rain drops are evaporating in an environment where the air is subsaturated. Under equilibrium conditions, they would vanish. An example is cloud drops growing by condensation in an environment where the vapor pressure is equal to the saturation vapor pressure over liquid. The affinity is the Gibbs free energy available to drive a process and unavailable to perform work. Notice the symmetry between the affinity declarations and the latent heat declarations.

If we apply Eq. (39) to adiabatic melting, condensation, and sublimation, we obtain

$$\begin{aligned} T(s_l - s_i) &= l_{il} + a_{il}, \\ T(s_v - s_l) &= l_{lv} + a_{lv}, \text{ and} \\ T(s_v - s_i) &= l_{iv} + a_{iv}. \end{aligned} \tag{40}$$

We can state that the total entropy is equal to the sum of the entropy in each system component:

$$S = m_ds_d + m_vs_v + m_ls_l + m_is_i, \tag{41}$$

where the dry air entropy is

$$s_d \equiv c_{pd}\,\ln\,T - R_d\,\ln\,p_d. \tag{42}$$

We have required that the only external fluxes are those of precipitation. Hence, we find

$$\begin{aligned} d_iS &= dS - d_eS, \\ &= dS - s_l d_e m_l - s_i d_e m_i, \\ &= d(m_d s_d + m_v s_v) + s_i d_i m_l \\ &\quad + s_i d_i m_i + m_l ds_l + m_i ds_i = \frac{A_{lv}}{T} d_i m_v - \frac{A_{il}}{T} d_i m_i + \frac{Q_i}{T}. \end{aligned} \tag{43}$$

Combining Eqs. (36) and (40) with (4) we obtain:

$$\begin{aligned} d\left(m_d s_d + \frac{m_v l_{lv}}{T} \right) - d_i\left[\frac{w_i l_{il}}{T} + m_v d\left(\frac{A_{lv}}{T} \right) \right] \\ - m_i d\left(\frac{A_{il}}{T} \right) + (m_v + m_l + m_i)\, ds_l = \frac{Q_i}{T}. \end{aligned} \tag{44}$$

If we now assume that the liquid is approximately incompressible, then $ds_l = c_l(dT/T)$. Defining mixing ratio of the jth constituent as:

$$r_j \equiv \frac{m_j}{m_d}, \tag{45}$$

and divide Eq. (44) by m_d. Then we obtain the following general relation:

$$\begin{aligned} d\left\{ c_{pd} \ln T - R_d \ln p_d + \frac{r_v l_{lv}}{T} \right\} - d_i\left(\frac{r_i l_{il}}{T} \right) + r_v d\left(\frac{A_{lv}}{T} \right) - r_i d\left(\frac{A_{il}}{T} \right) \\ + (r_v + r_l + r_i) c_l \frac{dT}{T} = \frac{q_i}{T}. \end{aligned} \tag{46}$$

The differentials account for the interrelationship between temperature change, pressure change, and liquid phase change. Additional diabatic heating tendencies by radiative transfer, molecular diffusion, and eddy diffusion can be accounted for with the internal heating term q_i. Neglected are the effects of chemical reaction on entropy, although the heating effect can be included in the heating term. Generally, these effects are small and can be neglected.

What is included are not only the latent heating effects of the equilibrium reaction but the effects on heating resulting from the entropy change in nonequilibrium reactions. Hence, since entropy change removes energy from that available for work, nonequilibrium heating from phase change modifies the enthalpy change resulting from phase change. The affinity terms account for these effects.

Notice that both affinity terms are inexact differentials, implying that their effect is irreversible. The heat storage term [last term on the left-hand side (LHS)] is also an exact differential if we neglect the variation of c_l and we assume the total water, defined as:

$$r_T = r_v + r_l + r_i, \tag{47}$$

is a constant. If there is a change, i.e., a diabatic loss or gain of moisture by precipitation, then the differential is inexact and the process is irreversible. This term also contains all of the net effects of diabatic fluxes of moisture for systems that are otherwise in equilibrium between phases. The heating term, q_i, is purely diabatic and defines a diabatic thermal forcing on the system such as by radiative transfer or molecular diffusion and turbulence.

Note that, although the first term on the LHS is written as a total derivative, the external derivative of the quantity in brackets is in fact zero. Hence all reversible terms appear as internal derivatives while the reversible terms contain both internal derivatives for moist adiabatic process and external derivatives for diabatic flux terms.

6 ENTHALPY FORM OF THE FIRST–SECOND LAW

We now derive what is perhaps a more commonly used form of Eq. (46). Whereas Eq. (46) is in a form representing entropy change, it can be rewritten as an enthalpy change. To do so we make some manipulations.

First, we make the following assumptions:

1. Neglect the curvature effects of droplets.
2. Neglect solution effects of droplets.

These assumptions are really quite unimportant for the macrosystem of fluid parcels. However, the assumptions would be critical when discussing the microsystem of the droplet itself, since these effects strongly influence nucleation and the early growth of very small droplets. With these assumptions, the chemical potential of the vapor is defined:

$$\mu_l = \mu_o + R_v T \ \ln \ e_v, \tag{48}$$

where e_v is the atmospheric partial pressure of the vapor. The chemical potential of the liquid and ice are defined as the chemical potential of vapor, which would be in equilibrium with a plane pure surface of the liquid or ice and are given by:

$$\mu_l = \mu_o + R_v T \ \ln \ e_s, \tag{49}$$

$$\mu_i = \mu_o + R_v T \ \ln \ e_{si}, \tag{50}$$

where e_s and e_{si} are the saturation vapor pressures of the ice and liquid defined by the temperature of the liquid or ice particle. This temperature need not be the same as the vapor temperature T for this equation.

Combining the equation of state applied to vapor and to dry air, it can be shown that:

$$d \ \ln p(R_d + r_v R_v) = R_d d \ \ln p_d + r_v R_v d \ \ln e_v. \tag{51}$$

Combining Kirchoff's equation (5), with Eqs. (46) to (51) and (18) and (19),

$$c_{pm} d \ \ln T - R_m d \ \ln p + \frac{l_{lv}}{T} dr_v - \frac{l_{il}}{T} d_i r_i = \frac{q_i}{T} \tag{52}$$

where $c_{pm} = c_p + r_v c_{vp} + r_i c_i + r_l c_l$ is the effective heat capacity of moist air and $R_m = R_d + r_v R_v$ is the moist gas constant (not to be confused with the gas constant of moist air).

Although Eq. (52) appears different from Eq. (46), it contains no additional approximations other than the neglect of the curvature and solution effects implicit in the assumed form of chemical potential. It is simpler and easier to solve than the other form because the affinity terms and latent heat storage terms are gone. Note, however, that there are some subtle inconveniences. In particular, each term is an *inexact differential* that means that they will not vanish for a cyclic process. This makes it more difficult to integrate Eq. (52) analytically. Nevertheless, it is a convenient form for applications such as a numerical integration of the temperature change during a thermodynamic process.

Some of the effects of precipitation falling into or out of the system are included in Eq. (52) implicitly. To see this look at the change in vapor. It is a total derivative because only internal changes are allowed. The ice change, on the other hand, is strictly written as an internal change. Hence it is the internal change that implies a phase change, and knowing the ice phase change and liquid phase change, the liquid phase change is implicitly determined since the total of all internal phase changes are zero. Since, by virtue of the assumption that a heterogenous system is composed of multiple homogenous systems, we assumed that hydrometeors falling into or out of the system all have the same enthalpy as those in the system itself, there is no explicit effect on temperature.

This assumption can have important implications. For instance, frontal fog forms when warm rain droplets fall into a cold parcel, hence providing an external flux of heat and moisture through the diabatic movement of the rain droplet relative to the parcel. We neglect this effect implicitly with eq. (52). By requiring $d_e H = \sum_j d_e(m_j h_j) = \sum_j h_j d_e m_j$, we only considered the external changes due to an external flux of water with the same enthalpy of the parcel. The neglect of these effects is consistent with the pseudo-adiabatic assumption that is often made. That assumption assumes condensed water immediately disappears from the system, and so the heat storage effects within the system and for parcels falling into or out of the system can be neglected. So far, the pseudo-adiabatic assumption has only been partially made because we still retain the heat storage terms of the liquid and ice phases within the system. It is unclear whether there is an advantage to retaining them only partially.

7 HUMIDITY VARIABLES

Water vapor, unlike the "dry" gases in the atmosphere exists in varying percentages of the total air mass. Obviously, defining the amount of vapor is critical to understanding the thermodynamics of the water–air system. We have developed a number of variables to define the vapor, liquid, and ice contents of the atmosphere. Below is a list of the variables used to define water content.

Vapor Pressure (e_v)

Vapor pressure represents the partial pressure of the vapor and is measured in pascals. The saturation vapor pressure over a plane surface of pure liquid water is defined to be e_s while the vapor pressure exerted by a plane surface of pure ice is e_{si}.

Mixing Ratio and Specific Humidity

We have already introduced *mixing ratio* to be $r_v = \rho_v/\rho_d$, and employing the equation of state, we can relate mixing ratio to vapor pressure:

$$r_v = \frac{\epsilon e_v}{p - e_v} \tag{53}$$

where $\epsilon \equiv M_d/M_v$. We define *specific humidity* to be the ratio of $q_v = \rho_v/\rho$. It follows that

$$q_v = \frac{r_v}{1 + r_v}$$

Then, similar to Eq. (53) we can write for specific humidity:

$$q_v = \frac{\epsilon e_v}{p} \frac{1}{(1 + (r_v/\epsilon))} \tag{54}$$

To a reasonable approximation, one can show that $e_v \ll p$ and hence:

$$q_v \sim r_v \sim \epsilon \frac{e_v}{p} \tag{55}$$

Relative Humidity

We define relative humidity to be the ratio of the vapor pressure to the vapor pressure exerted by a plain surface of pure water. There is a relative humidity for liquid (H_l) and a relative humidity over ice (H_i):

$$H_l = \frac{e_v}{e_s}$$

$$H_i = \frac{e_v}{e_{si}}$$

which is approximately equal to

$$H_{l'} \sim \frac{r_v}{r_s} \tag{56}$$

$$H_{i'} \sim \frac{r_v}{r_{si}} \tag{57}$$

where r_s and r_{si} are the saturation mixing ratios over liquid and ice.

8 TEMPERATURE VARIABLES

Virtual Temperature (T_v)

We now apply the ideal gas law to the mixture of air and vapor. Applying the total pressure is given by Dalton's law, $p = p_d + e_v$, to the equation of state separately to the vapor and dry air components, we obtain the modified equation of state:

$$p = \rho R_d T \frac{\left(1 + \dfrac{M_d}{M_v} r_v\right)}{1 + r_v} = \rho R_d T_v \tag{58}$$

where $T_v \equiv T(1.0 + 0.61 r_v)$ is the virtual temperature. Note the effect of adding moisture is to increase the virtual temperature over the temperature. Since the total density is a function of the pressure and virtual temperature, the addition of moisture actually lowers the air density. This tends to be opposite to the common perception that humid air is "heavy."

Dew Point Temperature (T_d)

This is the temperature to which moist air must be cooled at constant pressure and vapor mixing ratio in order to become saturated over a plane surface of pure water.

Dew point temperature can be calculated with the following algorithm:

$$T_d = \frac{35.86 \ln e_s - 4947.2325}{\ln e_s - 23.6837} \tag{59}$$

where e_s is in millibars and T_d is in Celsius.

Wet-Bulb Temperature (T_w)

The wet bulb temperature is the temperature that a ventillated thermometer wrapped in a wet cloth will have due to evaporation from the cloth. This will be colder than the air temperature.

The wet-bulb temperature (T_w) may be defined by the isobaric or the adiabatic process.

Isobaric Process T_w is the temperature to which air will cool by evaporating water into it at constant pressure until it is saturated. The latent heat is assumed to be supplied by the air. Note that r_v is not kept constant and so T_w differs from T_d.

When measured by a *psychrometer*, the air is caused to move rapidly past two thermometer bulbs, one dry and the other shrouded by a water-soaked cloth. When thermal equilibrium is reached on the wet bulb, the loss of heat by air flowing past the wet bulb must equal the sensible heat, which is transformed to latent heat. Hence,

$$(T - T_w)(c_p + r_v c_{pv}) - [r_s(T_w, p) - r_v] l_{lv}, \tag{60}$$

where T is the temperature of the air approaching the wet bulb.

Given temperature and mixing ratio, and a suitable relation for obtaining $r_s(T_w)$ and l_{lv}, one can solve for T_w. Alternatively, one can measure T and T_w directly with a psychrometer, and knowing pressure solve for r_v.

Adiabatic Process One can find T_w graphically with the aid of a thermodynamic diagram using the following steps:

1. Begin with pressure and mixing ratio.
2. Reduce pressure dry-adiabatically until saturation is reached to find temperature and pressure of lifting condensation level (LCL).
3. Increase pressure moist-adiabatically from LCL to original pressure.
4. The temperature at the original pressure is T_w.

This is sometimes called the *wet-adiabatic wet-bulb temperature*. It differs at most a few tenths of a degree from the other wet-bulb temperature.

9 ENTROPY VARIABLES FOR MOIST AIR

Potential Temperature for Moist Air (θ)

We earlier derived the Poisson's equation for θ relative to a dry air parcel. Later we pointed out that the conservation of θ in a dry parcel, was equivalent to the conservation of *entropy* for a dry adiabatic, reversible process. Hence, we could show that

$$c_{pd} d \ \ln \ \theta = ds. \tag{61}$$

We now extend our definition of θ to a system containing vapor gas as well as dry air. To derive this, we return to Eq. (52) and assume an adiabatic reversible system with no condensation or sublimation. We then can define θ_m for an adiabatic process:

$$c_{pm} d \ \ln \ \theta_m = d(c_{pm} \ \ln \ T) - d(R_m \ \ln \ p) = \frac{q_i}{T}. \tag{62}$$

Assuming that c_{pm} and R_m are constant for an adiabatic process, we now define θ_m to be

$$\theta_m \equiv T\left(\frac{p_{oo}}{p}\right)^{R_m/c_{pm}}. \tag{63}$$

Now, Eq. (63) can be written as:

$$d\theta_m = \frac{\theta_m}{T}\frac{q_i}{c_{pm}} \tag{64}$$

The term θ_m differs only about 1% from the θ defined earlier. As a result, distinction between θ and θ_m is usually neglected in meteorological applications and the simpler θ is used as the standard form or potential temperature.

The term θ, unlike T, takes into account the natural cooling of air as it rises from pressure change so that one can compare air parcels at different elevations to determine which parcel is warmer or colder when brought to a common elevation. Our intuitive concepts such as "cold air sinks" or "warm air rises" do not work over deep atmospheric depths because of this. Hence it normally gets colder with height, but the low-level air does not start to rise. Our intuitive concepts do work, however, when we use θ as our temperature variable. If warmer θ occurs below colder θ air, the underlying warm air will, in fact, rise spontaneously. Features such as a cold air mass do appear as cold dense flowing masses when viewed with θ instead of temperature. Figure 6 depicts a frontal system flowing southward over the central United States. Note that the cold air dams up against the Rockies to the west. One can see the wavelike feature on the eastern side of the cold air mass representing warm front and cold frontal features.

Using θ in atmospheric science naturally makes the air conceptually easier to understand as a fluid. It also makes formulations depicting the dynamics and evolution of the flow easier and more straightforward. Since air will tend to conserve its potential temperature unless diabatic effects are occurring, air naturally tends to move along θ surfaces. As a consequence, there is a better relationship between

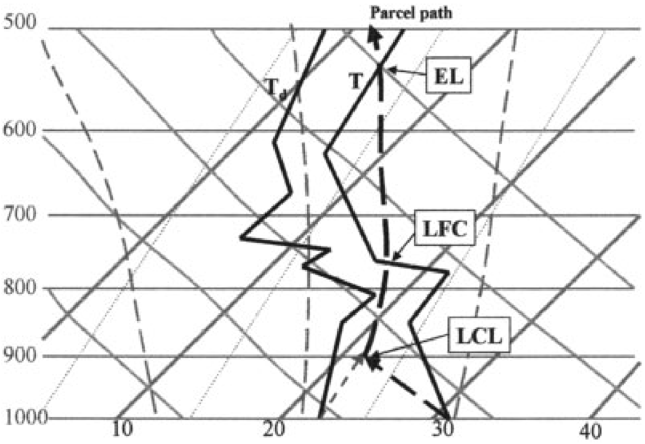


Figure 6 273 K θ surface the March, 1993 Storm of the Century. Note that the cold θ surface, flows along the ground like a heavy fluid and even dams up against the mountains. The wave like feature in the southeastern quadrant are the warm front and cold front created by the cyclone moving up the east coast at the time of this drawing.

air flows along a θ surface than along say a horizontal surface, or even a topographical surface.

Equivalent Potential Temperature (θ_e)

We will now find the *equivalent potential temperature* for an air parcel undergoing phase transition. We again assume an adiabatic, reversible system with multiple phases. Strictly, this is possible only at the triple point temperature, otherwise the ice and liquid cannot both be in equilibrium with vapor at the same time. Hence, we will find diabatic sources to any equivalent potential temperature that we define when both liquid and ice are present at other than the triple point. Equation (46) depicts entropy change for such an irreversible system. Note that there is not a general condition of equilibrium at any temperature to reduce this equation to an exact differential. Since entropy is a state variable, the entropy of the final state is determined by the state parameters of the final state, which themselves are dependent on path. Hence we *cannot* integrate Eq. (46) as we did with (62) to find a moist entropy similar to Eq. (64).

Nevertheless, as previously stated, entropy is an absolute, and according to Eq. (41), it is a sum of the entropies of each component, which for specific entropy is written

$$s = s_d + r_v s_v + r_l s_l + r_i s_i. \tag{65}$$

Employing the definition of specific entropy we can write

$$s_d = c_{pv} \ \ln \ T - R_d \ \ln \ p_d, \tag{66}$$

$$s_v = c_{pv} \ \ln \ T - R_v \ \ln \ e_v, \tag{67}$$

$$s_{lv} = c_{pv} \ \ln \ T - R_v \ \ln \ e_s, \tag{68}$$

$$s_{iv} = c_{pv} \ \ln \ T - R_v \ \ln \ e_{si}. \tag{69}$$

where s_{lv} is the equilibrium entropy of the liquid surface and s_{iv} is the equilibrium entropy over the ice surface, defined by the entropy of vapor at saturation vapor pressure as given by the Clausius–Clapeyron equation. The entropies for pure liquid and ice are, respectively,

$$s_l = c_l \ \ln \ T \tag{70}$$

$$s_i = c_i \ \ln \ T \tag{71}$$

Now substituting Eqs. (40), (47), and (66) to (71) into Eq. (65), we obtain an equation describing the total entropy of a mixed-phase system:

$$s = (c_{pd} + r_T c_l) \ \ln \ T - R_d \ \ln \ p_d + r_v \frac{l_{lv}}{T} + \frac{a_{lv}}{T} - r_v \frac{l_{il}}{T} + \frac{a_{il}}{T}. \tag{72}$$

Although the effects of curvature, solution, chemical changes, and other fairly minor considerations have been neglected, this equation accurately defines the total specific entropy of a parcel. Under equilibrium conditions, and except for the minor omissions described, we would expect s to be invariant for moist adiabatic processes. We can readily identify the differential of Eq. (72) within Eq. (46) and so determine the source of entropy to be

$$ds = d_e\left(\frac{l_{il}}{T} r_i\right) - \frac{A_{lv}}{T} dr_v + \frac{A_{il}}{T} dr_i - c_l \frac{dT}{T} d_e r_T + \frac{q_i}{T}. \tag{73}$$

As expected, the irreversible (diabatic) effect of nonequilibrium is proportional to the amount of phase transition occurring under nonequilibrium conditions in addition to external mass fluxes of water and diabatic heat sources. Hence Eq. (73) describes the change in moist entropy resulting from irreversible processes.

We can now follow the same procedure as we did with Eqs. (61) and (62) to define an equivalent potential temperature (θ_e) for a moist adiabatic process to be

$$c_{pl} d \ \ln \ \theta_e \equiv ds \tag{74}$$

where $c_{pl} = c_{pd} + r_T c_l$. Integrating Eq. (74) and defining the arbitrary constant we get

$$c_{pl} \ \ln \ \theta_e \equiv s + R_d \ \ln \ P_{oo} \tag{75}$$

Substituting Eq. (72) into Eq. (75), we obtain the expression:

$$\theta_e = T\left(\frac{P_{oo}}{p_d}\right)^{R_d/c_{pl}} (H_l)^{[(r_v+r_i)R_v)]/c_{pl}} (H_i)^{-r_v R_v/c_{pl}} e^{[l_{lv}r_v - l_{il}r_i]/c_{pl}T}. \tag{76}$$

Similar to θ_m for dry adiabatic reversible processes in moist air, θ_e is conserved for all moist adiabatic processes carried out at equilibrium. We can relate the changes of θ_e to the irreversible changes in entropy by rewriting Eq. (73) as:

$$\begin{aligned} d\theta_e = \frac{\theta_e}{c_{pl}T}[l_{il}d_e r_i + R_v T \ \ln \ H_l(dr_v - d_i r_i) - R_v T \ \ln \ H_i d_i r_i \\ -c_l T \ \ln \ T d_e(r_l + r_i) + q_i] \end{aligned} \tag{77}$$

Note the inclusion of the external derivative for ice. This demonstrates that for the case of an adiabatic system in equilibrium, and neglect of heat capacity of liquid, θ_e is conserved.

Note that the more vapor in the air the greater the θ_e. It is proportional to the amount of potential energy in the air by the effects of temperature, latent heat, and even geopotential energy combined. Hence θ_e tends to be high on humid warm days and low on cool and/or dry days. It also tends to be greater at high elevations than low elevations for the same temperature because of the lower pressure present at higher elevations. The statement: "The higher the θ_e is at low levels, the greater the potential for strong moist convection," is analogous to the statement: "The higher the θ at low levels, the greater the potential for dry convection." Whereas, we look at a tank of relatively incompressible water and say warm (cold) water rises (sinks), we look at the dry atmosphere and say warm (cold) θ rises (sinks) and look at the moist saturated atmosphere and we state warm (cold) θ_e rises (sinks).

If we ignore diabatic heating effects such as radiation, friction, and heat conduction at the surface, θ_e is nearly perfectly conserved in the atmosphere! θ_e acts like a tracer such that one can identify an air parcel from its θ_e and trace where it came from, even in the midst of moist processes. Figure 7 depicts the 345 K θ_e surface for a supercell thunderstorm viewed from the northeast. A thunderstorm occurs when warm θ_e builds up under a region of colder θ_e, creating the condition where the warm θ_e air will rise if a cloud forms and the moisture within the warm θ_e air condenses. Also, the condition evolves, where the colder θ_e air at middle levels will sink if evaporation takes place. Both of these processes are occurring in Figure 7. The rising currents of warm θ_e air form a "tree trunk"-like structure of the 345 K θ_e, very analogous to rising plume in a lava lamp. To the west (right in Fig. 7) rain is falling into the cold θ_e air at middle levels where it is evaporating causing the cold θ_e

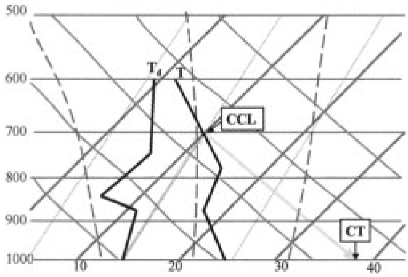


Figure 7 345 K θ_e surface of a model simulation of supercell thunderstorm observed over Montana on 2 August, 1981. The view is from the northeast. Generally θ_e increases with height, i.e., the warmest air is on top. But summertime and tropical conditions allow warm θ_e to be produced near the surface creating a situation where plumes of warm θ_e rise and cold plumes sink as can be seen in the figure. The surface plane is colored with surface temperature where cold air from evaporational cooling is found in the western sector of the storm resulting from cold evaporating downdrafts carrying cold θ_e from middle levels to the surface.

air to sink to the surface. The surface, colored according to its temperature, is cold in the middle of the pool of cold θ_e air sinking to the ground. The power of using equivalent potential temperature to understand the inner structures of complex moist convective weather systems is demonstrated in this figure. Many other weather systems evolve from energy released by phase change including tropical cyclones and some middle latitude cyclones.

10 PSEUDO-ADIABATIC PROCESS (θ_{ep})

Because entropy is a function of pressure, temperature, and liquid and ice water, a multiphase process cannot be represented on a two-dimensional thermodynamic diagram. It is convenient to define a *pseudo-adiabatic* process, where the heat capacity of liquid and ice is neglected.

Omitting the liquid water term (and references to ice) from Eq. (72), and differentiating we can write

$$ds_p = (c_{pd} + r_v c_l) d \ \ln \ T - R_d d \ \ln \ p_d + d\left(\frac{r_v l_{lv}}{T}\right) - d(r_v R_v \ \ln \ H_l) \qquad (78)$$

Since the first term on the RHS now contains r_v instead of r_T, an exact differential is not possible. Bolton integrated numerically and derived the following expression for the *pseudo-equivalent potential temperature* θ_{ep}, as:

$$\theta_{\mathrm{ep}} = T\left(\frac{p_{oo}}{p}\right)^{0.2854(1-0.28r_v)} \exp\left[r_v(1+0.81r_v)\left(\frac{3376}{T_{\mathrm{sat}}} - 2.54\right)\right] \qquad (79)$$

where T_{sat} is a saturation temperature defined as one of:

$$T_{\text{sat}} = \frac{2840}{3.5\ \ln\ T - \ln\ e - 4.805} + 55, \tag{80}$$

or

$$T_{\text{sat}} = \frac{1}{\dfrac{1}{T-55} - \dfrac{\ln(H_l)}{2840}} + 55. \tag{81}$$

T_{sat} can also be determined graphically to be the temperature of the lowest condensation level (LCL).

The pseudo-adiabatic (θ_{ep}) isentropes parallel the *adiabatic wet-bulb potential temperature* (θ_w). The θ_{ep} is related to θ_w by:

$$\theta_{\text{ep}} = \theta_w \exp\left[r_{v'}(1 + 0.81 r_{v'})\left(\frac{3376}{\theta_w} - 2.54\right)\right], \tag{82}$$

where

$$r_v \equiv r_{sl}(p_{oo},\ \theta_w). \tag{83}$$

The pseudo-equivalent potential temperature is interpreted as a two-step process being pseudo-adiabatic ascent to zero pressure followed by dry adiabatic descent to $p_{oo} = 1000\,\text{mb}$. This is how we label the moist adiabats on a thermodynamic diagram.

11 NEGLECTING HEAT STORAGE

A common approximation made to the first–second law is to neglect the heat storage by water. This greatly simplifies the heat capacity and gas constant terms such that $c_{pm} \simeq c_{pd}$ and $R_m \simeq R_d$. This effectively reduces Eq. (52) to:

$$c_{pd} d\ \ln\ T - R_d d\ \ln\ p + \frac{l_{lv}}{T} dr_v - \frac{l_{il}}{T} d_i r_i = \frac{q_i}{T} \tag{84}$$

Generally the effects of this approximation are negligible for meteorological applications. As a result of this approximation, $\theta_m \simeq \theta$ and

$$\theta_e \simeq \theta\ \exp \frac{l_{lv} r_v - l_{il} r_i}{c_{pd} T_{\text{LCL}}} \tag{85}$$

where T_{LCL} is the temperature at the lowest condensation level at which the relative humidity (H_l) is 1. It also follows that:

$$d \ln \theta - \frac{l_{iv}}{c_{pd}T} dr_v + \frac{l_{il}}{c_{pd}} d_i r_i = q_i \tag{86}$$

These approximations are standard in most applications.

12 HYDROSTATIC BALANCE AND HYPSOMETRIC EQUATION

It is sometimes useful to consider the force balance responsible for suspending an air parcel above the surface. The force upward results from the pressure change across the parcel per height change, which gives T net force per parcel volume. The downward force per parcel volume is gravity multiplied by parcel density. Setting these forces equal we obtain the hydrostatic balance:

$$dp = -\rho g \, dz \tag{87}$$

where z is the height coordinate.

Strict hydrostatic balance is not adhered to in the atmosphere, otherwise there would not be vertical acceleration. But the balance is usually very close, and nearly exact over a time average. Assuming the balance, variations in pressure can be converted to variations in height.

We can substitute the equation of state for density to obtain the hypsometric equation:

$$\frac{d \ln p}{dz} = -\frac{g}{R_d T_v} \tag{88}$$

The hypsometric can be integrated to give

$$p = p_0 \exp \frac{-g\Delta z}{R_d \overline{T_v}} \tag{89}$$

where the subscript 0 refers to an initial value, the Δz is the height change from the initial value, and the bar represents an average value during the change from the initial state. This equation is commonly used to find the height change between two pressures or the pressure change between two heights. A common use of this formula is to find the pressure the atmosphere would have at sea level given a pressure measured at a surface location above sea level.

13 DRY AND MOIST ADIABATIC LAPSE RATES (Γ_d AND Γ_m)

A by-product of assuming hydrostatic balance is that we can determine how temperature would change with height for different thermodynamic processes. For an adiabatic process, where $d\theta = 0$, it is easily shown that Eq. (88) requires that the *dry lapse rate* of temperature (Γ_d) be

$$\Gamma_d = -\frac{dT}{dz} = \frac{g}{c_{pd}} = 9.8^\circ\mathrm{C/K}. \tag{90}$$

Similar equations can be derived for a *moist lapse rate* Γ_m occurring in a saturated (w.r.t. liquid) atmosphere. Then applying Eq. (84) for a no-ice system:

$$\frac{dT}{dz} = \frac{\dfrac{g}{c_{pd}}}{1 + \dfrac{L_{lv}}{c_{pd}T}\dfrac{dr_{vs}}{dT}}, \tag{91}$$

where r_{vs} is the saturation mixing ratio, and where we have assumed that the air parcel remains at 0% supersaturation. Notice that the moist adiabatic lapse rate is reduced from the dry adiabatic lapse rate because the latent heating of condensation counters a portion of the expansional cooling. This effect weakens with decreasing temperature as the rate of change of saturation mixing ratio with height decreases.

Dry and Moist Static Energy (h and h_d)

It is sometimes convenient to work with a thermodynamic variable that is directly related to enthalpy. Recall that we looked at this in the first section where we defined gravitational potential energy. It is useful to do the same here, further subdividing the contributing energies. In most applications of static energy, direct measurement of condensate is not possible and so we cannot consider the energies stored in the liquid or ice phases.

The partial work term $-R_d d \ln P$ found in the first law is conceptually difficult. We learned that the term comes from the work performed by expansion that depletes the parcel of total enthalpy. This can either reduce its temperature or slow it down, i.e., reduce its kinetic energy. For large-scale flows, we can generally neglect the effect on the parcel's kinetic energy and assume a nonaccelerating, hydrostatic balance. Then we can employ the hydrostatic Eq. (87) and transform $R_d d \ln P$ into $g\, dz$. Under the hydrostatic assumption, the partial work term is thus shown to be equivalent to a conversion of enthalpy to geopotential energy.

Ignoring the effects of ice, and diabatic processes, Eq. (87) then is written:

$$d(c_{pd}\, dT - g\, dz + l_{lv}\, dr_v) = q. \tag{92}$$

We now define the dry and moist static energies to be

$$h_m \equiv c_{pd}T - g\ dz + lr_v \text{ and}$$
$$h \equiv c_{pd}T - g\ dz,$$

respectively. Notice that the dry static energy is closely related to θ and the moist static energy is closely related to θ_e. The three energy storage terms involved are the sensible heat ($c_{pd}T$), the geopotential energy (gz), and the latent heat ($L_{vl}r_v$). It is now convenient to study the total energy budget of a large-scale atmospheric system to and assess conversions between kinetic energy and static energy.

REFERENCES

Dutton, J. (1976). *The Ceaseless Wind*, McGraw-Hill.

Emanuel, K. A. (1994). *Atmospheric Convection*, Oxford University Press.

Hess, S. L. (1959). *Introduction to Theoretical Meteorology*, Holt, Reinhardt, and Winston.

Iribarne, J. V., and W. L. Godson (1973). *Atmospheric Thermodynamics*, D. Reidel Publishing Co.

Sears, F. W. (1953). *Thermodynamics*, Addison-Wesley.

Wallace, J. M., and P. V. Hobbs, (1977). *Atmospheric Science: An Introductory Survey*, Academic Press.

CHAPTER 17

THERMODYNAMIC ANALYSIS IN THE ATMOSPHERE

AMANDA S. ADAMS

1 ATMOSPHERIC THERMODYNAMIC DIAGRAMS

Before computations were made on computers, atmospheric scientists regularly employed graphical techniques for evaluating atmospheric processes of all kinds including radiation processes, dynamical processes, and thermodynamic processes. Over the last few decades, the use of graphical techniques has almost completely disappeared with the exception of thermodynamic diagrams and conserved variable thermodynamic diagrams.

Classical Thermodynamic Diagrams

The thermodynamic diagram is used to graphically display thermodynamic processes that occur in the atmosphere. The diagram's abscissa and ordinate are designed to represent two of the three state variables, usually a pressure function on one and a thermodynamic function on another. Any dry atmospheric state may be plotted. Unfortunately, any moist state cannot be plotted as a unique point since that must depend on the values of r_v, r_l, and r_i. However, vapor content can be inferred by plotting dewpoint temperature, and moist processes can be accounted for by assuming certain characteristics of the moist process such as if the process is pseudo-adiabatic.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

There are three characteristics of a thermodynamic diagram that are of paramount importance. They are:

1. *Area is proportional to the energy of a process or the work done by the process*. An important function of a thermodynamic diagram is to find the energy involved in a process. If the diagram is constructed properly, the energy can be implied, by the area under a curve or the area between two curves representing a process.
2. *Fundamental lines are straight*. Keeping the fundamental lines straight aids in use of the thermodynamic diagram.
3. *Angle between isotherms (T) and isentropes (θ) are to be as large as possible*. One of the major functions of the thermodynamic diagram is to plot an observed environmental sounding and then compare its lapse rate to the dry adiabatic lapse rate. Since small differences are important, the larger the angle between an isotherm and an isentrope, the more these small differences stand out.

Over the years there have been several different thermodynamic diagrams designed. While they all have the same basic function of representing thermodynamic processes, they have a few fundamental differences. Before these differences can be discussed, it is important to have an understanding of the variables and lines found on a thermodynamic diagram.

Dry Adiabatic Lapse Rate and Dry Adiabats

One of the lines on a thermodynamic diagram, known as the dry adiabat, is representative of a constant potential temperature (θ). Following along a dry adiabat the potential temperature will remain constant, while the temperature will cool at the dry adiabatic lapse rate. The dry adiabatic lapse rate represents the rate at which an unsaturated parcel will cool as it rises in the atmosphere, assuming the parcel does not exchange heat or mass with the air around it, hence behaving adiabatically. The dry adiabatic lapse rate can be derived from a combination of the first law of thermodynamics, the ideal gas law, and the hydrostatic approximation (see chapters 15 and 16), given by:

$$-\frac{dT}{dz} = \frac{g}{cp}$$

Thus the dry adiabatic lapse rate for an ascending parcel is simply g/c_p and is approximately equal to 9.8°C/km or 9.8 K/km. The dry adiabat lines on a thermodynamic diagram are beneficial in that the user may ascend an unsaturated parcel along the dry adiabats and simply read off the new temperature.

Moist Adiabatic Lapse Rate and Pseudo-Adiabats

When a parcel is saturated, it does not suffice to use the dry adiabatic lapse rate. A parcel that is saturated will experience condensation of the water vapor in the parcel. As water vapor condenses the process releases latent heat. This release of latent heat keeps the parcel from cooling at the dry adiabatic lapse rate. The parcel cools at a slower rate, which is dependent on temperature. This reduced rate of temperature decrease has a slope approximately parallel to the moist adiabat and exactly parallel to a line of constant equivalent potential temperature. The term pseudo-adiabat is often used to describe the moist adiabat on a thermodynamic diagram. The moist adiabat on a thermodynamic diagram ignores the contribution by ice, which is more difficult to represent because freezing and melting do not typically occur in equilibrium as does condensation. In a pseudo-adiabatic process the liquid water that condenses is assumed to be immediately removed by idealized instantaneous precipitation. Thus the word *pseudo-adiabat* is used since the line on a thermodynamic diagram is not truly moist adiabatic. If ice is neglected, an expression for the moist adiabatic lapse rate can be derived in a manner similar to the dry adiabatic lapse rate, by using the first law for moist adiabatic hydrostatic ascent. The moist adiabatic lapse rate can never be greater than the dry adiabatic lapse rate. At very cold temperatures, around $-40°\mathrm{C}$, the moist adiabatic lapse rate will approach the dry adiabatic lapse rate. In the lower troposphere the moist adiabatic lapse rate may be as small as $5.5°\mathrm{C/km}$. The equation for the moist adiabatic lapse rate is given as:

$$-\frac{dT}{dz}=\frac{g}{cp}\left[\frac{1+\dfrac{L_{vl}r_{sl}}{R_d\,T}}{1+\dfrac{L_{vl}}{R_d}\dfrac{r_{sl}}{T}\dfrac{\varepsilon L_{vl}}{c_pT}}\right]$$

Types of Thermodynamic Diagrams

There are several different thermodynamic diagrams. All the diagrams serve the same basic function; however, there are differences among these diagrams that should be understood before choosing a diagram. The *emagram* was named by Refsdal as an abbreviation for "energy per unit mass diagram" (Hess, 1959). The emagram uses temperature along the abscissa and the log of pressure along the ordinate. The isobars (lines of constant pressure) and the isotherms (lines of constant temperature) are straight and at a right angle to each other. With pressure plotted on a logarithmic scale, the diagram is terminated at 400 mb, well before the tropopause.

On the *tephigram* one coordinate is the natural log of potential temperature (θ) and the other is temperature. Thus, on a tephigram an isobar is a logarithmic curve sloping upward to the right. The slope decreases with increasing temperature. In the range of meteorological observations, the isobars are only gently sloped. When using the tephigram, the diagram is usually rotated so that the isobars are horizontal with decreasing pressure upward. The pseudo-adiabats are quite curved, but lines of

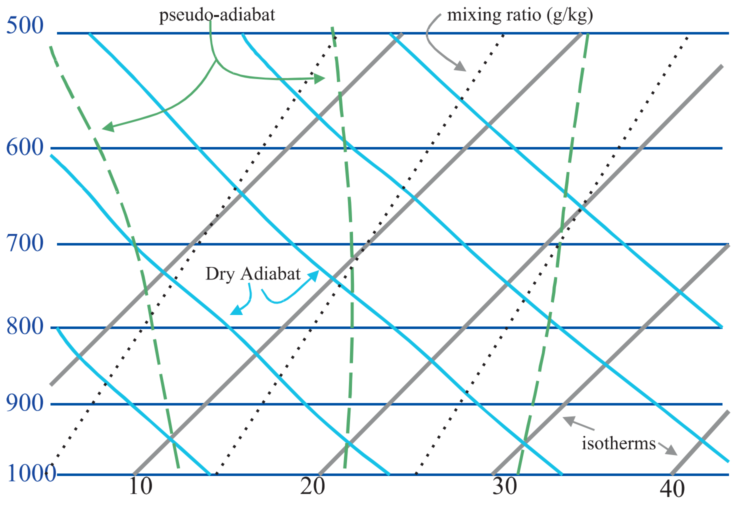


Figure 1 Skew *T*–log *P* diagram. See ftp site for color image.

constant saturation mixing ratio tend to be straight. The angle between the isotherms and isentropes is 90°, making this type of thermodynamic diagram particularly good for looking at variations in stability of an environmental sounding.

The skew *T*–log *P* diagram is now the most commonly used thermodynamic diagram in the meteorological community (Fig. 1). This diagram first suggested by Herlofson in 1947 was intended as a way to increase the angle between the isotherms and isentropes on an emagram (Hess, 1959). The isobars are plotted on a log scale, similar to the emagram. The isotherms slope upward to the right at 45° to the isobars. This angle of the isotherms is what gives this diagram its name, since the temperature is skewed to the right. The isentropes are skewed to the left, at approximately a 45° angle to the isobars. However, the isentropes are not perfectly straight, although they remain nearly straight in the meteorological range typically used. With both the isotherms and isentropes (dry adiabats) at approximately 45° to the isobars, the angle between the isotherms and isentropes is close to 90°. The pseudo-adiabats are distinctly curved on this diagram. In order for the pseudo-adiabats to be straight, the energy–area proportionality of the diagram would need to be sacrificed.

The Stüve diagram is another type of thermodynamic diagram. The Stüve diagram has pressure in mb along the *x* axis and temperature along the *y* axis. This coordinate system allows the dry adiabats (isentropes) to be straight lines. However, this diagram does not have area proportional to energy. The uses of the Stüve diagram are limited when compared to the uses of energy conservation thermodynamic diagrams. This diagram is still occasionally used in the meteorological community, but because of familiarity rather than usefulness.

The characteristics of the thermodynamic diagrams discussed is summarized in the following table:

Attribute	Emagram	Tephigram	Skew T–log P	Stüve
Area α energy	Yes	Yes	Yes	No
T vs. θ angle	45°	90°	Almost 90°	45°
p	Straight	Gently curved	Straight	Straight
T	Straight	Straight	Straight	Straight
θ	Gently curved	Straight	Gently curved	Straight
θ_{ep}	Curved	Curved	Curved	Curved
r_s	Gently curved	Straight	Straight	Straight

With multiple thermodynamic diagrams available it may seem overwhelming to a new user to determine the appropriate diagram for the task. The skew *T*–log *P* diagram and tephigram are the most commonly used thermodynamic diagrams. Their strength lies in the angle between the isentropes and isotherms. The tephigram is most widely used by tropical meteorologists because of the ability of the tephigram to show small changes in stability. In the tropics convective potential may be modified by only small changes in the vertical stability of the environment, and the tephigram best captures small changes in the environmental lapse rate. The skew *T*–log *P* diagram is the most commonly used thermodynamic diagram in the midlatitudes. The straight isobars and the straight up vertical temperature profile make the diagram intuitively easy to comprehend. The ultimate decision of which diagram to use relies on the familiarity of the diagram to the user. While a summer air mass in the middle latitudes may best show instability on a tephigram, a meteorologist may prefer to use the familiar skew *T*–log *P* diagram.

2 ATMOSPHERIC STATIC STABILITY AND APPLICATIONS OF THERMODYNAMIC DIAGRAMS TO THE ATMOSPHERE

Early atmospheric scientists sent up kites with thermometers attached to acquire an understanding of the vertical profile (atmospheric sounding) of the atmosphere. Today, atmospheric scientists use balloons rather than kites and radiosondes rather than thermometers in an effort to understand the vertical structure of the atmosphere. Radiosondes collect temperature, moisture, wind speed, and wind direction at various pressure levels in the atmosphere. The data gained from radiosondes, when displayed on a thermodynamic diagram, provide forecasters and scientific investigators with information that can be used to diagnose not only the current state of the atmosphere but also the recent history of the local atmosphere and the likelihood of future evolution. An air mass originating over a land mass will have inherently different characteristics to its thermodynamic profile than an air mass originating over the oceans. An air mass influenced by motions associated with an approaching weather system will acquire characteristics associated with those circulations. A trained meteorologist can read an atmospheric sounding plotted on a thermodynamic diagram like a book, revealing the detailed recent history of the air mass.

Radiosondes are launched twice daily around the world, at 0000 Universal Time, Coordinated (UTC), also known as Greenwich Mean Time (GMT) or Zulu time (Z), and 1200 UTC. While the data from radiosondes are available only twice daily, when the data is plotted on a thermodynamic diagram, they can be used to study potential changes in the atmosphere. Thermodynamic diagrams can be used to examine the potential warmth of the atmosphere near the surface during the day, the potential cooling of temperature at night, and the potential for fog. The potential for cloudiness can be found due to the movements induced by approaching or withdrawing weather systems or by the development of unstable rising air parcels. When the vertical thermodynamic structure is combined with the vertical wind structure, information on the structure, and severity of possible cumulus clouds can be ascertained. The sounding can also describe the potential for local circulations such as sea breezes, lake effect snow, downslope windstorms, and upslope clouds and precipitation.

An atmospheric sounding plotted on a thermodynamic diagram is one of the most powerful diagnostic tools available to meteorologists. The applications of the thermodynamic diagrams are far too many to be adequately covered in this discussion. This discussion will focus on some of the basic interpretive concepts that can be applied generally to atmospheric soundings plotted on thermodynamic diagrams.

Environmental Structure of the Atmosphere

The sounding plotted on a thermodynamic diagram represents a particular atmospheric state. The state may be observed by radiosonde, satellite derived, or forecasted by a computer atmospheric numerical model. The data plotted on the diagram represent the temperature and dew-point temperature at various pressure levels throughout the atmosphere. Individual observations are plotted, and then the points are connected forming the dew point and temperature curves. The dew-point temperature is always less than or equal to the temperature for any given pressure level. The plotted curves represent the *environmental* temperature and dew-point temperature, as a function of pressure.

The environmental temperature and dew-point temperature curves will divulge much information to a well-trained observer who can deduce where the sounding is from, as well as identify air masses and the processes involved in their formation. Air masses originate from specific source regions and carry the characteristics of that source region.

Arctic Air Mass Air masses that originate near one of the poles have a nearly isothermal temperature profile. In the polar regions, the main process driving the temperature profile is radiational cooling. Light winds allow very little vertical mixing of the air. When coupled with the lack of solar radiation for much of the year, radiational cooling of the surface is the remaining process. Initially the air near Earth’s surface cools faster. The rate at which energy is emitted is dependent on temperature to the fourth power (Stefan–Boltzmann equation), and therefore the warmer air above will then cool faster than the colder air at the surface. This results in the atmosphere moving toward a constant temperature, and this result is

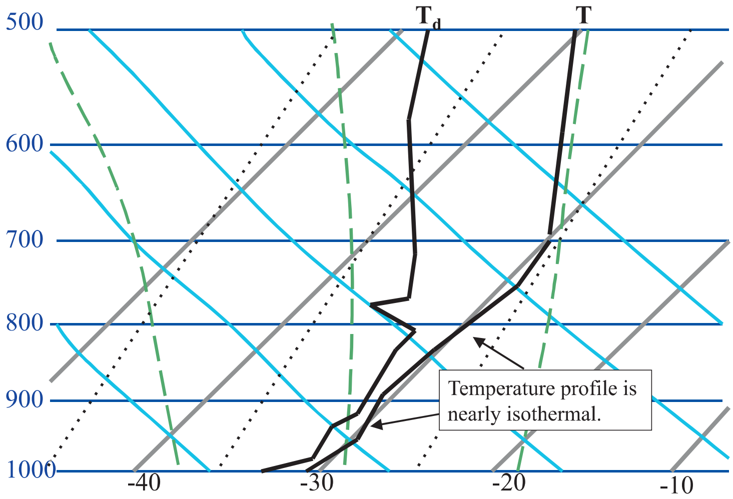


Figure 2 Characteristic arctic air mass. See ftp site for color image.

observable on a skew-*T* diagram by the isothermal profile that begins at the surface (Fig. 2). The tropopause is found at a much lower height near the poles. An arctic sounding will illustrate the low tropopause. Both maritime and continental arctic air masses will have an isothermal temperature profile; the difference is found in the degree of saturation of the air.

Marine Tropical Air Mass In a marine tropical environment (Fig. 3) the temperature profile is nearly moist adiabatic (parallels the pseudo-adiabats). The air in this region is close to saturation near the surface and thus very humid. Above the planetary boundary layer, the air is dominated by sinking motion between cumulus clouds and will tend to be dry except for some moisture mixed from the cumulus clouds. Since saturated air parcels in cumulus updrafts force much of the sinking motion outside clouds, the temperature profile closely parallels the pseudo-adiabat. A dry air maximum resulting from the sinking around 700 mb characterizes the dew-point profile of a marine tropical environment. Below 700 mb surface moisture mixes upward driven by solar heating and solar-powered surface evaporation. The dry air maximum is the feature that truly distinguishes the maritime tropical air mass from other air masses.

Well-Mixed Air Mass Perhaps one of the easiest air masses to identify is one that is well mixed (Fig. 4). If a layer of the atmosphere has sufficient turbulence, the layer will become well mixed with the moisture and potential temperature evenly distributed.

The temperature profile will take on the slope of the adiabatic lapse rate and the dew-point temperature will follow a constant mixing ratio value. Mixing ratio is a measure of the actual amount of water vapor in the air, and in an environment with a

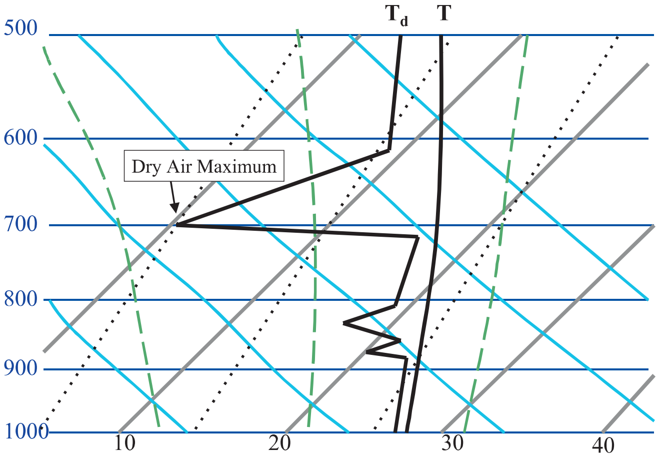


Figure 3 Marine tropical air mass. See ftp site for color image.

lot of turbulent mixing the water vapor will become evenly distributed. Well-mixed layers generally form in a capped boundary layer. However, a well-mixed layer may form in the boundary layer of one region and then be advected into another region. If the region that the well-mixed layer is advected into is at a lower elevation, the well-mixed layer may ride over the existing boundary layer, resulting in an elevated mixed layer. This is commonly observed in the Great Plains of the United States, where a well-mixed layer is advected off the Mexican Plateau or Rocky Mountains, over the boundary layer of the Plains, which is very moist due to flow from the Gulf of Mexico.

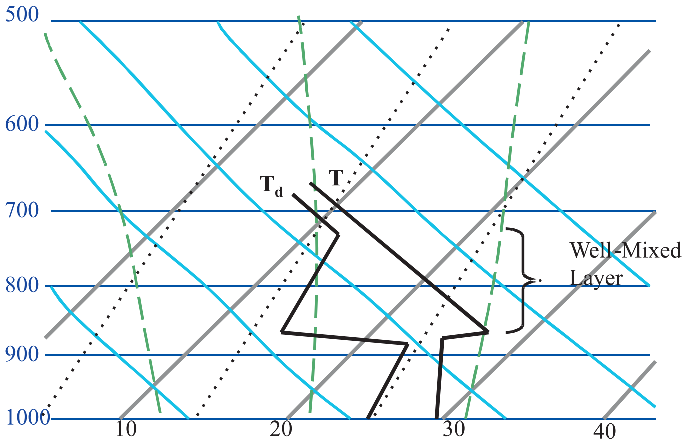


Figure 4 Well-mixed layer in the atmosphere. See ftp site for color image.

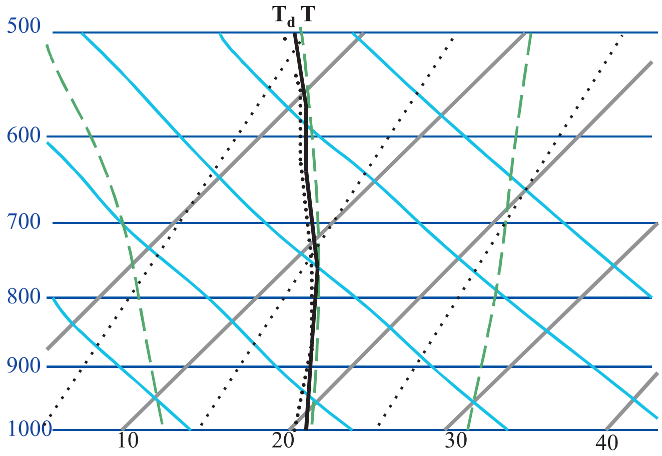


Figure 5 Deep convection. See ftp site for color image.

Deep Convection Layers of the atmosphere that include clouds are identifiable because the temperature and dew-point temperatures will be close together indicating the air is saturated. In the case of deep convection (Fig. 5), the dew-point temperature and temperature, within the updraft, are so close together that they are almost indistinguishable from each other. They are close together throughout a deep layer of the atmosphere, typically the surface up to the tropopause. The temperature profile of the environment will closely parallel the moist adiabat in the case of deep convection. In deep convection, where the atmosphere is basically saturated throughout, all rising parcels will cool at the moist adiabatic lapse rate.

Critical Levels on a Thermodynamic Diagram

The information on a thermodynamic diagram may be used not only to determine where air masses originated from but to also determine what processes the atmosphere may undergo. There are many different critical levels that can be diagnosed from a sounding plotted on a thermodynamic diagram. A thermodynamic diagram may be used to determine where a parcel will become either saturated and/or buoyant. The level at which the parcel becomes saturated depends on not only the moisture content but the processes the atmosphere is undergoing, for this reason there are multiple condensation levels dependent on the process occurring.

Lifting Condensation Level (LCL) This is the pressure level at which a parcel lifted dry adiabatically will become saturated. This level is important for two reasons. First, since the parcel becomes saturated at this point, it represents the height of cloud base. Second, since the parcel is saturated, if it continues to rise above this level, the parcel will cool at the moist adiabatic lapse rate. In a conditionally unstable atmosphere, a parcel is unstable only if saturated. A conditionally unstable parcel lifted to its LCL will typically be cooler than the environment at the LCL. Further lifting will cause the parcel to cool at a slower rate than the environment and eventually will become warmer than the environment. To find the LCL on a thermodynamic diagram first determine from which level you wish to raise a parcel. (Typically the LCL is found for a parcel taken from the surface, but in certain atmospheric conditions, it may be more pertinent for a meteorologist to raise a parcel from a level other than the surface, especially if there is strong convergence at another level.) Initially the parcel is assumed to have the same temperature and moisture content (dew-point temperature) as the environment. Starting from the temperature follow the dry adiabat (constant potential temperature), since the parcel temperature will change at the dry adiabatic lapse rate until it becomes saturated. From the dew-point temperature, follow the intersecting line of constant mixing ratio. The actual grams per kilogram of water vapor in the air is assumed to not change, so long as the parcel rises the mixing ratio remains constant. Note that while the actual amount of water vapor remains constant, the relative humidity will increase. The pressure level at which the mixing ratio value of the parcel intersects the potential temperature of the parcel is the level at which the parcel is saturated. This level is the lifting condensation level (Fig. 6).

Level of Free Convection (LFC) In the atmosphere there are many mechanisms that may force a parcel to rise, and thus initiate convection. The level of free convection is the level above which a parcel becomes buoyant and thus will continue to rise without any additional lifting. Typically, the LFC is found at a height above the LCL. To determine the LFC of a parcel, one need only follow the parcel path until it crosses the environmental temperature profile (Fig. 6). The path the parcel will follow is along the dry adiabat until the LCL is reached, and along the pseudo-adiabat above the LCL. Due to the existence of layers of increased stability (low

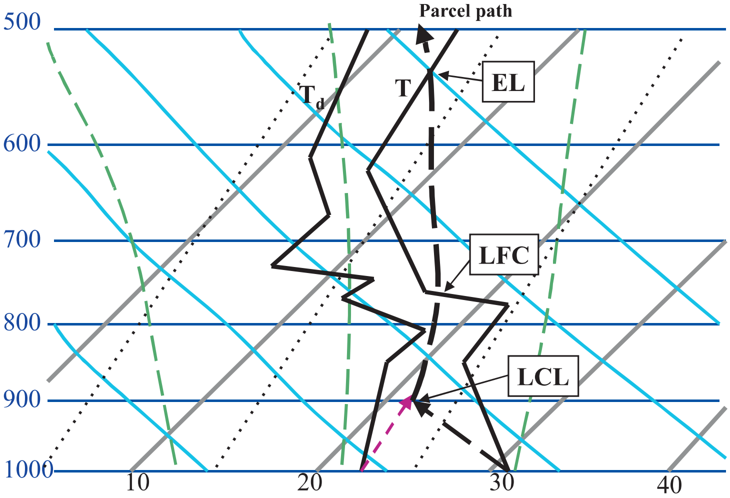


Figure 6 Determining the LCL, LFC, and EL of a parcel. See ftp site for color image.

lapse rates) in the atmosphere, a parcel may have multiple levels of free convection. When using the LFC to determine how much lifting or vertical motion is needed to produce free convection, the LFC at the lowest pressure level (highest above the ground) is typically used.

Equilibrium Level (EL) The equilibrium level is the level at which a parcel buoyant relative to the environment is no longer buoyant. At the equilibrium level, the parcel and the environment have the same temperature. If a parcel rises above the equilibrium level, it will become colder than the environment and thus negatively buoyant. To determine the EL, follow the parcel path from the LFC until it intersects the environmental temperature (Fig. 6). Often the equilibrium level will be found near the beginning of the tropopause. The tropopause is isothermal in nature and thus very stable. On a sounding with multiple levels of free convection, there will also be multiple equilibrium levels. In an extremely stable environment, with no level of free convection, there will also be no equilibrium level. It is noteworthy to mention that while the equilibrium level represents where a parcel is no longer positively buoyant, it does not mean the parcel cannot continue rising. Above the equilibrium level, while the parcel is negatively buoyant, a parcel that reaches this level will have a certain amount of kinetic energy due to its vertical motion. This kinetic energy will allow the parcel to continue rising until the energy associated with the negative buoyancy balances the kinetic energy.

Convective Condensation Level (CCL) The convective condensation level can be used as a proxy for the level at which the base of convective clouds will begin. The CCL differs from the LCL in that the LCL assumes some sort of lifting, and the CCL assumes the parcel is rising due to convection alone. To determine the convective condensation level on a thermodynamic diagram use the dew-point temperature and follow the mixing ratio line that intersects that dew-point temperature until it crosses the environmental temperature (Fig. 7). This represents the level at which a parcel rising solely due to convection will become saturated and form the base of a cloud.

Convective Temperature (CT) The convective temperature is the temperature that must be reached at the surface to form purely convective clouds. If daytime heating warms the surface to the convective temperature, thermals will begin to rise and, upon reaching the CCL, be not only saturated but also positively buoyant. The convective temperature is determined by first locating the CCL. The dry adiabat, which intersects the convective condensation level, is followed down to the surface (Fig. 7). The temperature at the surface represents the convective temperature. If the surface is able to warm to the convective temperature, then the LCL and CCL are the same.

Mixing Condensation Level (MCL) Clouds can form as a result of turbulent mixing rather than lifting or convection. The height at which a cloud will form due to mixing can be determined on a thermodynamic diagram and is referred to as the

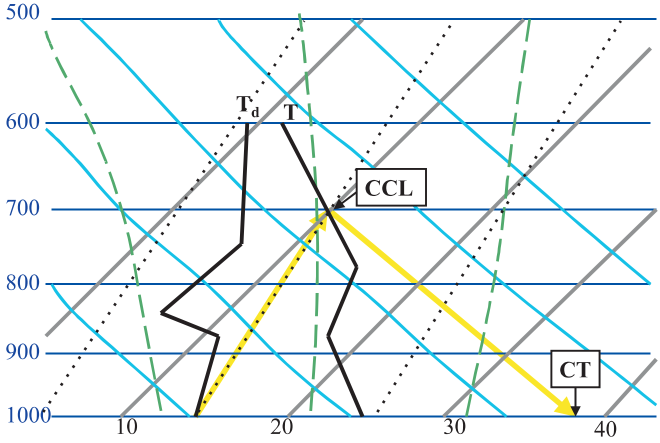


Figure 7 Determining the CCL and CT. See ftp site for color image.

mixing condensation level (Fig. 8). The MCL is determined on a thermodynamic diagram by estimating the average water content (mixing ratio) and the average potential temperature (θ) within the layer. The pressure level at which the average mixing ratio and average potential temperature intersect is the mixing condensation level. The idea is that in a well-mixed layer the amount of moisture will be constant with height, as will the potential temperature. This process often occurs in a shallow strongly capped boundary layer with wind speeds over 10 m/s. The mixing process that forms the cloud is the same process that allows one to "see their breath" on a cold day. While the boundary layer may initially be unsaturated throughout, turbulent mixing can redistribute the moisture and result in saturation near the top of the boundary layer.

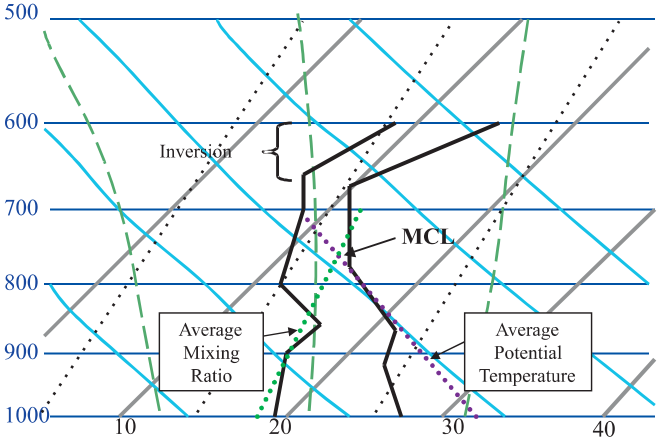


Figure 8 Determining the MCL. See ftp site for color image.

Diagnosing Stability and Parcel Path

The thermodynamic profile of the atmosphere is indicative of the stability of the atmosphere. Understanding the stability of various layers of the troposphere is important in forecasting what processes may occur. To discuss stability, the environment is compared to a parcel. A parcel can be thought of as an entity of the atmosphere with specific temperature and moisture content, which is assumed to not interact with the air around it, and thus undergoes purely adiabatic processes. With this assumption about a parcel, it is easy to ascertain on a thermodynamic diagram the temperature a parcel would have as a result of being either raised or lowered in the atmosphere. This temperature of the parcel at various pressure levels can be drawn on a thermodynamic diagram and is referred to as the parcel path. In general, it is assumed that an unsaturated parcel will change temperature at the dry adiabatic lapse rate, while a saturated parcel will change temperature at the pseudoadiabatic (moist) lapse rate, as it moves up or down from its initial location.

Stability is determined by comparing the temperature of a parcel to the temperature of the environment to which the parcel rises or sinks (Fig. 9). A parcel that is warmer than the environment is unstable and will rise. A parcel colder than the environment is stable and will sink. However, it is important to remember that a parcel will cool as it rises. Therefore, the stability at a single point is not as important as stability throughout the layer. By comparing the environmental lapse rate to the lapse rate of the parcel, stability can be determined. A layer is stable if the environmental lapse rate is greater than the parcel's lapse rate. A layer is unstable if the environmental lapse rate is less than the parcel's lapse rate. The lapse rate of the parcel is dependent on whether the parcel is unsaturated or saturated. An unsaturated parcel will cool at the dry adiabatic lapse rate (DALR), of approximately 9.8°C/km in the troposphere. A saturated parcel will cool at the moist adiabatic lapse rate. The moist adiabatic lapse rate (MALR) is not constant (hence the use of a thermody-

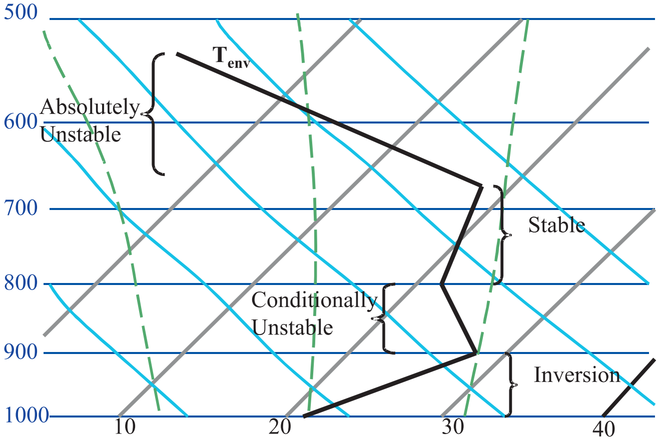


Figure 9 Using the environmental temperature lapse rate to determine stability. See ftp site for color image.

namic diagram for ease), but averages about 6°C/km in the troposphere. The moist adiabatic lapse rate is always less than the dry adiabatic lapse rate. Therefore, if the environmental lapse rate (ELR) is greater than both the moist and dry adiabatic lapse rates, the parcel is *absolutely stable* in that its stability is not dependent on whether the parcel is unsaturated or saturated. A layer is *absolutely unstable* when the environmental lapse rate is less than both the dry and moist adiabatic lapse rates. In the case of an absolutely unstable layer, both dry and saturated parcels will be unstable relative to the environment and will rise. An absolutely unstable layer is rarely observed in a sounding because when a layer is absolutely unstable it will instantly overturn and mix the air in the layer. If an absolutely unstable layer is plotted on a thermodynamic diagram, it is likely that the observation is an error, with the possible exception being when it is plotted in the lowest 50 m near the surface. When the environmental lapse rate is less than the dry adiabatic lapse rate, but greater than the moist adiabatic lapse rate, the atmosphere in that layer is *conditionally unstable*. In a conditionally unstable environment a saturated parcel will be unstable, while an unsaturated parcel is stable. The average environmental lapse rate of the troposphere is conditionally unstable:

DALR > MALR > ELR	Absolutely stable
DALR > ELR > MALR	Conditionally unstable
ELR > DALR > MALR	Absolutely unstable

As mentioned previously in this chapter, one of the benefits of a thermodynamic diagram is that area is proportional to energy. By comparing the parcel path to the environmental temperature profile on one of these diagrams, the *convective available potential energy* (CAPE) and *convective inhibition* (CIN) can be ascertained (Fig. 10). The equation for CAPE is found by integrating from the LFC to the EL to determine the area where the parcel is positively buoyant:

$$\text{CAPE} = g \int_{LFC}^{EL} \frac{\theta_{parcel} - \theta_{env}}{\theta_{\text{env}}} dz$$

If a parcel reaches the LFC, the maximum updraft speed that could develop can be estimated by converting the convective available potential energy into kinetic energy. This gives an equation for the updraft speed of

$$w = \sqrt{2 * \text{CAPE}}$$

The CIN represents the area from the surface to the LFC where the parcel is colder than the environment. Whereas the CAPE provides an idea of how much energy is available once the LFC is reached, the CIN gives an approximation of how much energy must be expended to lift a parcel to the LFC. CAPE and CIN play an important role in determining the possibility for severe weather to occur. Convective inhibition is important for severe weather because it allows CAPE above the bound-

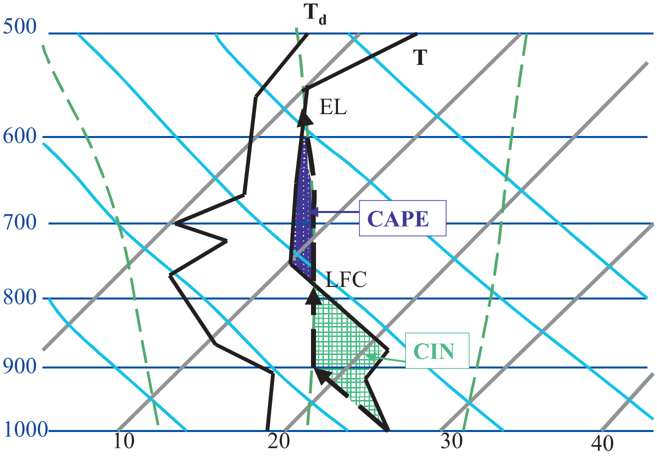


Figure 10 Path of parcel rising from the surface is denoted by the arrows. See ftp site for color image.

ary layer to build. Convective inhibition also keeps parcels from freely rising everywhere, and thus the convection will be more focused along a specific forcing mechanism when significant CIN is present. Large values of CAPE are conducive to severe weather because the more CAPE the more energy a potential storm will have.

Inversions

When the environmental lapse rate is such that the temperature actually increases with height, the layer in which this occurs is called an inversion. An inversion is a very stable layer. The high stability of an inversion can act to “cap” a layer of the atmosphere, preventing parcels from rising above the inversion. Inversions can form as a result of several different processes. While the environmental temperature profile is the same for different types of inversions, the environmental dew-point profile will appear different depending on the processes involved.

Subsidence Inversion The most distinct characteristic of a subsidence inversion (Fig. 11) is the dryness of the inversion. A subsidence inversion occurs as the result of widespread sinking air. The air aloft is initially cooler, and thus has less water vapor. As the air sinks, compressional heating warms it, but no additional water vapor is added. The dew-point temperature will decrease throughout the inversion, with the driest air at the top of the inversion. A subsidence inversion will commonly occur due to the position of the jet streak, causing wide-scale subsidence.

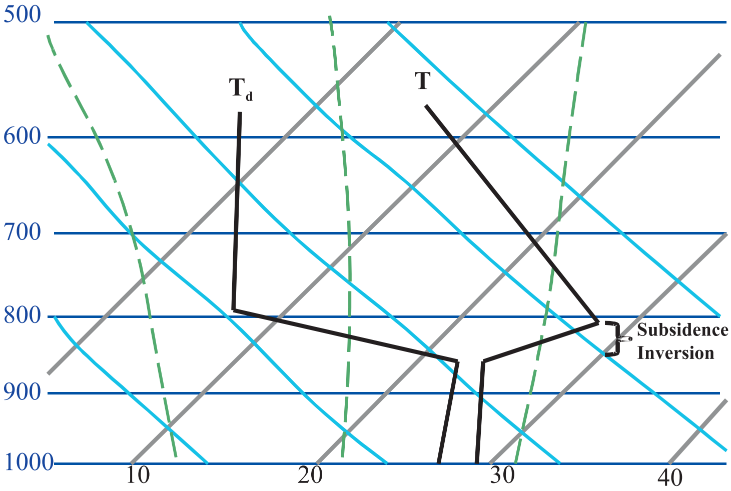


Figure 11 Subsidence inversion. See ftp site for color image.

Radiational Inversion Strong radiational cooling can act to form an inversion. The dew-point temperature in this inversion will typically follow the slope of the temperature profile. However, the dew-point temperature profile does not have to follow the temperature profile slope in the case of a radiational inversion. See Figure 12. An inversion formed by radiational cooling is typically deduced by where in the vertical the inversion occurs. Nighttime cooling of Earth's surface will form an inversion near the surface visible on morning soundings. A radiational inversion can also be observed at the top of a cloud layer due to the cloud top cooling.

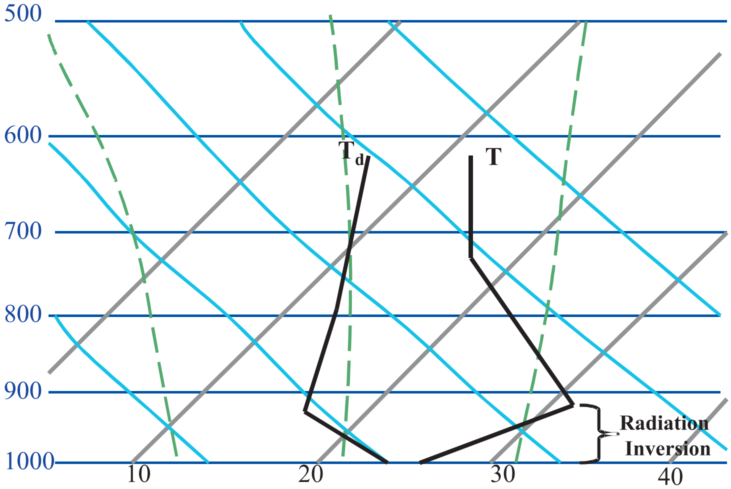


Figure 12 Radiation inversion. See ftp site for color image.

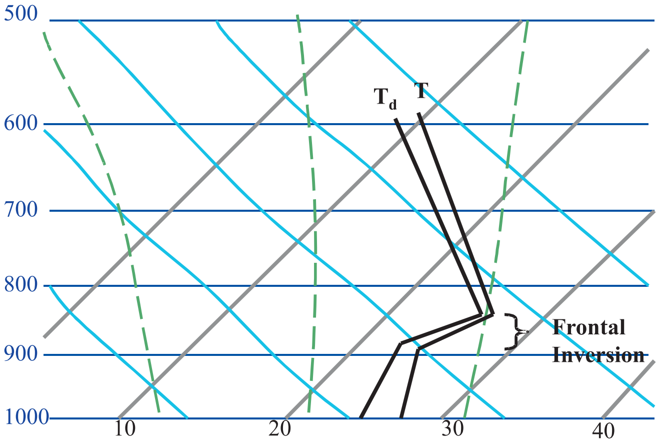


Figure 13 Frontal inversion. See ftp site for color image.

Frontal Inversion In the case of a frontal inversion the dew-point temperature will generally increase with height through the inversion (Djuric, 1994) (Fig. 13). Along a frontal boundary warm, less dense, air will override the colder denser air at the surface. This results in the air aloft being warmer than the surface, and this is depicted in the environmental temperature profile as an inversion. A frontal inversion is differentiated from a radiational inversion in that it is typically much deeper than a radiational inversion. Also, meteorologists with many tools at their disposal will be able to determine the location of the front with surface and upper air analyses.

Inversions, even shallow ones, can play an important role in thermodynamic processes. The height of the LFC, the amount of CAPE, and the amount of convective inhibition a sounding possesses is very dependent on the strength and height of any inversions. The processes involved in the creation of an inversion are also important to the stability of the atmosphere. While a frontal inversion is representative of a stable layer, a trained meteorologist may infer that there is lifting in advance of the front, which could force convection.

BIBLIOGRAPHY

Bohren, C. F., and Albrecht, B. A. (1998). *Atmospheric Thermodynamics*, Oxford University Press, New York, NY.

Djuric', D. (1994). *Weather Analysis*, Prentice Hall Inc., Englewood Cliffs, New Jersey.

Emanuel, Kerry A. (1994) Atmospheric Convection, Oxford University Press, New York, NY.

Hess, Seymour L., (1959) Introduction to Theoretical Meteorology, Krieger Publishing Company, Malabar, FL.

Glossary of Meteorology, (2nd ed.) (2000). American Meteorological Society, Boston, MA.

Wallace, J. M., and P. V. Hobbs (1977). *Atmospheric Science An Introductory Survey*, Academic, San Diego, CA.

CHAPTER 18

MICROPHYSICAL PROCESSES IN THE ATMOSPHERE

ROBERT M. RAUBER

1 INTRODUCTION

Viewed from space, the most distinct features of Earth are its clouds. Most of the world's clouds form within rising air currents forced by atmospheric circulations that can be local or extend over thousands of kilometers. Clouds and fogs also form when air cools to saturation either from radiative cooling or from conduction of heat from atmosphere to Earth's surface. Clouds can also develop as air masses with different thermal and moisture properties mix together. In all cases, clouds have direct influence on atmospheric motions through latent heat exchange, absorption and scattering of solar and terrestrial radiative energy, and redistribution of water vapor. They modulate Earth's climate, reducing the amount of solar radiation reaching Earth and trapping terrestrial radiation. Particle scavenging, chemical reactions, and precipitation processes within clouds continuously alter the trace gas and aerosol composition of the atmosphere. Clouds are an important component of Earth's hydrological cycle.

Clouds are broadly classified according to their altitude, the strength, orientation, and extent of their updrafts, their visual shape, and whether or not they are precipitating. The internationally accepted cloud classification first proposed by Pharmacist Luke Howard in 1803 identifies four broad categories: cumulus (clouds with vertical development), stratus (layer clouds), cirrus (high, fibrous clouds), and nimbus (precipitating clouds), and numerous secondary categories. For example, clouds forming at the crest of waves generated by airflow over mountains are called altocumulus lenticularis because they are often lens shaped. Excellent

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

photographs of clouds with their classifications can be found in the International Cloud Atlas (World Meteorological Organization, 1987).

Clouds are composed of ensembles of water and ice particles, most of which form on aerosol particles that range in diameter from about 10^{-8} to 10^{-5} m. Raindrops typically grow to diameters between 10^{-4} and 5×10^{-3} m, while large hailstones reach diameters of about 5×10^{-2} m. Clouds typically have dimensions between 10^5 and 10^7 m, and individual particle paths through clouds may extend 10^4 to 10^5 m. As water drops and ice particles follow these paths, they can encounter temperatures that can range from 30 to −50°C, pressures from 1050 to 250 mb, and humidity conditions from supersaturation with respect to water to nearly dry air. The complexities of the ever-changing cloud environment and this enormous range of scales must all be considered when investigating the physics governing cloud and precipitation processes.

This chapter explores the fundamental principles and key issues within the discipline of *cloud microphysics*, a discipline specifically concerned with the formation, growth, and fallout of cloud and precipitation particles. An extensive body of cloud microphysics literature has been published over the last century. Limited references are provided in this chapter. Readers who wish to explore individual topics in greater depth, or quickly refer to the scientific literature summarized in this chapter, should consult Pruppacher and Klett (1997), the most authoritative treatise on cloud microphysics currently available.

2 ATMOSPHERIC AEROSOL

The atmosphere consists of a mixture of gases that support a suspension of solid and liquid particles called the atmospheric aerosol. Chemical reactions between gases, aerosol, and cloud particles continually modify the chemical structure, concentration, and distribution of aerosol in the atmosphere. Cloud droplets and ice crystals form on specific aerosol called cloud condensation nuclei and ice nuclei. Once formed, cloud droplets and ice crystals scavenge atmospheric gases and aerosol and provide an environment for additional chemical reactions.

Sources, Sinks, and Formation Mechanisms

Aerosol particles (AP) form during chemical reactions between gases (gas-to-particle conversion), as droplets containing dissolved or solid material evaporate (drop-to-particle conversion), and through mechanical or chemical interactions between the earth or ocean surface and the atmosphere (bulk-to-particle conversion). Gas-to-particle conversion primarily involves the transformation of sulfur oxides, nitrogen oxides, and gaseous hydrocarbons to sulfates, nitrates, and solid hydrocarbon particles. These reactions typically involve water vapor, often require solar radiation, and usually include intermediate states involving sulfuric, nitric, or other acids. Drop-to-particle conversion occurs when cloud drops containing dissolved or suspended material evaporate. Tiny droplets originating at the sea surface when air bubbles break also produce AP when they evaporate. Bulk-to-particle conversion involves wind erosion of rocks and soils and decay of biological material.

Mineral dust from Earth's land surfaces, salt particles from the ocean, organic material and gas emissions from aquatic and terrestrial plants, combustion products from human activities and natural fires, volcanoes, and even meteor bombardment all are sources of atmospheric aerosol. The continents are a much larger source than the oceans, and urbanized areas are a larger source than rural areas. Consequently, the highest concentration of aerosol in the lower atmosphere can normally be found over cities and the lowest concentrations over the open ocean. The concentration of AP decreases rapidly with height, with approximately 80% of AP contained in the lowest kilometer of the atmosphere. Aerosol concentrations are reduced through self-coagulation, precipitation processes, and gravitational settling. The residence time of aerosol in the atmosphere depends on the size and composition of the particles and the elevation where they reside (Fig. 1). Smaller particles ($<0.01\,\mu$m radius) collide rather quickly due to thermal diffusion and coagulate, while large particles ($>10\,\mu$m radius) fall out quickly due to their increased fall velocity. Particles with radii between 0.01 and 10 μm have the longest residence times, typically from 1 to 10 days in the lower troposphere, weeks in the upper troposphere, and months to years at altitudes above the tropopause.

Number Concentration, Mass, and Size Distribution

Aerosol particles range in size from molecular clusters consisting of a few molecules to about 100 μm. Aerosol spectra fall into four size groups: Aitken particles (dry radii $r < 0.1\,\mu$m), large particles ($0.1 < r < 1\,\mu$m), giant particles ($1 < r < 10\,\mu$m), and ultragiant particles ($r > 10\,\mu$m). Whitby (1978) showed that the aerosol size distribution is comprised of three modes, each related to different physical processes (Fig. 2). The *nuclei mode*, which consists of the smallest particles, develops during chemical reactions associated with gas-to-particle conversion. This mode is large in polluted regions and essentially absent in pristine environments. The larger *accumulation mode* forms through coagulation of smaller particles and continued growth of existing particles during vapor condensation and chemical reactions. The accumulation mode develops as a result of aging of the aerosol population. The largest mode, the *coarse particle mode*, is comprised of particles that originate at Earth's surface, either through mechanical disintegration of the solid earth or through evaporation of tiny droplets produced during bubble breakup at the ocean surface. These aerosols differ in chemical composition from those comprising the accumulation mode.

The number concentration of aerosol in the troposphere varies substantially from location to location. In very polluted air over cities, the number concentration may approach $10^6\,\mathrm{cm}^{-3}$. Over land, average concentrations range from 10^3 to $10^5\,\mathrm{cm}^{-3}$ while over the oceans, average concentrations typically range from a few hundred to $10^3\,\mathrm{cm}^{-3}$. The mass concentration varies in a similar way. Over cities, the mass concentration of aerosol typically ranges from about 100 to 200 $\mu\mathrm{g/m}^{-3}$, while over

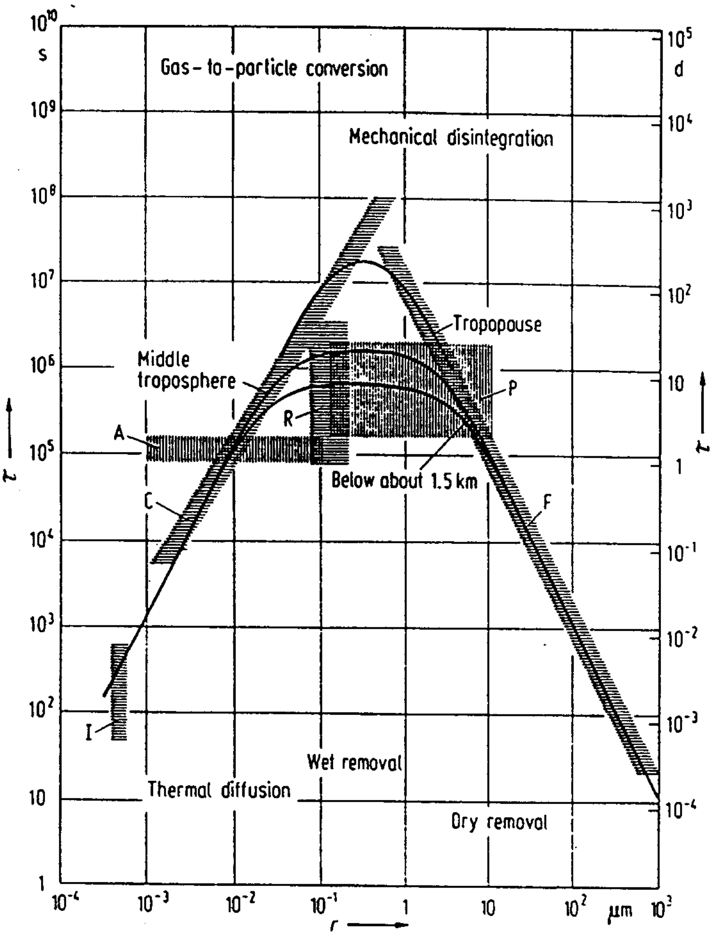


Figure 1 Residence time of aerosol particles as a function of their radius and altitude. I, small ions; A, Aitken particles; C, from thermal diffusion of aerosol particles; R, based on radioactivity data; P, removal by precipitation; F, removal by sedimentation. Solid lines are empirical fits for three atmospheric levels (Jaenicke, 1988).

the oceans, mass concentrations are nearly an order of magnitude smaller (15 to 30 $\mu g/m^{-3}$). The number concentration of Aitken particles decreases exponentially with height between the surface and ~5 km with a scale height of about 1 km (Fig. 3). Between 5 km and the tropopause, the Aitken particle concentration is nearly constant. In the lower stratosphere, to a height of about 20 km, the Aitken particle concentration decreases slowly with a scale height of about 7 km. The number concentration of "large" particles decreases by about 3 orders of magnitude over the 4 lowest kilometers of the atmosphere, and reaches a minimum near the tropopause. Unlike Aitken particles, the concentration of large particles increases

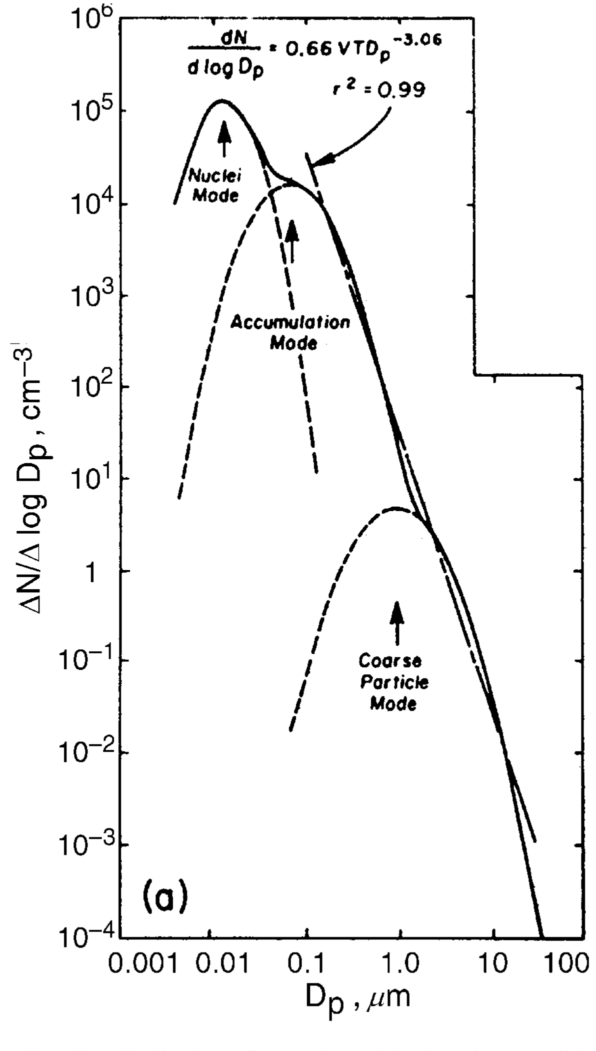


Figure 2 Number size distribution of aerosol particles in urban conditions. N, number of particles; D_p, diameter of particles; V_T, total volume concentration; r, correlation coefficient between the power law in the figure and experimental data (Whitby, 1978).

with height in the stratosphere, with maximum concentrations occurring at an altitude of about 20 km. This stratospheric aerosol layer, called the *Junge layer* after its discoverer, consists primarily of sulfate particles. The number concentration of giant particles, such as sea salt particles, decreases rapidly with height above the surface. The effects of gravitational settling and cloud scavenging reduce the concentration of sea salt aerosol to near insignificance above about 3 km.

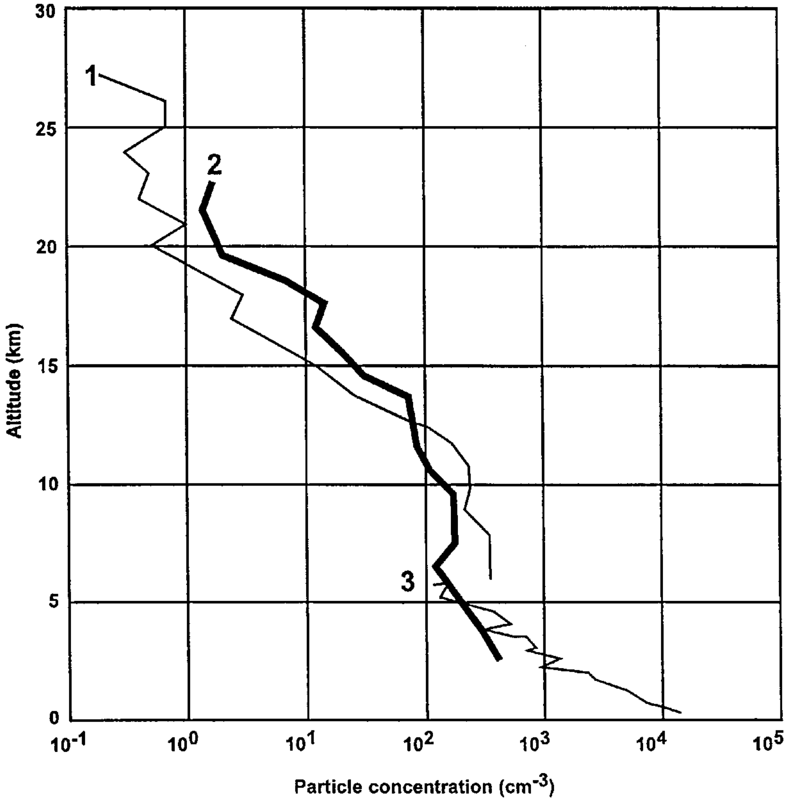


Figure 3 Vertical profiles of Aitken particles: (1) Average concentration per reciprocal cubic centimeter from seven flights over Sioux Falls, South Dakota (44°N); (2) average concentration per cubic centimeter from four flights over Hyderabad, India (17°N); (3) average concentration per cubic centimeter based on flights over the Northeast United States. (Adapted from Junge, 1963.)

Because of the wide range of sizes of aerosol particles, it has been customary to use logarithmic size intervals to characterize the aerosol particle size distribution. If we express the total concentration of aerosol particles with sizes greater than r as

$$N(r) = \int_{\log(r)}^{\infty} n(r) d \log(r) \tag{1}$$

then the number of particles in the size range r, $r + d \log r$ is given by

$$n(r) = -\frac{dN}{d \log(r)} \tag{2}$$

The corresponding volume, surface, and mass distributions, expressed in log radius, are given respectively by:

$$\frac{dV(r)}{d\log(r)}=\frac{4\pi r^3}{3}\frac{dN}{d\log(r)} \tag{3}$$

$$\frac{dS(r)}{d\log(r)}=4\pi r^2\frac{dN}{d\log(r)} \tag{4}$$

$$\frac{dM(r)}{d\log(r)}=\frac{4\pi r^3\rho(r)}{3}\frac{dN}{d\log(r)} \tag{5}$$

where $\rho(r)$ is the particle density. Figure 4 shows examples of aerosol number, surface, and volume distributions from different environments. The accumulation and course particle modes discussed by Whitby (1978) appear in many of the volume distributions. The nuclei mode appears only in the number distributions since these particles contribute negligibly to the volume because of their small

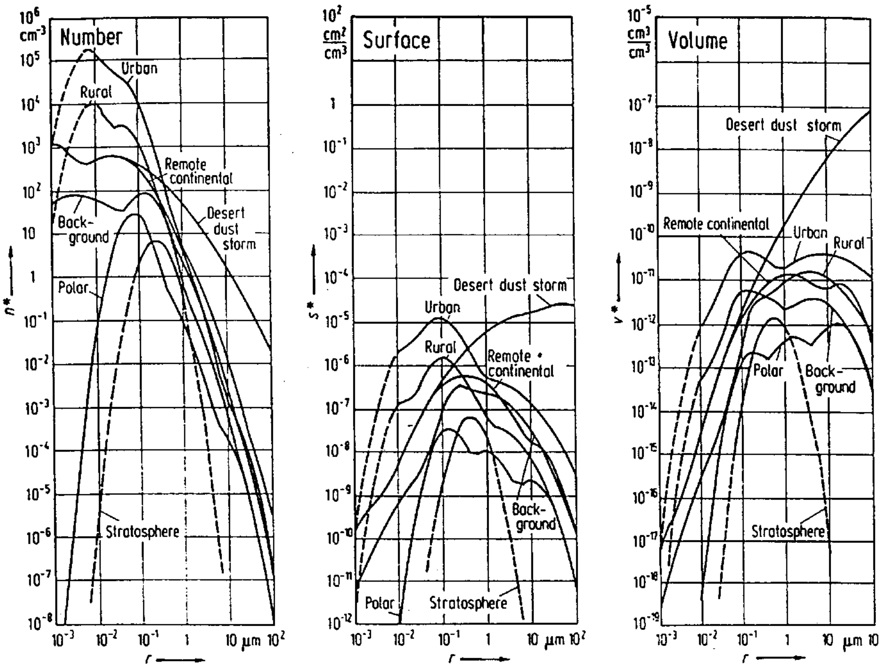


Figure 4 Number [$n^*(r)=dN/d\ \log(r)$], surface [$s^*(r)=dS/d\ \log(r)$], and volume [$v^*(r)=dV/d\ \log(r)$] size distributions for various aerosols (Jaenicke, 1988).

radius. Note that the concentration of aerosol with $r > 0.1$ μm decreases with increasing size such that the aerosol number distribution can be approximated by:

$$\frac{dN}{d\,\log(r)} = Cr^{-\beta} \tag{6}$$

where β generally falls between a value of 3 and 4.

3 FORMATION OF INDIVIDUAL CLOUD DROPLETS AND ICE CRYSTALS IN THE ATMOSPHERE

Cloud droplets and ice crystals form in the atmosphere through a spontaneous process called *nucleation*. *Homogeneous nucleation* is said to occur when a droplet or ice crystal forms from water vapor in the absence of any foreign substance. Homogeneous nucleation of ice crystals also occurs when water droplets freeze without the action of an ice nucleus. The water droplets may be very highly diluted solutions (cloud droplets) or concentrated solutions (haze droplets). *Heterogeneous nucleation* is said to occur when a condensation nucleus is involved in the formation of a water droplet or an ice nucleus is involved in the formation of an ice particle.

Homogeneous Nucleation of Water Droplets in Humid Air

The change in free energy during the homogeneous nucleation of a water droplet in humid air has two contributions: a negative change per unit volume associated with the creation of the new phase and a positive change associated with the creation of the surface interface between the phases. The former is proportional to the volume, while the latter is proportional to the surface area. Because the surface-to-volume ratio is large for small droplets, the surface term dominates until the embryo becomes sufficiently large. The thin surface layer of the drop has unique properties. Molecules within this layer find themselves in an asymmetric force field, attracted only to neighboring molecules located in the interior. This net attractive force causes the surface of the drop to be in a state of tension described by the surface tension, $\sigma_{w,v}$, which has units of energy/unit area. When humid air is cooled beyond saturation, a metastable state develops where the water vapor becomes supersaturated without condensing to form the new phase. The metastable state exists because the path to the lower energy state (a water droplet) initially involves an increase in free energy associated with the droplet surface. Homogeneous nucleation requires the spontaneous collision and aggregation of a sufficient number of water molecules so that the embryonic droplet will be large enough to overcome this "energy barrier." Although statistical mechanics are required to strictly describe this process, classical thermodynamic approaches that assume the embryos are distributed according to the Boltzmann law and are water spheres that have macroscopic densities and surface tensions provide an adequate description. With these assumptions, classical thermodynamics predicts the energy barrier to be

$$\Delta F = -\left(\frac{4\pi r^3 \rho_w}{3M_w}\right) R^* T \ln\left(\frac{e_r}{e_{s,\infty}}\right) + 4\pi r^2 \sigma_{w,v} \tag{7}$$

where ΔF is the Helmholtz free energy change, e_r is the water vapor pressure over a spherically curved surface of radius r, $e_{s,\infty}$ is the saturation vapor pressure over a plane water surface, T is temperature, R^* is the universal gas constant, M_w is the molecular weight of water, ρ_w is the density of water, r is the radius of the embryo, and $\sigma_{w,v}$ is the surface tension at the water–vapor interface. The ratio $e_r/e_{s,\infty}$ is the saturation ratio, $S_{w,v}$, over the droplet surface. The first term on the right is the decrease in free energy associated with the creation of the droplet, and the second term is the increase in free energy associated with the creation of the droplet's surface. Taking the limit $\Delta F/\Delta r = 0$ to determine the equilibrium radius r_{eq} gives the Kelvin equation:

$$S_{w,v} = \exp\left(\frac{2M_w \sigma_{w,v}}{R^* T \rho_w r_{\text{eq}}}\right) \tag{8}$$

Substituting r_{eq} into (7) gives the height of the energy barrier:

$$\Delta F_{\max} = \frac{16\pi M_w^2 \sigma_{w,v}^3}{3[R^* T \rho_w \ln(S_{w,v})]^2} \tag{9}$$

The Kelvin equation relates the equilibrium radius of a droplet to the environmental saturation ratio. A droplet must grow by chance collisions of water molecules to a radius equal to r_{eq} to be stable. Numerical evaluation of the Kelvin equation shows that an embryo consisting of 2.8×10^5 molecules would be in equilibrium at an environmental supersaturation of 10% ($S_{v,w} = 1.10$). A supersaturation of 500% would be required for an embryo consisting of 58 molecules to be stable (Rogers and Yau, 1989). The rate at which embryos of a critical radius form at a given supersaturation, the nucleation rate, J, has been determined using statistical mechanics to be

$$\begin{aligned} J &= \frac{\beta}{\rho_w}\left(\frac{2N_a^3 M_w \sigma_{w,v}}{\pi}\right)^{1/2}\left(\frac{e_{s,\infty}}{R^* T}\right)^2 S_{w,v} \exp\left(\frac{-\Delta F_{\max}}{kT}\right) \\ &\approx 10^{25}\,\text{cm}^{-3}/\text{s} \ \exp\left(\frac{-\Delta F_{\max}}{kT}\right) \end{aligned} \tag{10}$$

where β is the condensation coefficient, N_a is Avogadro's number, and k is the Boltzmann constant. Numerical evaluations of J for different $S_{w,v}$ show that J varies over 115 orders of magnitude as the cloud supersaturation increases from

200 to 600%. The threshold for nucleation is nominally considered to be $J = 1$ droplet cm^{-3}/s. Supersaturations exceeding 550% are required before $J = 0.1$ droplet cm^{-3}/s. Since supersaturations in natural clouds rarely exceed a few percent, we conclude from consideration of J and the Kelvin equation that homogeneous nucleation of water droplets does not occur in Earth's atmosphere.

Heterogeneous Nucleation of Water Droplets

Homogeneous nucleation of cloud droplets directly from vapor cannot occur in natural clouds because supersaturations in clouds rarely exceed a few percent. Cloud droplets form through a heterogeneous nucleation process that involves aerosol particles. In cloud chambers, where supersaturations of several hundred percent can be achieved, nearly all aerosol will initiate droplets. Aerosol that nucleate droplets under these high supersaturation conditions are called *condensation nuclei.* Aerosol that nucleate droplets in the low supersaturation conditions of natural clouds are called *cloud condensation nuclei* (CCN). The number of CCN is a function of cloud supersaturation (s_w) as well as the mass, composition, and concentration of the aerosol population in the region where the clouds are forming. The supersaturation required to activate a cloud droplet is a strong inverse function of particle mass, so only larger aerosol act as CCN. Figure 5 shows the concentration of CCN as a function of supersaturation in maritime and continental environments worldwide. The concentration of CCN typically varies from about 20 to $300\,\text{cm}^{-3}$ in maritime air and from about 200 to $1000\,\text{cm}^{-3}$ in continental air. Typically about 50% of maritime aerosol and 1% of continental aerosol act as CCN at $s_w = 1\%$. The relationship between CCN concentrations and supersaturation (%) in Figure 5 can be approximated as a power law of the form

$$N_{\text{CCN}} = C(S_w)^k \tag{11}$$

Measurements in unpolluted maritime environments have found values of C and k ranging from $25 \leq C \leq 250\,\text{cm}^{-3}$ and $0.3 \leq k \leq 1.4$. Values of C and k measured in continental environments are $600 \leq C \leq 5000$ and $0.4 \leq k \leq 0.9$.

Aerosols that act as CCN are normally hygroscopic and totally or partially water soluble. The chemical composition of CCN depends on the proximity to sources. The primary components of marine CCN are non-sea-salt sulfates (NSS). These aerosols are produced from organic gases such as dimethylsulfide (DMS) and methanesulfonate during gas-to-particle reactions. Measurements suggest that DMS, which is produced primarily by marine algae, is emitted into the atmosphere from the ocean at a global rate of 34 to 56 Tg/yr. The second component of the marine CCN spectra is sea salt. Sea salt is injected into the atmosphere by bubble bursting and spray. Studies suggest that submicron CCN in the marine atmosphere are primarily NSS, while supermicron particles are sea salt, and that sea salt particles contribute only about 1% to the total CCN concentration over the ocean.

The higher CCN concentrations found in continental air masses arise primarily from natural sources rather than anthropogenic activities. The production rate of

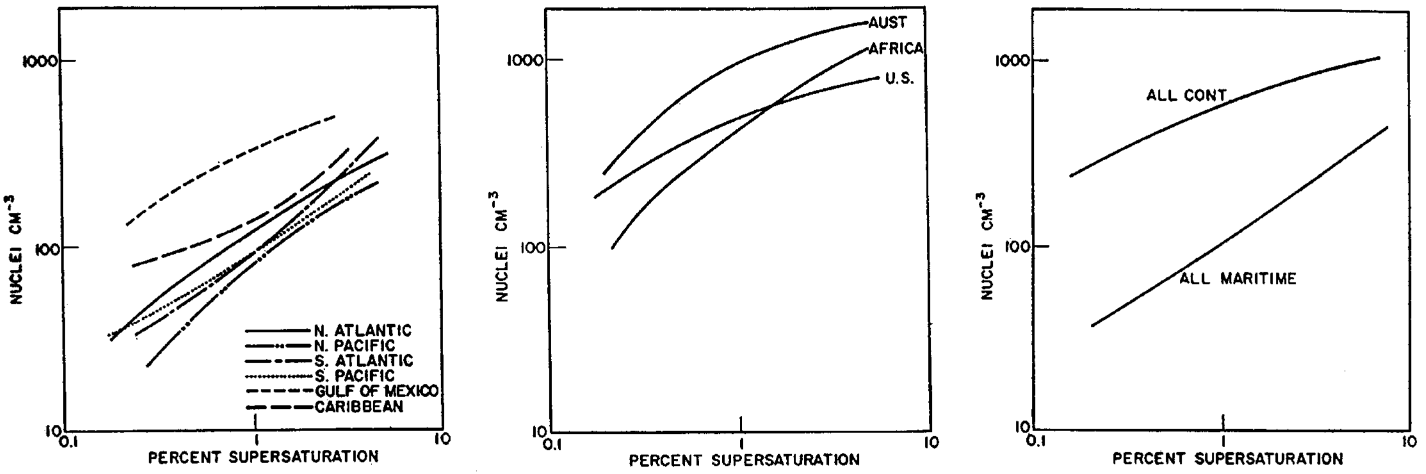


Figure 5 Median concentrations of CCN for (left) different oceanic regions, (center) different continents, and (right) all continents and oceans (Twomey and Wojciechowski, 1969).

CCN due to human activities in the United States has been estimated at about 14% of that from natural sources. Cloud and fog droplets sampled over land frequently contain residue of combustion products and organic material, often of biogenic origin. Forest fires and sugar cane burns, for example, produce large concentrations of CCN. Studies of subequatorial and Saharan air over West Africa suggest that CCN are produced from bush fire smoke, bacterial decomposition of plants and associated sulfur-containing gases, and emission of droplets rich in soluble substances by plants.

The surface of a pure water droplet consists of a layer of water molecules, each with the potential to evaporate and enter the humid air above the surface. Vapor molecules within the humid air also collide with and enter the drop. When a nonvolatile solute is present in the droplet, solute molecules or ions occupy some sites on the droplet surface. Thus, fewer water molecules are available to evaporate into the air. However, vapor molecules can enter the solution at the same rate as before. Equilibrium can only be established when the vapor pressure decreases over the solution so that the rate that water molecules leave the solution equals the rate at which they enter the solution from the air. Raoult's law,

$$\frac{e'_r}{e_r} = \frac{n_0}{n_s + n_0} = \frac{\left(\frac{4}{3}\pi r_{\text{eq}}^3 \rho_w - m_s\right)/M_w}{\left(\frac{4}{3}\pi r_{\text{eq}}^3 \rho_w - m_s\right)/M_w + i m_s/M_s} = \left[1 + \frac{i m_s M_w}{M_s\left(\frac{4}{3}\pi r_{\text{eq}}^3 \rho_w - m_s\right)}\right]^{-1} \tag{12}$$

which can be derived from principles of equilibrium thermodynamics, describes this process formally. Raoult's law states that the vapor pressure over a solution droplet (e'_r) is reduced from that over a pure water droplet (e_r) by an amount equal to the mole fraction of the solvent. In (12), n_0 is the number of molecules of water in the droplet, n_s the number of molecules of solute, m_s is the mass of the solute, and M_s is the molecular weight of the solute. In nature, soluble particles are typically salts or other chemicals that dissociate into ions when they dissolve. For solutions in which the dissolved molecules dissociate, the number of moles of solute, n_s, must be multiplied by the factor i, the degree of ionic dissociation. Unfortunately, the factor i, called the van't Hoff factor, has not been determined for many substances. Other quantities such as the rational activity coefficient, the mean activity coefficient, and the molal or practical osmotic coefficient have been used as alternate expressions to the van't Hoff factor and are tabulated in many chemical reference books (Pruppacher and Klett, 1997).

The Kelvin equation (8) describes the equilibrium vapor pressure over a curved water surface, e_r, and Raoult's law (12) describes the reduction in vapor pressure e'_r/e_r over a solution droplet. Multiplying (8) by (12) to eliminate e_r, gives

$$S_{s,v} = \frac{e'_r}{e_{s,\infty}} = \left[1 + \frac{i m_s M_w}{M_s\left(\frac{4}{3}\pi r_{\text{eq}}^3 \rho_w\right) - m_s}\right]^{-1} \exp\left(\frac{2 M_w \sigma_{w,v}}{R^* T \rho_w r_{\text{eq}}}\right) \tag{13}$$

which describes the equilibrium saturation ratio over a solution droplet. For a sufficiently dilute solution, this equation can be simplified by approximating $e^x \approx 1 + x$ and $(1 + y)^{-1} \approx 1 - y$. Ignoring the small product $x \times y$, and m_s compared to the mass of water, one obtains

$$S_{s,v} = 1 + \frac{a}{r_{\mathrm{eq}}} - \frac{b}{r_{\mathrm{eq}}^3} \tag{14}$$

where $a = 2M_{\mathrm{w}}\sigma_{w,v}/R^*T\rho_w = 3.3 \times 10^{-5}/T$, $b = 3im_sM_w/4\pi M_s\rho_w = 4.3im_s/M_s$ in cgs units with T in kelvins. Equations (13) and (14) are forms of the Köhler equation, first derived by the Swedish meteorologist H. Köhler in the 1920s. Curves 2 to 6 in Figure 6 show solutions of the Köhler equation for droplets containing fixed masses of NaCl and NH_4SO_4, common components of CCN. Curve 1 shows the solution for a pure droplet, the Kelvin equation. For a given $S_{s,v}$, droplets on the left side of the maximum in the Köhler curves are in stable equilibrium. If a droplet in equilibrium on the left side of the curve experiences a small increase in radius due to chance collection of vapor molecules, the droplet would find the vapor pressure around itself less than that required for equilibrium at its new radius. Physically, less water molecules would be striking the droplet from the vapor field than would be evaporating from the droplet. As a result, the droplet would shrink, returning to its position on the equilibrium curve. The opposite would happen if the droplet lost water molecules—it would grow back to its equilibrium size. Note that when $S_{s,v} \leq 1$, all droplets growing from CCN remain on the left side of the curves. The small droplets in stable equilibrium on the left side of the curve are called *haze droplets.* Over some cities, where soluble particles are abundant and relative humidities high, haze droplets can severely restrict visibility.

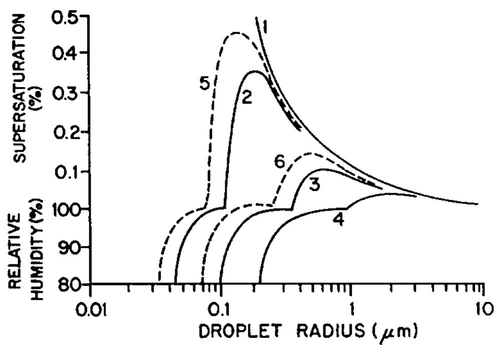


Figure 6 Variations of the relative humidity and supersaturation of air adjacent to droplets of (1) pure water and solution droplets containing the following fixed masses of salt: (2) 10^{-19} kg of NaCl, (3) 10^{-18} kg of NaCl, (4) 10^{-17} kg of NaCl, (5) 10^{-19} kg of $(NH_4)_2SO_4$, and (6) 10^{-18} kg of $(NH_4)_2SO_4$ (Wallace and Hobbs, 1977).

A droplet at the peak of a curve, gaining a few molecules by random collisions, would find the vapor pressure around it greater than that required for equilibrium. More molecules would strike the droplet than evaporate. Vapor would rapidly deposit on the droplet and it would quickly grow into cloud droplet. At the peak, and on the right side of the curves, droplets are in unstable equilibrium. Droplets that reach the peak in their Köhler curves are said to be *activated*. Such droplets are called *cloud droplets*. The peak of a curve (the critical radius) is the transition point between the effect of the solute and drop curvature. When the supersaturation $(S_{s,v} - 1)$ in the atmosphere exceeds a critical value, droplets containing solute of a critical mass will rapidly grow from haze droplets into cloud droplets. As indicated in Figure 6, larger nuclei, which produce stronger solution droplets, are much more likely to become cloud droplets because they require lower supersaturation to activate.

Homogeneous Nucleation of Ice Crystals in Humid Air

Equations analogous to (8) and (10) can be derived for the homogeneous nucleation of ice in humid air by assuming that the embryonic ice particle is spherical. In (8), for example, $S_{w,v}$, $\sigma_{w,v}$, and ρ_w are replaced by $S_{i,v} = e_{r,i}/e_{si,\infty}$, the saturation ratio with respect to ice, $\sigma_{i,v}$, the surface tension at the ice surface, and ρ_i, the density of ice. Numerical evaluation of the Kelvin equation for ice embryos shows that the required supersaturations exceed those for water droplet nucleation. Calculations of J for ice nucleation and the Kelvin equation both show that this process does not occur in the atmosphere.

Homogeneous Nucleation of Ice Particles in Supercooled Water Droplets

Homogeneous nucleation of ice in a water droplet requires that a stable icelike molecular structure form within the droplet through statistical fluctuations in the arrangement of the water molecules. The development of such an ice embryo is favored compared to homogeneous nucleation of ice in humid air because the water molecules in the droplet will be in direct contact with any icelike molecular structures created through statistical fluctuations. Unlike homogeneous nucleation of ice in moist air, the process of homogeneous nucleation in water involves two energy barriers. The first, ΔF_i, is associated with the increase in free energy at the ice crystal–liquid surface interface. The second, $\Delta F'$, exists because energy is required to break the bonds between individual water molecules before they can realign themselves to join the ice embryo. The nucleation rate of ice in water therefore depends on the number of liquid molecules per unit volume per unit time that will contact an ice structure of critical size, the probability that an icelike structure will exist in a droplet, and the probability that one of these liquid molecules will overcome the energy barrier and become free to attach to the ice structure. The rate equation becomes

$$J = 2N_c\left(\frac{\rho_w kT}{\rho_i h}\right)\left(\frac{\sigma_{w,i}}{kT}\right)^{1/2}\exp\left(-\frac{\Delta F'}{R^*T}+\frac{\Delta F_i}{kT}\right) \tag{15}$$

where h is Planck's constant, $\sigma_{w,i}$ is the interface energy per unit area between the ice surface and the water, and N_c is the number of water molecules in contact with a unit surface area of ice. Figure 7 shows measurements of J from several experiments. J increases from about 1 to $1020\,\text{cm}^{-3}/\text{s}$ as the temperature decreases from −32 to −40°C. Until recently, insufficient information was available on the properties of supercooled water and on $\Delta F'$. Pruppacher (1995) showed that incorrect extrapolations of data for these properties to large supercoolings, and a lack of understanding of the behavior of water molecules at large supercoolings led to the disagreement between J calculated from classical theory and experimental data (Fig. 7).

Pruppacher (1995) resolved the discrepancies between the classical nucleation equation, laboratory data, and the results of an earlier molecular theory. Pruppacher noted that because water molecules become increasingly bonded at colder temperatures, $\Delta F'$ might be expected to increase with decreasing temperature. However, earlier cloud chamber experiments found that $\Delta F'$ decreases sharply at temperatures

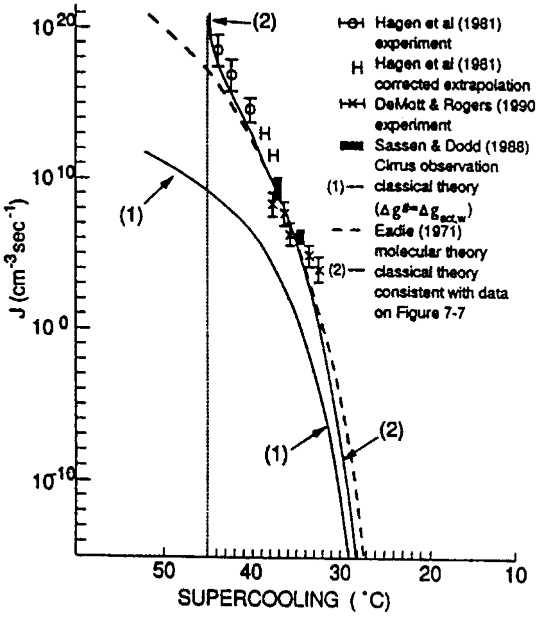


Figure 7 Variation of the rate of homogeneous nucleation in supercooled water. Data are from different experiments. The solid line (1) is from early classical theory. The solid line (2) is from the revision of classical theory discussed by Pruppacher (1995). The dashed line is from molecular theory (Pruppacher, 1995).

colder than −32°C. Pruppacher used nucleation rates measured in laboratory experiments and inferred from field observations, and recent measurements of the physical properties of supercooled water, to calculate $\Delta F'$ and showed using the classical nucleation equation that $\Delta F'$ indeed decreases at temperatures colder than −32°C. This behavior was attributed to the transfer of clusters of water molecules, rather than individual water molecules, across the ice–water interface. With clusters, the only hydrogen bonds that must be broken are those at the periphery of the cluster.

Pruppacher (1995) summarized the experiments of many investigators to determine the homogeneous nucleation temperature threshold for ice in water droplets. He reasoned that (1) the larger the volume of a droplet, the larger is the probability of a density fluctuation in the droplet and the larger the probability that an ice embryo will be produced and (2) the probability for ice formation in a given sized droplet increases with increasing time of exposure of the droplet to a given range of temperatures. Reexamining the available experimental data, he showed that the lowest temperature at which virtually all pure water droplets froze was a function of droplet diameter, with the spread in the data attributable to the different techniques used in the experiments to support the drops (Fig. 8). Figure 8 shows that freezing occurs at −40°C for 1-μm droplets and −35°C for 100-μm droplets. Pruppacher's results apply to cloud droplets, which are nearly pure water droplets, their solute concentrations typically less than 10^{-3} mol/L.

Unactivated haze droplets consist of much stronger solutions whose chemical consistencies depend on the parent CCN. Recent interest in understanding the importance of polar stratospheric clouds in ozone depletion and the role of cirrus in climate change has fueled investigations of the homogeneous nucleation of haze droplets. The homogeneous nucleation temperature in strong solution haze droplets

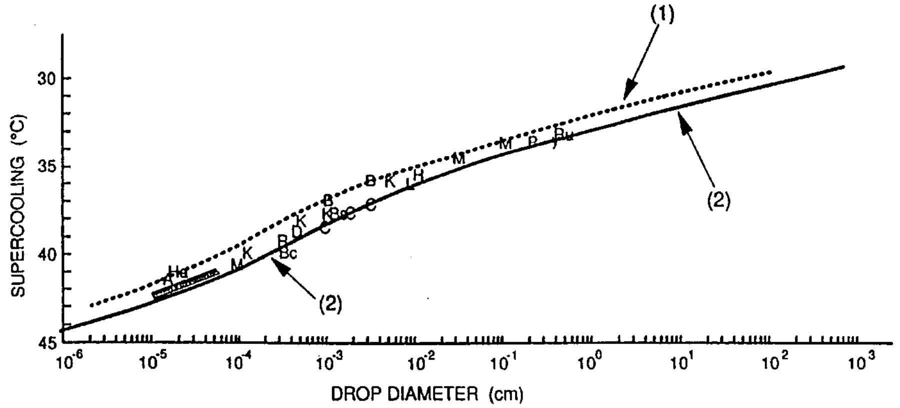


Figure 8 Lowest temperature to which extra pure water drops of a given size and exposed to cooling rates between 1°C/min and 1°C/s have been cooled in various laboratory experiments, indicated by different letters. Lines 1 and 2: Temperature at which 99.99% of a population of uniform-sized water drops freezes when exposed to cooling rates of 1°C/min and 1°C/s, respectively (Pruppacher, 1995).

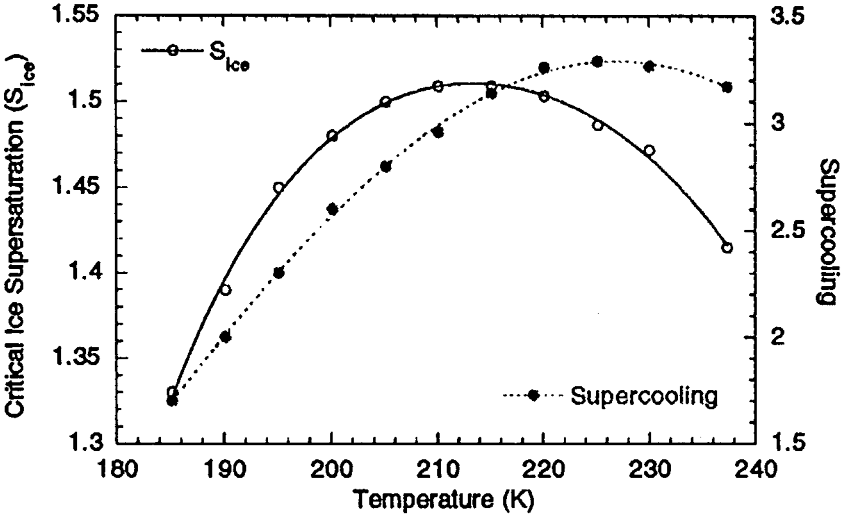


Figure 9 Critical ice supersaturation and the extent of supercooling required to achieve a nucleation rate $J = 1\ \text{cm}^{-3}/\text{s}$ for strong H_2SO_4 solution droplets. The ice supersaturation is defined as the ratio of the ambient water vapor pressure over the ice saturation vapor pressure when $J = 1\ \text{cm}^{-3}/\text{s}$. Supercooling is defined as the critical nucleation temperature minus the temperature of ice that has a vapor pressure equal to the ambient water vapor pressure (Tabazadeh and Jensen, 1997).

is depressed in proportion (nonlinearly) with increased solution molality. The equilibrium size of a solution haze droplet increases and solution molality decreases with increasing ambient relative humidity. As a result, the homogeneous nucleation temperature is depressed most at low relative humidities. Two factors of interest are the ice supersaturation where homogeneous nucleation occurs and the extent to which a sulfate solution can be supercooled above ice saturation over the solution drop before ice nucleation is possible. Figure 9 shows calculations of homogeneous nucleation of droplets containing sulfate aerosols made by Tabazadeh and Jensen (1997). Their results, which correspond to $J = 1\ \text{cm}^{-3}/\text{s}$, show that supercoolings of about 3 K (below the equilibrium freezing temperature of a strong H_2SO_4 solution) and ice supersaturation between 40 and 50% are required for homogeneous nucleation at ambient temperatures between −33 and −63°C (240 and 210 K). These conditions can exist in upper tropospheric clouds, suggesting that the high ice particle concentrations observed in some cirrus clouds are due to homogeneous nucleation of haze droplets.

Heterogeneous Nucleation of Ice Crystals

Ice particles form in the atmosphere through homogeneous nucleation of supercooled water droplets, heterogeneous nucleation processes that involves aerosol particles called *ice nuclei* (IN), and shattering of existing ice particles during collisions or droplet freezing events. Homogeneous nucleation of ice particles in supercooled water is limited to temperatures colder than about −33°C. At warmer temperatures, primary ice particle formation requires ice nuclei.

Most ice nuclei are composed of clay particles such as vermiculite, kaolinite, and illite and enter the atmosphere during wind erosion of soils. Combustion, volcanic eruptions, and airborne microorganisms are also sources of IN. Ice nuclei function in four modes: (1) *deposition* or *sorption-nuclei* adsorb water vapor directly on their surfaces to form ice; (2) *condensation-freezing nuclei* act first as CCN to form drops and then as IN to freeze the drops; (3) *immersion nuclei* become incorporated into a drop at $T > 0°C$ and act to initiate freezing after the drop has been transported into a colder region; and (4) *contact nuclei* initiate freezing of a supercooled drop on contact with the drop surface. The fact that IN can function in these different modes has made their measurement difficult. Instruments designed to measure IN, which include rapid and slow expansion chambers, mixing chambers, thermal precipitation devices, membrane filters, and other devices, typically create conditions favoring one mode, and often only measure the dependence of IN concentration on one variable, such as temperature [see reviews by Vali (1985) and Beard (1992)]. In addition, the time scales over which ice nucleation can occur in natural clouds often differs substantially from those characterizing the measurements. Measurements made with different instruments at workshops in controlled conditions with air samples drawn from the same source have shown considerable scatter. Absolute values of concentrations of ice nuclei should therefore be viewed with some caution.

The concentration of ice nuclei in the atmosphere is highly variable (Fig. 10). In general, ice nuclei concentrations increase by an order of magnitude with each 4°C decrease in temperature but can vary by nearly an order of magnitude at any temperature. The equation

$$N_{\mathrm{IN}} = A \exp(\beta \, \Delta T) \tag{16}$$

where $A = 10^{-5}$ per liter, $\beta = 0.6/°\mathrm{C}$, N_{IN} the number of active ice nuclei per liter active at temperatures warmer than T and $\Delta T = T_0 - T$ is the supercooling in degrees centigrade, reasonably approximates this behavior. Ice nucleation can occur at relative humidities below water saturation, provided that the air is supersaturated with respect to ice. Experiments have shown that at a given temperature, N_{IN} increases with increasing supersaturation with respect to ice (s_i) according to the relationship

$$N_{\mathrm{IN}} = C^k s_i \tag{17}$$

where the values of C and k depend on the air mass. The concentration of IN can undergo orders of magnitude variations at a single location from day to day. Explanations forwarded for these rapid changes include advection of dust from desert windstorms, downward transport of stratospheric IN created from meteor bombardment, IN production from evaporation of cloud and precipitation particles, and preactivation of IN in the cold upper troposphere followed by transport to the surface. Studies of the vertical profile of IN concentrations have found that

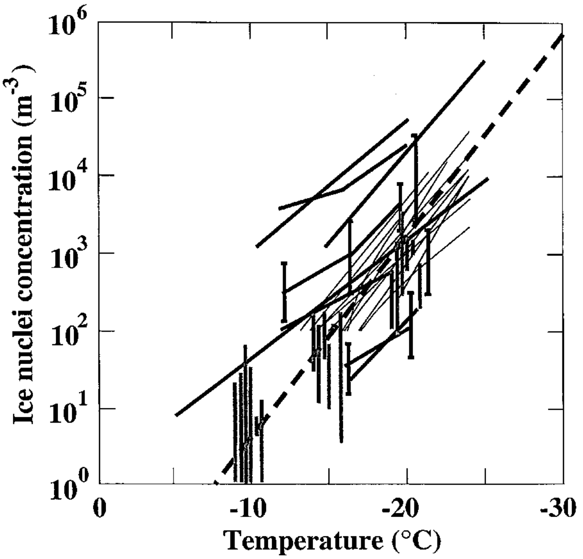


Figure 10 Mean or median worldwide measurements of ice nuclei concentrations from Bigg and Stevenson (1970) (vertical gray bars), compilations of 11 studies by various authors in Götz et al. (1992) (thick lines), and compilations of 10 additional studies by Pruppacher and Klett (1997) (thin lines). Bigg and Stevenson's data at −10, −15 and −20°C are spread over a small temperature range for clarity. The heavy dashed line is $N_{\mathrm{IN}} = 10^{-5}\ \exp(0.6\Delta T)$.

concentrations generally decrease with height in the troposphere above the surface mixed layer, although evidence exists for higher concentrations of IN in the vicinity of the jet stream.

Because ice nuclei must provide a stable solid substrate during the growth of ice embryos, they are normally water insoluble. Nucleation occurs when an ice crystal with a specific lattice structure first forms on a substrate that has a different lattice structure. The probability of an ice nucleation event increases when the lattices of atoms composing the ice crystal and ice nuclei closely align. For the IN atoms and ice atoms to bond, a strain in the bonds must be accommodated between the out-of-place atoms in the IN and ice lattices. The surface free energy of the interface increases in response to increased elastic deformation. The greater the lattice mismatch, the higher the surface free energy and the less likely the ice crystal will form. Particle surfaces normally contain contaminants, cracks, crevices, and particular growth patterns and may possess specific electrical properties due to ions or polar molecules. Certain of these locations are much more effective at adsorbing water molecules onto the surface of the aerosol and enhance the nucleating capability of the substance. Since the water molecule is polar, and the ice lattice is held together by hydrogen bonding, substances that exhibit hydrogen bonds on their surfaces act as more effective nucleants. For this reason, some organics exhibit strong ice nucleating behavior. Ice nuclei are large aerosols, typically having radii larger than 0.1 μm.

The theory governing the nucleation rate, J, of an ice embryo in a saturated vapor, or of an ice embryo in a supercooled water drop, proceeds in a similar way to the theory concerning heterogeneous nucleation of a drop. The theory assumes an ice embryo forms a spherical cap with a contact angle with the surface substrate. Hydrophobic substances have large contact angles and act as poor ice nucleants. The equations are analogous to those used for nucleation of a water droplet, except that all terms applying to water and vapor now apply to ice and vapor, or ice and water. A discussion of this theory, and more complicated extensions of the classical theory, is presented by Pruppacher and Klett (1997). The theory predicts that at -5°C, particles with radii smaller than 0.035 μm and a contact angle of 0 will not be effective as ice nuclei. The threshold is 0.0092 μm at -20°C. Few if any particles will exhibit contact angles of zero, so actual ice nuclei will have to be somewhat larger than these values.

Inside a water droplet, nuclei sizes can be somewhat smaller, with the threshold at least 0.010 and 0.0024 μm at -5 and -20°C, respectively. Experiments have shown that the exact value of the cutoff is also dependent on the chemical composition of the particle and on its mode of action (deposition, freezing, or contact). In the case of deposition nuclei, it also depends on the level of supersaturation with respect to ice. Some organic chemicals have been found to have somewhat smaller sizes and still act as ice nuclei.

There is considerable experimental evidence showing that atmospheric ice nuclei can be preactivated. Preactivation describes a process where an ice nucleus initiates the growth of an ice crystal, is subjected to an environment where complete sublimation occurs, and then is involved in another nucleation event. The particle is said to be preactivated if the second nucleation event occurs at a significantly warmer temperature or lower supersaturation. Ice nuclei can also be deactivated, that is lose their ice nucleating ability. This effect is due to adsorption of certain gases on to the surface of the nucleus. Pollutants such as NO_2, SO_2, and NH_3 have been found to decrease the nucleation ability of certain aerosols. There has also been evidence from laboratory and field experiments that the nucleating ability of silver iodide particles decreases when the aerosols are exposed to sunlight.

The very poor correspondence between ice nucleus measurements and ice particle concentrations in clouds has yet to be adequately explained. Hypotheses forwarded to explain these observations generally focus on more effective contact nucleation, particularly in mixed regions of clouds where evaporation can lead to the formation of giant ice nuclei, and secondary ice particle production, which involves shattering of existing ice particles during collisions or droplet freezing events. The relative importance of each of these mechanisms in real clouds is still uncertain.

4 FORMATION OF RAIN IN WARM CLOUDS

Raindrops form through one of two microphysical paths. The first occurs when ice particles from high, cold regions of clouds fall through the melting level and become raindrops. In cloud physics literature, clouds that support this process are called "cold" clouds. "Warm" clouds are clouds that develop rain in the absence of ice processes. In the tropics, shallow clouds such as trade wind cumulus and stratocumulus produce rain entirely through warm rain processes. Deep tropical convective clouds and summertime midlatitude convection also support active warm rain processes, although melting ice also contributes to rainfall.

The warm rain process occurs in three steps: (1) activation of droplets on CCN, (2) growth by condensation, and (3) growth by collision and coalescence of droplets. The rapid production of warm rain in both maritime and continental clouds remains one of the major unsolved problems in cloud physics. The central problem lies in the transition from step 2 to 3. Theory predicts that growth by condensation will create narrow drop spectra in clouds. Growth by coalescence, on the other hand, requires large and small cloud droplets within droplet spectra. Determining how broad droplet spectra develop in clouds has been a central theme of cloud physics research.

Growth of a Single Droplet by Condensation

Cloud droplets form when the supersaturation in the atmosphere becomes sufficiently large so that haze droplets reach their critical radii and begin to grow in the unstable regime to the right of the Köhler curves (Fig. 6). The equation for the growth of a pure water droplet, first derived by Maxwell in 1890, is obtained by assuming that (1) heat transfer and vapor concentration in the vicinity of the droplet both satisfy the diffusion equation, (2) an energy balance exists at the surface of the droplet such that the latent heat added during condensation balances the diffusion of heat away from the droplet, and that (3) the concentration of droplets in the vapor field is independent of direction outwards from the droplet. Modifications must be made to account for curvature and solute effects. Additionally, one must account for kinetic effects near the drop surface. These include vapor molecules striking the surface of the droplet and rebounding rather than condensing, and the direct heat transfer that occurs at the droplet–vapor interface as water molecules cross the interface between the liquid and gas phases.

Accounting for these effects, the equation describing the diffusional growth of a droplet is

$$r\frac{dr}{dt} = \frac{S_{w,v} - 1 - a/r + b/r^3}{((L_v/R_vT) - 1)(L_v\rho_w/KT^2f(\alpha)) + (\rho_wR_vT/De_{s,\infty}g(\beta))} \tag{18}$$

where a and b are defined in Eq. (14), L_v is the latent heat of vaporization, R_v is the gas constant for water vapor, D is the diffusion coefficient, and K, the thermal conductivity. The quantities $f(\alpha)$ and $g(\beta)$ are given by:

$$f(\alpha) = \frac{r}{r + [(K/\alpha p)(2\pi R_dT)^{1/2}/(C_v + (R_d/2))]} = \frac{r}{r + l_\alpha} \tag{19}$$

$$g(\beta) = \frac{r}{r + (D/\beta)(2\pi/R_vT)^{1/2}} = \frac{r}{r + l_b} \tag{20}$$

In Eqs. (19) and (20), the accommodation coefficient, α, characterizes the transfer of heat by molecules arriving at and leaving the interface between the liquid and vapor phase, and the condensation coefficient, β, is the fraction of vapor molecules hitting the droplet surface that actually condense. In these equations, C_v is the specific heat at constant volume, p is pressure, and R, the gas constant of dry air. The terms l_α and l_β can be considered length scales. As r becomes large such that $r \gg l_\alpha$, l_β, $f(\alpha)$ and $g(\beta)$ approach unity. This essentially is the case by the time a droplet reaches $r = 5$ μm, as is apparent from Figure 11, which compares growth by condensation with and without kinetic effects.

Equation (18) relates the radius of a droplet growing by vapor diffusion to the supersaturation in the atmosphere, and the environmental pressure and temperature. It is clear that the rate of growth of the radius of a droplet is inversely proportional to its size. Solution and curvature effects are most important when the droplets are very small, as indicated by the $1/r$ and $1/r^3$ dependence. Table 1 gives calculated growth times in seconds for droplets with different nuclear masses of NaCl at $T = 273$ K, $P = 900$ mb and a supersaturation of 0.05%. Note that for a droplet with a large nucleus, growth to $r = 20$ μm takes 5900 s or 1.63 h. Growth to a radius of 50 μm takes 41,500 s, or approximately a half-day. A very small drizzle droplet is about 50 μm radius. Individual cloud parcels are unlikely to exist in a supersaturated environment for half a day. Clearly, diffusional growth alone is inadequate to explain the formation of rain in clouds.

Equation (18) is valid for a stationary droplet. As droplets grow, they fall through the air. Air flowing around the droplet causes the field of vapor molecules to interact differently with the droplet surface than for a still droplet. The net result is that the growth rate (or evaporation rate if the relative humidity <100%) of the droplet increases. The effect of ventilation has been studied using both numerical simulations and by making direct measurements in wind tunnels. For droplets with

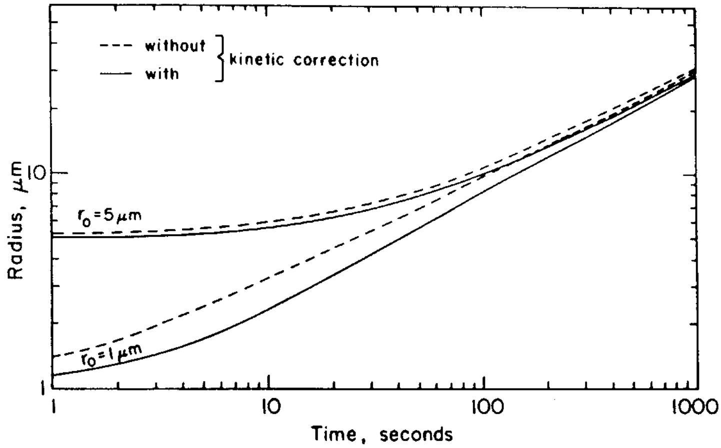


Figure 11 Comparison of condensation growth, with and without kinetic effects, for droplets of initial radii of 1 and 5 μm (Rogers and Yau, 1989).

TABLE 1 Time (s) required for Droplet to Grow from Initial Radius of 0.75 μm to Specified Size at 273 K, Pressure of 900 mb, and Supersaturation $[100(S_{w,v} - 1)] = 0.05$ for Salt Nuclei with 3 Nuclear Masses

Radius (μm)	10^{-14} g	10^{-13} g	10^{-12} g
2	130	7	1
5	1,000	320	62
10	2,700	1,800	870
15	5,200	4,200	2,900
20	8,500	7,400	5,900
25	12,500	11,500	9,700
30	17,500	16,000	14,500
35	23,000	22,000	20,000

From Best (1951).

$r \leq 50$ μm, ventilation can be ignored, but for droplets of greater sizes, the ventilation effect increases with droplet size. In the growth equation, ventilation is accounted for by multiplying the numerator on the right by f_v, the ventilation coefficient.

Growth of Population of Droplets by Condensation

Cloud droplets will activate on CCN when a parcel of air rises through cloud base. The larger, more soluble CCN activate first, followed by the more numerous, but less effective CCN. The new droplets each extract water vapor from the parcel. The maximum supersaturation occurs, and the droplet concentration is determined, when the rate of vapor deposition on the drops equals the rate at which vapor is supplied through cooling of the rising parcel. Above this point in a cloud, the supersaturation remains below its peak value and no additional droplets are activated. The cloud droplet concentration therefore depends on the spectrum of CCN entering the cloud, the updraft strength, and the cloud base temperature.

Equation (18) shows that the growth rate of activated cloud droplets decreases with radius. This equation predicts that the width of a spectrum of different sized droplets narrows with time, since smaller droplets in a rising parcel grow faster than larger droplets. This is illustrated by the measurements and computations reported by Fitzgerald (1972). Fitzgerald found close agreement between cloud droplet spectra measured 200 to 300 m above cloud base with that predicted by condensation theory including kinetic effects (see Fig. 12).

For collisions between cloud droplets to occur, the droplet spectrum must evolve so that droplets possess a range of sizes and fall speeds. Calculations suggest that the collision efficiency of two falling drops only becomes appreciable (>10%) when the radius of the larger collector drop (R) and smaller collected droplet (r) reach $R = 25$ μm and $r = 7.5$ μm. Equation (18) predicts that a droplet with a large nucleus will require 1.6 h to grow to $r = 20$ μm, yet many tropical maritime clouds produce

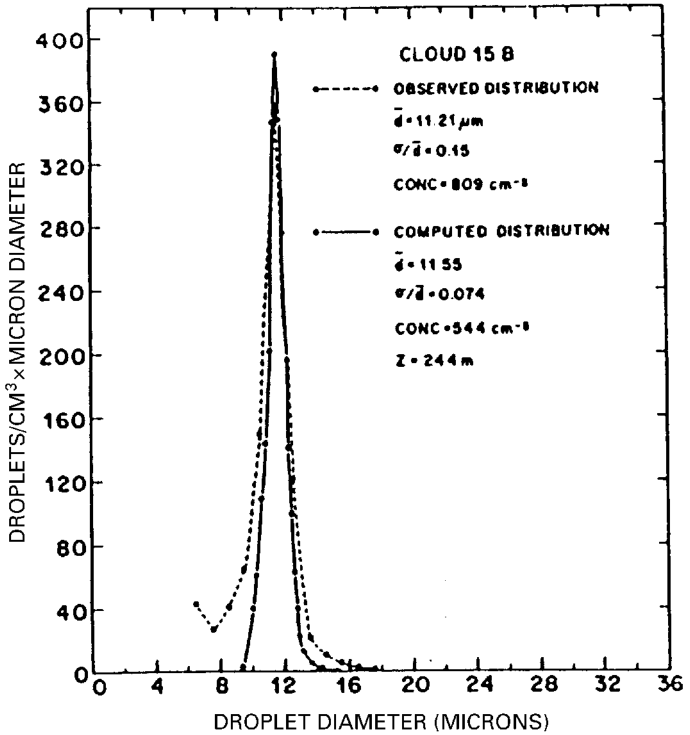


Figure 12 Computed and measured cloud droplet distributions at a height of 244 m above cloud base (Fitzgerald, 1972).

rain in about 20 min. Continental clouds, which have much higher droplet concentrations, also produce rain in short times compared to the prediction of Eq. (18). Although continued condensation within a parcel may eventually generate droplets larger than about 25 μm radius to initiate coalescence, the time for this process to produce rain is generally much too long.

Proposed Mechanisms for Broadening of Droplet Spectra

Several mechanisms have been suggested to explain the production of large cloud droplets capable of initiating coalescence. These include (1) favorable CCN spectra, such as low concentrations of smaller nuclei and high concentrations of larger nuclei; (2) mechanisms that enhance condensation growth of a few larger cloud droplets through mixing of air parcels; (3) stochastic condensation, the mixing of droplets (rather than parcels) that have different supersaturation histories; and (4) fluctuations in supersaturation or collision rates induced by turbulence.

One of the earliest mechanisms hypothesized to explain the onset of precipitation is that very large (>10 μm radius) wettable or soluble nuclei exist below cloud base that, when carried through cloud base, can almost immediately begin to collect cloud droplets. Measurements in marine environments show that giant salt particles can often occur in significant concentrations. Within the moist region below cloud base such salt particles will deliquesce into much larger solution droplets [e.g., a NaCl particle of 1 ng (5 μm) has a 25-μm equilibrium radius at 99% relative humidity]. Thus, there are often significant concentrations of coalescence nuclei that can enter the base of maritime clouds. Numerical parcel models initialized with CCN spectra from observations have been used to evaluate the role of giant CCN in the production of warm rain. Calculations show that raindrops can form on these CCN in time scales comparable to observed cloud lifetimes. Drop trajectory calculations also suggest that accretion of cloud water on giant and ultragiant nuclei can account for the formation of rain in observed time scales when these large particles are present at cloud base. Nevertheless, questions remain as to how important these large salt particles really are to the warm rain process in maritime clouds. Uncertainties about the growth rate of these particles in updrafts remain, since their growth is highly nonequilibrium. Also, the residence time of larger salt particles in the atmosphere is short, with the highest concentrations located close to the ocean surface. Much less is known about the role giant particles play in initiating coalescence in continental clouds. Measurements show that 50-μm-diameter particles exist in concentrations of 100 to 1000 m^{-3} in the boundary layer over the High Plains of the United States. These are typically insoluble soil particles that will not grow as they pass through cloud base. They are large enough, however, to make effective coalescence nuclei without undergoing growth by condensation.

A second mechanism to explain the onset of precipitation involves broadening of the drop size distribution during the mixing process. Early studies of parcels subjected to velocity fluctuations were unable to reproduce the appreciable broadening of the cloud droplet spectra that occurs in warm cumulus. These studies invoked the process of *homogeneous* mixing in which dry air is mixed into a cloudy parcel of air such that all droplets are exposed to the dry air and evaporate until the parcel is again saturated. This process has been shown to produce a constant modal size, but a decrease in mean size because of a broadening toward smaller sizes. Absent in this process is the large drop tail to the spectrum required for coalescence growth.

Laboratory experiments by Latham and Reed (1977) raised questions about the applicability of the homogeneous mixing process. Their experiments suggested that the mixing process occurs in such a way that some portions of a cloudy parcel are completely evaporated by mixing with unsaturated air while other portions remain unaffected. Broadening of the droplet spectrum toward larger sizes will occur by subsequent condensation of the parcel if it continues to ascend, since there is reduced competition for vapor among the remaining droplets. Further studies have showed that an idealized inhomogeneous mixing process will produce appreciable broadening compared to homogeneous mixing. However, other studies have found that special dynamical sequences of vertical mixing can achieve the same spectral broadening without invoking inhomogeneous mixing concepts. Observations in clouds over the High Plains of the United States found that the large drop peak was more consistent with a closed parcel environment than a mixed one and that there was no evidence that mixing or cloud age increased the size or concentration of the largest drops. Considerable uncertainty remains concerning the importance of mixing to the broadening of the droplet spectrum.

The stochastic condensation mechanism invokes the idea that droplets can have different supersaturation histories. This process considers the mixing of droplets, rather than parcels. Cooper (1989) describes the theory of stochastic condensation. Regions of clouds having fine-scale structure, such as mixed regions of clouds outside of adiabatic cores, would best support this type of process. Measurements within warm, orographic clouds in Hawaii have shown little broadening in the laminar clouds, but appreciable broadening in the breaking wave regions of the clouds near cloud top, suggesting the importance of stochastic condensation in producing large droplets.

A fourth hypothesis is that the breadth of the droplet spectra in clouds can be increased by turbulent fluctuations. Conceptually, turbulence may be thought to induce vertical velocity fluctuations that induce fluctuations in supersaturation. These fluctuations, in turn, lead to the production of new droplets at locations throughout the cloud. Turbulent fluctuations can also lead to drop clustering, which can create opportunities for favorable supersaturation histories and/or enhanced collision rates. A general approach has been to examine the evolution of a distribution of droplets exposed to a supersaturation that varied with a known distribution, such as a normal distribution, and to derive analytic expressions for the droplet size distribution as a function of time. This approach predicts continued dispersion of the spectra with time. This approach has been criticized on the basis that updrafts and supersaturation are highly correlated. A droplet that experiences a high supersaturation is likely to be in a strong updraft and will arrive at a given position in a cloud faster, which means there will be less time for growth. Conversely, a droplet that experiences a low supersaturation will grow slower but have a longer time to grow. The net result is that the droplets arrive at the same place with approximately the same size. Supersaturation fluctuations in clouds can arise not only from fluctuations in the vertical velocity, but also from fluctuations in integral radius (mean radius multiplied by the droplet concentration) of the droplet spectra. Realistic droplet spectra can be obtained in model simulations when the fluctuations in mean radius are negatively correlated with vertical motions. Unfortunately, experimental data in cumulus show the opposite behavior—the largest droplets occur in upward moving parcels.

Recent studies have provided conflicting evidence regarding the role of turbulence in spectral broadening. Studies show that turbulence can lead to significant trajectory deviations for smaller droplets in clouds, leading to larger relative vector velocities for droplets and collector drops and enhanced collision rates. Results indicate that cloud turbulence supports spectral broadening and more rapid production of warm rain. Studies have also examined whether turbulence can create regions of preferential concentration in conditions typical of cumulus clouds, and whether nonuniformity in the spatial distribution of droplets and/or variable vertical velocity in a turbulent medium will contribute to the broadening of the drop size distribution.

These studies suggest that turbulence severely *decreases* the broadening of the size distribution compared to numerical experiments performed in the absence of turbulence. The importance of turbulence in drop spectral broadening is still a matter of debate. Villiancourt and Yau (2000) provides a review of recent understanding of the role of turbulence in spectral broadening.

Growth by Coalescence

When sufficiently large (>25 μm radius) cloud droplets exist in a cloud, their growth is accelerated by collision and coalescence with neighboring droplets. The nonlinear behavior of the flow around a drop affects both the drop's terminal velocity and the manner in which a smaller drop is swept around a larger drop as the two approach in free fall.

Single Drop Hydrodynamics The Reynolds number, given by $N_{\mathrm{Re}} = 2V_T r/v$, where V_T is the drop terminal velocity, and v is the kinematic viscosity, is the ratio of inertial to viscous forces and is typically used as a scaling parameter to describe the flow regimes around a drop. The characteristics of flow around a sphere as a function of the Reynolds number are well documented. When N_{Re} is very small (<0.4), the droplet is in the so-called Stokes flow regime where viscous forces dominate and the flow field around a droplet is symmetric and laminar. At about $N_{\mathrm{Re}} = 2$ (about $r = 50$ μm at standard temperature and pressure), a wake appears behind the droplet. At about $N_{\mathrm{Re}} = 20$ (about $r = 150$ μm), an eddy of rotating fluid appears behind the drop. This eddy grows as N_{Re} increases to 130 ($r = 350$ μm), at which point the wake becomes unstable. At $N_{\mathrm{Re}} = 400$ ($r = 650$ μm) eddies are carried downstream by the turbulent wake.

The problem of determining the flow around a drop is complicated by shape deformation due to hydrodynamic and sometimes electrical forces. The nature of the deformation is primarily related to the droplet size. Observations in wind tunnel experiments and model calculations have provided information on the nature of drop deformation. Numerical modeling has further clarified the nature of drop deformation under the influence of external hydrostatic pressure. Chuang and Beard (1990) summarize the effect of hydrodynamic and electric forces on both charged and uncharged drops. Figure 13 from their study shows the shapes drops can assume as they fall in various electric fields characteristic of thunderstorms.

The drag force on the droplet and the force of gravity determine the terminal velocity of a droplet. For Stoke's flow, determining the balance between these forces is relatively straightforward, but in the more complicated flow regimes, a single accurate analytical expression for the terminal velocity is not available. Rogers and Yau (1989) suggest the following general expression:

$$V_t = ar^b \tag{21}$$

for drops that remain spherical, where the terms a and b have different values depending on the Reynold's number. Gunn and Kinzer (1949) provide the most

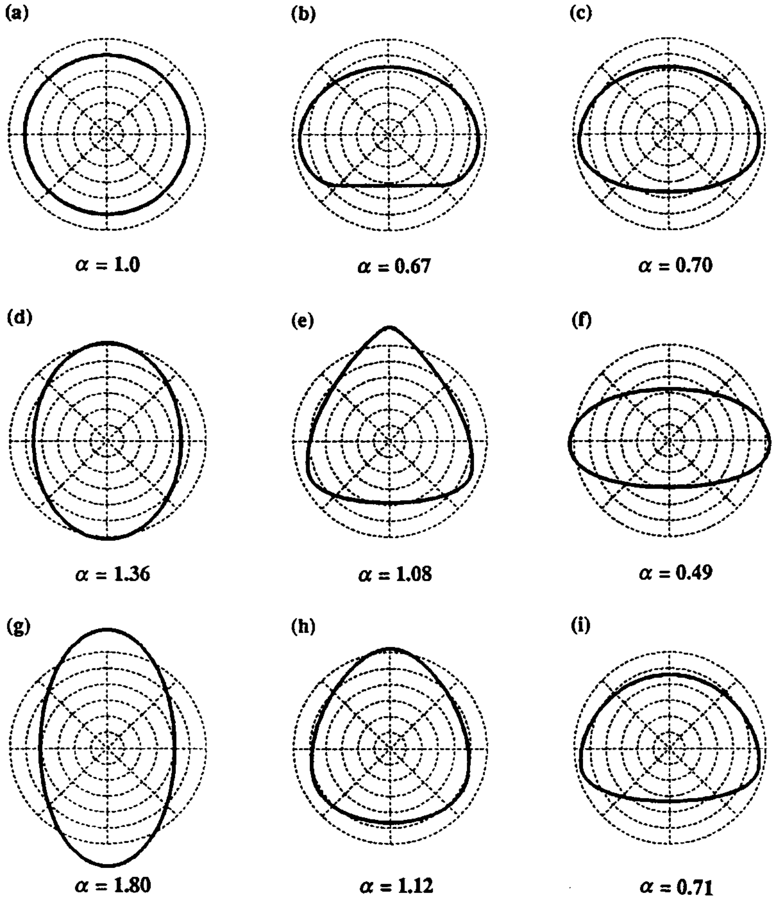


Figure 13 Modeled drop shapes and axis ratios ($\alpha =$ vertical/horizontal dimension) for 5-mm diameter drops with various distortion effects: (*a*) stationary drop (surface tension only), (*b*) drop resting on a flat surface, (*c*) raindrop falling at terminal velocity, (*d*) stationary drop in strong vertical electric field, (*e*) raindrop falling in a strong vertical electric field, (*f*) highly charged drop falling at terminal velocity, (*g*) stationary drop in the maximum vertical electric field before disintegration, (*h*) charged raindrop falling in maximum thunderstorm electrical field with upward directed electric force, and (*j*) same as (*i*), but with downward electric force (Chuang and Beard, 1990).

accurate observational data over a large range of drop sizes. Their measurements, made at 1013 mb and 20°C, are shown in Figure 14.

Collection Efficiency When two droplets interact in free fall, the outcome of the interaction will depend on the droplet's sizes and the offset, x, between the drop centers. Inside some critical offset, x_0 (see Fig. 15), the outcome will be a collision, while outside x_0 the outcome will be a miss. For drops acting as rigid spheres, the

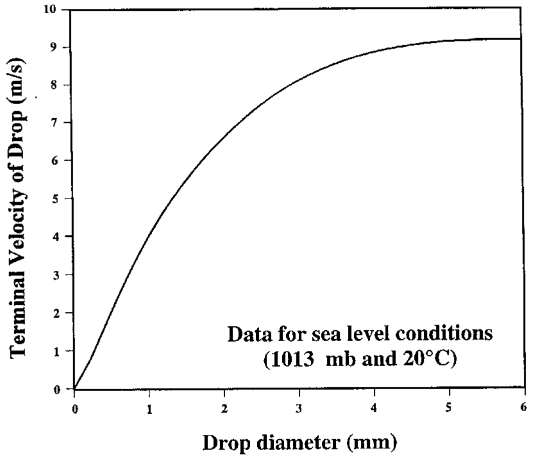


Figure 14 Terminal velocity of a raindrop as a function of its diameter at 1013 mb and 20°C (adapted from Gunn and Kinzer, 1949).

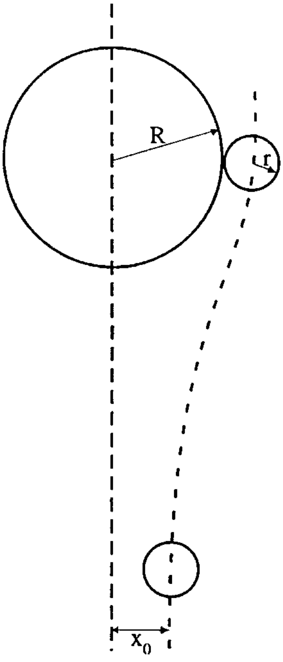


Figure 15 Collision geometry for two droplets.

collision efficiency, E_{col}, is defined as

$$E_{\text{col}} = \frac{(x_0)^2}{(R + r)^2} \tag{22}$$

the ratio of the collection cross section to the geometric cross section for the large and small drops. Theoretical calculations for collector drops with $R < 75\ \mu\text{m}$ are shown in Figure 16. Typically most newly formed cloud droplets will range in size from about $4 < r < 8\ \mu\text{m}$. A 25-μm-radius collector drop will have a collection efficiency of about 10% for these droplets. For this reason, the 25-μm radius is generally considered a threshold size for initiation of coalescence. Theoretical collision efficiencies for larger R and r/R ratios rapidly approach unity and can even exceed unity when r/R approaches 1 and $R > 40\ \mu\text{m}$ due to the possibility of capture

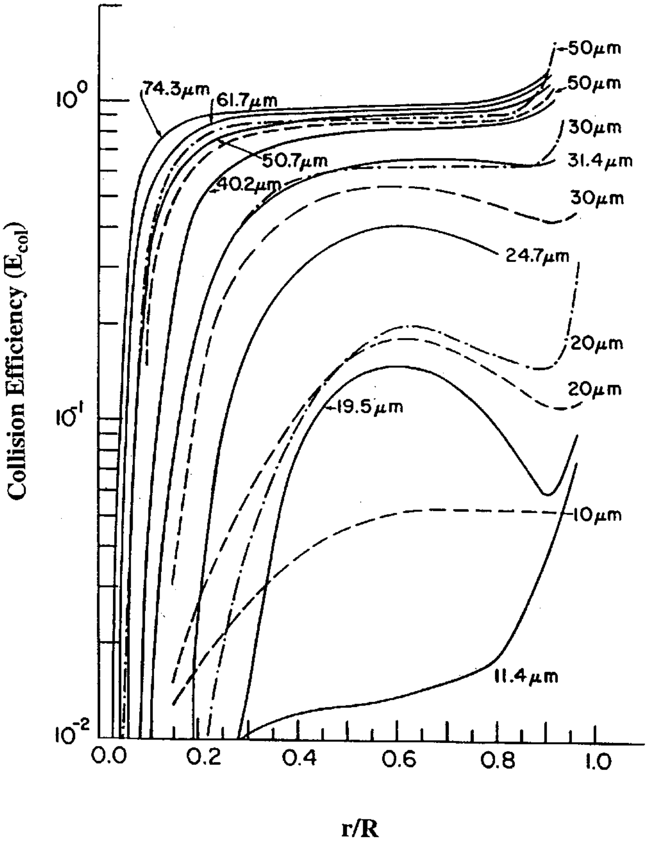


Figure 16 Collision efficiencies for small spheres as a function of the ratio of their radii. Solid lines from Schlamp et al. (1976), dashed lines from Klett and Davis (1973), dashed-dot lines from Lin and Lee (1975). (Adapted from Schlamp et al., 1976.)

of the small droplet in the large droplet wake. Unfortunately, no reliable laboratory studies exist for $R < 40$ μm, so the critical efficiencies for the coalescence threshold remain untested.

The calculations of E_{col} in Figure 16 assume that the drops are rigid spheres. Drizzle-sized drops ($R > 100$ μm) deform upon approach to another droplet as the air film between the drops drains. Studies of small precipitation-sized drops in the laboratory have shown that the outcome of collisions can be complicated, with the possibility for coalescence, temporary coalescence, and satellite production. These outcomes depend on R and r, the drop separation, charge, and the ambient relative humidity. Coalescence depends upon the drop–droplet interaction time and their impact energy. Formulas to calculate coalescence efficiencies, E_{coal}, over a wide range of drop sizes are now available. The collection efficiency, E, is the product of E_{col} and E_{coal}. Beard and Ochs (1984) show that as E_{col} approaches unity, E_{coal} decreases, leading to a maximum E for drop–droplet pairs in the size range $R = 100$, $r = 12$ μm (see Fig. 17).

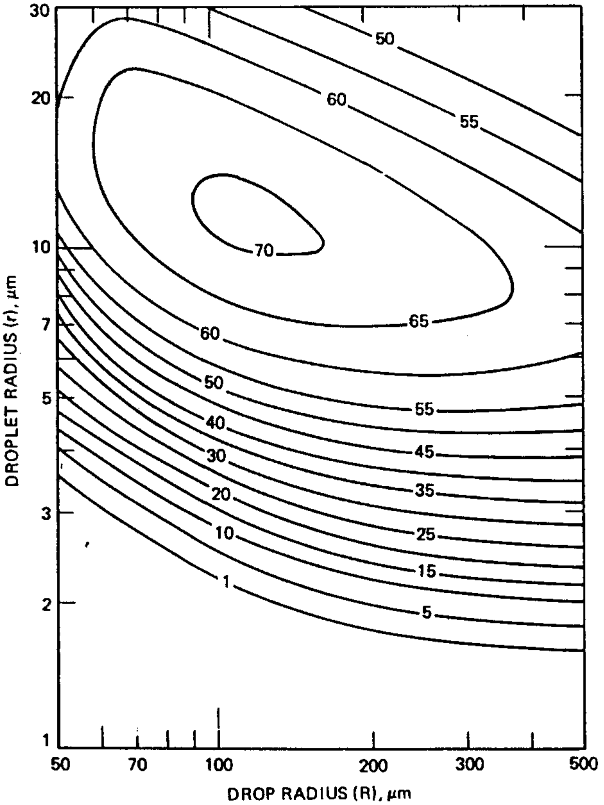


Figure 17 Contours of collection efficiency (in percent) as a function of drop and droplet radius (Beard and Ochs, 1984).

Growth by Coalescence Two methods have been used for calculating growth rates of cloud droplets by coalescence. The first, called *continuous collection*, is used to calculate the growth of single drops that pass through a cloud of smaller droplets. These single drops are assumed not to interact with drops of similar size so as to lose their identity. The continuous collection equation is

$$\frac{dM}{dt} = \sum K(R, r) w_L(r) \Delta r \tag{23}$$

where M is the mass of the drop, K is the collection kernel, and w_L is the liquid water content per unit size interval Δr. The collection kernel, the volume of air swept out per unit time by the larger drop, is given by $K = \pi E(R + r)^2[V_T(R) - V_T(r)]$, and $w_L = n(r)4\pi r^3 \rho_w/3$, where E is the collection efficiency, $V_T(R)$ and $V_T(r)$ are the terminal velocities of the large and small drops, ρ_w is the liquid water content, and n is the small droplet concentration.

Calculations show that small raindrops can form in about 20 to 30 min provided coalescence nuclei ($R > 50\,\mu$m) are present in the cloud and the smaller cloud droplets have radii of at least 7 μm. Szumowski et al. (1999) applied the continuous growth equation in marine cumulus using a Lagrangian drop-growth trajectory model in wind fields derived from dual-Doppler radar analyses. The evolution of the radar-observed clouds in this study provided a strong constraint on the growth of raindrops. Raindrops were found to grow to sizes ranging from 1 to 5 mm from coalescence nuclei in about 15 to 20 min. Their calculations suggest that coalescence growth is efficient in marine clouds that contain giant and ultragiant sea salt aerosol and low (100 to 300 cm^{-3}) droplet concentrations. In continental clouds, which generally have narrow droplet distributions and high (800 to 1500 cm^{-3}) droplet concentrations, continuous growth may take significantly longer.

In nature, the droplets composing a cloud have a continuous distribution of sizes, and droplets of all sizes have the potential to collide with drops of all other sizes. The drops of most importance to the precipitation process occupy the large end tail of the drop size distribution. To consider rain formation, the statistical aspects of interactions between droplets must be considered. There is an element of probability associated with the collection process, and, occasionally, collisions will occur between larger droplets. A general method for calculating coalescence growth is to assess the probability that droplets of one size will collect droplets of another size for all possible droplet size pairs in the droplet distribution. The term *stochastic collection* is applied when the collection probability is used to determine the number of droplets of a given size, r, that are created by collisions of smaller pairs, $r - \Delta r_1$ and $r - \Delta r_2$, and destroyed by collisions of r with other droplets. Stochastic collection considers rare collisions between larger droplets and consequently predicts accelerated coalescence in warm clouds.

Possible interactions between drops include small droplets collecting small droplets (*autoconversion*), large drops collecting small drops (*accretion*), and large drops collecting large drops (*large hydrometeor self-collection).* Bulk microphysical cloud models often parameterize these three processes. Another outcome is *droplet breakup* where two drops collide with the result being a number of smaller drops.

Collisions between large drops will result in a dramatic broadening of the tail of the distribution and an increase in the size and number of precipitating drops. However, these drops are in such low concentration that collisions of this type must be considered a rare event from a statistical point of view. Simulations of the collection process using stochastic approaches must accurately consider the most rare events.

The continuous form of the stochastic coalescence equation is given by:

$$\begin{aligned}\frac{\partial N(m,t)}{\partial t} &= \frac{1}{2}\int_0^m N(m',t)N(m-m',t)K(m',m-m')dm \\ &\quad - \int_0^\infty N(m,t)N(m',t)K(m,m')dm' \qquad (24)\end{aligned}$$

where N is the concentration of drops of mass m at time t and K is the collection kernel. The first term describes the coalescence of two drops of mass m' and $m - m'$ to form a droplet of mass m. The average concentration of drops of mass m will increase for every coalescence of two smaller drops of masses m' and $m - m'$. For every two drops "destroyed" one is created, thus the factor of $\frac{1}{2}$. The second term describes coalescence of drops of mass m with drops of mass m'. This term represents the "destruction" of drops of mass m, since any mass m drop that combines with any other drop will no longer be mass m.

Care must be taken to avoid numerical diffusion when integrating the stochastic collection equation. Considerable efforts have been made to develop integration schemes that accurately portray the evolution of the droplet spectra. Berry and Reinhardt's (1974) study indicated that the formation of raindrops in warm clouds can be fairly rapid. In their numerical experiments, initial droplet sizes were specified using gamma distributions of their masses, with the liquid water content as 1 g/m^3 and mean drop sizes ranging from 10 to 18 μm radius (see example in Fig. 18). Drizzle drops were produced in 10 to 22 min and raindrops in 20 to 30 min in different experiments. Faster growth occurred for larger mean initial sizes. The growth of drops after they reached drizzle drop size was nearly independent of the drop spectrum. Berry and Reinhardt's studies show that growth of raindrops in warm clouds can occur by stochastic processes within time scales observed in nature, provided the initial spectra has sufficiently large mean radii and the spectra contain droplets with radii of at least 30 μm radius.

5 GROWTH OF ICE PARTICLES IN ATMOSPHERE

Ice crystals form in the atmosphere through either homogeneous or heterogeneous nucleation. Once formed, ice particles grow by three mechanisms: diffusion of water vapor, accretion of supercooled droplets, and aggregation with neighboring crystals.

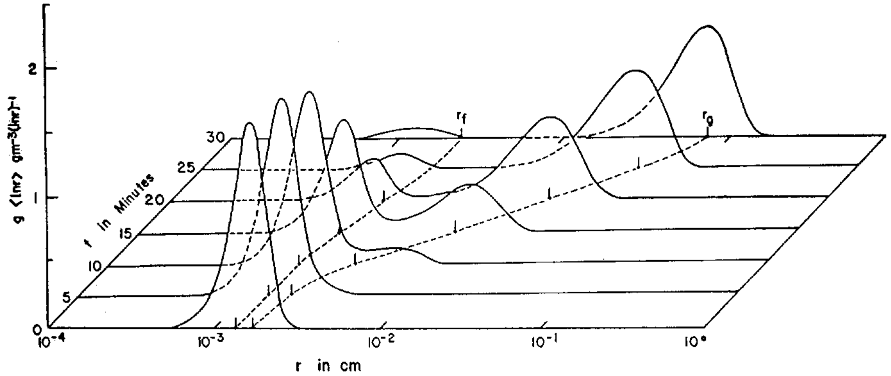


Figure 18 Example of development of a droplet spectrum using stochastic coalescence (from Berry and Reinhardt, 1974). The radius r_f denotes the mean mass, which follows the cloud droplet peak, and r_g denotes the radius of mean-square mass, which follows the raindrop peak.

Accretion leads to the formation of graupel and hail, while aggregation produces snowflakes. Ice particles sometimes shatter during growth or evaporation, increasing particle concentrations. Mechanisms that enhance ice particle concentrations in clouds in this manner are called *secondary* ice particle production mechanisms.

Diffusional Growth of Ice Particles

The Clausius–Clapeyron equation describes the equilibrium condition between two phases. By comparing the solution of the Clausius–Clapeyron equation for the cases of water and vapor, and ice and vapor, it can be easily shown (e.g., Rogers and Yau, 1989; Pruppacher and Klett, 1997) that the saturation vapor pressure over water, $e_{s,\infty}$, will always exceed the saturation vapor pressure over ice, $e_{si,\infty}$, for $T < 0°\text{C}$. This is a consequence of the fact that the latent heat of sublimation, L_s, is larger than L_v. This difference has major implications for the growth of ice in clouds. Air saturated with respect to ice will always be subsaturated with respect to water. Conversely, air saturated with respect to water will always be supersaturated with respect to ice. As a consequence, supercooled water droplets and ice crystals can *never exist in a cloud together in equilibrium.* Ice crystals, always subject to greater supersaturations, will rapidly grow, extracting vapor from the surroundings, while coexisting supercooled droplets will evaporate, supplying vapor to the surroundings. The differences in relative humidity with respect to ice and water as a function of temperature can be found by taking the ratios $e/e_{s,\infty}$ and $e/e_{si,\infty}$ and plotting them as a function of temperature. Figures 19 and 20 show that, at cold temperatures, air can be significantly below saturation with respect to water (relative humidity with respect to water, $\text{RH}_w \approx 65$ to 70%) and still be supersaturated with respect to ice. High in the atmosphere, at cirrus cloud levels, ice crystals can grow at moderate relative humidities with respect to water. Even at lower altitudes, particularly in

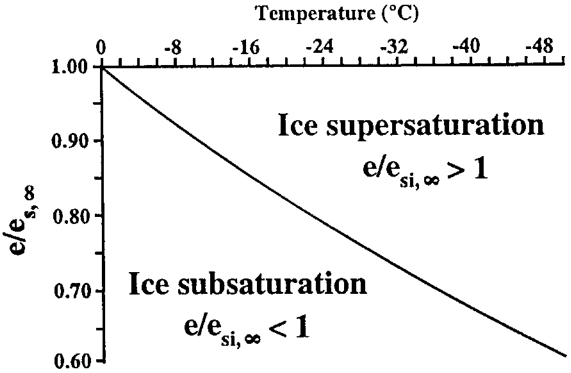


Figure 19 Ice saturation as a function of saturation ratio with respect to water (adapted from Pruppacher and Klett, 1997).

wintertime, ice particles can grow at $RH_w < 100\%$. On the other hand, a cloud that has a relative humidity with respect to water of 100% will be supersaturated with respect to ice by a substantial amount. For example, at −10°C, the relative humidity with respect to ice will be about 112%. At −20°C, it will be 120%. In an environment where $RH_w = 100\%$ exactly, water drops will not grow, yet ice crystals in this same environment will experience enormous supersaturations. As a consequence, they grow very rapidly to become large crystals.

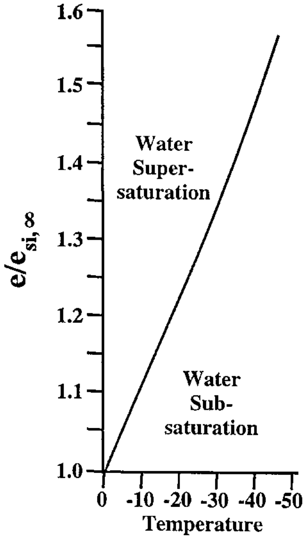


Figure 20 Water saturation as a function of saturation ratio with respect to ice (adapted from Pruppacher and Klett, 1997).

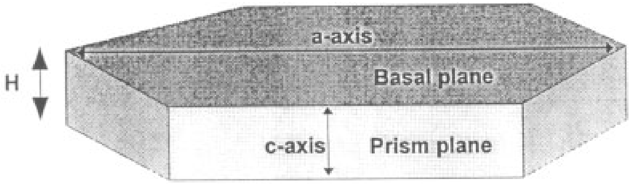


Figure 21 Geometry of a simple ice crystal.

From a crystallographic point of view, ice crystal growth normally occurs on two planes, the *basal* and *prism* planes. The basal plane of an ice crystal refers to the hexagonal plane illustrated in Figure 21. The axis that passes from the outside edge of the crystal through the center along the basal plane is called the *a* axis. The *a* axis can be thought of as the diameter of the circle that circumscribes the basal plane. The prism plane refers to any plane along the vertical axis of the crystal, the *c* axis. Prism planes appear rectangular in Figure 21. The height of a crystal, *H*, is shown in Figure 21. Growth along the basal plane implies that the *c* axis will lengthen and *H* will become larger. Growth along the prism plane implies that the *a* axis will lengthen and the diameter of the crystal will become larger.

Laboratory experiments have shown that the rate of growth along the *c* axis relative to growth along the *a* axis of crystals varies significantly as a function of temperature and supersaturation. These variations occur in a systematic manner, resulting in the characteristic shapes, or "habits," of ice crystals, summarized in Figure 22. As temperature decreases, the preferred growth axis changes from the *a* axis (platelike crystals, 0 to −4°C) to the *c* axis (columnar crystals, −4 to −9°C), to the *a* axis (platelike crystals, −9 to about −22°C), to the *c* axis (columnar crystals,

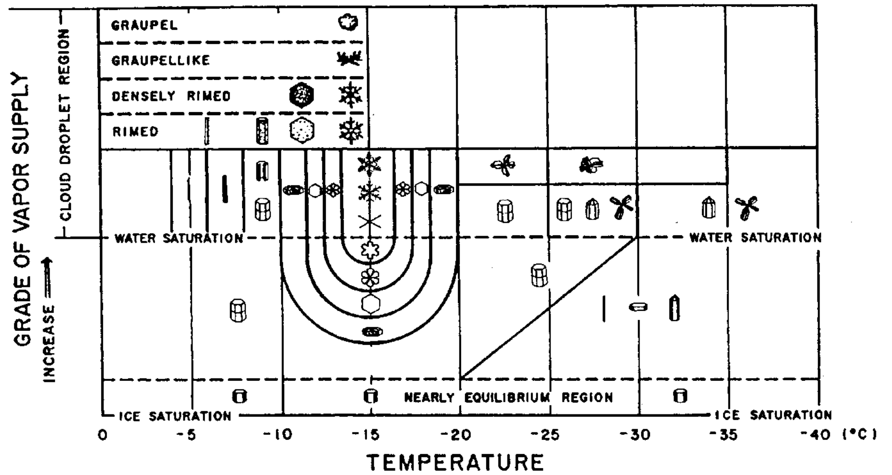


Figure 22 Natural ice crystal habits that form when crystals grow in different temperature and humidity conditions (Magono and Lee, 1966).

-22°C and colder). The first two transitions are well defined, while the latter is diffuse, occurring between about -18 and -22°C. At high supersaturations with respect to ice (water supersaturated conditions), the transition as a function of temperature is, in the same order as above, a plate, a needle, a sheath, a sector, a dendrite, a sector, and a sheath. The temperatures of these transitions are evident from Figure 22. At low supersaturations with respect to ice (below water saturation), the transition in habit as a function of temperature is from a plate, to a hollow (or solid) column, to a thin (or thick) plate to a hollow (or solid) column, depending on the level of saturation. Observations of ice particles in the atmosphere show that the same basic habits appear in nature, although the structure of the particles is often more complicated. This can be understood when one considers that the particles fall through a wide range of temperatures and conditions of saturation.

Ice particles possess complex shapes and, as a result, are more difficult to model than spherical droplets. However, the diffusional growth of simple ice crystals can be treated in a similar way as drops by making use of an analogy between the governing equations and boundary conditions for electrostatic and diffusion problems. The electrostatic analogy is the result of the similarity between the equations that describe the electrostatic potential about a conductor and the vapor field about a droplet. Both the electrostatic potential function outside a charged conducting body and the vapor density field around a growing or evaporating droplet satisfy Laplace's equation. The derivation of the growth equation for ice crystals proceeds exactly as that for a water droplet, except the radius of a droplet, r, is replaced by the electrostatic capacitance, C, the latent heat is the latent heat of sublimation, L_s, and the saturation is that over ice, $S_{i,v}$, rather than water. In the case of a spherical ice crystal $C = r$. The result is

$$\frac{dM}{dt} = \frac{4\pi C(S_{i,v} - 1)}{L_s/KT\big((L_s/R_vT) - 1\big) + R_vT/De_{si,\infty}} \tag{25}$$

The problem in using this equation is determining the capacitance factor C, which varies with the shape of the conductor. For a few shapes, C has a simple form. These shapes are not the same as ice crystals but have been used to approximate some ice crystal geometries. To use Eq. (25) to determine the axial growth rate of ice particles, the term dM/dt must be reduced to a form containing each of the axes. In general, this will depend on the crystal geometry. To solve for a or c, an additional relationship is required between da/dt and dc/dt. This relationship is generally obtained from measurements of the axial relationships determined from measurements of a large number of crystals. Laboratory experiments conducted to determine the importance of surface kinetic effects in controlling ice crystal growth have shown that surface kinetic effects are important in controlling crystal habit since kinetic effects control how vapor molecules are transported across the crystal surface and incorporated into the crystal lattice. Ryan et al. (1976) has measured the growth rates of ice crystals in the laboratory. Figure 23 from their study shows that the a axis

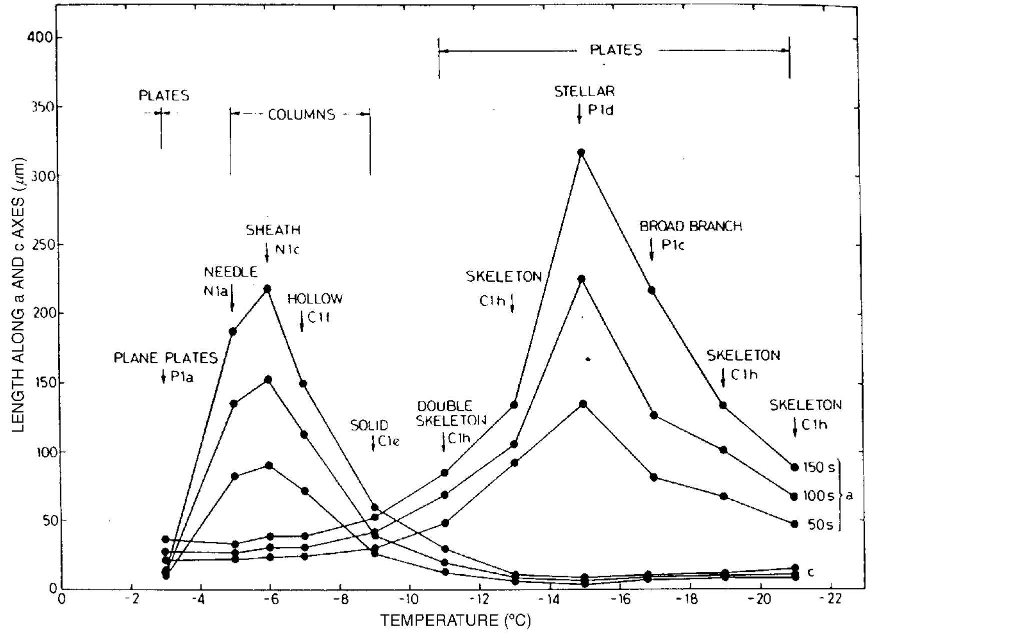


Figure 23 Variation of crystal axial dimensions with temperature for 50, 100, and 150 s after seeding (Ryan et al., 1976).

grows most rapidly at $T > -9°\text{C}$, with a peak at $-15°\text{C}$, while the c axis grows most rapidly at $T < -9°\text{C}$, with a peak at $-6°\text{C}$.

Ice Particle Growth by Accretion

Growth by accretion occurs when ice particles collide with supercooled water droplets, causing them to freeze on to the ice surface. For supercooled water to be present in substantial quantities in a mixed-phase cloud, it is necessary for the condensate supply rate to exceed the bulk diffusional growth rate of the ice particles, after accounting for any dry air entrainment that may be occurring. High condensate supply rates require sustained large vertical velocities. For this reason, accretion is an important growth process in clouds with localized regions of sustained high vertical velocities.

The most common type of cloud that meets this requirement is a cumulus cloud, particularly cumulus clouds with large vertical development. Regions of other types of cloud systems also have sustained high vertical velocities. Examples include orographic clouds, particularly near the maximum terrain slope, and convective cores of frontal bands. A second way supercooled water can appear in a cloud is through vertical transport of droplets upward through the 0°C level. This process can occur in any type of cloud with a moderate updraft near the melting level. The supercooled water in this case will generally be confined to the warmer range of subfreezing temperatures. Clouds that generally do not contain much supercooled water are those with weak vertical velocities, such as stratus clouds. In such clouds, the accretion process will not be as important, if it occurs at all.

Studies of accretion have focused primarily on the conditions required for riming onset, and the late stages of riming, when the particles reach the graupel or hail stage of growth. The onset of riming has been studied theoretically and in field and laboratory experiments. These studies generally show that platelike crystals must grow to a threshold size of about 150 μm radius before any droplets accrete to their surfaces. Droplets must be larger than about 5 μm radius before they are collected. Collection efficiencies increase with increasing ice particle and cloud droplet size, reaching about 0.9 when the ice particle radius approaches 400 μm and the droplet radius reaches 35 μm.

Graupel develops in a cloud when small supercooled cloud droplets collect in large numbers on falling ice particles. The parent ice particle, or embryo, of a graupel particle can have various shapes. Certain types of ice particles serve as graupel embryos more effectively than others. Studies have shown that aggregates of ice crystals can provide excellent graupel embryos in hailstorms because of their ability to rapidly collect water droplets. Aggregate embryos can be entrained into the main updraft core of hailstorms from the debris clouds of previously active convective towers. Large drops present in the core updraft can also serve as graupel embryos upon freezing. Graupel characteristics, including density, mass, and terminal velocities have been measured in a number of field studies. In general, the data show a great deal of spread, since the observations come from very different cloud systems.

Hail develops when extreme growth by accretion occurs in thunderstorms. Hail growth occurs in one of two ways, depending on the surface temperature of the hailstone and the rate of supercooled water accumulation. As a hailstone collects supercooled water droplets while falling through a thunderstorm, latent heat is released at the hailstone's surface as the droplets freeze, raising the surface temperature, T_s, of the hailstone. If $T_s < 0°\text{C}$, droplets continue to freeze immediately on the hailstone's surface. If, however, T_s reaches 0°C, water impacting on the hailstone will no longer immediately freeze. Water then spreads across the surface of the stone, drains into porous regions, and can shed from the stone's surface. The former situation, when $T_s < 0°\text{C}$, is called the *dry growth regime*. The latter situation, when $T_s = 0°\text{C}$, is called the *wet growth regime*. For dry growth, the growth rate of the spherical hailstone due to collection of water drops can be approximated by the continuous collection equation:

$$\left(\frac{dM}{dt}\right)_{\text{dry}} = E\pi R_H^2 V_{T,H} w_L \tag{26}$$

where R_H is the radius of the hailstone and $V_{T,H}$ is its terminal velocity. The terminal velocity and radius of cloud drops are neglected because they are small relative to R_H and $V_{T,H}$ [compare Eqs. (23) and (26)]. The same equation applies in the wet growth regime provided the hailstone does not shed water. If the hailstone grows in the wet growth regime, and sheds excess liquid water from its surface, its growth rate will be determined by the rate at which the collected water can be frozen. Calculations show that at $-5°C$, the effective liquid water content for wet growth is less than $1\ g/m^3$. At $-10°C$ the effective liquid water content is near $1\ g/m^3$. Higher liquid water content is required for wet growth at colder temperatures. The threshold values of liquid water content are well within the range of values found in cumulonimbus, implying that shedding of drops and wet growth are important processes in thunderstorms.

Hailstones exhibit alternating layers of higher and lower density. The density variations are due to variations in the concentration of bubbles trapped in the stone. Generally bubbles are grouped in concentric layers of higher and lower density, which in turn are associated with the different growth regimes. Field studies have shown that hailstone embryos can be conical graupel, large frozen drops, and occasionally large crystals. Hailstones sometimes exhibit irregular structure while others have distinct conical or spherical shapes. The largest known hailstone, collected near Coffeyville, Kansas, weighed 766 g and had a circumference of 44 cm.

Ice Particle Aggregation

Snowflakes form through aggregation of ice crystals. Several mechanisms have been proposed to explain the adhesion of ice particles to one another. Snowflakes of sizes greater than 1 cm are often composed of planar and spatial dendritic crystals, needlelike crystals or, at higher cloud levels, radiating assemblages of bulletlike crystals. All of these crystals have shapes that easily interlock during collisions, suggesting that mechanical interlocking is initially important for crystals to attach to each other. Two mechanisms have been identified that bond ice particles together once in contact. The first, called sintering, occurs because ice particles in point contact form a nonequilibrium system. To minimize the total free surface energy of the system, a requirement for equilibrium, water molecules must diffuse from ice surfaces toward the point of contact, strengthening the bond between the particles. The second bonding mechanism appears to be associated with the presence of a liquid layer on the surface of ice crystals. This layer is thought to enhance a crystal's sticking efficiency. Experiments have shown that the surface electrical conductivity of ice increases significantly at temperatures warmer than $-10°C$, suggesting the surface contains a quasi-liquid layer. Studies using optical and magnetic resonance techniques support the existence of the layer. However, the existence of a liquid layer on the surface of ice crystals at subfreezing temperatures is still not universally accepted.

Larger aggregates typically first develop between -15 and $-12°C$, the temperature range where dendritic crystals form, although aggregates have been observed at colder temperatures in cirrus. Aggregates form most rapidly near the melting level. Enhanced aggregation near the melting level is apparently due to increased collection efficiency as a liquid layer develops on the surface of the crystals. The rate of aggregation in clouds also depends on ice particle concentrations. Calculations and field observations both suggest that aggregation is enhanced in the presence of high crystal concentrations.

Ice Particle Concentrations and Evidence for Ice Multiplication

Ice particles first appear in clouds when ice nucleation occurs. Measurements of ice nuclei suggest that they should be active primarily at cold temperatures (<-20°C) and that the concentration of active nuclei should be an exponential function of temperature. According to measurements of ice nuclei (Fig. 10), few ice particles should present in clouds whose tops are warmer than -10°C.

Measurements of ice particle concentrations in clouds have not shown this expected relationship. Ice particle concentrations often exceed by several orders of magnitude the concentrations expected on the basis of ice nuclei measurements. This is particularly true at warm temperatures (>-10°C). Conversely, at temperatures between -20°C and the homogeneous nucleation threshold (-35 to -40°C) ice particle concentrations are often less than expected on the basis of ice nuclei measurements.

Determining the origin of ice particles in clouds remains one of the major unsolved problems in cloud physics. It is possible that ice nucleation occurs in clouds under conditions that are not well understood. It is also possible that secondary mechanisms for ice enhancement occur. Three mechanisms have been studied extensively: (1) fragmentation of colliding ice particles, (2) ice splintering during the riming process, and (3) shattering of drops during freezing.

Ice particles that fragment during collisions are typically more delicate crystals such as dendrites, sectors, side planes, and needles and sheaths. Particles, that seldom fragment are thick and thin plates, solid and hollow columns, and scrolls. Experiments to examine fragmentation show that dendritic crystals readily fragment in conditions expected in clouds. On the other hand, breakup of rime accumulated on crystals during collisions is unlikely in clouds. Overall, data suggest that only clouds containing dendritic particles are likely to have a significant increase in measured ice particle concentrations due to fragmentation during collisions.

The high ice particle concentrations observed in clouds provoked a large number of research groups in the early 1970s to try to find a strong ice multiplication mechanism to explain the observations. A breakthrough on the problem was reported by Hallett and Mossop (1974), who showed that splintering of ice during the riming process led to copious ice production under a limited range of conditions. Subsequent studies determined the conditions required for splinter production and clarified the mechanism for splintering.

These studies together showed that copious numbers of ice splinters could be produced during riming when: (1) the temperature was between -3 and -8°C; (2) large ice particles were present in the cloud to collect cloud droplets, (3) droplets with diameters $>24\,\mu$m were present, and (4) droplets smaller than 13 μm diameter were present. The small droplets, upon freezing onto a falling ice particle, provide sites for the larger droplets to come into point contact with the ice, attach, and freeze. Laboratory experiments indicate that supercooled drops accreting on to another already frozen droplet in this manner will freeze in distinct modes that depended on temperature. In the temperature range characteristic of ice multiplication, droplets freeze such that the outer shell of the droplet freezes first. The expanding ice builds internal pressure on the liquid in the interior of the drop until the shell ruptures at its weakest point ejecting liquid, which freezes creating a protuberance or a fragment. This process, called the Hallett–Mossop process after its discoverers, is believed to be important in many cloud systems.

The final mechanism that has been studied extensively is droplet fragmentation by shattering or splinter ejection during freezing. Droplet shattering during freezing appears to be limited to droplets $>50\,\mu\mathrm{m}$ diameter. Larger drops are more likely to fragment than smaller drops. Even when droplets do not fragment, the nature of the freezing process can cause them to eject an ice splinter as the outer shell ruptures. Because of the large droplet criteria, this process tends to be favored in maritime clouds with larger drop distributions, and much less favored in continental clouds, particularly those with large updrafts, such as cumulus.

Despite the fact that these ice multiplication mechanisms are well understood, they still cannot account for the observed high ice particle concentrations in many clouds. One attractive hypothesis that fits the observations is that rapid nucleation of ice occurs at cloud boundaries, particularly cloud top. It has been shown observationally and theoretically that a narrow layer of supercooled water frequently exists at the top of cold clouds, even at temperatures $<-20^\circ\mathrm{C}$. This liquid layer is believed to exist because shear, radiation, and entrainment enhance vertical velocities near cloud top, and because ice crystals are naturally very small near cloud top. Because of these factors, the condensate supply rate frequently exceeds the bulk ice particle growth rate and liquid is produced. Broad droplet spectra are often present in these regions. Entrainment of dry air leads to evaporation of droplets near cloud top. Contact nucleation may be favored in these conditions. Also, evaporation of drops can lead to giant aerosol particles that, in deep cumulus, may carry significant charge. These particles may be more effective as ice nuclei. At present, the importance of each of these ice-forming mechanisms in different clouds is speculative at best. Clearly, more research is required before we understand all the possibilities and processes associated with the formation of the ice phase in clouds.

LIST OF SYMBOLS

A	constant in IN concentration equation
a	ice crystal axis along the basal plane
C	constant in CCN, IN concentration equations, capacitance factor
C_v	specific heat of air at constant volume
c	ice crystal axis along prism plane
D	diffusion coefficient
E_{col}	collision efficiency
E_{coal}	coalescence efficiency
E	collection efficiency

e'_r	vapor pressure over a solution droplet
e_r	vapor pressure over a spherically curved water surface of radius r
$e_{r,i}$	vapor pressure over a spherically curved ice surface of radius r
$e_{s,\infty}$	saturation vapor pressure over a plane water surface
$e_{si,\infty}$	saturation vapor pressure over a plane ice surface
F	Helmholz free energy
ΔF_i	free energy change at ice–liquid interface
$\Delta F'$	free energy change associated with molecular realignment
f_v	ventilation coefficient
h	Planck's constant
i	van't Hoff factor
J	nucleation rate
K	thermal conductivity, collision kernel
k	Boltzmann constant
k	constants in CCN, IN concentration equations
L_v	latent heat of vaporization
L_s	latent heat of sublimation
$n(r)$	number of aerosol particles in the size range r, $r + d \log r$
m	mass of a drop
m_s	mass of solute
$M(r)$	total mass of aerosol particles with radii greater than r
M_w	molecular weight of water
M_s	molecular weight of the solute
N	number concentration
N_a	Avagadro's number
N_c	number of water molecules in contact with ice surface
N_{CCN}	number of CCN active at a specified supersaturation
N_{IN}	number of ice nuclei
$N(r)$	total number of aerosol particles with radii greater than r
N_{Re}	Reynolds number
n_0	number of molecules of water
n_s	number of molecules of solute
p	pressure
R	large drop radius
R^*	universal gas constant
R_d	gas constant for dry air
R_v	gas constant for water vapor
RH_w	relative humidity with respect to water
r	particle radius
r_{eq}	equilibrium radius of a droplet
$S(r)$	total surface area of aerosol particles with radii greater than r
$S_{w,v}$	saturation ratio over a water surface
$S_{s,v}$	saturation ratio over a solution droplet
$S_{i,v}$	saturation ratio over an ice surface
s_w	supersaturation with respect to water

s_i	supersaturation with respect to ice
T	temperature
T_s	surface temperature of a hailstone
t	time
$V(r)$	total volume of aerosol particles with radii greater than r
V_T	terminal velocity
$V_{T,H}$	terminal velocity of a hailstone
x_0	collision cross section
α	accomodation coefficient
β	condensation coefficient, constant in IN concentration equation
ν	kinematic viscosity
ρ	aerosol particle density
ρ_w	density of water
$\sigma_{w,v}$	surface tension at a water–vapor interface
$\sigma_{w,i}$	surface tension at a water–ice interface
$\sigma_{i,v}$	surface tension at an ice–vapor surface
ρ_i	density of ice

REFERENCES

Beard, K. V. (1992). Ice initiation in warm-base convective clouds: An assessment of microphysical mechanisms, *Atmos. Res.* **28**, 125–152.

Beard, K. V., and H. T. Ochs III (1984). Collection and coalescence efficiencies for accretion, *J. Geophys. Res.* **89**, 7165–7169.

Berry, E. X., and R. L. Reinhardt (1974). An analysis of cloud drop growth by collection: Parts I–IV, *J. Atmos. Sci.* **31**, 1814–1831, 2118–2135.

Best, A. C. (1951). The size of cloud droplets in layer-type cloud, *Quart. J. Roy. Met. Soc.* **77**, 241–248.

Bigg, E. K., and C. M. Stevenson (1970). Comparison of concentrations of ice nuclei in different parts of the world, *J. Rech. Atmos.* **4**, 41–58.

Chuang, C. C., and K. V. Beard (1990). A numerical model for the equilibrium shape of raindrops, *J. Atmos. Sci.* **47**, 1374–1389.

Cooper, W. A. (1989). Effects of variable droplet growth histories on droplet size distributions. Part I: Theory, *J. Atmos. Sci.* **46**, 1301–1311.

Fitzgerald, J. W. (1972). A study of the initial phase of cloud droplet growth by condensation: Comparison between theory and observation, Ph.D. Dissertation, Dept. of Geophys. Sci., University of Chicago.

Gunn, R., and G. D. Kinzer (1949). The terminal velocity of fall for water droplets in stagnant air. *J. Meteor.* **6**, 243–248.

Hallett, J., and S. C. Mossop (1974). Production of secondary particles during the riming process, *Nature* **249**, 26–28.

Jaenicke, R. (1988). In G. Fischer (Ed.), *Numerical Data and Functional Relationships in Science and Technology*, Landolt-Börnstein New Series, V: Geophysics and Space

Research, 4: Meteorology, b: Physical and Chemical Properties of Air, Berlin, Springer, pp. 391–457.

Junge, C. E. (1963). Large scale distribution of condensation nuclei in the troposphere*, J. Rech. Atmos.* **1**, 185–189.

Klett, J. D., and M. H. Davis (1973). Theoretical collision efficiencies of cloud droplets at small Reynolds numbers, *J. Atmos. Sci.* **30**, 107–117.

Latham, J., and R. L. Reed (1977). Laboratory studies of the effects of mixing on the evolution of cloud droplet spectra, *Quart. J. Roy. Meteor. Soc.* **103**, 297–306.

Lin, C. L., and S. C. Lee (1975). Collision efficiency of water drops in the atmosphere, *J. Atmos. Sci.* **32**, 1412–1418.

Magono, C., and C. W. Lee (1966). Meteorological classification of natural snow crystals, *J. Fac. Sci.* **7**(2), 321–362.

Pruppacher, H. R. (1995). A new look at homogeneous ice nucleation in supercooled water drops, *J. Atmos. Sci.* **52**, 1924–1933.

Pruppacher, H. R., and J. D. Klett (1997). *Microphysics and Clouds and Precipitation*, 2nd ed., Dordrecht, Kluwer Academic.

Rogers, R. R., and M. K. Yau (1989). *A Short Course in Cloud Physics*, Pergamon Press, Oxford, England.

Ryan, B. F., E. R. Wishart, and D. E. Shaw (1976). The growth rates and densities of ice crystals between $-3°C$ and $-21°C$, *J. Atmos. Sci.* **33**, 842–850.

Schlamp, R. J., S. N. Grover, H. R. Pruppacher, and A. E. Hamilec (1976). A numerical investigation of the effect of electric charges and vertical external electric fields on the collision efficiency of cloud drops, *J. Atmos. Sci.* **33**, 1747–1750.

Szumowski, M. J., R. M. Rauber, and H. T. Ochs III (1999). The microphysical structure and evolution of Hawaiian rainband clouds. Part III: A test of the ultragiant nuclei hypothesis, *J. Atmos. Sci.* **56**, 1980–2003.

Tabazadeh, A., and E. J. Jensen (1997). A model description for cirrus cloud nucleation from homogeneous freezing of sulfate aerosols, *J. Geophys. Res.* **102**, D20, 23845–23850.

Twomey, S., and T. A. Wojciechowski (1969). Observations of the geographical variation of cloud nuclei, *J. Atmos. Sci.* **26**, 684–688.

Vaillancourt, P. A., and M. K. Yau (2000). Review of particle-turbulence interactions and consequences for cloud physics, *Bull. Amer. Met. Soc.* **81**, 285–298.

Vali, G. (1985). Atmospheric ice nucleation—a review, *J. Rech. Atmos.* **19**, 105–115.

Wallace, J. M., and P. V. Hobbs (1977). *Atmospheric Science, An Introductory Survey*, Academic Press, Orlando, Florida, p. 163.

Whitby, K. T. (1978). The physical characteristics of sulfur aerosols, *Atmos. Environ.* **12**, 135–159.

World Meteorological Organization (1987). *International Cloud Atlas*, Vol. II, Geneva, World Meteorological Organization.

CHAPTER 19

RADIATION IN THE ATMOSPHERE: FOUNDATIONS

ROBERT PINCUS AND STEVEN A. ACKERMANN

1 OVERVIEW

“What is the weather like?” Meteorologists answer this question in thousands of locations every day when we make observations to determine the state of the atmosphere. The answer is summarized in a few carefully chosen quantities, typically wind speed and direction, relative humidity, and temperature. In the language of physics these roughly correspond to momentum, mass (of water), and energy. Weather forecasting is the science and craft of predicting how these interrelated quantities will change with time.

Energy is transferred through the atmosphere via five processes: convection, advection, conduction, phase change, and radiation. The first two processes involve the movement of mass from one place to another; conduction occurs when two bodies at different temperatures are in contact; phase changes release latent heat. Radiation is fundamentally different because it allows energy to be transferred between two locations without any intervening material.

Why Study Radiation?

Because radiation can transport energy even without a medium, it is the only way in which Earth interacts with the rest of the universe. It is radiation, in fact, that determines Earth’s climate, since in the long term the planet must shed as much energy as it absorbs. Even within the Earth–atmosphere system radiation can be a powerful player in determining the local energy budget. Radiation is why it is usually

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

warmer during the day, when the sun shines, than at night, and why the surface air temperature is higher on cloudy nights than clear ones.

Radiation is also the basis for remote sensing, the ability to measure the state of the atmosphere from a remote location (usually the ground or outer space). Remote sensing includes everything from simple cloud imagery, through radar estimates of precipitation, to ground-based sounding. Understanding the capabilities and limits of remote-sensing measurements requires learning something about radiation.

Earth's atmosphere is primarily made of gases, but it also contains liquids and solids in the form of aerosols and clouds. Radiation interacts in fundamentally different ways with gases and condensed materials. Gases, as we will see, interact with radiation in just a few ways but have complicated spectral structure. Clouds and aerosols, on the other hand, affect the radiation in each spectral region in about the same way but make the mathematics much more complicated.

Nature of Radiation

In the atmospheric sciences, the word *radiation* means "electromagnetic radiation," which is different than the particle radiation emitted from radioactive decay. Visible light is one kind of electromagnetic radiation, as are gamma- and x-rays, ultraviolet, infrared, and microwave radiation, and radio waves.

When radiation is measured using very sensitive instruments at extremely low light levels, it is observed that the energy does not arrive continuously but rather in small, finite amounts. These and other observations described in physics texts, including the photoelectric and Compton effects, suggest that radiation can be thought of as a collection of photons, tiny but discrete packets of energy traveling at the speed of light. This is the particle view of radiation.

We can also describe radiation as an electromagnetic phenomenon. The interactions among electric and magnetic fields, matter, charges, and currents are described by Maxwell's equations. The only nonzero solution to these equations in empty space is a traveling wave. Constants in Maxwell's equations predict the wave velocity, which is exactly the speed of light. Light behaves as a wave in many circumstances, too, diffracting when passed through a slit and reflecting from discontinuities in the medium. These observations are the motivation to describe radiation in purely electromagnetic terms.

How do we reconcile these two views? It is tempting to say that light can be both wave and particle, but this is not quite accurate; rather, there are circumstances in which light behaves like a wave and others in which it behaves like a particle. In this and the following chapter we will primarily use the wave model, which is usually more useful in the context of meteorology.

Radiation is also the single aspect of atmospheric science in which quantum mechanics plays a role. This theory, developed in the first few decades of the twentieth century, is based on the idea that the world is not continuous at very small scales, but is divided up ("quantized") into discrete elements: An electron's angular momentum, for example, can take on only certain values. As we will see, a complete description of radiation requires us to invoke ideas from quantum mechanics several times.

2 FOUNDATIONS

Geometry

The directional aspects of radiative transfer are most naturally expressed in spherical coordinates, illustrated in Figure 1. Location is specified by radius r from the origin, zenith angle θ, and azimuthal angle φ, with differential increments dr, $r\,d\theta$, and $r\ \sin\theta\,d\varphi$. The amount of radiation depends on direction, which in three dimensions is specified with solid angle Ω, measured in steradians (str). The differential element $d\Omega$ is the product of the polar and azimuthal angle differentials. There are 4π steradians in a sphere:

$$\int_0^{2\pi}\int_0^{\pi}\sin\theta\,d\theta\,d\varphi = \int_{\Omega} d\Omega = 4\pi \tag{1}$$

Polar angle is often replaced with its cosine $\mu = \cos\theta$, $d\mu = \sin\theta\,d\theta$, so that $d\Omega = d\mu\,d\varphi$. Polar angle is measured relative to a beam directed upwards, so $\theta = \pi$ and $\mu = -1$ for a beam pointing straight down, and $\mu > 0$ for radiation traveling upwards.

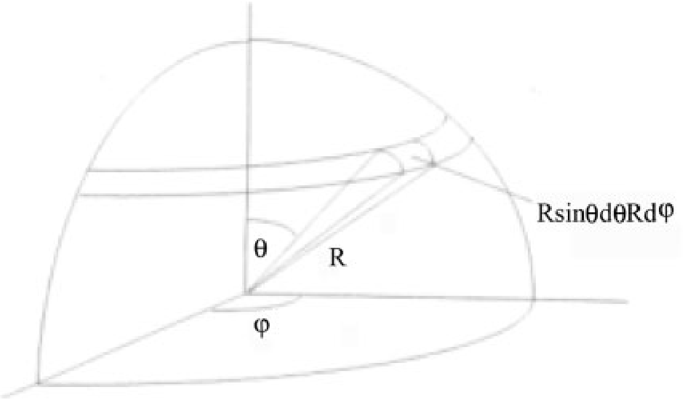


Figure 1 Geometry in polar coordinates. Radius r is measured from the origin, zenith angle θ from the vertical, and azimuthal angle φ from the south.

Describing Electromagnetic Waves

Electromagnetic waves can be characterized in terms of their velocity c, the frequency of oscillation v, and the wavelength λ. These quantities are related:

$$c = \lambda v \tag{2}$$

Wavelength is the distance between successive maxima of field strength and has dimensions of length, while frequency has dimensions of inverse time, so velocity is measured in distance per time. In a vacuum the speed of light $c = 3 \times 10^8\ \mathrm{m/s}$. This value can change in other materials depending on the index of refraction m, which varies with frequency. In air $m \approx 1$ so c is nearly unchanged; in water at visible wavelengths $m \approx 1.33$, so within cloud drops the speed of light is diminished by about 25%. The inverse of wavelength is the wavenumber $k = 1/\lambda$. In atmospheric applications wavelength is commonly measured in microns ($1\ \mu\mathrm{m} = 10^{-6}\ \mathrm{m}$), nanometers ($1\ \mathrm{nm} = 10^{-9}\ \mathrm{m}$), or angstroms ($1\ \mathrm{\AA} = 10^{-10}\ \mathrm{m}$), with frequency in megahertz ($1\ \mathrm{MHz} = 10^6\ \mathrm{s}^{-1} = 0^6\ \mathrm{Hz}$) or gigahertz ($1\ \mathrm{GHz} = 10^9\ \mathrm{Hz}$) and wavenumber expressed in inverse centimeters.

Name	*Spectral Region*
X-ray	$\lambda < 10\ \mathrm{nm}$
Ultraviolet (UV)	$10 < \lambda < 400\ \mathrm{nm}$
Visible	$0.4 < \lambda < 0.7\ \mu\mathrm{m}$
Near-infrared (near-IR)	$0.7 < \lambda < 3.5\ \mu\mathrm{m}$
Middle-IR	$3.5 < \lambda < 30\ \mu\mathrm{m}$
Far-IR	$30 < \lambda < 100\ \mu\mathrm{m}$
Microwave	$1\ \mathrm{mm} < \lambda < 1\ \mathrm{m}$

The plane in which the electric field oscillates determines the polarization of the radiation. In the atmosphere, though, this plane is rarely constant (i.e., radiation in the atmosphere is usually unpolarized) so we'll ignore this aspect.

Describing Radiation

If we stand in a field on a clear day and look around the sky, we will notice that lots of light comes from the sun and little from the other directions. It is brighter in the open field than in the shadow of a tree, and darker at night than at noon. The amount of radiation, then, depends on space, time, direction, and wavelength. Because light travels so fast, we usually ignore the time dependence.

The fundamental measure of radiation is the amount of energy traveling in a given direction at a certain wavelength. This measure is called spectral intensity I_λ and has dimensions of power per unit area per solid angle per spectral interval, or units of $\mathrm{W/m^{-2}\,str\,\mu m}$. Spectral intensity is assumed to be monochromatic, or consisting of exactly one wavelength, and depends on position.

If we want to know the total amount of energy traveling in a given direction (say, the amount of energy entering a camera lens or a satellite detector), we must integrate I_λ across some portion of the spectrum to compute intensity (or broadband intensity) I:

$$I = \int I_\lambda \, d\lambda \tag{3}$$

The limits of integration in (3) depend on the application. If the film in our camera is sensitive only to visible light, for example, the integration is over the visible portion of the spectrum, while if the lens is behind a colored filter we include only those wavelengths the filter passes. Intensity has units of W/m^2 str; we use the term broadband intensity when the integration is over a large part of the spectrum (e.g., all the infrared or all the visible). Unless the radiation interacts with the medium neither I_λ nor I change with distance.

Imagine next a sheet of back plastic placed on the ground in the sunlight. How much energy does the sheet absorb? If we start with just one wavelength, we see that the sheet absorbs energy at a rate F_λ

$$F_\lambda = \int_{-1}^{1} \int_{0}^{2\pi} I_\lambda |\mu| \, d\varphi \, d\mu \tag{4}$$

Weighting by μ accounts for geometry: A ray encountering the surface at an angle is spread over a wider area than a beam coming straight at the surface. F_λ is called the spectral flux and has units of W/m^2 μm. Flux is traditionally divided into upward- and downward-going components:

$$\begin{aligned} F_\lambda^\downarrow &= \int_{-1}^{0} \int_{0}^{2\pi} I_\lambda |\mu| \, d\varphi \, d\mu \\ F_\lambda^\uparrow &= \int_{0}^{1} \int_{0}^{2\pi} I_\lambda |\mu| \, d\varphi \, d\mu \end{aligned} \tag{5}$$

The black sheet absorbs at all wavelengths, so the total amount of energy absorbed F is computed by integrating over both solid angle and spectral interval:

$$F^\downarrow = \int_{-1}^{0} \int_{0}^{2\pi} \int I_\lambda |\mu| \, d\lambda \, d\varphi \, d\mu = \int_{-1}^{0} \int_{0}^{2\pi} I |\mu| \, d\lambda \, d\varphi \, d\mu = \int F_\lambda^\downarrow \, d\lambda \tag{6}$$

Flux (or broadband flux) has units of W/m^2. The terms *radiance* and *irradiance* are also used in textbooks and the technical literature; these correspond to our terms *intensity* and *flux*.

Why are so many different quantities used to describe radiation? Because the two main applications of radiation, remote sensing and energy budget computations, require fundamentally different kinds of information. Remote-sensing instruments, for example, usually have a finite field of view and a finite spectral sensitivity; interpreting measurements from these sensors therefore requires calculations of intensity, which may be broadband, narrowband, or essentially monochromatic, depending on the detector.

Imagine, though, wanting to compute the rate at which radiation heats or cools a layer of air in the atmosphere. Here we need to know the net amount of energy remaining in the layer, which we can calculate by considering the decrease in both upwelling and downwelling fluxes as they cross a layer between z_1 and z_2:

$$\begin{aligned} E_{\text{in}} &= F^{\uparrow}(z_1) - F^{\uparrow}(z_2) + F^{\downarrow}(z_2) - F^{\downarrow}(z_1) \\ &= F^{\uparrow}(z_1) - F^{\downarrow}(z_1) - [F^{\uparrow}(z_2) - F^{\downarrow}(z_2)] \\ &= F^{\text{net}}(z_1) - F^{\text{net}}(z_2) \end{aligned} \tag{7}$$

where we have defined the net flux as $F^{\text{net}}(z) = F^{\uparrow}(z) - F^{\downarrow}(z)$. As z_1 and z_2 get very close together, the difference in (7) becomes a differential, and the rate of heating can be related to the divergence of the radiative flux though the air density ρ and heat capacity c_p:

$$\frac{dT(z)}{dt} = -\frac{1}{\rho c_p}\frac{dF^{\text{net}}(z)}{dz} \tag{8}$$

3 SOURCES OF RADIATION

Imagine setting a cast-iron skillet over a really powerful burner and turning on the gas. At first the pan appears unchanged, but as it heats it begins to glow a dull red. As the pan becomes hotter it glows more brightly, and the color of the light changes too, becoming less red and more white.

In the idealized world inhabited by theoretical physicists, the cast-iron skillet is replaced by a block. The block has a cavity inside it and a small hole opening into the cavity. The block can be heated to any temperature. Measurements of the radiation emerging from this cavity show that:

1. The spectral intensity B_λ emerging from the cavity at each wavelength depends only on the temperature T of the block, and not on the material or the shape of the cavity.
2. The total amount of energy emitted by the cavity increases as the temperature increases.
3. The wavelength at which B_λ reaches its maximum value decreases with temperature.

The Planck Function

A theoretical explanation for cavity radiation was the single most compelling unsolved problem in physics at the beginning of the twentieth century. The best fit to observations had been suggested by Wien:

$$B_\lambda(T) = \frac{c_1}{\lambda^5}\frac{1}{e^{c_2/\lambda T}} \tag{9}$$

where c_1 and c_2 were determined experimentally. In October 1900 Max Planck proposed a small modification to this relationship, namely

$$B_\lambda(T) = \frac{c_1}{\lambda^5}\frac{1}{e^{c_2/\lambda T} - 1} \tag{10}$$

Planck's relationship, now called the *Planck function*, was a better fit to the data but was not an explanation for why radiation behaves this way. To develop a theoretical model, Planck imagined that the atoms in the cavity walls behave like tiny oscillators, each with its own characteristic frequency, each emitting radiation into the cavity and absorbing energy from it. But it proved impossible to derive the properties of cavity radiation until he made two radical assumptions:

1. The atomic oscillators cannot take on arbitrary values of energy E. Instead, the oscillators have only values of E that satisfy $E = nhv$ where n is an integer called the quantum number, v the oscillator frequency, and h a constant.
2. Radiation occurs only when an oscillator changes from one of its possible energy levels to another; that is, when n changes value. This implies that the oscillators cannot radiate energy continuously but only in discrete packets, or quanta.

By December of 1900 Planck had worked out a complete theory for cavity radiation, including the values of c_1 and c_2:

$$c_1 = 2hc^2 \qquad c_2 = \frac{hc}{k} \tag{11}$$

In these equations k is Boltzmann's constant, which appears in statistical mechanics and thermodynamics, and h is Planck's constant. Planck used observations of $B_\lambda(T)$ to determine the values $h = 6.63 \times 10^{-34}$ J/s and $k = 1.38 \times 10^{-23}$ J/K. Equations (10) and (11) were the basis for the development of quantum mechanics, which fundamentally changed the way physicists looked at the world.

Blackbody Radiation and Its Implications

In the atmospheric sciences we refer not to cavity radiation but to *blackbody radiation*, a blackbody being anything that radiates according to (10). Blackbody radiation is isotropic, meaning that it is emitted in equal amounts in all directions.

Wavelength of Maximum Emission

The wavelength at which blackbody radiation reaches its maximum intensity can be found by taking the derivative of $B_\lambda(T)$ with respect to wavelength, setting the result to 0, and solving for λ. That is, we solve

$$\frac{\partial B_\lambda(T)}{\partial \lambda} = 0 \tag{12}$$

for the wavelength λ_{max}, which yields *Wien's displacement law*

$$\lambda_{\text{max}} = 2898\,\mu\text{m}\;K/T \tag{13}$$

At temperatures typical of Earth's atmosphere (say, $T = 288$ K), the maximum wavelength is about 10 μm in the infrared portion of the spectrum, while the sun (at a temperature about 5777 K) is brightest at about 0.5 μm, where visible light is green.

Total Amount of Energy Emitted

How much energy does a blackbody radiate at a given temperature? This quantity is the blackbody broadband intensity and is computed by integrating over all wavelengths:

$$B(T) = \int_0^\infty B_\lambda(T)\,d\lambda = \frac{\sigma T^4}{\pi} \tag{14}$$

Equation (14) is the *Stefan–Boltzmann relation* and makes use of the Stefan–Boltzmann constant $\sigma = 5.67 \times 10^{-8}\,\text{W/m}^2\,\text{K}^{-4}$. Because blackbody radiation is isotropic, the total amount of radiation lost by an object into each hemisphere is

$$F(T) = \int_0^{2\pi} \int_0^1 B(T)\mu\,d\mu\,d\varphi = \sigma T^4 \tag{15}$$

Figure 2 shows $B_\lambda(\lambda, T)$ plotted for two values of T, roughly corresponding to the average surface temperatures of Earth and sun. The curves are each normalized, since otherwise the much greater intensity produced at solar temperatures would swamp the intensity produced by terrestrial objects. Almost all the sun's energy is produced at wavelengths less than about 4 μm, while almost all the energy emitted

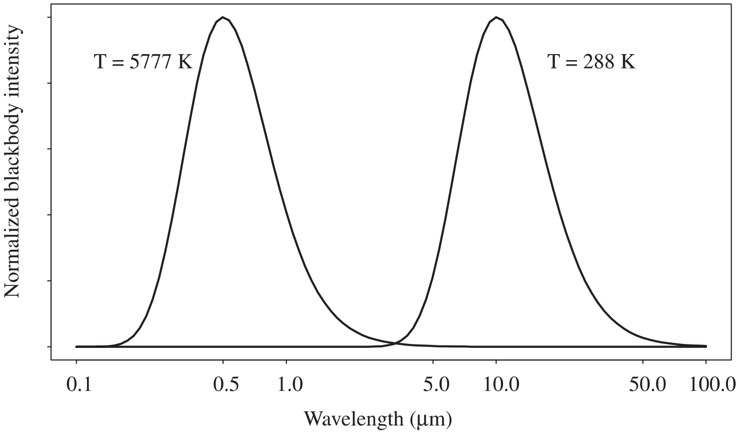


Figure 2 Blackbody intensity B_λ as a function of wavelength for temperatures corresponding to the surfaces of Earth and sun. The curves are arbitrarily normalized.

by Earth is produced at wavelengths longer than 4 μm; this convenient break lets us treat solar radiation and terrestrial radiation as independent.

Blackbody radiation helps us understand the light produced by our cast-iron skillet. The pan emits radiation even at room temperature, although the emission is mostly in the infrared portion of the spectrum. As the skillet is heated, the total amount of radiation emitted increases (as per the Stefan–Boltzmann relation) and the wavelength of maximum emission gets shorter (as per Wien's displacement law) until some of the emission is in the longer (red) part of the visible spectrum. If the pan were made even hotter, it would glow white, like the filament in an incandescent light bulb.

Emissivity, Energy Conservation, Brightness Temperature

The radiation emitted by any object can be related to the blackbody radiation

$$I_\lambda(T) = \varepsilon_\lambda B_\lambda(T) \tag{16}$$

where ε_λ is the emissivity of the object, which varies between 0 and 1. If ε_λ does not depend on wavelength, we say that the object is a gray body; a blackbody has a value of $\varepsilon_\lambda = 1$.

How are emission and absorption related? Imagine an object that absorbs perfectly at one wavelength but not at all at any other wavelength (i.e., an object with $\varepsilon_\lambda = 1$ at one value of $\lambda = \lambda^*$ and $\varepsilon_\lambda = 0$ everywhere else), illuminated by broadband blackbody radiation from a second body. The object absorbs the incident radiation and warms; as it warms the emission (which occurs only at λ^*, remember) increases. Equilibrium is reached when the amount of energy emitted by the particle E_{out} is equal to the amount absorbed E_{in}. At wavelength λ the body acts as a blackbody, so emission depends only on the equilibrium temperature and $E_{\text{out}} = I_{\lambda^*}(T) = B_{\lambda^*}(T)$. The body is exposed to broadband blackbody radiation, so $E_{\text{in}} = B_{\lambda^*}(T)$. But since equilibrium implies that $E_{\text{in}} = E_{\text{out}}$, the absorption at every wavelength other than λ^* must be zero. This chain of reasoning, known as *Kirchoff's law*, tells us that the absorptivity and emissivity of objects is the same at every wavelength.

The Planck function finds another application in the computation of brightness temperature. If we make measurements of monochromatic intensity I_m at some wavelength λ, and assume that $\varepsilon_\lambda = 1$, we can invert the Plank function to find the temperature T_b at which a blackbody would have to be in order to produce the measured intensity

$$T_b = \frac{c_2}{\lambda} \frac{1}{\ln(c_1/I_m \lambda^5 + 1)} \tag{17}$$

where T_b is called the equivalent blackbody temperature or, more commonly, *brightness temperature*. It provides a more physically recognizable way to describe intensity.

4 ABSORPTION AND EMISSION IN GASES: SPECTROSCOPY

The theory of blackbody radiation was developed with solids in mind, and in the atmospheric sciences is most applicable to solid and liquid materials such as ocean and land surfaces and cloud particles. The emissivity of surfaces and suspended particles is generally high in the thermal part of the spectrum, and tends to change fairly slowly with wavelength.

In gases, however, emissivity and absorptivity change rapidly with wavelength, so blackbody radiation is not a useful model. Instead, it is helpful to consider how individual molecules of gas interact with radiation; then generalize this understanding by asking about the behavior of the collection of molecules in a volume of gas. We will find that the wavelengths at which gases absorb and emit efficiently are those that correspond to transitions between various states of the molecule; understanding the location and strength of these transitions is the subject of spectroscopy.

Spectral Lines: Wavelengths of Absorption and Emission

Imagine the simplest possible atom, a single electron of mass m_e circling a single proton. In a mechanical view the electrostatic attraction between the unlike charges balances the centripetal acceleration of the electron, so the velocity v of the electron, and therefore its angular momentum, are related to the distance r between the electron and the proton. In 1913 Niels Bohr postulated that the angular momentum of the electron is quantized, and can take on only certain values such that

$$m_e vr = \frac{nh}{2\pi} \tag{18}$$

where n is any positive integer. This implies that the electron can only take on certain values of v and r, so the total energy (kinetic plus potential) stored in the atom is quantized as well. If the energy levels of the atom are discrete, the changes between the levels must also be quantized.

It is these specific changes in energy levels that determine the *absorption and emission lines*, the wavelengths at which the atom absorbs and emits radiation. The energy of each photon is related to its frequency and so its wavelength

$$E = h\nu = \frac{hc}{\lambda} \tag{19}$$

A photon striking an atom may be absorbed if the energy in the photon corresponds to the difference in energy between the current state and another allowed state. In a population of atoms or molecules (i.e., in a volume of gas) collisions between molecules mean that there are molecules in many states, so a volume of gas has many absorption lines.

In the simple Bohr atom the energy of the atom depends only on the state of the electron. Polyatomic molecules can contain energy in their electronic state, as well as in their vibrational and rotational state. The energy in each of these modes is quantized, and photons may be absorbed when their energy matches the difference between two allowable states in any of the modes.

The largest energy differences (highest frequencies and shortest wavelengths, with absorption/emission lines in the visible and ultraviolet) are associated with transitions in the electronic state of the molecules. At the extreme, very energetic photons can completely strip electrons from a molecule. Photodissociation of ozone, for example, is the mechanism for stratospheric absorption of ultraviolet radiation.

Energy is also stored in the vibration of atoms bound together in a stable molecule. Vibrational transitions give rise to lines in near-infrared and infrared, between those associated with electronic and rotational transitions. The amount of energy in each vibrational mode of a molecule depends on the way the individual atoms are arranged within the molecule, on the mass of the atoms, on the strength of the bonds holding them together, and on the way the molecule vibrates. Vibrational motion can be decomposed into normal modes, patterns of motion that are orthogonal to one another. In a linear symmetric molecule such as CO_2, for example, the patterns are symmetric stretch, bending, and antisymmetric stretch, as shown in Figure 3. The state of each normal mode is quantized separately.

The way atoms are arranged within a molecule also plays a role in how energy is stored in rotational modes. Carbon dioxide, for example, contains three atoms arranged in a straight line, and thus has only one distinct mode of rotation about

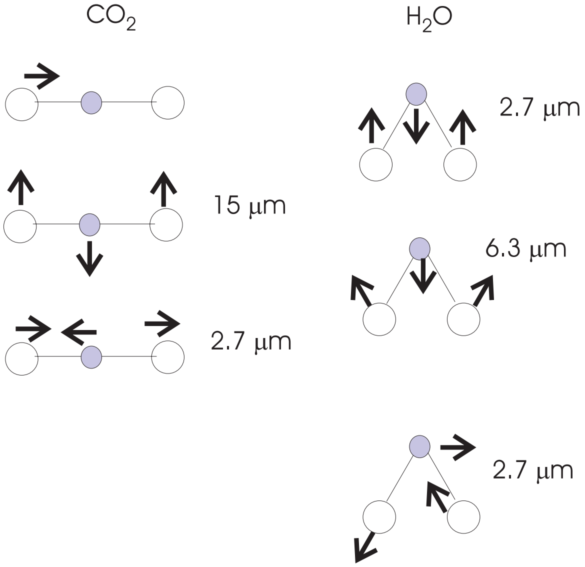


Figure 3 Vibrational normal modes of CO_2 and H_2O molecules. Any pattern of vibration can be projected on to these three modes, which are all orthogonal to one another. Also shown are the wavelengths corresponding to each vibrational mode. See ftp site for color image.

the center molecule and one quantum number associated with rotation. Complicated molecules like ozone or water vapor, though, have three distinct axes of rotation, and so three modes of rotation and three quantum numbers associated with the rotational state, each of which may change independently of the others. For this reason the absorption spectrum of water is much more complicated than that of CO_2.

Changes in vibrational and rotational states may also occur simultaneously, leading to a very rich set of absorption lines. Because the changes in energy associated with rotation are so much smaller than those associated with vibration, the absorption spectrum appears as a central line with many lines (each corresponding to a particular rotational change) clustered around it, as shown in Figure 4. More energetic lines (with higher wavenumbers and shorter wavelengths) are associated with simultaneous increases in the rotational and vibrational state, while the lines to the left of the central line occur when rotational energy is decreased while vibrational energy increases.

What gives rise to the family of absorption lines flanking the central wavenumber in Figure 4? Carbon dioxide has only one axis of rotation, so the different energies are all associated with the same mode, but the family of lines implies that there is a set of rotational transitions, each with a different energy. If J is the quantum number describing the rotational states, and ΔJ the change between two states, there are two possibilities. The different lines might correspond to different values of ΔJ, but this is prohibited by quantum mechanics. In fact, each line corresponds to $\Delta J = 1$, but neighboring states differ in energy by various amounts depending on J.

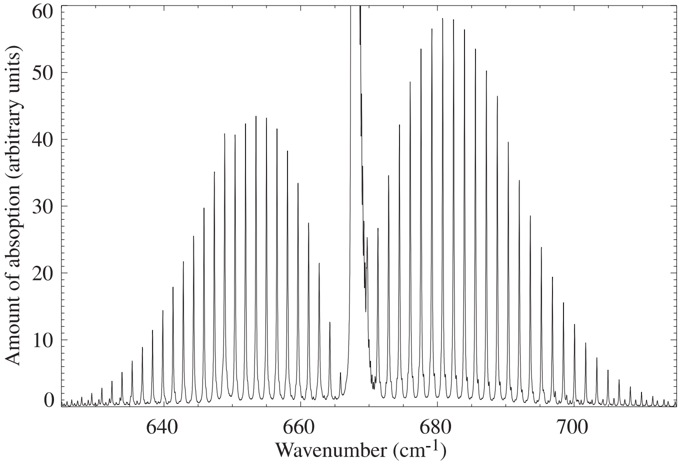


Figure 4 Pure vibration and vibration-rotation lines near the 15 μm (666 cm^{-1}) absorption line of carbon dioxide. The spectrum is computed for gases at 300 K and 1000 mb.

Line Strength

The quantum mechanics of electronic, vibrational, and rotational transitions determines the location of absorption and emission lines in gases. Molecules do not float freely in the atmosphere, however, but rather are members of a large ensemble, or population. The temperature of the gas determines the mean energy of the molecules, and random collisions between them redistribute the energy among the possible states. The amount of absorption due to any particular transition, then, is the product of how likely the transition is (as determined by quantum mechanics) and the fraction of molecules in the originating state. The effectiveness of absorption is called the *line strength* S and has dimensions of area per mass or per molecule.

Line strength depends in part on temperature. The atmosphere is almost everywhere in thermal equilibrium, which means that the average amount of energy per molecule is determined by the temperature, but this energy is constantly being redistributed among molecules by frequent collisions. The collisions also transfer energy between different states, so that a fast-moving molecule may leave a collision moving more slowly but rotating more rapidly. The abundance of molecules in a state with energy E depends on the ratio of E to the average kinetic energy kT. This explains the increase and decrease of line strengths to either side of the central line in Figure 4: The most common rotational state is $J = 7$ or $J = 8$, and the number of molecules in other states, and so the line strength, decreases the more J differs from the most common value.

Line Shape

Each molecular transition, be it electronic, vibrational, or rotational, corresponds to a particular change in energy, which implies that absorption and emission take place at exactly one frequency. But as Figure 4 shows, absorption and emission occur in a narrow range of wavelengths surrounding the wavelength associated with the transition. We say that absorption lines are affected by natural broadening, pressure (or collision) broadening, and Doppler broadening. We describe these effects in terms of a line shape function $f(\nu - \nu_0)$, which describes the relative amount of absorption at a frequency ν near the central transition frequency ν_0. Line broadening affects the spectral distribution of absorption but not line strength.

Natural Broadening of Absorption Lines

Natural broadening of absorption lines has a quantum mechanical underpinning. Because molecules spend a finite amount of time in each state, the Heisenberg uncertainty principle tells us that the energy of the state must be somewhat uncertain. This blurring of energy levels implies a similar blurring in the transition energies between levels, though the effect is small compared to other broadening mechanisms.

Doppler Broadening of Absorption Lines

If we stand beside a racetrack and listen to an approaching car, the whine of the engine will seem to increase in pitch, then decrease as the car passes us. This change in frequency is known as the Doppler shift and occurs whenever an object and observer move relative to one another. The frequency ν of the emitted sound or light appears to increase or decrease from its unperturbed value ν_0 depending on how fast the object moves v relative to the speed of the wave c:

$$\nu = \nu_0\left(1 \pm \frac{v}{c}\right) \tag{20}$$

Because the atmosphere is almost always in thermal equilibrium, the velocities of a collection of gas molecules follow a Maxwell–Boltzmann distribution. The probability $p(v)$ that a molecule will have a radial velocity v depends on the temperature T, since molecules move faster at higher temperatures, and on the mass m of the molecules

$$p(v) = \sqrt{\frac{m}{2\pi kT}}\exp\left(-\frac{mv^2}{2kT}\right) \tag{21}$$

We can compute the Doppler line shape $f_D(\nu - \nu_0)$ around an absorption line by combining (20) and (21):

$$f_D(\nu - \nu_0) = \frac{1}{\alpha_D\sqrt{\pi}}\exp\left(-\frac{(\nu - \nu_0)^2}{\alpha_D^2}\right) \tag{22}$$

where we have defined the Doppler line width

$$\alpha_D = \frac{v_0}{c}\sqrt{\frac{2kT}{m}} \tag{23}$$

Absorption lines associated with heavier molecules are broadened less than those associated with light molecules, since for the same mean thermal energy (as measured by temperature) heavier molecules move more sluggishly than light ones. Doppler broadening also increases in importance at shorter wavelengths. It has the greatest effect high in the atmosphere, where pressure is low.

Pressure Broadening of Absorption Lines

When pressure is high, however, collision or pressure broadening is the most important process determining the shape of absorption and emission lines. The exact mechanisms causing pressure broadening are so complicated that there is no exact theory. What is clear, however, is that the width of the absorption line increases as the frequency of collisions between molecules increases. Pressure-broadened spectral lines are well-characterized by the Lorentz line profile:

$$f_L(v - v_0) = \frac{\alpha_L}{\pi}\frac{1}{(v - v_0)^2 + \alpha_L^2} \tag{24}$$

The Lorentz line width α_L depends on the mean time t between collisions:

$$\alpha_L = \frac{1}{2\pi t} \tag{25}$$

In practice the value of α_L is determined for some standard pressure and temperature p_0 and T_0, then scaled to the required pressure and temperature using kinetic theory:

$$\alpha_L = \alpha_{L0}\frac{p}{p_0}\left(\frac{T_0}{T}\right)^n \tag{26}$$

where n varies but is near $\frac{1}{2}$. When molecules collide with others of the same species, we say that the line is self-broadened; when the collisions are primarily with other gases we say the line is subject to foreign broadening. For the same width pressure broadening affects the line wings, those frequencies further from the central frequency, more dramatically than Doppler broadening, as Figure 5 demonstrates.

Both pressure and Doppler broadening can occur simultaneously. The line shape that results, the convolution of $f_L(v - v_0)$ and $f_D(v - v_0)$, is called the Voigt line shape, which does not have an exact analytic form. It is instructive, though, to

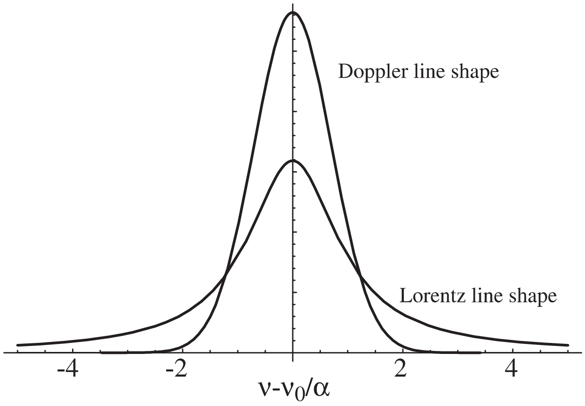


Figure 5 Lorentz and Doppler line shapes for equivalent line strengths and widths. Line wings are more strongly affected by pressure than Doppler broadening.

examine the relative importance of Doppler and pressure broadening before invoking anything more complicated. A useful approximation is

$$\frac{\alpha_L}{\alpha_D} \approx 10^{-12} \frac{\nu_0}{p} \tag{27}$$

when ν_0 is measured in hertz and p in millibars. Throughout most of the spectrum and most of the atmosphere this ratio is much less than one, telling us that the Lorentz line profile is usually a pretty good description of line shape.

Practical Applications

Absorption spectra almost never need to be determined directly from spectroscopy. Spectroscopic databases (HITRAN is popular) have been compiled for all the gases in Earth's atmosphere. These databases contain the line centers and parameters describing line shape as a function of temperature and pressure, and can be accessed with standard programs (e.g., LBLRTM) that provide the amount of absorption at very high spectral resolution.

Radiative Transfer Equation for Absorption

We are at last ready to compute the fate of radiation as it travels through a medium. Radiation may be absorbed or emitted by the medium, and may also be *scattered* or redirected without a change in intensity. We will develop the radiative transfer equation by adding each process in turn in order to keep the mathematics clear.

We will begin by considering a medium that absorbs but does not emit or scatter radiation. This is the framework we might use, for example, to describe the absorption of ultraviolet or visible light by gases in the atmosphere, where relatively low temperatures mean that emission is effectively zero.

Imagine a pencil of monochromatic radiation crossing a small distance *ds* between two points S_1 and S_2, as illustrated in Figure 6. The amount of radiation absorbed along this path depends linearly on the amount of incident radiation (more incoming photons means a greater likelihood that a photon will strike a molecule), the gas density (higher density increases the number of molecules encountered), and on how effectively the molecules absorb radiation at this wavelength. These are the three factors that appear in the radiative transfer equation for absorption:

$$\frac{dI_\lambda}{ds} = -I_\lambda k_\lambda \rho \tag{28}$$

where ρ is the density of the absorbing gas and k_λ the mass absorption coefficient (m^2/kg). Notice that k_λ has the same units as absorption line strength *S*. In fact, the absorption coefficient at some wavelength due to one particular line *i* is the product of the line shape and the line strength:

$$k_i(v) = S_i f(v - v_0) \tag{29}$$

and k_λ is the sum contributions from all lines from all gases.

The radiative transfer equation can be integrated along the path between S_1 and S_2:

$$\int_{S_1}^{S_2} \frac{1}{I_\lambda} dI_\lambda = \ln I_\lambda \Big|_{S_1}^{S_2} = -\int_{S_1}^{S_2} k_\lambda \rho \, ds \tag{30}$$

$$I_\lambda(S_2) = I_\lambda(S_1) \exp\left(-\int_{S_1}^{S_2} k_\lambda \rho \, ds\right) \tag{31}$$

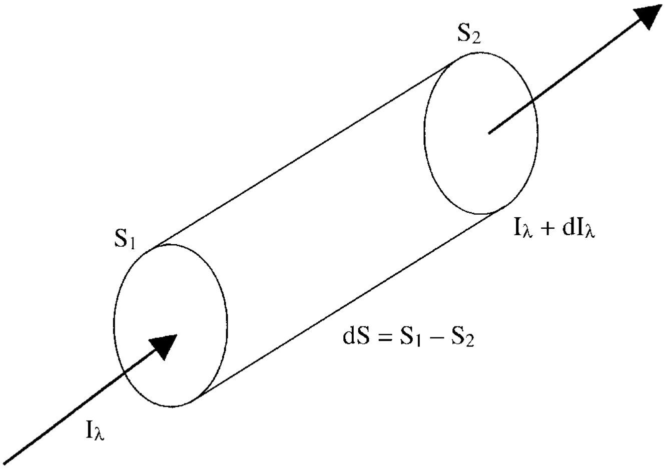


Figure 6 Beam of radiation passing through an absorbing medium.

where we leave the integral over the path unevaluated because ρ and k_λ may change with distance. We insist that the light be monochromatic since the integration does not make sense if k_λ is variable. Equation (31) is called *Beer's law*, or more formally the Beer–Lambert–Bouguer law.

How can we relate a path between arbitrary points S_1 and S_2 to the atmosphere? Two assumptions are usually made: that the atmosphere is much more variable in the vertical than in the horizontal, and that Earth is so large that the surface can be considered flat. These simplifications lead to the plane-parallel coordinate system, in which directions are specified through zenith angle θ and azimuthal angle φ but the medium varies only with vertical position z. In this system the path between S_1 and S_2 is related to the vertical displacement by $1/\mu$, so we may write

$$I_\lambda(z_2) = I_\lambda(z_1)\exp\left(-\frac{1}{\mu}\int_{z_1}^{z_2} k_\lambda \rho \, dz\right) \tag{32}$$

The integral in the exponential defines the *optical thickness* τ, also called the optical depth:

$$\tau = -\int_{z_1}^{z_2} k_\lambda \rho \, dz \tag{33}$$

The minus sign in (33) is an unfortunate hangover from the days when radiative transfer was dominated by astrophysicists. Since they were thinking about other stars, they set up a coordinate system in which $\tau = 0$ at the top of the atmosphere, so that $\mu \, d\tau < 0$. Though it is a little confusing, (33) does lead to much simpler forms of both the radiative transfer equation

$$\mu \frac{dI_\lambda}{d\tau} = I_\lambda \tag{34}$$

and to a more transparent form of Beer's law

$$I_\lambda(\tau_2) = I_\lambda(\tau_1)\exp\left[\frac{(\tau_2 - \tau_1)}{\mu}\right] = I_\lambda(\tau_1)\exp\left[\frac{-(\tau_2 - \tau_1)}{|\mu|}\right] \tag{35}$$

For downwelling radiation $\tau_2 > \tau_1$ and $\mu < 0$ and for upwelling radiation both signs are reversed, so intensity always decreases along the path. According to Beer's law, then, intensity in an absorbing medium falls off exponentially with optical depth.

The transmissivity T_λ of the layer between z_1 and z_2 is defined as

$$T_\lambda = \frac{I_\lambda(z_2)}{I_\lambda(z_1)} \tag{36}$$

which for an absorbing medium T_λ can be computed as

$$T_\lambda = \exp\left(\frac{-|\tau_2 - \tau_1|}{|\mu|}\right) \tag{37}$$

In a strictly absorbing medium the light that is not transmitted must be absorbed; so the absorptivity a_λ of the medium is

$$a_\lambda = 1 - T_\lambda = 1 - \exp\left(\frac{|\tau_2 - \tau_1|}{|\mu|}\right) \tag{38}$$

Computing Optical Depth along Inhomogeneous Paths

Optical depth is formally defined by (33), but computing its value is not always straightforward since both density and the absorption coefficient may change with height in the atmosphere. We can roughly account for the change in density with height by defining the specific gas ratio $q = \rho/\rho_{\text{air}}$ and invoking the hydrostatic equation $dp/dz = -\rho g$ in the definition of optical mass u:

$$u(p_1, p_2) = -\frac{1}{g}\int_{p_2}^{p_1} q\,dp \tag{39}$$

which has dimensions of mass per unit area.

The value of k_λ may also change with height if, for example, the wavelength in question is on the wing of an absorption line, so that the local strength changes as pressure and temperature change. We could adjust k_λ to some average value along the path. It is more common, however, to scale the optical mass to the reference temperature T_0 and pressure p_0 at which the mass absorption coefficient is measured. There are various approaches to scaling, though a common tactic is to compute a scaled absorber amount $\tilde{u}$ as

$$\tilde{u} = u\left(\frac{p_e}{p_0}\right)^m \left(\frac{T_0}{T_e}\right)^n \tag{40}$$

where m and n depend on the gas, and p_e and T_e are the effective pressure and temperature along the path, computed for temperature as

$$T_e = \frac{\int T\,du}{\int du} \tag{41}$$

and similarly for pressure.

Radiative Transfer Equation for Emission and Absorption

Imagine next a material that emits and absorbs at a wavelength λ. This might apply, for example, to the transfer of infrared radiation through Earth's atmosphere: Vibrational transitions in common gases give rise to absorption lines in the infrared, where temperatures through most of the atmosphere give rise to strong emission.

Kirchoff's law tells us that the emission and absorption per amount of material (or per unit of optical depth) must be the same, but emission adds to the intensity while absorption reduces it. Since $\mu\, d\tau < 0$, the radiative transfer equation is

$$\mu \frac{dI_\lambda}{d\tau} = I_\lambda - B_\lambda \tag{42}$$

This relation, known as Schwartzchild's equation, holds strictly for any medium that does not scatter radiation at the wavelength in question; (34) is a useful approximation when temperatures are such that B_λ is very small at the wavelength in question.

We can solve (42) using an integrating factor: We multiply both sides of the equation by $e^{-\tau/\mu}/\mu$ and collect terms in I_λ:

$$\mu e^{-\tau/\mu} \frac{dI_\lambda}{d\tau} - e^{-\tau/\mu} I_\lambda = -e^{-\tau/\mu} B_\lambda \tag{43}$$

or

$$\mu \frac{d}{d\tau} I_\lambda e^{-\tau/\mu} = -B_\lambda e^{-\tau/\mu} \tag{44}$$

To find the intensity propagating in direction μ at some vertical optical depth τ, we integrate (44) starting at one boundary, being careful to account for the sign of μ and for the fact that B_λ may vary with height. The downwelling intensity $I_\lambda^\downarrow$, at optical depth τ, for example, is

$$I_\lambda^\downarrow(\tau) e^{-\tau/\mu} - I_\lambda^\downarrow(0) = \frac{-1}{\mu} \int_0^\tau B_\lambda(\tau') e^{-\tau'/\mu}\, d\tau' \tag{45}$$

or

$$I_\lambda^\downarrow(\tau) = I_\lambda^\downarrow(0) e^{\tau/\mu} - \frac{1}{\mu} \int_0^\tau B_\lambda(\tau') e^{(\tau-\tau')/\mu}\, d\tau' \tag{46}$$

while the upwelling intensity is

$$I_\lambda^\uparrow(\tau) = I_\lambda^\uparrow(\tau^*) e^{-(\tau^*-\tau)/\mu} + \frac{1}{\mu} \int_\tau^{\tau^*} B_\lambda(\tau') e^{(\tau'-\tau^*)/\mu}\, d\tau' \tag{47}$$

where τ* refers to the bottom of the atmosphere.

Equations (46) and (47) tell us that the downwelling intensity at some height τ consists of the intensity incident at the boundary and attenuated by the intervening medium (the first term on the right-hand side) plus contributions from every other part of the medium (the integral over the blackbody contribution at every height), each attenuated by the medium between the source at τ' and the observation location τ.

As an example, imagine a satellite in orbit at the top of the atmosphere, looking straight down at the ground, which has emissivity 1 and is at temperature 294 K. At 10 μm, blackbody emission from the ground is found with (10) and (11) as $B_\lambda(294\,\mathrm{K}) = 2.85\,\mathrm{W/m^{-2}\,str\,\mu m}$. The clear sky is nearly transparent ($\tau \approx 0$) at 10 μm, so the upwelling nadir-directed ($\mu = 1$) intensity at the top of the atmosphere is essentially the same as the intensity at the ground. But if a thin ($\tau = 1$ at 10 μm), cold ($T = 195\,\mathrm{K}$) cirrus cloud drifts below the satellite, the outgoing intensity will be reduced. If we assume that the cloud has constant temperature, we can find the upwelling intensity with (47), using $\tau = 0$ at the top of the cloud, $\tau^* = 1$ at the cloud base, and $B_\lambda(\tau') = B_\lambda(195\,\mathrm{K})$ for $0 < \tau' < \tau^*$:

$$I_\lambda^{\uparrow}(\tau) = B_\lambda(294\,\mathrm{K})e^{-1} + B_\lambda(195\,\mathrm{K})(1 - e^{-1}) = 1.2\,\mathrm{W/m^{-2}\,str\,\mu m} \qquad (48)$$

which corresponds to a brightness temperature of $T_b = 250\,\mathrm{K}$. This effect is clear in infrared satellite imagery: Thick cirrus clouds appear much colder (i.e., have lower brightness temperatures) than thin ones at the same level.

It seems on the face of things that computing radiative transfer in the infrared is not that hard, since we can now predict the intensity and flux using (46) and (47) if we know the boundary conditions and the state of the atmosphere. Life, alas, is not that simple. Any practical use of radiative transfer involves integration over some spectral interval, and spectral integration in the infrared is where things become difficult. We learned in the section on spectroscopy that the absorption and emission characteristics of gases change very rapidly with wavelength, being large near absorption lines and small elsewhere. The atmosphere is composed of many gases, so the absorption structure as a function of wavelength is extremely rich. Brute force spectral integration, while theoretically possible, is computationally prohibitive in practice. We will address more practical methods for spectral integration, including band models and k distributions, in the next chapter.

5 FULL RADIATIVE TRANSFER EQUATION, INCLUDING ABSORPTION, EMISSION, AND SCATTERING

What is Scattering?

The absorption of radiation by a gas molecule is a two-step process. First, the photon must pass close enough to the molecule for an interaction to occur, and second, the photon's energy must match the difference between the molecule's current state and another allowed state. But what happens to those photons that interact with molecules but whose energies do not match an allowed transition? These photons essentially bounce off the molecule and are redirected; we call this process *scattering*.

Gases do scatter radiation, but the vast majority of scattering in the atmosphere occurs when light interacts with condensed materials, primarily clouds and aerosols. Formally, we say that photons that are absorbed undergo inelastic interactions with the medium, while elastic collisions cause scattering. The likelihood of each kind of interaction need not be the same; that is, the extinction coefficients for scattering and absorption are not identical.

The radiative transfer equation as we have been writing it describes intensity, a quantity associated with a particular direction of propagation. When we discuss scattering, the direction of the beam becomes more important than when considering only absorption and emission because photons can be scattered both out of the beam into other directions and into the beam from radiation traveling in any other direction.

Accounting for Scattering

When we first wrote the radiative transfer equation, we assumed that the medium absorbed but did not emit radiation at the wavelength in question, as occurs during the absorption of solar radiation in Earth's atmosphere. In going from (34) to (42) we included the effects of emission, as when considering the transfer of infrared radiation in the atmosphere. The blackbody emission contribution in (42) is called a source term. Scattering from the beam into other directions is an additional reduction in intensity, while scattering into the beam from other directions adds a second source term.

To write the complete radiative transfer equation, we must distinguish the amount of absorption and emission along the path from the amount of scattering. We do so by introducing a mass scattering coefficient $k_{s\lambda}$ with dimensions of area per mass, as an analog to the mass absorption coefficient $k_{a\lambda}$. Both absorption and scattering diminish the beam, while scattering of radiation traveling in any other direction into the beam can add to the intensity. The full radiative transfer equation is therefore

$$\begin{aligned}\frac{dI_\lambda(\mu,\varphi)}{ds} &= -(k_{s\lambda}+k_{a\lambda})\rho I_\lambda(\mu,\varphi)+k_{a\lambda}\rho B_\lambda + \\ &\frac{k_{s\lambda}\rho}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu',\varphi'\to\mu,\varphi)I(\mu',\varphi')\,d\mu'\,d\varphi'\end{aligned} \tag{49}$$

where we have made explicit the direction of the beam (specified by μ, φ). The last term in (49) accounts for the scattering of radiation into the beam traveling in direction μ,φ from every other direction. The quantity $P(\mu',\varphi'\to\mu,\varphi)$ is called the *single scattering phase function*, or often simply the phase function, and describes how likely it is that radiation traveling in the μ',φ' direction will be scattered into the μ,φ direction. The phase function is reciprocal, so that

$P(\mu',\varphi' \rightarrow \mu,\varphi) = P(\mu,\varphi \rightarrow \mu',\varphi')$, and is defined such that the integral over the entire sphere is 4π.

We divide both sides of the equation by $(k_{s\lambda} + k_{a\lambda})\rho$ and relate path length differential to the vertical displacement to obtain

$$\mu \frac{dI_\lambda(\mu,\varphi)}{d\tau} = I_\lambda(\mu,\varphi) - (1-\omega_0)B_\lambda - \frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu',\varphi' \rightarrow \mu,\varphi) I(\mu',\varphi')\, d\mu'\, d\varphi' \tag{50}$$

where we have now defined the *single scattering albedo* $\omega_0 = k_{s\lambda}/(k_{s\lambda} + k_{a\lambda})$, which is the likelihood that a photon is scattered rather than absorbed at each interaction. Single scattering albedo varies between zero and one; the lower limit corresponds to complete absorption and the upper to complete scattering.

Equation (50) is the plane parallel, unpolarized, monochromatic radiative transfer equation in full detail. Despite its length, it describes only four processes: extinction by absorption and by scattering out of the beam into other directions, emission into the beam, and scattering into the beam from every other direction. The equation is, unfortunately, quite difficult to solve because it is an integrodifferential equation for intensity; that is, intensity appears both in the differential on the left-hand side and as part of the integral on the right-hand side of the equation.

Before we can even begin to solve this equation, we have to come to grips with the way particles scatter light. When we consider absorption and emission, we need only to determine the mass absorption coefficient $k_{a\lambda}$. When we include scattering in the radiative transfer equation, though, we require three additional pieces of information: the mass scattering coefficient $k_{s\lambda}$, along with the phase function $P(\mu',\varphi' \rightarrow \mu,\varphi)$ and single scattering albedo ω_0.

6 SINGLE SCATTERING

When we are concerned with emission and absorption, it is spectroscopy, a combination of quantum mechanics and statistical mechanics, that lets us determine $k_{a\lambda}$ from knowledge of the temperature, pressure, and chemical composition of the atmosphere. To compute the single scattering parameters within a volume, we begin with knowledge of the way light interacts with individual particles, which comes from solutions to Maxwell's equations; this knowledge is then combined with information about the statistics of different particle types within the volume.

Scattering from a particle is most naturally computed in a frame of reference centered on the particle. In particular, it is easiest to describe the phase function in terms of a *scattering angle* Θ between the incident and scattered radiation. The phase function is often summarized using the *asymmetry parameter* g:

$$g = \frac{1}{2}\int_{-1}^{1} \cos\Theta P(\Theta) d\ \cos\Theta \tag{51}$$

which is the average cosine of the scattering angle. When $g > 0$, more light is scattered into the forward than backward direction. $P(\mu', \varphi' \rightarrow \mu, \varphi)$ is, of course, related to $P(\Theta)$:

$$\cos\Theta = \mu'\mu + \sqrt{(1-\mu'^2)}\sqrt{(1-\mu^2)}\cos(\varphi' - \varphi) \tag{52}$$

Computing Scattering from a Single Particle

The task of computing the intensity scattered from a single particle is conceptually straightforward: Maxwell's equations are solved inside and outside the particle subject to the boundary conditions on the particle's surface. In practice this is such a difficult feat that it is possible only in circumstances when geometric or size considerations are favorable, or when it is possible to make simplifying assumptions of one kind or another.

In the electromagnetic terms of Maxwell's equations, cloud drops and aerosol drops differ from the gaseous atmosphere only in their index of refraction $m = n_r + in_i$. The index of refraction is complex: The real part primarily determines the speed of light within the medium, while the imaginary part determines the amount of absorption per amount of material. The value of m varies with wavelength and can also depend on the state of the material; the indices of refraction of water and ice, for example, can differ dramatically at certain wavelengths, as Figure 7 shows.

Imagine a particle with monochromatic radiation I_{inc} incident upon it. The total surface area projected in the direction of the beam's origin is called the particle's geometric cross section C_{geo}, and the power incident on the particle is $P_{\text{inc}} = C_{\text{geo}} I_{\text{inc}}$. We can generalize this idea to define scattering and absorption cross sections for the particle though the rate at which each process removes energy from the beam: $C_{\text{sca}} = P_{\text{sca}}/I_{\text{inc}}$, $C_{\text{abs}} = P_{\text{abs}}/I_{\text{inc}}$. Extinction includes both scattering and absorption, so the extinction cross section C_{ext} is the sum of C_{sca} and C_{ext}. The radiative cross sections of a particle are related to its geometric cross section but also depend on the particle shape and index of refraction. This makes it useful to unscramble the two influences, defining efficiencies $Q_j = C_j/C_{\text{geo}}$, where j denotes one of scattering, absorption, or extinction.

The single scattering parameters of a particle depend on the particle's size, composition, and shape, and on the wavelength of radiation being scattered through the index of refraction. The relative sizes of the particle and the radiation determine the methods that must be used to compute single scattering parameters. The relationship is embodied in the *size parameter* x, the ratio between some characteristic radius r and the wavelength; for spheres, for example, the size parameter is defined as $x = 2\pi r/\lambda$.

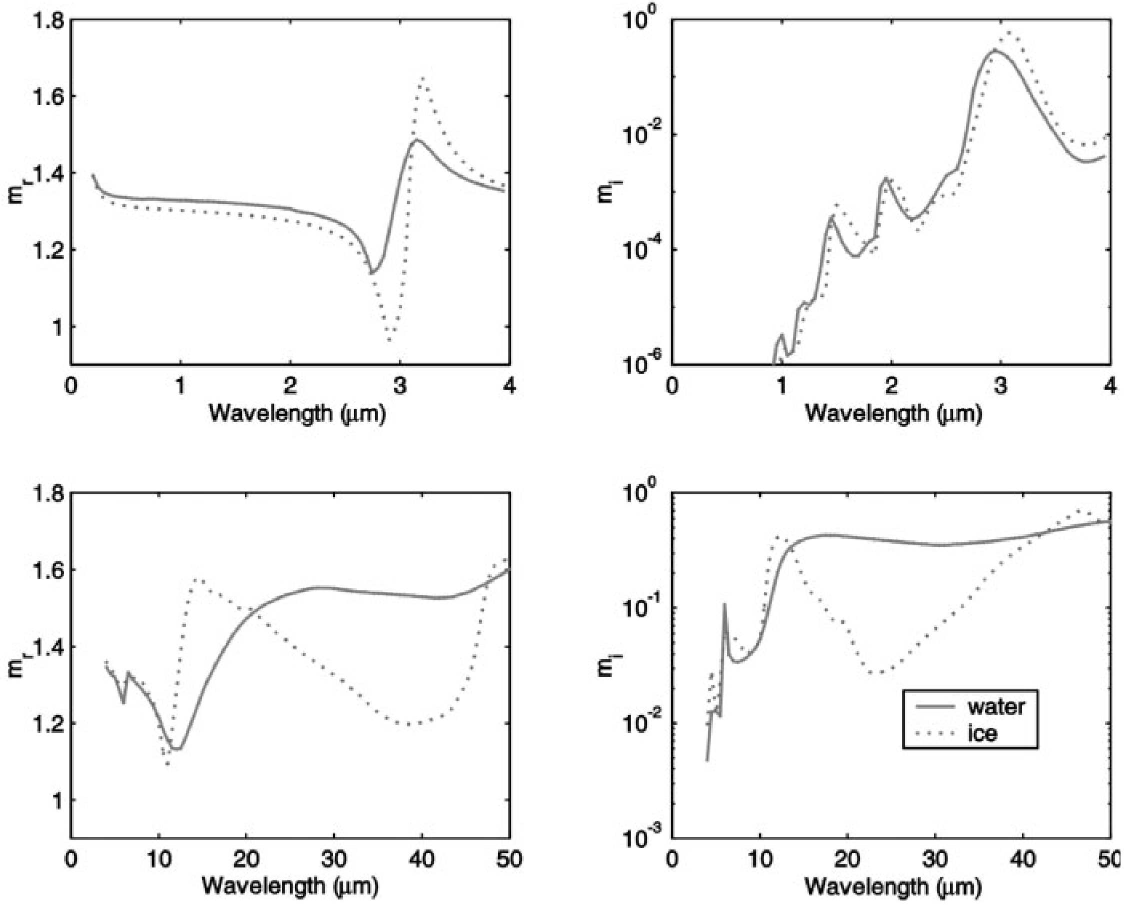


Figure 7 Indices of refraction for liquid water and ice as a function of wavelength. Upper set of panels is a closer look at the left-hand portion of the lower set of panels. See ftp site for color image.

Scattering by Small Particles: Rayleigh Theory

When materials are placed within a constant electric field, the molecular charges can become displaced from one another so that the material is polarized. In a traveling wave the electric field varies in time and space; so, in general, different parts of a particle subject to radiation are polarized differently. But if the particle is very small compared to the wavelength of radiation (i.e., if $x \ll 1$), the entire particle is subject to the same field at any given moment. This allows us to treat the radiation scattered from small particles as if it were emitted from a dipole oscillating at the same frequency as the incident wave.

The theory of scattering from small particles is named after its developer, Lord Rayleigh, who published the work in 1871. Rayleigh used an elegant and succinct dimensional argument to show that the scattering efficiency varies as the size parameter to the fourth power, which for particles of roughly constant size means that scattering depends strongly on wavelength. The full expression for scattering efficiency is

$$Q_{\text{sca}} = \frac{8}{3} x^4 \left| \frac{m^2 - 1}{m^2 + 2} \right|^2 \tag{53}$$

which implies that scattering cross section depends on the size of the particle to the sixth power and (if m does not depend too strongly on wavelength) on λ^{-4}. The theory provides a simple phase function for small particles shown in Figure 9.

Microwave radiation encountering cloud drops is subject to Rayleigh scattering, which is why radar beams reflect so strongly from large drops and precipitation. And Rayleigh scattering of visible light by gas molecules is why skies are blue and sunsets red: Gas molecules are all about the same size, but blue light is scattered from the incoming sunbeam into the open sky much more efficiently than red light.

Scattering by Round Particles: Lorenz–Mie Theory

Some of the most dramatic sights in the atmosphere come from the scattering of visible light by clouds. Cloud drops are typically about 10 μm in size, and ice crystals are an order of magnitude larger. For visible light this yields size parameters much larger than 1, so Rayleigh theory is not applicable. In warm clouds, though, surface tension acts to minimize the particle surface and make the drops round. *Lorenz–Mie theory*, developed around the turn of the twentieth century, takes advantage of this symmetry to develop an exact solution for scattering from homogeneous spheres. The technique computes the radiation field by finding a series solution to the wave equation for the scattered wave in spherical coordinates centered on the particle, then expanding the incident radiation in the same coordinates and matching the boundary conditions.

The application of Lorenz–Mie theory is routine, and codes are freely available in several computer languages. The calculation requires the relative index of refraction of the particle and its size parameter and provides the phase function, extinction efficiency, and single scattering albedo; examples are shown in Figures 8, 9, and 10. Because the number of terms used to expand the incoming wave increases with particle size, calculations require more terms as the size parameter increases.

When x is very small, the extinction efficiency increases rapidly (echoing Rayleigh theory) at a rate that depends on the index of refraction. The efficiency, plotted in Figure 8, then oscillates as interference between the scattered and incident radiation changes from constructive to destructive with small changes in particle size. At very large values of x (not shown) Q_{ext} approaches 2, implying that an area twice as large as the particle's cross section is removed from the beam. This is called the extinction paradox and highlights the role of diffraction in particle scattering. We might think that a large particle casts a shadow exactly as large as its cross section, and that this shadow corresponds to the amount of extinction. But every particle has an edge, and the light passing near this edge is diffracted, or diverted very slightly from its initial direction. Half of the extinction is due to diffraction and half to absorption and scattering into other directions. Diffraction contributes to a large forward peak in scattering phase function of moderately sized, weakly absorbing spherical particles such as cloud drops in the visible; this peak can be five or more orders of magnitude larger than other parts of the phase function, as Figure 9 shows.

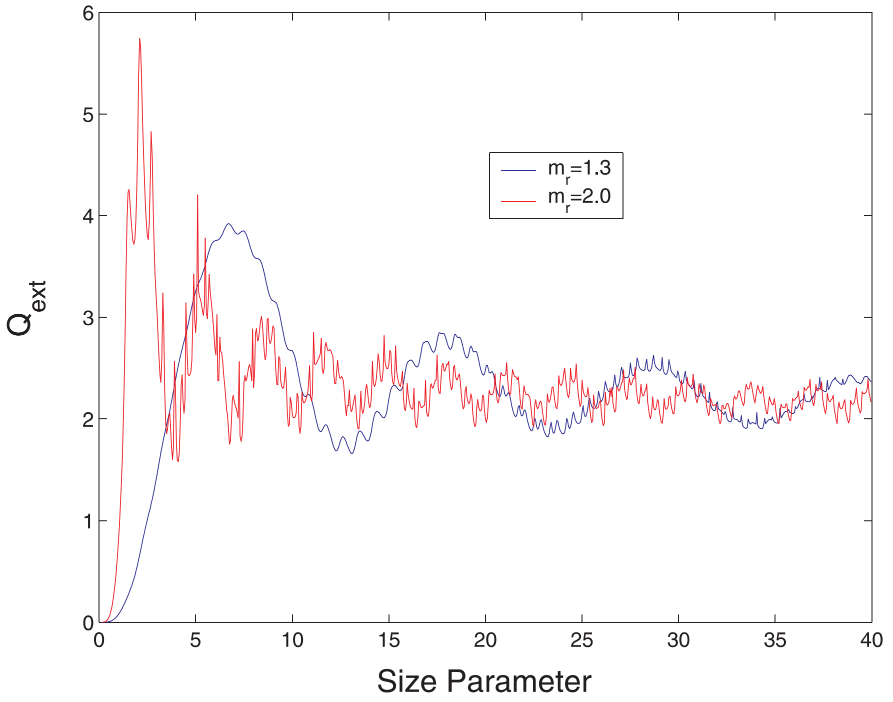


Figure 8 Extinction efficiency, as computed with Mie theory for spherical particles, as a function of size parameter and index of refraction. The value asymptotes to 2 (the "extinction paradox") for large particles as both refraction and diffraction act. See ftp site for color image.

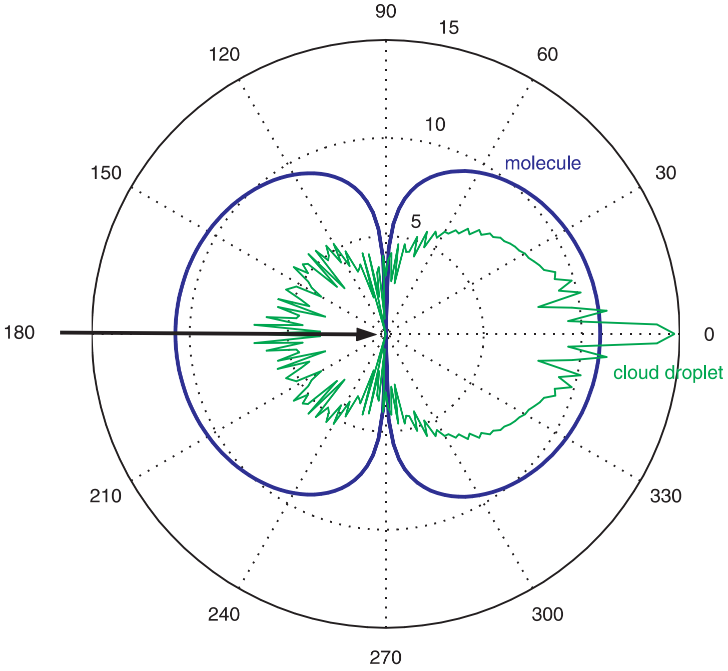


Figure 9 Scattering phase functions for small particles computed using Rayleigh theory (dark line) and a spherical particle of moderate size parameter. The radial axis is on a logarithmic scale. See ftp site for color image.

Regardless of drop size or details of the drop size distribution, in fact, the value of the asymmetry parameter *g* is always quite near to 0.86 in water clouds.

Scattering by Arbitrary Particles

Mie theory and Rayleigh theory are the two most common methods of computing scattering from particles in the atmosphere. Neither is applicable to the scattering of visible light by either ice crystals or mineral aerosols. Other techniques have been developed for irregular particles, but these tend to be more complicated and difficult to use.

Very large particles (size parameters greater than about 50) are said to be in the "geometric optics limit," and their single scattering properties can be computed by ray tracing. The particle is oriented in space and a series of infinitely thin rays are assumed to illuminate it. The direction and intensity is computed using the Fresnel relations for reflection and transmission each time the ray encounters an interface, and the ray is absorbed as it travels through the medium; the total radiation field is the sum of all the reflected and transmitted components for all the rays. Diffraction is computed separately.

Irregularly shaped particles of intermediate size are the biggest challenge. If the phase function is not required, we can use anomalous diffraction theory (ADT) to find the extinction efficiency and single scattering albedo. ADT applies to large particles with an index of refraction near one. It makes the simplifying assumption that extinction is due primarily to interference between waves slowed by their passage through the medium, and those diffracted around the particle edge. Though it ignores internal reflection and refraction, it is analytically tractable and reasonably accurate.

Relatively small particles can be treated using the discrete dipole approximation (DDA), which breaks the particle up in small volumes relative to the wavelength of radiation, treats each volume as a dipole, and computes the interaction among each dipole pair. This gets very computationally expensive as the particle size increases, since the number of dipoles increases as the volume cubed and the number of interactions as the number of dipoles squared; the current range of applicability is to size parameters less than 5 to 10.

An alternative to DDA is the finite-difference time-domain (FDTD) technique. Here the particle is discretized in space; then electric and magnetic fields that vary in time and space are imposed and the solution to Maxwell's equations is integrated forward in time. The incident wave can be monochromatic, but is more commonly a

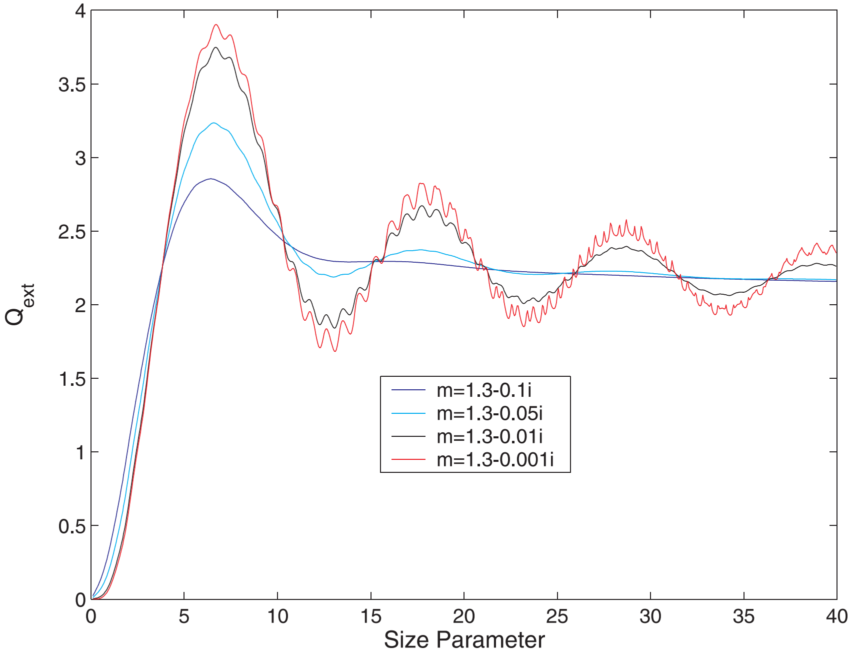


Figure 10 Extinction efficiencies of spherical particles for absorbing particles. In more absorptive materials the variation of Q_{ext} with size parameter is damped. See ftp site for color image.

pulse, since the latter can provide information about many frequencies at once through a Fourier analysis. FDTD can be used for size parameters less than 10 to 20.

Integrating over a Particle Size Distribution

Individual photons are scattered by individual particles, but in treating a beam of radiation we have to consider the many different particles encountered along a differential path $d\tau$, so that the phase function, single scattering albedo, and extinction coefficient represent the entire distribution of drops encountered by the beam. Because cloud drops and aerosols are separated by distances much greater than their characteristic size, we can treat each interaction as independent, and simply add up the contributions from different kinds of particles according to their relative abundance. We have to account for all classes of particles that scatter light uniquely and must weight the effects according to their contributions to the change in intensity.

Let us make this concrete by asking ourselves how light is scattered within warm clouds. The drops within a cloud are all round and have the same index of refraction, but vary in radius r according to the droplet size distribution $n(r)$, which has dimensions of number per volume per radius increment. If we add up all the drops in the distribution, we find the total droplet number concentration

$$\int_0^\infty n(r)\,dr = N \tag{54}$$

where N is measured in number per volume. Drop size distributions within clouds are often represented with the two-parameter gamma distribution:

$$n(r) = \frac{N}{\Gamma(\alpha) r_n}\left(\frac{r}{r_n}\right)^{\alpha-1} \exp\left(-\frac{r}{r_n}\right) \tag{55}$$

in which $\Gamma(\alpha)$ is the Euler gamma function, r_n a characteristic radius, and α is related to the variance of the distribution. The gamma distribution is useful because moments of the drop size distribution can be found analytically.

The scattering properties of the drops depend on the drop radius. Each drop has an extinction cross section $C_{\text{ext}}(r)$, in units of area; if we add up the contributions from each drop size, we obtain the volume extinction coefficient k_{ext} in units of area per volume, or inverse length

$$k_{\text{ext}} = \int_0^\infty C_{\text{ext}}(r) n(r)\,dr = \int_0^\infty \pi r^2 Q_{\text{ext}}(r) n(r)\,dr \tag{56}$$

To compute the average single scattering albedo $\langle\omega_0\rangle$ we divide the average amount of scattering by the average amount of extinction:

$$\langle\omega_0\rangle = \frac{\int_0^\infty C_{\text{sca}}(r) n(r)\,dr}{\int_0^\infty C_{\text{ext}}(r) n(r)\,dr} \tag{57}$$

while the average phase function is weighted according to the amount of light scattered by each drop size:

$$\langle P(\Theta)\rangle = \frac{\int_0^\infty P(\Theta)C_{\text{sca}}(r)n(r)\,dr}{\int_0^\infty C_{\text{sca}}(r)n(r)\,dr} \tag{58}$$

Single scattering parameters are not particularly sensitive to the exact drop size distribution, but are controlled primarily by the total surface area of drops within the cloud. The radius associated with the average surface area of the drops is called the *effective radius* r_e:

$$r_e = \frac{\int_0^\infty r^3 n(r)\,dr}{\int_0^\infty r^2 n(r)\,dr} \tag{59}$$

In most water clouds r_e is of order 10 μm. The phase function, single scattering albedo, and extinction coefficient vary smoothly with effective radius, in part because integration over the size distribution smoothes out rapid oscillations like those in Figure 8.

The drops within water clouds vary only in their radius, while a collection of aerosols or ice crystals might contain particles with differing shapes or indices of refraction. If this is the case the sums (integrals) in (56) though (58) must be extended to account for every combination of shape, size, and material, but the idea remains unchanged.

Relating Cloud Optical and Physical Parameters

Imagine a cloud of thickness z made up of drops following a size distribution $n(r)$. We can compute the optical thickness of this cloud by integrating the extinction coefficient, given by (56), to find

$$\tau = \int_0^z k_{\text{ext}}\,dz = \int_0^z \int_0^\infty \pi r^2 Q_{\text{ext}}(r)n(r)\,dr\,dz \approx \int_0^z \int_0^\infty 2\pi r^2 n(r)\,dr\,dz \tag{60}$$

where we can make the last approximation in the visible, where cloud drops have large size parameters.

How can this be related to the cloud physical properties? The liquid water content (LWC) of this cloud is the sum of the water mass in each drop:

$$\text{LWC} = \int_0^\infty \rho_w \frac{4}{3}\pi r^3 n(r)\,dr \tag{61}$$

and the liquid water path (LWP) is the liquid water content integrated through the depth of the cloud:

$$\text{LWP} = \int_0^z \text{LWC}\,dz = \int_0^z \int_0^\infty \rho_w \frac{4}{3}\pi r^3 n(r)\,dr\,dz \tag{62}$$

We multiply (62) by $\frac{3}{2}\rho_w$ and use the definition of effective radius r_e from (59):

$$\begin{aligned}\frac{3}{2\rho_w}\frac{\text{LWP}}{r_e} &= \int_0^z \int_0^\infty 2\pi r^3 n(r)\,dr\,dz \frac{\int_0^\infty r^2 n(r)\,dr}{\int_0^\infty r^3 n(r)\,dr} \\ &\approx \tau\end{aligned} \tag{63}$$

The last relationship is exact if $n(r)$ is constant with height; the approximation holds to good accuracy for most distributions and is very useful.

7 SIMPLIFYING THE RADIATIVE TRANSFER EQUATION

The full radiative transfer equation is usually applied to solar radiation, for which we know the boundary condition at the top of the atmosphere, namely that the atmosphere is illuminated by the sun from a particular direction denoted μ_0,φ_0:

$$I(\tau = 0, \mu < 0, \varphi) = F_s \delta(\mu_0 - \mu)\delta(\varphi_0 - \varphi) \tag{64}$$

Here F_s is the solar flux at the top of the atmosphere and $\delta(x)$ is the Kroenecker delta function, with is infinite when its argument is zero and zero otherwise, but integrates to one.

If sunlight were not scattered, it would obey Beer's law and intensity would decrease exponentially with optical depth. This implies that energy traveling in any direction other than μ_0, φ_0 has been scattered at least once. We will therefore decompose the total intensity field into direct and diffuse components, depending on whether the photons have or have not been scattered at least once:

$$I(\tau, \mu, \varphi) = I_{\text{dir}}(\tau, \mu, \varphi) + I_{\text{dif}}(\tau, \mu, \varphi) \tag{65}$$

The boundary condition in (64) applies to the direct intensity, which is depleted by any scattering and any absorption, and so follows Beer's law. There is no incoming diffuse intensity at the top of the atmosphere, but the diffuse intensity can be increased by scattering from either itself or from the direct beam. The equation describing diffuse intensity is therefore:

$$\begin{aligned}\mu\frac{dI_{\text{dif}}(\mu,\varphi)}{d\tau} =& I_{\text{dif}}(\mu,\varphi) - (1-\omega_0)B - \frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu',\varphi' \to \mu,\varphi) I_{\text{dif}}(\mu',\varphi')d\mu' d\varphi' \\ & - \frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu',\varphi' \to \mu,\varphi) I_{\text{dir}}(\mu',\varphi')d\mu' d\varphi' \\ =& I_{\text{dif}}(\mu,\varphi) - (1-\omega_0)B - \frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu',\varphi' \to \mu,\varphi) I_{\text{dif}}(\mu',\varphi')d\mu' d\varphi' \\ & - \frac{\omega_0}{4\pi} P(\mu_0,\varphi_0 \to \mu,\varphi) F_s \exp\left(-\frac{\tau}{\mu_0}\right)\end{aligned} \tag{66}$$

The last term in (66) is called the single scattering source term, and the penultimate contribution is called the multiple scattering source term, which is the redistribution of diffuse intensity from one direction into another.

In many applications the full details of the intensity field are more information than we need. One of the easiest ways we can simplify the radiative transfer equation is to average it over azimuth φ. We will begin by defining the azimuthally averaged phase function P_0:

$$P_0(\mu' \to \mu) = \frac{1}{2\pi}\int_0^{2\pi} P(\mu', \varphi' \to \mu, \varphi)\, d\varphi \tag{67}$$

and azimuthally average intensity I_0:

$$I_0(\mu) = \frac{1}{2\pi}\int_0^{2\pi} I(\mu, \varphi)\, d\varphi \tag{68}$$

We then average (66) by integrating both sides over 2π radians in φ and dividing the entire equation by 2π. Since the blackbody source term is isotropic, this results in

$$\begin{aligned}\mu\frac{dI_0}{d\tau} = I_0 - (1-\omega_0)B - \frac{1}{2\pi}\int_0^{2\pi}\frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu', \varphi' \to \mu, \varphi) I(\mu', \varphi')\, d\mu'\, d\varphi'\, d\varphi \\ - \frac{\omega_0}{4\pi} P(\mu_0, \varphi_0 \to \mu, \varphi) F_s \exp\left(-\frac{\tau}{\mu_0}\right)\end{aligned} \tag{69}$$

We can simplify the second to last term by recalling that the phase function is reciprocal:

$$\begin{aligned}&\frac{1}{2\pi}\int_0^{2\pi}\frac{\omega_0}{4\pi}\int_0^{2\pi}\int_{-1}^{1} P(\mu', \varphi' \to \mu, \varphi) I(\mu', \varphi')\, d\mu'\, d\varphi'\, d\varphi \\ &\quad = \frac{\omega_0}{2}\frac{1}{2\pi}\int_0^{2\pi}\int_{-1}^{1} P_0(\mu' \to \mu) I(\mu', \varphi')\, d\mu'\, d\varphi' \\ &\quad = \frac{\omega_0}{2}\int_{-1}^{1} P_0(\mu' \to \mu) I_0(\mu')\, d\mu'\end{aligned} \tag{70}$$

which yields the azimuthally averaged radiative transfer equation, from which we have dropped the subscript:

$$\mu\frac{dI}{d\tau} = I - (1-\omega_0)B - \frac{\omega_0}{2}\int_{-1}^{1} P(\mu' \to \mu) I(\mu')\, d\mu' - \frac{\omega_0}{4\pi} P(\mu_0 \to \mu) F_s \exp\left(\frac{-\tau}{\mu_0}\right) \tag{71}$$

In the remainder of this chapter we will focus on solving (71) rather than any of the more involved and complete versions of the radiative transfer equation. The choice is a simplification, but it is a very useful one: Most methods for solving the azimuthally dependent version of the radiative transfer equation use a Fourier expansion in azimuth, and (71) is the lowest order moment in such a treatment.

Delta Scaling

When discussing single scattering, we found that diffraction from particle edges has as big an impact on extinction as refraction within the particle itself, but that this light is diverted only slightly from its initial direction. Phase functions as strongly peaked as those in Figure 9 cause numerical headaches in solving the radiative transfer equation. And in fact, the width of the diffraction peak for cloud particles is so narrow that the light may as well not be scattered at all.

In most problems in solar radiative transfer we replace the original phase function, which contains a large, narrow forward peak, with two components: A delta function in the forward direction and a smoother, scaled phase function. We define f as the fraction of light scattered directly forward; then we approximate

$$P(\cos\Theta) \approx 2f\partial(1 - \cos\Theta) + (1 - f)P'(\cos\Theta) \tag{72}$$

We choose the asymmetry parameter of the scaled phase function so that the asymmetry parameter of the original phase function is unchanged:

$$g = \frac{1}{2}\int_{-1}^{1} P(\cos\Theta)\, d\,\cos\Theta = f + (1 - f)g' \tag{73}$$

The value of f can be chosen in a variety of ways, depending on how much information is available. One approach is to define $f = g$ so that $g' = 0$; the more common delta-Eddington approximation sets $f = g^2$.

We could apply this scaling to any form of the radiative transfer equation that includes scattering. For the purposes of illustration we will substitute (72) into the azimuthally averaged radiative transfer equation (71), omitting the source terms for clarity:

$$\mu\frac{dI}{d\tau} = I - f\omega_0 I - \frac{(1 - f)\omega_0}{2}\int_{-1}^{1} P'(\mu' \to \mu)I(\mu')\, d\mu' \tag{74}$$

Dividing both sides of this equation by $(1 - \omega_0 f)$ yields

$$\mu\frac{dI}{(1 - \omega_0 f)\, d\tau} = I - \frac{1 - f}{1 - \omega_0 f}\frac{\omega_0}{2}\int_{-1}^{1} P'(\mu' \to \mu)I(\mu')\, d\mu' \tag{75}$$

This is exactly the same form as the original equation if we scale variables:

$$\tau' = (1 - \omega_0 f)\tau \qquad \omega_0' = \frac{(1 - f)\omega_0}{1 - \omega_0 f} \qquad g' = \frac{g - f}{1 - f} \tag{76}$$

so that any technique we have for solving the original radiative transfer equation works on the scaled version as well. In the scaled system both optical thickness and the asymmetry parameter are reduced, so less forward scattering combines with less extinction to produce the same reflection, transmission, and absorption as the unscaled system. The systems agree, of course, when the direct and diffuse beams are added together. In practice, almost all calculations of radiative transfer in the solar system are made using scaled versions of the radiative transfer equation.

8 SOLVING THE RADIATIVE TRANSFER EQUATION SIMPLY

How can we approach the solution of even the azimuthally averaged radiative transfer equation? Intensity may vary with both optical depth and polar angle μ, and ignoring the vertical variation would mean we could not compute even such simple quantities as radiative heating rates. We can, however, try to find the intensity at only a few angles. In fact, computing the intensity field at just two angles, one each in the upward and downward hemispheres, is a lot like computing upward and downward fluxes defined in (5).

Two-stream methods are those that describe the radiation field with just two numbers. They have the advantage of being analytically soluble, which makes them very fast and thus suitable for, say, use in a numerical climate model. They are generally good at computing fluxes and therefore useful in heating rate calculations, but they cannot be used when the angular details of the intensity field are important. There are more than a few two-stream methods, and in every one simplifications need to be made about both the intensity and the phase function.

The following examples illustrate the computation of fluxes in the visible part of the spectrum, where there is no emission, but two-stream models are also applicable to calculations in absorbing and emitting atmosphere.

Eddington's Solution

In the Eddington approximation we expand both intensity and phase function to first order in polar angle. That is, we assume that each varies linearly with μ:

$$I(\mu) = I_0 + I_1\mu, \qquad P(\mu \to \mu') = 1 + 3g\mu\mu' \tag{77}$$

This means we can compute the upward and downward fluxes analytically:

$$F^+ = \pi(I_0 - 2I_1/3) \qquad F^- = \pi(I_0 + 2I_1/3) \tag{78}$$

To find the intensity we substitute (77) into (71):

$$\begin{aligned}\mu\frac{d(I_0 + I_1\mu)}{d\tau} = I_0 + I_1\mu - \frac{\omega_0}{2}\int_{-1}^{1}(1 + 3g\mu\mu')(I_0 + I_1\mu')\,d\mu' \\ - \frac{\omega_0}{4\pi}(1 - 3g\mu\mu_0)F_s\exp\left(\frac{-\tau}{\mu_0}\right)\end{aligned} \tag{79}$$

Now we can evaluate the scattering integral

$$\mu\frac{d(I_0+I_1\mu)}{d\tau}=I_0+I_1\mu-\omega_0(I_0+I_1g\mu)-\frac{\omega_0}{4\pi}(1-3g\mu\mu_0)F_s\exp\left(\frac{-\tau}{\mu_0}\right) \tag{80}$$

and rearrange terms

$$\mu\frac{dI_0}{d\tau}+\mu^2\frac{dI_1}{d\tau}=I_0(1-\omega_0)+I_1\mu(1-\omega_0 g)-\frac{\omega_0}{4\pi}(1-3g\mu\mu_0)F_s\exp\left(\frac{-\tau}{\mu_0}\right) \tag{81}$$

We can break (81) into two pieces by observing that it contains both odd and even powers of μ. What we will do, then, is integrate the equation over μ from -1 to 1, which will leave only terms in even powers of μ. We then multiply (81) by μ and repeat the integration. The two resulting equations are

$$\frac{dI_1}{d\tau}=3(1-\omega_0)I_0-\frac{3\omega_0}{4\pi}F_s\exp\left(\frac{-\tau}{\mu_0}\right) \tag{82}$$

$$\frac{dI_0}{d\tau}=(1-\omega_0 g)I_1+\frac{3\omega_0}{4\pi}g\mu_0F_s\exp\left(\frac{-\tau}{\mu_0}\right) \tag{83}$$

Equations (82) and (83) are a set of two first-order coupled linear differential equations. We can uncouple the equations and find a solution by differentiating one equation with respect to optical depth and substituting the other. This yields, for example

$$\frac{d^2I_0}{d\tau^2}=k^2I_0-\frac{3\omega_0}{4\pi}F_s\exp\left(\frac{-\tau}{\mu_0}\right)(1+g-\omega_0 g) \tag{84}$$

where we have defined the eigenvalue $k^2=3(1-\omega_0)\,(1-\omega_0 g)$. The solutions to (84) and the analogous equation for I_1 are a sum of exponentials in $k\tau$, i.e.,

$$I_0=Ae^{k\tau}+Be^{-k\tau}+\psi e^{-\tau/\mu_0} \tag{85}$$

where A, B, and ψ are determined from the boundary conditions at the top and bottom of the medium and from the particular solution.

The general solution is very complicated but is tractable in some limits. The simplest case is a single homogeneous layer of total optical depth τ^* over a nonreflective surface. If the layer is optically thin ($\tau^*\ll 1$), the reflected and transmitted fluxes are

$$R=\omega_0\left(\frac{1}{2}-3g\mu_0/4\right)\frac{\tau^*}{\mu_0}\qquad T=1-R-\left(\frac{\tau^*}{\mu_0}\right)(1-\omega_0) \tag{86}$$

If the layer does not absorb (i.e., if $\omega_0 = 0$), the reflected flux is

$$R = \frac{(1-g)\tau^* + \left(\frac{2}{3} - \mu_0\right)\left(1 - e^{-\tau^*/\mu_0}\right)}{\frac{4}{3} + (1-g)\tau^*} \tag{87}$$

Reflectance increases with optical thickness, rapidly at first and then more slowly as shown in Figure 11. Reflectance at a given optical depth also increases as μ_0 decreases (this is not shown), since the radiation must change direction less drastically to be reflected into the upward hemisphere. Decreases in the asymmetry parameter increase the amount of reflection, since photons are more likely to be scattered backwards, and since most photons, especially near cloud top, are headed downward. Even relatively optically thick clouds transmit at least 20% of the flux incident on them, which is why it is not pitch dark at the surface even on very cloudy days.

Computing Flux: Two-Stream Model

Eddington's approach to the radiative transfer equation is to expand both the intensity and the phase function to first order in angle; the flux can be computed once the approximate intensity field is known. In the two-stream model we first average the radiative transfer equation and the phase function to get a differential equation for fluxes; then we compute the solution. The solutions are quite similar, as we expect them to be, and the choice of exactly which method to use is a little arbitrary.

Let us begin with the azimuthally averaged radiative transfer equation (71) and integrate it over each hemisphere to find the flux. Ignoring emission, we have for the downward flux

$$\begin{aligned}\int_{-1}^{0} \mu \frac{dI}{d\tau}\, d\mu = \int_{-1}^{0} I\, d\mu - \frac{\omega_0}{2}\int_{-1}^{0}\int_{-1}^{1} P(\mu' \to \mu) I(\mu')\, d\mu'\, d\mu \\ - \frac{\omega_0}{4\pi}\int_{0}^{1} P(\mu_0 \to \mu) F_s \exp(-\tau/\mu_0)\, d\mu\end{aligned} \tag{88}$$

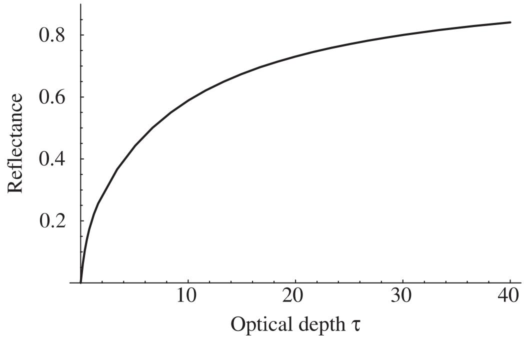


Figure 11 Reflectance of a nonabsorbing layer as computed from the Eddington approximation. The solar zenith angle is 60°.

or

$$
\begin{aligned}
\bar{\mu}\frac{dF^{\downarrow}}{d\tau} = F^{\downarrow} &- \frac{\omega_0}{2}\int_{-1}^{0}\int_{-1}^{0} P(\mu' \to \mu) I^{\downarrow}(\mu')\, d\mu'\, d\mu \\
&- \frac{\omega_0}{2}\int_{-1}^{0}\int_{0}^{1} P(\mu' \to \mu) I^{\uparrow}(\mu')\, d\mu'\, d\mu \\
&- \frac{\omega_0}{4\pi}\int_{0}^{1} P(\mu_0 \to \mu) F_s \exp\left(\frac{-\tau}{\mu_0}\right) d\mu
\end{aligned} \tag{89}
$$

Implicit in going from (88) to (89) is a choice about the ratio:

$$
\bar{\mu} = \frac{\int_{-1}^{0} \mu \frac{dI}{d\tau}\, d\mu}{\int_{-1}^{0} \frac{dI}{d\tau}\, d\mu} \tag{90}
$$

Different two-stream approximations differ in this choice. We then make use of the reciprocity of the phase function and define the *backscattering coefficients*:

$$
b(\mu) = \frac{1}{2}\int_{-1}^{0} P(\mu' \to \mu)\, d\mu' \qquad b = \frac{1}{2}\int_{0}^{1} b(\mu)\, d\mu \tag{91}
$$

which define the fraction of flux scattered into the opposite hemisphere. Applying these definitions, and doing the analogous calculation for the upward flux, we arrive at the two-stream equations:

$$
\bar{\mu}\frac{dF^{\downarrow}}{d\tau} = F^{\downarrow} - \omega_0(1-b)F^{\downarrow} + \omega_0 b F^{\uparrow} - \frac{\omega_0}{2\pi}[1 - b(\mu_0)]F_s \exp\left(\frac{-\tau}{\mu_0}\right) \tag{92}
$$

$$
\bar{\mu}\frac{dF^{\uparrow}}{d\tau} = F^{\uparrow} - \omega_0(1-b)F^{\uparrow} + \omega_0 b F^{\downarrow} - \frac{\omega_0}{2\pi} b(\mu_0) F_s \exp\left(\frac{-\tau}{\mu_0}\right) \tag{93}
$$

These are two first-order, linear, coupled differential equations with constant coefficients. They can be uncoupled by differentiating one with respect to optical thickness, then substituting the other. As with the Eddington approximation, the complete solution for reflected or transmitted flux is the sum of exponentials in optical thickness, with coefficients determined by the boundary conditions.

9 SOLVING RADIATIVE TRANSFER EQUATION COMPLETELY

Analytic approximations like the Eddington and two-stream methods are useful and often accurate enough for flux calculations. If intensity is required, though, we must find numerical ways to solve the radiative transfer equation. In numerical weather prediction, the continuous Navier–Stokes equations of motion are approximated by finite differences between points on a spatial grid. The integrals in the radiative transfer equation are over direction, so the discretization is over angle, turning the continuous equation into one at discrete ordinates or directions. There are two popular approaches to solving this equation, which of course give the same numerical results.

Adding–Doubling Method

Imagine two layers, one of which overlies the other. The upper layer has flux transmittance and reflectance T_1 and R_1, and the lower layer T_2 and R_2. How much total flux R_T is reflected from the combination of layers? Some flux is reflected from the first layer (R_1); some of the flux transmitted through the first layer is reflected from the second layer and transmitted through the first layer ($T_1R_2T_1$); some of this flux reflected from the second layer is reflected back downwards, where some portion is reflected back up ($T_1R_2R_1R_2T_1$), and so on. Because reflection and transmission are both less than one, we can use the summation formula for geometric series to compute the total reflection:

$$\begin{aligned} R_T &= R_1 + T_1R_2T_1 + T_1R_2R_1R_2T_1 + T_1R_2R_1R_2R_1R_2T_1 + \cdots \\ &= R_1 + T_1R_2[1 + R_1R_2 + R_1R_2R_1R_2 + \cdots]T_1 \\ &= R_1 + T_1R_2[1 + R_1R_2]^{-1}T_1 \end{aligned} \tag{94}$$

where the term in square brackets accounts for the multiple reflections between the layers. There is a similar relation for transmission.

Equation (94) and its analog for transmission may be combined with the transmission and reflection results from the Eddington or two-stream solutions to compute the transmittance and reflectance of layered atmospheres, in which the single scattering albedo or asymmetry parameter changes with optical thickness, or to account for a reflecting surface beneath the atmosphere.

The adding–doubling method extends this idea to intensity by replacing the reflection and transmission terms in (94) with matrices that account for the forward and backward scattering from one polar angle into another. Rather than beginning with analytical results for arbitrary layers, we find R and T for a layer of optical depth $d\tau \ll 1$, assuming that photons are scattered no more than once. The reflection and transmission of a layer of $2d\tau$ can be computed using (94) and the reflection and transmission matrices for the original layer. Repeating this process we can find the reflection and transmission of arbitrarily thick layers. Arbitrarily complicated atmospheres can be built up by superimposing layers with different properties and computing the reflection and transmission matrices for the combination. Reflection from surfaces fits naturally into the adding framework.

Eigenvector Solutions

As an alternative to the adding–doubling method we can approximate the continuous radiative transfer equation with a version that is discrete in polar angle

$$\frac{d}{d\tau}\begin{pmatrix}\mathbf{I}^+\\ \mathbf{I}^-\end{pmatrix}=\begin{pmatrix}-t & -r\\ r & t\end{pmatrix}\begin{pmatrix}\mathbf{I}^+\\ \mathbf{I}^-\end{pmatrix}-\begin{pmatrix}\mathbf{S}^+\\ \mathbf{S}^-\end{pmatrix} \tag{95}$$

where $\mathbf{I}^{\pm}$ are vectors containing the intensity $I(\mu_j)$ at each of N angles (each discrete ordinate). The t and r matrices contain the phase function information:

$$t_{j,j'}=\frac{1}{\mu_j}\left[\frac{\omega_0}{2}P(\mu_j\to\mu_{j'})-1\right]\qquad r_{j,j'}=\frac{1}{\mu_j}\frac{\omega_0}{2}P(\mu_j\to-\mu_{j'}) \tag{96}$$

and the source vector is

$$\mathbf{S}^{\pm}=\frac{\pm 1}{\mu_j}\frac{\omega_0}{4\pi}P(\mu_0\to\pm\mu_j)F_s e^{-\tau/\mu} \tag{97}$$

To solve the matrix version (95) of the radiative transfer equation, we treat the homogeneous equation as an eigenvector problem. That is, we set the source term to zero and substitute solutions of the form $\mathbf{I}^{\pm}=\mathbf{G}^{\pm}e^{-k\tau}$. This results in

$$\begin{pmatrix}-t & -r\\ r & t\end{pmatrix}\begin{pmatrix}\mathbf{G}^+\\ \mathbf{G}^-\end{pmatrix}=k^2\begin{pmatrix}\mathbf{G}^+\\ \mathbf{G}^-\end{pmatrix} \tag{98}$$

The solution for intensity is then a sum of the eigenvectors and the particular solution for the source term, with weights determined from the boundary conditions. The solution can be extended to more than one layer by ensuring that intensities match at the layer boundaries. This technique is usually called "discrete ordinates" in the scientific literature, though many solutions use discrete forms of the radiative transfer equation.

Radiative Transfer in Two and Three Dimensions

In our development of the radiative transfer equation we assumed that the atmosphere varies only in the vertical, which leads to a radiation field that depends only on direction and vertical position. But the real atmosphere varies in the horizontal as well as the vertical. In particular, cloud optical properties can change dramatically over relatively short distances. A full description of a radiation field depends on horizontal position as well as vertical position and direction, and in the three-dimensional radiative transfer equation the derivative becomes a gradient operator to account for transfers in the horizontal as well as the vertical.

Traditionally, radiative transfer in two and three dimensions has been computed using Monte Carlo integration techniques. The domain is divided into discrete volumes, within which the optical properties (extinction coefficient, single scattering albedo, and phase function) are considered constant. The paths of many, many photons are then traced from their source until they leave the domain. In solar radiation calculations, for example, the photons are typically introduced at the top of the domain, then travel a random distance depending on the amount of extinction they encounter along the trajectory. They then undergo a scattering and/or absorption event, and continue in a new direction related to the scattering phase function; this process continues until they exit the domain. Monte Carlo methods are best suited to the computation of domain-averaged quantities, but are very flexible. An alternative is to estimate the radiation field at each grid point, then iterate until the field is self-consistent.

Three-dimensional radiative transfer is much more computationally expensive than one-dimensional calculations, and it is difficult to measure atmospheric properties well enough to provide useful input fields. Three-dimensional effects can be significant, however, especially in remote-sensing applications.

10 FROM THEORY TO APPLICATIONS

In this chapter we have laid out the foundations of radiative transfer. We have discussed the physical mechanisms that underlie absorption and scattering and talked about ways to compute the quantities needed for remote-sensing and

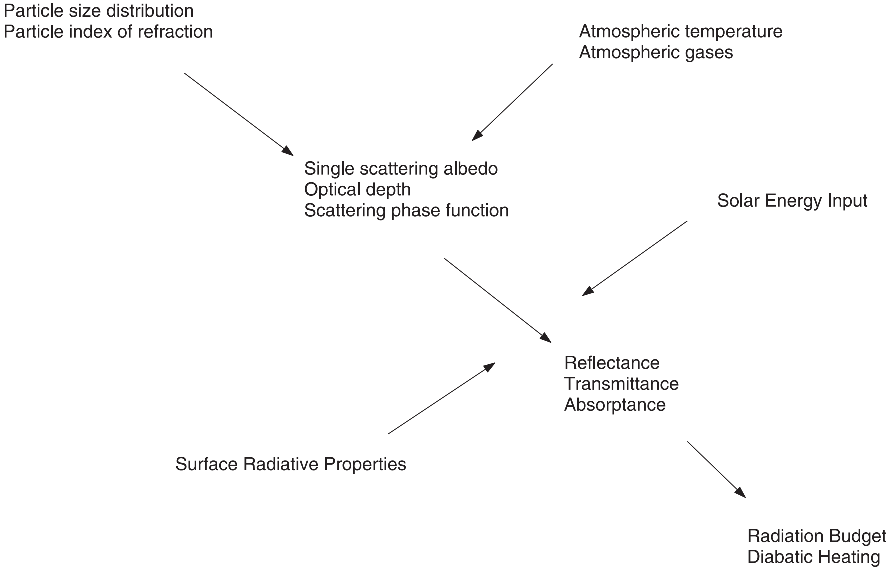


Figure 12 Topics in radiative transfer and remote sensing. This chapter has covered the upper half of the diagram, as well as the computation of reflectance, transmittance, and absorptance; the other topics are discussed in the next chapter.

energy budget applications. The relationship between the various topics, and those to be covered in the following chapter, is shown in Figure 12.

In day-to-day practice much of this work is routine. There are codes available to compute the absorption spectrum of an arbitrary atmosphere, to compute the single scattering parameters of individual particles, and any number of models available for solving the radiative transfer equation, each of which is tailored to a specific application. An understanding of the underlying theory is essential for a critical assessment, but most current research in radiative transfer is in the applications, that we discuss in the next chapter.

CHAPTER 20

RADIATION IN THE ATMOSPHERE: OBSERVATIONS AND APPLICATIONS

STEVEN A. ACKERMANN AND ROBERT PINCUS

1 OVERVIEW

The previous chapter discussed the physical mechanisms that underlie absorption and scattering and presented approaches to compute the quantities needed for remote sensing and energy budget applications. In this chapter we discuss observations of radiative fluxes and apply the equations to remote sensing the atmosphere. The boundary conditions at the top and bottom of the atmosphere are required to compute the transfer of radiation through the atmosphere. Thus, the chapter begins by describing the boundary conditions at the top of atmosphere and at the surface. These conditions are then used to calculate the radiative heating of the atmosphere in the next section. After discussing radiative heating profiles, observations of the energy budget at the top of atmosphere are presented, followed by a discussion of the greenhouse effect.

Satellite observations are routinely used to help answer the question "What is the weather like?" In addition, satellite observations are routinely used to determine the state of Earth's surface as well as cloud properties. Section 6 provides examples of satellite remote sensing the surface and atmospheric conditions.

2 BOUNDARY CONDITIONS

The boundary conditions at the top and bottom of the atmosphere include the properties of the surface and the incoming solar radiation at the top of the

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

atmosphere. To determine the amount of incoming solar radiation, we must consider the sun–Earth geometric relationships.

The Sun and Its Relationship to Earth

Earth receives energy from the sun. This energy drives atmospheric and oceanic circulations. In this section we briefly discuss some of the properties of the sun before presenting methods of computing sun–Earth astronomical relationships.

The sun, which primarily consists of hydrogen and helium, is 4.6 billion years old and is approximately 1.5×10^8 km from Earth. The radius of the sun is approximately 700,000 km with a mass of 2×10^{35} g. The temperature of the sun is about 5×10^6 K at center decreasing to about 5800 K at the surface.

As with the Earth's atmosphere, the sun can be categorized into different layers, including the photosphere, corona, and the chromosphere. The photosphere is the visible region, or the *surface* of the sun with a thickness of approximately 500 km within which the temperature decreases from 8000 to 4000 K. The photosphere is often marked by features called sunspots, which are associated with convection of the sun's hot gases. Sunspots do not occur in the polar regions of the sun. The sunspot minima and maxima occur in a cycle that has a period of approximately 11 years. During sunspot maxima the sun is disturbed with particle outbursts and solar flares. During sunspot minima the sun is quiet or less active. Pairs of sunspots often have opposite magnetic polarities. For a given sunspot cycle, the polarity of the leading spot is always the same for a given hemisphere. With each new sunspot cycle the polarities reverse; thus, the sunspot cycle is often thought of as a 22-year period. Sunspots appear as dark spots on the surface because these areas are at lower temperatures than the surrounding surface; however, satellite measurements show that more radiation is emitted during a sunspot maximum than a sunspot minimum.

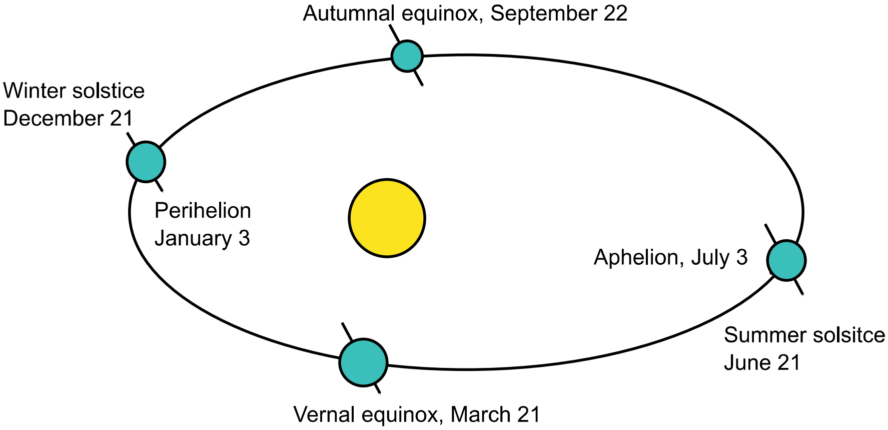


Figure 1 Earth's orbit around the sun. See ftp site for color image.

This may result from increased emission by the faculae, which are bright areas around the sunspots.

The chromosphere extends up to 5000 km above the photosphere. The temperature ranges from 4000 to 6000 K near the photosphere and increases to 10^6 K at 5000 km. The corona, which can be thought of as the sun's atmosphere, extends from the solar disk outward many millions of kilometers. The corona is visible during total solar eclipses. The stream of gas that flows out of the corona and into the solar system is called the solar wind, which provides the energy to produce Earth's aurora borealis and aurora australis.

Sun–Earth Astronomical Relationships

To calculate the solar flux distribution within the atmosphere and ocean, we need to know where the sun is with respect to the geographical region of interest. Such relationships are depicted in Figures 1 through 3 and formulated below.

The mean sun–Earth distance, r_0, is defined to be one astronomical unit (1 AU). The minimum sun–Earth distance (perihelion) is about 0.983 AU and occurs on approximately January 3. The maximum distance (aphelion) is 1.1017 AU and occurs on approximately July 3. The sun–Earth distance, r, is known accurately for every day of the year and is published in the *American Ephemeris and Nautical Almanac*; however, for applications in atmospheric radiative transfer, mathematical approximations are simpler to use. What is required is the reciprocal of the square of

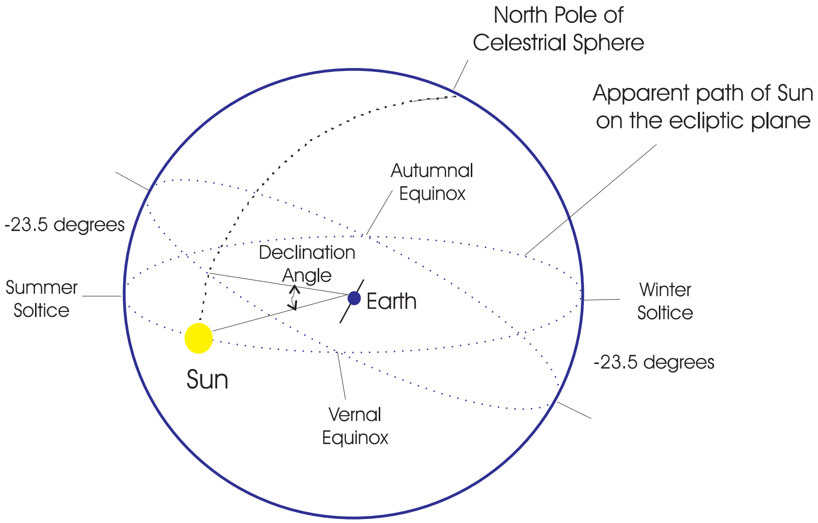


Figure 2 Definition of Sun-Earth astronomical relations. See ftp site for color image.

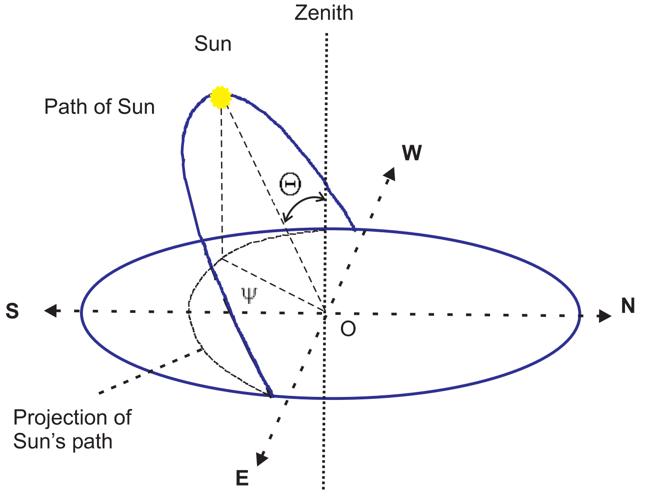


Figure 3 Definition of Earth's azimuth and zenith angles. See ftp site for color image.

the radius vector of Earth, or $(r_0/r)^2$, which is an eccentricity correction factor of Earth's orbit. A common approximation is

$$\left(\frac{r_0}{r}\right)^2 = 1.000110 + 0.034221\cos\Gamma + 0.001280\sin\Gamma + 0.000719\cos 2\Gamma + 0.000077\sin 2\Gamma \qquad (1)$$

where $\Gamma = 2\pi(d_n - 1)/365$ and d_n is the day of the year, which ranges from 1 on January 1 to 365 (366 on a leap year) on December 31.

To determine the solar flux within Earth's atmosphere requires defining some fundamental parameters that describe the position of the sun relative to a location on Earth. The ecliptic plane is the plane of revolution of Earth around the sun. Earth rotates around its polar axis, which is inclined at approximately 23.5° from the normal to the ecliptic plane. The angle between the polar axis and the normal to the ecliptic plane remains unchanged as does the angle between Earth's celestial equator plane and the ecliptic plane (Fig. 1). The angle between the line joining the centers of the sun and Earth to the equatorial plane changes and is known as the solar declination, δ. It is zero at the vernal and autumnal equinoxes and approximately 23.5° at the summer solstice and −23.5° at the winter solstice. A useful approximate equation to express the declination in degrees is

$$\delta = 0.006918 - 0.399912\cos\Gamma + 0.070257\sin\Gamma - 0.006758\cos 2\Gamma + 0.000907\sin 2\Gamma \qquad (2)$$

The solar zenith angle θ_0 is the angle between the local zenith and the line joining the observer and the sun. The solar zenith angle ranges between 0° and 90° and is determined from

$$\cos\theta_0 = \sin\delta\sin\phi + \cos\delta\cos\phi\cos\omega \tag{3}$$

where ϕ is the geographic latitude. The hour angle ω is measured at the celestial pole between the observer's longitude and the solar longitude. The hour angle is zero at noon and positive in the morning, changing 15° per hour.

The solar azimuth ψ is the angle at the local zenith between the plane of the observer's meridian and the plane of a great circle passing through the zenith and the sun. It is zero at the south and measured positive to the east and varies between ±180° and is calculated from

$$\cos\Psi = \frac{(\cos\theta_0\sin\phi - \sin\delta)}{\sin^{-1}[\cos\theta_0]\cos\phi} \tag{4}$$

and $0° \leq \Psi \leq 90°$ when $\cos\Psi \geq 0$ and $90° \leq \Psi \leq 180°$ when $\cos\Psi \leq 0$.

Incoming Solar Radiation at the Top of the Atmosphere

Earth's rotation causes daily changes in the incoming solar radiation while the position of its axis relative to the sun causes the seasonal changes.

The Solar Constant

The flux at the top of Earth's atmosphere on a horizontal surface is

$$F = S_0\left(\frac{r_0}{r}\right)^2\cos\theta_0 \tag{5}$$

where S_0 is the solar constant, the rate of total solar energy at all wavelengths incident on a unit area exposed normally to rays of the sun at 1 AU.

The solar constant can be measured from satellites or derived from the conservation of energy. The power emitted by the sun is approximately 3.9×10^{26} W. The radius of the sun is approximately 7×10^8 m, so the flux at the surface of the sun is $F_{\text{sun}} \approx 6.3 \times 10^7\ \text{W/m}^{-2}$. What is the irradiance reaching Earth (S_0)? Assuming the power from the sun is constant, by conservation of energy the total power crossing a sphere at a radius equivalent to the Earth–sun distance must be equal to the power emitted at the sun's surface, so

$$S_0(4\pi R_{\text{es}}^2) = F_{\text{sun}}(4\pi R_s^2) \tag{6}$$

$$S_0 = F_{\text{sun}}\left[\frac{R_s^2}{R_{\text{es}}^2}\right] \approx 1368\ \text{W/m}^2 \tag{7}$$

The amount of solar energy reaching Earth is not constant, but a function of distance from the sun, and therefore time of year, as indicated by Eq. (5).

The spectral distribution of the solar flux at the top of the atmosphere is shown in Figure 4. While the sun is often considered to have a radiative temperature of approximately 5777 K, Figure 4 demonstrates that the sun is not a perfect blackbody. Also shown in Figure 4 is the percentage of solar energy below a given wavelength (dotted line) confirming that most of the sun's energy resides at wavelengths less than 4 μm.

The daily radiation on a horizontal surface in joules per square meter per day ($\mathrm{J/m^2}$ day) is

$$F_{\text{day}} = \int_{\text{sunrise}}^{\text{sunset}} F\,dt = 2\int_0^{\omega_s} F\,d\omega \frac{24}{2\pi} \tag{8}$$

After converting dt to hour angle,

$$\frac{d\omega}{dt} = \frac{2\pi \text{ radians}}{24 \text{ hour}}$$

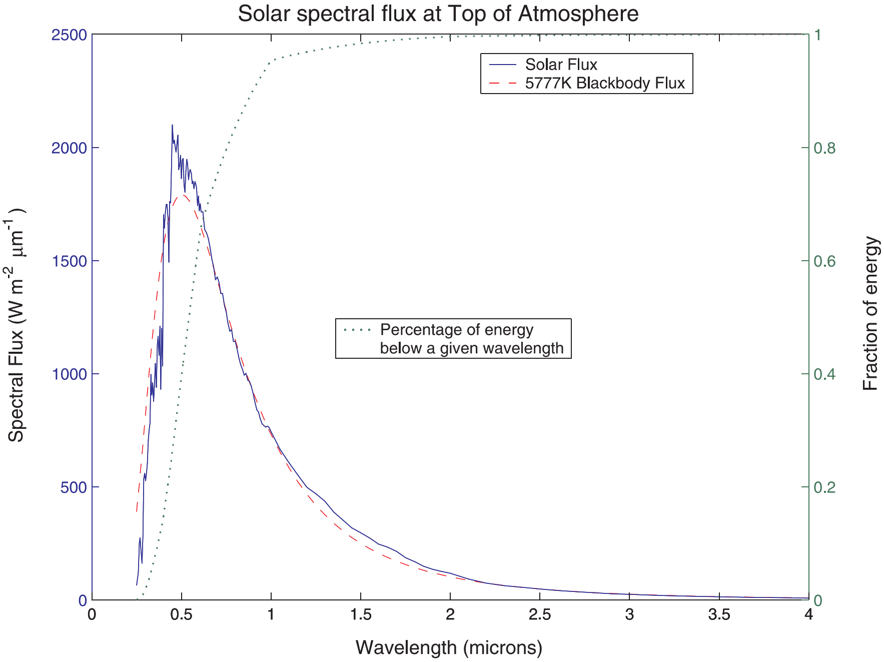


Figure 4 Spectral distribution of solar energy at the top of the atmosphere. Solid line is incoming solar spectral flux, and dashed line is the radiation emitted by a blackbody with a temperature of 5777 K. See ftp site for color image.

we obtain the average daily insolation on a level surface at the top of the atmosphere as

$$F_{\text{day}} = \frac{S_0 (r_0/r)^2 (\omega_s \sin\delta \sin\phi + \cos\delta \cos\phi \sin\omega_s)}{\pi} \tag{9}$$

In polar regions during summer, when the sun is always above the surface, ω_s equals π and the extraterrestrial daily flux is

$$F_{\text{day}} = S_0 \left(\frac{r_0}{r}\right)^2 \sin\delta \sin\phi$$

The daily variation of insolation at the top of the atmosphere as a function of latitude and day of year is depicted in Figure 5. Since Earth is closer to the sun in January, the Southern Hemisphere maximum insolation during summer is about 7% higher than the maximum insolation during Northern Hemisphere summer. The insolation of the polar regions is greater than that near the equator during the summer solstice. This results from the longer days, despite the high solar zenith angles at these high

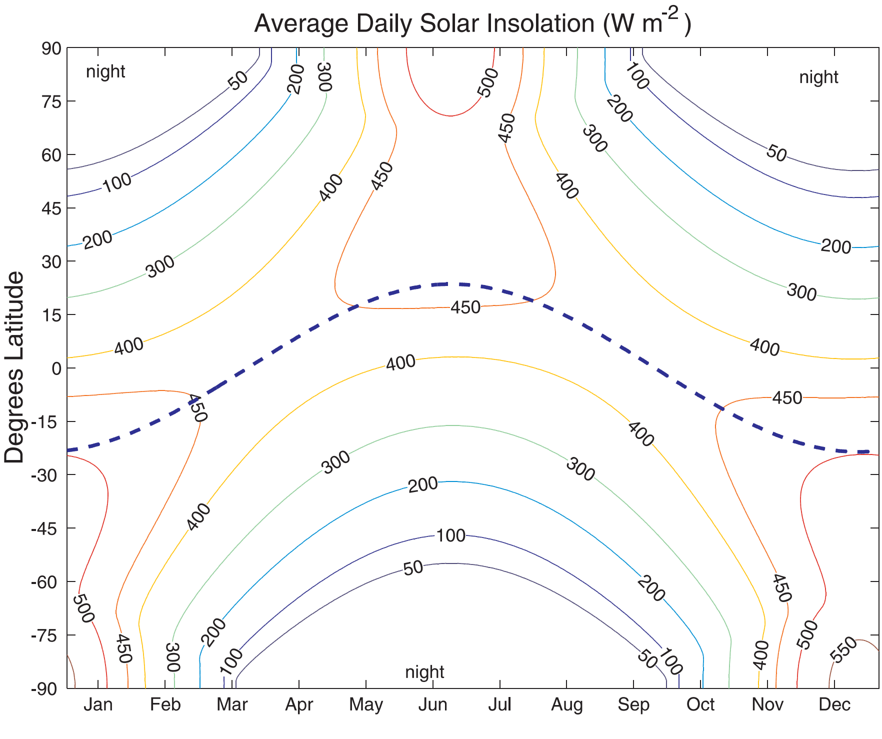


Figure 5 Daily insolation at top of atmosphere as function of latitude and day of year. Also shown is the solar declination angle (thick dashed line). See ftp site for color image.

latitudes. However, the annual average insolation at the top of the atmosphere at the poles is less than half the annual average insolation at the equator.

Surface Radiative Properties

The surface albedo ρ_g is defined as the ratio of the radiation reflected from the surface to the radiation incident on the surface. The surface albedo varies from approximately 5% for calm, deep water to over 90% for fresh snow. The surface albedo over land depends on the type and condition of the vegetation or bare ground. Thus, over land, the surface albedo varies from location to location and with time. Over water, the surface albedo is also a strong function of solar zenith angle.

The surface albedo depends on the wavelength of the incident radiation. Figure 6 is an example of spectral reflection of various surfaces. Snow is very reflective at visible wavelengths (0.4 to 0.7 μm) and less reflective an the near-infrared wavelengths. Plants have higher reflectances in the near-infrared than in the visible. Photosynthesis is effective at absorbing visible energy. When plants dry out, their chlorophyll content decreases and the reflectance at visible wavelength increases.

In the longwave spectral region we generally speak of surface emissivity, or emittance, $\varepsilon_g = 1 - \rho_g$. It is common in models to assume that the surface emittance

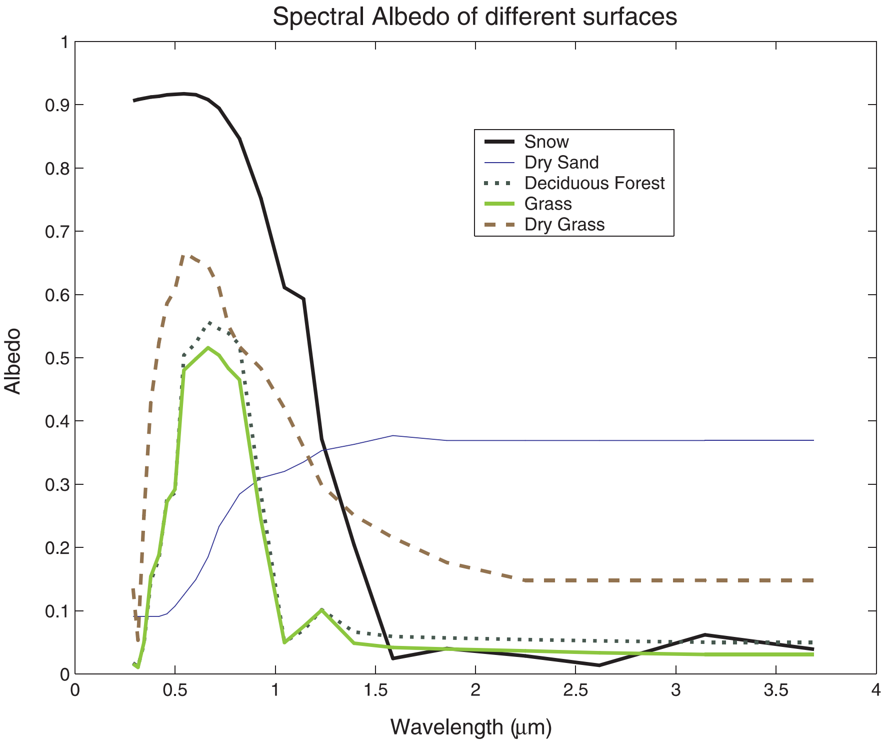


Figure 6 Examples of the spectral albedo of different surfaces. See ftp site for color image.

TABLE 1 Infrared Emissivities of Some Surfaces

Water	0.92–0.96
Ice	0.96
Fresh snow	0.82–0.99
Dry sand	0.89–0.90
Wet sand	0.96
Desert	0.90–0.91
Dry concrete	0.71–0.88
Pine forest	0.90
Grass	0.90

in the infrared (ε_g) is spectrally independent and equal to 1, even though the infrared emissivity of most surfaces is between 0.9 and 0.98 (Table 1). Assuming that the surface emissivity is 1 leads to approximately a 5% error in upward fluxes. The emittance of some surfaces is also wavelength dependent. In particular, sand and desert surface emissivity varies in the 8- to 12-μm window regime between 0.8 and 0.95.

3 RADIATIVE HEATING RATES

This section provides examples of radiative heating profiles for clear and cloud conditions. If more radiative energy enters the layer than leaves the layer, a heating of the layer results. Since the atmosphere is too cold to emit much radiation below 3 μm, atmospheric solar radiative heating rates must all be positive. Since the atmosphere emits and absorbs longwave radiation, longwave radiative heating rates can be positive (net energy gain) or negative (net energy loss).

Radiatively Active Gases in the Atmosphere

The three major absorbing and emitting gases in the stratosphere and troposphere are ozone, carbon dioxide, and water vapor. There are also important minor constituents such as chlorofluorocarbons (CFCs) and methane. While N_2 and O_2 are the most abundant gases in the atmosphere, from an atmospheric energetics point of view they are of small importance. The solar energy below 0.2 μm is absorbed by O, NO, O_2, and N_2 before it reaches the stratosphere.

Ozone (O_3) primarily absorbs in the ultraviolet and in the 9.6-μm region. Solar radiation in the Hartley spectral band (0.2 to 0.3 μm) is absorbed in the upper stratosphere and mesosphere by ozone. Absorption by O_3 in the Huggins band (0.3 to 0.36 μm) is not as strong as in the Hartley bands. Ozone absorbs weakly in the 44- to 0.76-μm region, and strongly around the 9.6-μm region, where radiation is emitted by the surface.

Carbon dioxide is generally a weak absorber in the solar spectrum with very weak absorption in the 2.0-, 1.6-, and 1.4-μm bands. The 2.7-μm band is strong enough that it should be included in calculations of solar absorption, though it overlaps with H_2O. The 4.3-μm band is important more in the infrared region due to the small amount of solar energy in this band. This 4-μm band is important for remote sensing atmospheric temperature profiles. CO_2 absorbs significantly in the 15-μm band from about 12.5 to 16.7 μm (600 to 800 cm^{-1}). It is these differences in the shortwave and infrared properties of CO_2 (and atmospheric water vapor) that lead to the greenhouse effect.

Water vapor absorbs in the vibrational and rotational bands (ground-state transitions). In terms of radiative transfer through the atmosphere, the important water vapor absorption bands in the solar spectrum are centered at 0.94, 1.1, 1.38, 1.87, 2.7, and 3.2 μm. In the infrared, H_2O has a strong vibrational-absorption band at 6.3 μm. The rotational band extends from approximately 13 μm to 1 mm. In the region between these two infrared water vapor bands is the continuum, 8 to 13 μm, known as the atmospheric window. The continuum enhances absorption in the lower regions of the moist tropical atmosphere.

Radiative Heating Rates under Clear Skies

The previous chapter discussed the procedures to calculate radiative fluxes and heating rates in the atmosphere. In this section we discuss examples of radiative heating and cooling under clear-sky conditions.

Atmospheric radiative heating rates due to absorption of solar energy are given in Figure 7 for different solar zenith angles and for tropical conditions. As the solar zenith angle decreases, the total heating of the atmosphere decreases as the solar energy incident on a horizontal surface at the top of atmosphere decreases. Figure 8 demonstrates the absorption by the individual gases if they existed in the atmosphere alone. The large heating in the stratosphere is due to absorption of solar energy by O_3. A minimum in the heating occurs in the upper troposphere. The increased heating in the lower troposphere is due to water vapor. CO_2 contributes little to the solar heating of the atmosphere.

Infrared heating rates are shown in Figure 9 for standard tropical, midlatitude summer, and subarctic summer conditions. Negative values indicate a cooling. The larger cooling rates in the lower troposphere for the tropical conditions arise due to the warmer temperatures and larger amounts of water vapor. The contributions by H_2O, CO_2, and O_3 to the cooling in a tropical atmosphere are shown in Figure 10.

In the tropical moist atmosphere, the water vapor continuum (8- to 13-μm region) makes significant contributions to the cooling to the lowest layers of the atmosphere. CO_2 accounts for the large cooling in the stratosphere. The positive radiative heating rates by O_3 in the stratosphere arise due to the large amounts of radiation in the 9.6-μm band.

Radiative Heating under Cloudy Conditions

Clouds significantly alter the radiative heating and cooling of the atmosphere and at Earth's surface. Clouds also undergo physical changes (e.g., particle size distribution, water content, and cloud top and base altitude) as they form, grow, and dissi-

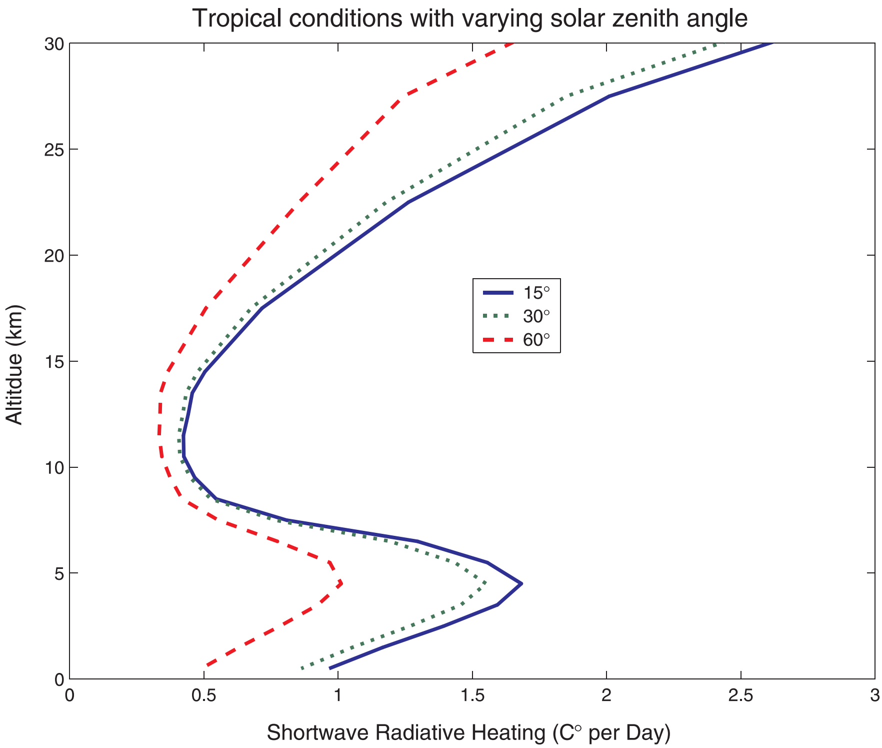


Figure 7 Shortwave radiative heating in degrees celsius per day for standard tropical conditions and three solar zenith angles. See ftp site for color image.

pate. These physical changes are capable of affecting the radiative characteristics of the cloud. Radiative processes have a strong influence on the convective structure and water budget of clouds and affect the particle growth rate.

The previous chapter examined the appropriate equations for calculating the radiative properties of clouds. In this section, we examine the radiative impact of clouds on the atmospheric heating profile. Figure 11 shows the impact of four clouds, each 1 km thick, with differing cloud microphysics on the radiative heating of a midlatitude summer atmosphere with a solar zenith angle of 15°. Three of the clouds have r_{eff} radius of 20 μm with differing ice water content (IWC). Heating rates are expressed in degrees Celsius per day. The larger the IWC the greater the energy convergence, expressed as a heating, within the cloud, as more solar radiation is absorbed. The larger the IWC, the more the heating below cloud base is reduced because of the reduced amount of solar energy below the cloud.

The presence of the cloud affects the heating profiles above the cloud. As the cloud optical depth increases, more energy is reflected upward, enabling absorption in gaseous absorption bands. The particle size also impacts the heating profile. Comparing the two clouds with the same IWC, but different r_{eff} indicates that the cloud with the large particle size absorbs less solar energy.

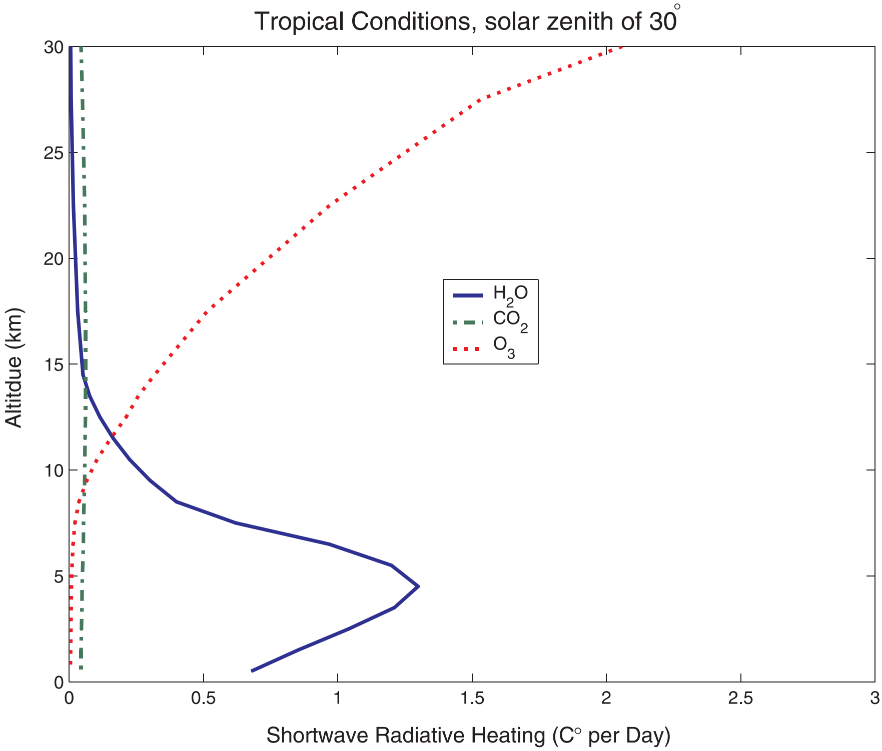


Figure 8 Shortwave radiative heating for standard tropical conditions by H_2O, CO_2, and O_3 and a solar zenith angle of 30°. See ftp site for color image.

Figure 12 depicts the impact of clouds on the longwave heating, for a midlatitude summer atmosphere. The cloud is 2 km thick with a top at 10 km. The top of the cloud has a net energy loss, expressed as a cooling. The cloud base can experience a radiative warming or cooling, depending on the IWC. Increasing the IWC reduces the cooling below cloud base as more energy is emitted by the cloud into this lower layer. Increasing the particle size reduces the cloud radiative heating, if the IWC is fixed.

We will now consider how radiation interacts with a cloudy layer, accounting for multiple scattering within the cloud. The previous chapter discussed how radiation interacts with a particle, or distribution of particles, in terms of the single scattering properties of the particles. There are some useful limits of the two stream model to consider:

$$\lim_{\delta \to 0} R = \frac{\omega_0 \delta}{\mu_0} \gamma_3 \tag{10}$$

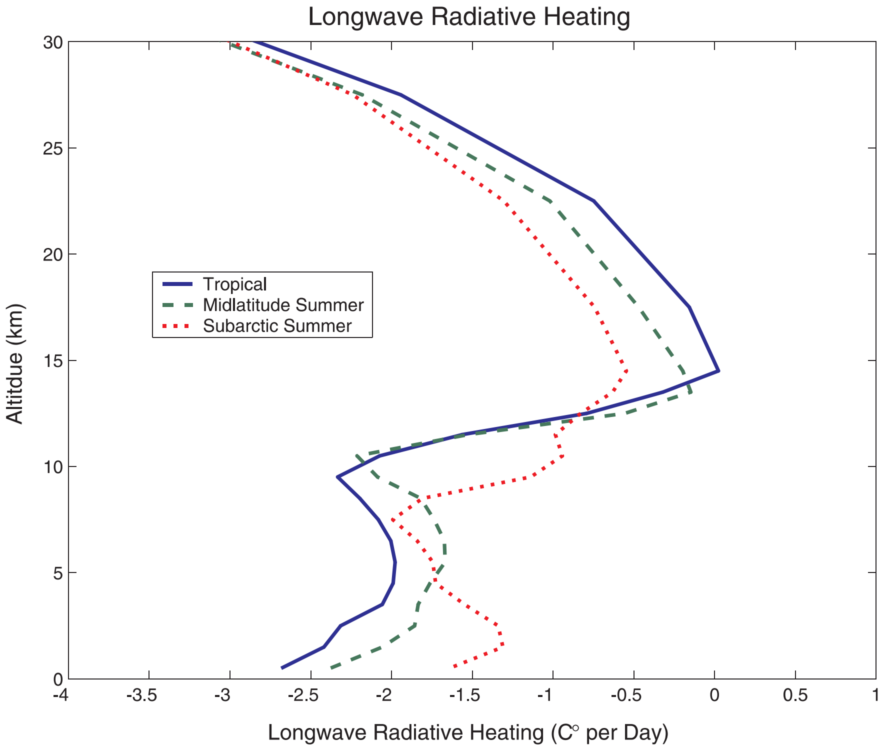


Figure 9 Longwave radiative heating for standard tropical, midlatitude summer, and subarctic summer conditions. See ftp site for color image.

$$\lim_{\delta \to 0} A = \frac{\delta}{\mu_0}(1 - \omega_0) \tag{11}$$

$$\lim_{\delta \to \infty} R = \frac{\omega_0(\alpha_2 + k\gamma_3)}{(1 + k\mu_0)(k + \gamma_1)} \tag{12}$$

These limits become useful in understanding the relationship between a cloud's microphysical properties and its radiative properties. For example, the reflectance is inversely proportional to the cosine of the solar zenith angle. Thus, as the sun gets lower in the sky, the reflectance of a cloud increases. Cloud albedo increases with optical depth, or the water path, though it does approach a limiting value. Reflectance is directly proportional to the single scattering albedo. As the particles composing a cloud get smaller, ω_0 increases, and so does the cloud reflectance.

Cloud absorption decreases with increasing solar zenith angle, the opposite of the case for gases. For optically thin clouds, and thin aerosol layers, the absorption is proportional to $(1 - \omega_0)$. The variation of absorption with increasing r_e is a combination of two competing factors: Droplet absorption $1 - \omega_0$ increases with increasing drop size, while the extinction, and thus optical depth, decreases with increasing

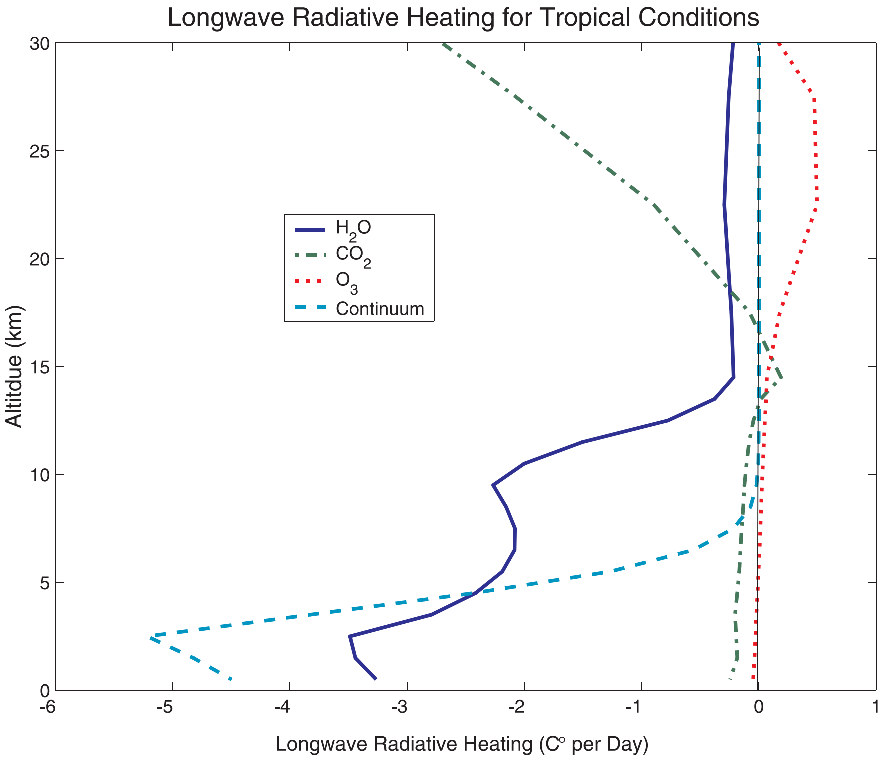


Figure 10 Longwave radiative heating for standard tropical conditions by individual gases. See ftp site for color image.

r_e. For low liquid water path (LWP) clouds and in the weak absorption regime, the combination of these effects renders A approximately independent of r_e. In the strong absorption regime, these two effects combine to produce an absorption that is a decreasing power of r_e. By contrast, reflectance varies primarily as a function of $1/r_e$ through the dependence of extinction on droplet size.

Figures 13 and 14 show the reflectance and transmittance of a water cloud as a function of effective radius for three different liquid water contents (LWC) and three different wavelengths. The red lines are for wavelengths of 0.5 μm and represents the case of no absorption. The blue and black lines are for wavelengths of 2.1 and 2.9 μm, which represent moderate and strong absorption, respectively. The previous chapter demonstrated the inverse dependence of σ_{ext} on r_e. Combining this with the two-stream model, we see that cloud reflectance decreases as an approximate inverse function of r_e for low LWP clouds. The sensitivity between R and r_e decreases with increasing LWP and is most acute at weakly absorbing wavelengths. The cloud transmittance increases with increasing r_e. The limit is a function of the index of refraction.

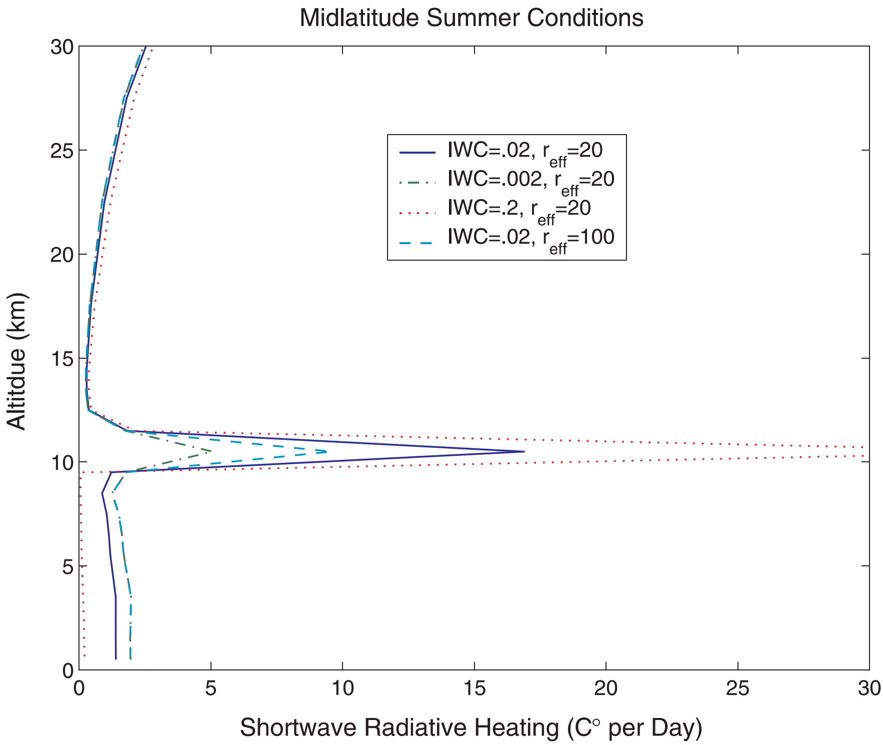


Figure 11 Impact of clouds with differing cloud microphysical properties on radiative heating in a midlatitude summer atmosphere. Clouds are 1 km thick and the solar zenith angle is 15°. See ftp site for color image.

Figure 15 shows the cloud absorptance as a function of effective radius for three different LWPs and two different wavelengths. This figure demonstrates that the absorption of solar radiation by a cloud layer varies with size spectra in a manner dependent on the cloud LWP. For deep clouds, with large LWPs, the absorption increases monotonically with r_e, due to the droplet radius dependence on $1 - \omega_0$. For thin cloud layers and low LWPs, the absorption depends on the combined effects associated with the variation of $1 - \omega_0$ and σ_{ext} with r_e. For small LWP, absorption decreases with increasing r_e.

Volcanic Aerosols

Volcanoes can inject gases and particles into the stratosphere, where the residence times for particles are on the order of 1 year, in contrast to a residence time of 1 week for tropospheric particles. Sulfur-bearing gases such as SO_2 are photochemically converted to sulfuric acid aerosols, which quickly establish themselves as the aerosol species. The dominant visible optical depth of the resulting particulate cloud depends

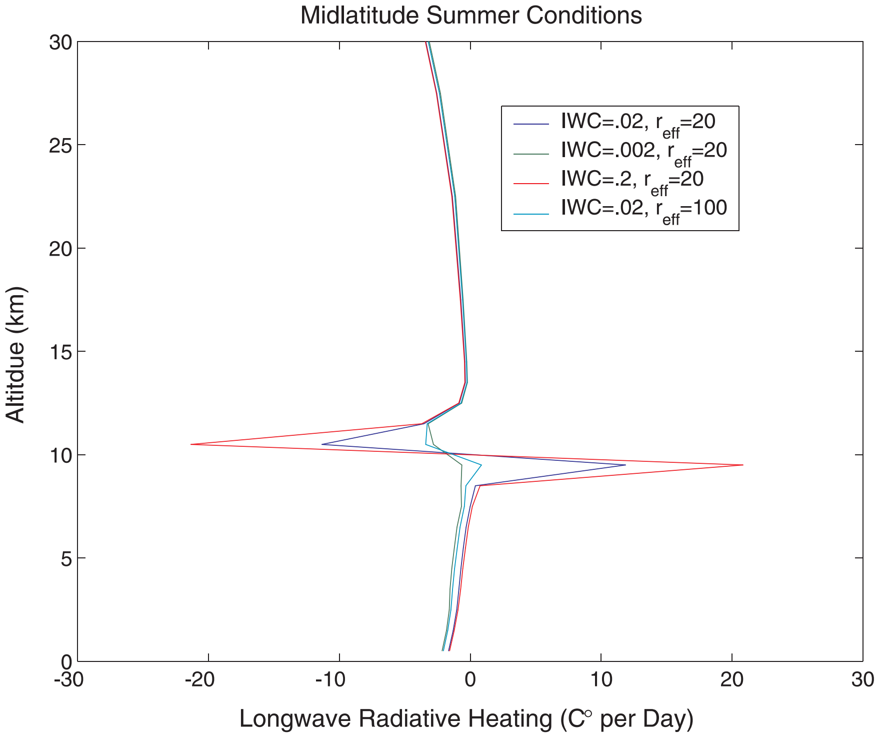


Figure 12 Impact of clouds with differing cloud microphysical properties on the longwave radiative heating in a midlatitude summer atmosphere. Cloud is 2-km thick with a top at 10 km. See ftp site for color image.

on the sulfur content of the volcanic effluents. Thus, although Mt. St. Helens and El Chichon injected similar quantities of ash into the stratosphere, the former caused an optical enhancement of 10^{-2} at 0.55 μm over the Northern Hemisphere during the first year, while the latter caused an enhancement in excess of 10^{-1} because it was a sulfur-rich volcano.

The single scattering albedo of the volcanic aerosols is quite close to unity (0.995) at visible wavelengths. Thus, these aerosols can be expected to cause Earth's albedo to increase, temperatures in the lower stratosphere to increase, and temperature in the troposphere to decrease. Measurements of the El Chicon cloud imply that the thermal effects of the volcanic aerosols are important for the heat balance of both the troposphere and stratosphere, although it dominates in the latter. The stratospheric cloud consisted of sulfuric acid droplets with very small amounts of volcanic ash. The particle size distribution had a mean mode radius of 0.3 μm. Spectral measurements of optical depth show a relatively flat distribution through the visible spectrum with a peak optical depth at 0.55 μm.

The sign and magnitude of thermal infrared heating due to a stratospheric volcanic aerosol of a given optical depth varies significantly with the height and

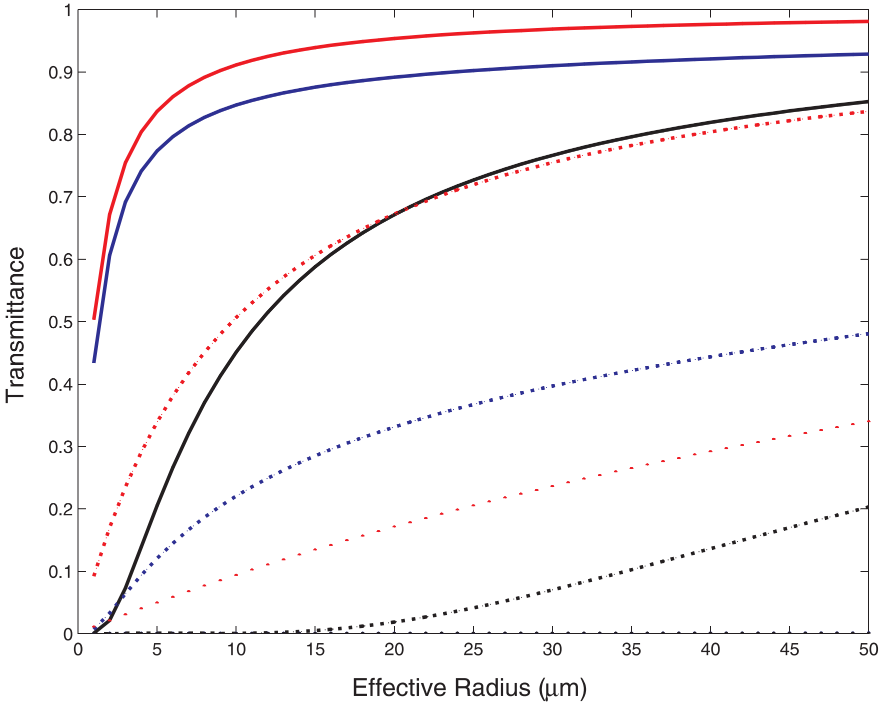


Figure 13 Impact of effective radius on transmittance as function of three wavelengths and three ice water contents. Three wavelengths are 0.5 μm (red), 2.1 μm (blue), and 2.9 μm (black). Liquid water content (LWC) of 0.01, 0.1, and 1 g^{-3} are represented by the solid, dashed, and dotted lines, respectively. See ftp site for color image.

latitude of the aerosols, and also with the height of the tropospheric cloud beneath the aerosol layer and its emissivity. The aerosol of El Chicon was mostly sulfuric acid particles and had significant opacity in the 760- to 1240-cm^{-1} window region. Thus, the stratospheric aerosols can cause a net heating by absorption of upwelling radiation from a warm surface; the aerosols own infrared emission is small at the cold stratospheric temperatures.

Modeling results indicate that these volcanic particles caused a warming by the lower stratosphere of several degrees and a cooling of the troposphere of a few tenths by a degree over their first year.

The largest volcanic eruption in recent history has been that of Mount Pinatubo. Observations of the Earth Radiative Budget Experiment (ERBE) have indicating that the aerosol is causing a net radiative cooling of Earth's atmosphere system. The Mount Pinatubo aerosols resulted in an enhancement of the Earth–atmosphere albedo, with a smaller shift in the lower outgoing longwave radiation.

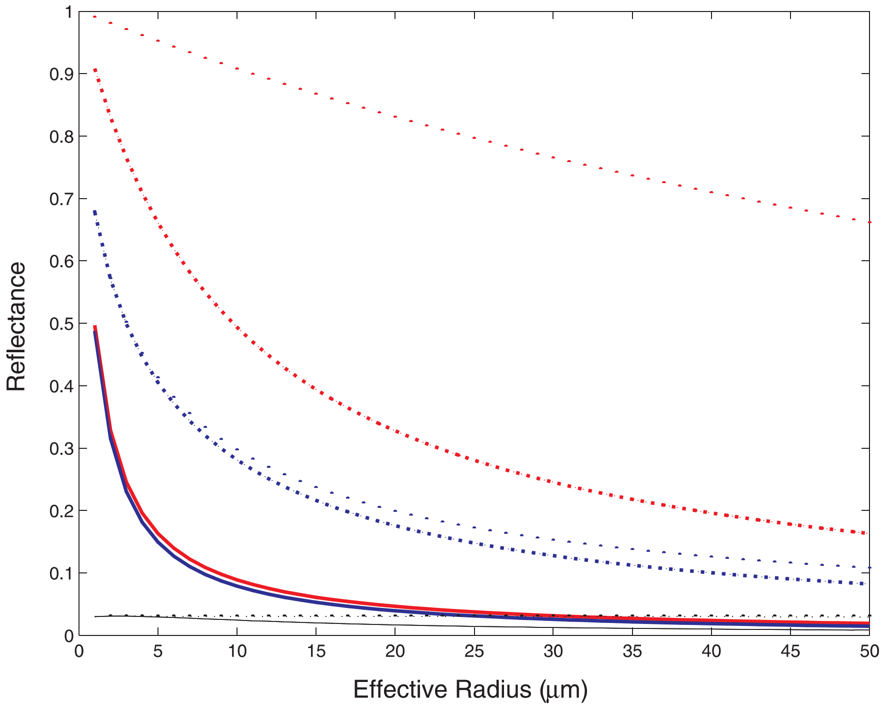


Figure 14 Impact of effective radius on reflectance as function of three wavelengths and three ice water contents. Three wavelengths are 0.5 μm (red), 2.1 μm (blue), and 2.9 μm (black). The LWC and 0.01, 0.1, and 1 g^{-3} are represented by the solid, dashed, and dotted lines, respectively. See ftp site for color image.

4 TOP-OF-ATMOSPHERE RADIATION BUDGETS

If we consider the planet as a system, Earth exchanges energy with its environment (the solar system) via radiation exchanges at the top of the atmosphere. The balance between radiative energy gains and radiative energy losses at the top of the atmosphere is referred to as the Earth radiation budget. The determination of Earth's radiation budget is essential to atmospheric modeling and climate studies as it determines net energy gains and losses of the planet. Radiation budget experiments have used satellites to measure the fundamental radiation parameters:

- Amount of solar energy received by the planet
- Planetary albedo (the portion of incoming solar radiation reflected back to space)
- Emitted terrestrial radiation [also referred to as outgoing longwave radiation (OLR)]
- Net planetary energy balance (difference between the absorbed solar energy and the OLR)

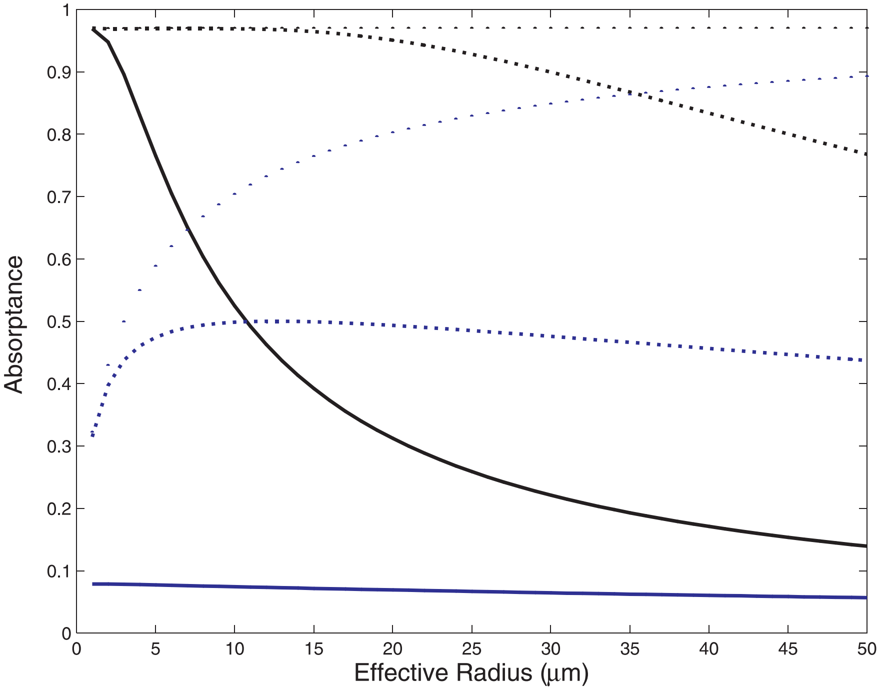


Figure 15 Impact of effective radius on absorption as function of three wavelengths and three ice water contents. Three wavelengths are 0.5 μm (red), 2.1 μm (blue), and 2.9 μm (black). The LWP of 10, 100, and 1000 g^{-2} are represented by the solid, dashed, and dotted lines, respectively. See ftp site for color image.

These elements are described in Figure 16.

Averaged over the globe for a year, the incoming shortwave flux is ($S_0/4$) where S_0 is the solar constant, 1368 W/m^2. Of this incoming energy, approximately 17%, or 82 W/m^2, is absorbed by the atmosphere, with about 103 W/m^2 (30%) reflected back to space (24% from the atmosphere due to clouds, aerosols, and Rayleigh scattering and 6% from the surface). The net shortwave flux at the surface is 157 W/m^2, or about 53% of that incident at the top of atmosphere. In the longwave, 239 W/m^2 leaves Earth to balance the shortwave gain. Of this 239 W/m^2, approximately 57% is due to emission by the atmosphere and 13% is due to surface emission that is transmitted through the atmosphere. Compared to the top of the atmosphere, the surface loses approximately 51 W/m^2 in the longwave, 88% of which is absorbed by the atmosphere. The surface also gains longwave energy emitted by the atmosphere. Thus, while the net top of atmosphere flux is zero, the surface balance is 21 W/m^2. To retain energy balance the surface must lose or store energy. Neglecting storage, the surface loses 21 W/m^2, which is transferred to the atmosphere to balance the net radiative loss. This is accomplished via latent and sensible heat fluxes. Thus, the atmospheric radiative cooling is balanced by the latent

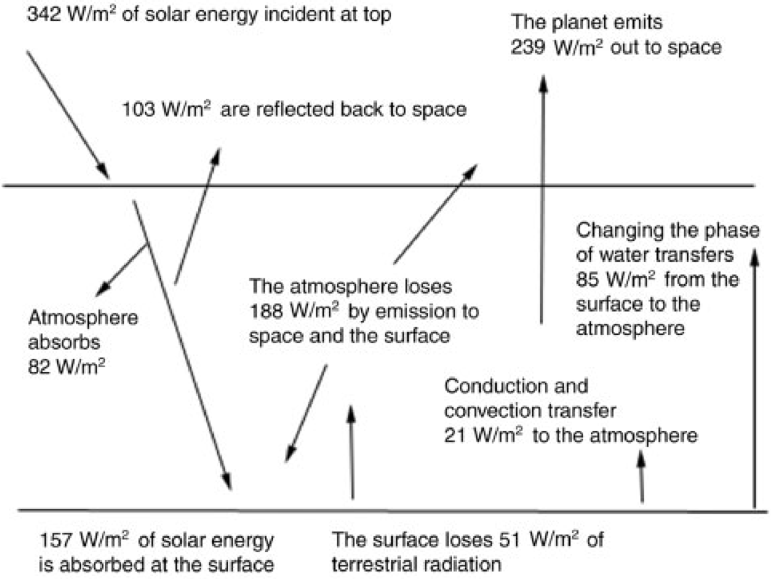

342 W/m² of solar energy incident at top
The planet emits 239 W/m² out to space
103 W/m² are reflected back to space
Changing the phase of water transfers 85 W/m² from the surface to the atmosphere
Atmosphere absorbs 82 W/m²
The atmosphere loses 188 W/m² by emission to space and the surface
Conduction and convection transfer 21 W/m² to the atmosphere
157 W/m² of solar energy is absorbed at the surface
The surface loses 51 W/m² of terrestrial radiation

Figure 16 Average annual global energy budget in W/m^2.

heat of condensation, which is released in regions of precipitation, and by conduction of sensible heat from the surface.

Earth's Top of the Atmosphere Radiation Budget

Independent satellite observations of the longwave emitted energy to space have indicated that on a global annual mean basis, the absorbed solar radiation is in balance with the outgoing longwave radiation (to within instrument noise). This balance exists on a hemispherical scale, indicating that there is little net cross-equatorial energy transport.

Seasonal and latitudinal variations in temperature are driven by the variations in the incoming solar radiation. Measurements of the solar insolation at the top of the atmosphere are important, since it is the primary climate external forcing mechanism. Methods of calculating these variations have already been discussed. Measurements from satellites indicate that there are small variations in the solar insolation, on the order of 0.1 to 0.3%.

The global annual averaged albedo is approximately 0.30. The globally and annually averaged planetary albedo is a key climate variable since it, combined with the solar insolation, determines the radiative energy input to the planet. Variations in the global mean albedo result from the eccentricity of Earth's orbit and geographical differences between the Northern and Southern Hemispheres. The annual average albedo of the Northern and Southern Hemispheres is nearly the same, demonstrating the important influence of clouds.

There are significant variations in the month-to-month global mean albedo, longwave, and net flux (Fig. 17). The annual cycle in the global monthly means are due to Earth's orbit about the sun and the geographical differences between the Northern and Southern Hemispheres. The range in the OLR throughout the year is approximately 10 W/m^2, with a maximum in July and August. This maximum results from there being more land in the Northern Hemisphere than the Southern. The annual average global albedo is approximately 30%, with an amplitude of approximately 2%. The planetary albedo reaches a maximum around October and November. This albedo variation reduces the impact of the annual variation in incoming solar radiation on the net radiation budget.

The amplitude of the annual cycle of globally averaged net flux is approximately 26 W/m^2. This is similar to the peak-to-peak amplitude in the external forcing associated with variations in the solar insolation due to Earth–sun geometry. The interannual variation of the hemispherical averaged net flux shows maximum heating during the summer with the largest changes occurring for the transition from solstice to equinox.

The zonal annual average absorbed solar radiation exceeds the outgoing longwave radiation in the tropical and subtropical regions, resulting in a net radiative heating of the surface–atmospheric column; while in the mid to polar latitudes there

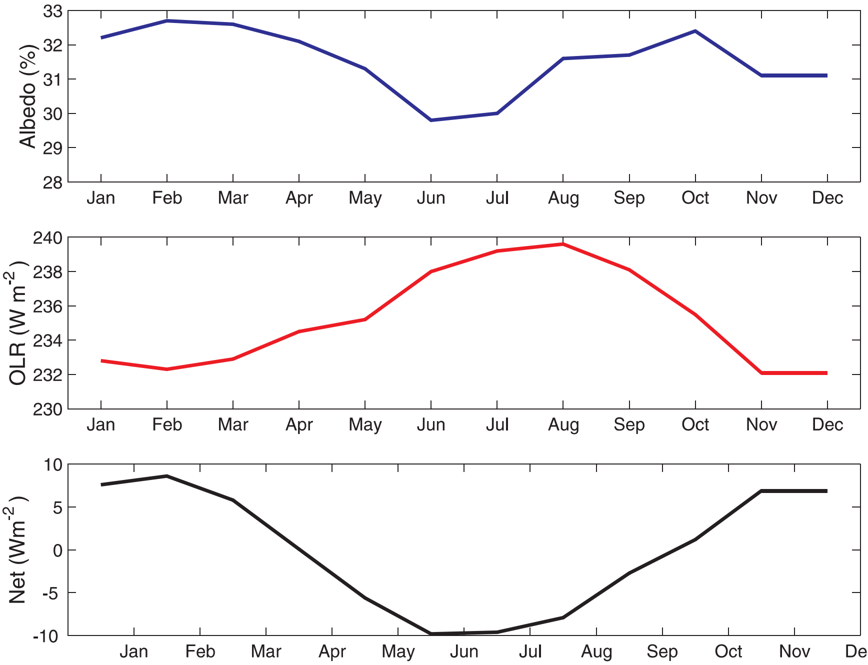


Figure 17 Annual variation in planet's energy budget components: albedo (top), OLR (middle), and net radiation (bottom). See ftp site for color image.

is a net divergence (Fig. 17). This equator-to-pole gradient in radiative energy is the primary mechanism that drives the atmospheric and oceanic circulations. On an annual and long-term basis no energy storage and no change in the global mean temperature occurs, so the zonal mean radiative budget must be balanced by meridional heat transport by the atmosphere and oceans.

The minimum in emitted flux by the planet located near the equator is due to the high cloud tops associated with the Inter-Tropical Convergence Zone (ITCZ) (Fig. 18). This minimum migrates about the equator as seen in the seasonal profiles and is seen as a maximum in albedo. Note also the large emission in the vicinity of the subtropical highs and the corresponding lower albedos. The lowest values of OLR are associated with the Antarctic plateau in winter. This region is very cold and the high altitude means that most of the surface-emitted radiation escapes to space. Maximum OLR occurs in the tropics. Throughout the year the OLR slowly increases toward the summer hemisphere.

The largest albedos are associated with the polar regions, which are snow covered and have high solar zenith angles. The increasing albdeo with latitude is, in general, due to the increasing solar zenith angle (Fig. 19). In the Northern Hemisphere the albedo is larger in summer than winter, due to the increase in cloud cover and optical

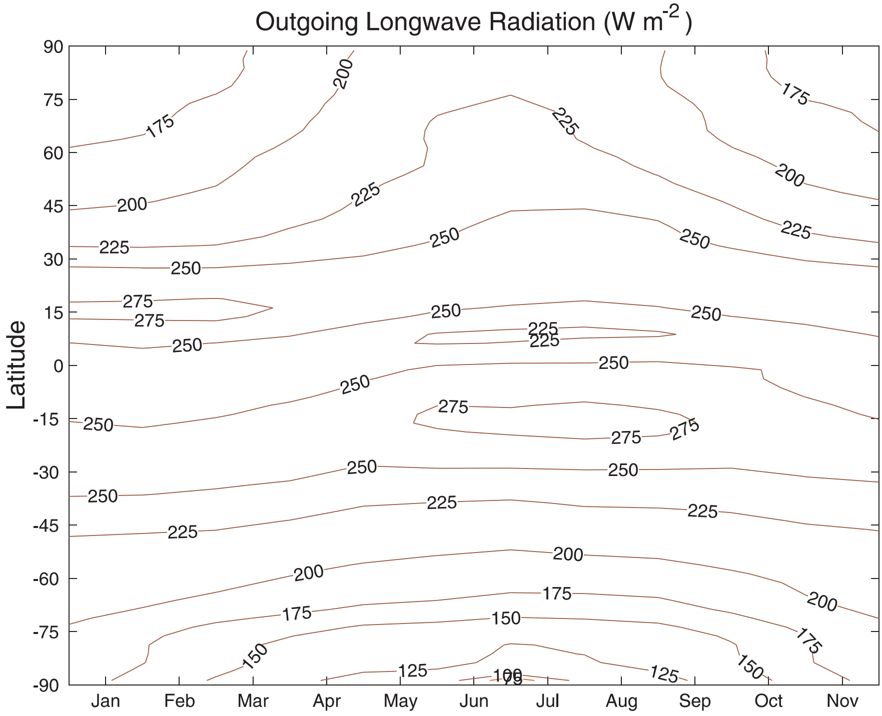


Figure 18 Annual cycle of monthly mean OLR in W/m^{-2} determined from ERBE observations. See ftp site for color image.

thickness. In the tropical regions the albedo variation is influenced primarily by weather disturbances and their associated cloud distributions. In the polar region, variations are due to the distribution of major ice sheets and the decreasing mean solar elevation angle with latitude. The annual variation in the zonal mean absorbed solar radiation follows the variation of the solar declination due to the annual variation of the incoming solar energy being greater than the annual variation of the albedo. The net energy also exhibits this same dependency.

There are net radiation energy gains in the tropics and subtropics with energy losses over the polar regions (Fig. 20). The region of net energy gains tracks the solar declination. Differences between the Northern and Southern Hemispheres result from differences in land distributions. While the minimum OLR occurs over Antarctica, the minimum net losses occur between approximate 60 S and 70 S. Maximum energy gains occur in the Southern Hemisphere subtropical regions during December and January. Differences between land and water are more evident in a regional analysis of Earth's radiation budget.

Figures 21 and 22 are the ERBE measured January and July OLR. Maximum OLR occurs over the deserts and cloud free ocean regions of the subtropical highs. Relative minimums in tropical regions result from high, thick cirrus associated with

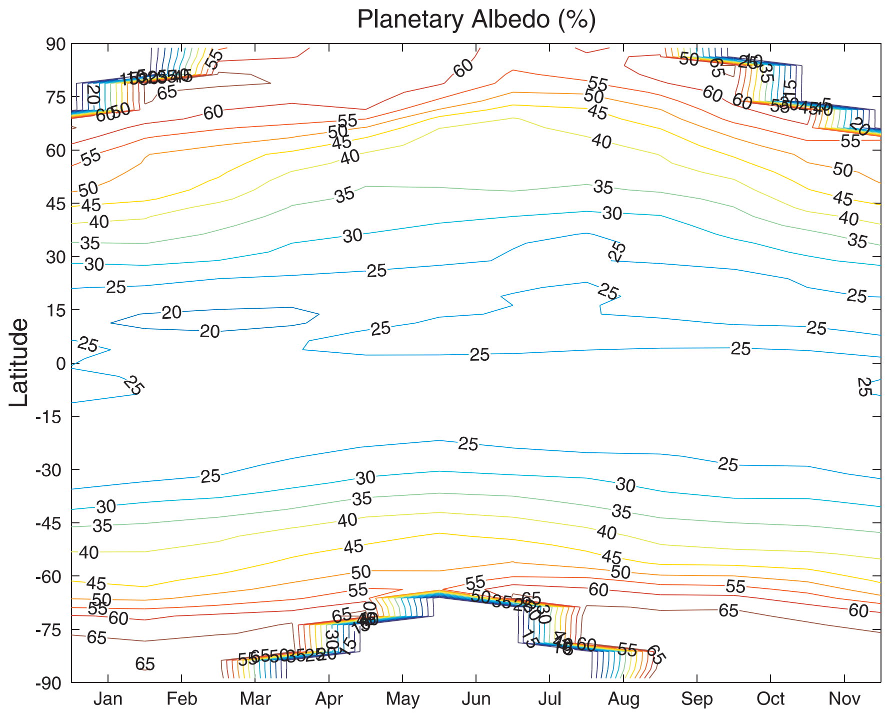


Figure 19 Annual cycle of monthly mean albedo in percent determined from ERBE observations. See ftp site for color image.

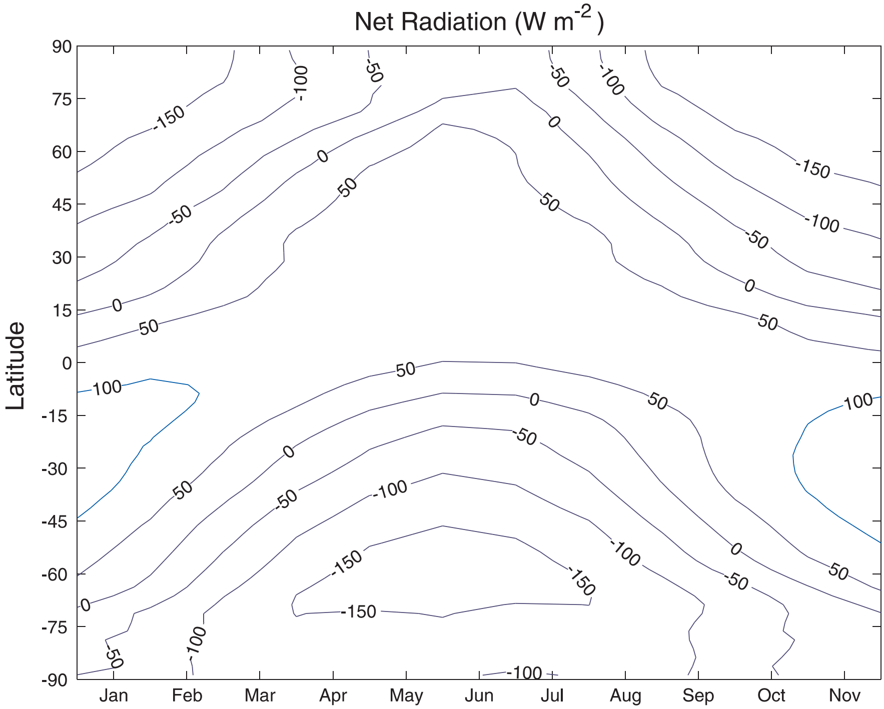


Figure 20 Annual cycle of monthly mean net radiation in W/m^{-2} determined from the ERBE observations. See ftp site for color image.

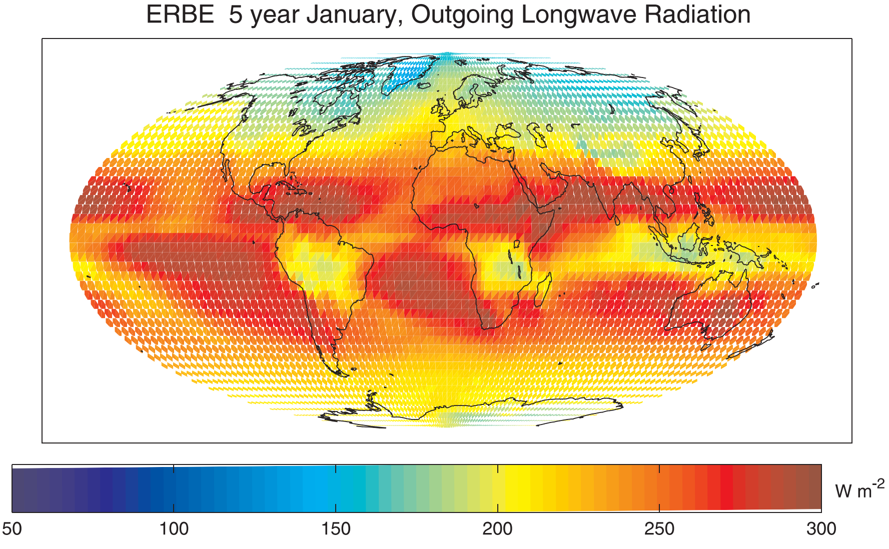


Figure 21 (see color insert) January mean OLR in W/m^{-2} determined from ERBE observations. See ftp site for color image.

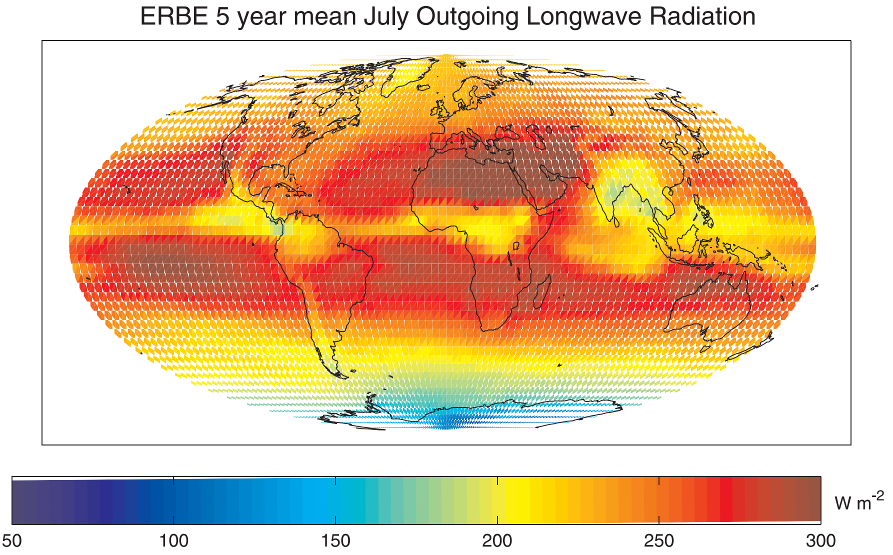


Figure 22 (see color insert) July mean OLR in W/m^{-2} determined from ERBE observations. See ftp site for color image.

tropical convection regions (e.g., Indonesia, Congo Basin, central South America). These thick clouds also yield local minima in the January and July averaged planetary albedos (Figs. 23 and 24). Regional albedo maps for January and July indicate the presence of maritime stratus regions located off the west coast of continents.

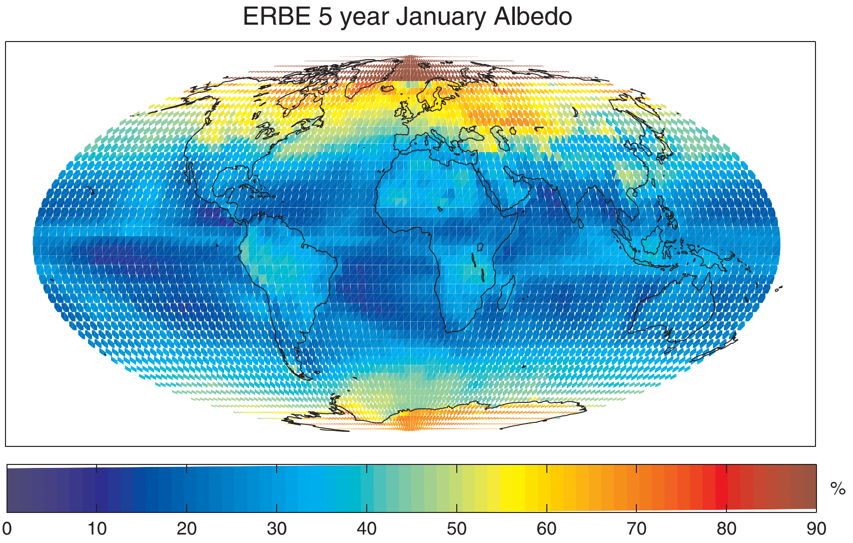


Figure 23 (see color insert) January mean albedo in percent determined from ERBE observations. See ftp site for color image.

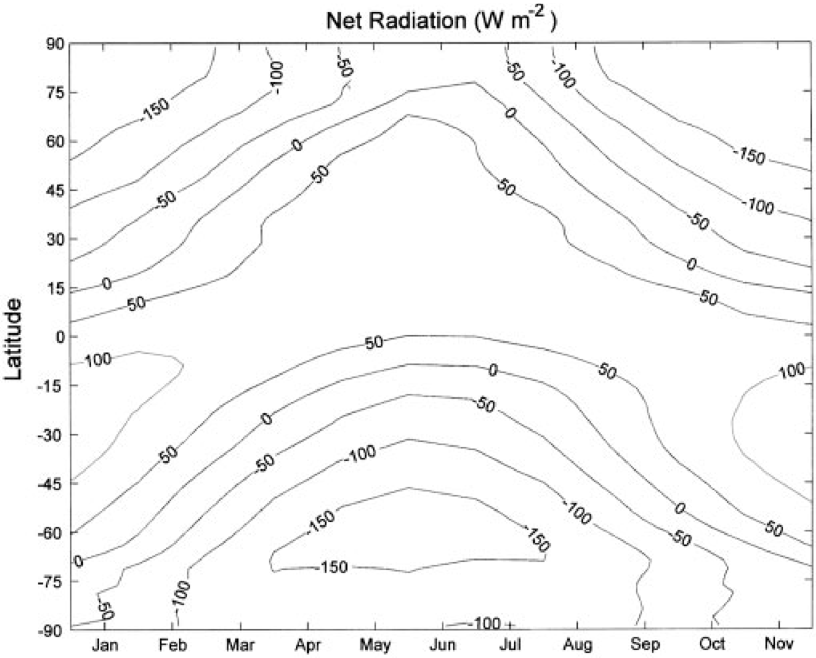


Figure 24 July mean albedo in percent determined from ERBE observations. See ftp site for color image.

These features do not appear on the OLR maps because of the similarity between the sea surface temperature and cloud effective temperature. The ITCZ appears in the albedo maps as a regional enhancement, while in maps of OLR it is a relative minimum.

Analysis of the regional distribution of the net radiative energy gains and losses for January and July is shown in Figures 25 and 26, respectively. These figures clearly indicate that the summer hemisphere gains radiative energy while the winter hemisphere is a net radiative sink. In July, the large desert regions of northern Africa and Saudi Arabia also exhibit a net radiative loss. This arises from the high surface albedos and high surface temperatures and overall cloud-free conditions. The largest energy gains are associated with cloud-free ocean regions in the summer hemisphere, due to the relatively low albedo and high solar input.

The measured outgoing longwave radiation and albedo also indicate regional forcing mechanisms. For example, in the tropics longitudinal variations can be as large as the zonal averages and are associated with east–west circulations. While tropical regions in general display a net radiative heating, the Sahara is radiatively cooling. This is due to the high surface albedo, the warm surface temperatures, and the dry and cloud-free atmosphere. The radiative cooling is maintained by subsidence warming, which also has a drying effect and therefore helps maintain the desert.

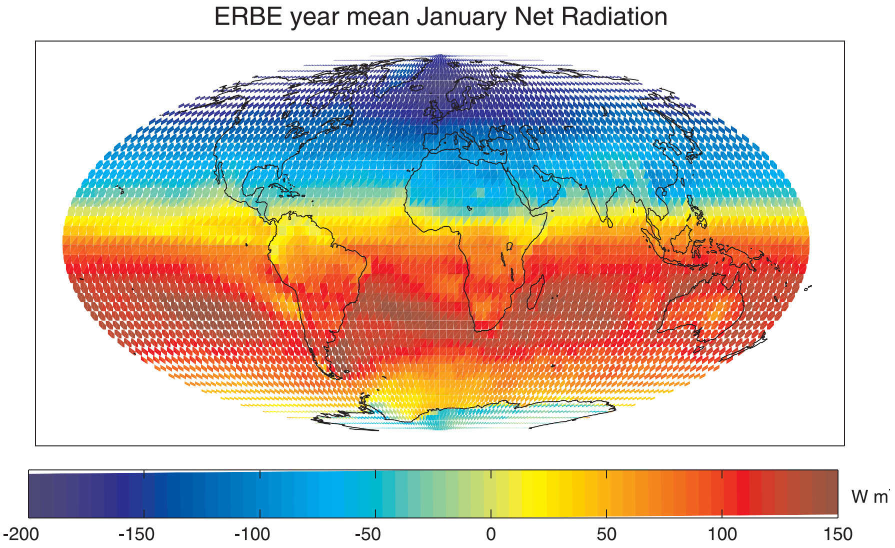


Figure 25 (see color insert) January mean net radiation in W/m^{-2} determined from ERBE observations. See ftp site for color image.

Cloud-Radiative Forcing

One of the major research problems in radiative transfer applications is how clouds affect the energy budget of the planet and thus impact climate. The net radiative heating H within a column is

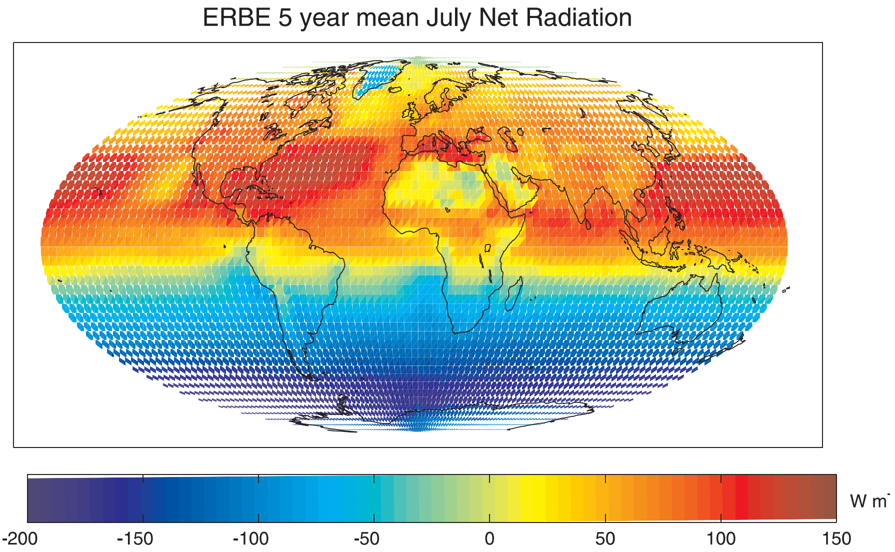


Figure 26 (see color insert) July mean net radiation in W/m^{-2} determined from ERBE observations. See ftp site for color image.

$$H = S_0(1 - \alpha) - \mathrm{OLR} \tag{13}$$

where $S_0(1 - \alpha)$ is the absorbed solar energy and OLR is the outgoing longwave energy. To describe the effects of clouds on H from observations we define the cloud forcing,

$$C = H - H_{\mathrm{clr}} \tag{14}$$

where H_{clr} is the clear-sky net heating. Thus, cloud-radiative forcing is the difference between the all-sky (cloudy and clear regions) and cloud-sky fluxes. We can write

$$C = C_{\mathrm{SW}} + C_{\mathrm{LW}} \tag{15}$$

where

$$C_{\mathrm{SW}} = S(\alpha_{\mathrm{clr}} - \alpha) \tag{16}$$

and

$$C_{\mathrm{LW}} = \mathrm{OLR}_{\mathrm{clr}} - \mathrm{OLR} \tag{17}$$

In the longwave, clouds generally reduce the LW emission to space and thus result in a heating. While in the SW, clouds reduce the absorbed solar radiation, due to a generally higher albedo, and thus result in a cooling of the planet. Results from ERBE indicate that in the global mean, clouds reduce the radiative heating of the planet. This cooling is a function of season and ranges from approximately -13 to $-21\,\mathrm{W/m^2}$, with an uncertainty of approximately $5\,\mathrm{W/m^2}$. On average, clouds reduce the globally absorbed solar radiation by approximately $50\,\mathrm{W/m^2}$, while reducing the OLR by about $30\,\mathrm{W/m^2}$. These values may be compared with the $4\,\mathrm{W/m^2}$ heating predicted by a doubling of CO_2 concentration.

Variations in net cloud forcing are associated with surface type, cloud type, and season. These dependencies can be seen in maps of January and July cloud net radiative forcing (Figs. 27 and 28). Maritime stratus tend toward a negative net cloud forcing as the shortwave effects dominate the longwave. The deserts of North Africa and Saudi Arabia have a positive cloud radiative forcing as the longwave dominates the cloud impact on the albedo.

5 RADIATION AND THE GREENHOUSE EFFECT

The energy that fuels the atmospheric and oceanic circulations originates from the sun. If the total energy content of the Earth–atmosphere system does not vary significantly with time, there must be a close balance between the incoming absorbed solar radiation and the outgoing terrestrial emitted thermal radiation. In

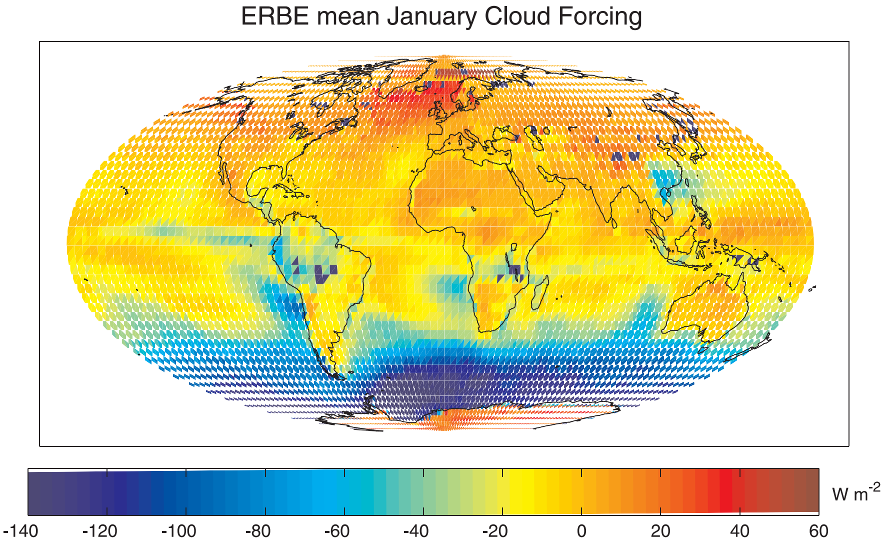


Figure 27 (see color insert) January cloud radiative forcing in W/m^{-2}. See ftp site for color image.

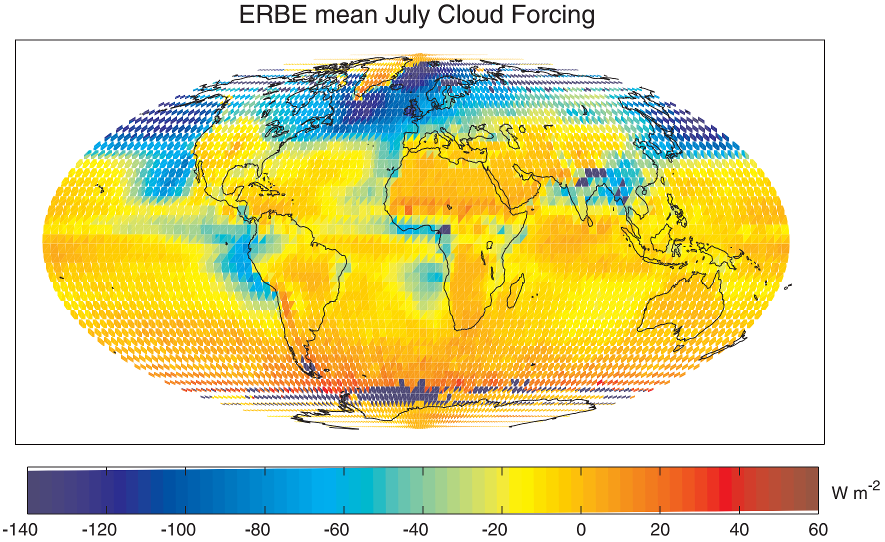


Figure 28 (see color insert) July cloud radiative forcing in W/m^{-2}. See ftp site for color image.

other words, when averaged over the entire globe for the year, the net energy gain equals the net loss

$$\text{Absorbed sunlight} = \text{terrestrial emission} \tag{18}$$

$$S_0 \pi R_e^2 (1 - \alpha_e) = 4\pi R_e^2 \sigma T_e^4 \tag{19}$$

where α_e is the albedo (broadband reflectance) of Earth and T_e is the effective temperature. For an albedo of $\alpha_e = 0.3$, $T_e \approx 255^\circ$, which is considerably cooler than the average surface temperature of 288 K.

The average surface temperature is greater than the effective radiative temperature because of the radiative properties of the atmosphere. The atmosphere is relatively transparent to solar radiation while absorbing and emitting longwave radiation effectively.

The spectral selectivity of absorption by atmospheric gases is the fundamental cause of the greenhouse effect. Much of the shortwave, or solar, energy passes through the atmosphere and warms the surface. While the atmosphere is transparent to shortwave radiation, it is efficient at absorbing terrestrial or longwave radiation emitted upward by the surface. So, while carbon dioxide and water vapor comprise only a very small percentage of the atmospheric gases, they are extremely important because of their radiative properties, i.e., their abilities to absorb this longwave radiation and emit it throughout the atmosphere.

Gases that are transparent to solar energy while absorbing terrestrial energy will warm the atmosphere because they allow solar energy to reach the surface and inhibit longwave radiation from reaching outer space. These radiatively active gases are called greenhouse gases. In addition to carbon dioxide, other important greenhouse gases are water vapor, methane (CH_4), and CFCs. Methane and CFCs are important because they absorb terrestrial radiation in the 10- to 12-μm infrared (IR) atmospheric window. The concentration of these latter two gases has also been increasing. Increases in the greenhouse gases with time can potentially result in a climate change, as the atmosphere becomes more effective at absorbing longwave energy emitted by the surface.

Greenhouse warming or the enhanced greenhouse effect are the terms used to explain the relationship between the observed rise in global temperatures and an increase in atmospheric carbon dioxide. In this chapter we have considered one aspect of the enhanced greenhouse effect, absorption and emission of radiation by certain atmospheric gases. How does this play a role in the enhanced greenhouse effect? Let us use carbon dioxide as an example. Increasing the carbon dioxide concentrations in the atmosphere does not appreciably affect the amount of solar energy that reaches the surface. However, since carbon dioxide absorbs longwave radiation, the amount of longwave energy emitted by the surface and absorbed by the atmosphere increases as the atmospheric concentration of carbon dioxide rises. The increased absorption increases the temperature of the atmosphere. The warmer atmospheric temperatures increase the amount of longwave energy emitted by the atmosphere toward the surface, which increases the energy gain of the surface, warming it. Thus, increased concentration of carbon dioxide may result in a warming of Earth's atmosphere and surface.

Water vapor is the most effective greenhouse gas because of its strong absorption of longwave energy emitted by the surface. A warmer atmosphere can mean more water vapor in the atmosphere and possibly more clouds.

Clouds and the Greenhouse Effect

Increased concentration of greenhouse gases can lead to a warming of the atmosphere. As the air temperature warms, the relative humidity initially decreases. Evaporation depends on relative humidity; so, as the atmosphere warms, more evaporation occurs, which adds more water to the atmosphere and enhances the greenhouse warming. Earth and atmosphere keep heating up until the energy emitted balances the amount of sunlight absorbed. But greenhouse gases are not the whole story of climate change. Clouds have a large impact on the solar and terrestrial energy gains of the atmosphere. Clouds reflect solar energy and reduce the amount of solar radiation reaching the surface, and thus cause a cooling of Earth. The thicker the cloud, the more energy reflected back to space, and the less solar energy available to warm the surface and atmosphere below the cloud. By reflecting solar energy back to space, clouds tend to cool the planet. Clouds are also very good emitters and absorbers of terrestrial radiation.

Clouds block the emission of longwave radiation to space. Thus, in the longwave, clouds act to warm the planet, much as greenhouse gases do. To complicate matters, the altitude of the cloud is important in determining how much they warm the planet. Cirrus are cold clouds. Thick cirrus therefore emit very little to space because of their cold temperature, while at the same time cirrus are effective at absorbing the surface-emitted energy. Thus, with respect to longwave radiation losses to space, cirrus tend to warm the planet. Stratus also warm the planet but not as much as cirrus. This is because stratus are low in the atmosphere and have temperatures that are more similar to the surface than cirrus clouds. Stratus absorb radiation emitted by the surface, but they emit similar amounts of terrestrial radiation to space as the surface. To complicate matters still further, how effective a cloud is at reflecting sunlight is a function of how large the cloud droplets or cloud ice crystals are.

Figure 29 depicts the dependence of cloud radiative forcing as a function of cloud temperature and the change in the planetary albedo due to the cloud. Thin, cold clouds tend to warm the planet while thick, warm clouds tend to cool. The zero line indicates those clouds that have no net effect on the top of the atmosphere energy budget. Cold, thick clouds, such as convective systems, have little impact on the energy budget at the top of the atmosphere.

So clouds can either act to cool or warm the planet, depending on how much of Earth they cover, how thick they are, how high they are, and how big the cloud particles are. Measurements indicate that on average, clouds' reflection of sunlight dominates the clouds' greenhouse warming. Thus, today's distribution of clouds tends to cool the planet. But this may not always be the case. As the atmosphere warms the distribution of cloud amount, cloud altitude, and cloud thickness may all

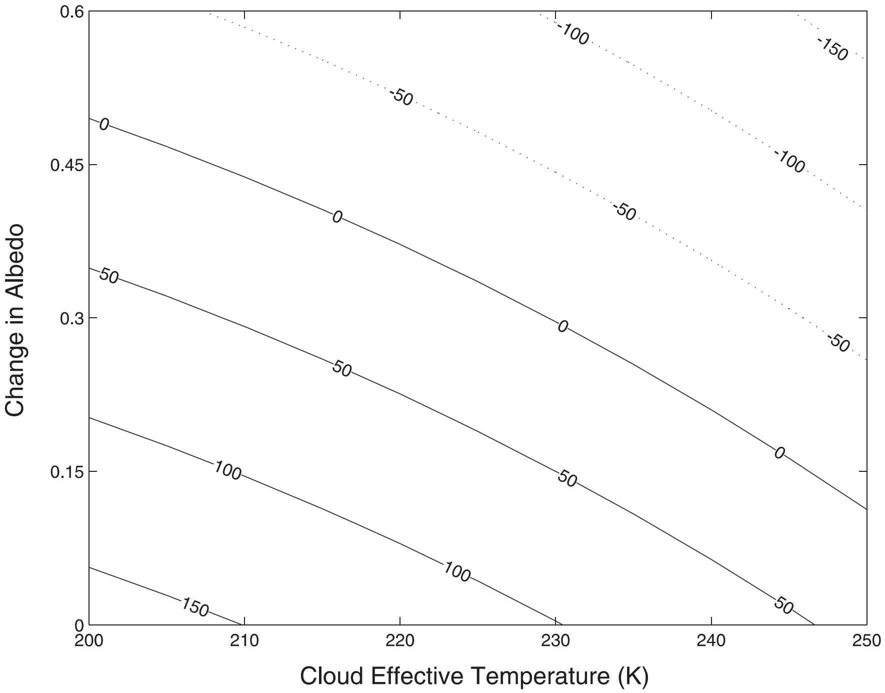


Figure 29 Change in net radiation at top of atmosphere as function of cloud effective temperature and change in planetary albedo due to the cloud. Contours are in W/m^{-2} (after D. Hartmann, (1994) Global Physical Climatology, Academic Press, NY).

change. We do not know what the effect of clouds will be on the surface temperatures if the global climate changes. Clouds could dampen any greenhouse warming by increasing cloud cover or decreasing cloud altitude. On the other hand, clouds could increase a warming if the cloud cover decreases or cloud altitude increases. Climate is so sensitive to how clouds might change that an accurate prediction of climate change hinges on correctly predicting cloud formation.

Radiative Equilibrium

A fundamental challenge of modern science is to predict climate. Recent concerns about global warming and the effect of greenhouse gases added to the atmosphere by humans have heightened the need to understand the processes that cause climate variations. Models are used to gain a better understanding of the atmosphere. We can use a simple model to derive the temperature distribution of an atmosphere in radiative equilibrium by assuming that the only source term is due to thermal emission: $B = \sigma T^4$. For demonstration, we also assume that the atmosphere absorbs at all wavelengths. The upwelling and downwelling fluxes are then

$$\frac{d(F\uparrow - F\downarrow)}{d\tau} = (F\uparrow + F\downarrow) + 2B \tag{20}$$

$$\frac{d(F\uparrow + F\downarrow)}{d\tau} = (F\uparrow - F\downarrow) \tag{21}$$

We can directly solve this set of equations to yield the emission as a function of optical depth:

$$B(\tau) = \frac{F\uparrow - F\downarrow}{2}\tau + B_0 \tag{22}$$

where B_0 is the emission of the first layer of atmosphere ($\tau = 0$).

For thermal equilibrium, the atmosphere and surface outgoing longwave flux must balance the incident solar flux, F_0, so that

$$F\uparrow (0) = F_0 \tag{23}$$

and

$$(F\uparrow - F\downarrow) = F_0 \tag{24}$$

yielding

$$B(\tau) = \frac{F_0}{2}(\tau + 1) \tag{25}$$

At the surface, the upwelling flux emitted by the surface must balance the absorbed solar flux and the downwelling flux emitted by the atmosphere [$F\downarrow (\tau_s)$].

$$B(T_s) = F_0 + F\downarrow(\tau_s) \tag{26}$$

which leads to

$$B(T_s) = B(\tau_s) + \frac{F_0}{2} \tag{27}$$

This simple model demonstrates that the atmospheric temperature profile is a function of its optical depth. It also shows that the surface temperature is warmer than the overlying air. The greater the optical depth the greater the difference between the surface temperature and the effective temperature of the planet.

6 SATELLITE REMOTE SENSING

Meteorologists use two basic methods of observing the atmosphere: in situ and remote-sensing methods. In situ methods, for "in place," measure the properties of the air in contact with an instrument. Remote-sensing methods obtain information without coming into physical contact with the region of the atmosphere being measured. Remote sensing the atmosphere is the emphasis of this section. In remote sensing, radiation measurements are used to infer the state of the atmosphere. The inference of the atmospheric state is often referred to as the retrieval process (see page 39 for discussion of retrieval process).

There are two basic types of remote sensing the atmosphere: active sensors and passive sensors. Active remote-sensing instruments emit energy, such as a radio wave or beam of light, into the atmosphere and then measure the energy that returns. Passive remote-sensing instruments measure radiation emitted by the atmosphere, Earth's surface, or the sun. Much of what we observe about the atmosphere using remote-sensing techniques results from how radiation interacts with molecules or objects, such as water drops, suspended in the atmosphere. The previous chapter discussed the principles on which remote-sensing methods are based, while this section looks at some applications of passive remote sensing from satellites.

Remote Sensing the Surface

In remote sensing the surface of Earth from a satellite, we select spectral regions, or channels, in which the atmosphere is transparent. These are called atmospheric windows. In the solar spectrum, or the shortwave, very little absorption occurs in the visible region; however, Rayleigh scattering is large. Rayleigh scattering is well understood and can be handled via modeling as described in the previous chapter. There are also windows in the near-infrared spectral regions where Rayleigh scattering is smaller.

Analysis of the spectral reflectance of different surfaces generally shows a distinct difference between the visible and near-IR regions. Observations in both the visible (such as 0.58 to 0.68 μm) and near-infrared (such as 0.725 to 1.1 μm) are useful for monitoring surface conditions. Vegetation regions generally have reflectances in the near-infrared (NIR) that range from 20 to 40%, while visible reflectances generally range from 5 to 15%. Soils also have a higher reflectivity in the NIR than in the visible while the opposite is true for snow.

A common method of monitoring surface vegetation is through the normalized difference vegetation index (NDVI):

$$\mathrm{NDVI} = \frac{R_{\mathrm{NIR}} - R_{\mathrm{VIS}}}{R_{\mathrm{NIR}} - R_{\mathrm{VIS}}} \tag{28}$$

This has long been used to monitor the vegetation, and changes in vegetation, of the entire Earth. NDVI for vegetation generally range from 0.3 to 0.8, with the larger values representing "greener" surfaces. Bare soils range from about 0.2 to 0.3.

Identifying snow cover is important for weather and hydrological forecasting. To detect the presence of snow, recent satellite instruments include observations at 0.66 and 1.6 μm (Fig. 30). The atmosphere is transparent at both these wavelengths, while snow is very reflective at 0.66 μm and not reflective at 1.6 μm. The normalized difference snow index (NDSI),

$$\mathrm{NDSI} = \frac{R_{0.66} - R_{1.6}}{R_{0.66} + R_{1.6}} \tag{29}$$

is used to monitor the extent of snow cover. At visible wavelengths (e.g., 0.66 μm), snow cover is just as bright as clouds and is therefore difficult to distinguish from cloud cover. However, at 1.6 μm, snow cover absorbs sunlight and therefore appears much darker than clouds. This allows the effective discrimination between snow cover and clouds. Values of NDSI <0.4 typically indicate the presence of snow. Figure 30 demonstrates the ability to separate clouds from snow using observations at these wavelengths.

Sea surface temperature (SST) is another surface property we are interested in from a meteorological perspective. An IR window in the atmosphere for sensing

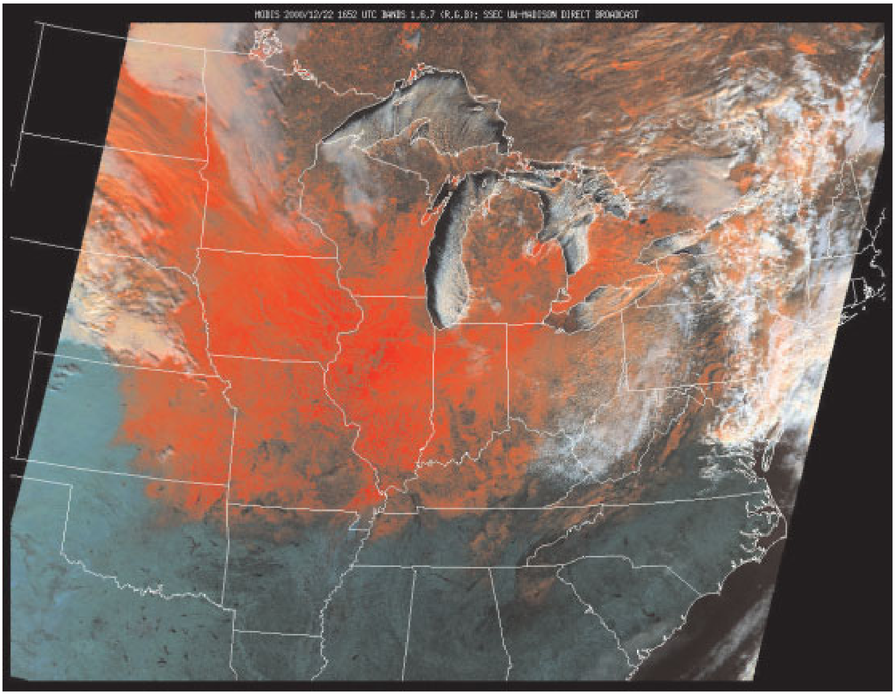

MODIS 2000/12/22 1652 UTC BANDS 1,6,7 (R,G,B); SSEC UW-MADISON DIRECT BROADCAST

Figure 30 (see color insert) False color image from a combination of observations at 0.66, 1.64, and 2.14 μm are used to detect snow. In this false color images, land surfaces are green, water surfaces are black, snow cover is red, and clouds are white. See ftp site for color image.

surface temperature is the 10- to 12-μm region where absorption by water vapor is weak. Figure 31 is a MODIS (moderate resolution imaging spectrometer) 11-μm image that clearly indicates changes of SST in the vicinity of the Gulf Stream. Most of the radiation in this band is emitted by the surface and transmitted through the atmosphere. In a warm moist atmosphere the difference between the SST and the brightness temperature at 11 μm (BT_{11}) can approach 10°C. This difference is often

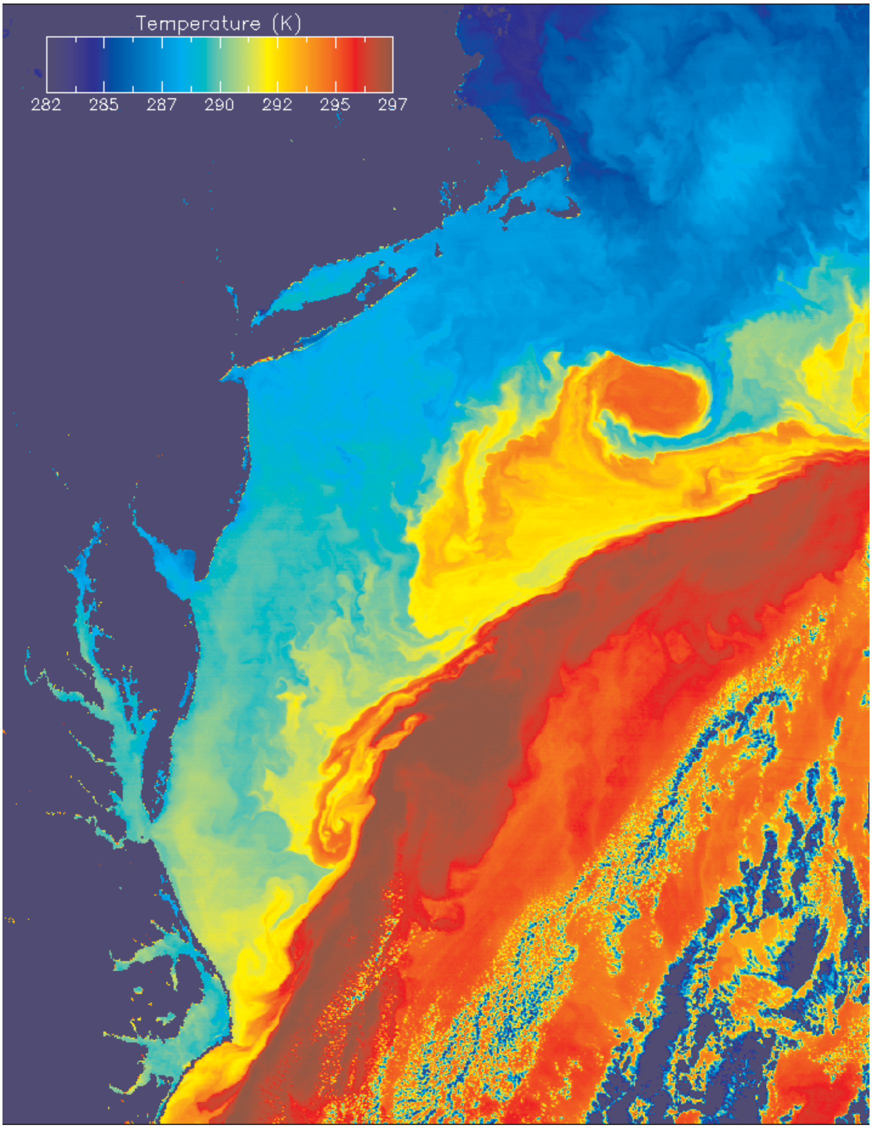


Figure 31 (see color insert) An 11-μm image from MODIS of Atlantic Ocean off east coast of North America. See ftp site for color image.

corrected for by making observations at more than one wavelength, such as 11, 12, 3.7, and 8.5 μm. Differences between these channels represent the total amount of water vapor in the column. For example, the 12-μm channel has more absorption and therefore ($BT_{11} - BT_{12}$) is positive; the greater this difference the larger the water vapor loading of the atmosphere. Observations at these wavelengths are used daily to derive SST. The SST from satellite observations is typically determined from a regression derived empirically using observations from drifting buoys.

Remote Sensing of Clouds

Clouds are generally characterized by higher reflectance and lower temperature than the underlying Earth surface. As such, simple visible and infrared window threshold approaches offer considerable skill in cloud detection. However, there are many surface conditions when this characterization of clouds is inappropriate, most notably over snow and ice. Additionally, some cloud types such as thin cirrus, low stratus at night, and small cumulus are difficult to detect because of insufficient contrast with the surface radiance. Cloud edges cause further difficulty since the instrument field of view will not always be completely cloudy or clear. There are many different methods of detecting clouds. In this section we review some of the more common approaches.

The simplest cloud measurement technique is the threshold method in which an equivalent blackbody temperature or a spectral reflectance threshold is selected that distinguishes between cloud and noncloud in infrared or visible satellite images. Information on cloud top temperature is obtained by comparing the observed brightness temperature with an atmospheric temperature profile—this approach usually underestimates the cloud height. Using a visible or near-infrared reflectance threshold works well for determining clear-sky ocean scenes that are free of sun glint.

Another straightforward approach employs two channels in combination. For example, the split window technique makes use of observations near 11 and 12 μm to detect clouds over oceans. Cloud classification is accomplished by considering the 11-μm blackbody temperature and the difference between 11 and 12 μm. Clear scenes have warm temperatures and brightness temperature differences that are negative, usually less than about $-1°$. Another simple two-channel technique uses visible and infrared observations. In this approach observed visible reflectance and equivalent blackbody temperature are compiled, and observations are then classified based on their relative brightness and temperature. For example, clear sky oceans would be warm and dark while convective clouds would be cold and bright. Classification of the cloud is accomplished by either assigning thresholds or by employing maximum-likelihood statistical techniques.

Once clear pixels can be determined, these observations can be combined with cloudy observations to derive cloud properties. One such approach is the CO_2 slicing technique. In this technique a cloud pressure function $G(\lambda_1, \lambda_2, p)$ is defined as an expression involving a pair of differences of two column radiances and [$I(\lambda^{\mathrm{clr}})$ and $I(\lambda^{\mathrm{cld}})$], one clear and one cloud contaminated with a cloud at pressure level p_c. If you express the observed column radiances in terms of differences, then the implied vertical integration need only go from the surface to the cloud top because the above cloud components subtract; then by taking the ratio of two observed radiances, you can remove the coefficient of cloud amount that results from expanding these terms into clear and cloudy portions. It is necessary to make an assumption that the cloud is infinitesimally thin, so that you need only work with a single transmittance function below the cloud. This method enables us to assign a quantitative cloud top pressure to a given cloud element using observed radiances for the CO_2 spectral bands. Defining the radiance from a partly cloudy scene as

$$I_\lambda = F I_\lambda^{\mathrm{cld}} + (1 - F) I_\lambda^{\mathrm{clr}} \tag{30}$$

where F is the fractional cloud cover. The cloud radiance is given by

$$I_\lambda^{\mathrm{cld}} = \varepsilon_\lambda I_\lambda^{\mathrm{bcld}} + (1 - \varepsilon_\lambda) I_\lambda^{\mathrm{clr}} \tag{31}$$

where ε is the cloud emissivity, and $I_\lambda{}^{\mathrm{bcld}}$ is the radiance from an opaque cloud. Thus

$$I_\lambda^{\mathrm{clr}} = B_\lambda(T_{ps})\tau\lambda(p_s) + \int_{ps}^{0} B_\lambda(T_p)\, d\tau_\lambda \tag{32}$$

$$I_\lambda^{\mathrm{bcld}} = B_\lambda(T_{pc})\tau\lambda(p_c) + \int_{pc}^{0} B_\lambda(T_p)\, d\tau_\lambda \tag{33}$$

where p_c is the cloud top pressure. Integrating by parts (e.g., $\int u\dot{v}\, dx = uv - \int v\dot{u}\, dx$) and subtracting the two terms yields

$$I_\lambda^{\mathrm{clr}} - I_\lambda^{\mathrm{bcld}} = \int_{p_c}^{p_s} \tau_\lambda(p)\, dB_\lambda(T_p) \tag{34}$$

and

$$I_\lambda - I_\lambda^{\mathrm{clr}} = F\varepsilon_\lambda \int_{p_c}^{p_s} \tau_\lambda(p)\, dB_\lambda(T_p) \tag{35}$$

If two wavelengths are chosen that are close to one another, then $\varepsilon_{\lambda_1} \approx \varepsilon_{\lambda_1}$, which leads to

$$\frac{I_{\lambda_1} - I_{\lambda_1}^{clr}}{I_{\lambda_2} - I_{\lambda_2}^{clr}} = \frac{\int_{p_c}^{p_s} \tau_{\lambda_1}(p)\, dB_{\lambda_1}(T_p)}{\int_{p_c}^{p_s} \tau_{\lambda_2}(p)\, dB_{\lambda_2}(T_p)} \tag{36}$$

In practice, the left-hand side is determined from observations, and the right-hand side is calculated from a known temperature profile and the profiles of atmospheric transmittance for the spectral channels as a function of p_c. The optimum cloud top pressure is determined when the difference between the right-hand and left-hand sides of the equation are a minimum.

Once the cloud height has been determined, the effective cloud amount $\eta = F\varepsilon$ is determined from a window channel observation.

$$F\varepsilon = \frac{I_w - I_w^{\mathrm{clr}}}{B_w(T_{pc}) - I_w^{\mathrm{clr}}} \tag{37}$$

The $\mathrm{CO_2}$ slicing approach is good for detecting clouds in the upper troposphere but fails for clouds in the lower troposphere.

Remote-Sensing Atmospheric Temperature Profiles

The retrieval of the atmospheric temperature and moisture profile is often accomplished using spectral observations in the infrared. The appropriate equation for the transfer of infrared radiation is

$$I_\lambda = B_\lambda(T_{\mathrm{sfc}})T_\lambda(0) + \int_{z=0}^{\infty} B_\lambda(T)\frac{dT_\lambda(z)}{dz}dz \tag{38}$$

where I_λ is the observed radiance at wavelength λ, T_{sfc} is the surface temperature, T_λ is the transmittance, and $B_\lambda(T)$ is the Planck function containing information on the atmospheric temperature. The term

$$\frac{dT_\lambda(z)}{dz} = W(z) \tag{39}$$

is referred to as the weighting function. The intensity measured by a satellite radiometer due to the emission from a layer in the atmosphere at location z, is determined from the layer blackbody emission $B_\lambda(T)$ weighted by the factor $W(z)$. The weighting function is of fundamental importance to vertical sounding the atmosphere from satellite observations. The weighting distribution depends on the strength and distribution of the absorbing gas.

To retrieve atmospheric temperature profiles, satellite radiometers make measurements in the carbon dioxide absorption bands because carbon dioxide is relatively uniformly mixed in the atmosphere, and thus the vertical distribution is known. Observations are made at spectral regions across the carbon dioxide absorption band, including weak and strong absorption regions. Figure 32 shows the weighting functions in the 12- to 15-μm spectral region of the Geostationary Orbiting Environmental Satellite (GOES) sounder radiometer. The 14.7-μm region is a strong absorption region of carbon dioxide and so the weighting function peaks in the stratosphere. The weighting function at 14.7 μm is near zero for pressures greater than 500 mb, indicating that radiance observations at this wavelength receive no contribution from the lower atmosphere and surface. The 13.4-μm spectral region is a weak absorption region and weighting function peaks at the surface, with only small

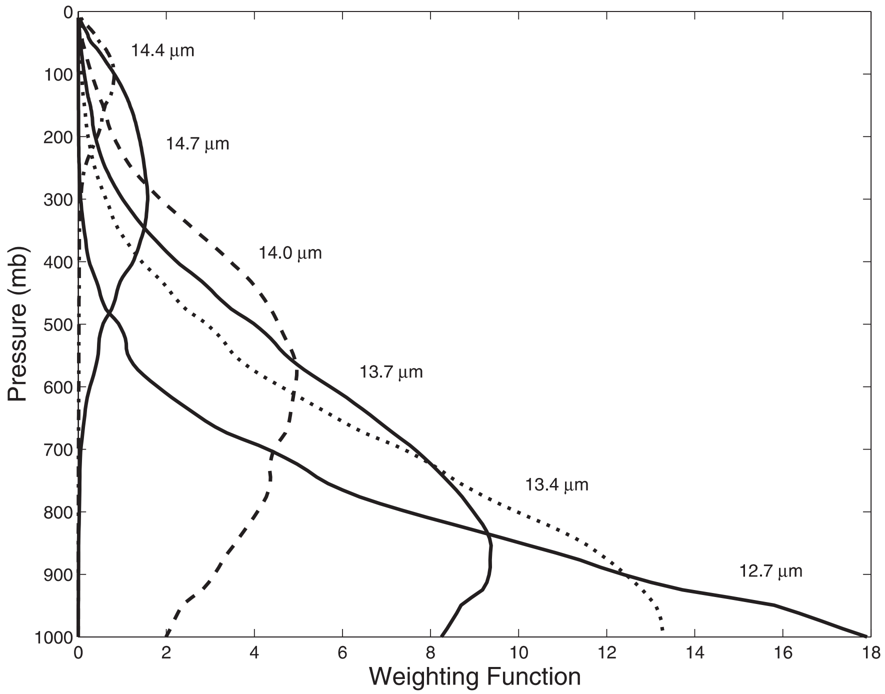


Figure 32 GOES sounder channels used in retrieval of temperature profiles.

contributions in the stratosphere. The location of peak in the weighting function indicates the region of the atmosphere that makes the largest contribution to the radiance being measured. The width of the weighting function characterizes the vertical resolution of the retrieval.

Given different spectral observations, I_λ, we wish to solve for the temperature profile, given the distribution of absorbing gas. This solution is not straightforward as the equation is nonlinear and the problem is underconstrained so that no unique solution exists. To simplify the problem we consider the discrete form of the radiative transfer equation

$$I_i = (B_0)_i T_i(z = 0) + \sum_{j=1}^{m} B_{i,j} K_{i,j} T_{i,j-1} - T_{i,j} \tag{40}$$

where i represents a spectral channel and j is the atmospheric level. Or

$$I_i = (B_0)_{iri}(z = 0) + \sum_{j=1}^{m} B_{i,j} K_{i,j} \tag{41}$$

The vector I_λ represents our spectral channel observations, K_{ij} is the discrete weighting function elements and includes the surface term, and the unknowns are the vector $B_{i,j}$. Our problem is to invert this equation to solve for $B_{i,j}$ from which the temperature profile follows. The problem is underconstrained: Given I_λ we cannot say anything about the individual values of $B_{i,j}$. We require a priori constraints, for example, by reducing the number of layers over which the profile is specified or by specifying the representation of $B_{i,j}$. The numerical methods used to retrieve temperature from the radiance measurements are described elsewhere. Retrievals often start with a first-guess profile, for example, from in situ radiosonde observations or a numerical forecast model. The first-guess profiles are then adjusted until calculated radiances match the observed radiances within some threshold. Because the observed radiances arise from deep and overlapping layers, as indicated in the weighting function, the retrieved temperature profiles are for layers of the atmosphere and do not resolve the sharp temperature gradients sometimes observed in radiosonde measurements. However, the satellite provides better spatial and temporal resolution and can be very useful for forecasting.

An example is monitoring the convective conditions of the atmosphere using the lifted index. The lifted index is the temperature difference found by subtracting the temperature of a parcel of air lifted from the surface to 500 mb from the existing

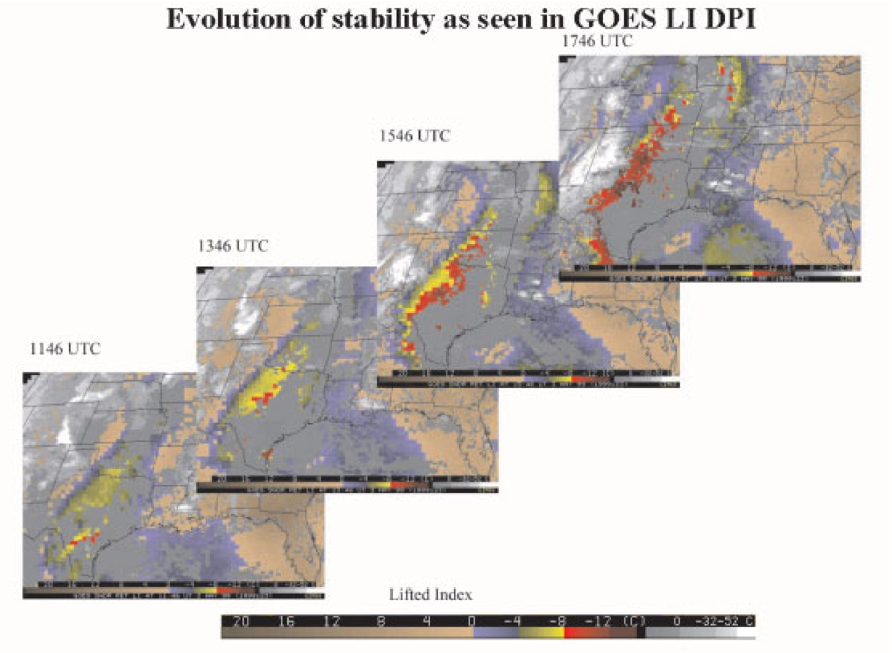

Evolution of stability as seen in GOES LI DPI
1746 UTC
1546 UTC
1346 UTC
1146 UTC
Lifted Index
20 16 12 8 4 0 -4 -8 -12 (C) 0 -32-52 C

Figure 33 (see color insert) Sequence of GOES-derived lifted index on May 3, 1999. Red regions indicate the potential areas of thunderstorm development. Strong tornadoes developed in southwest Oklahoma before 22 UTC. The lifted index versus color code is given in the legend. See ftp site for color image.

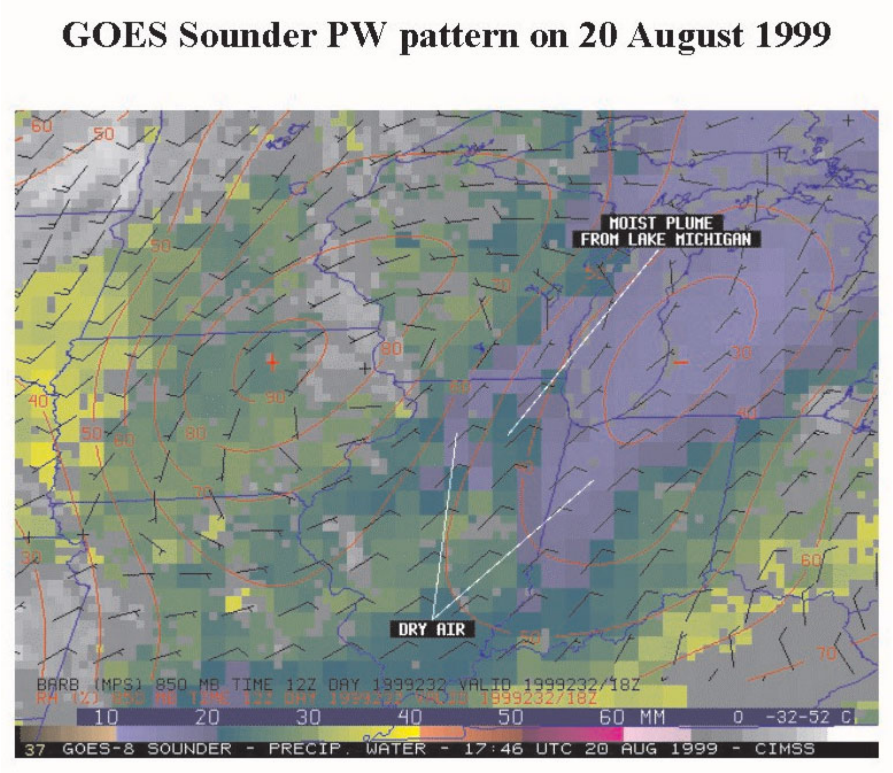

GOES Sounder PW pattern on 20 August 1999
MOIST PLUME FROM LAKE MICHIGAN
DRY AIR
BARB (MPS) 850 MB TIME 12Z DAY 1999232 VALID 1999232/18Z
RH (%) 850 MB TIME 12Z DAY 1999232 VALID 1999232/18Z
10 20 30 40 50 60 MM 0 -32-52 C
37 GOES-8 SOUNDER - PRECIP. WATER - 17:46 UTC 20 AUG 1999 - CIMSS

Figure 34 (see color insert) GOES-8-derived precipitable water on April 20, 1999. Blue regions indicate regions of low PW and green relative high amounts of PW. The satellite analysis clearly shows dry region south of Great Lakes. See ftp site for color image.

temperature at 500 mb. The lifted index numbers are related to thunderstorm severity. Values less than or equal to −6 indicates conditions are very favorable for development of thunderstorms with a high likelihood that if they occur, they would be severe with high winds and hail. The satellite-observed lifted index on May 3, 1999, over the south central United States is shown in Figure 33. The time sequence of the satellite-derived lifted index clearly shows (see the red region) the region favorable for the development of severe thunderstorms.

Remote-Sensing Atmospheric Moisture

Once the temperature profile of the atmosphere has been determined, infrared observations in water vapor absorption bands can be used to infer atmospheric water content. GOES observations at water vapor absorption bands are routinely used by the National Oceanic and Atmospheric Administration (NOAA) to derive the vertically integrated water vapor, or precipitable water (PW). An example is shown

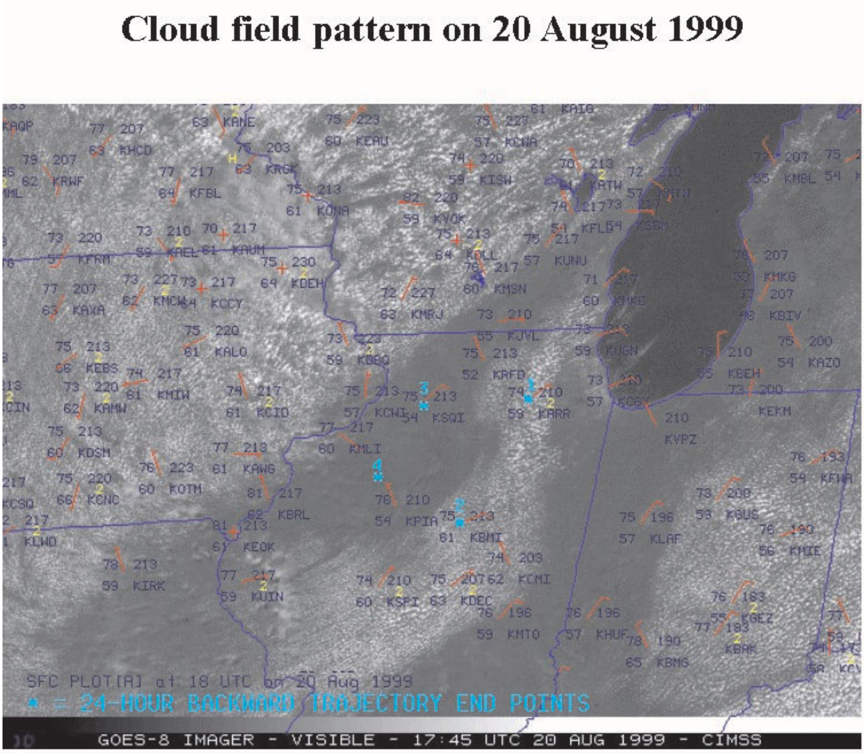

Cloud field pattern on 20 August 1999
SFC PLOT[A] at 18 UTC on 20 Aug 1999
* = 24-HOUR BACKWARD TRAJECTORY END POINTS
GOES-8 IMAGER - VISIBLE - 17:45 UTC 20 AUG 1999 - CIMSS

Figure 35 GOES-8-visible image on April 20, 1999 at 1715 UTC. Cumulus clouds are absent in the regions of low PW. See ftp site for color image.

in Figure 34. The GOES sequence of observations captures the large region of relatively dry air with less than 20 mm of PW (blue enhancement) in the western Great Lakes region on August 20, 1999. During the late morning and afternoon hours, the GOES-derived PW shows an elongated moist region with PW greater than 20 mm (green enhancement) across Illinois.

This moisture distribution impacted the cloud formation on this day. Cumulus formation was suppressed on either side of this moist plume, as seen in the GOES visible image. The GOES-derived PW remained over these cloud-free areas during the entire day (see Fig. 35).

CHAPTER 21

THE CLASSIFICATION OF CLOUDS

ARTHUR L. RANGNO

1 INTRODUCTION

Official synoptic weather observations have contained information of the coverage of various types of clouds since 1930. These cloud observations are based on a classification system that was largely in place by the late 1890s (Brooks, 1951). In recent years, these cloud observations have also had increased value. Besides their traditional role in helping to assess the current condition of the atmosphere and what weather may lie ahead, they are also helping to provide a long-term record from which changes in cloud coverage and type associated with climate change might be discerned that might not be detectable in the relatively short record of satellite data (Warren et al., 1991). This article discusses what a cloud is, the origin of the classification system of clouds, and contains photographs of the most commonly observed clouds.

2 WHAT IS A CLOUD?

As defined by the World Meteorological Organization (1969), a cloud is an aggregate of minute suspended particles of water or ice, or both, that are in sufficient concentrations to be visible—a collection of "hydrometeors," a term that also includes in some cases, due to perspective, the precipitation particles that fall from them.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

Clouds are tenuous and transitory. No single cloud element, even within an extensive cloud shield, exists for more than a few hours, and most small clouds in the lower atmosphere exist for only a few minutes. In precise numbers, the demarcation between a cloud and clear air is hard to define. How many cloud drops per liter constitute a cloud? When are ice crystals and snow termed "clouds" rather than precipitation? When are drops or ice crystals too large to be considered "cloud" particles, but rather "precipitation" particles?

These questions are difficult for scientists to answer in unanimity because the difference between cloud particles and precipitation particles, for example, is not black and white; rather they represent a continuum of fallspeeds. For some scientists, a 50-μm diameter drop represents a "drizzle" drop because it likely has formed from collisions with other drops, but for others it may be termed a "cloud" drop because it falls too slowly to produce noticeable precipitation, and evaporates almost immediately after exiting the bottom of the cloud. Also, the farther an observer is from falling precipitation, the more it appears to be a "cloud" as a result of perspective. For example, many of the higher "clouds" above us, such as cirrus and altostratus clouds, are composed mainly of ice crystals and even snowflakes that are settling toward the Earth; they would not be considered a "cloud" by an observer inside them on Mt. Everest, for example, but rather a very light snowfall. Some of the ambiguities and problems associated with cloud classification by ground observers were discussed by Howell (1951).

3 ORIGIN OF THE PRESENT-DAY CLOUD CLASSIFICATION SYSTEM

The classification system for clouds is based on what we see above us. At about the same time at the turn of the 19th century, the process of classifying objectively the many shapes and sizes of something as ephemeral as a cloud was first accomplished by an English chemist, Luke Howard, in 1803, and a French naturalist, Jean Baptiste Lamarck, in 1802 (Hamblyn, 2001). Both published systems of cloud classifications. However, because Howard used Latin descriptors of the type that scientists were already using in other fields, his descriptions appeared to resemble much of what people saw, and because he published his results in a relatively well-read journal, *Tilloch's Philosophical Magazine*, Howard's system became accepted and was reproduced in books and encyclopedias soon afterward (Howard, 1803).

Howard observed, as had Lamarck before him, that there were three basic cloud regimes. There were fibrous and wispy clouds, which he called *cirrus* (Latin for hair), sheet-like laminar clouds that covered much or all of the sky, which he referred to as *stratus* (meaning flat), and clouds that were less pervasive but had a strong vertical architecture, which he called *cumulus* (meaning heaped up). Howard used an additional Latin term, *nimbus* (Latin for cloud), meaning in this case, a cloud or system of clouds from which precipitation fell. Today, *nimbus* itself is not a cloud, but rather a prefix or suffix to denote the two main precipitating clouds, *nimbostratus* and *cumulonimbus*. The question over clouds and their types generated such enthusiasm among naturalists in the 19th century that an ardent observer and member of the British Royal Meteorological Society, Ralph Abercromby, took two voyages around the world to make sure that no cloud type had been overlooked!

The emerging idea that clouds preferred just two or three levels in the atmosphere was supported by measurements using theodolites and photogrammetry to measure cloud height at Uppsala, Sweden, as well as at sites in Germany and in the United States in the 1880s. These measurements eventually led H. Hildebrandsson, Director of the Uppsala Observatory, and Abercromby to place the "low," "middle," and "high" cloud groupings of Howard more systematically in their own 1887 cloud classification. At this time, *cumulus* and *cumulonimbus* clouds were placed in a fourth distinct category representing clouds with appreciable updrafts and vertical development.

Howard's modified classification system was re-examined at the International Meteorological Conference at Munich in 1891 followed by the publication of a color cloud atlas in 1896 (Hildebrandsson et al., 1896). At this point, the definitions of clouds were close to their modern forms. Additional international committees made minor modifications to this system in 1926 that were realized with the publication of the 1932 International Cloud Atlas. Little change has been made since that time.

The most comprehensive version of the classification system was published in two volumes (International Cloud Atlas) by the World Meteorological Organization in 1956 (WMO, 1956). Volume I contained the cloud morphology and Volume II consisted of photographs. An abridged Atlas published in 1969 consisted of combined morphology and photographs. The descriptions of clouds and their classifications (Volume I) were published again in 1975 by the WMO (WMO, 1975). In 1987, a revised Volume II of photographs (WMO, 1975) was published that included photographs of clouds from more disparate places than in the previous volumes.

4 THE CLASSIFICATION OF CLOUDS

There are ten main categories or genera into which clouds are classified for official observations (e.g., British Meteorological Office, 1982; Houze, 1993): *cirrus*, *cirrostratus*, *cirrocumulus*, *altostratus*, *altocumulus*, *nimbostratus*, *stratocumulus*, *stratus*, *cumulus*, and *cumulonimbus*. Table 1 is a partial list of the nomenclature used to describe the most commonly seen species and varieties of these clouds. Figures 1 to 13 illustrate these main forms and their most common varieties or species.

Within these ten categories are three cloud base altitude regimes: "high" clouds, those with bases generally above 7 km above ground level (AGL); "middle-level" clouds, those with bases between 2 and about 7 km AGL; and "low" clouds, those with bases at or below 2 km AGL. The word "about" is used because clouds with certain visual attributes that make them, for example, a middle-level cloud, may actually have a base that is above 7 km. Similarly, in winter or in the Arctic, high clouds with cirriform attributes (fibrous and wispy) may be found at heights below 7 km. Also, some clouds that are still considered low clouds (e.g., *cumulus* clouds) can have bases that are a km or more above the general low cloud upper base limit of

TABLE 1 The Ten Cloud Types and Their Most Common Species and Varieties.

Genera	Species	Varieties
Cirrus (Ci)	Uncinus, fibratus, spissatus, castellanus	Intortus, radiatus, vertebratus
Cirrostratus (Cs)	Nebulosus, fibratus	
Cirrocumulus (Cc)	Castellanus, floccus lenticularis	Undulatus
Altocumulus (Ac)	Castellanus, floccus, lenticularis	Translucidus, opacus, undulatus, perlucidus
Altostratus (As)	None	Translucidus, opacus
Nimbostratus (Ns)	None	None
Stratocumulus (Sc)	Castellanus, lenticularis	Perlucidus, translucidus opacus
Stratus (St)	Fractus, nebulosus	
Cumulonimbus (Cb)	Calvus, capillatus	
Cumulus (Cu)	Fractus, humilis, mediocris, congestus	

Letters in parentheses denote accepted abbreviations.
Source: From World Meteorological Organization, 1975.

2 km AGL. Therefore, these cloud base height boundaries should be considered somewhat flexible. Note, too, that what is classified as an *altocumulus* layer when seen from sea level will be termed a *stratocumulus* layer when the same cloud is seen by an observer at the top of a high mountain because the apparent size of the cloud elements, part of the definition of these clouds, becomes larger the nearer one is to the cloud layer. The definitions are made on the basis of the distance of the observer from the cloud.

The classification of clouds is also dependent on their composition. This is because the composition of a cloud, all liquid, all ice, or a mixture of both, determines many of its visual attributes on which the classifications are founded (e.g., luminance, texture, color, opacity, and the level of detail of the cloud elements). For example, an *altocumulus* cloud cannot contain too many ice crystals and still be recognizable as an *altocumulus* cloud. It must always be composed largely of water drops to retain its sharp-edged compact appearance. Thus, it cannot be too high and cold. On the other hand, wispy trails of ice crystals comprising *cirrus* clouds cannot be too low (and thus, too warm). Therefore, having the ability to assess the composition of clouds (i.e., ice vs. liquid water) visually can help in the determination of a cloud's height.

Other important attributes for identifying a cloud are: How much of the sky does it cover? Does it obscure the sun's disk? If the sun's position is visible, is its disk sharply defined or diffuse? Does the cloud display a particular pattern such as small cloud elements, rows, billows, or undulations? Is rain or snow falling from it? If so, is the rain or snow falling from it concentrated in a narrow shaft, suggesting heaped cloud tops above, or is the precipitation widespread with little gradation, a characteristic that suggests uniform cloud tops? Answering these questions will allow the best categorization of clouds into their ten basic types.

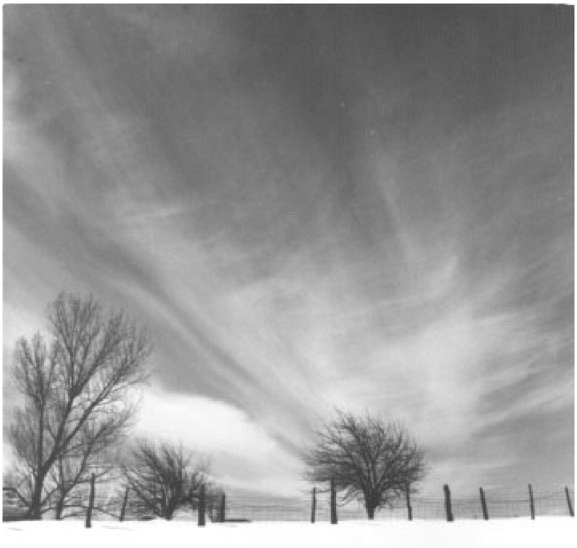


Figure 1 Cirrus (*fibratus*) with a small patch of *cirrus spissatus* left between the trees.

4.1 High Clouds

Cirrus, *cirrostratus*, and *cirrocumulus* clouds (Figs. 1, 2, and 3, respectively) comprise high clouds. By WMO definition, they are not dense enough to produce shading except when the sun is near the horizon, with the single exception of a thick patchy *cirrus* species called *cirrus spissatus* in which gray shading is allowable.[1] *Cirrus* and *cirrostratus* clouds are composed of ice crystals with, perhaps, a few momentary exceptions just after forming and when the temperature is higher than −40°C. In this case, droplets may be briefly present at the instant of formation. The bases of *cirrus* and *cirrostratus* clouds, composed of generally low concentrations of ice crystals that are about to evaporate, are usually colder than −20°C. The coldest cirriform clouds (i.e., *cirrus* and *cirrostratus*) can be −80° C or lower in deep storms with high cloud tops ($>$ 15 km above sea level) such as in thinning anvils associated with high-level outflow of thunderstorms.

Because of their icy composition, *cirrus* and *cirrostratus* clouds are fibrous, wispy, and diffuse. This wispy and diffuse attribute is because the ice crystals that comprise them are overall in much lower concentrations (often an order of magnitude or more) than are the droplet concentrations in liquid water clouds. In contrast, droplet clouds look hard and sharp-edged with the details of the tiniest elements clearly visible. The long filaments that often comprise cirriform clouds are caused by

[1] Many users of satellite data refer to *cirrus* or *cirriform* those clouds with cold tops in the upper troposphere without regard to whether they produce shading as seen from below. However, many such clouds so described would actually be classified as *altostratus* clouds by ground observers. This is because such clouds are usually thick enough to produce gray shading and cannot, therefore, be technically classified as a form of cirrus clouds from the ground.

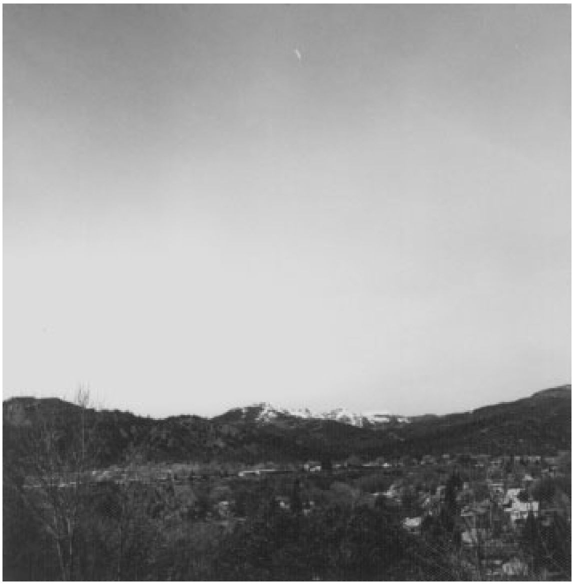


Figure 2 Cirrostratus (*nebulosus*), a relatively featureless ice cloud that may merely turn the sky from blue to white.

the growth of larger ice crystals that fall out in narrow, sloping shafts in the presence of changing wind speeds and directions below the parent cloud. As a result of the flow settling of ice crystals soon after they form, mature *cirrus* and *cirrostratus* clouds are, surprisingly, often 1 km or more thick, although the sun may not be appreciably dimmed (e.g., Planck et al., 1955; Heymsfield, 1993).

Cirrus and *cirrostratus* clouds often produce haloes when viewed from the ground whereas thicker, mainly ice clouds, such as altostratus clouds (see below) cannot. This is because the cirriform clouds usually consist of small, hexagonal prism crystals such as thick plates, short solid columns—simple crystals that refract the sun's light as it passes through them. On the other hand, *altostratus* clouds are much deeper and are therefore composed of much larger, complicated ice crystals and snowflakes that do not permit simple refraction of the sun's light, or, if a halo exists in its upper regions, it cannot be seen due to the opacity of the *altostratus* cloud. The appearance of a widespread sheet of cirrostratus clouds in wintertime in the middle and northern latitudes has long been identified as a precursor to steady rain or snow (e.g., Brooks, 1951).

Cirrocumulus clouds are patchy, finely granulated clouds. Owing to a definition that allows no shading, they are very thin (less than 200 m thick), and usually very short-lived, often appearing and disappearing in minutes. The largest of the visible cloud elements can be no larger than the width of a finger held skyward when observed from the ground; otherwise it is classified as an *altocumulus* or, if lower, a *stratocumulus* cloud.

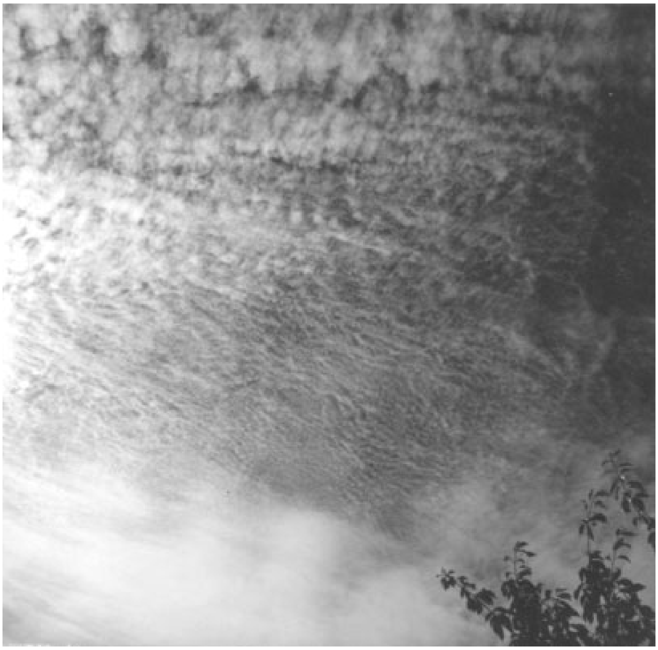


Figure 3 Cirrocumulus: the tiniest granules of this cloud can offer the illusion of looking at the cloud streets on a satellite photo.

Cirrocumulus clouds are composed mostly or completely of water droplets. The liquid phase of these clouds can usually be deduced when they are near the sun; a corona or irisation (also called iridescence) is produced due to the diffraction of sunlight by the cloud's tiny (< 10-μm diameter) droplets. However, many *cirrocumulus* clouds that form at low temperatures (< -30°C) migrate to fibrous *cirriform* clouds within a few minutes, causing the granulated appearance to disappear as the droplets evaporate or freeze to become longer-lived ice crystals that fall out and can spread away from the original tiny cloudlet.

4.2 Middle-Level Clouds

Altocumulus, *altostratus*, and *nimbostratus* clouds (Figs. 4, 5, 6, and 7, respectively) are considered middle-level clouds because their bases are located between about 2 and 7 km AGL (see discussion concerning the variable bases of *nimbostratus* clouds below). These clouds are the product of slow updrafts (centimeters per second) often taking place in the middle troposphere over an area of thousands of square kilometers or more. Gray shading is expected in *altostratus*, and is generally present in *altocumulus* clouds. *Nimbostratus* clouds by definition are dark gray and the sun cannot be seen through them. It is this property of shading that differentiates these clouds from high clouds, which as a rule can have no shading.

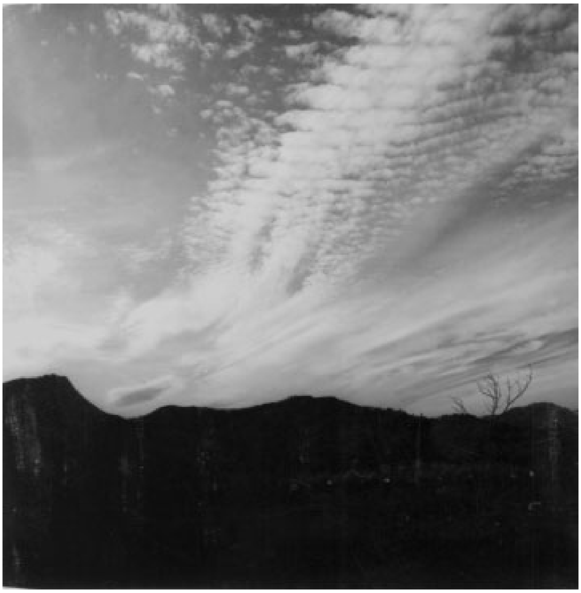


Figure 4 Altocumulus translucidus is racing toward the observed with an altocumulus lenticularis cloud in the distance above the horizon. Cirrostratus nebulosus is also present above the altocumulus clouds.

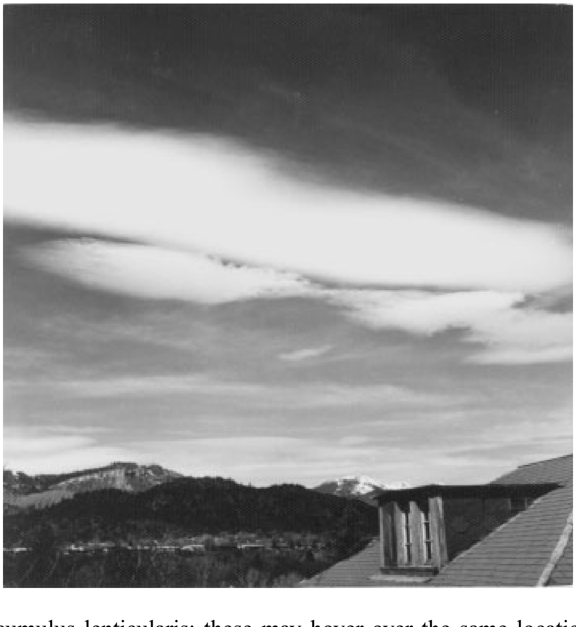


Figure 5 Altocumulus lenticularis: these may hover over the same location for hours at a time or for just a few minutes. They expand and shrink as the grade of moisture in the lifted air stream waxes and wanes.

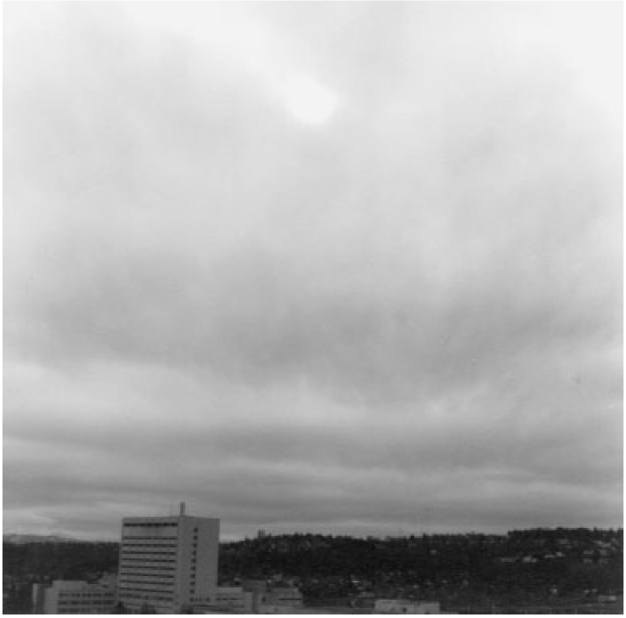


Figure 6 Altostratus translucidus overspreads the sky as a warm-frontal system approaches Seattle. The darker regions show where the snow falling from this layer (virga) has reached levels that are somewhat lower than from regions nearby.

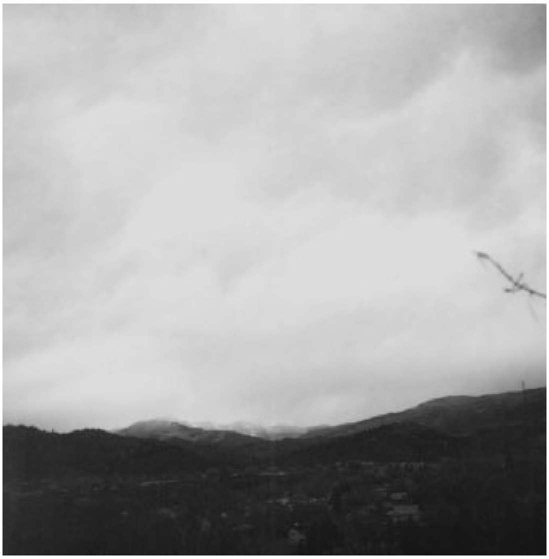


Figure 7 Nimbostratus on a day with freezing level at the ground. Snowfall is beginning to obscure the mountain peaks nearby and snow reached the lower valley in the foreground only minutes later.

Altostratus and *altocumulus* are different from one another in the same way that *cirrus* and *cirrostratus* clouds are different from *cirrocumulus* clouds; in *altocumulus* clouds, droplets predominate, giving them a crisp, sharper-edged look. With *altostratus* clouds, ice crystals and snowflakes dominate or comprise the entire cloud, giving it a diffuse, fibrous look. The observation of *altocumulus* and *altostratus* clouds in the sky has long been a marker for deteriorating weather in the hours ahead.

In spite of its name, *altocumulus* clouds are generally rather flat clouds that strongly resemble *stratocumulus* clouds. An exception to this overall laminar architecture is in those species of *altocumulus* called *castellanus* and *floccus*. In these forms, *altocumulus* clouds resemble miniature, lofted *cumulus* clouds that usually occur in rows or patches rather than in widespread layers.

Altocumulus clouds are distinguished from *cirrocumulus* because they are lower and the cloud elements in *altocumulus* are, or appear to be from the ground, several times larger than those in *cirrocumulus* clouds. For example, the elements of an *altocumulus* cloud are typically the width of three fingers held skyward at the ground. Also, shading toward the center of the thicker elements is usually present in *altocumulus* clouds, a property that is not allowed in the classification of *cirrocumulus* clouds. *Altocumulus* clouds are distinguished from *stratocumulus* because they are higher above ground level than *stratocumulus* (at least 2 km) and because the individual cloud elements in *altocumulus* are, or appear to be from the ground, smaller than those in *stratocumulus*.

In spite of the gray shading that may be present, *altocumulus* clouds are rarely more than 1 km thick. This is because the concentrations of drops in them are relatively high (typically 50,000 or more per liter) compared with fibrous ice clouds whose particle concentrations may only be only tens to a few thousand per liter. This density of particles produces an optical depth in which the sun's disk can be obscured (optical depth of 4 or more) by an *altocumulus* cloud layer only 300–500 m thick.

Altocumulus clouds sometimes sport patchy "virga." *Virga* is light precipitation that falls from a cloud but does not reach the ground because it evaporates. Because virga is almost always caused by falling snow, it appears fibrous, often with striations or long filaments that often far surpass the depth of the cloud from which it is falling. Virga, because it is comprised of falling snow, can appear to be quite dense. *Altocumulus* clouds with virga are predominantly those clouds whose temperatures are lower than -10° C (Heymsfield, 1993). However, at the same time, they are rarely colder than about -30° C. This is because at very low temperatures they are likely to take on the attributes of ice clouds such as *cirrus* or its thicker brethren, *altostratus*.

The species of *altocumulus* clouds called *altocumulus castellanus* has always had a special significance in meteorology because these clouds reveal a conditionally unstable lapse rate in the middle troposphere. Instability of this sort has been viewed as a marker for likely releases of deeper convection in the hours ahead. Occasionally, *castellanus* clouds group and enlarge into higher based *cumulus* and *cumulonimbus* clouds.

When the winds are relatively strong aloft (greater than about $20\,\mathrm{m\,s^{-1}}$) and moderately moist, but stable lapse rate conditions are present, a species of *altocumulus* called *lenticularis* (lens or almond-shaped) clouds (Fig. 5) may form over or downwind of mountains. *Altocumulus lenticularis* clouds can hover over the same location for minutes to hours while expanding and shrinking in response to fluctuations in the relative humidity of the air mass being lifted over the terrain. Because the conditions under which these clouds form are most often associated with advancing short wave troughs in the middle and upper atmosphere and their accompanying regions of low pressure, *lenticularis* clouds are usually precursors to deteriorating weather.

With *altostratus* clouds (Fig. 6), the dominance of ice causes a diffuse, amorphous appearance with striations or fallstreaks (virga) on the bottom; an observer is usually viewing relatively low concentrations of precipitation particles rather than a cloud per se. *Altostratus* clouds are rarely less than 2 km thick and often have tops at the same heights as *cirrus* and *cirrostratus* clouds. Because of this great altitude range, they are considerably colder and span a much greater temperature range than do *altocumulus* clouds. They also, by definition, cover the entire sky or at least wide portions of it; they are not patchy clouds. Precipitation is usually imminent when *altostratus* clouds are moving in because they are a sign that a widespread slab of air is being lifted, usually that due to an approaching cyclone and its frontal system (Brooks, 1951).

The relatively low concentrations of large particles in some *altostratus* clouds (often tens per liter) can allow the sun's position to be seen as though looking through ground or fogged glass: the sun's position is apparent, but the outline of its disk is not. This may be the case (*translucidus* variety) even when the cloud layer is two or more kilometers thick.

Somewhat surprisingly, when the top of an *altostratus* cloud layer is warmer than about -30 to $-35°\mathrm{C}$, it is not uncommon to find that the top is comprised of a thin droplet cloud virtually identical to an *altocumulus* cloud layer that is producing virga. The survival, growth, aggregation, and breakup of ice crystals spawned by these cold, liquid clouds over a great depth usually obscures the ice-producing droplet cloud top in these cases. These kinds of situations were dubbed "the upside-down storm" when first noticed in the mid-1950s because the coldest part of the cloud (the top) was liquid and the warmer regions below were comprised of ice (Cunningham, 1957).

Optical phenomena seen from the ground with *altostratus* clouds are limited to *parhelia* ("sun dogs"). They are usually observed in thin portions of *altostratus* when the sun is low in the sky. Parhelia are bright, colored highlights, which sometimes rival the brightness of the sun, that are located 22° from the sun's position. They are most noticeable in the morning or before sunset.

However, since the composition of the uppermost regions of the deepest *altostratus* clouds are virtually identical to *cirriform* ice clouds with simpler, smaller ice crystals, haloes are often observed in the uppermost portions of deep *altostratus* clouds.

Nimbostratus clouds (Fig. 7) are virtually identical to *altostratus* clouds in their composition except that their bases are usually perceived from the ground as lower than in *altostratus* (from which it has usually derived due to a downward thickening). Therefore, they often appear somewhat darker than *altostratus* clouds and, by definition, do not allow the sun to be seen through them. The perceived base of *nimbostratus* is caused by snowflakes that are melting into raindrops. This apparent base of the cloud is a result of the greater opacity of snow particles giving the impression of a bottom or sharp increase in thickness of the cloud at the melting level. Thus, the base of an amorphous, precipitating *nimbostratus* cloud might be perceived at mid-levels on a day when the freezing level is high (above 2 km) such as in southern latitudes or the tropics, or be perceived as low when the freezing level is low as in northern latitudes in the winter. Generally, the bottom of *nimbostratus* clouds is obscured by lower detached clouds such as *stratus fractus*, or *stratocumulus*.

Nimbostratus clouds produce relatively steady precipitation that often continues for hours at a time. They are not clouds responsible for passing showers with periods of sun in between. The tops of dark, steadily precipitating *nimbostratus* clouds can be as shallow as 2–3 km and even be above freezing in temperature, or they may reach into the upper regions of the troposphere (to *cirriform* cloud levels) and be as cold as −80°C. At elevations above the freezing level *nimbostratus* is largely composed of ice crystals and snowflakes, though embedded thin (supercoooled) droplet cloud layers similar to *altocumulus* clouds are relatively common. Also, similar to *altostratus* clouds, when the temperature at the top of *nimbostratus* clouds is above about −30° to −35°C, a thin droplet cloud layer may be found in which the first ice crystals form.

A broken-to-overcast layer of shallow *stratus* or *stratocumulus* clouds often resides at the bottom of *nimbostratus* clouds. However, while usually not precipitating themselves, these lower cloud layers are important in enhancing the amount of rain or snow that falls from *nimbostratus* clouds. This enhancement occurs because of the accretion or riming of the relatively small cloud drops in the lower clouds by the rapidly falling precipitation-sized particles. This enhancement is especially evident in hilly or mountainous regions. However, just the existence of lower clouds even with drops too small to be collected by the faster falling precipitation indicates enhanced precipitation at those locations compared with cloud-free locations. This is because where there are no lower clouds, the precipitation is likely to be subject to a degree of evaporation, and the drops or snowflakes are slightly smaller in comparison to those locations where clouds exist below the *nimbostratus* and there is no evaporation through the depth of the cloud.

Cumulonimbus clouds (see below) may also be embedded in *nimbostratus* clouds. The presence of such clouds within *nimbostratus* is evident by sudden gushes of much heavier rain and sometimes lightning within a context of relatively steady rain.

4.3 Low Stratiform Clouds

Stratocumulus and *stratus* clouds (Figs. 8 and 9, respectively) are low-based, shallow stratiform clouds, almost always less than 1 km thick. They are composed of droplets unless the cloud top is cooler than about -5 to -10° C in which case ice crystals may form (Hobbs and Rangno, 1985). *Stratocumulus* clouds differ from *stratus* clouds because they have a rather lumpy appearance at cloud base with darker and lighter regions due to embedded weak convection. Also, their bases tend to be higher, and more irregular in height than those of *stratus* clouds. *Stratus* presents a smoother, lower, more uniform sky than do *stratocumulus* clouds because the internal convective overturning that produces lighter and darker regions is slight. Drizzle (precipitation comprised of drops < 500-μm diameter that nearly float in the air) is likely to form when the cloud droplet concentrations are lower than about $100\,\text{cm}^{-3}$. Therefore, drizzle is common from both *stratus* and *stratocumulus* clouds at sea and along and inland of coastlines in onshore flow because in such situations the air is clean and there are relatively few cloud condensation nuclei on which droplets can form. However, recent measurements have also shown that drizzle and light rain producing shallow clouds with low droplet concentrations are much more common at inland locations in winter than was once believed (e.g. Huffman and Norman, 1988). Since such clouds are often supercooled in winter, they pose a severe potential for aircraft icing and freezing rain or drizzle at the ground.

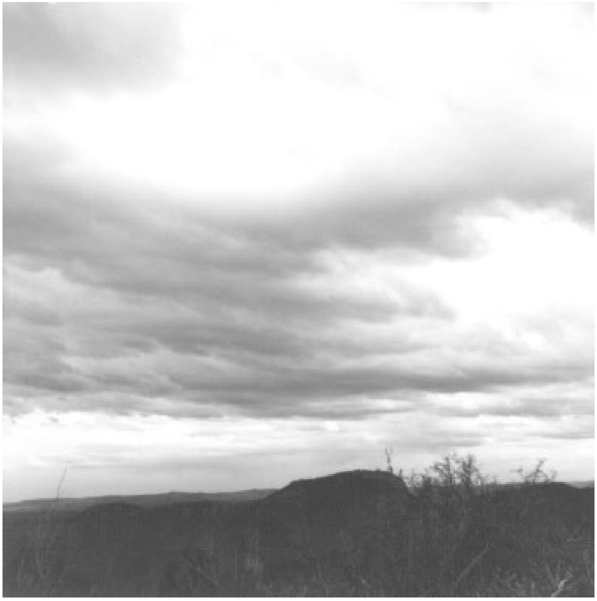


Figure 8 Stratocumulus: note the uneven bases of this cloud layer. The darker regions show where there is enhanced modest convection and higher liquid water content.

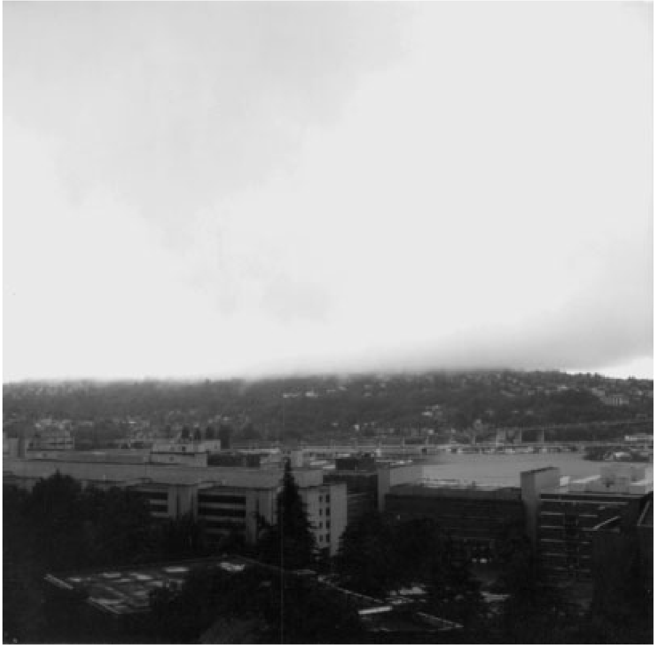


Figure 9 Stratus: the clouds intercept hills only a few hundred meters above sea level. Note here the uniformity of this cloud layer compared with *stratocumulus* in Fig. 8.

4.4 Convective Clouds

Cumulus and *cumulonimbus* clouds (Figs. 10, 11, 12, and 13, respectively) are convective clouds created when the temperature decreases rather rapidly with increasing height. Differential heating and converging air currents in this circumstance can therefore send plumes of warmer air skyward with relative ease. Convective clouds are limited in coverage compared with stratiform clouds and, except for the anvil portions of *cumulonimbus* clouds, rarely cover the entire sky or do so only for short periods. This coverage characteristic differentiates, for example, *stratocumulus* clouds with their linked cloud bases covering large portions of the sky, and similar-sized *cumulus* clouds that by definition must be relatively scattered into isolated clouds or small clusters with large sky openings.

Cumulus clouds have a size spectrum of their own that ranges from *cumulus fractus*, those first cloud shreds that appear at the top of the convective boundary layer, to *congestus* size (more than about 2 km deep). Between these sizes are *cumulus humilis* and *cumulus mediocris* clouds, clouds that range between about 1 and 2 km in depth, respectively. The tops of these larger clouds are marked by sprouting portions called turrets that represent the growing and usually warmer parts of the cloud. Individual turrets are generally one to a few kilometers wide, although in strong storms individual turrets may coalesce into groups of many turrets to form

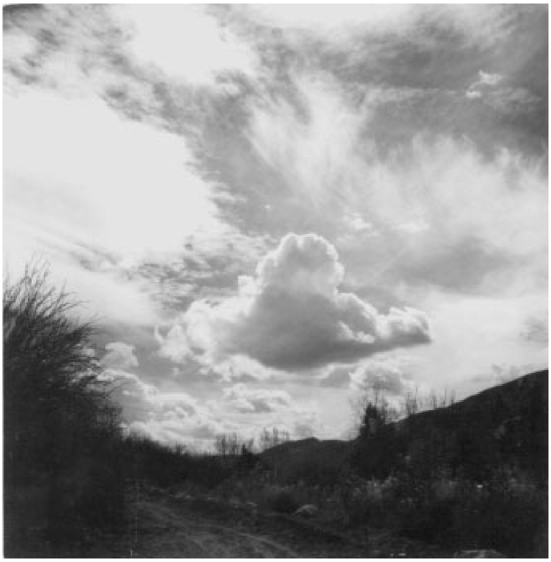


Figure 10 In this busy sky, *cumulus mediocris* is at the center, with *cumulus humilis* in the distance, and *cumulus congestus* on the horizon at left. Also present in this photo is *altocumulus translucidus* (left center), *cirrocumulus* (top center), and *cirrus uncinus* (center above *cumulus* and right center).

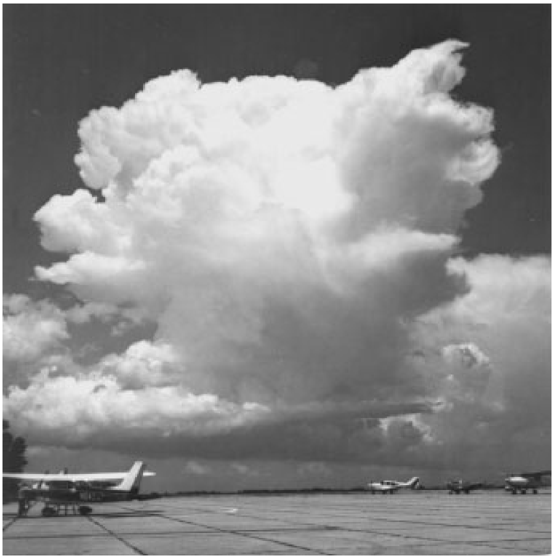


Figure 11 Cumulonimbus calvus: this rapidly changing and expanding *cumulonimbus* cloud is in the short-lived calvus stage where fibrousness of the top is just beginning to be apparent. No strong rainshaft has yet appeared below base, though the careful eye can see that considerable ice has already formed aloft (right half of cloud). An intense rainshaft emerged below cloud base a few minutes after this photograph was taken.

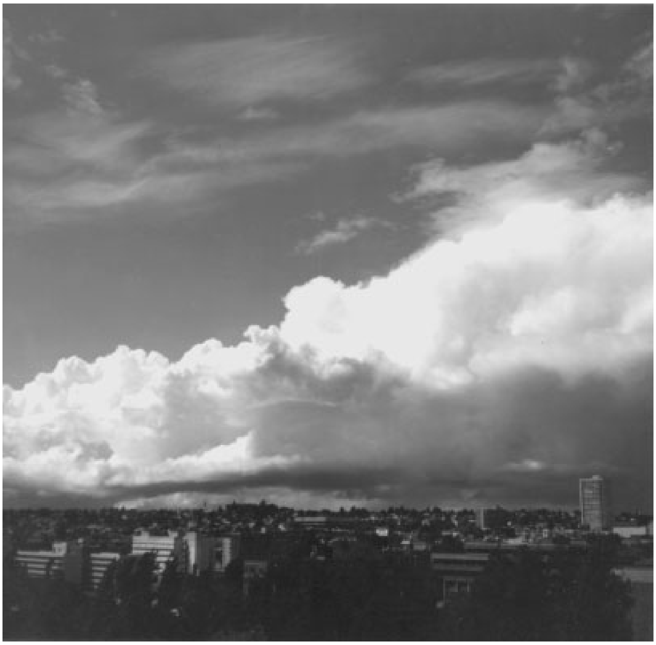


Figure 12 Cumulus congestus, *cumulonimbus calvus*, and *cumulonimbus capillatus*. In this line of convection north of Seattle, WA, (nonprecipitating) *cumulus congestus* clouds comprise the left side of the line, while in the center a taller single turret has emerged and reached the *cumulonimbus calvus* stage. The right third of the photograph shows a *cumulonimbus capillatus*, the stage that follows the calvus stage. In the capillatus stage, in this instance consisting of a conglomerate of turrets, the tops have lost their compact cumuliform look and are clearly fibrous and icy. Strands of precipitation are falling below the bases of both types of *cumulonimbus* clouds.

a large, tightly packed, and hard-appearing cauliflower mass that roils upward with little turret differentiation.

Prior to reaching the *cumulonimbus* stage, *cumulus* clouds are therefore composed of droplets and contain very few precipitation-sized particles. Precipitation, however, usually begins to develop in *cumulus congestus* clouds if they are more than about 3 km thick over land and about 1.5 to 2 km thick over the oceans (Ludlum, 1980; Wallace and Hobbs, 1977). The precipitation that falls may be caused by collisions with coalescence of cloud drops in the upper portions of the cloud or it may be a result of the formation of ice particles in clouds with cooler bases. However, in the winter, even small *cumulus* clouds with tops colder than about −10 to −15° C can produce virga, snow flurries, or even accumulating amounts of snow. These kinds of small, cold-based, and precipitating *cumulus* clouds are found in winter in such locations as the Great Lakes of the United States, off the east coasts of the continents, and over high mountains or desert regions (Rangno and Hobbs, 1994).

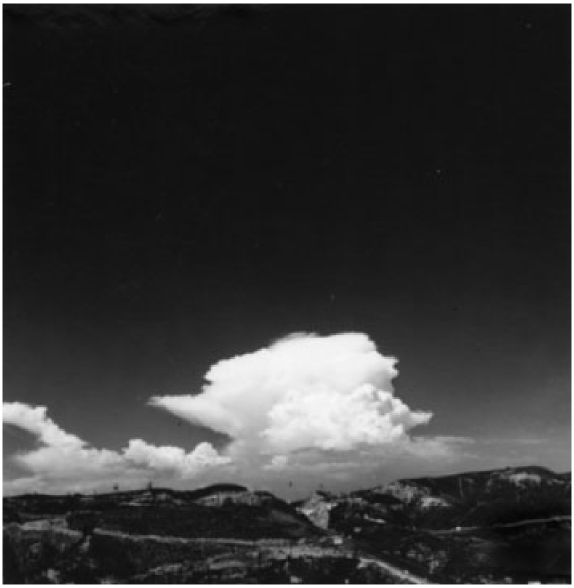


Figure 13 Cumulonimbus capillatus incus: this well-known species of *cumulonimbus* cloud also has a fibrous and icy top, but it is marked by a noticeable flattening there. Incus translates to "anvil." These types of *cumulonimbus* clouds suggest that the convection has spanned the entire troposphere.

If significant precipitation begins to develop in deep *cumulus* clouds, they quickly take on the visual attributes of *cumulonimbus* clouds (Figs. 11–13)—a strong precipitation shaft is seen below cloud base with a cloud top that is soft, fibrous, fraying, or wispy. The visual transition to a softer, fibrous appearance in the upper portion of *cumulus* clouds is caused by the lowering of the concentrations of the particles from hundreds of thousands per liter of relatively small cloud droplets (< 50-μm diameter), to only tens to hundreds per liter of much larger (millimeter-sized) precipitation-sized particles (rain drops or ice particles). These larger particles tend to fall in filaments and help produce a striated appearance.

In the period while this transformation is taking place and the fibrousness is just becoming visually apparent in the upper portions of a *cumulus congestus* cloud, the cloud is entering a short-lived period of its lifecycle when it is referred to as a *cumulonimbus calvus* ("bald") cloud. At this same time, a concentrated precipitation shaft may not be present yet or is just emerging below the cloud base (Fig. 11).

When the fibrousness of the upper portion of the cloud is fully apparent, the *cumulonimbus* cloud has transitioned to a *cumulonimbus capillatus* (hair) in which most or all of its upper portion consists of ice crystals and snowflakes (Figs. 12 and 13). In the tropics or in warm humid air masses, this visual transformation also occurs but can be due solely to the evaporation of the smaller drops leaving drizzle and raindrops rather than ice and snow in smaller *cumulonimbus* clouds. *Cumulo-*

nimbus capillatus clouds span a wide range of depths, from miniature versions only about 2 km deep in polar air masses over the oceans, to as much as 20 km in the most severe thunderstorms in equatorial regions, eastern China in the summer, and the plains and southeast regions of the United States. If a pronounced flattening of the top develops into a spreading anvil then the cloud has achieved the status of a *cumulonimbus capillatus incus* (incus meaning "anvil").

Hail or graupel (soft hail) are usually found, if not at the ground, then aloft in virtually all *cumulonimbus* clouds that reach above freezing level. Updrafts may reach tens of meters per second in *cumulus* and *cumulonimbus* clouds, particularly in warm air masses. These updrafts lead to large amounts of condensation and liquid water content. Depending on how warm the cloud base is, the middle and upper building portions of deep *cumulus* clouds might contain 1 to 5 $\mathrm{g\,m^{-3}}$ of condensed water in the form of cloud droplets and raindrops. Supercooled water concentrations of these magnitudes are sufficient to cause a buildup of about 1 cm or more of ice buildup on an airframe for every one to two minutes in cloud. Therefore, they are avoided by aircraft. *Cumulonimbus* clouds are the only clouds, by definition, that produce lightning. If lightning is observed, the cloud type producing it is automatically designated a *cumulonimbus*.

REFERENCES

British Meteorological Office (1982). *Cloud Types for Observers*, Met. O.716. London: Her Majesty's Stationery Office, 38 pp.

Brooks, C. F. (1951). The use of clouds in forecasting. In *Compendium of Meteorology*, Boston: American Meteorological Society, 1167–1178.

Cunningham, R. M. (1957). A discussion of generating cell observations with respect to the existence of freezing or sublimation nuclei. In *Artificial Stimulation of Rain*, Ed. H. Weickmann, New York: Pergamon Press.

Hamblyn, R. (2001). *The Invention of Clouds*, New York: Farrar, Straus and Giroux, 244 pp.

Heymsfield, A. J. (1993). Microphysical structures of stratiform and cirrus clouds. In *Aerosol–Cloud–Climate Interactions*, New York: Academic Press, 97–121.

Hildebrandsson, H., Riggenbach, A., and L. Teeisserenc de Bort, (1896). *Atlas International des Nuages*, Comité Météorologique International.

Hobbs, P. V., and A. L. Rangno, (1985). Ice particle concentrations in clouds, *J. Atmos. Sci.*, **42**, 2523–2549.

Houze, R. A., Jr. (1993). *Cloud Dynamics*, New York: Academic Press, 573 pp.

Howard, L. (1803). On the modifications of clouds, and on the principles of their production, suspension, and destruction. *Phil. Mag.*, **16**, 97–107; 344–357.

Howell, W. E. (1951). The classification of cloud forms. In *Compendium of Meteorology*, Boston: American Meteorological Society, 1161–1166.

Huffman, G. J., and G. A. Norman, Jr., 1988. The supercooled warm rain process and the specification of freezing precipitation. Monthly Wea. Rev., *116*, 2172–2182.

Ludlum, F. H. (1980). *Clouds and Storms: The Behavior and Effect of Water in the Atmosphere*. The Pennsylvania State University Press, University Park, 123–128.

Plank, V. G., Atlas, D., and W. H. Paulsen, (1955). The nature and detectability of clouds and precipitation by 1.25 cm radar. *J. Meteor.*, **12**, 358–378.

Rangno, A. L., and P. V. Hobbs, (1994). Ice particle concentrations and precipitation development in small continental cumuliform clouds. *Quart. J. Roy. Meteor. Soc.*, **120**, 573–601.

Wallace, J. M., and P. V. Hobbs, (1977). *Atmospheric Science: An Introductory Survey*, New York: Academic Press, 216–218.

Warren, S. G., Hahn, C. J., and J. London, (1991). *Analysis of cloud information from surface weather reports*. World Climate Research Programme Report on Clouds, Radiative Transfer, and the Hydrological Cycle, World Meteorological Organization, 61–68.

World Meteorological Organization (1956). *International Cloud Atlas (Complete Atlas)*, Vol. I, Geneva: World Meteorological Organization, 175 pp.

World Meteorological Organization (1956). *International Cloud Atlas (Complete Atlas)*. Vol. II, Geneva: World Meteorological Organization, 124 pp.

World Meteorological Organization (1975). *International Cloud Atlas*, Vol. I, *Manual on the Observations of Clouds and Other Meteors*, Geneva: World Meteorological Organization, 155 pp.

World Meteorological Organization (1987). *International Cloud Atlas*. Vol. II, Geneva: World Meteorological Organization, 212 pp.

CHAPTER 22

ATMOSPHERIC ELECTRICITY AND LIGHTNING

WALTER A. LYONS AND EARLE R. WILLIAMS

1 GLOBAL ELECTRICAL CIRCUIT

The atmosphere is the thin veneer of gas surrounding our planet and is one of the most highly electrically insulating regions in the universe. The electrical conductivity of air near Earth's surface is some 12 orders of magnitude smaller than both Earth's crust and the conductive ionosphere. In such a medium, electric charges may be physically separated by both air and particle motions to create large electric fields. The laws of electrostatics apply in this situation, and, though the charges involved are seldom static, the application of these laws in the atmosphere (in fair weather regions, in electrified clouds which may produce lightning, and in the upper atmosphere where sprites are found) is the science of atmospheric electricity.

Earth itself, whose surface is composed of seawater and water-impregnated crustal rock, is a good electrical conductor. This conducting sphere is known to carry a net negative charge of half a million coulombs. The corresponding downward-directed electric field at the conductor's surface in fair weather regions is of the order of 100 V/m. The conductivity of highly insulating air near Earth's surface is of the order of 10^{-14} mho/m and, according to Ohm's law, invoking a linear relationship between electric field and current density, a downward-directed current is flowing with a current density of order 1 pA/m^2. Integrated over the globe, this amounts to a total current of about 1 kA.

The origin of Earth's persistent fair weather electrification remained a mystery for more than 100 years. In the 1920s, C.T.R. Wilson proposed that the electrification manifest in fair weather regions was caused by the action of electrified clouds and

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

thunderstorms worldwide. The fair weather region provided the return current path for electrified weather and is the collective battery for the global circuit. Nominally, 1000 storms are simultaneously active worldwide, each supplying a current of the order of 1 A, which flows upward from the cloud top into the conductive upper atmosphere where the current spreads out laterally and then turns downward in fair weather regions to the conductive earth to complete the circuit.

The finite conductivity of the lower atmosphere is primarily the result of ionization of neutral nitrogen and oxygen molecules by energetic cosmic radiation from space. The atmosphere attenuates the incoming ionizing radiation, and this causes the conductivity to increase exponentially with altitude toward the highly conductive ionosphere, with a scale height of about 5 km. Since the fair weather return current is uniform with altitude, Ohm's law guarantees that the electric field will decline exponentially with altitude from its maximum nominal value of 100 V/m at the surface, with the same scale height. The integral of the fair weather electric field with altitude from the conductive earth to the conductive upper atmosphere is referred to as the ionospheric potential, and amounts to some 250,000 V. This quantity is a global invariant and the standard measure of the global electrical circuit. The ionospheric potential and the net charge on Earth are not constant but vary by approximately ±15% over the course of every day in response to the systematic modulation of electrified weather by solar heating, centered in the tropics. Three major zones of deep convection—the Maritime Continent (including the Indonesian archipelago, Southeast Asia, and Northern Australia), Africa, and the Americas—exhibit strong diurnal variations, and the integrated effect lends a systematic variation to the electrification of Earth and the ionospheric potential in universal time. The consistent phase behavior between the action of the three tropical "chimney" regions and Earth's general electrification is generally regarded as key evidence in support of C.T.R. Wilson's hypothesis.

According to this general idea, the negative charge on Earth is maintained by the action of electrified clouds. One direct agent is cloud-to-ground lightning, which, as explained below, is most frequently negative in polarity. The less frequent but more energetic positive ground flashes, which are linked with sprites in the mesosphere should actually deplete the negative charge in Earth. Possibly more important (though more difficult to evaluate) agents for maintaining the negatively charged Earth are point discharge currents that flow from plants, trees, and other elevated objects in the kilovolt-per-meter level electric fields beneath electrified clouds, and precipitation currents carried to Earth by rain and graupel particles charged by microphysical processes within storms.

The global electrical circuit outlined here provides a natural framework for monitoring electrified weather on a worldwide basis. The expectation that lightning activity is temperature-dependent has spurred interest in the use of the global circuit as a monitor for global temperature change and for upper atmospheric water vapor. A second manifestation of the global circuit, sometimes referred to as the "AC" global circuit, is provided for by the same insulating air layer between two conductors. This configuration is an electromagnetic waveguide for a phenomenon known as Schumann resonances and maintained continuously by worldwide lightning activity. The wavelength for the fundamental resonant mode at 8 Hz is equal to the circumference of Earth. One advantage afforded by Schumann resonances over ionospheric potential as a measure of the global circuit is the capability for continuous monitoring from ground level. The signatures of the three major tropical zones referred to earlier are distinguishable in daily recordings from a single measurement station. Furthermore, variations in Schumann resonance intensity on longer time scale (semiannual, annual, interannual), for which global temperature variations are recognized, have also been documented.

The evidence that giant positive lightning can simultaneously create sprites in the mesosphere and “ring” Earth’s Schumann resonances to levels higher than the contribution of all other lightning combined has also spurred interest in this aspect of the global circuit.

2 WHAT IS LIGHTNING?

Lightning is a transient, high-current atmospheric discharge whose path length is measured in kilometers (Uman and Krider, 1989). It is, in effect, a big spark. Benjamin Franklin’s legendary (and incredibly risky) kite flying experiment in 1752 documented that, indeed, Thor’s bolts were “merely” electricity. The number of scientists killed or injured in subsequent years attempting to duplicate Franklin’s experiment attests to the threat to life and property inherent in a lightning discharge.

Lightning represents an electrical breakdown of the air within a cloud between separated pockets of positive and negative electrical space charge. What actually causes the formation and separation of the electrical charges within the thunderstorm? While considerable progress is being made in this research area, a concise theory has yet to emerge (Williams, 1988; MacGorman and Rust, 1998). Suffice it to say that in almost all cases of natural lightning, charge separation results from the interactions between ice particles, graupel, and supercooled droplets within thunderstorm clouds (Fig. 1). The complex updrafts and downdrafts separate the positive and negative charge pools allowing the electric field to attain ever-higher values until dielectric breakdown occurs. Storm updrafts penetrate well above the freezing (0°C) level. It is generally thought that significant updrafts ($>\sim 8\,\mathrm{m/s}$) in the layers where supercooled water droplets and ice particles coexist (0° to −40°C) are necessary though not sufficient for most lightning-generating scenarios. Lightning discharges also occur within the stratiform precipitation regions of large mesoscale convective systems. Though the sources of the charge centers have been thought caused by advection from the convectively active parts of the storm, there are also strong arguments favoring in situ charge generation. Occasional reports of lightning occurring in warm clouds (entirely above 0°C) have never been convincingly substantiated.

Several terminologies are employed to describe lightning. Discharges occurring within the cloud are called intracloud (IC) events. A cloud-to-ground event is called a CG. In most storms ICs greatly outnumber CGs. The typical ratio may be 4:1, but the ratio may approach 100:1 or more in some especially intense storms. The IC:CG

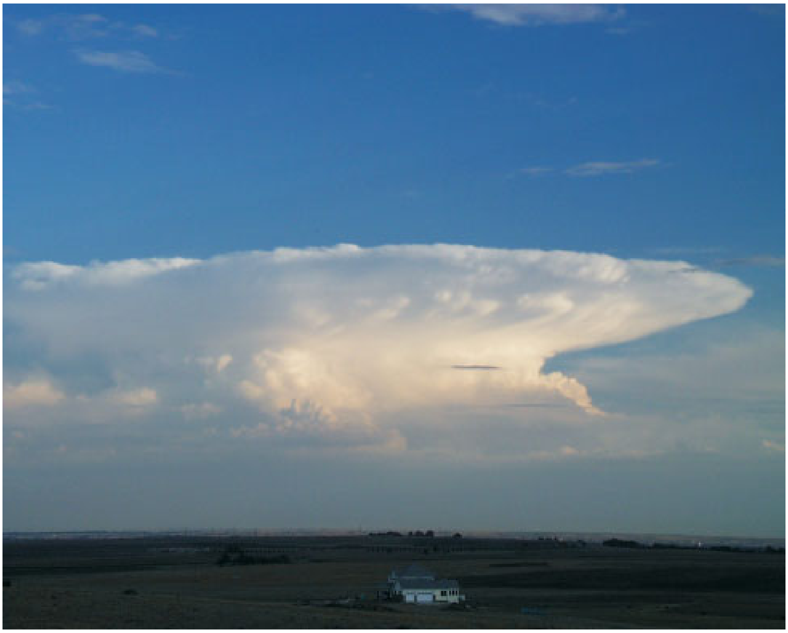


Figure 1 Cumulonimbus cloud over the U.S. High Plains unleashes a thunderstorm. See ftp site for color image.

ratio also shows geographical variations. It has been maintained that the ratio is lowest at more northerly latitudes. Recent satellite measurements have shown that, within the United States, certain regions, notably the High Plains and West Coast have much higher IC:CG ratios than other parts of the country such as Florida.

Cloud-to-ground events typically lower negative charge to earth (the negative CG, or −CG). CG discharges are sometimes initiated by leaders that are positively charged, and the resulting return stroke essentially lowers positive charge to ground (the +CG). Once thought to be very rare, the +CG is now thought to constitute about 10% of all CGs. For reasons not well understood, many severe local storms tend to have higher percentages of +CG flashes. Negative CGs typically lower on the order of 5 to 20 C of charge to ground, while +CGs can often be associated with much larger amounts of charge transfer.

When viewed with the naked eye, a CG flash often appears to flicker. This is a consequence of the fact that CG flashes are often composed of a sequence of multiple events called strokes. While flashes can consist of a single stroke (which is the norm for +CGs), the average stroke multiplicity for −CGs is around 4. Flashes with greater than 10 strokes are not uncommon, and there have been up to 47 strokes documented within a single flash. Strokes within the same flash tend to attach themselves to the same point on the surface, but this is not necessarily always the case. One Florida study (Rakov and Uman, 1990) found that about 50% of the flashes showed multiple terminations per flash to the ground. The multiple attach points within single flashes were separated by 0.3 to 7.3 km, with a mean of 1.3 km (Thottappillil et al., 1992). A very small percentage of flashes are initiated from the tops of tall towers, buildings, or mountains. Unlike conventional CG discharges, their channels branch upwards and outwards.

The CG flash is part of a complex series of events (Golde, 1977; Holle and Lopez, 1993; Uman, 1987; MacGorman and Rust, 1998). A typical flash begins with one or more negatively charged channel(s), called stepped leaders, originating deep within the cumulonimbus cloud and may emerge from the base of the cloud. Depending on the electrical charge distribution between cloud and ground, the leader proceeds erratically downward in a series of luminous steps traveling tens of meters in around a microsecond. A pause of about 50 μs occurs between each step. As the stepped leader approaches the ground, it is associated with electrical potentials on the order of 100 million volts. The intense electric field that develops between the front of the leader and the ground causes upward-moving discharges called streamers from one or more objects attached to the ground. When one of these streamers connects with the leader, usually about 100 m above the ground, the circuit is closed, and a current wave propagates up the newly completed channel and charge is transferred to the ground. This last process is called the return stroke. This is the brilliant flash seen by the naked eye, even though it lasts only tens to perhaps a few hundred microseconds. The peak current, which is typically on the order of 30 kA, is attained within about 1 μs. After the current ceases to flow through the leader-created channel, there is a pause ranging from 10 to 150 ms. Another type of leader, called a dart leader, can propagate down the same ionized channel, followed by a subsequent return stroke. The entire lightning discharge process can last, in extreme cases, for over 3 s, with the series of individual strokes comprising the CG flash sequence sometimes extending over a second. Certain phases of the lightning discharge, particularly the return stroke, proceed at speeds of more than one half the speed of light, while other discharge processes travel through the clouds up to two orders of magnitude more slowly.

While many CG return strokes are very brief (sub-100 μs), additionally many are followed by a long lasting (tens to many hundreds of milliseconds) continuing current. This behavior is particularly true of +CGs. This, along with the high initial peak currents often found in some +CGs, explains why such flashes are thought to ignite a substantial number of forest fires and cause other property damage (Fuquay et al., 1972). Since the temperature within the lightning channel has been estimated to reach 30,000 K, extending the duration of the current flow allows for greater transfer of heat energy and thus combustion.

There is great variability in the amount of peak current from stroke to stroke. While a typical peak current is in the 25 to 30 kA range, much smaller and larger currents can occur. Recent data suggest that peak currents of less than 10 kA may be more common than once believed. One survey of 60 million lightning flashes found 2.3% had peak currents of >75 kA; the largest positive CG reaching 580 kA and the largest negative CG 960 kA (Lyons et al., 1998). It has long been assumed that +CGs on the average had larger peak currents than −CGs, but again recent studies in the United States suggest that, while there are certainly many large peak current

+CGs, that for all classes between 75 and 400 kA, negative CGs outnumbered the positives (Lyons et al., 1998). It has generally been assumed that the first stroke in a flash contains the highest peak current. While on the average this is true (typically the first stroke averages several times that of subsequent strokes), recent research has found that subsequent strokes with higher peak currents are not that unusual (Thottappillil et al., 1992).

There are many reports of lightning strikes out of an apparently clear sky overhead. This is the so-called bolt from the blue. Powerful lightning discharges can leap for distances of 10 km (and sometimes more) beyond the physical cloud boundary. Under hazy conditions or with the horizon obscured by trees or buildings, the perception of a bolt from the blue is entirely understandable. Lightning tracking equipment at NASA's Kennedy Space Center has documented one discharge that came to ground more than 50 km from its point of origin within a local thunderstorm.

Lightning has also been associated with smoke plumes from massive forest fires, likely from the thunderstorm clouds induced by the intense local heat source (Fuquay et al., 1972). Plumes from volcanic eruptions have also produced lightning discharges (Uman, 1987). Lightning has also been associated with large underwater explosions as well as atmospheric thermonuclear detonations. Lightning is not confined to Earth's atmosphere. Space probes have detected lightning within the clouds of Jupiter, and possibly in Saturn and Uranus. Considerable additional information on the physics and chemistry of lightning can be found in Uman (1987), Golde (1977), Uman and Krider (1989), the National Research Council (1986), and Williams (1988).

3 HUMAN COST OF LIGHTNING

Lightning is a major cause of severe-weather-related deaths in the United States. According to the U.S. Department of Commerce's "Storm Data," during the period 1940–1981, lightning killed more people in the United States (7583) than hurricanes and tornadoes combined (7071). Between 1959 and 1995, the toll was 3322 reported deaths and 10,346 injuries. Thus while there was a downward trend in lightning casualties during the later part of the twentieth century in the United States, in many years lightning deaths still continue to outnumber those associated with tornadoes and hurricanes. Floods, however, are the number one killer. Between 1959 and 1987, the most lightning casualties were recorded in Florida, the nation's "lightning capital." The states of Michigan, North Carolina, Pennsylvania, Ohio, and New York followed in the rankings. Though having lower lightning frequencies, the large populations of these states translated into large numbers of casualties. Lightning casualties are less well documented in other parts of the world. It has been estimated by Andrews et al. (1992) that the annual worldwide lightning casualty figures are about 1000 fatalities and 2500 injuries.

Many of the published U.S. and worldwide losses from lightning may be significant underestimates (Lopez et al., 1993). In Colorado, a region with an efficient emergency response and reporting system, detailed examinations of hospital records found that U.S. government figures still underestimated lightning deaths by at least 28% and injuries by 40%. It appears many deaths ultimately caused by lightning (by heart attacks or in lightning-triggered fires) are not included within the casualty totals. Cooper (1995) suggests actual U.S. lightning deaths are probably 50% higher than reported, with as many as 500 injuries, many of them unreported or erroneously classified. Thus the lightning threat tends to be underestimated by public safety officials. This may be in part also due to the tendency of lightning to kill or injure in small numbers. It rarely receives the press notice accorded more "spectacular" weather disasters such as hurricanes, tornadoes, and flash floods. In recent years, however, there have been increasing reports of "mass casualties." In July 1991, at least 22 people were injured when lightning struck a crowded beach in Potterville, Michigan. In 1980 a high school football team practicing in Wickliffe, Ohio, was struck by lightning. All the players were knocked over and one was injured. In the Congo, 11 members of one soccer team were killed as a flash struck the playing field. In Nigeria, 12 were killed as a storm struck an outdoor funeral service. In the Colorado Rockies, a herd of 56 elk were found dead within a radius of 100 yards, apparently the result of a lightning strike.

4 ECONOMIC COSTS OF LIGHTNING

While some summaries of lightning property damages in the United States have listed relatively modest annual losses (under $100 million) recent research suggests the damages are many times greater. One study of insurance records (Holle et al., 1996) estimated about 1 out of every 55 Colorado CG flashes resulted in a damage claim. A major property insurer, covering about 25% of all U.S. homes, reports paying out 304,000 lightning-related claims annually. The average insurance lightning damage claim was $916. This suggests that U.S. homeowner losses alone may exceed $1 billion annually. Lightning has been responsible for a number of major disasters, including a substantial fraction of the world's wildfires in forests and grasslands. During the 1990s in the United States, some 100,000 lightning-caused forest fires consumed millions of acres of forests. Lightning is the leading cause of electric utility outages and equipment damage throughout the world, estimated to account for 30 to 50% of all power outages. On July 13, 1977, a lightning strike to a power line in upstate New York triggered a series of chain-reaction power outages, ultimately blacking out parts of New York City for up to 24 h. The resulting civil disturbances resulted in over $1 billion in property losses. Though the actual total of lightning's economic effects are unknown, the National Lightning Safety Institute estimates it could be as high as $4 billion to $5 billion per year in the United States alone. Given the substantial noninsured and poorly documented losses occurring in forestry, aviation, data processing, and telecommunications plus the lost wages due to power disruptions and stoppages of outdoor economic activities, this estimate seems plausible.

On December 8, 1963, a lightning strike ignited fuel vapors in the reserve tank of a Pan American airliner while the plane was circling in a holding pattern during a thunderstorm. The Boeing 707 crashed in Elkton, Maryland, with a loss of 81 lives. Engineering improvements make such deadly airliner incidents much less likely today. Spacecraft launches, however, still must contend with the lightning hazard. The 1969 flight of Apollo 12 almost ended in tragedy rather than the second moon landing when the vehicle "triggered" two lightning discharges during the initial moments of ascent. The event upset much of the instrumentation and caused some equipment damage, but the crew fortunately was able to recover and proceed with the mission. On March 26, 1987, the U.S. Air Force launched an Atlas/Centaur rocket from the Kennedy Space Center. Forty-eight seconds after launch the vehicle was "struck" by lightning—apparently triggered by the ionized exhaust plume trailing behind the rocket (Uman and Krider, 1989). The uninsured cost to U.S. taxpayers was $162 million. Ground facilities can fare poorly as well. On July 10, 1926, lightning struck a U.S. Navy ammunition depot. The resulting explosions and fires killed 19 people and caused $81 million in property losses. A lightning strike to a Denver warehouse in 1997 resulted in a $50 million fire.

Blaming lightning for equipment failure has become commonplace, and perhaps too much so. One estimate suggests that nearly one third of all insurance lightning damage claims are inaccurate or clearly fraudulent. Damage claim verification using data from lightning detection networks is now becoming commonplace in the United States.

5 CLIMATOLOGY OF LIGHTNING

An average of 2000 thunderstorm cells are estimated present on Earth at any one time (Uman, 1987). Over the past several decades, the worldwide lightning flash rate has been variously estimated at between 25 and 400 times per second. Based upon more recent satellite monitoring, the consensus is emerging that global lightning frequency probably averages somewhat less than 50 IC and CG discharges per second. Lightning is strongly biased to continental regions, with up to 10 times more lightning found over or in the immediate vicinity of land areas.

Before the advent of the National Lightning Detection Network (NLDN), various techniques were used to estimate CG flash density (expressed in flashes/km^2 year unless otherwise stated). One measure frequently employed is the isokeraunic level, derived from an analysis of the number of thunderstorm days per year at a point. A thunderstorm day occurs when a trained observer at a weather station reports hearing at least one peal of thunder (typically audible over a range of 5 to 10 km). Changery (1981) published a map of thunder days across the United States using subjective reports of thunder from weather observers at 450 stations over the period 1948–1977. A very rough rule of thumb proposed that for each 10 thunderstorm days there were between 1 and 2 flashes/km^2 year.

Needed, however, was a more accurate determination of the CG flash density, for a variety of purposes, including the design of electrical transmission line protection systems. During the 1980s, as discussed below, lightning detection networks capable of locating CG events began operating in the United States. By the mid-1990s, a reasonably stable annual pattern had begun to emerge (Orville and Silver, 1997). Approximately 20 to 25 million flashes per year strike the lower 48 states. The expected flash density maximum was found in central Florida ($\sim$10 to 15 flashes/km^2 year). Other regional maxima include values around 6 flashes/km^2 year along the Gulf Coast and around 5 flashes/km^2 year in the Midwest and Ohio River region. More than half the United States has a flash density of 4 flashes/km^2 year or greater. In any given year the highest flash densities may not be found in central Florida. For instance, the highest flash densities during 1993 were found in the Mississippi and Ohio valleys, in association with the frequent and massive thunderstorms leading to the great floods of that summer. Two annual maxima occurred in 1995, in southern Louisiana and near the Kentucky–Illinois border. In the intermountain west, values around 1.0 flashes/km^2 year are common with West Coast densities being less than 0.5 flashes/km^2 year.

Different thunderstorm types can produce lightning at very different rates. A small, isolated air mass shower may produce a dozen IC discharges and just one or two CGs. At the other extreme, the largest convective system, aside from the tropical cyclone, is the mesoscale convective complex (MCC). These frequent the central United States during the warm season as well as many other parts of the world including portions of South America, Africa, and Asia. The MCC, which can cover 100,000 km^2 and last for 12 h or more, has been known to generate CG strikes at rates exceeding 10,000 per hour. Goodman and MacGorman (1986) have noted that the passage of the active portion of a single MCC can result in 25% of a region's annual CG total.

There are distinct regional differences in such parameters as the percentage of flashes that are of positive polarity and also of peak current. The U.S. region with the largest number of positive polarity CGs ($>$12.5% of the total) is concentrated in a broad belt in the interior of the United States, stretching from Texas north to Minnesota. A study of the climatology of CGs with large peak currents (defined as $>$75 kA) found powerful +CGs were similarly clustered in a band stretching from eastern New Mexico and Colorado northeastward into Minnesota (Lyons et al., 1998). By contrast, large peak current $-$CGs were largely confined to the overocean regions of the northern Gulf of Mexico and along the southeastern U.S. coastline and the adjacent Atlantic ocean. The implications of these regional differences in flashes with large peak current upon facility design and protection as well as hazards to human safety are yet to be explored.

The geographic distribution of lightning flash density on a global basis is becoming better known. A combination of proliferating land-based CG detection networks and satellite observations should allow a more robust global climatology to emerge over the next decade or so. For now we must rely on observations of total lightning (IC + CG events) made by polar orbiting satellites such as NASA's Optical Transient Detector (OTD). Figure 2 shows the elevated lightning densities over parts of North America, the Amazon, central Africa, and the Maritime Continent of Southeast Asia.

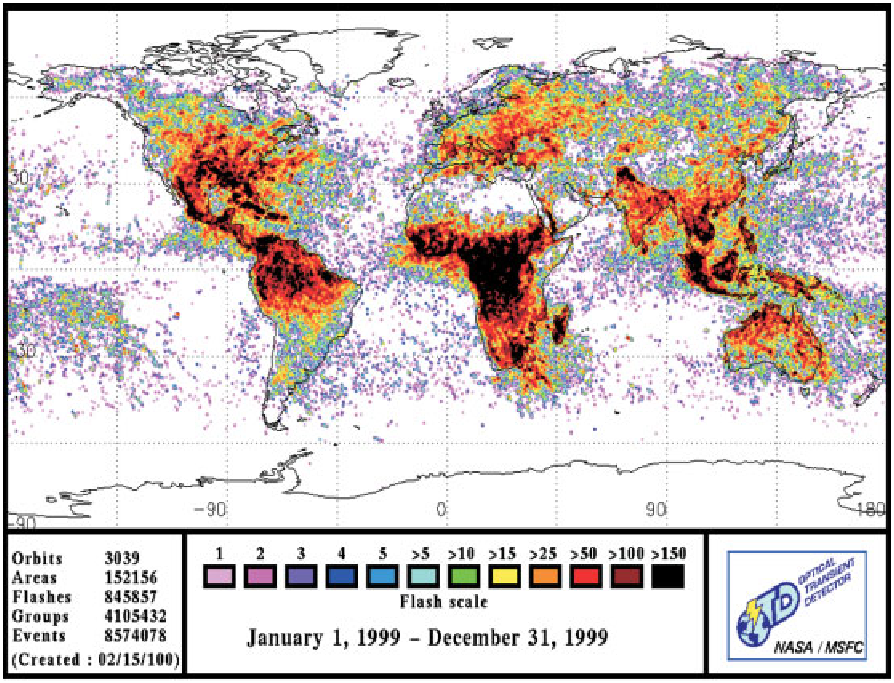


Figure 2 (see color insert) Map of global lightning detected over a multiyear period by NASA's Optical Transient Detector flying in polar orbit. See ftp site for color image.

6 LIGHTNING DETECTION

During the 1980s, a major breakthrough in lightning detection occurred with the development of lightning detection networks. Two distinct approaches were developed to detect and locate CGs. One was based on magnetic direction finding (MDF) and the second on time-of-arrival (TOA) approaches. Krider et al. (1980) described one of the first applications of an MDF-based system, the location of CG flashes in remote parts of Alaska, which might start forest fires. The TOA technique was applied to research and operations in the United States in the late 1980s (Lyons et al., 1985, 1989). As various MDF and TOA networks proliferated, it soon was evident that a merger of the two techniques was both technologically and economically desirable. The current U.S. National Lightning Detection Network (NLDN) (Cummins et al., 1998) consists of a hybrid system using the best of both approaches.

The NLDN uses more than 100 sensors communicating via satellite links to a central control facility in Tucson, Arizona, to provide real-time information on CG events over the United States. While employing several different configurations, it has been operating continuously on a nationwide basis since 1989. Currently, the mean flash location accuracy is approaching 500 m and the flash detection efficiency ranges between 80 and 90%, varying slightly by region (Cummins et al., 1998). It can provide coverage for several hundred kilometers beyond the U.S. coastline, although the detection efficiency drops off with range. The networks are designed to detect individual CG strokes, which can then be combined to flashes. The NLDN provides data on the time (nearest millisecond), location (latitude and longitude), polarity, and peak current of each stroke in a flash, along with estimates of the locational accuracy. National and regional networks using similar technology are gradually evolving in many areas, including Japan, Europe, Asia, Australia, and Brazil. The Canadian and U.S. networks have effectively been merged into a North American network (NALDN).

In the United States, lightning data are available by commercial subscription in a variety of formats. Users can obtain real-time flash data via satellite. Summaries of ongoing lightning-producing storms are distributed through a variety of websites. Pagers are being employed to alert lightning-sensitive facilities, such as golf courses, to the approach of CGs. Historical flash and stroke data can also be retrieved from archives (see www.LightningStorm.com). These are used for a wide variety of scientific research programs, electrical system design studies, fault finding, and insurance claim adjusting.

Future detection developments may include three-dimensional lightning mapping arrays (LMAs). These systems, consisting of a network of very high frequency (VHF) radio receivers, can locate the myriad of electrical emissions produced by a discharge as it wends its way through the atmosphere. Combined with the existing CG mapping of the NLDN, total lightning will be recorded and displayed in real time. Such detection systems are now being tested by several private firms, universities, and government agencies.

A new long-range lightning detection technique employs emissions produced by CG lightning events in the extremely low frequency (ELF) range of the spectrum. In the so-called Schumann resonance bands (8, 14, 20, 26, . . . Hz) occasional very large transients occur that stand out against the background "hum" of all the world's lightning events. These transients are apparently the result of atypical lightning flashes that transfer massive amounts of charge (>100 C) to ground. It is suspected many of these flashes also produce mesospheric sprites (see below) (Huang et al., 1999). Since the ELF signatures travel very long distances within Earth–ionosphere waveguide, just a few ELF receivers are potentially capable of detecting and locating the unusual events on a global basis.

7 LIGHTNING PROTECTION

Continued improvement in the detection of lightning using surface and space-based systems is a long-term goal of the atmospheric sciences. This must also be accompanied by improving alerts of people and facilities in harm's way, more effective medical treatment for those struck by lightning, and improved engineering practice in the hardening of physical facilities to withstand a lightning strike.

The U.S. National Weather Service (NWS) does not issue warnings specifically for lightning, although recent policy is to include it within special weather statements and warnings for other hazards (tornado, hail, high winds). Thus members of the public are often left to their own discretion as to what safety measures to take. When should personnel begin to take avoidance procedures? The "flash-to-bang method" as it is sometimes called is based on the fact that while the optical signature of lightning travels at the speed of light, thunder travels at 1.6 km every 5 s. By counting the time interval between seeing the flash and hearing the thunder, one can estimate the distance rather accurately. The question is then how far away might the next CG flash jump from its predecessor's location? Within small Florida air mass thunderstorms, the average value of the distance between successive CG strikes is 3.5 km (Uman and Krider, 1989). However, recently statistics from larger midwestern storms show the mean distance between successive flashes can often be 10 km or more ($\sim$30 s flash-to-bang or more). According to Holle et al. (1992), lightning from receding storms can be as deadly as that from approaching ones. People are struck more often near the end of storms than during the height of the storm (when the highest flash densities and maximum property impacts occur). It appears that people take shelter when a storm approaches and remain inside while rainfall is most intense. They fail to appreciate, however, that the lightning threat can continue after the rain has diminished or even ceased, for up to a half hour. Also, the later stages of many storms are characterized by especially powerful and deadly +CG flashes. Thus, has emerged the 30:30 rule (Holle et al., 1999). If the flash-to-bang duration is less than 30 s, shelter should be sought. Based upon recent research in lightning casualties, persons should ideally remain sheltered for about 30 min after the cessation of audible thunder at the end of the storm.

Where are people struck by lightning? According to one survey, 45% were in open fields, 23% were under or near tall or isolated trees, 14% were in or near water, 7% were on golf courses, and 5% were using open-cab vehicles. Another survey noted that up to 4% of injuries occurred with persons talking on (noncordless) telephones or using radio transmitters. In Colorado some 40% of lightning deaths occur while people are hiking or mountain climbing during storms. The greatest risks occur to those who are among the highest points in an open field or in a boat, standing near tall or isolated trees or similar objects, or in contact with conducting objects such as plumbing or wires connected to outside conductors. The safest location is inside a substantial enclosed building. The cab of a metal vehicle with the windows closed is also relatively safe. If struck, enclosed structures tend to shield the occupants in the manner of a Faraday cage. Only rarely are people killed directly by lightning while inside buildings. These include persons in contact with conductors such as a plumber leaning on a pipe, a broadcaster talking into a microphone, persons on the telephone, or an electrician working on a power panel.

A lightning strike is not necessarily fatal. According to Cooper (1995) about 3 to 10% of those persons struck by lightning are fatally injured. Of those struck, fully 25% of the survivors suffered serious long-term after effects including memory loss, sleep disturbance, attention deficits, dizziness, numbness/paralysis, and depression. The medical profession has been rather slow to recognize the frequency of lightning-related injuries and to develop treatment strategies. There has, fortunately, been an increasing interest paid to this topic over the past two decades (Cooper, 1983, 1995; Andrews et al., 1988, 1992). Persons struck by lightning who appear "dead" can often be revived by the prompt application of cardiopulmonary resuscitation (CPR). The lack of external physical injury does not preclude the possibility of severe internal injuries.

Lightning can and does strike the same place more than once. It is common for the same tall object to be struck many times in a year, with New York City's Empire State Building being struck 23 times per year on average (Uman, 1987). From a risk management point of view, it should be assumed that lightning cannot be "stopped" or prevented. Its effects, however, can be greatly minimized by diversion of the current (providing a controlled path for the current to follow to ground) and by shielding.

According to the U.S. Department of Energy (1996), first-level protection of structures is provided by the lightning grounding system. The lightning leader is not influenced by the object that is about to be struck until it is only tens of meters away. The lightning rod (or air terminal) developed by Benjamin Franklin remains a key component for the protection of structures. Its key function is to initiate an upward-connecting streamer discharge when the stepped leader approaches within striking distance. Air terminals do not attract significantly more strikes to the structure than the structure would receive in their absence. They do create, however, a localized preferred strike point that then channels the current to lightning conductors and safely down into the ground. To be effective, a lightning protection system must be properly designed, installed, and maintained. The grounding must be adequate. Sharp bends (less than a 20-cm radius) can defeat the conductor's function. The bonding of the components should be thermal, not mechanical, whenever possible. Recent findings suggest that air terminals with rounded points may be more effective than those with sharp points.

Sensitive electronic and electrical equipment should be shielded. Switching to auxiliary generators, surge protectors, and uninterruptible power supplies (UPSs) can help minimize damage from direct lightning strikes or power surges propagating down utility lines from distant strikes. In some cases, the most practical action is to simply disconnect valuable assets from line power until the storm has passed.

There is still much to be learned about atmospheric electrical phenomena and their impacts. The characteristics of lightning wave forms and the depth such discharges can penetrate into the ground to affect buried cables are still under investigation. Ongoing tests using rocket-triggered lightning are being conducted by the University of Florida at its Camp Blanding facility in order to test the hardening of electrical systems for the Electric Power Research Institute.

The notion that lightning is an "act of God" against which there is little if any defense is beginning to fade. Not only can we forecast conditions conducive to lightning strikes and monitor their progress, but we are beginning to understand how to prevent loss of life, treat the injuries of those harmed by a strike, and decrease the vulnerability of physical assets to lightning.

8 IMPACTS OF LIGHTNING UPON THE MIDDLE ATMOSPHERE

Science often advances at a deliberate and cautious pace. Over 100 years passed before persistent reports of luminous events in the stratosphere and mesosphere, associated with tropospheric lightning, were accepted by the scientific community. Since 1886, dozens of eyewitness accounts, mostly published in obscure meteorological publications, have been accompanied by articles describing meteorological esoterica such as half-meter wide snow flakes and toads falling during rain showers. The phenomena were variously described using terms such "cloud-to-space lightning" and "rocket lightning." A typical description might read, "In its most typical form it consists of flames appearing to shoot up from the top of the cloud or, if the cloud is out of sight, the flames seem to rise from the horizon." Such eyewitness reports were largely ignored by the nascent atmospheric electricity community—even when they were posted by a Nobel Prize winning physicist such as C.T.R. Wilson (1956). During the last three decades, several compendia of similar subjective reports from credible witnesses worldwide were prepared by Otha H. Vaughan (NASA Marshall) and the late Bernard Vonnegut (The State University of New York—Albany). The events were widely dispersed geographically from equatorial regions to above 50° latitude. About 75% of the observations were made over land. The eyewitness descriptions shared one common characteristic—they were perceived as highly atypical of "normal" lightning (Lyons and Williams, 1993). The reaction of the atmospheric science community could be summarized as casual interest at best. Then, as so often happens in science, serendipity intervened.

The air of mystery began to dissipate in July 1989. Scientists from the University of Minnesota, led by Prof. John R. Winckler, were testing a low-light camera system (LLTV) for an upcoming rocket flight at an observatory in central Minnesota (Franz et al., 1990). The resulting tape contained, quite by accident, two fields of video that provided the first hard evidence for what are now called sprites. The twin pillars of light were assumed to originate with a thunderstorm system some 250 km to the north along the Canadian border. The storm system, while not especially intense, did contain a larger than average number of positive polarity cloud-to-ground (+CG) lightning flashes. From this single observation emanated a decade of fruitful research into the electrodynamics of the middle atmosphere.

Spurred by this initial discovery, in the early 1990s NASA scientists searched videotapes from the Space Shuttle's LLTV camera archives and confirmed at least 17 apparent events above storm clouds occurring worldwide (Boeck et al., 1998). By 1993, NASA's Shuttle Safety Office had developed concerns that this newly discovered "cloud-to-space lightning" might be fairly common and thus pose a potential threat to Space Shuttle missions especially during launch or recovery. Based upon the available evidence, the author's hunt for these elusive events was directed above the stratiform regions of large mesoscale convective systems (MCSs), known to generate relatively few but often very energetic lightning discharges. On the night of July 7, 1993, an LLTV was deployed for the first time at the Yucca Ridge Field Station (YRFS), on high terrain about 20 km east of Fort Collins, Colorado. Exploiting an uninterrupted view of the skies above the High Plains to the east, the LLTV was trained above a large nocturnal MCS in Kansas, some 400 km distant (Lyons, 1994). Once again, good fortune intervened as 248 sprites were imaged over the next 4 h. Analyses revealed that almost all the sprites were associated with +CG flashes, and assumed an amazing variety of shapes (Fig. 3). Within 24 h, in a totally independent research effort, sprites were imaged by a University of Alaska team onboard the NASA DC8 aircraft over Iowa (Sentman and Wescott, 1993). The following summer, the University of Alaska's flights provided the first color videos detailing the red sprite body with bluish, downward extending tendrils. The same series of flights documented the unexpected and very strange blue jets (Wescott et al., 1995).

By 1994 it had become apparent that there was a rapidly developing problem with the nomenclature being used to describe the various findings in the scientific literature. The name sprite was selected to avoid employing a term that might presume more about the physics of the phenomena than our knowledge warranted. Sprite replaced terms such as "cloud-to-space" lightning and "cloud-to-ionosphere discharge" and similar appellations that were initially used. Today a host of phenomena have been named: sprites, blue jets, blue starters, elves, sprite halos, and trolls, with perhaps others remaining to be discovered. Collectively they have been termed transient luminous events (TLEs).

Since the first sprite observations in 1989, the scientific community's misperception of the middle atmosphere above thunderstorms as "uninteresting" has completely changed. Much has been learned about the morphology of TLEs in recent years. Sprites can extend vertically from less than 30 km to about 95 km. While telescopic investigations reveal that individual tendril elements may be of the order of 10 m across, the envelope of the illuminated volume can exceed 10^4 km^3. Sprites are

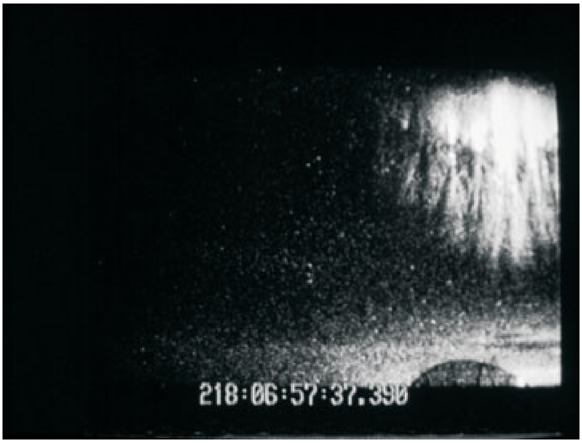


Figure 3 One of the many shapes assumed by sprites in the mesosphere. Top may be near 95 km altitude with tendril-like streamers extending downward toward 40 km or lower. Horizon is illuminated by the flash of the parent lightning discharge. Image obtained using a low-light camera system at the Yucca Ridge Field Station near Ft. Collins, CO. Sprite is some 400 km away.

almost always preceded by +CG flashes, with time lags of less than 1 to over 100 ms. To date, there are only two documented cases of sprites associated with negative polarity CGs. The sprite parent +CG peak currents range widely, from under 10 kA to over 200 kA, though on average the sprite +CG peak current is 50% higher than other +CGs in the same storm. High-speed video images (1000 fps) suggest that many sprites usually initiate around 70 to 75 km from a small point, and first extend downward and then upward, with development at speeds around 10^7 m/s. Sprite luminosity on typical LLTV videos can endure for tens of milliseconds. Photometry suggests, however, that the brightest elements usually persist for a few milliseconds, though occasionally small, bright "hot spots" linger for tens of milliseconds. By 1995, sprite spectral measurements by Hampton et al. (1996) and Mende et al. (1995) confirmed the presence of the N_2 first positive emission lines. In 1996, photometry provided clear evidence of ionization in some sprites associated with blue emissions within the tendrils and sometimes the sprite body (Armstrong et al., 2000). Peak brightness within sprites is on the order of 1000 to 35,000 kR. In 7 years of observations at Yucca Ridge, sprites were typically associated with larger storms ($>10^4$ km^2 radar echo), especially those exhibiting substantial regions of stratiform precipitation (Lyons, 1996). The TLE-generating phase of High Plains storms averages about 3 h. The probability of optical detection of TLEs from the ground in Colorado is highest between 0400 and 0700 UTC. It is suspected that sprite activity maximizes around local midnight for many storms around the world. The TLE counts observed from single storm systems has ranged from 1 to 776, with 48 as an average count. Sustained rates as high as once every 12 s have been noted, but more typical intervals are on the order of 2 to 5 min.

Can sprites be detected with the naked eye? The answer is a qualified yes. Most sprites do not surpass the threshold of detection of the dark-adapted human eye, but some indeed do. Naked-eye observations require a dark (usually rural) location, no moon, very clean air (such as visibilities typical of the western United States) and a dark-adapted eye (5 min or more). A storm located 100 to 300 km distant is ideal if it contains a large stratiform precipitation area with +CGs. The observer should stare at the region located some 3 to 5 times the height of the storm cloud. It is best to shield the eye from the lightning flashing within the parent storm. Often sprites are best seen out of the corner of the eye. The event is so transient that often observers cannot be sure of what they may have seen. The perceived color may not always appear "salmon red" to any given individual. Given the human eye's limitations in discerning color at very low light levels, some report seeing sprites in their natural color, but others see them as white or even green.

While most TLE discoveries came as a surprise, one was predicted in advance from theoretical arguments. In the early 1990s, Stanford University researchers proposed that the electromagnetic pulse (EMP) from CG flashes could induce a transient glow at the lower ledge of the ionosphere between 80 and 100 km altitudes (Inan et al., 1997). Evidence for this was first noted in 1994 using LLTVs at Yucca Ridge. Elves were confirmed the following year by photometric arrays deployed at Yucca Ridge by Tohoku University (Fukunishi et al., 1996). Elves, as they are now called, are believed to be expanding quasi-torroidal structures that attain an integrated width of several hundred kilometers. (The singular is elve, rather than elf, in order to avoid confusion with ELF radio waves, which are used intensively in TLE studies.) Photometric measurements suggest the elve's intrinsic color is red due to strong N_2 first positive emissions. While relatively bright (1000 kR), their duration is $<500\,\mu s$. These are usually followed by $\sim 300\,\mu s$ very high peak current (often >100 kA) CGs, most of which are positive in polarity. Stanford University researchers, using sensitive photometric arrays, documented the outward and downward expansion of the elve's disk. They also suggest many more dim elves occur than are detected with conventional LLTVs. It has been suggested that these fainter elves are more evenly distributed between positive and negative polarity CGs.

Recently it has been determined that some sprites are preceded by a diffuse disk-shaped glow that lasts several milliseconds and superficially resembles elves. These structures, now called "halos," are less than 100 km wide, and propagate downward from about 85 to 70 km altitude. Sprite elements sometimes emerge from the lower portion of the sprite halo's concave disk.

Blue jets are rarely observed from ground-based observatories, in part due to atmospheric scattering of the shorter wavelengths. LLTV video from aircraft missions revealed blue jets emerging from the tops of electrically active thunderstorms. The jets propagate upwards at speeds of ~ 100 km/s reaching terminal altitudes around 40 km. Their estimated brightness is on the order of 1000 kR. Blue jets do not appear to be associated with specific CG flashes. Some blue jets appear not to extend very far above the clouds, only propagating as bright channels for a few kilometers above the storm tops. These nascent blue jets have been termed blue starters. During the 2000 observational campaign at Yucca Ridge, the first blue starters ever imaged from the ground were noted. They were accompanied by very bright, short-lived (~ 20 ms) "dots" of light at the top of MCS anvil clouds. A NASA ER2 pilot flying over the Dominican Republic high above Hurricane Georges in 1998 described seeing luminous structures that resembled blue jets.

The troll is the most recent addition to the TLE family. In LLTV videos, trolls superficially resemble blue jets, yet they clearly contain significant red emissions. Moreover, they occur after an especially bright sprite in which tendrils have extended downward to near cloud tops. The trolls exhibit a luminous head leading a faint trail moving upwards initially around 150 km/s, then gradually decelerating and disappearing by 50 km. It is still not known whether the preceding sprite tendrils actually extend to the physical cloud tops or if the trolls emerge from the storm cloud per se.

Worldwide, a variety of storm types have been associated with TLEs. These include the larger midlatitude MCSs, tornadic squall lines, tropical deep convection, tropical cyclones, and winter snow squalls over the Sea of Japan. It appears that the central United States may be home to some of the most prolific TLE producers, even though only a minority of High Plains thunderstorms produce TLEs. Some convective regimes, such as supercells, have yet to be observed producing many TLEs and the few TLEs are mostly confined to any stratiform precipitation region that may develop during the late mature and decaying stages. Furthermore, the vast majority of +CGs, even many with peak currents above 50 kA, produce neither sprites nor elves, which are detectable using standard LLTV systems. While large peak current

+CGs populate both MCSs and supercells, only certain +CGs possess characteristics that generate sprites or elves.

Monitoring ELF radio emissions in the Schumann resonance bands (8 to 120 Hz) has provided a clue to what differentiates the TLE parent CG from "normal" flashes. The background Schumann resonance signal is produced from the multitude of lightning flashes occurring worldwide. It is generally a slowly varying signal, but occasionally brief amplitude spikes, called Q-bursts, are noted. Their origin was a matter of conjecture for several decades. In 1994, visual sprite observations at Yucca Ridge were coordinated in real time with ELF transients (Q-bursts) detected at a Rhode Island receiver station operated by the Massachusetts Institute of Technology (Boccippio et al., 1995). This experiment, repeated many times since, clearly demonstrated that Q-bursts are companions to the +CG flashes generating both sprites and elves. ELF measurements have shown that sprite parent +CGs are associated with exceptionally large charge moments (300 to >2000 C-km). The sprite +CG ELF waveform spectral color is "red," that is, peaked toward the fundamental Schumann resonance mode at 8 Hz. Lightning charge transfers of hundreds of coulombs may be required for consistency with theories for sprite optical intensity and to account for the ELF Q-burst intensity. Lightning causal to elves has a much flatter ("white") ELF spectrum, and though associated with the very highest peak current +CGs (often >150 kA), exhibits much smaller charge moments (<300 C-km) (Huang et al., 1999).

Recent studies of High Plains MCSs confirm that their electrical and lightning characteristics are radically different from the textbook "dipole" thunderstorm model, derived largely from studies of rather small convective storms. Several horizontal laminae of positive charge are found, one often near the 0°C layer, and these structures persist for several hours over spatial scales of ~ 100 km. With positive charge densities of 1 to 3 $\mathrm{nC/m^3}$, even relatively shallow layers (order 500 m) covering 10^4 to $10^5\,\mathrm{km}^2$ can contain thousands of coulombs. Some 75 years ago, C.T.R. Wilson (1956) postulated that large charge transfers and particularly large charge moments from CG lightning appear to be a necessary condition for conventional breakdown that could produce middle atmospheric optical emissions. Sprites occur most readily above MCS stratiform precipitation regions with radar echo areas larger than $\sim 10^4\,\mathrm{km}^2$. It is not uncommon to observe rapid-fire sequences of sprites propagating above storm tops, apparently in synchrony with a large underlying horizontal lightning discharge. One such "dancer" included a succession of eight individual sprites within 700 ms along a 200-km-long corridor. This suggests a propagation speed of the underlying "forcing function" of $\sim 3 \times 10^5$ m/s. This is consistent with the propagation speed of "spider" lightning—vast horizontal dendritic channels tapping extensive charge pools once a +CG channel with a long continuing current becomes established (Williams, 1998).

It is suspected that only the larger MCS, which contain large stratiform precipitation regions, give rise to the +CGs associated with the spider lightning networks able to lower the necessary charge to ground. The majority of sprite parent +CGs are concentrated in the trailing MCS stratiform regions. The radar reflectivities associated with the parent +CGs are relatively modest, 30 to 40 dBZ or less. Only a

TABLE 1 Current Ideas on TLE Storm/Lightning Parameters in Selected Storm Types

	Core of Supercells[a]	MCS Stratiform Region	"Ordinary" Convection
$+$CG peak currents	$> \sim 40$ kA	$> \sim 60$ kA	$\sim$30 kA
Storm dimension	10–20 km	10–500 km	<20–100 km
Spider discharges	Few/small	Many/large	Some
Continuing current	Short if any	Longest, strongest	Low intensity
$+$CG channel height	10–15 km (?)	5 km (?)	10 km (?)
Sprites occur?	No (except at end)	Many	Rare (?)
Elves occur?	No (except at end)	Many	Rare (?)
Blue jets occur?	Yes (?)	Rare (?)	Rare (?)

[a] Some supercells may generate a few sprites during their final phase when/if extensive stratiform develops.

small subregion of trailing stratiform area produces sprite and elves. It would appear that this portion of the MCS possesses, for several hours, the requisite dynamical and microphysical processes favorable for the unique electrical discharges, which drive TLEs. Tables 1 and 2 summarize the relationships between lightning and the major TLE types

9 THEORIES ON TRANSIENT LUMINOUS EVENTS

Transient luminous events have captured the interest of many theoreticians (Rowland, 1998; Pasko et al., 1995). Several basic mechanisms have been postulated to explain the observed luminous structures. These include sprite excitation by a quasi-electrostatic (QE) mechanism. Sprite production by runaway electrons in the strong electric field above storms has been suggested. The formation of elves from lightning electromagnetic pulses (EMP) is now generally accepted. More than one mechanism may be operating, but on different temporal and spatial scales, which in turn produce the bewildering variety of TLE shapes and sizes. Absent from almost all theoretical modeling efforts are specific data on key parameters characterizing lightning flashes that actually produce TLEs. Many modelers refer to standard reference texts, which, in turn, tend to compile data taken in storm types and locales that are not representative of the nocturnal High Plains. Specifically, many invoke the conventional view that the positive charge reservoir for the lightning is found in the upper portion of the cloud at altitudes of $\sim$10 km. But the positive dipole (or tripole) storm model has been found wanting in many midcontinental storms. We surveyed the range of lightning parameters used in over a dozen theoretical modeling studies. While the proposed heights of the vertical $+$CG channel ranges from 4 to 20 km, there is a clear preference for 10 km and above. The amount of charge lowered varies over three orders of magnitude, as does the time scale over which the charge transfer occurs. Only a few studies consider the possible role of horizontal

TABLE 2 Current Speculations as to Characteristics of TLEs and Their Parent Lightning (None for Blue Jets)

	Sprite	Elve	Blue Jet
Color of emission	Red top/blue base	Red?	Deep blue
Polarity of parent CG	Positive (almost all)	Positive (mostly)	None
+CG peak current	$> \sim 50$ kA	$> \sim 100$ kA	None
Charge transferred (C)	Largest	Large	N/A
Charge moment (C-km)	Largest (>300)	Large (<300)	N/A
Parent CG location	Stratiform area	Stratiform area (?)	N/A
Parent CG vertical channel height	5–7 km (?)	10 km (?)	N/A
di/dt value	Moderate	Very large (?)	N/A
Total flash duration	Very long	Short (?)	N/A
Spider involved	Yes (?)	No (?)	N/A
Continuing current duration	Very long (?)	Short in any (?)	N/A
VLF/ELF slow tail	Distinct	Yes	N/A
ELF spectral color	Red	White	None
VLF audio character	Low freq.	Higher freq.	None
Duration of TLE	1–150 ms	0.5 ms	100–200 ms
Altitude range of TLE	25–95 km	75–105 km	Cloud-40 km
Onset delay after CG	1–100 ms	~0.3 ms	N/A
Brightness	50–35,000 kR	1000 kR	1000 kR
Horizontal size of emission	100 m–100 km	100–400 km	~1–2 km

components of the parent discharge. The charge moment (in C-km), not the peak current as measured by the NLDN, is the key parameter in the basic QE conventional breakdown mechanism first proposed by Wilson. The key physics of the problem appear to involve the altitude and magnitude of the removed charge and the time scale on which this occurs—parameters about which little agreement exists. Many theorists note that even with an assumed tall +CG channel (~10 km) this still requires extremely large (~100 C) charge transfers, typically 10 times larger than in "conventional" lightning. Some models yield a 1000-fold enhancement in optical intensity at 75 km for a doubling of the altitude of lightning charge removal from 5 to 10 km. The use of shorter channels to ground, say 5 km, would imply truly large charge transfers. Yet evidence is accumulating that indeed such may be the case.

While the various models simulate optical emissions bearing some (though in many cases rather minimal) resemblance to the observations, such wide ranges in the lightning source term parameters do not appear physically realistic. If, in fact, such a range of lightning characteristics could produce sprites, why does only a very small subset of +CGs (<1:20 even in active storms) actually produce observable TLEs (with current sensors)? It appears that most models have made assumptions about the lightning to produce something resembling a TLE—rather than starting with hard physical constraints on the source term. The reason, of course, is that there are very little data on the actual CGs that generate specific TLE occurrences.

During the summer of 2000, an extensive field research effort was conducted over the High Plains of Colorado, Kansas, and Nebraska. Several radars, research aircraft, and mobile storm monitor teams were deployed, along with a three-dimensional lightning-mapping array. As sprite-bearing storms passed over the LMA, LLTV systems at Yucca Ridge (some 300 km to the northwest) was able to document the sprites and other TLEs coincident with a variety of detailed measurements of the parent lightning discharges. Data analyses have just begun, but initial findings suggest that +CGs with very large charge moments ($>$300 C-km) are indeed associated with sprite events above their parent storms.

10 WHY STUDY TLEs?

Aside from their intrinsic scientific interest, there may be some rather practical reasons to explore TLEs in more depth. It has been suggested that there may be significant production of NO_x in the middle atmosphere by sprites. This becomes even more interesting in light of recent observations that regional smoke palls from biomass burns radically enhance the percentage of +CGs within storms, and thus increase sprite counts (and middle atmosphere NO_x production?). Currently, no global chemical model accounts for any potential effects of TLEs (Lyons and Armstrong, 1997). Once a better estimate of NO_x production per sprite is obtained, it will be necessary to know the global frequency and distribution of sprites. It has been demonstrated that several Schumann resonance monitoring sites working in tandem are capable of obtaining a worldwide TLE census.

There is growing interest in determining the sources of unusual infrasound emissions detected above sprite-capable MCSs as determined by the National Oceanic and Atmospheric Administration's (NOAA's) Environmental Technology Laboratory near Boulder, Colorado. TLEs thus produce optical, radio frequency (RF), and acoustic emissions that have the potential of mimicking or masking signatures from clandestine nuclear tests. Such findings may have important implications for global monitoring efforts supporting the Comprehensive Test Ban Treaty.

Transient luminous events may contribute in ways not yet understood to the maintenance of the global electrical circuit. To quantify the impacts of TLEs, we require information on their global frequency (now roughly estimated between 1 and 10 per minute) and their geographic distribution. It has been proposed (Williams, 1992) that the Schumann resonance can be used in the manner of worldwide tropical thermometer on the assumption that as global warming occurs, the amount of lightning may rise rapidly. The implications for sprite production of any global warming are uncertain.

ACKNOWLEDGMENTS

This material is based upon work supported in part with funds from the National Aeronautics and Space Administration (Office of Space Sciences), the U.S. Department of Energy, and the U.S. National Science Foundation (Contract number ATM-0000569).

11 SOME WEATHER AND LIGHTNING-RELATED WEBSITES

National Lightning Safety Institute: *http://www.lightningsafety.com*

Global Atmospherics, Inc.—National Lightning Detection Network: *http://www.LightningStorm.com*

Sprites and Elves: *http://www.FMA-Research.com*

Lightning Injury Research: *http://tigger.uic.edu/ ~ macooper/cindex.htm*

Kennedy Space Center/Patrick Air Force Base: *www.patrick.af.mil/45og/45ws/ws1.htm*

The National Severe Storms Laboratory—Norman, OK: *http://www.nssl.noaa.gov*

U.S. National Oceanic and Atmospheric Administration: *http://weather.noaa.gov* and *http://www.srh.noaa.gov/ftproot/ssd/html/lightnin.htm*

The World Meteorological Organization: *http://www.wmo.ch*

National Center for Atmospheric Research: *http://www.rap.ucar.edu/weather*

Federal Emergency Management Agency: *http://www.fema.gov*

NASA Marshall Space Flight Center: *http://thunder.msfc.nasa.gov*

REFERENCES

Andrews, C. J., M. Darveniza, and D. Makerras (1988). A review of medical aspects of lightning injuries, Proceedings, Intl Aerospace and Ground Conference on Lightning and Static Electricity, Oklahoma City, NOAA Special Report, OK 231–250.

Andrews, C. J., M. A. Cooper, M. Darveniza, and D. Mackerras (1992). *Lightning Injuries: Electrical, Medical and Legal Aspects*, Boca Raton, FL, CRC Press.

Armstrong, R. A., D. M. Suszcynsky, W. A. Lyons, and T. E. Nelson (2000). Multi-color photometric measurements of ionization and energies in sprites, *Geophys. Res. Lett.* **27**, 653–656.

Boccippio, D. J., E. R. Williams, S. J. Heckman, W. A. Lyons, I. T. Baker, and R. Boldi (1995). Sprites, ELF transients, and positive ground strokes, *Science* **269**, 1088–1091.

Boeck, W. L., O. H. Vaughan, Jr., R. Blakeslee, B. Vonnegut, and M. Brook (1998). The role of the space shuttle videotapes in the discovery of sprites, jets and elves, *J. Atmos. Solar-Terr. Phys.* **60**, 669–677.

Changery, M. J. (1981). National thunderstorm frequencies for the contiguous United States. U.S. Nuclear Regulatory Commission, NUREG/CR-22452, Washington, DC.

Cooper, M. A. (1983). Lightning injuries, *Em. Med. Clin. N. Am.* **3**, 639.

Cooper, M. A., (1995). Myths, miracles, and mirages, *Seminars in Neurology* **15**, 358–361.

Cummins, K. L., M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer (1998). A combined TOA/MDF technology upgrade of the U.S. National Lightning Detection Network, *J. Geophys. Res.* **103**, D8, 9035–9044.

Franz, R. C., R. J. Nemzek, and J. R. Winckler (1990). Television image of a large upward electrical discharge above a thunderstorm system, *Science* **249**, 48–51.

Fukunishi, H., Y. Takahashi, M. Kubota, K. Sakanoi, U. S. Inan, and W. A. Lyons (1996). Elves: Lightning-induced transient luminous events in the lower ionosphere, *Geophys. Res. Lett.* **23**, 2157–2160.

Fuquay, D. M., A. R. Taylor, R. G. Hawe, and C. W. Schmid, Jr. (1972). Lightning discharges that caused forest fires, *J. Geophys. Res.* **77**, 2156–2158.

Golde, R. H. (1977). *Lightning, Vol. 1, Physics of Lightning*, London, Academic.

Goodman, S. J., and D. R. MacGorman (1986). Cloud-to-ground lightning activity in mesoscale convective complexes, *Mon. Wea. Rev.* **114**, 2320–2328.

Hampton, D. L., M. J. Heavner, E. M. Wescott, and D. D. Sentman (1996). Optical spectra characteristics of sprites. *Geophys. Res. Lett.* **23**, 89–92.

Holle, R. H., and R. E. Lopez (1993). Overview of Real-time lightning detection systems and their meteorological uses, NOAA Technical Memorandum ERL NSSL-102, National Severe Storms Laboratory.

Holle, R. L., R. E. Lopez, R. Ortiz, A. I. Watson, D. L. Smith, D. M. Decker, and C. H. Paxton (1992). Cloud-to-ground lightning related to deaths, injuries and property damage in central Florida. Proceedings, Intl. Conf. on Lightning and Static Electricity, Atlantic City, NJ, FAA Report DOT/FAA/CT-92/20,66-1-66-12.

Holle, R. L., R. E. Lopez, L. J. Arnold, and J. Endres (1996). Insured lightning-caused property damage in three western states, *J. Appl. Meteor.* **35**, 1344–1351.

Holle, R. L., R. E. Lopez, and C. Zimmermann (1999). Updated recommendations for lightning safety—1998, *Bull. Am. Meteor. Soc.* **80**, 2035–2041.

Huang, E., E. Williams, R. Boldi, S. Heckman, W. Lyons, M. Taylor, T. Nelson, and C. Wong (1999). Criteria for sprites and elves based on Schumann resonance observations, *J. Geophys. Res.* **104**, 16943–16964.

Inan, U. S., C. Barrington-Lee, S. Hansen, V. S. Glukhov, T. F. Bell, and R. Rairden (1997). Rapid lateral expansion of optical luminosity in lightning-induced ionospheric flashes referred to as "elves," *Geophys. Res. Lett.* **24**, 583–586.

Krider, E. P., R. C. Noggle, A. E. Pifer, and D. L. Vance, 1980: Lightning direction-finding systems for forest fire detection, *Bull. Am. Meteor. Soc.* **61**, 980–986.

Lopez, R. E., R. L. Holle, T. A. Heitkamp, M. Boyson, M. Cherington, and K. Langford (1993). The underreporting of lightning injuries and deaths, Preprints, 17th Conf. on Severe Local Storms, Conf. on Atmospheric Electricity, St. Louis, American Meteorological Society, pp 775–778.

Lyons, W. A. (1994). Characteristics of luminous structures in the stratosphere above thunderstorms as imaged by low-light video, *Geophys. Res. Letts.* **21**, 875–878.

Lyons, W. A. (1996). Sprite observations above the U.S. High Plains in relation to their parent thunderstorm systems, *J. Geophys. Res.* **101**, 29641–29652.

Lyons, W. A., and R. A. Armstrong (1997). NO_x Production within and above Thunderstorms: The contribution of lightning and sprites, Preprints, 3rd Conf. on Atmospheric Chemistry, Long Beach, American Meteorological Society.

Lyons, W. A., and E. R. Williams (1993). Preliminary investigations of the phenomenology of cloud-to-stratosphere lightning discharges. Preprints, Conference on Atmospheric Electricity, St. Louis, American Meteorological Society, pp 725–732.

Lyons, W. A., K. G. Bauer, R. B. Bent, and W. H. Highlands (1985). Wide area real-time thunderstorm mapping using LPATS—the Lightning Position and Tracking System, Preprints, Second Conf. on the Aviation Weather System, Montreal American Meteorological Society.

Lyons, W. A., K. G. Bauer, A. C. Eustis, D. A. Moon, N. J. Petit, and J. A. Schuh (1989). R·Scan's National Lightning Detection Network: The first year progress report. Preprints, Fifth Intl. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, Anaheim, American Meteorological Society.

Lyons, W. A., M. Uliasz, and T. E. Nelson (1998). Climatology of large peak current cloud-to-ground lightning flashes in the contiguous United States, *Mon. Wea. Rev.* **126**, 2217–2233.

MacGorman, D. R., and W. D. Rust (1998). *The Electrical Nature of Storms*, New York, Oxford University Press.

Mende, S. N., R. L. Rairden, G. R. Swenson, and W. A. Lyons (1995). Sprite spectra: N2 first positive band identification, *Geophys. Res. Lett.* **22**, 2633–2636.

National Research Council (1986). *The Earth's Electrical Environment*, Studies in Geophysics, Washington, DC, National Academy Press.

Orville, R. E., and A. C. Silver (1997). Annual Summary: Lightning ground flash density in the contiguous United States: 1992–95, *Mon. Wea. Rev.* **125**, 631–638.

Pasko, V. P., U. S. Inan, Y. N. Taranenko, and T. F. Bell (1995). Heating, ionization and upward discharges in the mesosphere due to intense quasi-static thundercloud fields, *Geophys. Res. Lett.* **22**, 365–368.

Rakov, V. A., and M. A. Uman (1990). Some properties of negative cloud-to-ground lightning flashes versus stroke order, *J. Geophys. Res.* **95**, 5447–5453.

Rowland, H. L. (1998). Theories and simulations of elves, sprites and blue jets, *J. Atmos. and Solar-Terrestrial Phys.* **60**, 831–844.

Sentman, D. D., and E. M. Wescott (1993). Observations of upper atmospheric optical flashes recorded from an aircraft, *Geophys. Res. Lett.* **20**, 2857–2860.

Thottappillil, R., V. A. Rakov, M. A. Uman, W. H. Beasley, M. J. Master, and D. V. Sheluykhin (1992). Lightning subsequent-stroke electric field peak greater than the first stroke peak and multiple ground terminations, *J. Geophys. Res.* **97**, 7503–7509.

Uman, M. A. (1987). *The Lightning Discharge*, International Geophysics Series, Vol. 39, Orlando, Academic.

Uman, M. A., and E. P. Krider (1989). Natural and artificially initiated lightning, *Science* **246**, 457–464.

U.S. Dept. of Energy (1996). *Lightning Safety*, Office of Nuclear and Facility Safety, USDOE, DOE/EH-0530, Washington, DC.

Wescott, E. M., D. Sentman, D. Osborne, D. Hampton, and M. Heavner (1995). Preliminary results from the Sprites94 aircraft campaign: 2. Blue jets, *Geophys. Res. Lett.* **22**, 1209–1212.

Williams, E. R. (1988). The electrification of thunderstorms, *Sci. Am.* **259**, 88–99.

Williams, E. R. (1992). The Schumann resonance: A global tropical thermometer, *Science* **256**, 1184–1186.

Williams, E. R. (1998). The positive charge reservoir for sprite-producing lightning, *J. Atmos. Solar-Terr. Phys.* **60**, 689–692.

Wilson, C. T. R. (1956). A theory of thundercloud electricity, *Proc., Royal Met. Soc., London* **236**, 297–317.

CHAPTER 23

WEATHER MODIFICATION

HAROLD D. ORVILLE

1 INTRODUCTION

Although much remains to be learned about the seeding of clouds, scientists and engineers have accomplished much in the 55 or so years since the discovery that dry ice and silver iodide are efficient ice nucleants, able to influence the natural precipitation processes. In addition, within the past 10 years renewed efforts using hygroscopic seeding materials have yielded very promising results in producing more rain from convective clouds. This chapter will discuss the basic physics and chemistry of the precipitation and hail processes and explain the primary methods for changing the various atmospheric precipitation components. Some of the key advances in weather modification in the past 25 years will be highlighted. The primary source material comes from a workshop report to the Board on Atmospheric Sciences and Climate (BASC) of the National Research Council (NRC) published in December 2000 (BASC, 2000). In addition, the most recent statement by the American Meteorological Society (AMS, 1998a) and the scientific background for the policy statement (AMS, 1998b) has much good information concerning the current status of weather modification.

Basic Physics and Chemistry of the Precipitation Processes

Only a brief review of these complex processes will be given here. Detailed treatments can be found in the textbooks or scientific studies listed in the references at the end of this chapter (e.g., Dennis, 1980; Rogers and Yau, 1989; Young, 1993). Concerning rain and snow, the basic problem of precipitation physics is: How do one million cloud droplets combine to form one raindrop or large snowflake in a

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

period of 10 to 20 min? Two primary processes are involved after the clouds have formed.

Atmospheric aerosols are important for the formation of clouds, the precursors of rain. The chemical content and distribution of the aerosol plus the cloud updraft speed determine the initial distribution and number concentration of the droplets. Normally there are ample suitable hygroscopic aerosols, called cloud condensation nuclei (CCN), for clouds to form at relative humidities very near 100% (at supersaturations of a few tenths of a percent only). The relative humidity with respect to a liquid or ice surface is defined as the ratio of the environmental vapor pressure to the saturation vapor pressure at the liquid or ice surface, respectively. Maritime and continental regions have very different aerosol populations, leading to characteristically different cloud droplet distributions.

For clouds with all liquid particles (certainly all those clouds warmer than 0°C) a process known as collection (a combination of collision and coalescence)—the *warm rain process*—is thought to be operative. This process depends on the coexistence of cloud droplets of different sizes and, consequently, of different fall velocities. Sizes range from a few micrometers in diameter to 30 or 40 μm. The larger droplets fall through a population of smaller droplets, collecting them as they fall. The larger particles become raindrop embryos. If conditions are right (depth of cloud, breadth of droplet distribution, cloud updraft, etc.), the embryos can form raindrops by this process in the necessary time period. Raindrops range in size from 200 μm to 5 mm diameter (the upper limit occurs because of raindrop breakup).

One might think that with billions and billions of cloud droplets in a cloud that it would be relatively easy for the droplets to collide, coalesce, and grow to raindrop size. However, even though very numerous, the droplets occupy less than one millionth of the volume of the cloud. Consequently, the droplets are relatively far apart and require special conditions to interact and grow. Indeed, even large thunderstorms and massive raining nimbostratus clouds are more than 99.99% clear space (dry air and vapor), but by the time precipitation particles have formed in such clouds the collection process is well developed and relatively efficient.

The second precipitation process is called the *cold rain process* and depends on the formation of ice in the cloud. This process also depends on nuclei (ice nuclei, very different particles from the CCN mentioned above), but in this case there is a shortage of ice-forming nuclei in the atmosphere so that liquid water does not freeze at 0°C. The water is then called *supercooled*. To understand this cold rain process, one other important fact is needed: For the same subzero temperature, the relative humidity will be much above 100% for the ice particle while staying at 100% over the liquid surface, due to the fact that the saturation vapor pressure is higher over the liquid surface than over the ice surface. The growth of the particles depends on the vapor pressure difference between the cloudy air and the particles' surface. The liquid droplets are generally much more numerous than ice crystals in a cloud and control the relative humidity (the vapor pressure in the cloudy air). The stage is then set for rapid growth of an ice particle if it is introduced into a cloud of supercooled liquid droplets. The ice crystal grows, depleting the water vapor in the cloudy environment, which is then replenished by water vapor from evaporating droplets trying to maintain 100% relative humidity with respect to liquid. The crystal rapidly increases in mass and falls through the cloud droplets, collecting them and further increasing the ice particle's mass. Snow or rain will appear at the ground depending on the temperature conditions of the lower atmosphere. Bergeron and Findeisen first described this process in the 1930s.

The formation of the ice particle by the ice nucleus can occur in at least four ways: (1) Deposition of water vapor molecules may occur directly on the nucleus. (2) Condensation may occur on a portion of the nucleus, followed by freezing of the liquid. (3) A drop or droplet may come in contact with a nucleus and freeze. (4) And finally a drop or droplet may have formed including a nucleus in its bulk water content, which then freezes the water when a low enough temperature is reached. How low? All of the processes mentioned above require a supercooling of from 15 to 20°C, that is, temperatures of −15 to −20°C in continental-type clouds (those clouds with a narrow drop size distribution and large number concentrations). Clouds with extremely strong updrafts may delay their primary ice formation to temperatures near −40°C (Rosenfeld and Woodley, 2000). There is some evidence accumulating that the ice nucleating temperature may depend on the type of cloud; maritime clouds with large and fewer droplets and a broad droplet distribution produce ice particles at temperatures as warm as −10°C. The more frequent collisions among large and small droplets may contribute to the early formation of ice in those clouds.

As mentioned above, in the absence of any ice nuclei, the water will remain in the liquid state until a temperature of −40°C is reached. The most effective nuclei are insoluble soil particles, kaolinites, and montmorillonites, but industrial and natural combustion products may also contribute to the ice-forming characteristics of clouds.

These ice processes refer to primary ice formation in a cloud. Secondary processes may also contribute (Mossop, 1985). Ice crystals may fracture or graupel, and droplets may interact to form additional ice particles, or ice particles may be transported from one cloud cell to another. A stormy, turbulent cloud situation has many ways to create and distribute ice among the clouds.

Basic Hail Processes

Hailstorms are normally large thunderstorms that produce hail. Hailstones are irregular ice particles larger than 5 mm in diameter (about one quarter of an inch). Large hailstones may grow to several centimeters in diameter and fall at many tens of meters per second. Hailstorms are characterized by large updrafts, usually greater than 15 m/s (30 mph) and sometimes as great as 50 m/s (100 mph). The large updrafts in a storm are closely associated with the largest hailstones in a storm. Such storms produce more than a billion dollars of crop damage and a like amount of property damage per year in the United States.

In addition to large updrafts to suspend the growing hailstones, a two-stage process is thought needed to produce hailstones. The first stage requires the formation of hailstone embryos. Either frozen raindrops or graupel particles may serve as the embryos. They range in size from 1 to 4 or 5 mm. The second stage involves the growth of the embryo to hailstone size. This requires the presence of ample supercooled water at sufficiently cold temperatures to grow the hailstone. The most efficient growth occurs between the temperatures of -10 to -30°C, at elevations of 5 to 8 or 9 km in the atmosphere. Unfortunately, the source regions of the embryos are not well known and can come from several locations in a storm. This makes it difficult to modify the hailstone growth process. This two-stage scenario indicates that the microphysical and dynamical processes in a storm must be matched in a special way to produce hail.

2 MODIFICATION METHODS

Rain and Snow

Knowledge and understanding of the precipitation processes lead to finding ways of modifying the processes. The two techniques used for rain and snow increases for cold clouds are called *microphysical* (or *static*) and *dynamic* seeding methods. The first method rests on the assumption that clouds lack the proper number of rain or ice embryos for the maximum precipitation to occur. Hence, great emphasis is placed on increasing the precipitation efficiency of the cloud, making sure that as much as possible of the cloud water is converted to precipitation. To affect the ice process, artificial nuclei, such as silver iodide, can be added or powdered dry ice (as cold as -80 to -100°C) can be dropped through the cloud, forming ice crystals by both homogeneous and heterogeneous nucleation. (Homogeneous freezing refers to the change of phase of water from vapor to ice without the assistance of a nucleus. Heterogeneous freezing occurs with the aid of a nucleus, which allows freezing at warmer temperatures than in homogeneous freezing.) The most striking effect of this microphysical seeding mode is the production of precipitation from marginal-type clouds, those on the brink of producing precipitation. However, the most effective precipitation increase may come from treating more vigorous but inefficiently raining or snowing clouds.

The concept of the second process, dynamic seeding, is that seeding a supercooled cloud with large enough quantities of artificial ice nuclei or a coolant such as solid carbon dioxide (dry ice) will cause rapid glaciation of the cloud. The resultant latent heat release from the freezing of supercooled drops (modified by the adjustment of saturation with respect to ice instead of with respect to liquid) will then increase the buoyancy of a cloud. The increased vigor of the cloud may result in a taller cloud, stronger updraft, more vapor and water flux, broader and longer lasting rain area, stronger downdrafts, and greater interaction with neighboring clouds, resulting in more cloud mergers and, hopefully, more rain. The reasoning sounds convincing, but there are many junctures where the process may go astray, and much research is needed to verify the hypothesis. Recent field studies of this seeding process in convective cloud systems in Texas (Rosenfeld and Woodley, 1993) have shown nearly twice as many mergers in the seeded complexes compared with the unseeded cloud complexes.

Theoretical results have indicated that dynamic seeding can also affect relatively dry stratus-type clouds as well as wet cumulus clouds (Orville et al., 1984, 1987). The action of the seeding is to cause embedded convection in the stratus clouds. One of the challenges with both of these seeding methods is to identify those clouds and cloud environments that will respond positively to the seeding.

The amount of seeding agent to use depends on the type of seeding agent, the seeding method, and the vigor of the cloud. One gram of silver iodide can produce about 10^{14} nuclei when completely activated. The activation starts at -4°C and becomes more efficient as the atmosphere cools, about one order of magnitude more activated nuclei for each 4° drop in temperature.

Normally the goal is to supply about one ice nucleus per liter of cloudy air to affect the precipitation process through the microphysical seeding method. Up to 100 times this number may be needed for the dynamic seeding method to be effective. The total amount of seeding material needed depends on the volume of cloud to be seeded, which will depend on the strength of the updraft. In general, a few tens of grams of silver iodide are needed for microphysical seeding and about 1 kg or more for dynamic seeding. Dry ice produces about 10^{12} ice crystals per gram of sublimed CO_2 in the supercooled portion of the cloud. Normally about 100 g of dry ice are used per kilometer of aircraft flight path for the microphysical seeding method, with only about 10 to 20% of the dry ice being sublimed before falling below the 0°C level in the cloud.

There are other determinants for the growth of precipitation-sized particles that can, in at least some cases, be artificially influenced. In the case of the collision–coalescence mechanism, opportunities sometimes exist for providing large hygroscopic nuclei to promote initial droplet growth or much larger salt particles to form raindrop embryos directly. The purpose of the hygroscopic seeding is thus to produce precipitation particles either directly or by enhancing the collision–coalescence mechanism. Two salt seeding methods are currently in use. One method applies hundreds of kilograms of salt particles (dry sizes are 10 to 30 μm in diameter) near cloud base to produce drizzle-size drops very soon after the salt particles enter the cloud. The second method uses salt flares to disperse 1 μm or smaller size particles into cloud updrafts, a method currently receiving renewed interest in cloud seeding efforts (Mather et al., 1996, 1997; Cooper et al., 1997; Bigg, 1997; Tzivion et al., 1994; Orville et al., 1998). The salt material is released from kilogram-size flares carried by aircraft; several flares are released per cloud cell. The salt particles change the size distribution of CCN in the updraft, creating a more maritime-type cloud. Coalescence is enhanced; rain forms in the seeded volume, eventually spreading throughout the cloud. This seeding thus accelerates the warm rain process. In addition, if the updraft lifts the rain to high enough altitudes, then the ice processes are also enhanced because of the larger drops and droplets, making it more likely that freezing of some drops will occur. Graupel and snow are more easily formed, increasing the precipitation efficiency of the cell. These hygroscopic seeding methods are thought to work only on continental-type clouds.

In the past, large water drops were added to clouds to accelerate the rain process, but this is not an economically viable way to increase warm rain.

Hail

The suppression of hail from a hailstorm is a much more complex matter than the increase of rain or snow from more benign cloud types. Two concepts appear to offer the most hope for the suppression of hail. They are called *beneficial competition* and *early rainout*. To suppress hail according to the beneficial competition theory, many more hail embryos must be introduced into the cloud cell than would occur naturally. (Cloud seeding with silver iodide is one way to provide the additional embryos.) According to this hypothesis, the sharing of the available supercooled water among a larger number of hailstones (i.e., their "competition" for the available water) reduces their size. If enough embryos are present, it should be possible to reduce the local supercooled liquid water content and hence the hailstone growth rates so that no particle grows large enough to survive without melting during fallout to the ground. Thus, less hail and more rain would be produced by the storm.

The consensus of many scientists is that this concept is most promising in storms containing large supercooled drops, since these drops when frozen (caused by the seeding) are large enough to act as efficient collectors and to provide the necessary competition. In storms containing only supercooled cloud droplets, this possibility is absent, and there is greater difficulty in creating graupel embryos in the correct place, time, and amount. Also, note that some storms may be naturally inefficient for the production of hail and the addition of potential hail embryos may actually increase hail production.

Creating early rainout from the feeder cells of an incipient hailstorm also appears to be an attractive method to suppress hail. Cloud seeding starts the precipitation process earlier than it would naturally. The weak updrafts in the feeder cells cannot support the rain particles and they fall out without participating in the hail formation process. This premature rainout removes liquid water and is accompanied by a reduction of the updraft strength due to the downward force components caused by both water loading in the lower part of the cloud and negative buoyancy caused by cooling resulting from melting of ice particles and the evaporation of precipitation beneath the cloud. These processes are thought to inhibit the hail generation process. Some operational projects use this method with encouraging results.

3 SOME SCIENTIFIC AND TECHNOLOGICAL ADVANCES IN THE PAST 25 YEARS

Weather modification as a scientific endeavor was last reviewed on a national basis in the late 1970s. At that time the Weather Modification Advisory Board (WMAB), established by an act of Congress in 1976, produced a report and a two-volume supporting document, detailing the status of weather modification in both research and operations (WMAB, 1978). Nearly $20 million per year was being spent on federal research in weather modification at the time of the Congressional act. An ambitious program was proposed to continue development of the technology, but little action was taken. Weather modification research funding dwindled to nearly zero in the ensuing years. However, operational weather modification continued and in recent years has shown signs of invigoration. Much of the operational work is being done in large areas of the midwestern and western United States (see Fig. 1).

Most of the following material comes from the BASC workshop summary (BASC, 2000).

Hygroscopic Seeding of Convective Clouds

An exciting breakthrough in the development of rain augmentation technology was made within the past decade. Three randomized experiments in three different parts of the world showed that hygroscopic seeding increased rainfall from convective clouds. Statistically significant increases in radar-estimated rainfall were achieved by hygroscopic flare seeding of cold convective clouds in South Africa, its replication on cold convective clouds in Mexico, and by hygroscopic particle seeding of warm convective clouds in Thailand. These efforts included physical studies as well as statistical experiments and resulted in strong evidence indicating that the increases in rainfall were due to seeded clouds lasting longer and producing rain over a larger area.

The common elements of the three randomized experiments were: (1) seeding with hygroscopic particles, (2) evaluation using a time-resolved estimate of storm rainfall based on radar measurements in conjunction with an objective software package for tracking individual storms (different software was used for each experiment), (3) statistically significant increases in radar-estimated rainfall, and (4) the necessity to invoke the occurrence of seeding-induced dynamic effects to explain the results.

The great significance of the Mexican experiment was that it replicated the results of the South African experiment in another area of the world. Several past programs in the world failed when attempts were made to replicate previously successful programs. Figure 2 displays a comparison of the quartile values of radar-derived rain mass in seeded (solid lines) and nonseeded (dashed lines) cases for the three different quartiles as a function of time after “decision time” (the time at which the treatment decision was made) for the South African (dark lines) and Mexican (gray lines) experiments.

The Mexican data were averaged over 5-min periods, while the South African data were averaged over 10 min. To be consistent, the combined results are plotted in Figure 2 as radar-derived “rain mass accumulated per minute.” The indicated differences in rain mass for both the South African and Mexican experiments were statistically significant after 20 to 30 min and remained significant for the remainder of the period. It is clear that the results from both experiments are in good agreement. The main difference is that the storms in South Africa tended to last somewhat longer than those in Mexico.

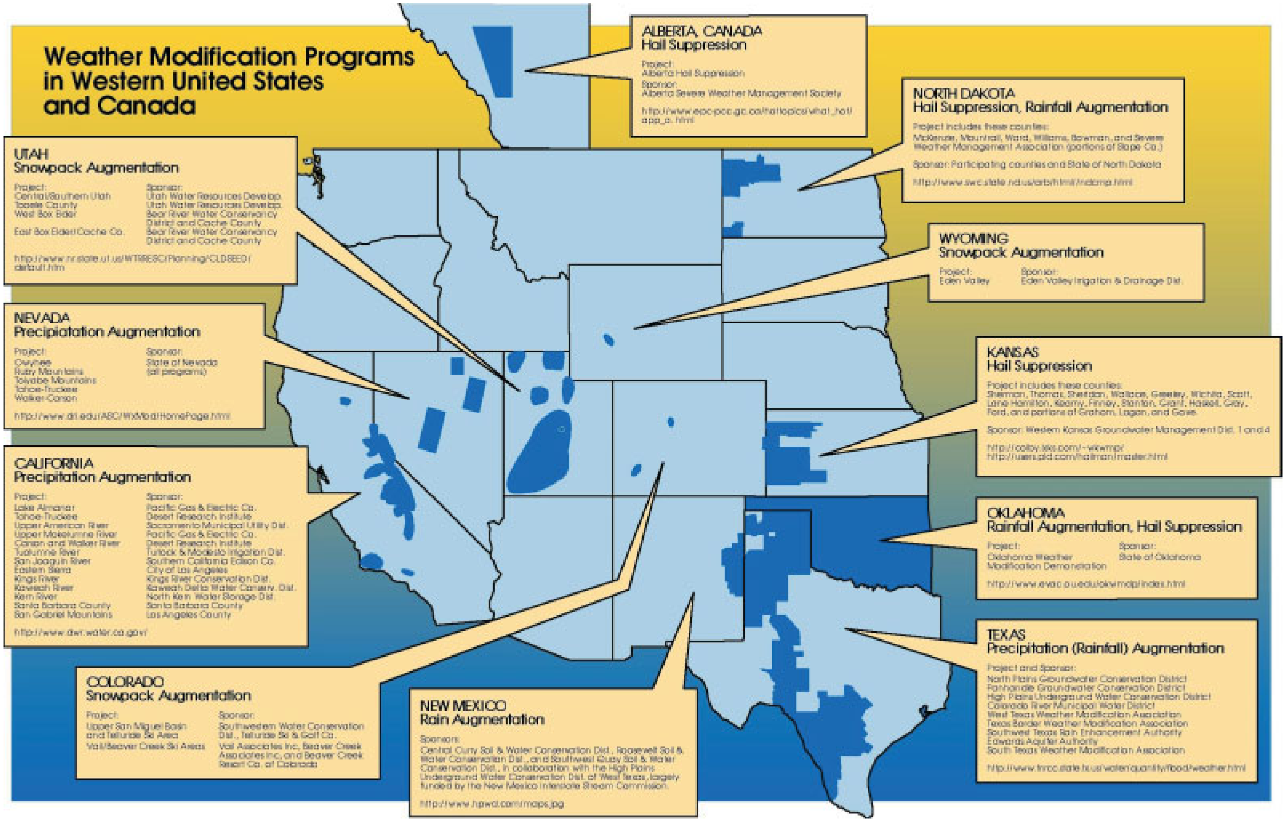

Weather Modification Programs in Western United States and Canada
ALBERTA, CANADA
Hail Suppression
Project:
Alberta Hail Suppression
Sponsor:
Alberta Severe Weather Management Society
http://www.epc-pcc.gc.ca/hailtopics/what_hail/app_a.html
NORTH DAKOTA
Hail Suppression, Rainfall Augmentation
Project includes these counties:
McKenzie, Mountrail, Ward, Williams, Bowman, and Severe Weather Management Association (portions of Slope Co.)
Sponsor: Participating counties and State of North Dakota
http://www.swc.state.nd.us/arb/html/ndcmp.html
WYOMING
Snowpack Augmentation
Project:
Eden Valley
Sponsor:
Eden Valley Irrigation & Drainage Dist.
KANSAS
Hail Suppression
Project includes these counties:
Sherman, Thomas, Sheridan, Wallace, Greeley, Wichita, Scott, Lane Hamilton, Kearny, Finney, Stanton, Grant, Haskell, Gray, Ford, and portions of Graham, Logan, and Gove.
Sponsor: Western Kansas Groundwater Management Dist. 1 and 4
http://colby.ixks.com/~wkwmp/
http://users.pld.com/hallman/master.html
OKLAHOMA
Rainfall Augmentation, Hail Suppression
Project:
Oklahoma Weather Modification Demonstration
Sponsor:
State of Oklahoma
http://www.evac.ou.edu/okwmdp/index.html
TEXAS
Precipitation (Rainfall) Augmentation
Project and Sponsor:
North Plains Groundwater Conservation District
Panhandle Groundwater Conservation District
High Plains Underground Water Conservation District
Colorado River Municipal Water District
West Texas Weather Modification Association
Texas Border Weather Modification Association
Southwest Texas Rain Enhancement Authority
Edwards Aquifer Authority
South Texas Weather Modification Association
http://www.tnrcc.state.tx.us/water/quantity/flood/weather.html
NEW MEXICO
Rain Augmentation
Sponsors:
Central Curry Soil & Water Conservation Dist., Roosevelt Soil & Water Conservation Dist., and Southwest Quay Soil & Water Conservation Dist., in collaboration with the High Plains Underground Water Conservation Dist. of West Texas, largely funded by the New Mexico Interstate Stream Commission.
http://www.hpwd.com/maps.jpg
COLORADO
Snowpack Augmentation
Project:
Upper San Miguel Basin and Telluride Ski Area
Vail/Beaver Creek Ski Areas
Sponsor:
Southwestern Water Conservation Dist., Telluride Ski & Golf Co.
Vail Associates Inc, Beaver Creek Associates Inc, and Beaver Creek Resort Co. of Colorado
CALIFORNIA
Precipitation Augmentation
Project:
Lake Almanor
Tahoe-Truckee
Upper American River
Upper Mokelumne River
Carson and Walker River
Tuolumne River
San Joaquin River
Eastern Sierra
Kings River
Kaweah River
Kern River
Santa Barbara County
San Gabriel Mountains
Sponsor:
Pacific Gas & Electric Co.
Desert Research Institute
Sacramento Municipal Utility Dist.
Pacific Gas & Electric Co.
Desert Research Institute
Turlock & Modesto Irrigation Dist.
Southern California Edison Co.
City of Los Angeles
Kings River Conservation Dist.
Kaweah Delta Water Conserv. Dist.
North Kern Water Storage Dist.
Santa Barbara County
Los Angeles County
http://www.dwr.water.ca.gov/
NEVADA
Precipiatation Augmentation
Project:
Owyhee
Ruby Mountains
Toiyabe Mountains
Tahoe-Truckee
Walker-Carson
Sponsor:
State of Nevada
(all programs)
http://www.dri.edu/ASC/WxMod/HomePage.html
UTAH
Snowpack Augmentation
Project:
Central/Southern Utah
Tooele County
West Box Elder
East Box Elder/Cache Co.
Sponsor:
Utah Water Resources Develop.
Utah Water Resources Develop.
Bear River Water Conservancy District and Cache County
Bear River Water Conservancy District and Cache County
http://www.nr.state.ut.us/WTRRESC/Planning/CLDSEED/default.htm

Figure 1 Active cloud seeding projects in the mid-western and western United States and Canada in the year 2000. (Figure supplied by Bruce Boe, Weather Modification, Inc.). See ftp site for color image.

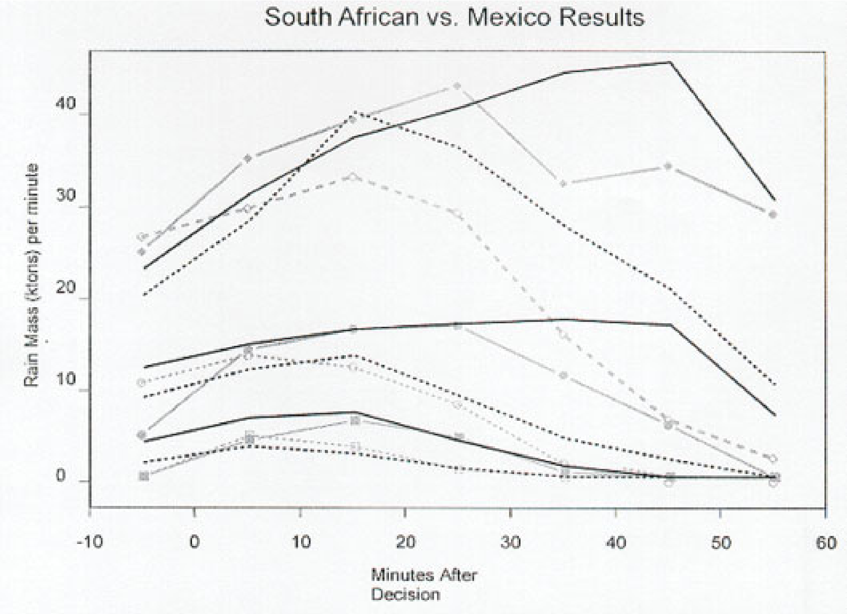


Figure 2 Quartile values of radar-derived rain mass (precipitation flux integrated over 1 min) versus time after "decision time" for seeded (solid lines) and nonseeded (dashed lines) cases for the South African (dark lines) and Coahuila (gray lines) experiments. First quartile is the value of rain mass that is larger than the value for 25% of the storms; second quartile value exceeds the value for 50% of the other storms; and third quartile value exceeds the value for 75% of the storms. (Figure supplied by Roelof Bruintjes, National Center for Atmospheric Research).

Glaciogenic Seeding Effects in Convective Clouds

The seeding results on these clouds are more controversial and more difficult to document (Rangno and Hobbs, 1995, 1997; Silverman, 2001; but also note Nicholls, 2001). Strong glaciogenic seeding signatures, which appear 2 to 3 min after initial seeding, have been documented in treated clouds by aircraft (Woodley and Rosenfeld, 2000) and, more recently, from space. Glaciation times after seeding are halved in both maritime and continental seeded clouds. These are consistent with the conceptual model of dynamic seeding guiding the glaciogenic seeding experiments. In addition, there is now limited field evidence and some numerical modeling results that indicate ice-phase seeding invigorates the internal cloud circulation concurrent with the growth of graupel to precipitation size and the depletion of the cloud water. These too are consistent with expectations from the conceptual model.

Radar estimation of the properties of treated and nontreated convective cells shows indications that under certain conditions the seeded cells produce more rainfall as shown by the increasing maximum radar reflectivities, maximum areas, maximum rain-volume rates, duration, clustering and merging of cells, and inferred maximum rainfall rates. The indicated effects on rainfall quantities require further validation by direct surface measurements. Earlier experimentation had also indicated that increases in the top height of the treated clouds accompanied the apparent rainfall increases. In recent years, however, this apparent height signal has been much weaker, due probably to a change in the method of estimating cloud top. In the early years, cloud tops were measured directly with a jet aircraft, while radar has been used in later years to estimate cloud top. Because radar may underestimate the tops of seeded clouds relative to nonseeded clouds as a consequence of the seeding-induced glaciation (glaciated clouds are less radar reflective than unglaciated clouds in the absence of hail), satellite measurements of cloud tops must be used to recheck this link in the conceptual chain.

Snowpack Augmentation

A significant accomplishment in recent snowpack augmentation research is the establishment of the direct link between the seeding activity and the water reaching the ground in the form of snow. The increases in precipitation rate caused by silver iodide seeding have been documented several times in the reviewed scientific literature (Reynolds, 1988; Super and Holroyd, 1997). The link has been established by physical and chemical techniques. The snow falling at particular targeted sites is connected directly to the seeding material (see Fig. 3).

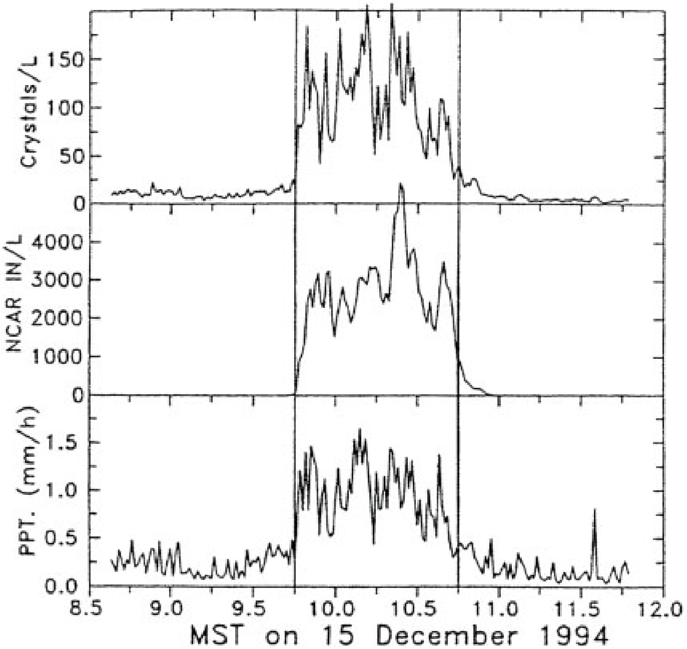


Figure 3 Observed concentrations of precipitating ice crystals, ice nuclei, and precipitation rate during one hour of AgI seeding between 0945 and 1045, December 15, 1994 in Utah (Super and Holroyd, 1997).

The methodologies used to establish this direct chemical linkage have been described by Warburton et al. (1985, 1994, 1995a, b), Super and Heimbach (1992), Chai et al. (1993), Stone and Huggins (1996), Super and Holroyd (1997), and McGurty (1999).

One big advantage of snowpack work is that the scientists are dealing with solid-state precipitation that can be sampled in fixed and targeted areas both during and after storm events and stored in the frozen state until analyzed.

Hail Suppression

In recent years, crop hail damage in the United States has typically been around $2.3 billion annually (Changnon, 1998). Susceptibility to damage depends upon the crop type, the stage of development, the size of the hail, and also the magnitude of any wind accompanying the hailfall.

Property damage from hail in recent years has been on the same order as crop hail damage, usually topping the $2 billion mark, sometimes more. A recent report by the Institute for Business and Home Safety (IBHS, 1998) indicated that losses from wind storms involving hail, from June 1, 1994 through June 30, 1997, totaled $13.2 billion. While some of this damage resulted from wind, hail certainly accounted for a significant fraction of the total damage.

Some of the recent high-dollar hailstorms include: Denver, 1990, $300 million; Calgary, 1991, $350 million; Dallas–Fort Worth, 1992, $750 million; Bismarck, ND, 1993, $40 million; Dallas–Fort Worth, 1995, $1 billion; and Calgary, 1996, $170 million. Wichita (Kansas), Orlando (Florida), and northern (Arlington) Virginia are just a few of the other U.S. locales that have recently been hard hit by hailstorms.

Results from North American hail suppression programs vary. The North Dakota Cloud Modification Project (NDCMP) reports reductions in crop hail damage on the order of 45% (Smith et al., 1997), while the Western Kansas Weather Modification Program (WKWMP) reports reduced crop hail losses of 27% (Eklund et al., 1999). Neither program reports any statistically significant changes in rainfall.

Both of these projects are operational, nonrandomized programs, and the evaluations are based upon analyses of crop hail insurance data. Projects elsewhere in the world (e.g., Dessens, 1986) have generally reported similar reductions in damage.

Considerable success has been achieved using numerical cloud models to simulate hailstorms and hail development (Farley et al., 1996; Orville, 1996). This has contributed significantly to the development of the contemporary conceptual models for hail suppression. Contemporary numerical models contain microphysical components, as well as cloud dynamics.

If a cloud model, after programming with actual atmospheric conditions, can successfully reproduce a cloud or storm like that actually observed in those same atmospheric conditions, the model has reinforced the physical concepts employed therein. Such cloud models can be used to test concepts for hail suppression. If the model employing a certain concept gets it right, this strengthens the confidence in the concept.

It is impossible to find two identical clouds, seed one, and leave the other untreated as a control, since no two clouds are exactly the same. However, the effects of seeding can be examined by modeling a cloud beginning with natural conditions and simulating seeding in a second run. Any differences in behavior can then be attributed to the seeding. In fact, this option is very attractive, as the timing (relative to the life cycle of the subject cloud), locations, and amounts of seeding can be varied in a succession of model runs to better understand the importance and effects of targeting.

Figure 4 gives an example of numerical model output, showing a decrease in hail in the large sizes and an increase in graupel size particles due to silver iodide cloud seeding. Total hailfall was decreased by 44%, hail impact energy by 58% in the seeded case as compared with the nonseeded case.

Advances in Technology

Instrumentation Unusual opportunities for determining, *by direct measurement*, the physical processes and water budgets of cloud systems and their changes due to purposeful and inadvertent cloud modification reside with new technologies, particularly remote-sensing technologies. Many other instruments and techniques have been developed that are applicable, including satellite systems and in situ systems including aircraft platforms. However, the main point to be noted is that *none* of these technologies were available during the period of about 1965–1985 when significant funding was directed toward weather modification research. There has been a paradigm shift in observational capabilities, and cloud modification can now be revisited with the new and emerging tools.

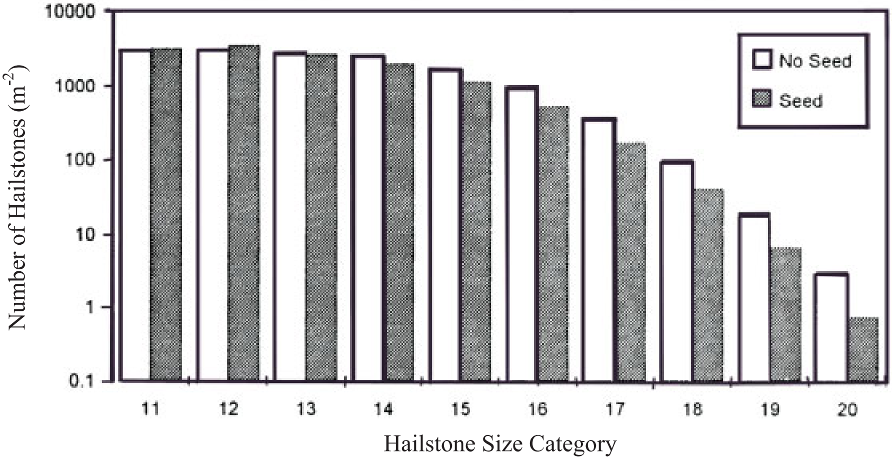


Figure 4 Size distributions of graupel and hail particles striking the ground (at the region of maximum hailfall from computer-simulated hailstorm). Categories 11, 14, 17, 20 represent 2-, 6-, 16-, 42-mm hailstones, respectively (Farley et al., 1996).

An abbreviated list of significant tools that have advanced our ability to observe cloud systems is summarized in Table 1. This table indicates that the microphysical and dynamic history of seeded and nonseeded clouds can now be documented and results compared between the two types of clouds.

Cloud Seeding Agents Practical capabilities have been established over the past 50 years, and especially since about 1980, for generating highly effective ice nucleating aerosols with well-characterized behaviors for both modeling and detecting their atmospheric fate after release for cloud modification. In the same period, our understanding of natural ice nucleation and our prospects for further elucidating ice formation processes have improved. Improvements in understanding how hygroscopic aerosols interact with clouds and how to use them to increase precipitation have also occurred in recent years.

By 1980, a clear need was recognized to account for the complex mechanisms of ice formation by specific ice nuclei used in weather modification field programs. In particular, the chemical and physical properties of aerosols were established to be very important in determining ice formation rates as well as efficiency. This recognition led to the development of new, highly efficient silver chloro-iodide ice nuclei (DeMott et al., 1983). These same nuclei can be generated with a soluble component to enhance the action of a fast condensation-freezing ice nucleation mechanism (Feng and Finnegan, 1989). The development of similar, fast-acting and highly efficient ice nuclei from pyrotechnic generation methods followed by the early 1990s (Fig. 5). These new nucleating agents represent substantial improvements

TABLE 1 Some Remote Sensors for Hydrometeorology*

Winds, air motions in clouds, clear air	Doppler radar, wind profiling radar, Doppler lidar
Temperature profiles	Radio acoustic sounding system (RASS)
Water vapor, cloud liquid water	Microwave radiometer
Cloud boundaries	Cloud (mm-wave) radar
Water vapor and liquid fluxes and profiles	(Combinations of the above)
Thermodynamic and microphysical cloud structures	"Retrievals" from Doppler radar data
Cloud phase, ice hydrometeor type and evolution	Dual-polarization cloud and precipitation radars, microwave radiometer, polarization microwave radiometer, polarization multi-field-of-view lidar
Transport and dispersion, mixing in clouds	Chaff wind tracking dual-circular-polarization radar, gaseous tracers
Precipitation trajectories	Chaff wind tracking and Doppler radar
Snowflake size, snowfall areal rate	Dual-wavelength radar
Rainfall rate, hail differentiation	Dual-polarization radar

*Table provided by Roger Reinking, National Oceanic and Atmospheric Administration.

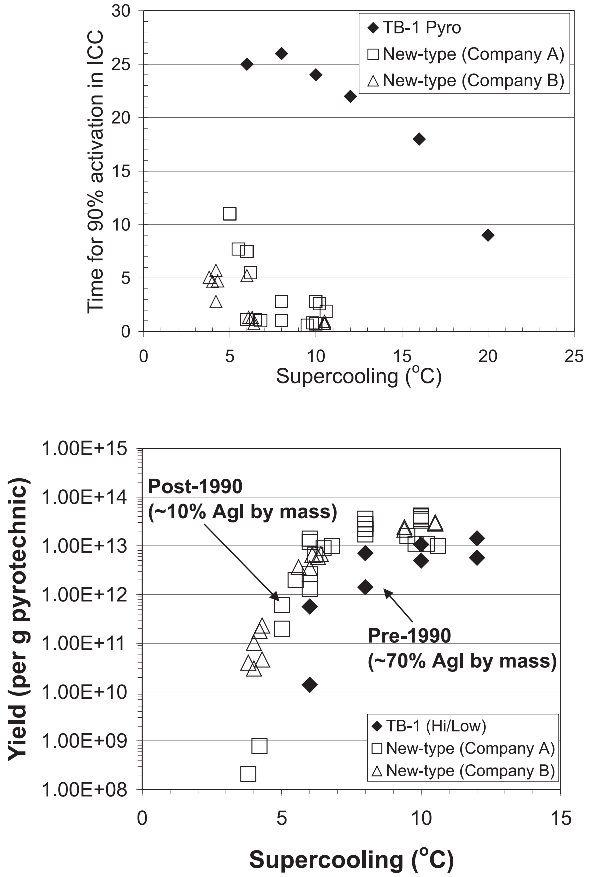


Figure 5 Yield (top panel) and rates (bottom panel) of ice formation by state-of-the-science pyrotechnic glaciogenic seeding generators, prior to (TB-1) and since about 1990. New-type pyrotechnics are more efficient in producing ice nuclei on a compositional basis, require less AgI (as $AgIO_3$), and "react" much faster in a water-saturated cloud. The new type pyrotechnics were developed in the former Yugoslavia but are now manufactured in North America. Results are from records of the Colorado State University isothermal cloud chamber (ICC) facility (Figure supplied by Paul DeMott, Colorado State University).

over prior nuclei generation capabilities and offer possibilities for engineering nuclei with specific and desired properties.

Numerical Modeling Capabilities

Simulations of Seeding Supercooled Orographic Clouds It is now possible to perform two- and three-dimensional simulations of clouds over individual mountains and entire mountain ranges with grid spacing on the order of a kilometer. For smaller mountain ridges, mixed-phase bin-resolving microphysics models can be used, while for larger mountain ranges, bulk microphysics is still needed. Such models can explicitly represent the transport and dispersion of seeding material, the primary and secondary ice nucleation of natural clouds, as well as nucleation of seeded material using laboratory-derived activity spectra.

An example of a numerical simulation of silver iodide transport over the Black Hills of South Dakota is given in Figure 6. Five generators were located in the simulation in the Northern Hills. The north winds spread the seeding material over the slopes and into the simulated orographic clouds. The AgI is activated in the clouds and forms snow at an earlier stage than the simulated natural nuclei (Farley et al., 1997).

Simulations of Seeding Cumulonimbi and Severe Convective Storms

It is now quite common to perform fully three-dimensional time-dependent simulations of cumulonimbi and mesoscale convective systems with grid spacing of about 1 km or even 0.5 km, with mixed-phase bulk microphysics models. For hailstorms, a hybrid bulk microphysics/bin-microphysics model with continuous accretion approximations has been implemented in several storm models (Farley et al., 1996; Johnson et al., 1993). With access to advanced, state-of-the-art computers, stochastic bin-resolving microphysics models such as Reisin et al. (1996) could be applied to the simulation of natural and seeded hailstorms. Such a model could be used to examine in some detail beneficial competition, trajectory lowering (including hygroscopic seeding), early rainout, and glaciation hail suppression concepts.

Other Advances and General Comments

Research in the past 10 years has shown the close association of the cloud dynamics with the cloud microphysics and precipitation processes. The total precipitation from a cloud system is often the product of both the precipitation processes and the dynamic airflow regime. Consequently, reasoning about seeding effects on precipitation should include the effects of the dynamics of the cloud and cloud system. Cloud and mesoscale numerical models run on supercomputer systems are needed to keep track of all of the possible interactions. And efficient field projects are needed to test the concepts and validate the models. These are tasks that should be taken more seriously so that the technology can be appropriately developed and applied.

A comment regarding the magnitude of the seeding effect is warranted here. Cloud seeding for weather modification does not cause floods nor does it create droughts. Floods are caused by natural conditions much more vigorous than those caused by seeding. The rain processes are fully developed and highly efficient in flood situations. Likewise, droughts are caused by large-scale weather conditions not influenced by cloud seeding. The use of cloud seeding to produce more precipitation is most helpful to store water in reservoirs before droughts occur or to cause as much precipitation as possible as weather conditions improve. Cloud seeding will not break a drought, but it may make the drought less severe by extracting as much of the atmospheric water supply as possible from the existing clouds. This aspect of

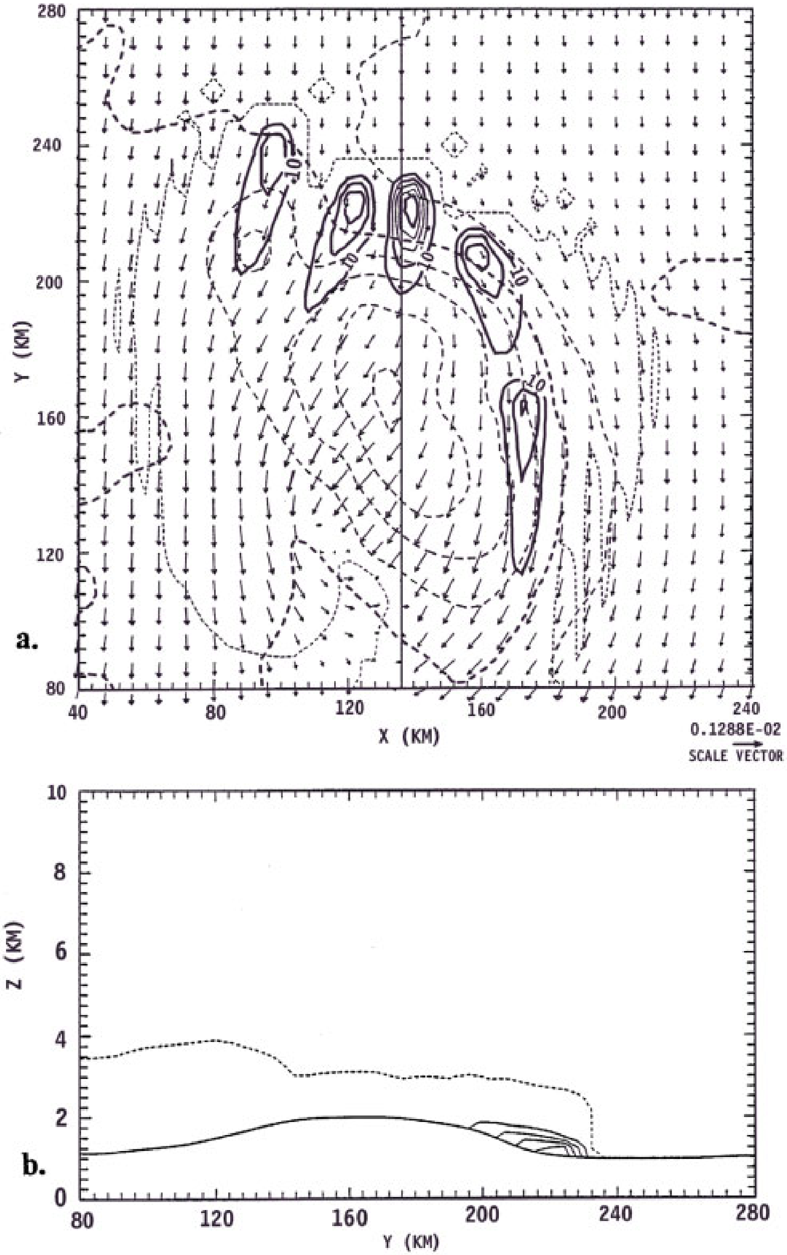


Figure 6 Seed case. (*a*) Surface locations where the inert tracer and AgI were released for day 3. (*b*) *Y–Z* cross section of the AgI field at $X = 136$ km. The contour interval is 0.1 g/kg for (*a*) and (*b*) (from Farley et al., 1997).

cloud seeding is still very much in the research stage. The research requires good cloud climatological data and well-conducted cloud seeding operations in drought conditions.

A common misconception follows from the fact that the desired result of cloud seeding for precipitation augmentation is to cause more rain or snow than might have occurred naturally. If the natural rainfall is far below normal, then the seeded rainfall will also be below normal, but not by as much, so the cloud seeding should not be blamed (but often is) for the below normal rainfall. This is another reason why more cloud seeding experiments are needed in various weather conditions, in the field and in computer models, to make more quantitative the effects of the cloud seeding and to develop methods that will distinguish the seeded rainfall from the natural component.

Nothing in this review has been said about fog suppression, the suppression of severe storms, and the reality of inadvertent weather modification, but the AMS statements give further information on these important topics (AMS, 1998a,b). Cold fog suppression is being practiced routinely in several airports around the world. The suppression of severe storms lacks critical hypotheses and should be tested in computer models before mounting field efforts. Large-scale land-use changes, widespread burning of forested lands, increased air pollution through industrialization, and extensive contrails caused by aircraft, have all contributed to inadvertent effects on the weather.

4 CONCLUDING REMARKS

Of great importance for proof of the efficacy of cloud seeding have been the significant strides made in the development of new or improved equipment, the use of more powerful and convenient computer power, and the development of more powerful statistical methods. Aircraft instrumentation, dual channel microwave radiometers, Doppler and multiparameter radars, satellite remote sensing, wind profilers, automated rain gauge networks, and mesoscale network stations have all contributed or have the potential to contribute to the detection of cloud seeding effects or to the identification of cloud seeding opportunities.

The increased computer power has made possible the greater use of cloud models in weather modification. Models have been developed that simulate explicitly the dispensing, transport, and nucleation effects of a seeding agent in convective and orographic clouds. Prediction and understanding of the seeding effects are products of the simulations. The results can also be used in evaluation and design of field projects.

The critical need for clean water in many parts of the world now, and even more so in the future, makes it imperative that the scientific basis of weather modification be strengthened. The conditions for increase, decrease, or redistribution of precipitation need to be determined. Stronger scientific and engineering projects in weather modification by cloud seeding give promise of more rapid progress in this scientifically important and socially relevant topic.

REFERENCES

AMS (1998a). Policy Statement: Planned and Inadvertent Weather Modification, *Bull. Am. Meteor. Soc.* **79**, 2771–2772.

AMS (1998b). Scientific background for the AMS policy statement on planned and inadvertent weather modification, *Bull. Am. Meteor. Soc.* **79**, 2773–2778.

BASC (2000). *New Opportunities in Weather Research, Focusing on Reducing Severe Weather Hazards and Providing Sustainable Water Resources*, Summary of the National Academy of Sciences Workshop for Assessing the Current State of Weather Modification Science as a Basis for Future Environmental Sustainability and Policy Development, prepared for the National Research Council Board on Atmospheirc Sciences and Climate (available at *ccrandal@taz.sdsmt.edu*).

Bigg, E. K. (1997). An independent evaluation of a South African hygroscopic cloud seeding experiment, 1991–1995, *Atm. Res.* **43**, 111–127.

Chai, S. K., W. G. Finnegan, and R. L. Pitter (1993). An interpretation of the mechanisms of ice crystal formation operative in the Lake Almanor cloud seeding program, *J. Appl. Meteor.* **32**, 1726–1732.

Changnon, S. A. (1998). In *Natural Hazards of North America*, National Geographic Maps, National Geographic Magazine, supplement, July, 1998, Washington, D.C.

Cooper, W. A., R. T. Bruintjes, and G. K. Mather (1997). Calculations pertaining to hygroscopic seeding with flares, *J. Appl. Meteor.* **36**, 1449–1469.

DeMott, P. J., W. C. Finnegan, and L. O. Grant (1983). An application of chemical kinetic theory and methodology to characterize the ice nucleating properties of aerosols used for weather modification, *J. Climate Appl. Meteor.* **22**, 1190–1203.

Dennis, A. S. (1980). *Weather Modification by Cloud Seeding*, New York, Academic.

Dessens, J. (1986). Hail in Southwestern France. II: Results of a 30-year hail prevention project with silver iodide seeding from the ground, *J. Climate Appl. Meteor.* **25**, 48–58.

Eklund, D. L., D. S. Jawa, and T. K. Rajala (1999). Evaluation of the Western Kansas weather modification program, *J. Wea. Mod.* **31**, 91–101.

Farley, R. D., H. Chen, H. D. Orville, and M. R. Hjelmfelt (1996). The numerical simulation of the effects of cloud seeding on hailstorms, Preprints, AMS 13th Conference on Planned and Inadvertent Weather Modification, Atlanta, GA, pp. 23–30.

Farley, R. D., D. L. Hjermstad, and H. D. Orville (1997). Numerical simulation of cloud seeding effects during a four-day storm period, *J. Wea. Mod.* **24**, 49–55.

Feng, D., and W. G. Finnegan (1989). An efficient, fast functioning nucleating agent—AgI•AgCl–4NaCl, *J. Wea. Mod.* **21**, 41–45.

IBHS (1998). *The Insured Cost of Natural Disasters: A Report on the IBHS Paid Loss Data Base*, Institute for Business and Home Safety, Tampa, FL (formerly Boston, MA).

Johnson, D. E., P. K. Wong, and J. M. Straka (1993). Numerical simulations of the 2 August 1981 CCOPE supercell storm with and without ice microphysics, *J. Appl. Meteor.* **32**, 745–759.

Mather, G. K., M. J. Dixon, and J. M. de Jager (1996). Assessing the potential for rain augmentation—the Nelspruit randomised convective cloud seeding experiment, *J. Appl. Meteor.* **35**, 1465–1482.

Mather, G. K., D. E. Terblanche, F. E. Steffens, and L. E. Fletcher (1997). Results of the South African cloud-seeding experiments using hygroscopic flares, *J. Appl. Meteor.* **36**, 1433–1447.

McGurty, B. M. (1999). Turning silver to gold: Measuring the benefits of cloud seeding, *Hydro. Rev.* April issue, 2–6.

Mossop , S. C. (1985). The origin and concentration of ice crystals in clouds, *Bull. Am. Meteor. Soc.* **66**, 264–273.

Nicholls, N. (2001). The insignificance of significance testing, *Bull. Am. Meteor. Soc.* **82**, 981–986.

Orville, H. D. (1996). A review of cloud modeling in weather modification, *Bull. Am. Meteor. Soc.* **77**, 1535–1555.

Orville, H. D., R. D. Farley, and J. H. Hirsch (1984). Some surprising results from simulated seeding of stratiform-type clouds, *J. Climate Appl. Meteor.* **23**, 1585–1600.

Orville, H. D., J. H. Hirsch, and R. D. Farley (1987). Further results on numerical cloud seeding simulations of stratiform-type clouds, *J. Wea. Modif.* **19**, 57–61.

Orville, H. D., C. Wang, and F. J. Kopp (1998). A simplified concept of hygroscopic seeding, *J. Wea. Mod.* **30**, 7–21.

Rangno, A. L. and P. V. Hobbs (1995). A new look at the Israeli cloud seeding experiments, *J. Appl. Meteor.* **34**, 1169–1193.

Rangno, A. L., and P. V. Hobbs (1997). Reply, *J. Appl. Meteor.* **36**, 272–276.

Reisin, T., S. Tzivion, and Z. Levin (1996). Seeding convective clouds with ice nuclei or hygroscopic particles: A numerical study using a model with detailed microphysics, *J. Appl. Meteor.* **35**, 1416–1434.

Reynolds, D. W. (1988). A report on winter snowpack augmentation, *Bull. Am. Meteor. Soc.* **69**, 1290–1300.

Rogers, R. R. and M. K. Yau (1989). *A Short Course in Cloud Physics*, 3rd ed., International Series in Natural Philosophy, Vol. 113, D. Ter Haar (Ed.), Butterworth-Heinemann, Woburn, MA.

Rosenfeld, D., and W. L. Woodley (1993). Effects of cloud seeding in west Texas: Additional results and new insights, *J. Appl. Meteor.* **32**, 1848–1866.

Rosenfeld, D. and W. L. Woodley (2000). Convective clouds with sustained highly supercooled liquid water down to $-37.5°$C, *Nature* **405**, 440–442.

Silverman, B. A. (2001). A critical assessment of glaciogenic seeding of convective clouds for rainfall enhancement, *Bull. Am. Meteor. Soc.* **82**, 903–923.

Smith, P. L., L. R. Johnson, D. L. Priegnitz, B. A. Boe, and P. J. Mielke, Jr. (1997). An exploratory analysis of crop hail insurance data for evidence of cloud seeding effects in North Dakota, *J. Appl. Meteor.* **36**, 463–473.

Stone, R. H., and A. W. Huggins (1996). The use of trace chemistry in conjunction with ice crystal measurements to assess wintertime cloud seeding experiments, 13th Conf. on Planned and Inadvertent Weather Modification, Atlanta, GA., Amer. Meteor. Soc., pp. 136–141.

Super, A. B., and J. A. Heimbach (1992). Investigations of the targeting of ground released silver iodide in Utah. Part I: Ground observations of silver in snow and ice nuclei, *J. Wea. Modif.* **24**, 19–34.

Super A. B., and E. W. Holroyd (1997). Some physical evidence of silver iodide and liquid propane seeding effects on Utah's Wasatch Plateau, *J. Wea. Modif.* **29**, 8–32.

Tzivion, S., T. Reisin, and Z. Levin (1994). Numerical simulation of hygroscopic seeding in a convective cloud, *J. Appl. Meteor.* **33**, 252–267.

Warburton, J. A., L. G. Young, M. S. Owens, and R. H. Stone (1985). The capture of ice nucleating and non ice-nucleating aerosols by ice-phase precipitationm, *J. Rech. Atmos*, **19**, (2–3), 249–255.

Warburton, J. A., L. G. Young, and R. H. Stone (1994). Assessment of seeding effects in snowpack augmentation programs: Ice nucleation and scavenging of seeding aerosols, *J. Appl. Meteor.* **33**, 121–130.

Warburton, J. A., R. H. Stone, and B. L. Marler (1995a). How the transport and dispersion of AgI aerosols may affect detectability of seeding effects by statistical methods, *J. Appl. Meteor.* **34**, 1930–1941.

Warburton, J. A., S. K. Chai, and L. G. Young (1995b). A new concept for assessing silver iodide cloud seeding effects in snow by physical and chemical methods, *Atmos. Res.* **36**, 171–176.

WMAB (1978). *The Management of Weather Resources.* Vol. I, Washington, DC, Dept. of Commerce.

Woodley, W. L., and D. Rosenfeld (2000). Evidence for changes in microphysical structure and cloud drafts following AgI Seeding, *J. Wea. Modif.* **32**, 53–67.

Young, K. C. (1993). *Microphysical Processes in Clouds*, New York, Oxford University Press.

CHAPTER 24

ATMOSPHERIC OPTICS

CRAIG F. BOHREN

1 INTRODUCTION

Atmospheric optics is nearly synonymous with light scattering, the only restrictions being that the scatterers inhabit the atmosphere and the primary source of their illumination is the sun. Essentially all light we see is scattered light, even that directly from the sun. When we say that such light is unscattered we really mean that it is scattered in the forward direction, hence it is *as if* it were unscattered. Scattered light is radiation from matter excited by an external source. When the source vanishes, so does the scattered light, as distinguished from light emitted by matter, which persists in the absence of external sources.

Atmospheric scatterers are either molecules or particles. A particle is an aggregation of sufficiently many molecules that it can be ascribed macroscopic properties such as temperature and refractive index. There is no canonical number of molecules that must unite to form a bona fide particle. Two molecules clearly do not a quorum make, but what about 10, 100, or 1000? The particle size corresponding to the largest of these numbers is about 10^{-3} μm. Particles this small of water substance would evaporate so rapidly that they could not exist long under conditions normally found in the atmosphere. As a practical matter, therefore, we need not worry unduly about hermaphrodite scatterers in the shadow region between molecule and particle.

A property of great relevance to scattering problems is *coherence*, both of the array of scatterers and of the incident light. At visible wavelengths air is an array of incoherent scatterers: The radiant power scattered by N molecules is N times that scattered by one (except in the forward direction). But when water vapor in air condenses, an incoherent array is transformed into a coherent array: Uncorrelated

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

water molecules become part of a single entity. Although a single droplet is a coherent array, a cloud of droplets taken together is incoherent.

Sunlight is incoherent but not in an absolute sense. Its lateral coherence length is tens of micrometers, which is why we can observe what are essentially interference patterns (e.g., coronas and glories) resulting from illumination of cloud droplets by sunlight.

This chapter begins with the color and brightness of a purely molecular atmosphere, including their variation across the vault of the sky. This naturally leads to the state of polarization of skylight. Because the atmosphere is rarely, if ever, entirely free of particles, the general characteristics of scattering by particles follow, setting the stage for atmospheric visibility.

Atmospheric refraction usually sits by itself, unjustly isolated from all those atmospheric phenomena embraced by the term *scattering*. Yet refraction is yet another manifestation of scattering, although in the forward direction, for which scattering is coherent.

Scattering by single water droplets and ice crystals, each discussed in turn, yields feasts for the eye as well as the mind. The curtain closes on the optical properties of clouds.

2 COLOR AND BRIGHTNESS OF MOLECULAR ATMOSPHERE

Brief History

Edward Nichols began his 1908 presidential address to the New York meeting of the American Physical Society as follows: “In asking your attention to-day, even briefly, to the consideration of the present state of our knowledge concerning the color of the sky it may be truly said that I am inviting you to leave the thronged thoroughfares of our science for some quiet side street where little is going on and you may even suspect that I am coaxing you into some blind alley, the inhabitants of which belong to the dead past.”

Despite this depreciatory statement, hoary with age, correct and complete explanations of the color of the sky still are hard to find. Indeed, all the faulty explanations lead active lives: the blue sky is the reflection of the blue sea; it is caused by water, either vapor or droplets or both; it is caused by dust. The true cause of the blue sky is not difficult to understand, requiring only a bit of critical thought stimulated by belief in the inherent fascination of all natural phenomena, even those made familiar by everyday occurrence.

Our contemplative prehistoric ancestors no doubt speculated on the origin of the blue sky, their musings having vanished into it. Yet it is curious that Aristotle, the most prolific speculator of early recorded history, makes no mention of it in his *Meteorologica* even though he delivered pronouncements on rainbows, halos, and mock suns and realized that “the sun looks red when seen through mist or smoke.” Historical discussions of the blue sky sometimes cite Leonardo da Vinci as the first to comment intelligently on the blue of the sky, although this reflects a European bias.

If history were to be written by a supremely disinterested observer, Arab philosophers would likely be given more credit for having had profound insights into the workings of nature many centuries before their European counterparts descended from the trees. Indeed, Möller (1972) begins his brief history of the blue sky with Jakub Ibn Ishak Al Kindi (800–870), who explained it as “a mixture of the darkness of the night with the light of the dust and haze particles in the air illuminated by the sun.”

Leonardo was a keen observer of light in nature even if his explanations sometimes fell short of the mark. Yet his hypothesis that “the blueness we see in the atmosphere is not intrinsic colour, but is caused by warm vapor evaporated in minute and insensible atoms on which the solar rays fall, rendering them luminous against the infinite darkness of the fiery sphere which lies beyond and includes it” would, with minor changes, stand critical scrutiny today. If we set aside Leonardo as *sui generis*, scientific attempts to unravel the origins of the blue sky may be said to have begun with Newton, that towering pioneer of optics, who, in time-honored fashion, reduced it to what he already had considered: interference colors in thin films. Almost two centuries elapsed before more pieces in the puzzle were contributed by the experimental investigations of von Brücke and Tyndall on light scattering by suspensions of particles. Around the same time Clausius added his bit in the form of a theory that scattering by minute bubbles causes the blueness of the sky. A better theory was not long in coming. It is associated with a man known to the world as Lord Rayleigh even though he was born John William Strutt.

Rayleigh’s paper of 1871 marks the beginning of a satisfactory explanation of the blue sky. His scattering law, the key to the blue sky, is perhaps the most famous result ever obtained by dimensional analysis. Rayleigh argued that the field $\mathbf{E}_s$ scattered by a particle small compared with the light illuminating it is proportional to its volume V and to the incident field $\mathbf{E}_i$. Radiant energy conservation requires that the scattered field diminish inversely as the distance r from the particle so that the scattered power diminishes as the square of r. To make this proportionality dimensionally homogeneous requires the inverse square of a quantity with the dimensions of length. The only plausible physical variable at hand is the wavelength of the incident light, which leads to

$$\mathbf{E}_s \propto \mathbf{E}_i \frac{V}{r\lambda^2} \tag{1}$$

When the field is squared to obtain the scattered power, the result is Rayleigh’s inverse fourth-power law. This law is really only an often—but not always—very good approximation. Missing from it are dimensionless properties of the particle such as its refractive index, which itself depends on wavelength. Because of this *dispersion*, therefore, nothing scatters exactly as the inverse fourth power.

Rayleigh’s 1871 paper did not give the complete explanation of the color and polarization of skylight. What he did that was not done by his predecessors was to give a law of scattering, which could be used to quantitatively test the hypothesis that selective scattering by atmospheric particles could transform white sunlight into blue skylight. But as far as giving the agent responsible for the blue sky, Rayleigh did not go essentially beyond Newton and Tyndall, who invoked particles. Rayleigh was circumspect about the nature of these particles, settling on salt as the most likely candidate. It was not until 1899 that he published the capstone to his work on skylight, arguing that air molecules themselves were the source of the blue sky. Tyndall cannot be given the credit for this because he considered air to be *optically empty*: When purged of all particles it scatters no light. This erroneous conclusion was a result of the small scale of his laboratory experiments. On the scale of the atmosphere, sufficient light is scattered by air molecules to be readily observable.

Molecular Scattering and the Blue of the Sky

Our illustrious predecessors all gave explanations of the blue sky requiring the presence of water in the atmosphere: Leonardo's "evaporated warm vapor," Newton's "globules of water," Clausius's bubbles. Small wonder, then, that water still is invoked as the cause of the blue sky. Yet a cause of something is that without which it would not occur, and the sky would be no less blue if the atmosphere were free of water.

A possible physical reason for attributing the blue sky to water vapor is that, because of selective *absorption*, liquid water (and ice) is blue upon transmission of white light over distances of order meters. Yet if all the water in the atmosphere at any instant were to be compressed into a liquid, the result would be a layer about 1 cm thick, which is not sufficient to transform white light into blue by selective absorption.

Water vapor does not compensate for its hundredfold lower abundance than nitrogen and oxygen by greater scattering per molecule. Indeed, scattering of visible light by a water molecule is slightly *less* than that by either nitrogen or oxygen.

Scattering by atmospheric molecules does not obey Rayleigh's inverse fourth-power law exactly. A least-squares fit over the visible spectrum from 400 to 700 nm of the *molecular scattering coefficient* of sea level air tabulated by Penndorf (1957) yields an inverse 4.089 scattering law.

The molecular scattering coefficient β, which plays important roles in following sections, may be written

$$\beta = N\sigma_s \tag{2}$$

where N is the number of molecules per unit volume and σ_s, the scattering cross section (an average because air is a mixture) per molecule, approximately obeys Rayleigh's law. The form of this expression betrays the incoherence of scattering by atmospheric molecules. The inverse of β is interpreted as the scattering *mean free path*, the average distance a photon must travel before being scattered.

To say that the sky is blue because of Rayleigh scattering, as is sometimes done, is to confuse an agent with a law. Moreover, as Young (1982) pointed out, the term *Rayleigh scattering* has many meanings. Particles small compared with the wavelength scatter according to the same law as do molecules. Both can be said to be

Rayleigh scatterers, but only molecules are necessary for the blue sky. Particles, even small ones, generally diminish the vividness of the blue sky.

Fluctuations are sometimes trumpeted as the "real" cause of the blue sky. Presumably, this stems from the fluctuation theory of light scattering by media in which the scatterers are separated by distances small compared with the wavelength. Using this theory, which is associated with Einstein and Smoluchowski, matter is taken to be continuous but characterized by a refractive index that is a random function of position. Einstein (1910) stated that "it is remarkable that our theory does not make *direct* use of the assumption of a discrete distribution of matter." That is, he circumvented a difficulty but realized it could have been met head on, as Zimm (1945) did years later.

The blue sky is really caused by scattering by molecules. To be more precise, scattering by bound electrons: free electrons do not scatter selectively. Because air molecules are separated by distances small compared with the wavelengths of visible light, it is not obvious that the power scattered by such molecules can be added. Yet if they are completely uncorrelated, as in an ideal gas (to good approximation the atmosphere is an ideal gas), scattering by N molecules is N times scattering by one. This is the only sense in which the blue sky can be attributed to scattering by fluctuations. Perfectly homogeneous matter does not exist. As stated pithily by Planck: "a chemically pure substance may be spoken of as a vacuum made turbid by the presence of molecules."

Spectrum and Color of Skylight

What is the spectrum of skylight? What is its color? These are two different questions. Answering the first answers the second but not the reverse. Knowing the color of skylight we cannot uniquely determine its spectrum because of *metamerism*: A given perceived color can in general be obtained in an indefinite number of ways.

Skylight is not blue (itself an imprecise term) in an absolute sense. When the visible spectrum of sunlight outside Earth's atmosphere is modulated by Rayleigh's scattering law, the result is a spectrum of scattered light that is neither solely blue nor even peaked in the blue (Fig. 1). Although blue does not predominate spectrally, it does predominate perceptually. We perceive the sky to be blue even though skylight contains light of all wavelengths.

Any source of light may be looked upon as a mixture of white light and light of a single wavelength called the *dominant wavelength.* The *purity* of the source is the relative amount of the monochromatic component in the mixture. The dominant wavelength of sunlight scattered according to Rayleigh's law is about 475 nm, which lies solidly in the blue if we take this to mean light with wavelengths between 450 and 490 nm. The purity of this scattered light, about 42%, is the upper limit for skylight. Blues of real skies are less pure.

Another way of conveying the color of a source of light is by its *color temperature*, the temperature of a blackbody having the same perceived color as the source. Since blackbodies do not span the entire gamut of colors, all sources of light cannot be assigned color temperatures. But many natural sources of light can. The color

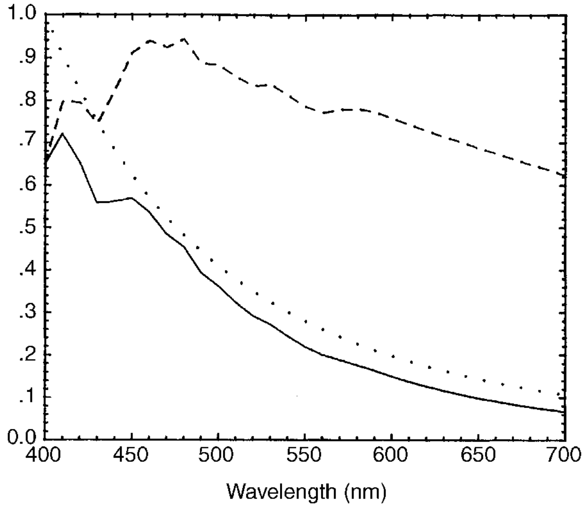


Figure 1 Rayleigh's scattering law (dots), the spectrum of sunlight outside the earth's atmosphere (dashes), and the product of the two (solid). The solar spectrum is taken from Thekaekara and Drummond (1971).

temperature of light scattered according to Rayleigh's law is infinite. This follows from Planck's spectral emission function $e_{b\lambda}$ in the limit of high temperature

$$e_{b\lambda} \simeq \frac{2\pi ckT}{\lambda^4} \qquad \frac{hc}{\lambda} \ll kT \tag{3}$$

where h is Planck's constant, k is Boltzmann's constant, c is the speed of light in vacuo, and T is absolute temperature. Thus the emission spectrum of a blackbody with an infinite temperature has the same functional form as Rayleigh's scattering law.

Variation of Sky Color and Brightness

Not only is skylight not pure blue, its color and brightness vary across the vault of the sky, with the best blues at zenith. Near the astronomical horizon the sky is brighter than overhead but of considerably lower purity. That this variation can be observed from an airplane flying at 10 km, well above most particles, suggests that the sky is inherently nonuniform in color and brightness (Fig. 2). To understand why requires invoking multiple scattering.

Multiple scattering gives rise to observable phenomena that cannot be explained solely by single-scattering arguments. This is easily demonstrated. Fill a blackened pan with clean water, then add a few drops of milk. The resulting dilute suspension illuminated by sunlight has a bluish cast. But when more milk is added, the suspen-

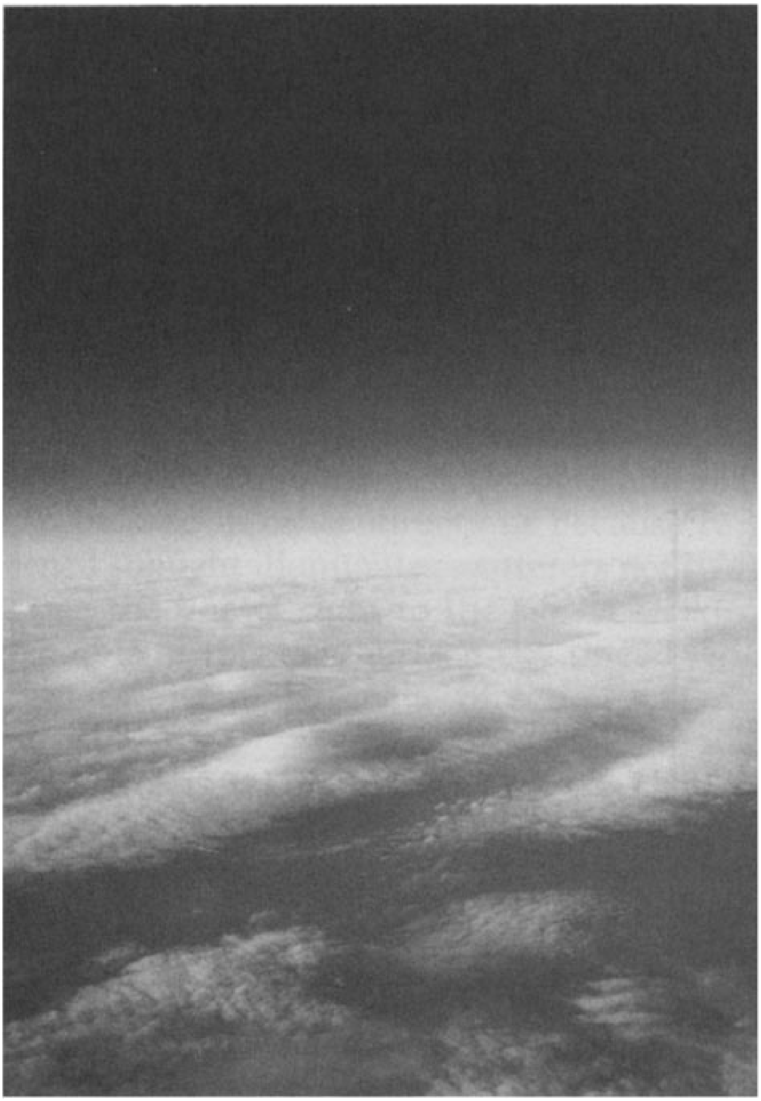


Figure 2 Even at an altitude of 10 km, well above most particles, the sky brightness increases markedly from the zenith to the astronomical horizon.

sion turns white. Yet the properties of the scatterers (fat globules) have not changed, only their *optical thickness*: the blue suspension being optically thin, the white being optically thick:

$$\tau = \int_1^2 \beta \, ds \tag{4}$$

Optical thickness is physical thickness in units of scattering mean free path, hence is dimensionless. The optical thickness τ between any two points connected by an arbitrary path in a medium populated by (incoherent) scatterers is an integral over the path: The *normal optical thickness* τ_n of the atmosphere is that along a radial path extending from the surface of Earth to infinity. Figure 3 shows τ_n over the visible spectrum for a purely molecular atmosphere. Because τ_n is generally small compared with unity, a photon from the sun traversing a radial path in the atmosphere is unlikely to be scattered more than once. But along a tangential path, the optical thickness is about 35 times greater (Fig. 4), which leads to several observable phenomena.

Even an intrinsically black object is luminous to an observer because of *airlight*, light scattered by all the molecules and particles along the line of sight from observer to object. Provided that this is uniformly illuminated by sunlight and that ground reflection is negligible, the airlight radiance L is approximately

$$L = GL_0(1 - e^{-\tau}) \tag{5}$$

where L_0 is the radiance of sunlight along the line of sight and τ is its optical thickness; G accounts for geometric reduction of radiance because of scattering of nearly monodirectional sunlight in all directions. If the line of sight is uniform in composition, $\tau = \beta d$, where β is the scattering coefficient and d is the physical distance to the black object.

If τ is small ($\ll 1$), $L \approx GL_0\tau$. In a purely molecular atmosphere, τ varies with wavelength according to Rayleigh's law; hence the distant black object in such an atmosphere is perceived to be bluish. As τ increases so does L but not proportionally.

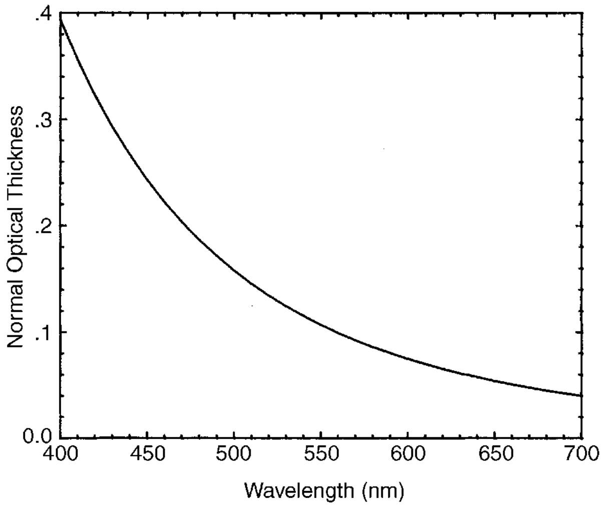


Figure 3 Normal optical thickness of a pure molecular atmosphere.

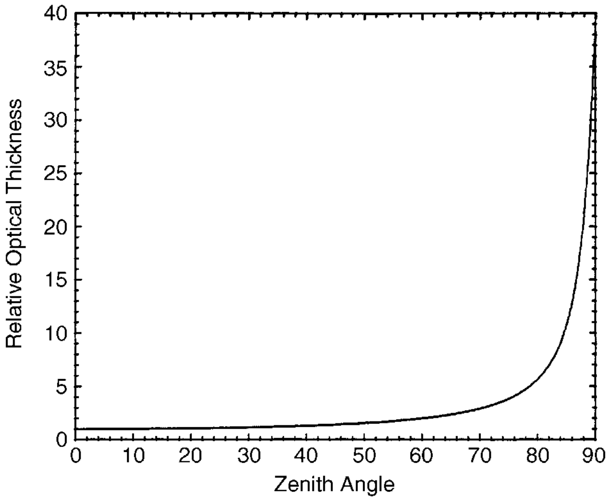


Figure 4 Optical thickness (relative to the normal optical thickness) of a molecular atmosphere along various paths with zenith angles between 0° (normal) and 90° (tangential).

Its limit is GL_0: The airlight radiance spectrum is that of the source of illumination. Only in the limit $d = 0$ is $L = 0$ and the black object is truly black.

Variation of the brightness and color of dark objects with distance was called *aerial perspective* by Leonardo. By means of it we estimate the distance of objects of unknown size such as mountains.

Aerial perspective belongs to the same family as the variation of color and brightness of the sky with zenith angle. Although the optical thickness along a path tangent to Earth is not infinite, it is sufficiently large (Figs. 3 and 4) that GL_0 is a good approximation for the radiance of the horizon sky. For isotropic scattering (a condition almost satisfied by molecules), G is around 10^{-5}, the ratio of the solid angle subtended by the sun to the solid angle of all directions (4π). Thus the horizon sky is not nearly so bright as direct sunlight.

Unlike in the milk experiment, what is observed when looking at the horizon sky is not multiply scattered light. Both have their origins in multiple scattering but manifested in different ways. Milk is white because it is weakly absorbing and optically thick, hence all components of incident white light are multiply scattered to the observer even though the blue component traverses a shorter average path in the suspension than the red component. White horizon light has escaped being multiply scattered, although multiple scattering is why this light is white (strictly, has the spectrum of the source). More light at the short wavelength end of the spectrum is scattered *toward* the observer than at the long wavelength end. But long wavelength light has the greater likelihood of being transmitted to the observer without being scattered *out of* the line of sight. For a long optical path, these two processes compensate, resulting in a horizon radiance spectrum of the source.

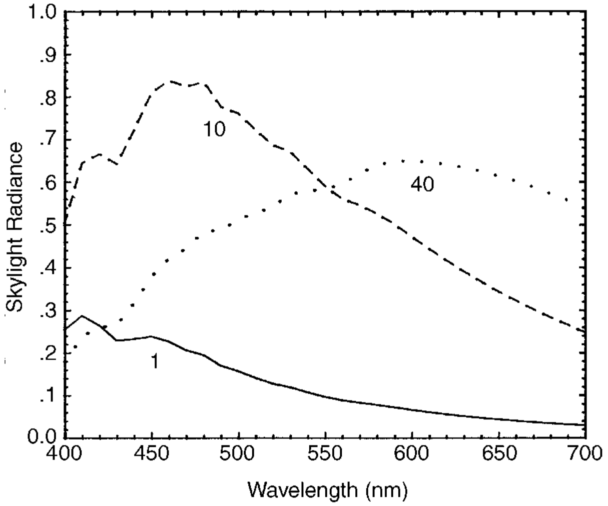


Figure 5 Spectrum of overhead skylight for the present molecular atmosphere (solid), as well as for hypothetical atmospheres 10 (dashes) and 40 (dots) times thicker.

Selective scattering by molecules is not sufficient for a blue sky. The atmosphere also must be optically thin, at least for most zenith angles (Fig. 4) (the blackness of space as a backdrop is taken for granted but also is necessary, which Leonardo recognized). A corollary of this is that the blue sky is not inevitable: An atmosphere composed entirely of nonabsorbing, selectively scattering molecules overlying a nonselectively reflecting Earth need not be blue. Figure 5 shows calculated spectra of the zenith sky over black ground for a molecular atmosphere with the present normal optical thickness as well as for hypothetical atmospheres 10 and 40 times thicker. What we take to be inevitable is accidental: If our atmosphere were much thicker, but identical in composition, the color of the sky would be quite different from what it is now.

Sunrise and Sunset

If short wavelength light is preferentially scattered out of direct sunlight, long wavelength light is preferentially transmitted in the direction of sunlight. Transmission is described by an exponential law (if light multiply scattered back into the direction of the sunlight is negligible):

$$L = L_o e^{-\tau} \tag{6}$$

where L is the radiance at the observer in the direction of the sun, L_o is the radiance of sunlight outside the atmosphere, and τ is the optical thickness along this path.

If the wavelength dependence of τ is given by Rayleigh's law, sunlight is *reddened* upon transmission: The spectrum of the transmitted light is comparatively richer than the incident spectrum in light at the long wavelength end of the visible spectrum. But to say that transmitted sunlight is reddened is not the same as saying it is red. The perceived color can be yellow, orange, or red, depending on the magnitude of the optical thickness. In a molecular atmosphere, the optical thickness along a path from the sun, even on or below the horizon, is not sufficient to give red light upon transmission. Although selective scattering by molecules yields a blue sky, reds are not possible in a molecular atmosphere, only yellows and oranges. This can be observed on clear days, when the horizon sky at sunset becomes successively tinged with yellow, then orange, but not red. Equation (6) applies to the radiance only in the direction of the sun. Oranges and reds can be seen in other directions because reddened sunlight illuminates scatterers not lying along the line of sight to the sun. A striking example of this is a horizon sky tinged with oranges and pinks in the direction *opposite* the sun.

The color and brightness of the sun changes as it arcs across the sky because the optical thickness along the line of sight changes with solar zenith angle Θ. If Earth were flat (as some still aver), the transmitted solar radiance would be

$$L = L_o \exp - \frac{\tau_n}{\cos \Theta} \tag{7}$$

This equation is a good approximation except near the horizon. On a flat Earth, the optical thickness is infinite for horizon paths. On a spherical Earth, optical thicknesses are finite although much larger for horizon than for radial paths.

The normal optical thickness of an atmosphere in which the number density of scatterers decreases exponentially with height z above the surface, $\exp(-z/H)$, is the same as that for a uniform atmosphere of finite thickness:

$$\tau_n = \int_0^\infty \beta \, dz = \beta_0 H \tag{8}$$

where H is the *scale height* and β_0 is the scattering coefficient at sea level. This equivalence yields a good approximation even for the tangential optical thickness. For any zenith angle, the optical thickness is given approximately by

$$\frac{\tau}{\tau_n} = \sqrt{\frac{R_e^2}{H^2}\cos^2 \Theta + \frac{2R_e}{H} + 1} - \frac{R_e}{H}\cos \Theta \tag{9}$$

where R_e is the radius of Earth. A flat Earth is one for which R_e is infinite, in which instance Eq. (9) yields the expected relation

$$\lim_{R_e \to \infty} \frac{\tau}{\tau_n} = \frac{1}{\cos \Theta} \tag{10}$$

For Earth's atmosphere, the molecular scale height is about 8 km. According to the approximate relation Eq. (9), therefore, the horizon optical thickness is about 39 times greater than the normal optical thickness. Taking the exponential decrease of molecular number density into account yields a value about 10% lower.

Variations on the theme of reds and oranges at sunrise and sunset can be seen even when the sun is overhead. The radiance at an observer an optical distance τ from a (horizon) cloud is the sum of cloudlight transmitted to the observer and airlight:

$$L = L_0 G(1 - e^{-\tau}) + L_0 G_c e^{-\tau} \tag{11}$$

where G_c is a geometrical factor that accounts for scattering of nearly monodirectional sunlight into a hemisphere of directions by the cloud. If it is approximated as an isotropic reflector with reflectance R and illuminated at an angle Φ, the geometrical factor G_c is $\Omega_s R \cos \Phi/\pi$. If $G_c > G$, the observed radiance is redder (i.e., enriched in light of longer wavelengths) than the incident radiance. If $G_c < G$, the observed radiance is bluer than the incident radiance. Thus distant horizon clouds can be reddish if they are bright or bluish if they are dark.

Underlying Eq. (11) is the implicit assumption that the line of sight is uniformly illuminated by sunlight. The first term in this equation is airlight, the second is transmitted cloudlight. Suppose, however, that the line of sight is shadowed from direct sunlight by clouds (that do not, of course, occlude the distant cloud of interest). This may reduce the first term in Eq. (11) so that the second term dominates. Thus under a partly overcast sky, distant horizon clouds may be reddish even when the sun is high in the sky.

The zenith sky at sunset and twilight is the exception to the general rule that molecular scattering is sufficient to account for the color of the sky. In the absence of molecular absorption, the spectrum of the zenith sky would be essentially that of the zenith sun (although greatly reduced in radiance), hence would not be the blue that is observed. This was pointed out by Hulburt (1953), who showed that absorption by ozone profoundly affects the color of the zenith sky when the sun is near the horizon. The Chappuis band of ozone extends from about 450 to 700 nm and peaks at around 600 nm. Preferential absorption of sunlight by ozone over long horizon paths gives the zenith sky its blueness when the sun is near the horizon. With the sun more than about 10° above the horizon, however, ozone has little effect on the color of the sky.

3 POLARIZATION OF LIGHT IN MOLECULAR ATMOSPHERE

Nature of Polarized Light

Unlike sound, light is a vector wave, an electromagnetic field lying in a plane normal to the propagation direction. The polarization state of such a wave is determined by the degree of correlation of any two orthogonal components into which its electric (or magnetic) field is resolved. Completely polarized light corresponds to complete correlation; completely unpolarized light corresponds to no correlation; partially polarized light corresponds to partial correlation.

If an electromagnetic wave is completely polarized, the tip of its oscillating electric field traces out a definite elliptical curve, the *vibration ellipse.* Lines and circles are special ellipses, the light being said to be linearly or circularly polarized, respectively. The general state of polarization is elliptical.

Any beam of light can be considered an incoherent superposition of two collinear beams, one unpolarized, the other completely polarized. The radiance of the polarized component relative to the total is defined as the *degree of polarization* (often multiplied by 100 and expressed as a percent). This can be measured for a source of light (e.g., light from different sky directions) by rotating a (linear) polarizing filter and noting the minimum and maximum radiances transmitted by it. The degree of (linear) polarization is defined as the difference between these two radiances divided by their sum.

Polarization by Molecular Scattering

Unpolarized light can be transformed into partially polarized light upon interaction with matter because of different changes in amplitude of the two orthogonal field components. An example of this is the partial polarization of sunlight upon scattering by atmospheric molecules, which can be detected by looking at the sky through a polarizing filter (e.g., polarizing sunglasses) while rotating it. Waxing and waning of the observed brightness indicates some degree of partial polarization.

In the analysis of any scattering problem a plane of reference is required. This is usually the *scattering plane*, determined by the directions of the incident and scattered waves, the angle between them being the *scattering angle.* Light polarized perpendicular (parallel) to the scattering plane is sometimes said to be vertically (horizontally) polarized. Vertical and horizontal in this context, however, are arbitrary terms indicating orthogonality and bear no relation, except by accident, to the direction of gravity.

The degree of polarization P of light scattered by a tiny sphere illuminated by unpolarized light is (Fig. 6)

$$P = \frac{1 - \cos^2\theta}{1 + \cos^2\theta} \tag{12}$$

where the scattering angle θ ranges from 0° (forward direction) to 180° (backward direction); the scattered light is partially linearly polarized perpendicular to the scattering plane. Although this equation is a first step toward understanding polarization of skylight, more often than not it also has been a false step, having led countless authors to assert that skylight is completely polarized at 90° from the sun. Although $P = 1$ at $\theta = 90°$ according to Eq. (12), skylight is never 100% polarized at this or any other angle, and for several reasons.

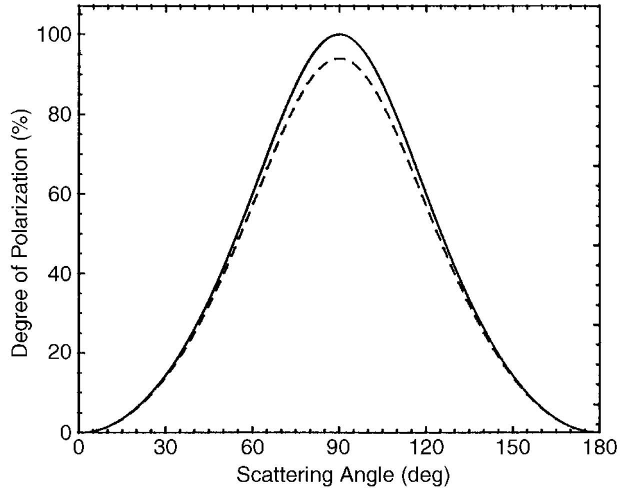


Figure 6 Degree of polarization of the light scattered by a small (compared with the wavelength) sphere for incident unpolarized light (solid). The dashed curve is for a small spheroid chosen such that the degree of polarization at 90° is that for air.

Although air molecules are very small compared with the wavelengths of visible light, a requirement underlying Eq. (12), the dominant constituents of air are not spherically symmetric.

The simplest model of an asymmetric molecule is a small spheroid. Although it is indeed possible to find a direction in which the light scattered by such a spheroid is 100% polarized, this direction depends on the spheroid's orientation. In an ensemble of randomly oriented spheroids each contributes its mite to the total radiance in a given direction, but each contribution is partially polarized to varying degrees between 0 and 100%. It is impossible for beams of light to be incoherently superposed in such a way that the degree of polarization of the resultant is greater than the degree of polarization of the most highly polarized beam. Because air is an ensemble of randomly oriented asymmetric molecules, sunlight scattered by air never is 100% polarized. The intrinsic departure from perfection is about 6%. ure 6 also includes a curve for light scattered by randomly oriented spheroids chosen to yield 94% polarization at 90°. This angle is so often singled out that it may deflect attention from nearby scattering angles. Yet the degree of polarization is greater than 50% for a range of scattering angles 70° wide centered about 90°.

Equation (12) applies to air, not to the atmosphere, the distinction being that in the atmosphere, as opposed to the laboratory, multiple scattering is not negligible. Also, atmospheric air is almost never free of particles and is illuminated by light reflected by the ground. We must take the atmosphere as it is, whereas in the laboratory we often can eliminate everything we consider extraneous.

Even if scattering by particles were negligible, because of both multiple scattering and ground reflection, light from any direction in the sky is not, in general, made up solely of light scattered in a single direction relative to the incident sunlight but is a superposition of beams with different scattering histories, hence different degrees of polarization. As a consequence, even if air molecules were perfect spheres and the atmosphere were completely free of particles, skylight would not be 100% polarized at 90° to the sun or at any other angle.

Reduction of the maximum degree of polarization is not the only consequence of multiple scattering. According to Figure 6, there should be two *neutral points* in the sky, directions in which skylight is unpolarized: directly toward and away from the sun. Because of multiple scattering, however, there are three such points. When the sun is higher than about 20° above the horizon there are neutral points within 20° of the sun, the *Babinet point* above it, the *Brewster point* below. They coincide when the sun is directly overhead and move apart as the sun descends. When the sun is lower than 20°, the *Arago point* is about 20° above the antisolar point, in the direction opposite the sun.

One consequence of the partial polarization of skylight is that the colors of distant objects may change when viewed through a rotated polarizing filter. If the sun is high in the sky, horizontal airlight will have a fairly high degree of polarization. According to the previous section, airlight is bluish. But if it also is partially polarized, its radiance can be diminished with a polarizing filter. Transmitted cloudlight, however, is unpolarized. Because the radiance of airlight can be reduced more than that of cloudlight, distant clouds may change from white to yellow to orange when viewed through a rotated polarizing filter.

4 SCATTERING BY PARTICLES

Up to this point we have considered only an atmosphere free of particles, an idealized state rarely achieved in nature. Particles still would inhabit the atmosphere even if the human race were to vanish from Earth. They are not simply byproducts of the "dark satanic mills" of civilization.

All molecules of the same substance are essentially identical. This is not true of particles: They vary in shape, size, and may be composed of one or more homogeneous regions.

Salient Differences Between Particles and Molecules

Magnitude of Scattering The distinction between scattering by molecules when widely separated and when packed together into a droplet is that between scattering by incoherent and coherent arrays. Isolated molecules are excited primarily by incident (external) light whereas the same molecules forming a droplet are excited by incident light and by each other's scattered fields. The total power scattered by an incoherent array of molecules is the sum of their scattered powers. The total power scattered by a coherent array is the square of the total scattered field, which in turn is the sum of all the fields scattered by the individual molecules. For an incoherent array we *may* ignore the wave nature of light, whereas for a coherent array we *must* take it into account.

Water vapor is a good example to ponder because it is a constituent of air and can condense to form cloud droplets. The difference between a sky containing water vapor and the same sky with the same amount of water but in the form of a cloud of droplets is dramatic.

According to Rayleigh's law, scattering by a particle small compared with the wavelength increases as the sixth power of its size (volume squared). A droplet of diameter 0.03 μm, for example, scatters about 10^{12} times more light than does one of its constituent molecules. Such a droplet contains about 10^7 molecules. Thus scattering per molecule as a consequence of condensation of water vapor into a coherent water droplet increases by about 10^5. Cloud droplets are much larger than 0.03 μm, a typical diameter being about 10 μm. Scattering per molecule in such a droplet is much greater than scattering by an isolated molecule, but not to the extent given by Rayleigh's law. Scattering increases as the sixth power of droplet diameter only when the molecules scatter coherently in phase. If a droplet is sufficiently small compared with the wavelength, each of its molecules is excited by essentially the same field and all the waves scattered by them interfere constructively. But when a droplet is comparable to or larger than the wavelength, interference can be constructive, destructive, and everything in between; hence scattering does not increase as rapidly with droplet size as predicted by Rayleigh's law.

The figure of merit for comparing scatterers of different size is their scattering cross section per unit volume, which, except for a multiplicative factor, is the scattering cross section per molecule. A scattering cross section may be looked upon as an effective area for removing radiant energy from a beam: The scattering cross section times the beam irradiance is the radiant power scattered in all directions.

The scattering cross section per unit volume for water droplets illuminated by visible light and varying in size from molecules (10^{-4} μm) to raindrops (10^3 μm) is shown in Figure 7. Scattering by a molecule that belongs to a cloud droplet is about 10^9 times greater than scattering by an isolated molecule, a striking example of the virtue of cooperation. Yet in molecular as in human societies there are limits beyond which cooperation becomes dysfunctional: Scattering by a molecule that belongs to a raindrop is about 100 times less than scattering by a molecule that belongs to a cloud droplet. This tremendous variation of scattering by water molecules depending on their state of aggregation has profound observational consequences. A cloud is optically so much different from the water vapor out of which it was born that the offspring bears no resemblance to its parents. We can see through tens of kilometers of air laden with water vapor, whereas a cloud a few tens of meters thick is enough to occult the sun. Yet a rain shaft born out of a cloud is considerably more translucent than its parent.

Wavelength Dependence of Scattering Regardless of their size and composition, particles scatter approximately as the inverse fourth power of wavelength if they are small compared with the wavelength and absorption is negligible, two important caveats. Failure to recognize them has led to errors, such as that yellow

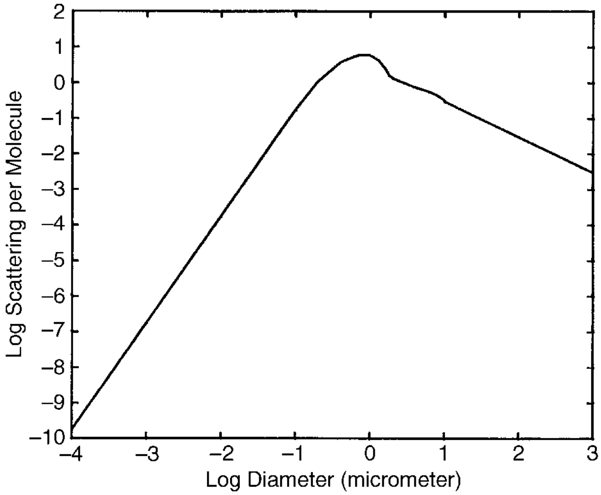


Figure 7 Scattering (per molecule) of visible light (arbitrary units) by water droplets varying in size from a single molecule to a raindrop.

light penetrates fog better because it is not scattered as much as light of shorter wavelengths. Although there may be perfectly sound reasons for choosing yellow instead of blue or green as the color of fog lights, greater transmission through fog is not one of them: Scattering by fog droplets is essentially independent of wavelength over the visible spectrum.

Small particles are selective scatterers, large particles are not. Particles neither small nor large give the reverse of what we have come to expect as normal. Figure 8 shows scattering of visible light by oil droplets with diameters 0.1, 0.8, and 10 μm. The smaller droplets scatter according to Rayleigh's law; the larger droplets (typical cloud droplet size) are nonselective. Between these two extremes are droplets (0.8 μm) that scatter long wavelength light more than short wavelength. Sunlight or moonlight seen through a thin cloud of these intermediate droplets would be bluish or greenish. This requires droplets of just the right size; hence it is a rare event, so rare that it occurs once in a blue moon. Astronomers, for unfathomable reasons, refer to the second full moon in a month as a blue moon, but if such a moon were blue it only would be by coincidence. The last reliably reported outbreak of blue and green suns and moons occurred in 1950 and was attributed to an oily smoke produced by Canadian forest fires.

Angular Dependence of Scattering The angular distribution of scattered light changes dramatically with the size of the scatterer. Molecules and particles that are small compared with the wavelength are nearly isotropic scatterers of unpolarized light, the ratio of maximum (at 0° and 180°) to minimum (at 90°) scattered radiance being only 2 for spheres, and slightly less for other spheroids. Although small particles scatter the same in the forward and backward hemispheres, scattering

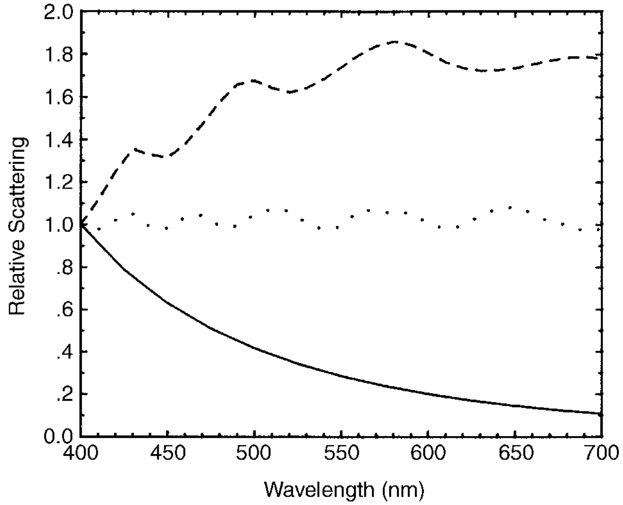


Figure 8 Scattering of visible light by oil droplets of diameter 0.1 μm (solid), 0.8 μm (dashes), and 10 μm (dots).

becomes markedly asymmetric for particles comparable to or larger than the wavelength. For example, forward scattering by a water droplet as small as 0.5 μm is about 100 times greater than backward scattering, and the ratio of forward to backward scattering increases more or less monotonically with size (Fig. 9).

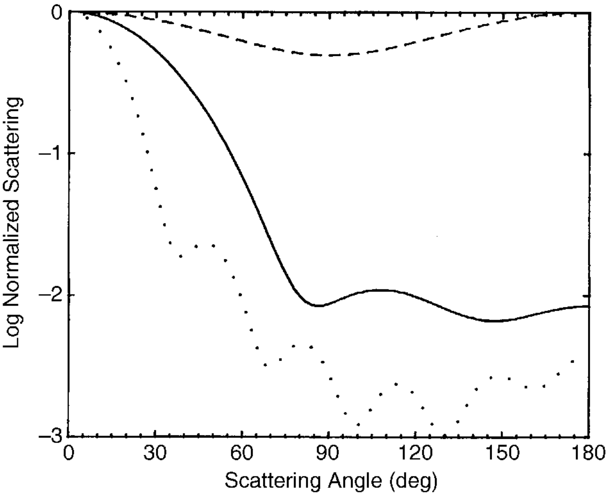


Figure 9 Angular dependence of scattering of visible light (0.55 μm) by water droplets small compared with the wavelength (dashes), diameter 0.5 μm (solid), and diameter 1.0 μm (dots).

The reason for this asymmetry is found in the singularity of the forward direction. In this direction, waves scattered by two or more scatterers excited solely by incident light (ignoring mutual excitation) are always in phase regardless of the wavelength and the separation of the scatterers. If we imagine a particle to be made up of N small subunits, scattering in the forward direction increases as N^2, the only direction for which this is always true. For other directions, the wavelets scattered by the subunits will not necessarily all be in phase. As a consequence, scattering in the forward direction increases with size (i.e., N) more rapidly than in any other direction.

Many common observable phenomena depend on this forward–backward asymmetry. Viewed toward the illuminating sun, glistening fog droplets on a spider's web warn us of its presence. But when we view the web with our backs to the sun, the web mysteriously disappears. A pattern of dew illuminated by the rising sun on a cold morning seems etched on a window pane. But if we go outside to look at the window, the pattern vanishes. Thin clouds sometimes hover over warm, moist heaps of dung, but may go unnoticed unless they lie between us and the source of illumination. These are but a few examples of the consequences of strongly asymmetric scattering by single particles comparable to or larger than the wavelength.

Degree of Polarization of Scattered Light All the simple rules about polarization upon scattering are broken when we turn from molecules and small particles to particles comparable to the wavelength. For example, the degree of polarization of light scattered by small particles is a simple function of scattering angle. But simplicity gives way to complexity as particles grow (Fig. 10), the scattered light being partially polarized parallel to the scattering plane for some scattering angles, perpendicular for others.

The degree of polarization of light scattered by molecules or by small particles is essentially independent of wavelength. But this is not true for particles comparable to or larger than the wavelength. Scattering by such particles exhibits *dispersion of polarization*: The degree of polarization at, say, 90° may vary considerably over the visible spectrum (Fig. 11).

In general, particles can act as polarizers or retarders or both. A polarizer transforms unpolarized light into partially polarized light. A retarder transforms polarized light of one form into that of another (e.g., linear into elliptical). Molecules and small particles, however, are restricted to roles as polarizers. If the atmosphere were inhabited solely by such scatterers, skylight could never be other than partially linearly polarized. Yet particles comparable to or larger than the wavelength often are present, hence skylight can acquire a degree of ellipticity upon multiple scattering: Incident unpolarized light is partially linearly polarized in the first scattering event, then transformed into partially elliptically polarized light in subsequent events.

Bees can navigate by polarized skylight. This statement, intended to evoke great awe for the photopolimetric powers of bees, is rarely accompanied by an important caveat: The sky must be clear. Figures 10 and 11 show two reasons—there are others—that bees, remarkable though they may be, cannot do the impossible. The simple wavelength-independent relation between the position of the sun and

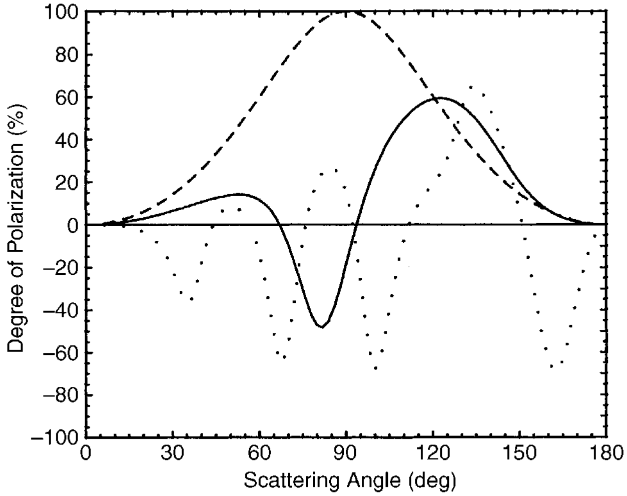


Figure 10 Degree of polarization of light scattered by water droplets illuminated by unpolarized visible light (0.55 μm). The dashed curve is for a droplet small compared with the wavelength; the solid curve is for a droplet of diameter 0.5 μm; the dotted curve is for a droplet of diameter 1.0 μm. Negative degrees of polarization indicate that the scattered light is partially polarized parallel to the scattering plane.

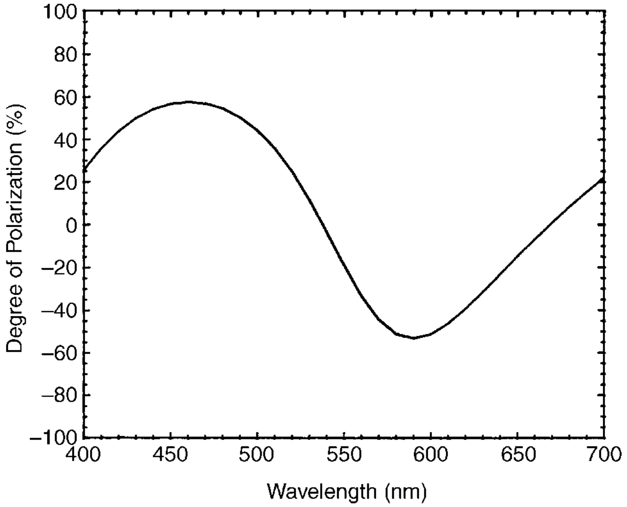


Figure 11 Degree of polarization at a scattering angle of 90° of light scattered by a water droplet of diameter 0.5 μm illuminated by unpolarized light.

the direction in which skylight is most highly polarized, an underlying necessity for navigating by means of polarized skylight, is obliterated when clouds cover the sky. This was recognized by the decoder of bee dances himself (von Frisch, 1971): "Sometimes a cloud would pass across the area of sky visible through the tube; when this happened the dances became disoriented, and the bees were unable to indicate the direction to the feeding place. Whatever phenomenon in the blue sky served to orient the dances, this experiment showed that it was seriously disturbed if the blue sky was covered by a cloud." But von Frisch's words often have been forgotten by disciples eager to spread the story about bee magic to those just as eager to believe what is charming even though untrue.

Vertical Distributions The scattering properties of particles are not only quite different, in general, from those of molecules, the different vertical distributions of particles and molecules by itself affects what is observed. The number density of molecules decreases more or less exponentially with height z above the surface: $\exp(-z/H_m)$, where the molecular scale height H_m is around 8 km. Although the decrease in number density of particles with height is also approximately exponential, the scale height for particles H_p is about 1 to 2 km. As a consequence, particles contribute disproportionately to optical thicknesses along near-horizon paths. Subject to the approximations underlying Eq. (9), the ratio of the tangential (horizon) optical thickness for particles τ_{tp} to that for molecules τ_{tm} is

$$\frac{\tau_{tp}}{\tau_{tm}} = \frac{\tau_{np}}{\tau_{nm}}\sqrt{\frac{H_m}{H_p}} \tag{13}$$

where the subscript t indicates a tangential path and n indicates a normal (radial) path. Because of the incoherence of scattering by atmospheric molecules and particles, scattering coefficients are additive, hence so are optical thicknesses. For equal normal optical thicknesses, the tangential optical thickness for particles is at least twice that for molecules. Molecules by themselves cannot give red sunrises and sunsets; molecules need the help of particles. For a fixed τ_{np}, the tangential optical thickness for particles is greater the more they are concentrated near the ground.

At the horizon the relative rate of change of transmission $T$of sunlight with zenith angle is

$$\frac{1}{T}\frac{dT}{d\Theta} = \tau_n \frac{R_e}{H} \tag{14}$$

where the scale height and normal optical thickness may be those for molecules or particles. Particles, being more concentrated near the surface, not only give disproportionate attenuation of sunlight on the horizon, they magnify the angular gradient of attenuation there. A perceptible change in color across the sun's disk (which subtends about 0.5°) on the horizon also requires the help of particles.

5 ATMOSPHERIC VISIBILITY

On a clear day can we really see forever? If not, how far can we see? To answer this question requires qualifying it by restricting viewing to more or less horizontal paths during daylight. Stars at staggering distances can be seen at night, partly because there is no skylight to reduce contrast, partly because stars overhead are seen in directions for which attenuation by the atmosphere is least.

The radiance in the direction of a black object is not zero because of light scattered along the line of sight (see discussion on variation of sky color and brightness under Section 2). At sufficiently large distances, this airlight is indistinguishable from the horizon sky. An example is a phalanx of parallel dark ridges, each ridge less distinct than those in front of it (Fig. 12). The farthest ridges blend into the horizon sky. Beyond some distance we cannot see ridges because of insufficient contrast.

Equation (5) gives the airlight radiance, a radiometric quantity that describes radiant power without taking into account the portion of it that stimulates the human eye or by what relative amount it does so at each wavelength. Luminance

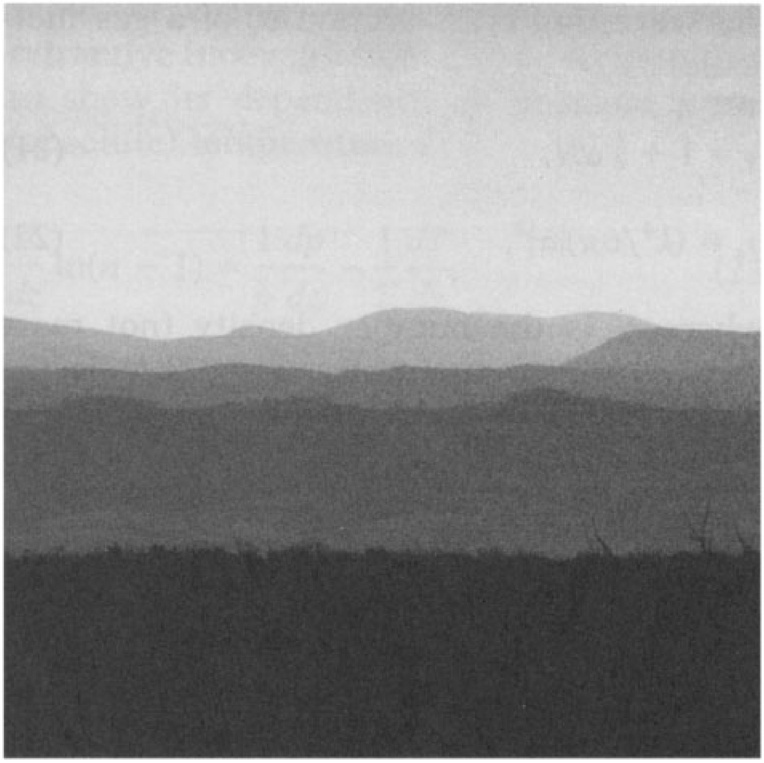


Figure 12 Because of scattering by molecules and particles along the line of sight, each successive ridge is brighter than the ones in front of it even though all of them are covered with the same dark vegetation.

(also sometimes called brightness) is the corresponding photometric quantity. Luminance and radiance are related by an integral over the visible spectrum:

$$B = \int K(\lambda)L(\lambda)\,d\lambda \tag{15}$$

where the luminous efficiency of the human eye K peaks at about 550 nm and vanishes outside the range 385 to 760 nm.

The *contrast* C between any object and the horizon sky is

$$C = \frac{B - B_\infty}{B_\infty} \tag{16}$$

where B_∞ is the luminance for an infinite horizon optical thickness. For a uniformly illuminated line of sight of length d, uniform in its scattering properties, and with a black backdrop, the contrast is

$$C = -\frac{\int KGL_0 \exp(-\beta d)\,d\lambda}{\int KGL_0\,d\lambda} \tag{17}$$

The ratio of integrals in this equation defines an average optical thickness:

$$C = -\exp(-\bar{\tau}) \tag{18}$$

This expression for contrast reduction with (optical) distance is mathematically, but not physically, identical to Eq. (6), which perhaps has engendered the misconception that atmospheric visibility is reduced because of attenuation. Yet, since there is no light from a black object to be attenuated, its finite visual range cannot be a consequence of attenuation.

The distance beyond which a dark object cannot be distinguished from the horizon sky is determined by the *contrast threshold*: the smallest contrast detectable by the human observer. Although this depends on the particular observer, the angular size of the object observed, the presence of nearby objects, and the absolute luminance, a contrast threshold of 0.02 is often taken as an average. This value in Eq. (18) gives

$$-\ln|C| = 3.9 = \bar{\tau} = \overline{\beta d} \tag{19}$$

To convert an optical distance into a physical distance requires the scattering coefficient. Because K is peaked at around 550 nm, we can obtain an approximate value of d from the scattering coefficient at this wavelength in Eq. (19). At sea level, the molecular scattering coefficient in the middle of the visible spectrum corresponds to about 330 km for "forever": the greatest distance a black object can be seen against the horizon sky assuming a contrast threshold of 0.02 and ignoring the curvature of Earth.

We also observe contrast between elements of the same scene, a hillside mottled with stands of trees and forest clearings, for example. The extent to which we can resolve details in such a scene depends on sun angle as well as distance.

The airlight radiance for a nonreflecting object is Eq. (5) with $G = p(\Theta)\Omega_s$, where $p(\Theta)$ is the probability (per unit solid angle) that light is scattered in a direction making an angle Θ with the incident sunlight and Ω_s is the solid angle subtended by the sun. When the sun is overhead, $\Theta = 90^\circ$; with the sun at the observer's back, $\Theta = 180^\circ$; for an observer looking directly into the sun $\Theta = 0^\circ$.

The radiance of an object with a finite reflectance R is given by Eq. (11). Equations (5) and (11) can be combined to obtain the contrast between reflecting and nonreflecting objects:

$$\begin{aligned} C &= \frac{Fe^{-\tau}}{1 + (F - 1)e^{-\tau}}, \\ F &= \frac{R\cos\Phi}{np(\Theta)} \end{aligned} \tag{20}$$

All else being equal, therefore, contrast decreases as $p(\Theta)$ increases. As shown in Figure 9, $p(\Theta)$ is more sharply peaked in the forward direction the larger the scatterer. Thus we expect the details of a distant scene to be less distinct when looking toward the sun than away from it if the optical thickness of the line of sight has an appreciable component contributed by particles comparable to or larger than the wavelength.

On humid, hazy days, visibility is often depressingly poor. Haze, however, is not water vapor but rather water that has ceased to be vapor. At high relative humidities, but still well below 100%, small soluble particles in the atmosphere accrete liquid water to become solution droplets (haze). Although these droplets are much smaller than cloud droplets, they markedly diminish visual range because of the sharp increase in scattering with particle size (Fig. 7). The same number of water molecules when aggregated in haze scatter vastly more than when apart.

6 ATMOSPHERIC REFRACTION

Atmospheric refraction is a consequence of molecular scattering, which is rarely stated given the historical accident that before light and matter were well understood refraction and scattering were locked in separate compartments and subsequently have been sequestered more rigidly than monks and nuns in neighboring cloisters. The connection between (lateral) scattering and refraction (forward scattering) can be divined from the expressions for the refractive index n of a gas and the scattering cross section σ_s of a gas molecule:

$$n = 1 + \frac{1}{2}\alpha N \tag{21}$$

$$\sigma_s = \frac{k^4}{6\pi}|\alpha|^2 \tag{22}$$

where N is the number density (not mass density) of gas molecules, $k = 2\pi/\lambda$ is the wavenumber of the incident light, and α is the polarizability of a molecule (induced dipole moment per unit inducing electric field). The appearance of the polarizability in Eq. (21) but its square in Eq. (22) is the clue that refraction is associated with electric fields whereas lateral scattering is associated with electric fields squared (powers). Scattering, without qualification, usually means scattering in all directions. Refraction, in a nutshell, is scattering in the forward direction. In this special direction incident and scattered fields superpose coherently to form the transmitted field, which is shifted in phase from that of the incident field by an amount determined by the polarizability and number density of scatterers.

Terrestrial Mirages

Mirages are not illusions, no more so than are reflections in a pond. Reflections of plants growing at its edge are not interpreted as plants growing into the water. If the water is ruffled by wind, the reflected images may be so distorted that they are no longer recognizable as those of plants. Yet we still would not call such distorted images illusions. And so is it with mirages. They are images noticeably different from what they would be in the absence of atmospheric refraction, creations of the atmosphere, not of the mind.

Mirages are vastly more common than is realized. Look and you shall see them. Contrary to popular opinion, they are not unique to deserts. Mirages can be seen frequently over ice-covered landscapes and highways flanked by deep snowbanks. Temperature per se is not what gives mirages but rather temperature gradients.

Because air is a mixture of gases, the polarizability for air in Eq. (21) is an average over all its molecular constituents, although their individual polarizabilities are about the same (at visible wavelengths). The vertical refractive index gradient can be written so as to show its dependence on pressure p and (absolute) temperature T:

$$\frac{d}{dz}\ln(n-1) = \frac{1}{p}\frac{dp}{dz} - \frac{1}{T}\frac{dT}{dz} \tag{23}$$

Pressure decreases approximately exponentially with height, where the scale height is around 8 km. Thus the first term on the right side of Eq. (23) is around 0.1/km. Temperature usually decreases with height in the atmosphere. An average lapse rate of temperature (i.e., its decrease with height) is around 6°C/km. The average temperature in the troposphere (within about 15 km of the surface) is around

280 K. Thus the magnitude of the second term in Eq. (23) is around 0.02/km. On average, therefore, the refractive index gradient is dominated by the vertical pressure gradient. But within a few meters of the surface, conditions are far from average. On a sun-baked highway your feet may be touching asphalt at 50°C while your nose is breathing air at 35°C, which corresponds to a lapse rate a thousand times the average. Moreover, near the surface, temperature can increase with height. In shallow surface layers, in which the pressure is nearly constant, the temperature gradient determines the refractive index gradient. It is in such shallow layers that mirages, which are caused by refractive index gradients, are seen.

Cartoonists by their fertile imaginations unfettered by science, and textbook writers by their carelessness have engendered the notion that atmospheric refraction can work wonders, lifting images of ships, for example, from the sea high into the sky. A back-of-the-envelope calculation dispels such notions. The refractive index of air at sea level is about 1.0003 (Fig. 13). Light from empty space incident at glancing incidence onto a uniform slab with this refractive index is displaced in angular position from where it would have been in the absence of refraction by

$$\delta = \sqrt{2(n-1)} \tag{24}$$

This yields an angular displacement of about 1.4°, which as we shall see is a rough upper limit.

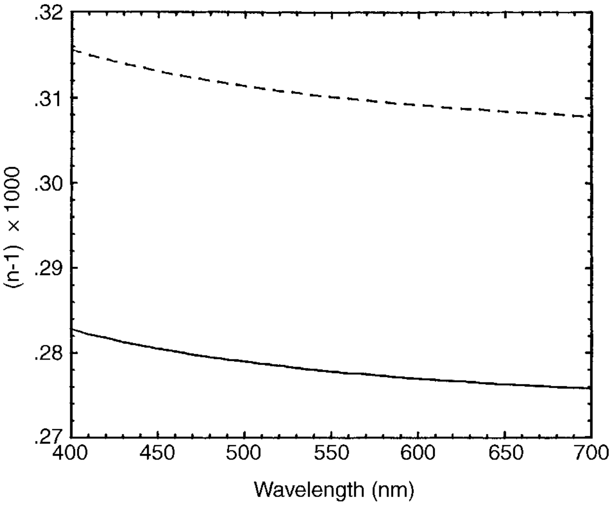


Figure 13 Sea-level refractive index versus wavelength at −15°C (dashes), and 15°C (solid). Data from Penndorf (1957).

Trajectories of light rays in nonuniform media can be expressed in different ways. According to Fermat's principle of least time (which ought to be extreme time), the actual path taken by a ray between two points is such that the path integral

$$\int_1^1 n\,ds \tag{25}$$

is an extremum over all possible paths. This principle has inspired piffle about the alleged efficiency of nature, which directs light over routes that minimize travel time, presumably freeing it to tend to important business at its destination.

The scale of mirages is such that in analyzing them we may pretend that Earth is flat. On such an Earth, with an atmosphere in which the refractive index varies only in the vertical, Fermat's principle yields a generalization of Snell's law

$$n \sin\theta = \text{constant} \tag{26}$$

where θ is the angle between the ray and the vertical direction. We could, of course, have bypassed Fermat's principle to obtain this result.

Under the assumption that θ is small compared to 1, Eq. (26) yields the following differential equation satisfied by a ray:

$$\frac{d^2z}{dy^2} = \frac{dn}{dz} \tag{27}$$

where y and z are its horizontal and vertical coordinates, respectively. For a constant refractive index gradient, which to good approximation occurs for a constant temperature gradient, Eq. (27) yields parabolas for ray trajectories. One such parabola for a constant temperature gradient about 100 times the average is shown in Figure 14. Note the vastly different horizontal and vertical scales. The image is displaced downward from what it would be in the absence of atmospheric refraction, hence the designation *inferior* mirage. This is the familiar highway mirage, seen over highways warmer than the air above them. The downward angular displacement is

$$\delta = \frac{1}{2} s \frac{dn}{dz} \tag{28}$$

where s is the horizontal distance between object and observer (image). Even for a temperature gradient 1000 times the tropospheric average, displacements of mirages are less than a degree at distances of a few kilometers.

If temperature increases with height, as it does, for example, in air over a cold sea, the resulting mirage is called a *superior* mirage. Inferior and superior are not designations of lower and higher caste but rather of displacements downward and upward.

For a constant temperature gradient, one and only one parabolic ray trajectory connects an object point to an image point. Multiple images therefore are not possible. But temperature gradients close to the ground are rarely linear. The

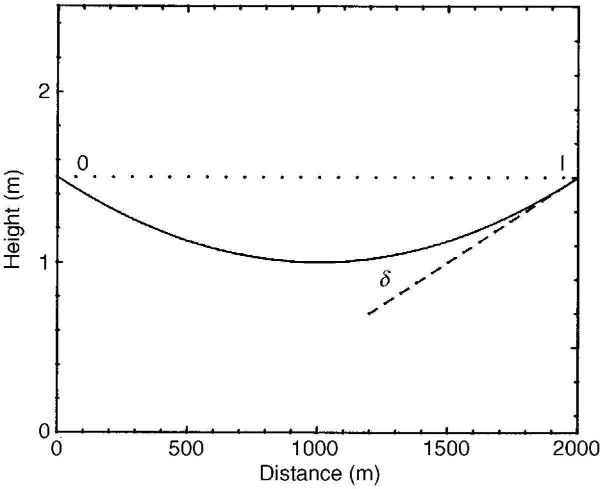


Figure 14 Parabolic ray paths in an atmosphere with a constant refractive index gradient (inferior mirage). Note the vastly different horizontal and vertical scales.

upward transport of energy from a hot surface occurs by molecular conduction through a stagnant boundary layer of air. Somewhat above the surface, however, energy is transported by air in motion. As a consequence, the temperature gradient steepens toward the ground if the energy flux is constant. This variable gradient can lead to two observable consequences: magnification and multiple images.

According to Eq. (28), all image points at a given horizontal distance are displaced downward by an amount proportional to the (constant) refractive index gradient. A corollary is that the closer an object point is to a surface where the temperature gradient is greatest, the greater the downward displacement of the corresponding image point. Thus nonlinear vertical temperature profiles may magnify images.

Multiple images are seen frequently on highways. What often appears to be water on the highway ahead but evaporates before it is reached is the inverted secondary image of either the horizon sky or of horizon objects lighter than dark asphalt.

Extraterrestrial Mirages

When we turn from mirages of terrestrial objects to those of extraterrestrial bodies, most notably the sun and moon, we can no longer pretend that Earth is flat. But we can pretend that the atmosphere is uniform and bounded. The total phase shift of a vertical ray from the surface to infinity is the same in an atmosphere with an exponentially decreasing molecular number density as in a hypothetical atmosphere with a uniform number density equal to the surface value up to height *H*.

A ray refracted along a horizon path by this hypothetical atmosphere and originating from outside it had to have been incident on it from an angle δ below the horizon:

$$\delta = \sqrt{\frac{2H}{R}} - \sqrt{\frac{2H}{R} - 2(n-1)} \tag{29}$$

where R is the radius of Earth. Thus, when the sun (or moon) is seen to be on the horizon, it is actually more than halfway below it, δ being about 0.36° whereas the angular width of the sun (and moon) is about 0.5°.

Extraterrestrial bodies seen near the horizon also are vertically compressed. The simplest way to estimate the amount of compression is from the rate of change of angle of refraction θ_r with angle of incidence θ_i for a uniform slab

$$\frac{d\theta_r}{d\theta_i} = \frac{\cos\theta_i}{\sqrt{n^2 - \sin^2\theta_i}} \tag{30}$$

where the angle of incidence is that for a curved but uniform atmosphere such that the refracted ray is horizontal. The result is

$$\frac{d\theta_r}{d\theta_i} = \sqrt{1 - \frac{R}{H}(n-1)} \tag{31}$$

according to which the sun near the horizon is distorted into an ellipse with aspect ratio about 0.87. We are unlikely to notice this distortion, however, because we expect the sun and moon to be circular, hence we see them that way.

The previous conclusions about the downward displacement and distortion of the sun were based on a refractive index profile determined mostly by the vertical pressure gradient. Near the ground, however, the temperature gradient is the prime determinant of the refractive index gradient, as a consequence of which the sun on the horizon can take on shapes more striking than a mere ellipse. For example, Figure 15 shows a nearly triangular sun with serrated edges. Assigning a cause to these serrations provides a lesson in the perils of jumping to conclusions. Obviously, the serrations are the result of sharp changes in the temperature gradient—or so one might think. Setting aside how such changes could be produced and maintained in a real atmosphere, a theorem of Fraser (1975) gives pause for thought. According to this theorem "in a horizontally (spherically) homogeneous atmosphere it is impossible for more than one image of an extraterrestrial object (sun) to be seen above the astronomical horizon." The serrations on the sun in Figure 15 are multiple images. But if the refractive index varies only vertically (i.e., along a radius), no matter how sharply, multiple images are not possible. Thus the serrations must owe their existence to horizontal variations of the refractive index, a consequence of gravity waves propagating along a temperature inversion.

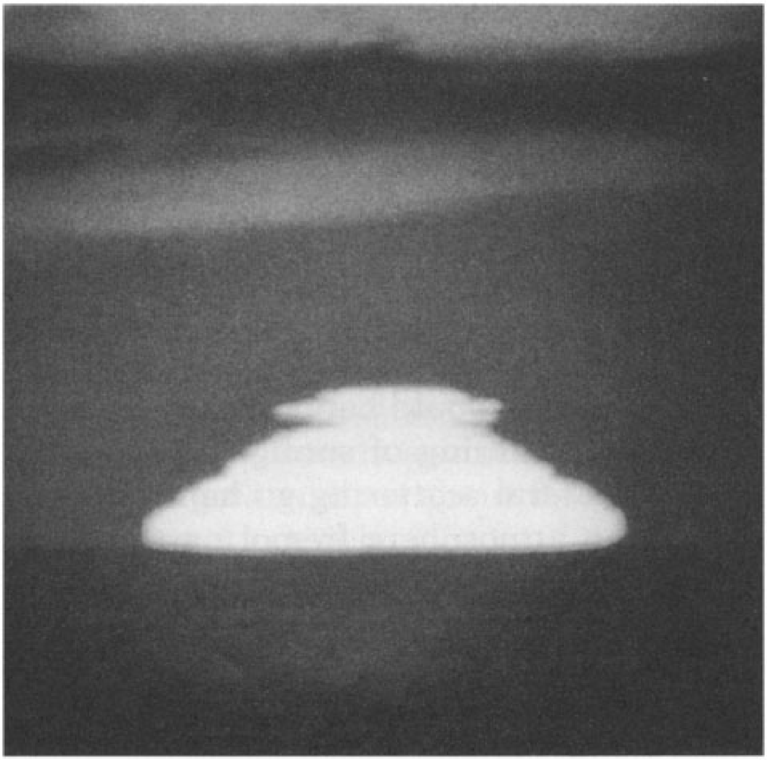


Figure 15 A nearly triangular sun on the horizon. The serrations are a consequence of horizontal variations in refractive index.

The Green Flash

Compared to the rainbow, the green flash is not a rare phenomenon. Before you dismiss this assertion as the ravings of a lunatic, consider that rainbows require raindrops as well as sunlight to illuminate them, whereas rainclouds often completely obscure the sun. Moreover, the sun must be below about 42°. As a consequence of these conditions, rainbows are not seen often, but often enough that they are taken as the paragon of color variation. Yet tinges of green on the upper rim of the sun can be seen every day at sunrise and sunset given a sufficiently low horizon and a cloudless sky. Thus the conditions for seeing a green flash are more easily met than those for seeing a rainbow. Why then is the green flash considered to be so rare? The distinction here is between a rarely observed phenomenon (the green flash) and a rarely observable one (the rainbow).

The sun may be considered to be a collection of disks, one for each visible wavelength. When the sun is overhead, each disk coincides and we see the sun as white. But as it descends in the sky, atmospheric refraction displaces the disks by slightly different amounts, the red less than the violet (see Fig. 13). Most of each disk overlaps all the others except for the disks at the extremes of the visible spectrum. As a consequence, the upper rim of the sun is violet or blue, its lower rim red, whereas its interior, the region in which all disks overlap, is still white.

This is what would happen in the absence of lateral scattering of sunlight. But refraction and lateral scattering go hand in hand, even in an atmosphere free of particles. Selective scattering by atmospheric molecules and particles causes the color of the sun to change. In particular, the violet-bluish upper rim of the low sun can be transformed to green.

According to Eq. (29) and Figure 13, the angular width of the green upper rim of the low sun is about 0.01°, too narrow to be resolved with the naked eye or even to be seen against its bright backdrop. But, depending on the temperature profile, the atmosphere itself can magnify the upper rim and yield a second image of it, thereby enabling it to be seen without the aid of a telescope or binoculars. Green rims, which require artificial magnification, can be seen more frequently than green flashes, which require natural magnification. Yet both can be seen often by those who know what to look for and are willing to look.

7 SCATTERING BY SINGLE WATER DROPLETS

All the colored atmospheric displays that result when water droplets (or ice crystals) are illuminated by sunlight have the same underlying cause: Light is scattered in different amounts in different directions by particles larger than the wavelength, and the directions in which scattering is greatest depends on wavelength. Thus, when particles are illuminated by white light, the result can be angular separation of colors even if scattering integrated over all directions is independent of wavelength (as it essentially is for cloud droplets and ice crystals). This description, although correct, is too general to be completely satisfying. We need something more specific, more quantitative, which requires theories of scattering.

Because superficially different theories have been used to describe different optical phenomena, the notion has become widespread that they are caused by these theories. For example, coronas are said to be caused by diffraction and rainbows by refraction. Yet both the corona and the rainbow can be described quantitatively to high accuracy with a theory (the Mie theory for scattering by a sphere) in which diffraction and refraction do not explicitly appear. No fundamentally impenetrable barrier separates scattering from (specular) reflection, refraction, and diffraction. Because these terms came into general use and were entombed in textbooks before the nature of light and matter was well understood, we are stuck with them. But if we insist that diffraction, for example, is somehow different from scattering, we do so at the expense of shattering the unity of the seemingly disparate observable phenomena that result when light interacts with matter. What is observed depends on the composition and disposition of the matter, not on which approximate theory in a hierarchy is used for quantitative description.

Atmospheric optical phenomena are best classified by the direction in which they are seen and by the agents responsible for them. Accordingly, the following sections are arranged in order of scattering direction, from forward to backward.

When a single water droplet is illuminated by white light and the scattered light projected onto a screen, the result is a set of colored rings. But in the atmosphere we see a mosaic to which individual droplets contribute. The scattering pattern of a single droplet is the same as the mosaic provided that multiple scattering is negligible.

Coronas and Iridescent Clouds

A cloud of droplets narrowly distributed in size and thinly veiling the sun (or moon) can yield a spectacular series of colored concentric rings around it. This corona is most easily described quantitatively by the Fraunhofer diffraction theory, a simple approximation valid for particles large compared with the wavelength and for scattering angles near the forward direction. According to this approximation, the differential scattering cross section (cross section for scattering into a unit solid angle) of a spherical droplet of radius a illuminated by light of wavenumber k is

$$\frac{|S|^2}{k^2} \tag{32}$$

where the scattering amplitude is

$$S = x^2 \frac{1 + \cos\theta}{2} \frac{J_1(x \sin\theta)}{x \sin\theta} \tag{33}$$

where J_1 is the Bessel function of first order and the size parameter $x = ka$. The quantity $(1 + \cos\theta)/2$ is usually approximated by 1 since only near-forward scattering angles θ are of interest.

The differential scattering cross section, which determines the angular distribution of the scattered light, has maxima for $x \sin\theta = 5.137, 8.417, 11.62, \ldots$ Thus the dispersion in the position of the first maximum is

$$\frac{d\theta}{d\lambda} \approx \frac{0.817}{a} \tag{34}$$

and is greater for higher-order maxima. This dispersion determines the upper limit on drop size such that a corona can be observed. For the total angular dispersion over the visible spectrum to be greater than the angular width of the sun (0.5°), the droplets cannot be larger than about 60 μm in diameter. Drops in rain, even in drizzle, are appreciably larger than this, which is why coronas are not seen through rain shafts. Scattering by a droplet of diameter 10 μm (Fig. 16), a typical cloud droplet size, gives sufficient dispersion to yield colored coronas.

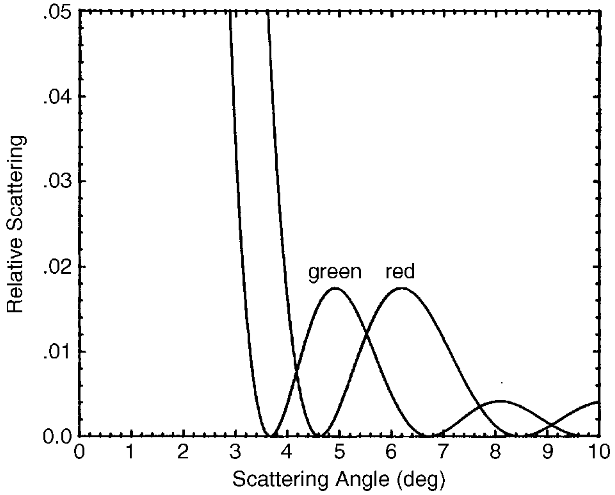


Figure 16 Scattering of light near the forward direction (according to Fraunhofer theory) by a sphere of diameter 10 μm illuminated by red and green light.

Suppose that the first angular maximum for blue light (0.47 μm) occurs for a droplet of radius a. For red light (0.66 μm) a maximum is obtained at the same angle for a droplet of radius $a + \Delta a$. That is, the two maxima, one for each wavelength, coincide. From this we conclude that coronas require narrow size distributions: If cloud droplets are distributed in radius with a relative variance $\Delta a/a$ greater than about 0.4, color separation is not possible.

Because of the stringent requirements for the occurrence of coronas, they are not observed often. Of greater occurrence are the corona's cousins, iridescent clouds, which display colors but usually not arranged in any obviously regular geometrical pattern. Iridescent patches in clouds can be seen even at the edges of thick clouds that occult the sun.

Coronas are not the unique signatures of spherical scatterers. Randomly oriented ice columns and plates give similar patterns according to Fraunhofer theory (Takano and Asano, 1983). As a practical matter, however, most coronas probably are caused by droplets. Many clouds at temperatures well below freezing contain subcooled water droplets. Only if a corona were seen in a cloud at a temperature lower than −40 °C could one assert with confidence that it must be an ice-crystal corona.

Rainbows

In contrast with coronas, which are seen looking toward the sun, rainbows are seen looking away from it and are caused by water drops much larger than those that give coronas. To treat the rainbow quantitatively we may pretend that light incident on a transparent sphere is composed of individual rays, each of which suffers a different fate determined only by the laws of specular reflection and refraction. Theoretical justification for this is provided by van de Hulst's (1957, p. 208) *localization principle*, according to which terms in the exact solution for scattering by a transparent sphere correspond to more or less localized rays.

Each incident ray splinters into an infinite number of scattered rays: externally reflected, transmitted without internal reflection, transmitted after one, two, and so on internal reflections. At any scattering angle θ, each splinter contributes to the scattered light. Accordingly, the differential scattering cross section is an infinite series with terms of the form

$$\frac{b(\theta)}{\sin\theta}\frac{db}{d\theta} \tag{35}$$

The *impact parameter* b is a sin Θ_i, where Θ_i is the angle between an incident ray and the normal to the sphere. Each term in the series corresponds to one of the splinters of an incident ray. A *rainbow angle* is a singularity (or *caustic*) of the differential scattering cross section at which the conditions

$$\frac{d\theta}{db} = 0 \qquad \frac{b}{\sin\theta} \neq 0 \tag{36}$$

are satisfied. Missing from Eq. (35) are various reflection and transmission coefficients (Fresnel coefficients), which display no singularities and hence do not determine rainbow angles.

A rainbow is not associated with rays externally reflected or transmitted without internal reflection. The succession of rainbow angles associated with one, two, three, ... internal reflections are called primary, secondary, tertiary,... rainbows. Aristotle recognized that "three or more rainbows are never seen, because even the second is dimmer than the first, and so the third reflection is altogether too feeble to reach the sun" (Aristotle's view was that light streams outward from the eye). Although he intuitively grasped that each successive ray is associated with ever-diminishing energy, his statement about the nonexistence of tertiary rainbows in nature is not quite true. Although reliable reports of such rainbows are rare (unreliable reports are as common as dirt), at least one observer who can be believed has seen one (Pledgley, 1986).

An incident ray undergoes a total angular deviation as a consequence of transmission into the drop, one or more internal reflections, and transmission out of the drop. Rainbow angles are angles of minimum deviation.

For a rainbow of any order to exist

$$\cos\Theta_i = \sqrt{\frac{n^2-1}{p(p+1)}} \tag{37}$$

must lie between 0 and 1, where Θ_i is the angle of incidence of a ray that gives a rainbow after p internal reflections and n is the refractive index of the drop. A primary bow therefore requires drops with refractive index less than 2; a secondary bow requires drops with refractive index less than 3. If raindrops were composed of titanium dioxide ($n \approx 3$), a commonly used opacifier for paints, primary rainbows would be absent from the sky and we would have to be content with only secondary bows.

If we take the refractive index of water to be 1.33, the scattering angle for the primary rainbow is about 138°. This is measured from the forward direction (solar point). Measured from the antisolar point (the direction toward which one must look in order to see rainbows in nature), this scattering angle corresponds to 42°, the basis for a previous assertion that rainbows (strictly, primary rainbows) cannot be seen when the sun is above 42°. The secondary rainbow is seen at about 51° from the antisolar point. Between these two rainbows is *Alexander's dark band*, a region into which no light is scattered according to geometrical optics.

The colors of rainbows are a consequence of sufficient dispersion of the refractive index over the visible spectrum to give a spread of rainbow angles that appreciably exceeds the width of the sun. The width of the primary bow from violet to red is about 1.7°; that of the secondary bow is about 3.1°.

Because of its band of colors arcing across the sky, the rainbow has become the paragon of color, the standard against which all other colors are compared. Lee and Fraser (1990) (see also Lee, 1991), however, challenged this status of the rainbow, pointing out that even the most vivid rainbows are colorimetrically far from pure.

Rainbows are almost invariably discussed as if they occurred literally in a vacuum. But real rainbows, as opposed to the pencil-and-paper variety, are necessarily observed in an atmosphere the molecules and particles of which scatter sunlight that adds to the light from the rainbow but subtracts from its purity of color.

Although geometrical optics yields the positions, widths, and color separation of rainbows, it yields little else. For example, geometrical optics is blind to *supernumerary bows*, a series of narrow bands sometimes seen below the primary bow. These bows are a consequence of interference, hence fall outside the province of geometrical optics. Since supernumerary bows are an interference phenomenon, they, unlike primary and secondary bows (according to geometrical optics), depend on drop size. This poses the question of how supernumerary bows can be seen in rain showers, the drops in which are widely distributed in size. In a nice piece of detective work, Fraser (1983) answered this question.

Raindrops falling in a vacuum are spherical. Those falling in air are distorted by aerodynamic forces, not, despite the depictions of countless artists, into tear drops but rather into nearly oblate spheroids with their axes more or less vertical. Fraser argued that supernumerary bows are caused by drops with a diameter of about 0.5 mm, at which diameter the angular position of the first (and second) supernumerary bow has a minimum: Interference causes the position of the supernumerary bow to increase with decreasing size whereas drop distortion causes it to increase with increasing size. Supernumerary patterns contributed by drops on either side of the minimum cancel leaving only the contribution from drops at the minimum. This cancellation occurs only near the tops of rainbow, where supernumerary bows are seen. In the vertical parts of a rainbow, a horizontal slice through a distorted drop is more or less circular; hence these drops do not exhibit a minimum supernumerary angle.

According to geometrical optics, all spherical drops, regardless of size, yield the same rainbow. But it is not necessary for a drop to be spherical for it to yield rainbows independent of its size. This merely requires that the plane defined by the incident and scattered rays intersect the drop in a circle. Even distorted drops satisfy this condition in the vertical part of a bow. As a consequence, the absence of supernumerary bows there is compensated for by more vivid colors of the primary and secondary bows (Fraser, 1972). Smaller drops are more likely to be spherical, but the smaller a drop, the less light it scatters. Thus the dominant contribution to the luminance of rainbows is from the larger drops. At the top of a bow, the plane defined by the incident and scattered rays intersects the large, distorted drops in an ellipse, yielding a range of rainbow angles varying with the amount of distortion, hence a pastel rainbow. To the knowledgeable observer, rainbows are no more uniform in color and brightness than is the sky.

Although geometrical optics predicts that all rainbows are equal (neglecting background light), real rainbows do not slavishly follow the dictates of this approximate theory. Rainbows in nature range from nearly colorless fog bows (or cloud bows) to the vividly colorful vertical portions of rainbows likely to have inspired myths about pots of gold.

The Glory

Continuing our sweep of scattering directions, from forward to backward, we arrive at the end of our journey: *the glory.* Because it is most easily seen from airplanes it sometimes is called the *pilot's bow.* Another name is *anticorona*, which signals that it is a corona around the antisolar point. Although glories and coronas share some common characteristics, there are differences between them other than direction of observation. Unlike coronas, which may be caused by nonspherical ice crystals, glories require spherical cloud droplets. And a greater number of colored rings may be seen in glories than in coronas because the decrease in luminance away from the backward direction is not as steep as that away from the forward direction. To see a glory from an airplane, look for colored rings around its shadow cast on clouds below. This shadow is not an essential part of the glory; it merely directs you to the antisolar point.

Like the rainbow, the glory may be looked upon as a singularity in the differential scattering cross section of Eq. (35). Equation (36) gives one set of conditions for a singularity; the second set is

$$\sin\theta = 0 \qquad b(\theta) \neq 0 \tag{38}$$

That is, the differential scattering cross section is infinite for nonzero impact parameters (corresponding to incident rays that do not intersect the center of the sphere) that give forward (0°) or backward (180°) scattering. The forward direction is excluded because this is the direction of intense scattering accounted for by the Fraunhofer theory.

For one internal reflection, Eq. (38) leads to the condition

$$\sin \Theta_i = \frac{n}{2}\sqrt{4 - n^2} \tag{39}$$

which is satisfied only for refractive indices between 1.414 and 2, the lower refractive index corresponding to a grazing incidence ray. The refractive index of water lies outside this range. Although a condition similar to Eq. (39) is satisfied for rays undergoing four or more internal reflections, insufficient energy is associated with such rays. Thus it seems that we have reached an impasse: The theoretical condition for a glory cannot be met by water droplets. Not so, says van de Hulst (1947) in a seminal work. He argues that 1.414 is close enough to 1.33 given that geometrical optics is, after all, an approximation. Cloud droplets are large compared with the wavelength, but not so large that geometrical optics is an infallible guide to their optical behavior. Support for the van de Hulstian interpretation of glories was provided by Bryant and Cox (1966), who showed that the dominant contribution to the glory is from the last terms in the exact series for scattering by a sphere. Each successive term in this series is associated with ever larger impact parameters. Thus the terms that give the glory are indeed those corresponding to grazing rays. Further unraveling of the glory and vindication of van de Hulst's conjectures about the glory were provided by Nussenzveig (1979).

It sometimes is asserted that geometrical optics is incapable of treating the glory. Yet the same can be said for the rainbow. Geometrical optics explains rainbows only in the sense that it predicts singularities for scattering in certain directions (rainbow angles). But it can predict only the angles of intense scattering not the amount. Indeed, the error is infinite. Geometrical optics also predicts a singularity in the backward direction. Again, this simple theory is powerless to predict more. Results from geometrical optics for both rainbows and glories are not the end but rather the beginning, an invitation to take a closer look with more powerful magnifying glasses.

8 SCATTERING BY SINGLE ICE CRYSTALS

Scattering by spherical water drops in the atmosphere gives rise to three distinct displays in the sky: coronas, rainbows, and glories. Ice particles (crystals) also can inhabit the atmosphere, and they introduce two new variables in addition to size: shape and orientation, the second a consequence of the first. Given this increase in the number of degrees of freedom, it is hardly cause for wonder that ice crystals are the source of a greater variety of displays than are water drops. As with rainbows, the gross features of ice-crystal phenomena can be described simply with geometrical optics, various phenomena arising from the various fates of rays incident on crystals. Colorless displays (e.g., sun pillars) are generally associated with reflected rays, colored displays (e.g., sun dogs and halos) with refracted rays. Because of the wealth of ice-crystal displays, it is not possible to treat all of them here, but one example should point the way toward understanding many of them.

Sun Dogs and Halos

Because of the hexagonal crystalline structure of ice, it can form as hexagonal plates in the atmosphere. The stable position of a plate falling in air is with the normal to its face more or less vertical, which is easy to demonstrate with an ordinary business card. When the card is dropped with its edge facing downward (the supposedly aerodynamic position that many people instinctively choose), the card somersaults in a helter-skelter path to the ground. But when the card is dropped with its face parallel to the ground, it rocks back and forth gently in descent.

A hexagonal ice plate falling through air and illuminated by a low sun is like a 60° prism illuminated normally to its sides (Fig. 17). Because there is no mechanism for orienting a plate within the horizontal plane, all plate orientations in this plane are equally probable. Stated another way, all angles of incidence for a fixed plate are equally probable. Yet all scattering angles (deviation angles) of rays refracted into and out of the plate are not equally probable.

Figure 18 shows the range of scattering angles corresponding to a range of rays incident on a 60° ice prism that is part of a hexagonal plate. For angles of incidence less than about 13° the transmitted ray is totally internally reflected in the prism. For angles of incidence greater than about 70°, the transmittance plunges. Thus the only rays of consequence are those incident between about 13° and 70°.

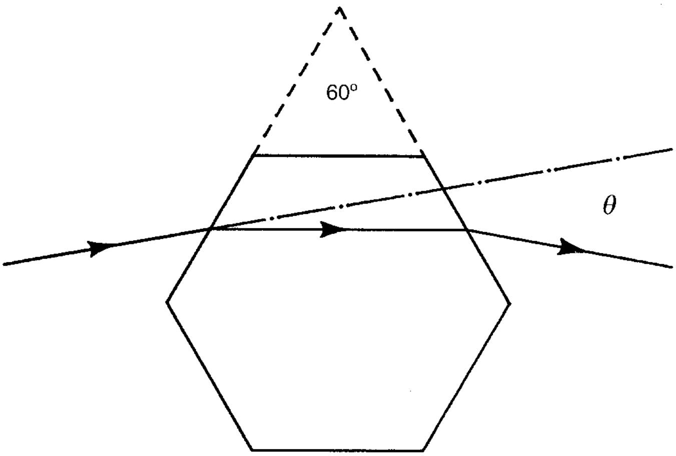


Figure 17 Scattering by a hexagonal ice plate illuminated by light parallel to its basal plane. The particular scattering angle θ shown is an angle of minimum deviation. The scattered light is that associated with two refractions by the plate.

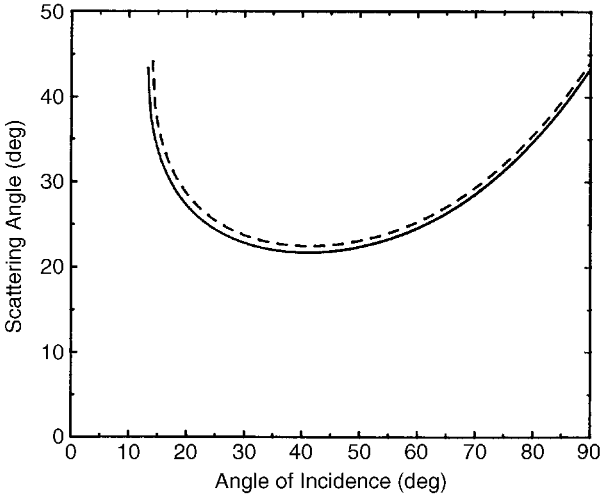


Figure 18 Scattering by a hexagonal ice plate (see Fig. 17) in various orientations (angles of incidence). The solid curve is for red light, the dashed curve is for blue light.

All scattering angles are not equally probable. The (uniform) probability distribution $p(\theta_i)$ of incidence angles θ_i is related to the probability distribution $P(\theta)$ of scattering angles θ by

$$P(\theta) = \frac{p(\theta_i)}{d\theta/d\theta_i} \tag{40}$$

At the incidence angle for which $d\theta/d\theta_i = 0$, $P(\theta)$ is infinite and scattered rays are intensely concentrated near the corresponding angle of minimum deviation.

The physical manifestation of this singularity (or caustic) at the angle of minimum deviation for a 60° hexagonal ice plate is a bright spot about 22° from either or both sides of a sun low in the sky. These bright spots are called *sun dogs* (because they accompany the sun) or *parhelia* or *mock suns.*

The angle of minimum deviation θ_m, hence the angular position of sun dogs, depends on the prism angle Δ (60° for the plates considered) and refractive index:

$$\theta_m = 2\sin^{-1}\left(n\sin\frac{\Delta}{2}\right) - \Delta \tag{41}$$

Because ice is dispersive, the separation between the angles of minimum deviation for red and blue light is about 0.7° (Fig. 18), somewhat greater than the angular width of the sun. As a consequence, sun dogs may be tinged with color, most noticeably toward the sun. Because the refractive index of ice is least at the red end of the spectrum, the red component of a sun dog is closest to the sun. Moreover, light of any two wavelengths has the same scattering angle for different angles of incidence if one of the wavelengths does not correspond to red. Thus red is the purest color seen in a sun dog. Away from its red inner edge a sun dog fades into whiteness.

With increasing solar elevation, sun dogs move away from the sun. A falling ice plate is roughly equivalent to a prism, the prism angle of which increases with solar elevation. From Eq. (41) it follows that the angle of minimum deviation, hence the sun dog position, also increases.

At this point you may be wondering why only the 60° prism portion of a hexagonal plate was singled out for attention. As evident from Figure 17, a hexagonal plate could be considered to be made up of 120° prisms. For a ray to be refracted twice, its angle of incidence at the second interface must be less than the critical angle. This imposes limitations on the prism angle. For a refractive index 1.31, all incident rays are totally internally reflected by prisms with angles greater than about 99.5°.

A close relative of the sun dog is the 22° halo, a ring of light approximately 22° from the sun (Fig. 19). Lunar halos are also possible and are observed frequently (although less frequently than solar halos); even moon dogs are possible. Until Fraser (1979) analyzed halos in detail, the conventional wisdom had been that they obviously were the result of randomly oriented crystals, yet another example of jumping to conclusions. By combining optics and aerodynamics, Fraser showed that if ice crystals are small enough to be randomly oriented by Brownian motion, they are too small to yield sharp scattering patterns.

But completely randomly oriented plates are not necessary to give halos, especially ones of nonuniform brightness. Each part of a halo is contributed to by plates with a different tip angle (angle between the normal to the plate and the vertical). The transition from oriented plates (zero tip angle) to randomly oriented plates occurs over a narrow range of sizes. In the transition region plates can be small enough to be partially oriented yet large enough to give a distinct contribution to the halo. Moreover, the mapping between tip angles and azimuthal angles on the halo depends on solar elevation. When the sun is near the horizon, plates can give a distinct halo over much of its azimuth.

When the sun is high in the sky, hexagonal plates cannot give a sharp halo but hexagonal columns—another possible form of atmospheric ice particles—can. The stable position of a falling column is with its long axis horizontal. When the sun is directly overhead, such columns can give a uniform halo even if they all lie in the horizontal plane. When the sun is not overhead but well above the horizon, columns also can give halos.

A corollary of Fraser's analysis is that halos are caused by crystals with a range of sizes between about 12 and 40 μm. Larger crystals are oriented; smaller particles are too small to yield distinct scattering patterns.

More or less uniformly bright halos with the sun neither high nor low in the sky could be caused by mixtures of hexagonal plates and columns or by clusters of bullets (rosettes). Fraser opines that the latter is more likely.

One of the byproducts of his analysis is an understanding of the relative rarity of the 46° halo. As we have seen, the angle of minimum deviation depends on the

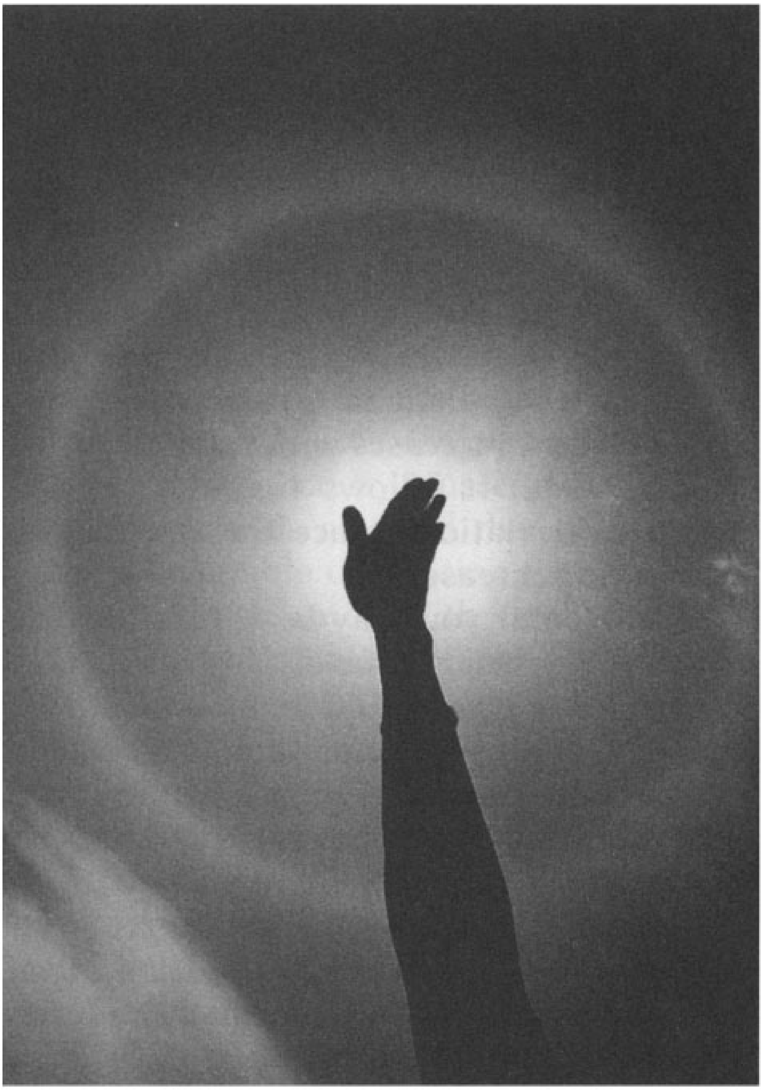


Figure 19 A 22° solar halo. The hand is not for artistic effect but rather to occlude the bright sun.

prism angle. Light can be incident on a hexagonal column such that the prism angle is 60° for rays incident on its side or 90° for rays incident on its end. For $n = 1.31$, Eq. (41) yields a minimum deviation angle of about 46° for $\Delta = 90°$. Yet, although 46° halos are possible, they are seen much less frequently than 22° halos. Plates cannot give distinct 46° halos although columns can. Yet they must be solid and most columns have hollow ends. Moreover, the range of sun elevations is restricted.

Like the green flash, ice-crystal phenomena are not intrinsically rare. Halos and sun dogs can be seen frequently—once you know what to look for. Neuberger (1951) reports that halos were observed in State College, Pennsylvania, an average of 74 days a year over a 16-year period, with extremes of 29 and 152 halos a year. Although the 22° halo was by far the most frequently seen display, ice-crystal displays of all kinds were seen, on average, more often than once every 4 days at a location not especially blessed with clear skies. Although thin clouds are necessary for ice-crystal displays, clouds thick enough to obscure the sun are their bane.

9 CLOUDS

Although scattering by isolated particles can be studied in the laboratory, particles in the atmosphere occur in crowds (sometimes called clouds). Implicit in the previous two sections is the assumption that each particle is illuminated solely by incident sunlight; the particles do not illuminate each other to an appreciable degree. That is, clouds of water droplets or ice grains were assumed to be optically thin, hence multiple scattering was negligible. Yet the term *cloud* evokes fluffy white objects in the sky or perhaps an overcast sky on a gloomy day. For such clouds, multiple scattering is not negligible, it is the major determinant of their appearance. And the quantity that determines the degree of multiple scattering is optical thickness (see earlier discussion in Section 2).

Cloud Optical Thickness

Despite their sometimes solid appearance, clouds are so flimsy as to be almost nonexistent—except optically. The fraction of the total cloud volume occupied by water substance (liquid or solid) is about 10^{-6} or less. Yet although the mass density of clouds is that of air to within a small fraction of a percent, their optical thickness (per unit physical thickness) is much greater. The number density of air molecules is vastly greater than that of water droplets in clouds, but scattering per molecule of a cloud droplet is also much greater than scattering per air molecule (see Fig. 7).

Because a typical cloud droplet is much larger than the wavelengths of visible light, its scattering cross section is to good approximation proportional to the square of its diameter. As a consequence, the scattering coefficient [see Eq. (2)] of a cloud having a volume fraction f of droplets is approximately

$$\beta = 3f \frac{\langle d^2 \rangle}{\langle d^3 \rangle} \tag{42}$$

where the brackets indicate an average over the distribution of droplet diameters d. Unlike molecules, cloud droplets are distributed in size. Although cloud particles can be ice particles as well as water droplets, none of the results in this and the following section hinge on the assumption of spherical particles.

The optical thickness along a cloud path of physical thickness h is βh for a cloud with uniform properties. The ratio $\langle d^3 \rangle / \langle d^2 \rangle$ defines a mean droplet diameter, a typical value for which is 10 μm. For this diameter and $f = 10^{-6}$, the optical thickness per unit meter of physical thickness is about the same as the normal optical thickness of the atmosphere in the middle of the visible spectrum (see Fig. 3). Thus a cloud only 1 m thick is equivalent optically to the entire gaseous atmosphere.

A cloud with (normal) optical thickness about 10 (i.e., a physical thickness of about 100 m) is sufficient to obscure the disk of the sun. But even the thickest cloud does not transform day into night. Clouds are usually translucent, not transparent, yet not completely opaque.

The scattering coefficient of cloud droplets, in contrast with that of air molecules, is more or less independent of wavelength. This is often invoked as the cause of the colorlessness of clouds. Yet wavelength independence of scattering by a single particle is only sufficient, not necessary, for wavelength independence of scattering by a cloud of particles (see discussion in Section 2). Any cloud that is optically thick and composed of particles for which absorption is negligible is white upon illumination by white light. Although absorption by water (liquid and solid) is not identically zero at visible wavelengths, and selective absorption by water can lead to observable consequences (e.g., colors of the sea and glaciers), the appearance of all but the thickest clouds is not determined by this selective absorption.

Equation (42) is the key to the vastly different optical characteristics of clouds and of the rain for which they are the progenitors. For a fixed amount of water (as specified by the quantity f), optical thickness is inversely proportional to mean diameter. Raindrops are about 100 times larger on average than cloud droplets; hence optical thicknesses of rain shafts are correspondingly smaller. We often can see through many kilometers of intense rain whereas a small patch of fog on a well-traveled highway can result in carnage.

Givers and Takers of Light

Scattering of visible light by a single water droplet is vastly greater in the forward ($\theta < 90°$) hemisphere than in the backward ($\theta > 90°$) hemisphere (Fig. 9). But water droplets in a thick cloud illuminated by sunlight collectively scatter much more in the backward hemisphere (reflected light) than in the forward hemisphere (transmitted light). In each scattering event, incident photons are deviated, on average, only slightly, but in many scattering events most photons are deviated enough to escape from the upper boundary of the cloud. Here is an example in which the properties of an ensemble are different from those of its individual members.

Clouds seen by passengers in an airplane can be dazzling; but, if the airplane were to descend through the cloud, these same passengers might describe the cloudy sky overhead as gloomy. Clouds are both givers and takers of light. This dual role is exemplified in Figure 20, which shows the calculated diffuse downward irradiance below clouds of varying optical thickness. On an airless planet the sky would be black in all directions (except directly toward the sun). But if the sky were to be filled from horizon to horizon with a thin cloud, the brightness overhead would markedly increase. This can be observed in a partly overcast sky, where gaps between clouds

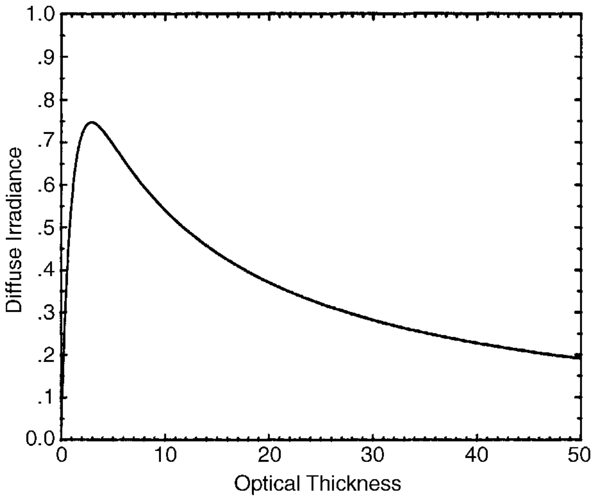


Figure 20 Computed diffuse downward irradiance below a cloud relative to the incident solar irradiance as a function of cloud optical thickness.

(blue sky) often are noticeably darker than their surroundings. As so often happens, more is not always better. Beyond a certain cloud optical thickness, the diffuse irradiance decreases. For a sufficiently thick cloud, the sky overhead can be darker than the clear sky.

Why are clouds bright? Why are they dark? No inclusive one-line answers can be given to these questions. Better to ask, Why is that particular cloud bright? Why is that particular cloud dark? Each observation must be treated individually, generalizations are risky. Moreover, we must keep in mind the difference between brightness and radiance when addressing the queries of human observers. Brightness is a sensation that is a property not only of the object observed but of its surroundings as well. If the luminance of an object is appreciably greater than that of its surroundings, we call the object bright. If the luminance is appreciably less, we call the object dark. But these are relative rather than absolute terms.

Two clouds, identical in all respects, including illumination, may still appear different because they are seen against different backgrounds, a cloud against the horizon sky appearing darker than when seen against the zenith sky.

Of two clouds under identical illumination, the smaller (optically) will be less bright. If an even larger cloud were to hove into view, the cloud that formerly had been described as white might be demoted to gray.

With the sun below the horizon, two identical clouds at markedly different elevations might appear quite different in brightness, the lower cloud being shadowed from direct illumination by sunlight.

A striking example of dark clouds can sometimes be seen well after the sun has set. Low-lying clouds that are not illuminated by direct sunlight but are seen against the faint twilight sky may be relatively so dark as to seem like ink blotches.

Because dark objects of our everyday lives usually owe their darkness to absorption, nonsense about dark clouds is rife: They are caused by pollution or soot. Yet of all the reasons that clouds are sometimes seen to be dark or even black, absorption is not among them.

GLOSSARY

Airlight: Light resulting from scattering by all atmospheric molecules and particles along a line of sight.

Antisolar Point: Direction opposite the sun.

Astronomical Horizon: Horizontal direction determined by a bubble level.

Brightness: Attribute of sensation by which an observer is aware of differences of luminace (definition recommended by the 1922 Optical Society of America Committee on Colorimetry).

Contrast Threshold: Minimum relative luminance difference that can be perceived by the human observer.

Inferior Mirage: Mirage in which images are displaced downward.

Irradiance: Radiant power crossing unit area in a hemisphere of directions.

Lapse Rate: Rate at which a physical property of the atmosphere (usually temperature) decreases with height.

Luminance: Radiance integrated over the visible spectrum and weighted by the spectral response of the human observer. Also sometimes called *photometric brightness*.

Mirage: Image appreciably different from what it would be in the absence of atmospheric refraction.

Neutral Point: Direction in the sky for which the light is unpolarized.

Normal Optical Thickness: Optical thickness along a radial path from the surface of Earth to infinity.

Optical Thickness: The thickness of a scattering medium measured in units of photon mean free paths. Optical thicknesses are dimensionless.

Radiance: Radiant power crossing a unit area and confined to a unit solid angle about a particular direction.

Scale Height: Vertical distance over which a physical property of the atmosphere is reduced to $1/e$ of its value.

Scattering Angle: Angle between incident and scattered waves.

Scattering Coefficient: Product of scattering cross section and number density of scatterers.

Scattering Cross Section: Effective area of a scatterer for removal of light from a beam by scattering.

Scattering Plane: Plane determined by incident and scattered waves.

Solar Point: Direction toward the sun.

Superior Mirage: Mirage in which images are displaced upward.

Tangential Optical Thickness: Optical thickness through the atmosphere along a horizon path.

REFERENCES

*Bryant, H. C., and A. J. Cox (1966). *J. Opt. Soc. Am.* **56**, 1529–1532.

Einstein, A. (1910). *Ann. der Physik* **33**, 175. English translation in J. Alexander (Ed.), *Colloid Chemistry*, Vol. I. New York, Chemical Catalog Company, 1926, p. 323.

Fraser, A. B. (1972). *J. Atmos. Sci.* **29**, 211–212.

*Fraser, A. B. (1975). *Atmosphere* **13**, 1–10.

*Fraser, A. B. (1979). *J. Opt. Soc. Am.* **69**, 1112–1118.

*Fraser, A. B. (1983). *J. Opt. Soc. Am.* **73**, 1626–1628.

Lee, R. (1991). *Appl. Opt.* **30**, 3401–3407.

Lee, R., and A. Fraser (1990). *New Scientist* **127** (1 September), 40–42.

*Hulburt, E. O. (1953). *J. Opt. Soc. Am.* **43**, 113–118.

Möller, F. (1972). "Radiation in the Atmosphere," in D. P. McIntyre (Ed.), *Meteorological Challenges: A History*, Ottawa, Information Canada, p. 43.

Neuberger, H. (1951). *Introduction to Physical Meteorology*, University Park, College of Mineral Industries, Pennsylvania State University.

*Nussenzveig, H. M. (1979). *J. Opt. Soc. Am.* **69**, 1068–1079.

*Penndorf, R. (1957). *J. Opt. Soc. Am.* **47**, 176–182.

Pledgley, E. (1986). *Weather* **41**, 401.

Takano, Y., and S. Asano (1983). *J. Meteor. Soc. Jpn.* **61**, 289–300.

Thekaekara, M. P., and A. J. Drummond (1971). *Nature Phys. Sci.* **229**, 6–9.

*van de Hulst, H. C. (1947). *J. Opt. Soc. Am.* **37**, 16–22.

van de Hulst, H. C. (1957). *Light Scattering by Small Particles*, New York, Wiley-Interscience.

von Frisch, K. (1971) *Bees: Their Vision, Chemical Senses, and Language*, Rev. Ed., Ithaca, New York, Cornell University Press, p. 116.

Young, A. T. (1982). *Physics Today*, January, 2–8.

Zimm, B. H. (1945). *J. Chem. Phys.* **13**, 141–145.9

BIBLIOGRAPHY

Many of the seminal studies in atmospheric optics, including those by Lord Rayleigh, are bound together in *Selected Papers on Scattering in the Atmosphere*, C. F. Bohren (Ed.), Bellingham, WA, SPIE Optical Engineering Press, 1989. References marked with an asterisk are in this collection.

M. Minnaert's *The Nature of Light and Colour in the Open Air*, New York, Dover, 1954, is the Bible for those interested in atmospheric optics. Like accounts of natural phenomena in the Bible, those in Minnaert's book are not always correct, despite which, again like the Bible, it has been and will continue to be a source of inspiration.

A history of light scattering, "From Leonardo to the Graser: Light Scattering in Historical Perspective," was published serially by J. D. Hey in *South African Journal of Science*, **79**, January 1983, 11–27; **79**, August 1983, 310–324; **81**, February 1985, 77–91; **81**, October 1985, 601–613; **82**, July 1986, 356–360. The history of the rainbow is recounted by C. B. Boyer, *The Rainbow*, Princeton, NJ, Princeton University Press, 1987.

Special issues of *Journal of the Optical Society of America* (August 1979 and December 1983) and *Applied Optics* (August 20, 1991) are devoted to atmospheric optics.

Several monographs on light scattering by particles are relevant to and contain examples drawn from atmospheric optics: H. C. van de Hulst, *Light Scattering by Small Particles*, New York, Wiley-Interscience, 1957 (reprinted by Dover, 1981); D. Deirmendjian, *Electromagnetic Scattering on Polydispersions*, New York, Elsevier, 1969; M. Kerker, *The Scattering of Light and Other Electromagnetic Radiation*, New York, Academic, 1969; C. F. Bohren and D. R. Huffman, *Light Scattering by Small Particles*, New York, Wiley-Interscience, 1983; H. M. Nussenzveig, *Diffraction Effects in Semiclassical Scattering*, Cambridge, Cambridge University Press, 1992.

The following books are devoted to a wide range of topics in atmospheric optics: R. A. R. Tricker, *Introduction to Meteorological Optics*, New York, Elsevier, 1970; E. J. McCartney, *Optics of the Atmosphere*, New York, Wiley, 1976; R. Greenler, *Rainbows, Halos, and Glories*, Cambridge, Cambridge University Press, 1980. Monographs of more limited scope are those by W. E. K. Middleton, *Vision Through the Atmosphere*, Toronto, University of Toronto Press, 1952; D. J. K. O'Connell, *The Green Flash and Other Low Sun Phenomena*, Amsterdam, North Holland, 1958; G. V. Rozenberg, *Twilight: A Study in Atmospheric Optics*, New York, Plenum, 1966; S. T. Henderson, *Daylight and Its Spectrum*, 2nd ed., New York, Wiley, 1977; R. A. R. Tricker, *Ice Crystal Haloes*, Washington, DC, Optical Society of America, 1979; G. P. Können, *Polarized Light in Nature*, Cambridge, Cambridge University Press, 1985.

Although not devoted exclusively to atmospheric optics, W. J. Humphreys, *Physics of the Air*, New York, Dover, 1964, contains a few relevant chapters. Two popular science books on simple experiments in atmospheric physics are heavily weighted toward atmospheric optics: C. F. Bohren, *Clouds in a Glass of Beer*, New York, Wiley, 1987; C. F. Bohren, *What Light Through Yonder Window Breaks?* New York, Wiley, 1991.

For an expository article on colors of the sky see C. F. Bohren and A. B. Fraser, *The Physics Teacher*, May, 267–272 (1985).

An elementary treatment of the coherence properties of light waves was given by A. T. Forrester, *Am. J. Phys.* **24**, 192–196 (1956). This journal also published an expository article on the observable consequences of multiple scattering of light: C. F. Bohren, *Am. J. Phys.*, **55**, 524–533 (1987).

Although a book devoted exclusively to atmospheric refraction has yet to be published, an elementary yet thorough treatment of mirages was given by A. B. Fraser and W. H. Mach in *Scientific American*, January, 102–111 (1976).

Colorimetry, the often (and unjustly) neglected component of atmospheric optics, is treated in, for example, *The Science of Color*, Washington, DC, Optical Society of America, 1963; F. W. Billmeyer and M. Saltzman, *Principles of Color Technology*, 2nd ed., New York, Wiley-Interscience, 1981; D. L. MacAdam, *Color Measurement*, 2nd ed., Berlin, Springer, 1985.

Understanding atmospheric optical phenomena is not possible without acquiring at least some knowledge of the properties of the particles responsible for them. To this end, the following are recommended: H. R. Pruppacher and J. D. Klett, *Microphysics of Clouds and Precipitation*, Dordrecht, Holland, Reidel, 1980; S. A. Twomey, *Atmospheric Aerosols*, New York, Elsevier, 1977.

SECTION 4

WEATHER SYSTEMS

Contributing Editor: John W. Nielsen-Gammon

CHAPTER 25

OVERVIEW OF WEATHER SYSTEMS

JOHN W. NIELSEN-GAMMON

1 INTRODUCTION

Meteorology is the only science in which it is common for a practitioner to appear every day on the evening news. Weather forecasting has immediate interest to people because it affects their day-to-day lives. The impact is great because weather itself changes from day to day. If weather were all good, or all bad, perhaps no one would notice—when's the last time you thought about the quality of the ground under your feet?

2 FOUNDATIONS OF WEATHER FORECASTING

The attempt to understand the weather is driven largely by the desire to be able to forecast the weather. However, the atmosphere is a complicated dynamical system, and an understanding of even the basic physical principles came only recently. Before that, scientists either attempted to understand the physical causes of weather events or recorded "weather signs," rules for weather forecasting. Because the physical understanding was incorrect, and most weather signs had no physical basis, neither approach was very successful. True progress in the study of the atmosphere that would lead to success in weather forecasting required a complete understanding of physical principles and a comprehensive set of observations of the atmosphere.

Aristotle's *Meteorologica*, written around 340 BC, was the leading book on meteorological principles in the Western world for two millenia. With his theories on the fundamental physical nature of the universe, Aristotle sought to explain a wide range of meteorological phenomena. For example, he argued that wind is

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

caused by the hot, dry exhalation of the earth when struck by sunlight. A leading book on weather signs, *De Signis Tempestatum*, written by Aristotle's pupil Theophrastus, focused on such maxims as "A dog rolling on the ground is a sign of violent storm" and "Reddish sky at sunrise foretells rain." Since some of these maxims are still in use today, it appears that Theophrastus's work has outlived that of Aristotle.

Astronomers during the Middle Ages often argued that weather could be predicted by careful examination of the sky, clouds, and stars. This method of weather forecasting is sometimes successful, but it more closely resembled astrology than modern-day meteorology. Astrological forecasting schemes grew more advanced and intricate with time, while the scientific study of weather systems made little or no progress.

Scientific progress required basic observational data for describing weather systems. Only with good observations could theories be tested and the correct nature of the atmosphere be determined. The primary tools for observation of weather conditions are the thermometer (for temperature), the barometer (for air pressure), and the hygrometer (for humidity). Prior to and during the Middle Ages, these basic weather instruments did not exist. These instruments were invented in the sixteenth and seventeenth centuries and refined throughout the eighteenth century. The resulting observations of the atmosphere eventually established the basis for meteorological breakthroughs of the nineteenth and twentieth centuries.

Meanwhile, Isaac Newton had laid the foundation for the development of physical laws governing the motion of objects. The great mathematician Leonhard Euler of Germany, in 1755, rewrote those laws in a form that applied to a continuous fluid such as air or water. But Euler's laws were incomplete. They described how air can be neither created nor destroyed and how air accelerates (and wind blows) in response to forces acting on it. Missing from Euler's equations were the crucial relationship between temperature and pressure and the consequences of evaporation and condensation.

With the new instruments, the study of air became an experimental science. Through numerous laboratory experiments, scientists gradually discovered the fundamental physical laws governing the behavior of gases. In particular, the missing key relationship, the first law of thermodynamics, became known from experiments during the first half of the nineteenth century.

Now that the basic physical laws were known, a comprehensive set of observations of the atmosphere was necessary for further progress. Weather observations from a single careful observer were insufficient, but dozens of observations taken simultaneously across Europe or the United States gave a much clearer picture of the distribution of weather elements within a winter storm. Such "synoptic" observations were also used directly in weather forecasting. As rapid long-distance communication became possible through the telegraph, individual weather systems could be tracked and therefore forecasted. The coordinated observations that have taken place since the middle of the nineteenth century form the foundation of modern weather forecasting and our understanding of weather systems of all scales.

3 DEVELOPING AN UNDERSTANDING OF WEATHER SYSTEMS

Progress in understanding meteorology in the nineteenth century was largely based on thermodynamics. James Pollard Espy of the United States conducted a variety of laboratory experiments involving the condensation of water vapor in air. In 1841, based on his experiments and those of others, Espy correctly described the basic principle of thunderstorm formation: Condensation of water vapor within ascending air causes the air to become warmer than its surroundings and thus to continue rising. Espy went on to assert that the resulting lower pressure near the ground accounts for the low pressure observed in large-scale storm systems.

Espy's theory lay dormant until later in the nineteenth century, when the complete first law of thermodynamics was established. With this development, scientists were able to put Espy's theory of convection on a solid mathematical footing. Karl Theodor Reye of Germany, for example, determined the specific conditions under which such ascending air would be unstable. Their ideas form our understanding of the fundamental nature of individual thunderstorms to this day.

Espy also suggested that upward motion and the release of latent heat by evaporation might be the driving force for large weather systems. The widespread development of synoptic observations finally established that surface winds spiral inward toward a large-scale low-pressure system. The strong temperature contrasts within low-pressure systems suggested that the ascent would be driven, at least initially, by a current of warm air impinging on a current of cold air and rising. The rising air, and the resulting latent heating from condensation, would cause surface pressures to fall. This theory for the cause of low-pressure systems gained rapid acceptance for both theoretical and observational reasons in the 1870s.

A leading proponent of this thermal theory of cyclones was the American William Ferrel. Earlier, Ferrel had completed the application of Euler's equations to the atmosphere by formally including the effect of Earth's rotation as the Coriolis force. Ferrel also saw a clear analog to the role of equatorial convection in the general circulation, which drives ascent in the tropics and descent in the subtropics.

Unfortunately for its proponents, the thermal theory was wrong. Coordinated mountaintop observations in Europe, analyzed by Julius von Hann, showed that above the ground the centers of low-pressure systems were actually cooler than the centers of high-pressure systems. This observation illustrates the self-limiting nature of moist convection: The downward motion outside the convective cell causes warming of that surrounding air, so that eventually the instability is eliminated.

4 WEATHER AND ENERGY

The above discussion leads to a key question for the understanding of weather systems: How do weather systems grow and maintain themselves? As noted above, the basic dynamics of ordinary moist convection were well established, both conceptually and mathematically, by the end of the nineteenth century.

Convection was also understood to be an important driver of the general circulation, or at least the tropical and subtropical circulation known as the Hadley cells.

By the turn of the century, the work of Hann and others showed that convection was not the energy source for extratropical cyclones, the midlatitude migratory low-pressure systems. Attention shifted from the role of latent heat release to the role of the large horizontal temperature gradients that were systematically observed within midlatitude low-pressure systems. In 1903, Max Margules, born in the Ukraine and working in Austria, calculated the amount of kinetic energy that could be obtained from the rising of hot air and the sinking of cold air in a low-pressure system. He found that the amount of energy that could be converted into kinetic energy was comparable to the actual amount of kinetic energy in a mature storm system. Margules had identified the correct energy source for extratropical cyclones. The process by which cyclones form and move was described by researchers working under Vilhelm Bjerknes in Bergen, Norway, shortly after World War I. Finally, a comprehensive theory by Jule Charney in 1947 explained that the structure and intensity of low-pressure systems was a consequence of the growth of unstable eddies on a large-scale horizontal temperature gradient.

Hurricanes did not have the prominent horizontal temperature variations found in midlatitude weather systems; indeed, many hurricanes seemed almost perfectly symmetrical. The prominence of convection throughout the hurricane, particularly within the eyewall, suggested that convection was fundamentally important in the development and maintenance of hurricanes. But how did the hurricane organize itself or maintain itself against large-scale subsidence within its environment? The currently accepted theory, published by Kerry Emanuel and Richard Rotunno in 1986 and 1987, relies on radiative cooling of the subsiding air at large distances from the hurricane. Emanuel also noted that as air spirals inward toward the eye it would become more unstable because it would be gaining heat and moisture from the sea surface at a progressively lower pressure.

Theories satisfactorily describing organized convection such as mesoscale convective systems and supercells also have been slow to develop. One reason for this is that no observing systems could accurately describe the structure of organized convection until the development of radar. Indeed, the term “mesoscale” was coined specifically to refer to in-between sizes (10 to 500 km) that were too large to be adequately observed at single locations and too small to be resolved by existing observing networks. The widespread use of weather radar in the 1950s helped fill the observation gap. The development of Doppler radar (for measuring winds within precipitating systems) for research purposes in the 1970s and as part of a national network in the 1990s helped even more. With comprehensive radar observations, it became clear that the long lifetime of organized convection was due to a storm keeping its updraft close to the leading edge of its cold, low-level outflow. The storm could then take advantage of the ascent caused by the cold air undercutting warm air without having its supply of warm air cut off completely. But even radar was not sufficient. The development of the first dynamical descriptions of supercells and squall lines in the 1980s, by Joseph Klemp, Richard Rotunno, Robert Wilhelmson, and Morris Weisman, relied upon numerical computer-generated simulations of the phenomena to provide an artificial data set unavailable from observations.

While the above discussion has focused on self-contained weather systems, certain "weather producers" may be thought of as byproducts of these weather systems. Many forms of severe weather, such as hail, lightning, and tornadoes, are essentially side effects of convection (organized or otherwise) but have little or no influence on the convection itself. Other mesoscale phenomena, such as sea breezes, upslope precipitation, and downslope windstorms, do not represent instabilities at all. Instead they are called "forced" weather phenomena because they are driven by such external features as solar heating gradients and topographic obstacles to the large-scale flow.

5 FORECASTING

The massive strides in weather forecasting during the past 50 years are due in large part to our growing understanding of the nature and dynamics of important weather systems. Forecasting of many phenomena has evolved from an exercise in extrapolation to specific predictions of the evolution of weather systems to computer simulations that accurately forecast the evolution of weather systems. Currently, the most skillful weather forecasts involve the prediction of large-scale extratropical weather systems.

One very important advance in our ability to forecast large-scale weather systems was the development of the omega equation, first presented in simplified form by Richard Sutcliffe in 1947. The omega equation is named after the Greek symbol that represents the change in pressure of an air parcel. In its original form, the omega equation is based upon a simplification of the equations of motion that assumes that all motions are large-scale and evolve slowly. When this approximation is made, it is possible to diagnose vertical motion entirely from large-scale pressure and temperature variations. Vertical motion is important for weather prediction because it is fundamentally related to clouds and precipitation, but large-scale vertical motion also can be used to infer the evolution of low-level and upper-level wind patterns. Vertical motion became the key to understanding the evolution of weather. In forecast offices, maps were designed and widely distributed that made it easy to look at the large-scale fields and diagnose vertical motion.

A second important advance was the development of numerical weather prediction, or NWP. NWP, discussed more extensively in the chapter by Kalnay in the dynamics part of this *Handbook*, has completely transformed the forecasting of large-scale weather systems. In the past, forecasters relied on their diagnosis of the current weather patterns to make forecasts. Nowadays, computers can make much more accurate forecasts than humans alone, and the best possible forecast is obtained by humans working with computer forecast output. The forecaster's task has become one of identifying likely errors in the model forecast, based on the forecaster's knowledge of systematic errors within the model, errors in the initial analysis, and computer forecast scenarios that run counter to experience.

In contrast to large-scale weather systems, individual convective storms are not directly simulated, nor accurately forecasted, by the numerical weather prediction models currently in use. Forecasting convective storms still relies heavily on extrapolation and a sound knowledge of what sorts of storms are likely to be produced by certain larger-scale weather conditions. The modern era of forecasting severe weather began in 1948, when two Air Force weather forecasters (Maj. Ernest Fawbush and Capt. Robert Miller) had responsibility for forecasting at Tinker Air Force Base near Oklahoma City. After a damaging tornado struck the base in 1948, the two meteorologists were given the task of identifying the days when tornadoes were likely to strike the base. We now know that forecasting the specific path of a tornado is essentially impossible, but when a similar low-pressure system evolved a few days later, the forecasters were compelled by the base commander to use a newly minted severe weather warning system to issue a forecast of a possible tornado at the base. Amazingly, the forecast came true, and preventive measures taken before the second tornado struck the base prevented considerable casualties to aircraft and personnel. This weather forecast, possibly the most serendipitous in the history of humankind, led directly to modern tools for diagnosing the likelihood of severe weather from large-scale conditions and laid the foundation for modern tornado forecasting.

6 CONCLUSION

This historical review has emphasized that progress in our understanding of the weather has required four ingredients: the need to understand the science of the atmosphere before one can hope to make accurate forecasts; a complete knowledge of the basic physical underpinnings of atmospheric behavior; observations adequate to test the theories and point the way toward new ones; and numerical models to provide simulations of the atmosphere more complete than any set of observations. The following chapters of this part of the *Handbook* discuss the current understanding of the weather phenomena discussed above, improvements in which have lead to vastly improved weather forecasts.

BIBLIOGRAPHY

Frisinger, H. H. (1983). *The History of Meteorology: To 1800*, Boston, American Meteorological Society.

Kutzbach, G. (1979). *The Thermal Theory of Cyclones: A History of Meteorological Thought in the Nineteenth Century*, Boston, American Meteorological Society.

Miller, R. C., and C. A. Crisp (1999). The First Operational Tornado Forecast—Twenty Million to One, *Weather and Forecasting*, **14**(4), 479–483.

CHAPTER 26

LARGE-SCALE ATMOSPHERIC SYSTEMS

JOHN W. NIELSEN-GAMMON

1 INTRODUCTION

The structure and evolution of large-scale extratropical weather systems is dominated by a fundamental contradiction: The airflow within such systems represents an almost exact balance among the forces affecting each air parcel, but the slight departures from balance are essential for vertical motion and the resulting clouds and precipitation, as well as changes in intensity of the systems. Furthermore, balanced flow could not be maintained without slight departures from balance. This section will explore the morphology and dynamics of large-scale weather systems and will never stray far from the concepts of balance and adjustment to balance.

The figures in this Chapter of the *Handbook* will typically use pressure as a vertical coordinate, in keeping with the standard practice of displaying all meteorological information above Earth's surface on levels of constant pressure. Constant-pressure surfaces are so nearly flat that they can be treated as horizontal for most purposes. Vertical motion in this coordinate system, represented as ω, is defined as dp/dt rather than dz/dt, and since pressure increases downward rather than upward, negative ω corresponds to upward motion.

2 HORIZONTAL BALANCE

In its simplest form, the balance of forces in the horizontal plane is geostrophic balance (see chapter by Salby in this *Handbook*) between the Coriolis force and the

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

horizontal pressure gradient force. Since the Coriolis force is proportional to the wind speed and directed to the right of the wind direction (to the left in the Southern Hemisphere), a balance of forces can only be attained if the horizontal pressure gradient force is directed to the left of the wind direction (right in the Southern Hemisphere), with sufficient speed that the magnitude of the Coriolis force equals the magnitude of the horizontal pressure gradient force. Thus, balanced flow is parallel to the isobars (in Cartesian coordinates) or height contours (in pressure coordinates) with a strength proportional to the horizontal pressure gradient. This balance (and the others described in this chapter) is generally stable, in the sense that air parcels initially out of balance will tend to approach balance, on a time scale of a few hours.

The full horizontal wind may be divided into geostrophic and ageostrophic components. The geostrophic wind is, by definition, that wind that would represent an exact balance between the horizontal pressure gradient force and the Coriolis force. For large-scale extratropical weather systems, the geostrophic wind very nearly equals the total wind. The ageostrophic wind is associated with an imbalance of forces and, therefore, acceleration. It may be thought of as that portion of the total wind whose associated Coriolis force is not balanced by a horizontal pressure gradient force. Thus, acceleration is proportional to the magnitude of the ageostrophic wind and is directed in the same direction as the Coriolis force, to the right of the ageostrophic wind (to the left in the Southern Hemisphere).

Other forms of balance may be defined that include nonzero ageostrophic winds. For example, circular flow represents a balance of three forces: the horizontal pressure gradient force, the Coriolis force, and the centrifugal force. In this circumstance, the Coriolis force associated with the geostrophic wind balances the horizontal pressure gradient force, and the Coriolis force associated with the ageostrophic wind balances the centrifugal force. The wind remains parallel to the isobars or height contours but is weaker (subgeostrophic) in a cyclonic vortex and stronger (supergeostrophic) in an anticyclonic vortex. Since ageostrophic winds can often be associated with balanced flow, it is often more useful to subdivide the wind as divergent and nondivergent rather than geostrophic and ageostrophic. The geostrophic wind is always nondivergent (except for effects due to the variation of the Coriolis parameter with latitude), and the divergent wind is directly related to vertical motion through the continuity equation.

3 VERTICAL BALANCE

The vertical balance of forces, known as hydrostratic balance, is between the vertical pressure gradient force and gravity. This balance may be assumed to be maintained exactly for extratropical weather systems. Indeed, the balance is so close that vertical accelerations must be deduced by indirect means, through diagnosis of the divergent wind.

As horizontal balance introduced a close connection between the instantaneous pressure and wind fields, vertical balance introduces a close connection between the instantaneous pressure and density/temperature fields. To balance gravity, the vertical pressure gradient force must be strong where air density is large and weak where density is small. Equivalently, the vertical separation between two given pressure surfaces (known as *thickness*) is small where the air is relatively cool and large where the air is relatively warm. Horizontal variations in temperature imply vertical differences in the horizontal variation of pressure; that is, the horizontal pressure gradient at one level will differ from the pressure gradient at another level if air of differing densities (temperatures) lies between the two levels. If one simultaneously assumes geostrophic (horizontal) and hydrostatic (vertical) balance, one obtains a relationship between temperature and wind known as the thermal wind law: If one represents the vertical derivative of the horizontal wind (vertical wind shear) by a vector, it will lie parallel to the isotherms with relatively warmer temperature to the right (to the left in the Southern Hemisphere). Thus, temperature, pressure, and wind are all constrained by each other in the limit of exact balance.

Vertical motion carries a strong influence on atmospheric temperature. Following an air parcel within which phase changes are not taking place and which is not exchanging heat with its surroundings, upward motion causes the parcel to cool at the dry adiabatic lapse rate, 9.8°C/km, and downward motion causes an identical amount of warming. In most circumstances away from the daytime boundary layer, the atmosphere is stably stratified, meaning that the instantaneous vertical derivative of temperature is greater than $-9.8\,\mathrm{K/km}$, so that, for example, an ascending air parcel replaces an air parcel that is warmer (and less dense) than itself. Consequently, at a given level, upward motion causes cooling at a rate proportional to the magnitude of upward motion and the degree of stratification. Similarly, downward motion causes warming. This stratification is called *stable* because an ascending air parcel, finding itself cooler and more dense than its surroundings, would tend to sink back down toward its original level, and a subsiding air parcel would be warmer than its surroundings and would tend to rise.

Extratropical large-scale weather systems can be directly affected by condensation of water vapor within ascending, cooling air parcels. This condensation releases latent heat, causing the air parcel to be warmer than it would have been without the release of latent heat. As a result, in areas of saturated ascent, the temperature at a given level does not fall as rapidly as it would in ascending dry air and may in fact rise if the atmosphere is unstable to moist convection. Condensation, and the generation of clouds and precipitation, represents the primary internal heat source for large-scale weather systems.

4 POTENTIAL VORTICITY

When the atmosphere is in approximate horizontal and vertical balance, the wind and mass fields are tightly interconnected. The distribution of a single mass or momentum variable may be used as a starting point to infer the distribution of all other such variables. We choose to focus on the variable known as potential vorticity (approximately equal to the vorticity times the stratification) for three reasons: (1) it tends to have a simple three-dimensional distribution; (2) it is conserved in the absence of diabatic and frictional processes, so it tends to have a simple evolution; (3) it is of direct relevance to the fundamental dynamics of atmospheric behavior.

Potential vorticity is related to the three-dimensional mass and wind fields through equations that involve three-dimensional inverse second-order operators whose exact form depends on the level of approximation. One consequence of the inverse Laplacian-like operator is that potential vorticity variations in one location are diagnostically related to variations in the height and wind fields at a distance. Indeed, the relationship between perturbation (deviation from some standard mean state) potential vorticity and perturbation height (or pressure) is mathematically and conceptually analogous to the relationship between electric charge and electric potential. Areas of locally high potential vorticity correspond to locally (and regionally) low heights, and therefore cyclones, while areas of locally low potential vorticity correspond to anticyclones.

While it is possible for isolated regions of high potential vorticity to form anywhere in the atmosphere, many potential vorticity anomalies are found at the tropopause. There, midlatitude potential vorticity changes suddenly from around $0.5 \times 10^{-6}\text{m}^2/\text{s K kg}$ (henceforth, 0.5 PVU) to around 5 PVU. Horizontal and vertical displacements in the position of the tropopause lead to sizable potential vorticity variations.

Figure 1 shows the wind and mass fields associated with a prototypical axially symmetric potential vorticity anomaly at the tropopause. This cyclonic anomaly (positive in the Northern Hemisphere) is associated with a cyclonic circulation that is strongest at the level of the anomaly and decreases downward and away from the anomaly. Temperatures are anomalously warm above the vortex (isentropic surfaces are deflected downward) and anomalously cold below the vortex. While consistent with the given simple potential vorticity distribution, the wind and mass fields are also in vertical and horizontal balance. The general characteristics of the

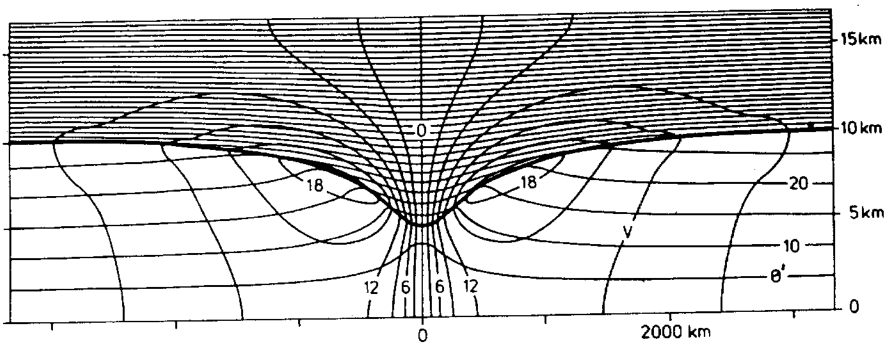


Figure 1 Wind (V) and potential temperature (θ') structure associated with a balanced axisymmetric vortex at the tropopause. Wind speeds are in m/s and are out of the page on the left and into the page on the right. Potential temperature (in K) is the departure from a uniform surface potential temperature. (From Thorpe, 1986).

wind and temperature distribution are common to cyclonic potential vorticity anomalies, with analogous atmospheric states for anticyclonic potential vorticity anomalies. Winds and temperatures associated with complex potential vorticity distributions may, to a first approximation, be recovered by treating the potential vorticity distribution as an assemblage of discrete anomalies, each with its own impact on the nearby wind and temperature fields.

5 FRONTS

Fronts are elongated zones of strong temperature gradient separating regions of relatively weak temperature gradient. Fronts can form as a consequence of deformation, convergence, and differential heating or cooling. Fronts produced by differential heating or cooling tend to be relatively small in scale, such as a gust front (produced by evaporative cooling) or a sea breeze front (produced by daytime heating over land). Such mesoscale fronts are discussed in Chapter 30 (Brooks et al). Here, we focus on synoptic-scale fronts, generally formed by a combination of deformation and convergence. Depending on their motion and structure, synoptic-scale fronts are called warm fronts, cold fronts, stationary fronts, or occluded fronts.

Characteristics of Fronts

While fronts are defined in terms of temperature, other atmospheric variables are also affected by the presence of a front. Often, the air on either side of a front originated from widely different locations, and a sharp humidity gradient will be present as well. Pressure and wind are also affected by the density contrast across a front. A pressure trough, wind shift, and vorticity maximum tend to be present along the warm side of the front. The temperature gradient also tends to be strongest at the warm edge of the front, and it is at this edge of the strong temperature gradient that the front itself is deemed to be located. The temperature gradient can be so strong that most of the temperature variation across the frontal zone takes place over a distance as small as 100 m. Under many circumstances, the leading edge of the frontal zone is marked by a discontinuity of low cloud or even a long, narrow band of low cloud apparent in visible satellite imagery and known as a rope cloud.

Cyclones (here used in the common meteorological sense to refer to synoptic-scale low-pressure systems) have a symbiotic relationship to fronts. Cyclones frequently form along preexisting surface frontal zones. If a preexisting frontal zone is not present, however, a cyclone is quite capable of generating its own surface fronts. A typical life cycle is shown in Figure 2. As the cyclone intensifies, both a warm and cold front are present, which may or may not intersect in the vicinity of the surface low. The warm front is marked by geostrophic wind from warm to cold air, and the cold front is marked by the opposite. As the cyclone matures, the cold front moves ahead of the low, and the former warm front between the cyclone and the

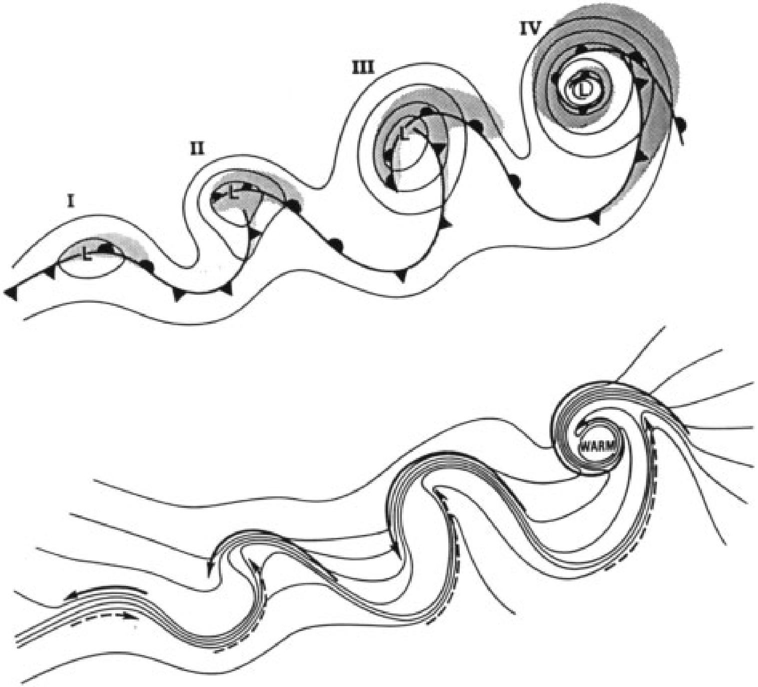


Figure 2 Typical life cycle of an intensifying offshore extratropical cyclone: (I) incipient frontal cyclone, (II) frontal fracture, (III) T-bone structure, and (IV) warm-core seclusion. *Top panel:* sea level pressure (solid), fronts (bold), and clouds (shaded). *Bottom panel:* temperature (solid), cold air currents (solid arrows), and warm air currents (dashed arrows). (From Shapiro and Keyser, 1990).

intersection of warm and cold fronts becomes the occluded front, marked with strong temperature gradients on both sides and the warmest air along the front.

Fronts are typically associated with a variety of clouds and precipitation. The precipitation may be ahead of the front, behind it, or both. Some distinguish between *anafronts*, with an updraft sloping upward toward the cold air and clouds and precipitation behind the front, and *katafronts*, with an updraft sloping upward toward the warm air and the bulk of the precipitation within the warm air. The distribution of clouds and precipitation associated with the surface fronts of a developing cyclone is shown in Figure 3.

Surface fronts tend to be strongest (i.e., possess the largest temperature gradient) near the ground and decrease in intensity upward. Typically, a surface front will lose its frontal characteristics by an altitude of 4 km or 625 mb (Fig. 4), although deeper fronts have been observed. Fronts are also favored at heights of 6 to 9 km. Such fronts are called upper-level fronts. An especially strong upper-level front is shown in Figure 5. On rare occasions, a surface and upper-level front may merge, yielding a single frontal zone stretching from the ground to the top of the troposphere (Fig. 6).

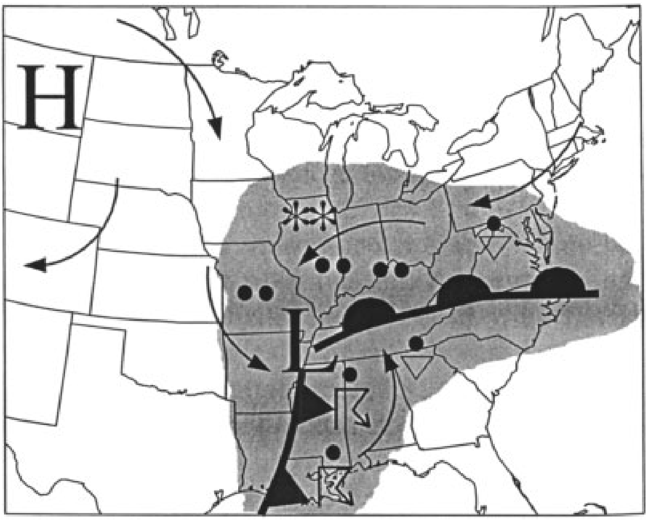


Figure 3 Distribution of clouds and fronts associated with a prototypical developing extratropical cyclone. Surface weather is represented by standard meteorological symbols.

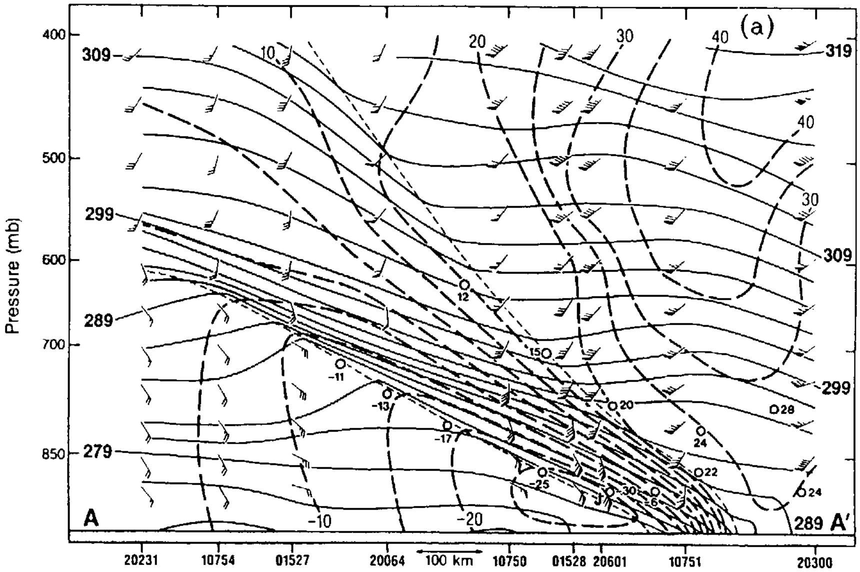


Figure 4 Cross section through strong front observed over the central Pacific on March 9, 1987. Data taken from dropwinsondes (potential temperature in K, solid lines; wind barbs, 1 long barb = 10 knots) and airborne Doppler radar (open circles with section-normal wind component in m/s beneath). Wind analyzed with bold contours in m/s; negative directed out of the page. (From Shapiro and Keyser, 1990).

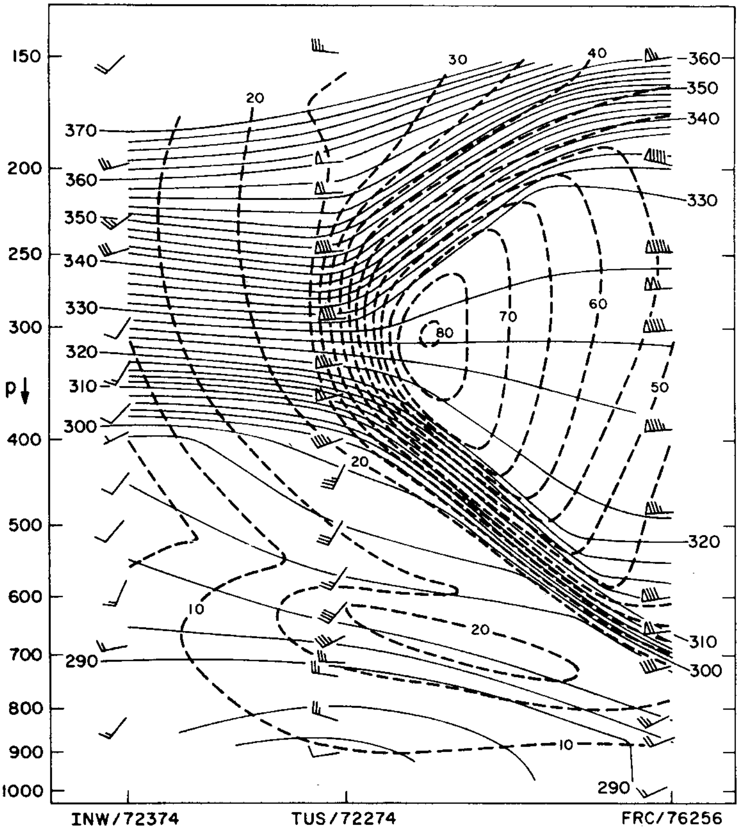


Figure 5 Cross section through a strong upper-level front, showing potential temperature (K, solid), observed wind (barbs, as in Fig. 4), and section-normal wind component (dashed, m/s, positive into the page). (From Shapiro, 1991).

In Figures 4 to 6, contours of potential temperature are shown rather than temperature. The fronts are indicated by the zones of sloping, tightly packed contours of potential temperature. At any given pressure level, potential temperature is proportional to temperature, so concentrated gradients of potential temperature imply concentrated gradients of temperature. One advantage of using potential temperature is that frontal zones consist of isentropes (lines of constant potential temperature) packed horizontally or vertically. A second advantage is that mixing can take place adiabatically along isentropes; cross-isentrope transport is only possible in conjunction with diabatic processes. Thus, the extent to which the front acts like an impenetrable wall is given by the extent to which the frontal zone is defined by a single isentrope.

In Figures 5 and 6, the frontal zone includes air from the stratosphere that has been drawn downward into the troposphere in a structure known as a tropopause

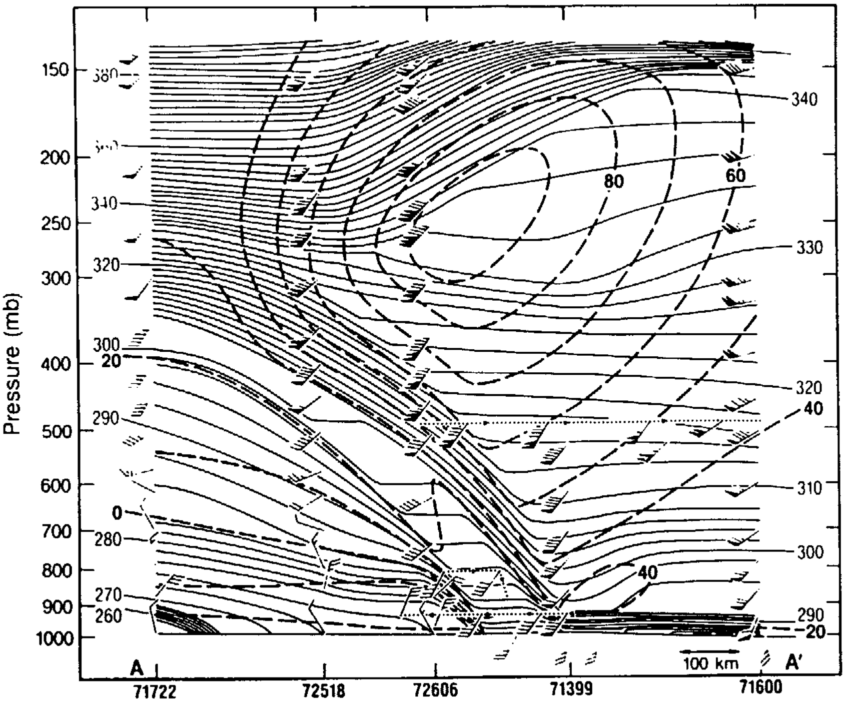


Figure 6 Cross section through a strong merged front, as in Fig. 5. Dashed line shows path of aircraft. (From Shapiro and Keyser, 1990).

fold. Tropopause folds are common to upper-level fronts; because of the large amounts of mixing that take place in the vicinity of upper-level fronts, tropopause folds are a primary mechanism for the exchange of air between the stratosphere and troposphere. The mixing associated with upper-level fronts is also one manifestation of clear-air turbulence. Extreme tropopause folding events can bring stratospheric air all the way to the ground. This is presently not thought to be dangerous, although it was a concern in the days of aboveground nuclear tests.

Dynamical Aspects of Fronts

Synoptic-scale surface fronts are often portrayed as zones of colliding air masses or war zones where battles are fought within thunderstorms. This portrayal is wrong, primarily because of the influence of the Coriolis force. Beneath a cold air mass will tend to be a region of high pressure, which in the absence of the Coriolis force would accelerate the cold air in the direction of the warm air and directly cause vertical motion. Instead, with the Coriolis force, the cold air comes into approximate balance and the wind is ultimately directed parallel to the isobars, with higher pressure to the right. Rather than two armies rushing forward toward each other, perhaps the appropriate image is of two armies nervously pacing sideways along a demilitarized zone.

In addition to the horizontal wind variations mentioned earlier, a front that is nearly in geostrophic and hydrostratic balance will be associated with significant vertical wind shear. Recall from the introduction to this section that because of the relationships between pressure and temperature on the one hand and pressure and wind on the other, a horizontal temperature gradient implies vertical shear of the horizontal wind. Since a front is a zone of concentrated temperature gradient, the vertical wind shear (or thermal wind) is large there too. The shear vector is oriented such that winds aloft will tend to blow parallel to the front with cold temperatures to the left and warm temperatures to the right. Because of the interrelationship between wind shear and temperature gradient, an upper-level jet stream is often located directly above a deep surface or upper-level front.

Because wind shear tends to improve the ability of midlatitude convection to organize itself into convective systems or supercells, the wind shear of a front is one reason severe weather tends to be located near fronts. A second reason for convection, and clouds in general, near fronts is the vertical motion that tends to be associated with fronts. This upward motion is not due to warm air being forced upward by cold air; it is a consequence of the atmosphere attempting to stay within balance.

In an intensifying synoptic-scale front, deformation and convergence is causing the magnitude of the temperature gradient to increase. At the same time, by means that will not be explained here, the deformation acts to reduce the vertical wind shear, even though the vertical wind shear would have to increase to remain within balance. The resulting imbalance between the pressure gradient force and Coriolis force produces accelerations: At low levels the air accelerates from the cold side of the front toward the warm side, and aloft the air accelerates from the warm side of the front toward the cold side. An ageostrophic circulation is thereby established, and mass continuity demands upward motion on the warm side of the front and downward motion on the cold side in order to complete the circulation cell. The horizontal ageostrophic flow, once it reaches finite magnitude, causes an acceleration to the right by the Coriolis force, eventually producing an increase in the vertical shear. Meanwhile, the vertical motion is acting to cool the air through ascent on the warm side of the front and warm it through descent on the cool side of the front, thereby acting to reduce the horizontal temperature gradient. This vertical circulation is called a *direct* circulation because relatively warm air rises and relatively cold air sinks; the opposite circulation, which is found when fronts are weakening, is called *indirect.*

Here, then, is the true “battle” within a front: The balanced flow is acting in one sense, and the unbalanced (ageostrophic) flow is acting in precisely the opposite sense! The net effect is that both the vertical wind shear and horizontal temperature gradient increase. By assuming that the ageostrophic circulation is precisely as strong as it has to be to maintain thermal wind balance, the horizontal ageostrophic and vertical motions can be diagnosed purely from the rate of frontogenesis (strengthening of the temperature gradient across an air parcel), using the so-called Sawyer–Eliassen equation. This constant adjustment, as the air attempts to keep up with thermal wind balance, is the primary cause of the vertical motion and resulting clouds and precipitation mainly on the warm side of active zones of strong temperature gradient.

The ageostrophic circulation also modifies the frontal structure. Since the vertical motions are constrained to be near zero at the ground and in the upper troposphere, the vertical motion serves to weaken the horizontal temperature gradient primarily in the middle troposphere, between 3 and 6 km, thus explaining the paucity of strong fronts at this level. Meanwhile, the low-level ageostrophic flow is directed from cold air toward warm air, where it ascends. The convergence on the warm air side of the front enhances the temperature gradient there, even as the gradient on the cold side of the front is being weakened by the same mechanism. This explains the tendency for surface fronts to have their strongest temperature gradient, or perhaps even a discontinuity, right at the warm-air edge of the frontal zone. By similar arguments, upper-level fronts ought to be strongest along the cold-air edge.

6 EXTRATROPICAL CYCLONES

Extratropical cyclones (also known as low-pressure systems) are large-scale circulations that develop spontaneously in midlatitudes. The sense of the circulation is cyclonic, or in the same direction as Earth's rotation, which is counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. Extratropical cyclones have scales ranging from hundreds to thousands of kilometers and can be associated with widespread rain and snow and high winds. Extratropical anticyclones, or high-pressure systems, are just as common and form by similar means, but receive less attention since they tend to be associated with fair weather and light winds.

Observed Characteristics

The life cycle of an extratropical cyclone has been shown in Figure 2. The general evolution shown in Figure 2 is typical for extratropical cyclones forming over the ocean and associated with a single upper-tropospheric mobile trough. Over land, warm fronts tend to be weaker and the presence of mountain ranges and coastlines strongly alters the structure of extratropical cyclones and their associated precipitation.

Rather than simply discuss the typical distribution of weather about a developing extratropical cyclone, the wide range of cyclone characteristics will be illustrated here. Weather maps and infrared satellite images showing four different extratropical cyclones affecting North America are shown in Figures 7 to 10. The first cyclone (Fig. 7) is a mature extratropical cyclone striking the coast of the northwestern United States, the second (Fig. 8) is a typical cyclone in the central United States, the third cyclone (Fig. 9) is of a type known as an *Alberta clipper*, and the fourth (Fig. 10) is a cyclone redeveloping along the East Coast of the United States.

Wind Typically, the strongest winds and pressure gradients associated with an offshore extratropical cyclone are in the northwest quadrant (southwest in the

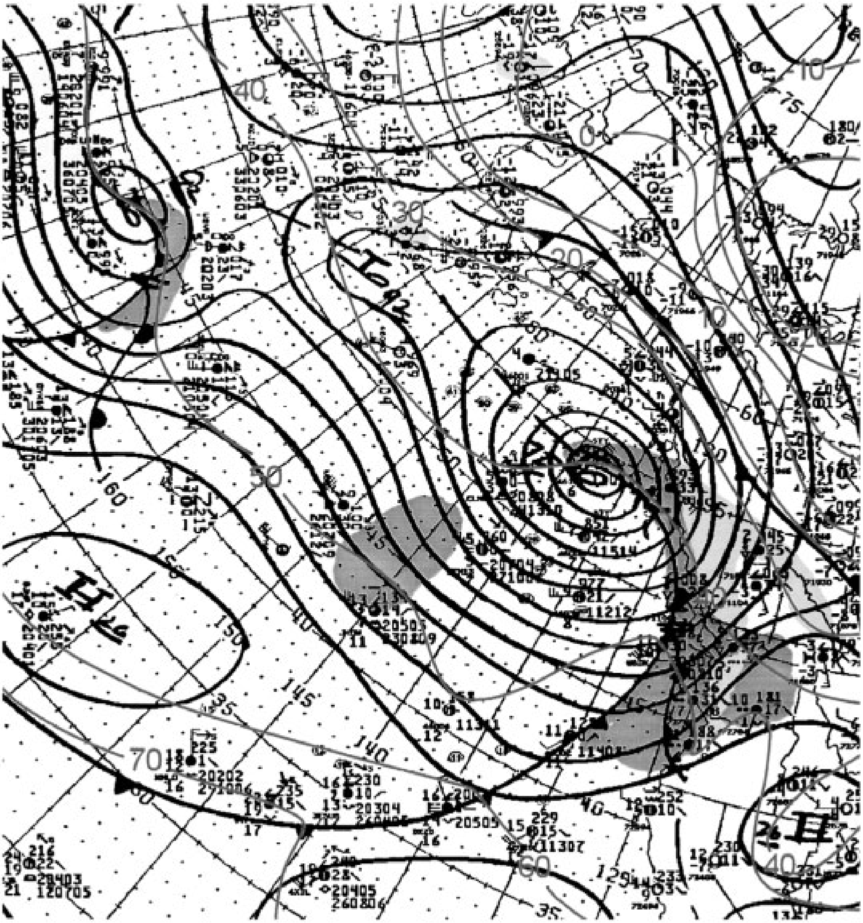
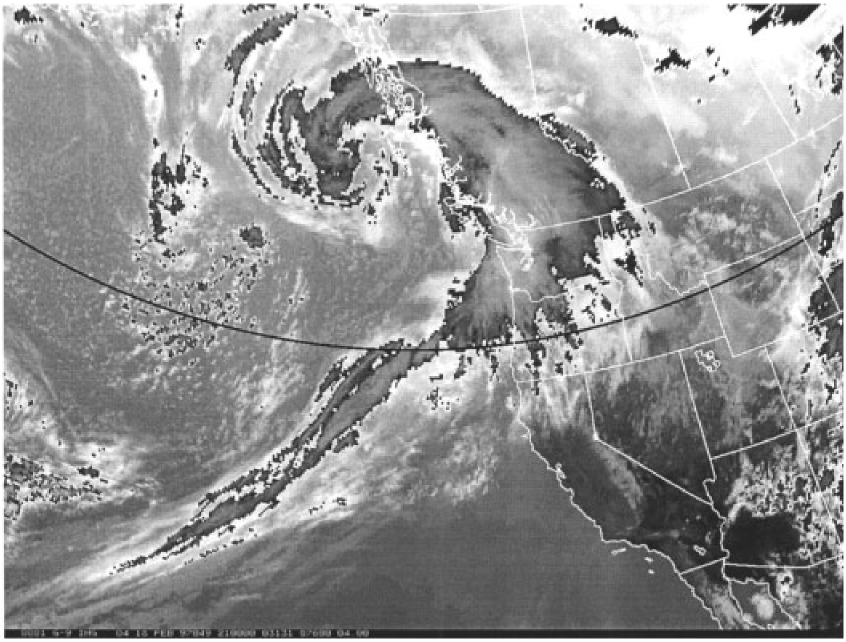


Figure 7 Mature extratropical cyclone in northwest United States, 0000 UTC Feb. 19, 1997. (*a*) Surface weather map with fronts (bold), sea level pressure (solid), temperature (°F, red), and areas of snow (light = light blue; moderate to heavy = dark blue), rain (light = light green; moderate to heavy = dark green), and sleet or freezing rain (light red). Surface data is plotted using conventional station model. Station temperatures and dewpoints are reported in Celsius in Fig. 7 and in Fahrenheit in Figs. 8 to 10. The temperature and precipitation analyses have been added by the author to the operational surface analyses generated by the National Centers for Environmental Prediction. See ftp site for color image.

Southern Hemisphere). Regional exceptions include the Pacific Northwest, where the strongest winds tend to be southerly ahead of the cyclone as it reaches the coast while high pressure remains in place inland (Fig. 7*a*), and the Northeast, where strong winds can often be found between the low-pressure center and the anticyclone in place ahead of it (Fig. 10*a*). Extratropical cyclones do not tend to have a tight



(b)

Figure 7 (*continued*) (*b*) Infrared satellite image, 2100 UTC Feb. 18, 1997, remapped to a polar stereographic projection similar to the surface weather map, with coldest temperatures (i.e., highest clouds) enhanced.

inner core like hurricanes, particularly over land, and the radius of maximum winds tends to be 100 to 500 km from the center.

Temperature The overall distribution of temperatures tends to be consistent with the locations of fronts. The warm sector, where the highest surface temperatures are found, is between the warm front and the cold front, generally the southeastern quadrant of the storm. Warmest temperatures are not in the center of the cyclone but tend to increase toward the southeast and south, with the warmest temperature at any given location typically occurring just prior to cold front passage. These patterns are modified by the history of air parcels within the extratropical cyclone. For example, cyclones in the northeastern Pacific tend to have small temperature variations at the surface (Fig. 7*a*), since the low-level air has been exchanging heat and moisture with the relatively uniform oceanic waters. Alberta clippers (Fig. 9*a*) and other cyclones close to the eastern edge of the Rocky Mountains often have their warmest temperatures to the southwest of the low in air that has descended over the mountains and warmed adiabatically. Along the East Coast (Fig. 10), relatively cold air ahead of the low often becomes trapped between the Appalachian Mountains to the west and the warm Atlantic Ocean to the east. The boundary between the cold air and the marine air is known as a coastal front and is analyzed as a trough in Figure 10.

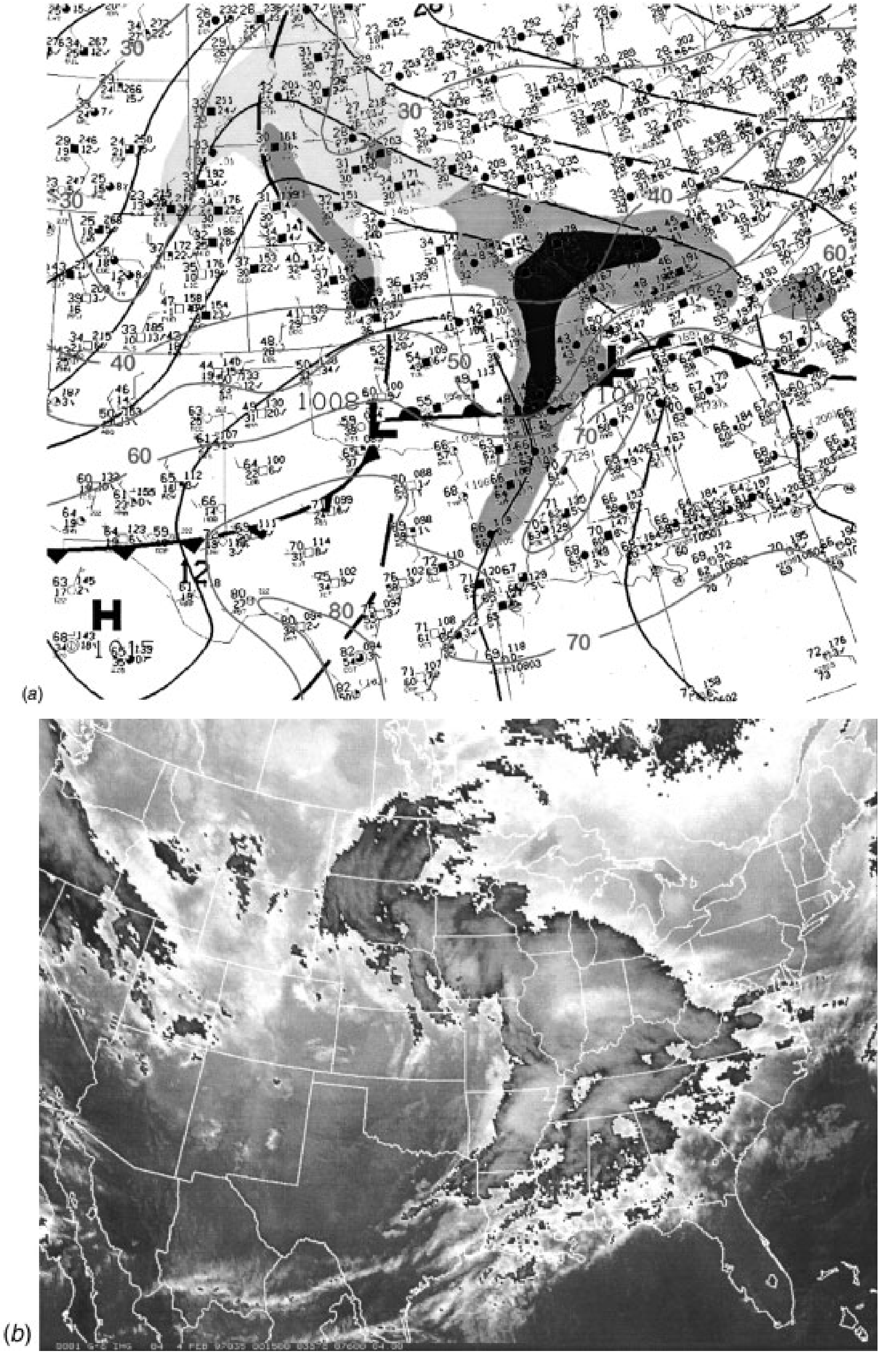


Figure 8 Typical cyclone in central United States, 0000 UTC Feb. 4, 1997. (*a*) Surface weather map. (*b*) Infrared satellite image, 0015 UTC Feb. 4, 1997. See Fig. 7 for details. See ftp site for color image.

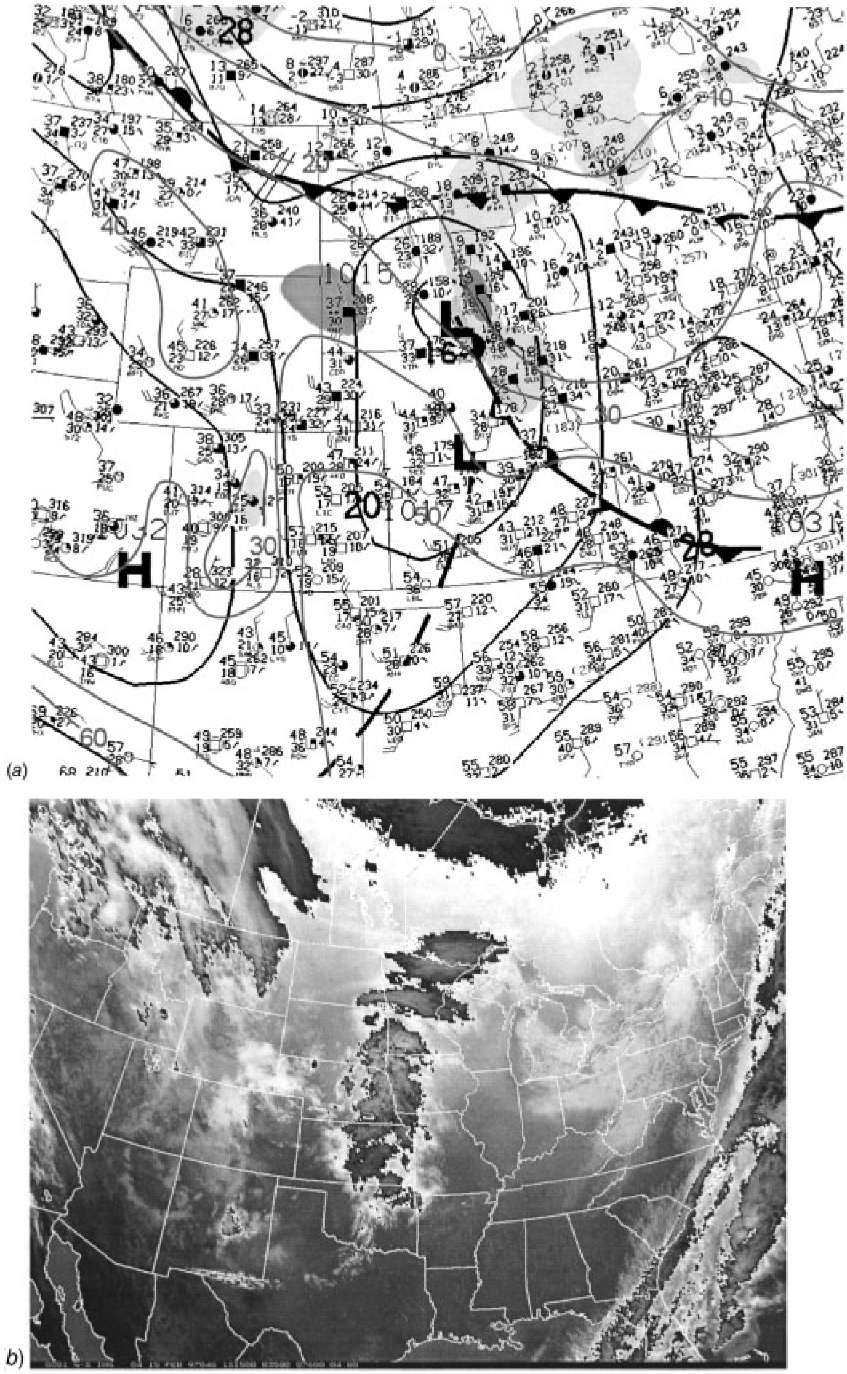


Figure 9 Alberta clipper cyclone, 1800 UTC Feb. 15, 1997. (*a*) Surface weather map. (*b*) Infrared satellite image, 1815 UTC Feb. 15, 1997. See ftp site for color image.

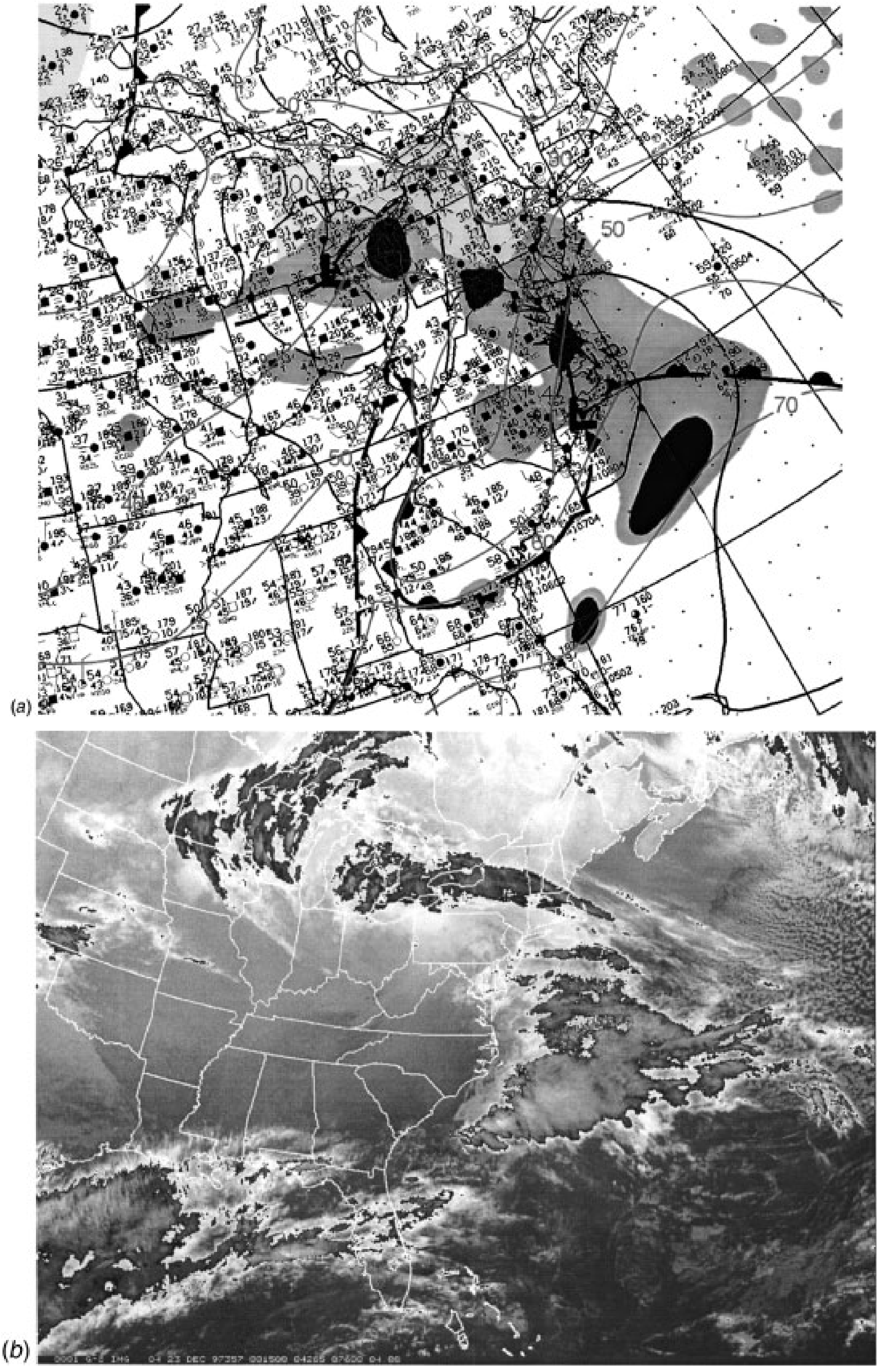


Figure 10 Redeveloping East Coast cyclone, 0000 UTC Dec. 23, 1997. (*a*) Surface weather map. (*b*) Infrared satellite image, 0015 UTC Dec. 23, 1997. See ftp site for color image.

Precipitation The distribution of precipitation is the most variable aspect of extratropical cyclone structure. The typical cyclone (Fig. 2) has extensive precipitation north of the warm front and in the vicinity of and behind the low, with the heaviest precipitation just ahead of the low center. Showers and rain bands tend to be present along the cold front, with scattered or widespread precipitation possible in the warm sector, especially near the low. The rain–snow line tends to be oriented roughly parallel to the track of the low, with a bulge northward near the low center, but its exact position depends on the temperature of the environment within which the extratropical cyclone is forming.

Precipitation requires both a moist air mass and a mechanism for lifting the air mass. The classical distribution of precipitation described above assumes a ready supply of moisture and a pattern of vertical motion determined by the dynamics of the cyclogenesis itself. The patterns can be strongly modified by both the air mass characteristics and the distribution of orography. Over water, where orography is nonexistent and moisture is readily available, precipitation tends to conform to the classical picture. However, the three cyclones over land in Figures 8 to 10 have markedly nonclassical precipitation patterns. The cyclone in Figure 8 possesses a rain band well ahead of the primary surface cold front, because air closer to the cold front has a history of descent over the Rocky Mountains and is relatively dry. Farther north of the cold front, another region of precipitation is being produced by upslope flow as air approaches the higher elevations of the Rocky Mountains and interacts with an inverted trough. Both of these areas of precipitation are common to cyclones in the central United States. The Alberta clipper in Figure 9 possesses no precipitation whatsoever within the warm sector or along the cold front because the entire air mass in the warm sector is dry after descending the mountains or originated as a cold air mass that has not passed over a large body of water and therefore remains dry. The storm along the East Coast in Figure 10 has an adequate supply of moisture from the adjacent Atlantic Ocean but is being affected by the Appalachian Mountains. Precipitation is relatively widespread on the eastern slope of the mountains where the air is warm and moist and is being forced to ascend over the mountains and over the cooler air trapped against them. Within the mountains, the trapped air often creates the necessary low-level temperature inversion for sleet or freezing rain. On the lee side of the mountains, precipitation is lighter.

The observations plotted in Figure 7 are too sparse to show the effects of the orography of the western United States on the distribution of precipitation. However, on scales of tens to thousands of kilometers, the orography modulates the precipitation by mechanically forcing ascent and descent and by altering the structure and tracks of weather systems. Precipitation tends to be high on the windward side of mountains and low on the leeward side. As air passes over successive mountain ranges, progressively less precipitation is produced.

The cloud patterns associated with extratropical cyclones are broadly similar to the precipitation patterns, but the determination of cloud top heights possible with infrared satellite imagery makes satellites a useful tool for diagnosing cyclone structure, particularly over oceans where other forms of data are scarce. Most common is a southwest to northeast oriented (in the Northern Hemisphere) band

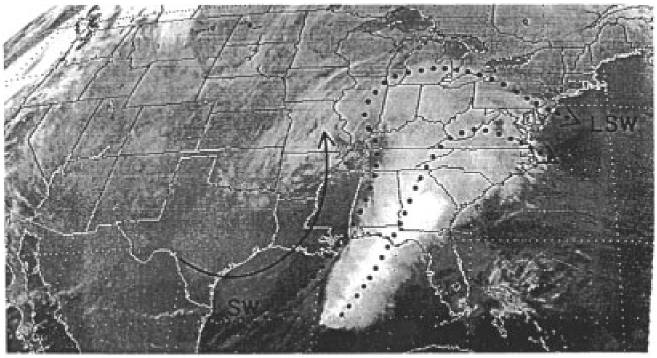


Figure 11 Infrared satellite images showing the location of the warm conveyor belt (dotted streamlines) and the dry airstream (solid streamline). "LSW" marks the leftmost streamline in the warm conveyor belt, which is known as the "limiting streamline" and typically is manifested as a sharp cloud boundary in satellite imagery. (From Carlson, 1991).

of high cloud with a sharp western edge. This cloud band represents the so-called warm conveyor belt, within which air from relatively warm latitudes is carried northward and upward through the storm roughly parallel to the cold front before curving anticyclonically beyond the warm front (Fig. 11). Precipitation beneath this conveyor belt may be convective, stratiform, or nonexistent. If the cyclone is sufficiently strong, the conveyor belt cloud band may curve cyclonically around the approximate position of the low center (as in Fig. 7*b*), or a separate cloud mass known as the cold conveyor belt may be present (as in Fig. 8*b* over the Dakotas and Fig. 10*b* over the Great Lakes and upper Midwest). This conveyor belt consists of air originating north of the warm front and ascending ahead of and on the poleward side of the cyclone before spreading northward or wrapping around the cyclone center. A third upper-tropospheric airstream, consisting of air descending behind the cyclone and curving cyclonically to the northeast, is known as the dry airstream and is indicated in Figure 11.

These cloud structures may be used to diagnose the location and intensity of the extratropical cyclone center, but they are most directly related to the horizontal and vertical wind fields in the upper troposphere and the distribution of precipitation. Because surface pressure features tend to be strongly modified by orography, the structure of the surface cyclone can be inferred reliably from cloud structures only over water. Notice the superficial similarity of Figures 7*b*, 8*b*, and 10*b* despite the wildly different low-level wind and pressure fields.

Statistical Climatology

The climatological distribution of extratropical cyclones reflects the influence of orography (Fig. 12). Cyclones are most common over the oceans between latitudes of 30°N and 60°N. In contrast, extratropical cyclones are rare over high orography.

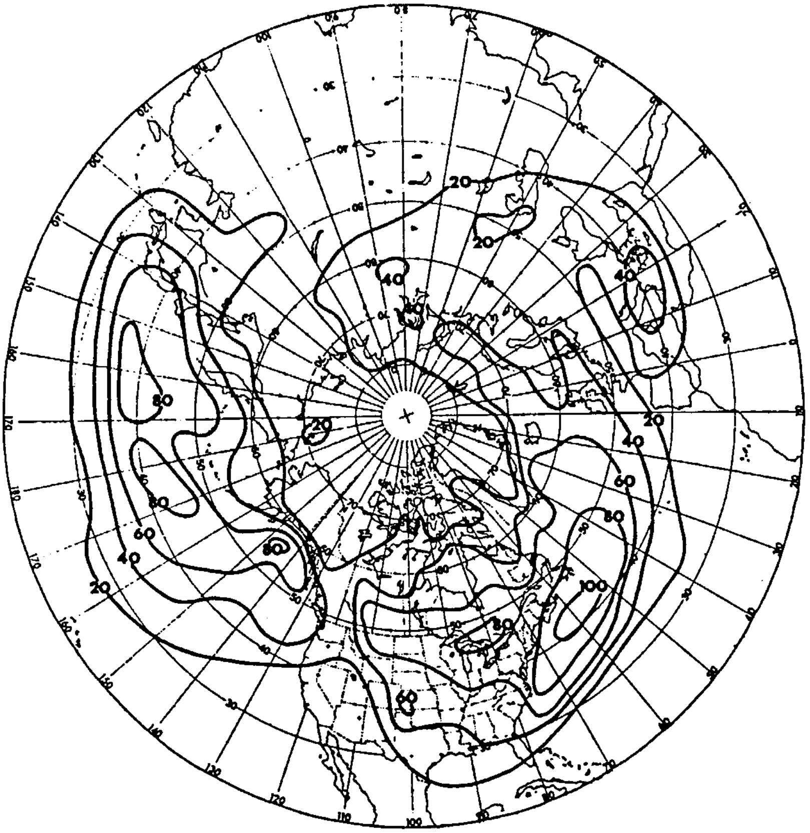


Figure 12 Distribution of total number of cyclones passing through 5 × 5 latitude–longitude boxes during 20 Januarys, 1958–1977. (From Whittaker and Horn, 1984).

Given the general tendency of cyclones to move toward the east or northeast, one can infer from Figure 12 the general distribution of cyclogenesis (formation and intensification of extratropical cyclones) and cyclolysis (weakening and dissipation of extratropical cyclones). Cyclones tend to form downstream of major mountain barriers such as the Rocky Mountains and the Alps, as well as over the midlatitude oceans. Semipermanent large-scale cyclones may be found in the extreme North Atlantic (the Icelandic low) and North Pacific (the Aleutian low). These lows are periodically reinvigorated as strong extratropical cyclones migrate northwestward across the jet and merge with them. At higher latitudes, small-scale intense extratropical cyclones known as *polar lows* sometimes form near the sea ice boundary. These polar lows are partly driven by deep convection as air masses are destabilized by the underlying warmer ocean surface, and share some of the characteristics of hurricanes.

The midlatitude extratropical cyclones that become particularly intense typically do so by deepening rapidly. Such explosively deepening cyclones, known as *bombs*,

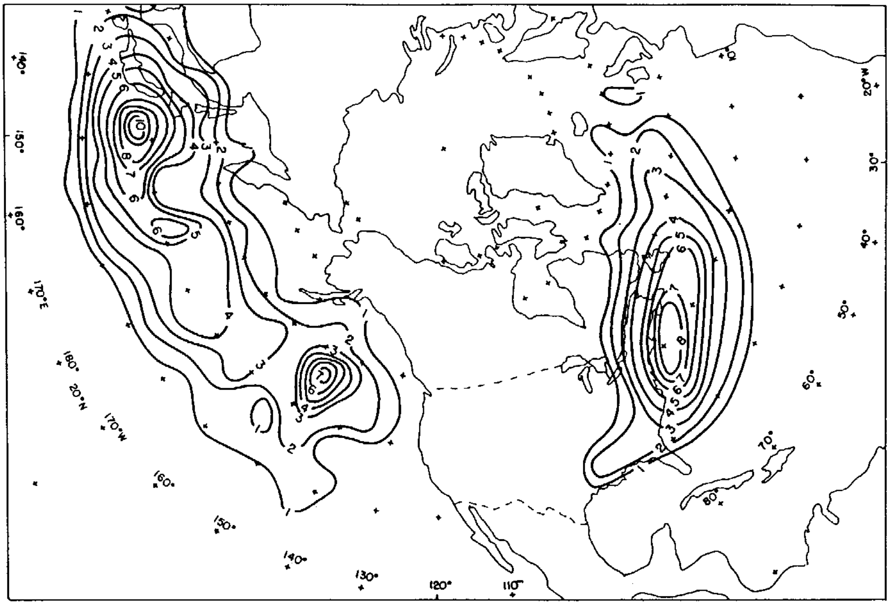


Figure 13 Distribution of the locations at which rapidly deepening extratropical cyclones undergo their most rapid intensification, expressed as total number within each 5 × 5 latitude–longitude box (normalized by latitude) for the period 1976–1982. (From Roebber, 1984).

develop almost exclusively over the oceans, preferentially near the eastern edge of continents (Fig. 13). Conditions favoring explosive deepening include a strong upper-level disturbance, a strong low-level horizontal temperature gradient, and a source of warm, moist air. The Kuroshio current (in the Pacific) and the Gulf Stream (in the Atlantic) help provide the latter two conditions by influencing the distribution of heat and moisture in the overlying atmosphere.

Vertical Structure and Dynamics of Extratropical Cyclones

For development, extratropical cyclones require vertical shear. This means that cyclones will typically develop beneath an upper-tropospheric jet stream. Since balanced vertical shear implies a horizontal temperature gradient, cyclones therefore develop within a large-scale temperature gradient or along a frontal zone. Discussing the vertical structure of extratropical cyclones requires the introduction of the terms upshear, meaning in the direction opposite the vertical shear vector (typically, the direction opposite the upper-tropospheric jet stream), and downshear, meaning in the same direction as the vertical shear vector (typically, downwind relative to the upper troposphere). These concepts are illustrated in Figure 14.

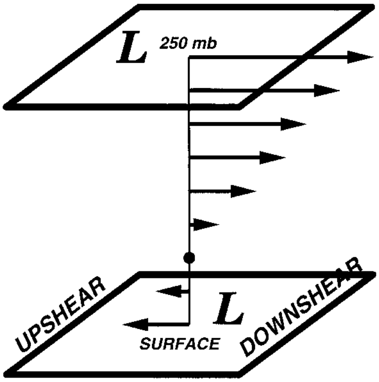


Figure 14 Schematic showing the definition of upshear and downshear for a simple, typical vertical distribution of horizontal wind. The low-pressure system in the figure tilts upshear with height.

An intensifying extratropical cyclone, as defined by the height and pressure distribution, typically tilts upshear with height. This means that the trough in the upper troposphere will be located upshear of the surface cyclone. The temperature distribution is tilted in the opposite sense, so that the warmest temperatures at the surface are just downshear of the surface cyclone position and the warmest temperatures in the upper troposphere are well downshear of the surface cyclone position. The largest variations in pressure and temperature are located in the upper troposphere and at the surface but are substantial at all levels in between. This pressure and temperature distribution is entirely consistent with thermal wind balance, so this particular pressure distribution implies this particular temperature distribution and vice versa. Winds, being in approximate geostrophic balance, vary in tandem with pressure.

The vertical motion is an integral part of the extratropical cyclone structure because without vertical motion cyclones would not intensify. Cyclone intensification can be thought of as an increase in the circulation about the cyclone center and, for large-scale frictionless flow, the fractional rate of increase of total circulation (including the circulation associated with Earth's rotation, which is essentially constant) is proportional to the horizontal convergence of the wind. Mass conservation implies upward motion in the middle troposphere wherever there is low-level convergence (excluding exotic effects associated with a sloping lower boundary), so vertical motion is not just a consequence of cyclogenesis; it is a necessary component of cyclogenesis. Since the stratosphere inhibits vertical motion, upper-tropospheric convergence (and intensification of the upper-tropospheric part of the cyclone) implies midtropospheric downward motion. In a typical intensifying cyclone, there is midtropospheric upward motion (with the associated clouds and precipitation) ahead of and over the surface cyclone, and downward motion beneath and behind the upper-tropospheric trough (Fig. 15).

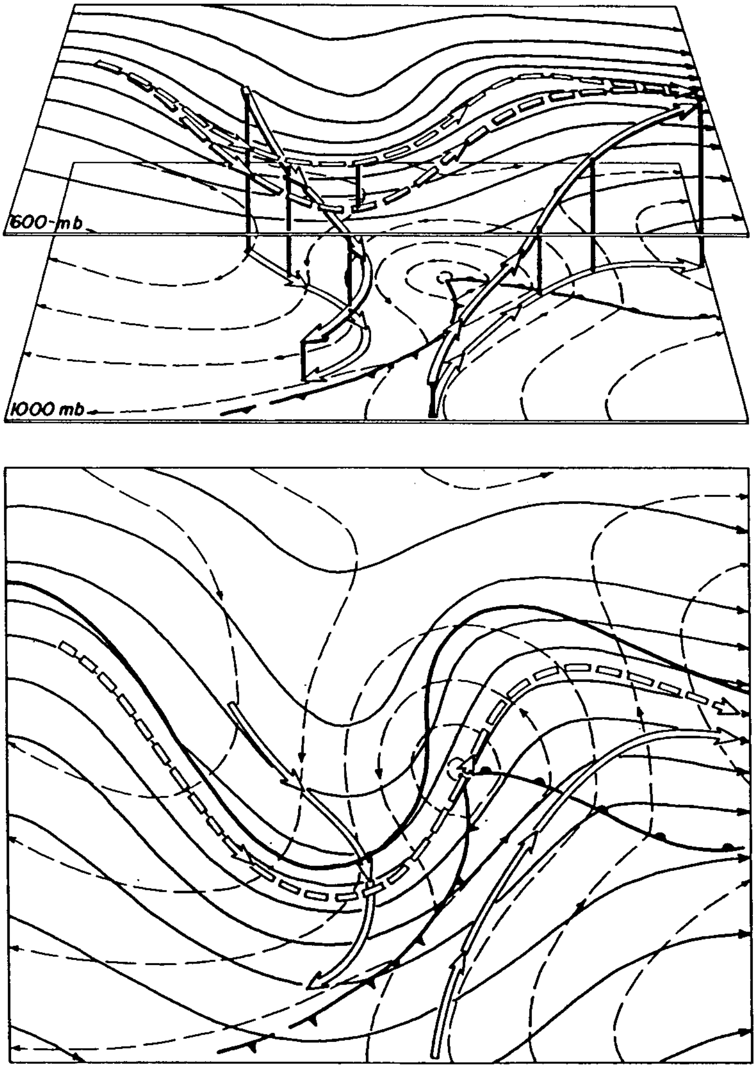


Figure 15 Distribution of upward and downward motion in an idealized extratropical cyclone. Depicted are a perspective view (top) and two-dimensional view (bottom) of height contours at 1000 mb (dashed) and 600 mb (thin solid), surface fronts, and ascending and descending airstreams (wide arrows) with their projections onto the 1000-mb or 600-mb surfaces. (From Palmen and Newton, 1969).

To diagnose and predict extratropical cyclogenesis, meteorologists use a relationship between the structure of the large-scale balanced winds and temperatures and the vertical motion known as the omega equation. The omega equation is a three-dimensional analog to the Sawyer–Eliassen equation and is based on the same dynamical principles of continuous adjustment to preserve nearly balanced flow. The *forcing*, the wind and temperature patterns that imply upward motion, is rather complicated when expressed mathematically, and several versions of the equation, some involving simplifying assumptions, are in use. The most common qualitative application roughly equates upward motion with the sum of two terms, one proportional to the vertical derivative of vorticity advection and the other proportional to the temperature advection. The first term tends to dominate in the upper troposphere and implies upward motion ahead of an upper-level trough and downward motion behind it. The second term tends to dominate in the lower troposphere and implies upward motion ahead of the surface cyclone (where warm advection is typically found) and downward motion behind it. Between the upper-level trough and the surface cyclone, it will be noted that the two terms are of opposite sign, leading to difficulty in inferring the vertical motion field there using this method. With the increasing use of computers, many of these simplifying assumptions are being discarded in favor of explicit mathematical calculation of the vertical motion forcing, but the qualitative interpretation is still essential for relating the vertical motion field to the large-scale features in the lower and upper troposphere.

In contrast to the simultaneous treatment of winds, temperatures, and vertical motion, cyclogenesis can also be understood from a dynamical point of view in terms of potential vorticity. Indeed, the basic theoretical paradigm for cyclogenesis, known as baroclinic instability, requires a potential vorticity gradient aloft opposite in orientation to the surface potential temperature gradient as a necessary condition for instability. Cyclogenesis itself, by any mechanism, requires an increase in the integrated perturbation potential vorticity and/or surface perturbation potential temperature in the vicinity of the surface cyclone. Baroclinic instability accomplishes this by having two waves exist simultaneously, one on the surface potential temperature gradient and the other on the upper-tropospheric potential vorticity gradient; if the two waves are of large enough horizontal scale and the upper one is upshear of the lower one, the circulation associated with each wave causes the other wave to intensify. Other mechanisms for growth that do not involve instability include rearrangement (specifically, compaction) of potential vorticity by horizontal shear, and a decrease in the vertical tilt between the upper and lower perturbations. Using the electrostatics analogy, the instability mechanism involves increasing the electric charge, while the other two mechanisms involve rearranging the electric charge into a compact area. The potential vorticity approach and the omega equation approach are mutually consistent and complementary.

7 UPPER-TROPOSPHERIC JETS AND TROUGHS

A meteorologist looks at the surface weather map to see the current weather but looks at an upper-tropospheric map to understand the current weather. The upper-tropospheric map, generally the 500-mb constant-pressure surface or above, shows the meteorologist the upper-level features associated with the current vertical motion field, the current motion of the upper-level features and related surface features, and the potential for intensification of surface cyclones and anticyclones. By examining the large-scale height field at or near jet stream level, the meteorologist can infer the likely distribution of surface temperatures and the likely path of any storms that might develop.

Jet Streams and Jet Streaks

The jet stream is often described as a band of strong winds in the upper troposphere (8 to 12 km above sea level) encircling the globe. Typically, however, the location of the jet stream is not so simple. There may be two or more jet streams at a given longitude, or the Northern Hemisphere jet stream may originate in the subtropical Atlantic, cross southern Asia, pass over the Pacific Ocean at about 35°N and the Atlantic at 45°N, and eventually weaken over northern Asia. Three examples of the wintertime Northern Hemisphere jet stream and one of the wintertime Southern Hemisphere jet stream are shown in Figure 16.

The first example (Fig. 16*a*) possesses many features common to the Northern Hemisphere wintertime circulation. The strongest jet is over the Pacific Ocean, although it is unusually far south in this example. Consistent with this departure from normal is an unusual northward displacement of the jet over North America; long-term departures from the average circulation tend to have wavelengths of about 8000 km. Over the Atlantic Ocean and over Asia there are at least two jets. The southern jet is called the subtropical jet and the northern jet is called the polar jet.

The second example (Fig. 16*b*) is unusual because the jet over the Atlantic is stronger than the jet over the Pacific. This unusually strong Atlantic jet corresponds to a location where the polar and subtropical jet streams appear to merge. Over the extreme northeastern Atlantic and northern Europe is a northward displacement of the jet stream known as a *block*, because weather disturbances are carried northward around the block rather than from west to east as would be normal. This particular block is known as an omega block because of the resemblance of the jet stream pattern to the Greek letter Ω.

The third example (Fig. 16*c*) occurred during an El Niño year and is also quite unusual. It features a subtropical jet that is essentially continuous around the globe, with a weaker polar jet that is nearly continuous except for its merger with the subtropical jet over the Pacific. Some resemblance may be found to the wintertime jet pattern over the Southern Hemisphere (Fig. 16*d*), which also features two nearly continuous jets. As a general rule, because the topographic and sea surface temperature variations in the Southern Hemisphere are comparatively small, the planetary-scale waves in the jet stream tend to be weaker there. Also (not shown), the jet streams actually tend to be stronger during the summer in the Southern Hemisphere than during the winter because the pole-to-subtropics temperature contrast increases during the summer season. In the Northern Hemisphere, temperatures are much more uniform during the summer and the upper-tropospheric jet speeds are correspondingly weaker.

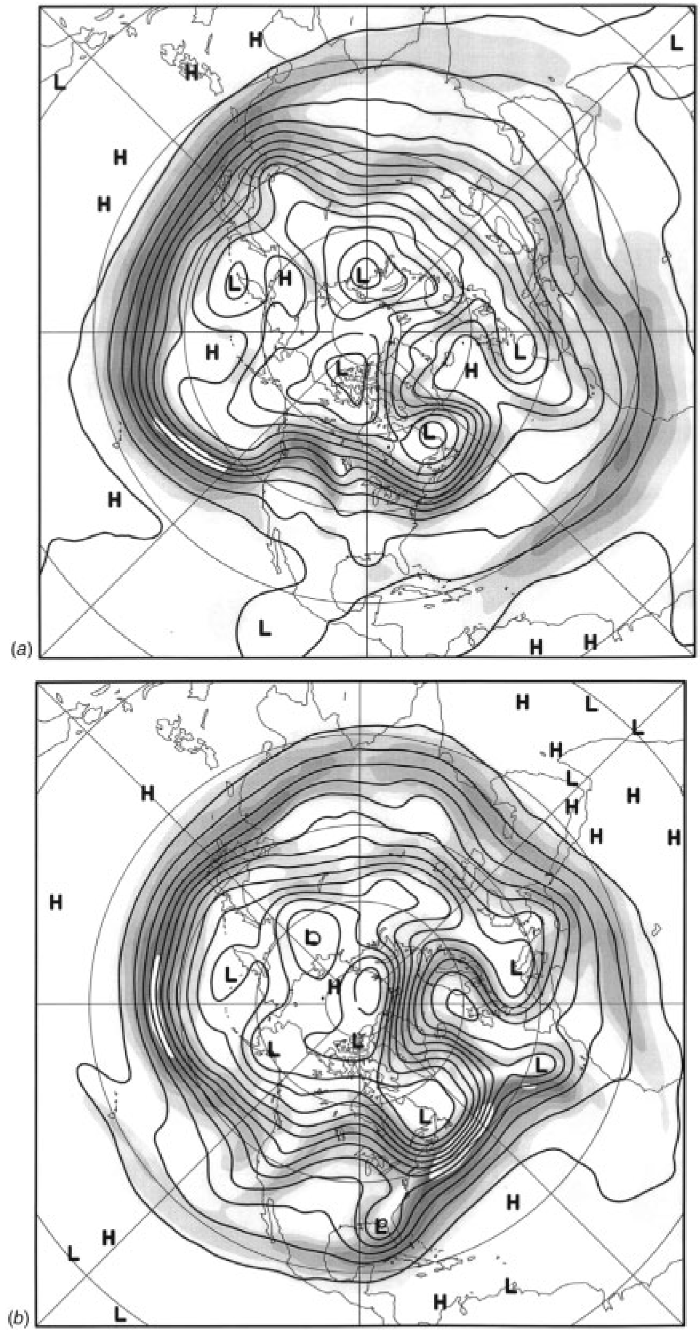


Figure 16 Wind speed (shaded, 10 m/s interval, beginning at 20 m/s) and height (solid, 120-m interval) at 250 mb for four wintertime hemispheres. (*a*) 0000 UTC Jan. 1, 1997. (*b*) 1200 UTC Dec. 15, 1997. (*c*) 0000 UTC Feb. 1, 1998. (*d*) 0000 UTC July 1, 1997.

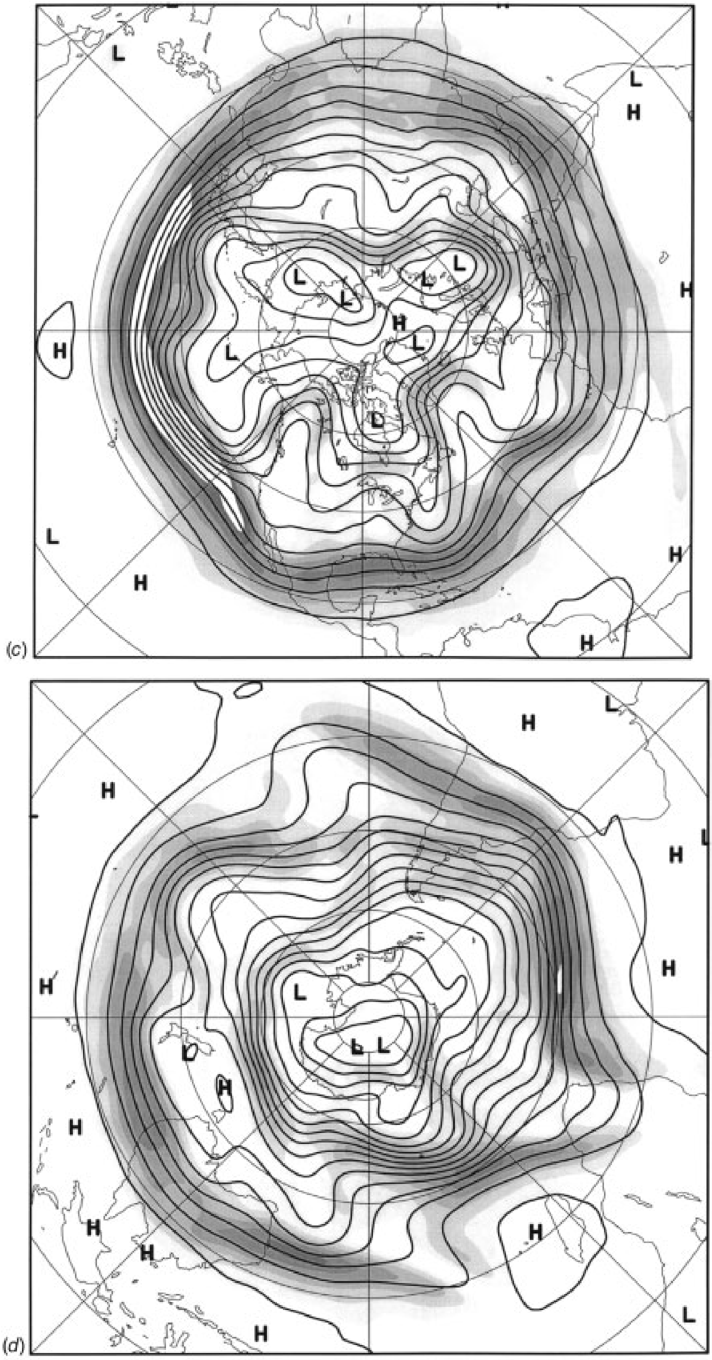


Figure 16 (*continued*)

In a cross section through a jet stream (Fig. 17), the close relationship between the horizontal derivative of temperature and the vertical derivative of wind speed is apparent. Beneath the jet is a strong horizontal temperature gradient, and the wind speed increases with height; above the jet the opposite is true. At jet steam level, the vertical derivative of wind speed is (by definition) zero and, to the extent that the airflow is balanced, the horizontal derivative of temperature is zero too. The height of the tropopause (marked by the transition between a rapid decrease of temperature with height and nearly uniform temperatures with height) also varies across the jet stream, being low on the poleward side and high on the equatorward side. Since potential vorticity is much higher in the stratosphere than the troposphere, a strong potential vorticity gradient would also be found in the vicinity of the jet.

A jet streak is essentially a local wind maximum within a jet stream. Being of smaller scale than a jet stream, air parcels frequently undergo rapid accelerations as they enter and exit a jet streak, implying strong ageostrophic winds and the likelihood of patterns of divergence and convergence. For example, as an air parcel enters a jet streak, it experiences a stronger pressure (or height) gradient than before, and accelerates to its left in the direction of lower heights. This downgradient-directed ageostrophic wind then implies a Coriolis force directed in the original direction of motion, causing the air parcel to speed up and ultimately approach balance with the height gradient. In terms of the ageostrophic wind, one expects a

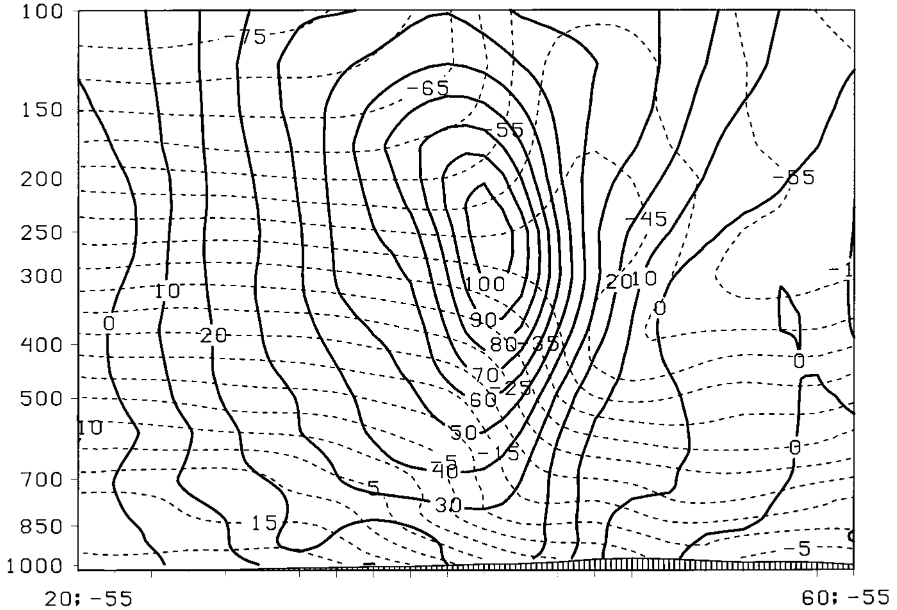


Figure 17 Vertical section taken through the strong Atlantic jet shown in Fig. 16*c* along 55°W. South is to the left. Solid contours show the component of wind into the page (m/s) and dashed contours show temperature (°C).

downgradient ageostrophic wind in the entrance region of a jet streak, and a corresponding upgradient ageostrophic wind in the exit region.

For an idealized straight jet streak (Fig. 18), this ageostrophic wind configuration produces characteristic patterns of convergence and divergence. Divergence is found in what is known as the right entrance region and the left exit region, with convergence in the left entrance region and the right exit region. The pattern is reversed in the Southern Hemisphere. And, since jet streaks are located near tropopause level and the overlying stratosphere inhibits vertical motion, the divergence and convergence imply rising and sinking motion beneath the jet streak in the middle troposphere. Consequently, clouds and precipitation are favored to the left or poleward side of an approaching jet streak and to the right or equatorward side of a receding jet streak.

An along-stream intensification of the height gradient, such as is found at the entrance region of a jet streak, implies confluence of the balanced geostrophic wind.

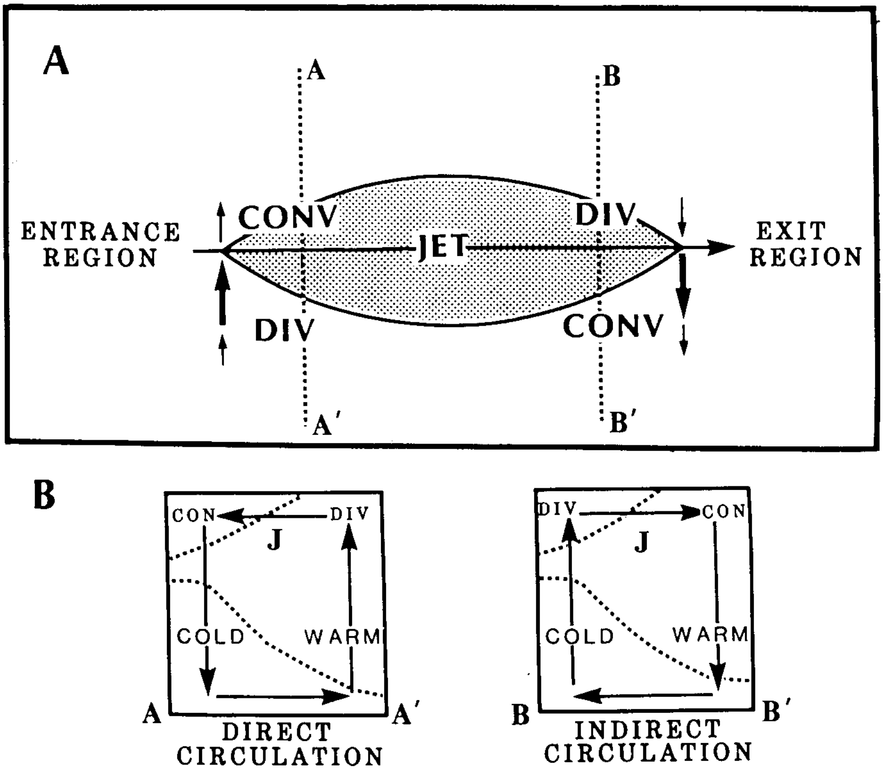


Figure 18 Idealized four-quadrant jet streak. (*a*) Patterns of divergence and convergence (DIV and CONV) and associated transverse ageostrophic winds (arrows) associated with a straight jet streak. (*b*) Vertical sections along A–A′ and B–B′ from (*a*) showing the vertical circulations in the entrance and exit regions. Also plotted in (*b*) are the jet position (J) and two representative isentropes (dotted lines). (From Kocin and Uccellini, 1990).

Beneath the jet streak is a horizontal temperature gradient, very often in the form of an upper-level front. The confluence implies frontogenesis, and the ageostrophic flow at jet streak level constitutes the upper branch of the direct circulation associated with frontogenesis.

Short and Long Waves

Just as straight, concentrated gradients of potential vorticity in the upper troposphere are associated with jets, waves in the potential vorticity gradient are associated with upper-tropospheric troughs and ridges. Indeed, the concentrated gradients serve as a medium along which the waves, known as Rossby waves on the synoptic scale and Rossby–Haurwitz waves on the planetary scale, propagate.

The propagation mechanism is fairly straightforward and can be applied also to the upper and lower waves in a baroclinic extratropical cyclone. Imagine that a straight jet has been perturbed into a periodic wave pattern (Fig. 19). Each equatorward excursion of the potential vorticity contours will be associated with a cyclonic circulation, by the principles outlined in the first part of this chapter, and each poleward excursion will be associated with an anticyclonic circulation. In a frame of reference moving with the large-scale westerlies, these circulations imply an alternating pattern of northerlies and southerlies, which act to redistribute the potential vorticity so as to cause the potential vorticity waves to propagate westward. The speed of propagation is proportional to the strength of the potential vorticity gradient and to the wavelength. As a general rule, the propagation speed of short waves (wavelengths less than 5000 km or so) tends to be much less than the speed of the westerlies in which they are embedded, so short waves (also known as mobile

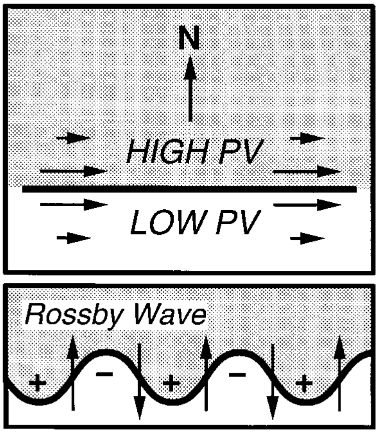


Figure 19 Schematic Rossby wave train. *Top panel:* undisturbed upper-level jet showing winds and PV variations. *Bottom panel:* Rossby waves along upper-level jet with associated north–south component of winds. The pluses and minuses represent cyclonic and anticyclonic potential vorticity anomalies, respectively.

troughs for this reason) tend to move in the same direction and slower speed as the upper-level winds in which they are embedded. Midlatitude long waves may move slowly, be stationary, or even move westward against the large-scale flow. Stationary ridges that persist for one or more weeks are known as blocks or blocking highs because extratropical cyclones follow the jet and migrate poleward around the ridge, and are blocked from entering the affected area.

Rossby waves are dispersive, and the group velocity is oriented opposite the phase velocity, so wave energy tends to propagate downstream at a speed faster than the jet stream winds. This phenomenon is illustrated schematically in Figure 20, which shows an isolated upper-tropospheric trough. The cyclonic winds associated with the trough act to move the trough westward, but at the same time they generate a ridge to the east of the trough. As time goes on, that ridge would then generate a trough farther downstream, and so on. Meanwhile, the original trough would weaken, unless it is involved in cyclogenesis or has some other energy source to allow it to maintain its strength. The process by which one wave triggers another wave downstream is known as downstream development. An example of downstream development in the atmosphere is shown in Figure 21.

Long waves may be formed by interaction between the jet and the underlying large-scale topography, by interaction between the jet and the embedded short waves, and by large-scale latent heating produced by convection. The latter mechanism is the primary means by which equatorial oceanic or atmospheric conditions affect the midlatitude and polar circulations. Equatorial tropospheric heating by convection and the associated upper-tropospheric divergence alters the potential vorticity pattern and initiates a wave train whose energy propagates poleward and eastward. Ideally, the waves follow a great circle path and ultimately return to the equator, but the exact path of the waves is influenced by the local wind and potential vorticity characteristics of the medium within which they are embedded. The wave train persists as long as the equatorial forcing persists.

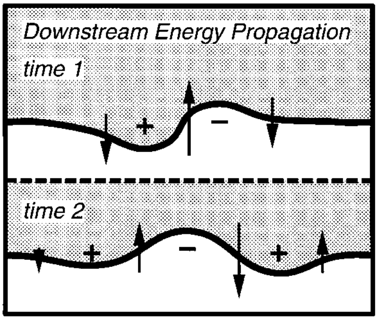


Figure 20 Schematic example of energy propagation associated with an isolated upper-level trough on a jet such as the one depicted in Fig. 19. As the pattern evolves, successive troughs and ridges appear downstream (to the east), while the original wave weakens. While individual crests and troughs propagate upstream, the wave packet as a whole propagates downstream.

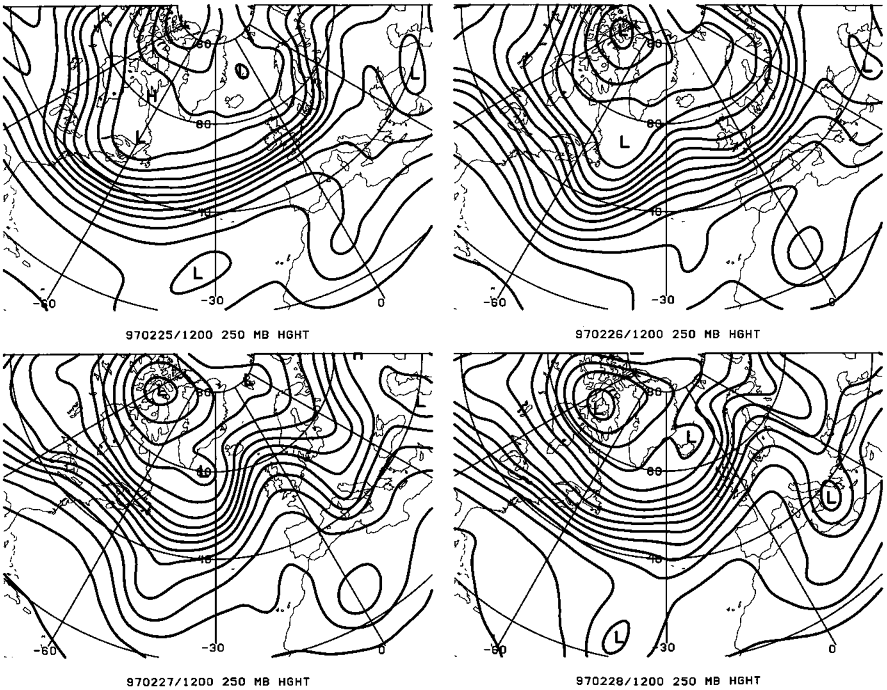


Figure 21 Example of downstream development in the atmosphere, Feb. 25–28, 1997. Contours represent 250-mb height at 120-m intervals. *First panel:* relatively straight jet across North Atlantic. *Second panel:* trough develops over west Atlantic associated with surface cyclogenesis (not shown); ridge begins forming in east Atlantic. *Third panel:* trough begins weakening over central Atlantic; ridge moves eastward and reaches maximum intensity over Great Britain; new trough forming over central Europe. *Fourth panel:* trough nearly gone over central Atlantic, ridge weakening over northern Europe, trough over central Europe attains maximum intensity.

8 WATER VAPOR IMAGERY

The primary satellite tool for detecting upper-tropospheric weather features is the water vapor image. This image is based on a near-infrared wavelength for which water vapor is a strong absorber and emitter, so that the radiation reaching the satellite typically comes from water vapor in the middle to upper troposphere. Where the upper troposphere is particularly dry, the radiation reaching the satellite originates from the lower troposphere, from water vapor at a higher temperature. The variation in intensity of radiation caused by the differing temperatures of the water vapor emitting the radiation is used to infer the distribution of moisture (or lack thereof) in the upper and middle troposphere.

A water vapor image can be used to pinpoint the locations of troughs and jets. Often, the jet stream is marked by a strong upper-level moisture gradient, with low moisture on the cyclonic side of the jet and higher moisture on the anticyclonic side of the jet. This pattern is partly related to the confluence of different air masses as air enters the jet, and partly related to the lower tropopause on the cyclonic side of the jet or in troughs, with comparatively low water vapor content in the stratosphere compared to the troposphere. However, similar features can be produced by ordinary advective processes and vertical motion, so loops of images or maps of upper-level height and wind must be consulted to properly interpret features appearing in water vapor images.

Figure 22 shows the water vapor satellite image for the time of the strong Atlantic jet shown in Figure 16*b*. The water vapor image features a band of moisture crossing Cuba and becoming brighter as it passes alongside the East Coast and moves south of Newfoundland. The sharp left edge of this band is associated with the axis of maximum winds at jet level. By contrast, the polar jet crossing into the Atlantic from Canada does not appear as a distinct feature in this water vapor image; as a general rule, southwesterly jets are easier to identify in water vapor imagery than northwesterly jets. The two upper-level vortices over Florida and off the coast of Spain are also prominent in the water vapor image. The dark area east of Newfoundland, north of the subtropical jet, indicates the position of an upper-level trough and is caused by the extremely low tropopause there. The trough appears to be initiating cyclogenesis, judging from the developing water vapor pattern just to its east (compare Fig. 22 to Fig. 10*b*). The much larger dark areas over the eastern Caribbean are associated

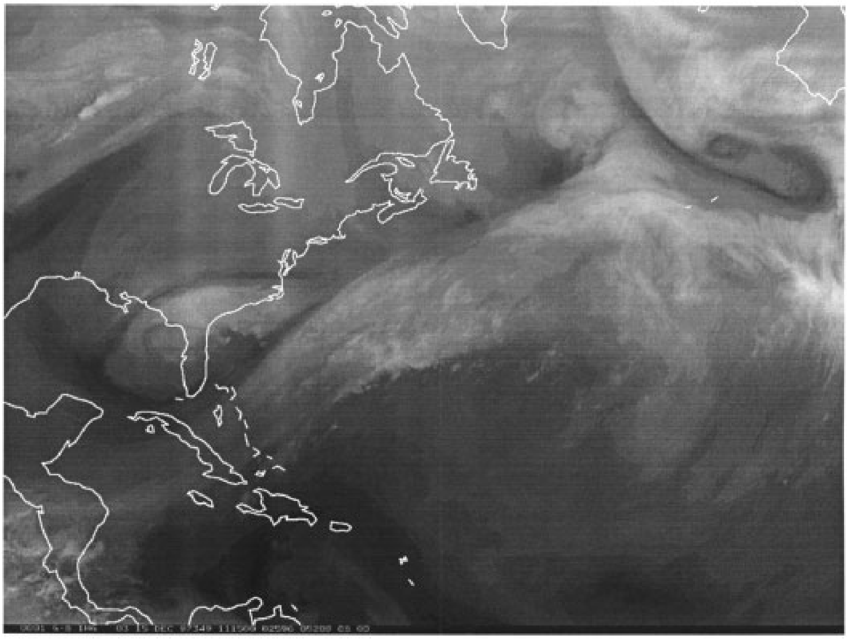


Figure 22 Water vapor image of strong Atlantic jet, 1115 UTC Dec. 15, 1997. Compare with Figs. 16*b* and 17.

with dry, subsiding air within the subtropics and are of no particular dynamical significance.

REFERENCES

Carlson, T. N. (1991). *Mid-Latitude Weather Systems*, London, HarperCollins.

Kocin, P. J., and L. W. Uccellini (1990). *Snowstorms Along the Northeastern Coast of the United States: 1955 to 1985*, *Meteor. Monogr.*, **22**, No. 44, Boston, American Meteorological Society, 280pp.

Nese, J. M., and L. M. Grenci, (1998). *A World of Weather: Fundamentals of Meteorology*, IA, Kendall/Hunt.

Roebber, P. J. (1984). Statistical analysis and updated climatology of explosive cyclones, *Mon. Wea. Rev.*, **112**, 1577–1489.

Shapiro, M. A. (1991). Frontogenesis and geostrophically forced secondary circulations in the vicinity of jet stream— frontal zone systems, *J. Atmos. Sci.*, **38**, 954–973.

Shapiro, M. A., and D. Keyser (1990). Fronts, jet streams, and the tropopause, *Extratropical Cyclones: The Erik Palmen Memorial Volume* (C. Newton and E. Holopainen, eds.), Boston, American Meterological Society, 167–191.

Thorpe, A. J. (1986). Synoptic scale disturbances with circular symmetry, *Mon. Wea. Rev.*, **114**, 1384–1389.

Whittaker, L. M., and L. H. Horn (1984). Northern hemisphere extratropical cyclone activity for four mid-season months, *J. Clim*, **4**, 297–310.

CHAPTER 27

WINTER WEATHER SYSTEMS

JOHN GYAKUM

1 INTRODUCTION

Winter weather in the extratropical latitudes is defined by the individual and collective contributions of cold air, wind, and precipitation. Each of these meteorological phenomena is determined by a combination of synoptic-scale weather systems and topography.

Economic impacts of winter weather can be substantial. Among the 48 weather disasters that cost at least $1 billion in the United States during the period 1980–2000, 11 were winter weather events (NOAA, 2000). These included 4 flooding/ heavy-rain events, three coastal cyclones, two freezes in Florida, and two ice storms.

This chapter focuses on the dynamics of winter weather systems and how topographic features affect these systems.

The weather maps used in this chapter show fields of sea level pressure (SLP), 500–1000 hPa thickness, and their anomalies from a climatological state. Each field is taken from the National Centers for Environmental Prediction (NCEP) global reanalysis of meteorological data (Kalnay et al., 1996). The thickness fields display the 500–1000 hPa layer mean virtual temperature. An incremental change of 60 m of 500–1000 hPa thickness represents an approximate temperature change of 3°C. The anomalies of this quantity are also displayed on these maps. We define an anomaly as the difference between the actual 500–1000 hPa thickness and the appropriate monthly climatological mean. This mean value is computed for the period 1963–1995. Therefore, a thickness anomaly field quantifies how much colder or warmer a particular region is, compared with its expected value for the season.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

2 COLD AIR

Winter is characterized by relatively cold temperatures in extratropical latitudes. These colder temperatures are a manifestation of equatorward displacements of the zonal jet stream and cold-surface anticyclones. One especially interesting component of winter weather is the production of cold-air outbreaks. These outbreaks are associated with anomalously cold temperatures, where an anomaly is defined as the difference between the actual temperature and the long-term climatological mean.

For a cold-air outbreak to have a large impact, cold air must travel rapidly equatorward from its source region in the higher latitudes of continents. The rapid displacement is essential for a strong cold anomaly to occur, because cold air will warm at lower latitudes owing to the warmer ground or ocean surface and to stronger solar insolation. A favorable upper-tropospheric pattern includes an anomalously strong wave in the jet stream with the ridge and trough being amplified relative to the mean climatological state.

A severe cold-air outbreak occurred during December 1983 with damage to the Florida citrus crop exceeding $2 billion. The structure of the highly amplified 1000–500 hPa thickness field (Fig. 1) included a strong ridge in Alaska with an anomaly

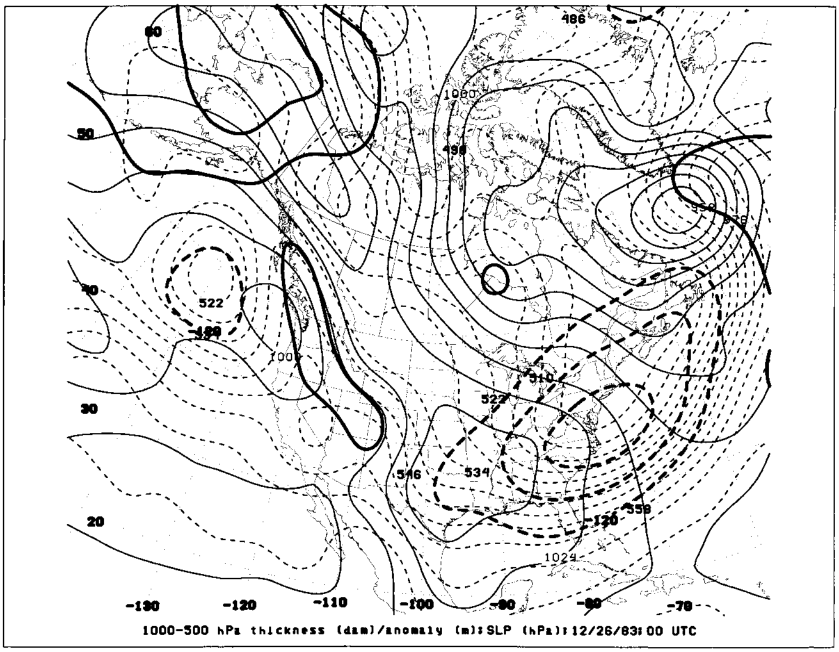


Figure 1 Sea-level pressure (light solid, interval of 8 hPa), 1000–500-hPa thickness (light dashed, interval of 6 dam), and 1000–500-hPa thickness anomaly (heavy solid/dashed dashed for positive/negative with interval of 120 m) for 0000 Universal Time Coordinated (UTC), December 26, 1983.

exceeding +24 dam and a cold anomaly (−44 dam) in the eastern middle Atlantic states. The structure shown by Figure 1 is an extreme case of a tropospheric field that persisted for much of December 1983 (e.g., Fig. 8*b* of Quiroz, 1984).

The qualitative structure of Figure 1 illustrates the necessary and sufficient conditions for a significant cold outbreak. The tropospheric ridging in the thickness field develops as a result of strong surface cyclogenesis in the North Pacific, as poleward flow to its east advects warm air into Alaska. This strong ridging in western North America provides anomalously strong equatorward steering flow for cold-surface anticyclones. A cold-surface ridge axis extends southeastward from Alaska to the Mexican border east of the Rockies, which provides a natural barrier from the moderating influence of the North Pacific. The combination of the surface anticyclones and a deep cyclone centered south of Greenland provides a northerly low-level geostrophic flow extending from the Arctic to the Caribbean Sea.

Topographic features, such as the Rockies, are often important to the evolution of the cold-air outbreak. A relatively warm sea surface temperature (SST) in the vicinity of the tropospheric thickness ridge would enhance its amplitude. A snow cover over the eastern part of North America acts to amplify the thickness trough.

As in North America, cold-air outbreaks in other regions of the globe are characterized by an upper-level pattern that favors a cold continental surface anticyclone traveling into the affected region. Significant cold-air outbreaks in western Europe, for example, are characterized by southwestward-traveling anticyclones from the continental regions of Russia.

3 WIND

High winds during the winter are often associated with blizzard conditions. The official definition of a blizzard includes large amounts of falling *or* blowing snow, with wind speeds greater than 56 km/h *and* visibilities less than 0.4 km for a least 3 h. Blizzards are substantial threats to North American east coastal regions because intense surface cyclones and their accompanying strong horizontal pressure gradients occur preferentially in these regions. Coastal residents are therefore especially vulnerable to the effects of these deep surface lows, which may include hurricane-force winds and storm surges comparable to those associated with land-falling hurricanes. Other locations in North America that experience blizzard conditions include the high-plains region extending east of the Rocky Mountains. Blizzard conditions are more likely to occur in these areas in association with high winds and blowing (with small amounts of falling snow) over relatively smooth terrain.

North American east coast cyclones develop in response to a middle-upper tropospheric mobile trough that typically travels along a northwesterly upper-level flow. The initial development of the coastal surface low may be in a preexisting zone of coastal frontogenesis that separates warm, moist marine air from cold, dry air that is dammed east of the Appalachians. The interaction of this lower-tropospheric cyclonic potential vorticity anomaly with its upper-level mobile counterpart defines the cyclogenesis.

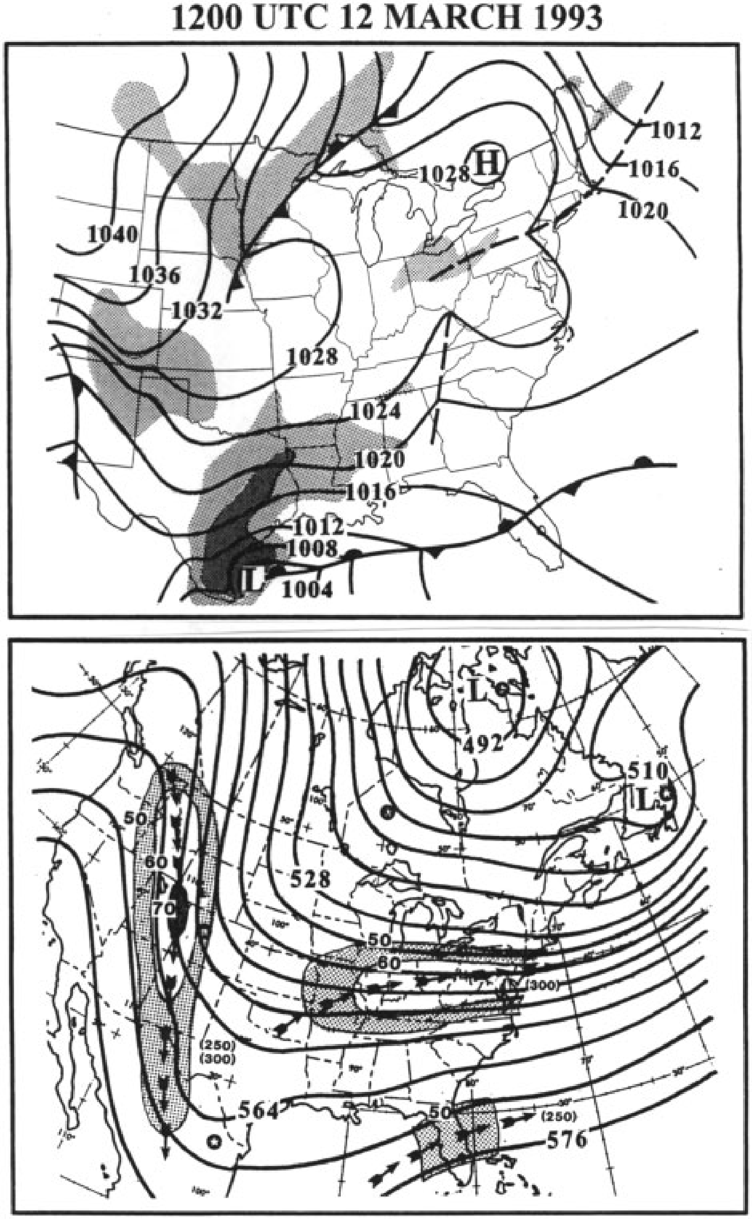


Figure 2 (*a*) Sea-level pressure (solid, interval of 4 hPa) for 1200 UTC, March 12, 1993. (*b*) Heights (solid, interval of 6 dam) for 1200 UTC, March 12, 1993. (From Kocin et al., 1995; reprinted by permission of AMS.)

An extreme example of North American cyclogenesis occurred in March 1993. The stage was set with a cold-air outbreak covering most of eastern North America southward into the Gulf of Mexico (Fig. 2*a*). Rapid cyclogenesis in the Gulf of Mexico occurred in response to multiple interactions with upper-tropospheric mobile troughs approaching from the west and northwest (Figs. 2*b* and 3). Additional intensification was associated with the latent heat release in embedded thunderstorms. Explosive intensification of 30 hPa during the next 24 h produced an unprecedented 968-hPa cyclone in central Georgia by 1200 Universal Time Coordinated (UTC), March 13 (Fig. 4*a*). At this time, the unusually large horizontal scale of the cyclonic circulation extended from the Caribbean Sea northward into New England and westward to the Mississippi River. The system continued to deepen to its minimum pressure of 962 hPa during the next 12 h, in a favorable location for interactions with the upper trough (Fig. 4*b*).

This Storm of the Century produced record low pressures at several locations along the North American coast, blizzard conditions to the north and west of its track, and a storm surge, comparable to that seen in hurricanes, along the Gulf of Mexico coast of 3 to 4 m. Its development was triggered by multiple troughs aloft, but the influence of topography was also substantial. First, the presence of the anomalously warm Gulf of Mexico in association with a cold-air outbreak destabilized the air. This destabilization enhanced the cyclogenesis by enhancing the interactions with upper-level troughs. Second, the warming and moistening of the air mass provided a favorable environment for the heavy thunderstorm activity occurring near the cyclone center. This enhanced thunderstorm activity aided the cyclogenesis through the addition of latent heat of condensation.

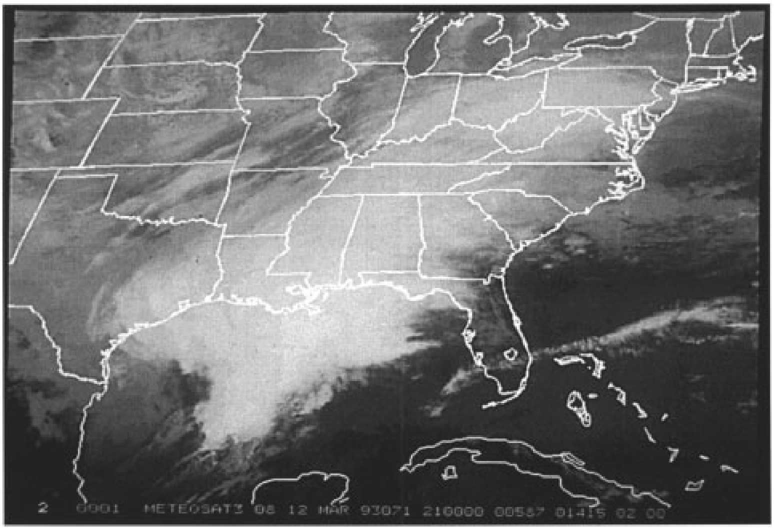


Figure 3 Infrared satellite image for 2100 UTC, March 12, 1993.

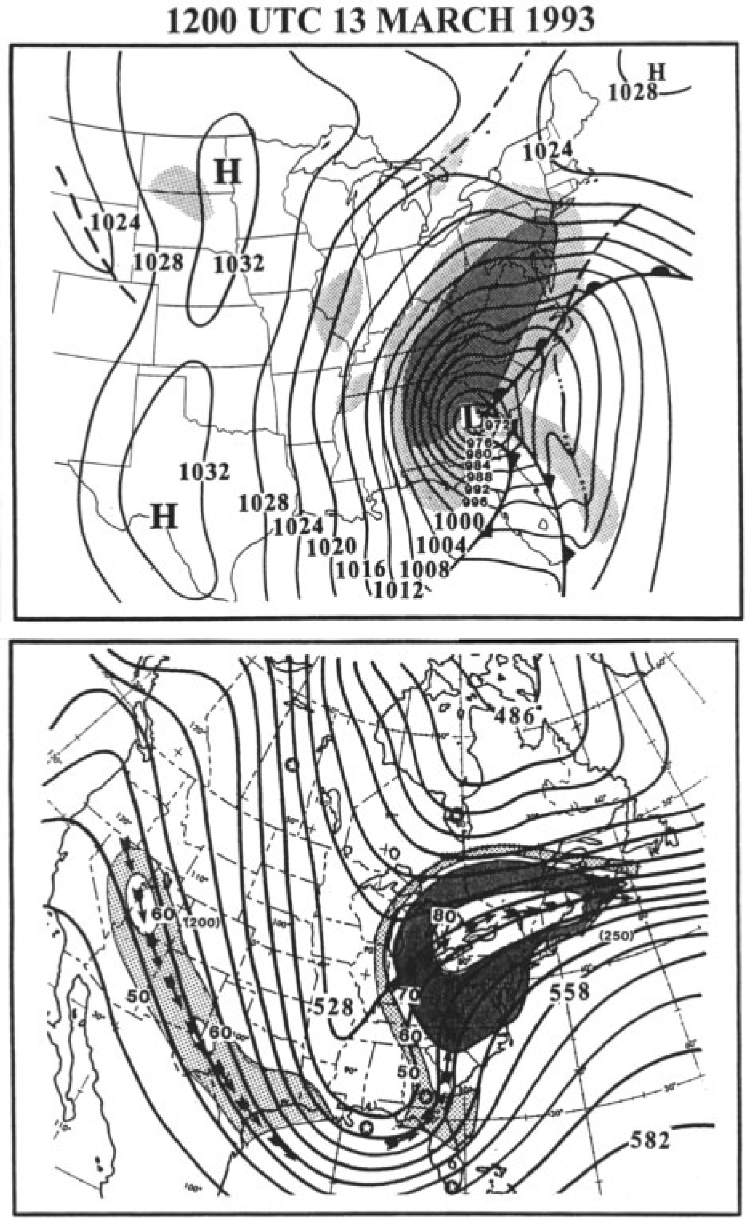


Figure 4 As for Fig. 2, except for 1200 UTC, March 13, 1993. (From Kocin et al., 1995; reprinted by permission of AMS).

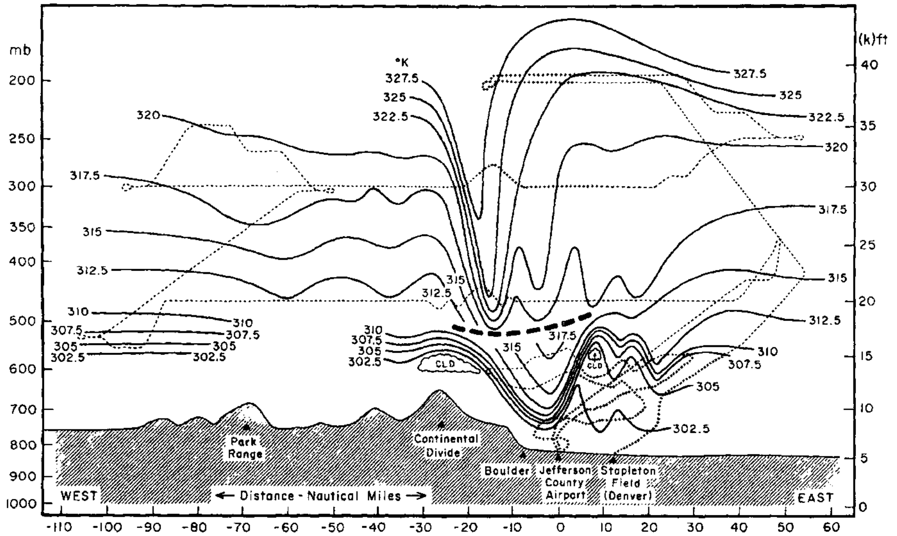


Figure 5 Cross section of potential temperature (K) along an east–west line through Boulder, as obtained from analysis of aircraft data on January 11, 1972. The isentropes are excellent indicators of streamlines for steady, adiabatic flow. Flight tracks are indicated by light dashed lines, with the heavy dashed line separating the tracks of the different research aircraft. (From Klemp and Lilly, 1975; reprinted by permission of AMS).

Strong winds also can occur during the cold season in the lee of mountain ranges. These wind storms, typically termed chinook winds, may contain hurricane-force wind speeds and produce extensive property damage. Klemp and Lilly (1975) studied a case that occurred in Boulder, Colorado, in which the phenomenon is found to be the surface signal of a standing gravity wave with a wavelength of ~60 km and an upwind temperature inversion approximately 60 hPa above the Continental Divide (Fig. 5). This extreme event was associated with surface wind gusts of 50 m/s and extensive property damage. The vertical cross section of wind speeds (Fig. 6) shows the core of extreme wind located 20 km to the east of the Divide. Generally, similar events with gusts in excess of hurricane force occur each winter. Such events occur when the westerlies are strong, the inversion above the mountains is also strong, and in the lee of where the westerlies arrive at the Divide unimpeded by upstream mountains.

4 PRECIPITATION

One of the more common synoptic-scale structures associated with cold-season precipitation along the west coast of North America is the so-called pineapple express. The synoptic-scale structure of this event is similar to that observed for most substantive cold-season rainfall events along the west coast of North America.

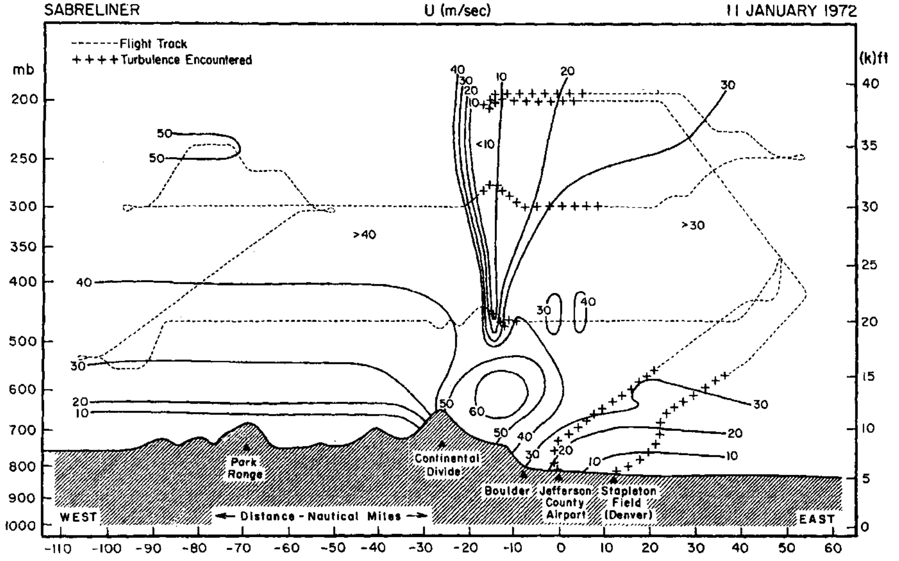


Figure 6 Horizontal velocity (m/s) contours along the cross section as in Fig. 5 and derived from research aircraft observations. (From Klemp and Lilly, 1975; reprinted by permission of AMS).

The dominant lifting mechanism is the effect of cyclonic thermal vorticity advection, in addition to anomalously large precipitable water amounts. Large amounts of water vapor are transported into the region from subtropical maritime regions near Hawaii by anomalously strong southwesterly tropospheric flow. These events can produce large amounts of precipitation, induce flooding and landslides, and generally produce great economic damage.

The meteorological effects of the pineapple express are amplified by the existence of the north–south oriented mountain ranges that exist along the west coastal regions of North America. The orographic lifting and cooling of the eastward-traveling moisture-laden subtropical air enhances the precipitation rates on the windward slopes of these mountains. The largest measured seasonal snowfalls in the world occur in these areas. Washington State's Paradise Ranger Station, at 1654 m on Mount Rainier in the Cascades, averages 1733 cm (682.4 in.) of snowfall each year. The Mount Baker ski resort, at 1286 m in the north Cascades of Washington, recorded a single-season total snowfall of 2896 cm (1140 in.) during 1998–1999, establishing a new world record [previously held by Paradise Ranger Station with 2850 cm (1122 in.) during 1971–1972].

The leeward regions to the east of the Cascades, in contrast, receive relatively small amounts of precipitation. This rain shadow effect is due to the compressional warming of the easterly traveling air as it subsides on the lee slopes of the mountains. A comparison of annual precipitation amounts illustrates this effect. Tacoma, Washington, at an elevation of 89 m, receives 939.8 mm (37.0 in.), whereas Coulee

Dam, to the east of the Cascades at an elevation of 518 m, receives only 274.3 mm (10.8 in.).

An example of the synoptic pattern relating to events in western Washington and Oregon is shown in Fig. 7 (Lackmann and Gyakum, 1999). The 46-case composite corresponds to days in which at least 12.5 mm of precipitation falls on each of the four stations shown in Fig. 7*d*, and in which the maximum temperature exceeded 10°C for each of the lowland stations and 5°C for Stampede Pass. The southwesterly geostrophic flow in the region of the precipitation exists from the surface to 500 hPa. The composite surface low in the Gulf of Alaska has an upper-level counterpart with a planetary-scale trough that extends throughout the Pacific Ocean. Figure 8 shows anomalously strong 500-hPa southwesterly flow prior, during, and after the event. This strong flow extends from the subtropical regions of the Pacific to Oregon, Washington, and British Columbia.

An extreme case of heavy rains along the western North American coast in late 1996 was part of a 2-week regime in which 250 to 1000 mm of rain fell, with economic damage of up to $3 billion. The large-scale SLP and 1000–500-hPa layer mean temperature fields for 1200 UTC, December 31, 1996 (Fig. 9), appear very similar to those shown in Figure 7, with a deep planetary-scale trough covering much of the south-central North Pacific Basin. The coldest tropospheric anomaly corresponds to the stratospheric intrusion of high-PV (potential vorticity) air seen as the dark region extending southwestward from the comma cloud of Figure 10. This long plume of large PV provides forcing for the 960-hPa surface cyclone seen in Figure 9. As pointed out by Lackmann and Gyakum (1999), these cyclonic disturbances are responsible for transporting the subtropical water vapor from marine areas near Hawaii directly toward the west coast of North America.

Snowfall may occur in ascending regions of surface cyclones and regions of frontogenesis. However, especially large snows are strongly affected by topography. One example of such a strong topographic influence on snowfall is that of the lake effect snowfalls. The most prominent region of North America that is affected by these snowfalls is that of the Great Lakes. However, any comparably large body of water in extratropical latitudes exerts a similar influence on the surrounding region. The Great Salt Lake in Utah and James Bay in Quebec are two other examples of North American lake effect regions.

Figure 11, derived from Eichenlaub's (1979) work and in turn printed in the review article by Niziol et al. (1995), shows the mean annual snowfall climatology for the Great Lakes. The amounts vary from less than 50 cm in the southern region to nearly 500 cm in regions east and south of Lake Superior. The existence of both the Great Lakes and orography controls the variability of these climatological snowfalls. The lake effect snowfall of December 17–19, 1985, which affected the Buffalo, New York, area, was typical of the associated large-scale conditions. The environment was characterized by a deep surface cyclone that traveled well to the north and east of the affected region (Fig. 12), in this case a 968-hPa low located between Greenland and Labrador. This low, combined with a 1040 surface anticyclone over the Dakotas, advected bitterly cold air directly over the Great Lakes. The 1000–500-hPa mean temperature over Lakes Superior, Michigan, and Huron was about 20°C colder

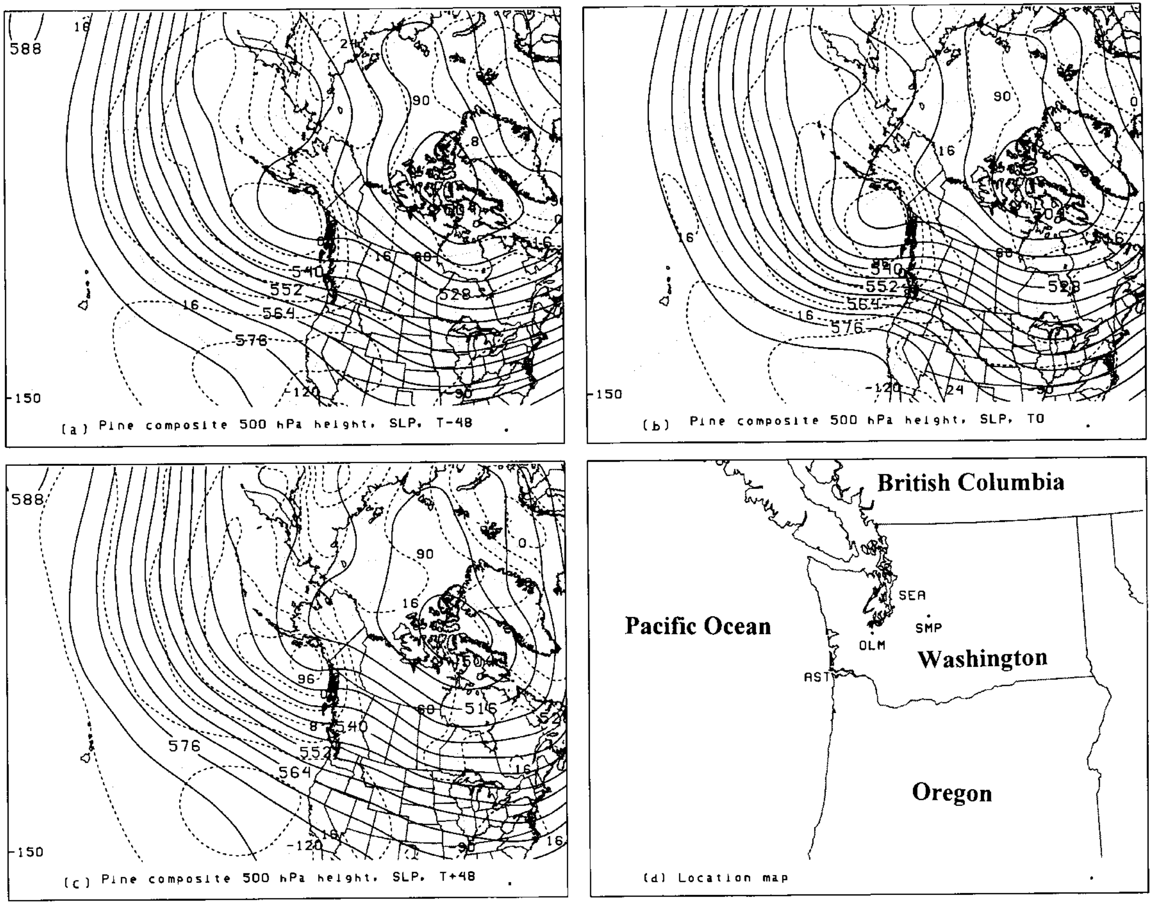


Figure 7 Pineapple Express composite of 500-hPa height (solid, interval of 6 dam) and sea-level pressure (dashed, interval of 4 hPa) of 46 cold-season cases for (*a*) 48 h prior, (*b*) the day of, and (*c*) 48 h after the event. Geographical references are displayed in (*d*), including Astoria, OR (AST); Olympia, WA (OLM); Seattle-Tacoma airport, WA (SEA); and Stampede Pass, WA (SMP). (From Lackmann and Gyakum, 1999; reprinted by permission of AMS.)

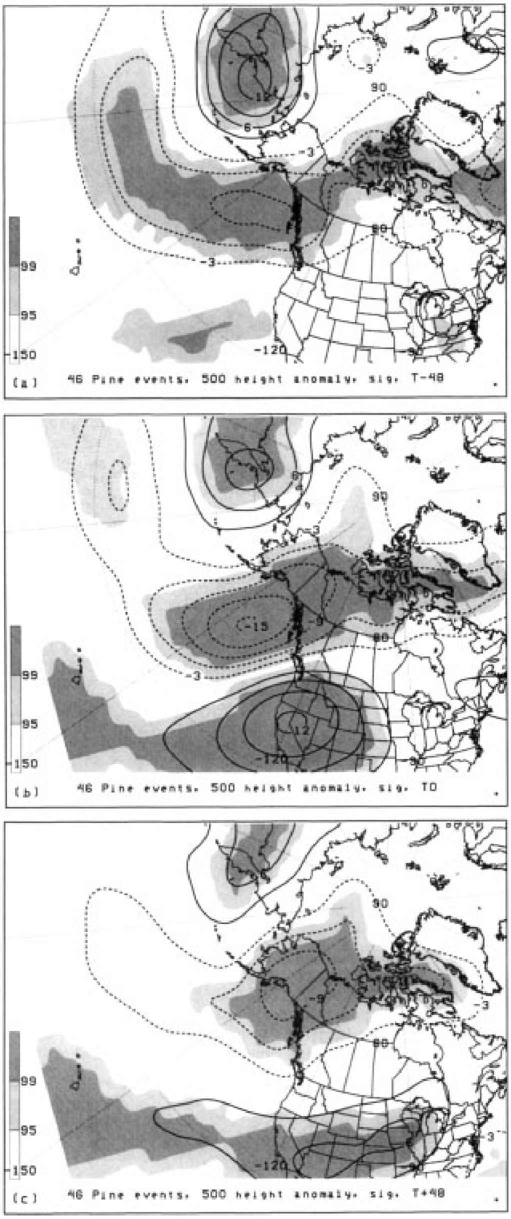

99
95
150
-3
-6
-12
-15
90
80
-120
(a) 46 Pine events, 500 height anomaly, sig, T-48
(b) 46 Pine events, 500 height anomaly, sig, T0
(c) 46 Pine events, 500 height anomaly, sig, T+48

Figure 8 Composite 500-hPa geopotential height anomaly [contour interval 3 dam, positive (negative) values solid (dashed), zero contour omitted] and statistical significance determined from a two-sided Student's *t*-test (shading intervals correspond to 95 and 99% confidence limits as shown in legend at left of panels): (*a*) 48 h prior, (*b*) the day of, and (*c*) 48 h after the event. The light (dark) shading denotes regions where there exists a greater than 95% (99%) probability that the composite belongs to a population distinct from that of climatology. (From Lackmann and Gyakum, 1999; reprinted by permission of AMS).

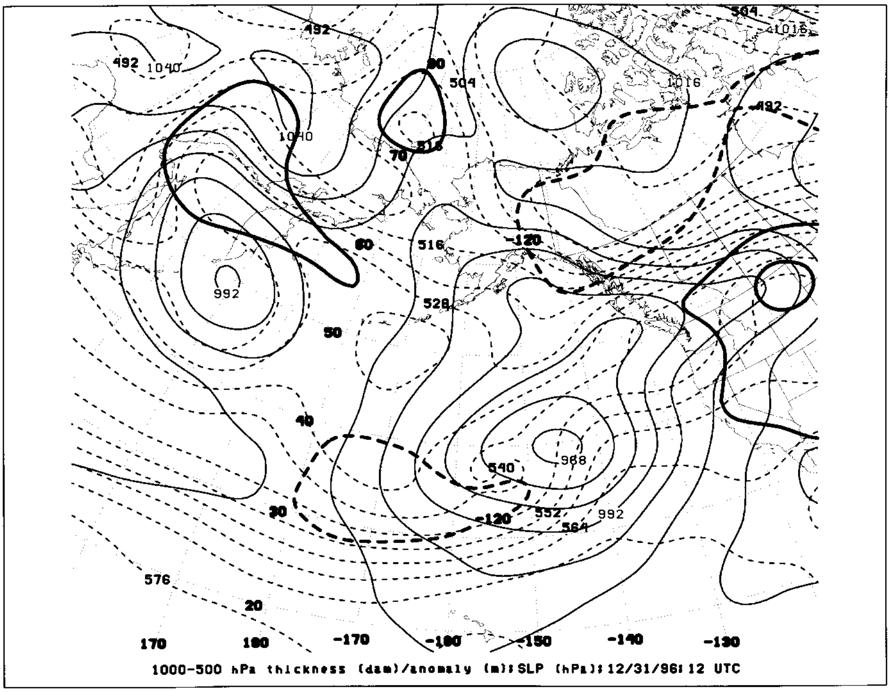


Figure 9 As for Fig. 1, except 1200 UTC, December 31, 1996.

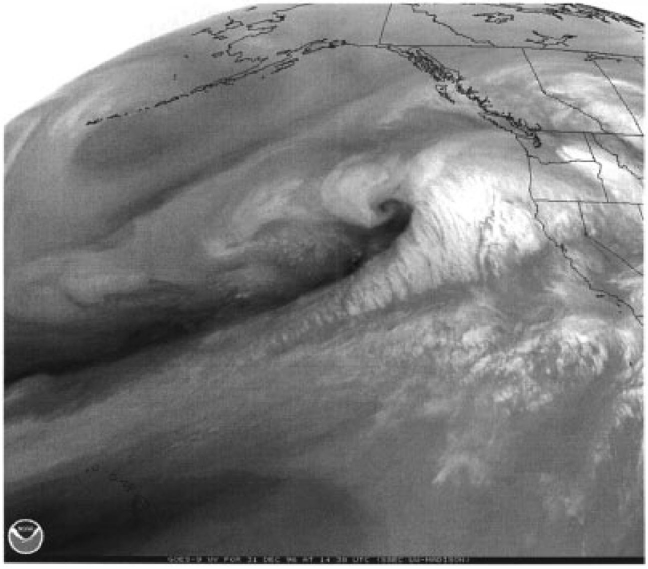


Figure 10 Water vapor image for 1430 UTC, December 31, 1996. (Courtesy of NOAA).

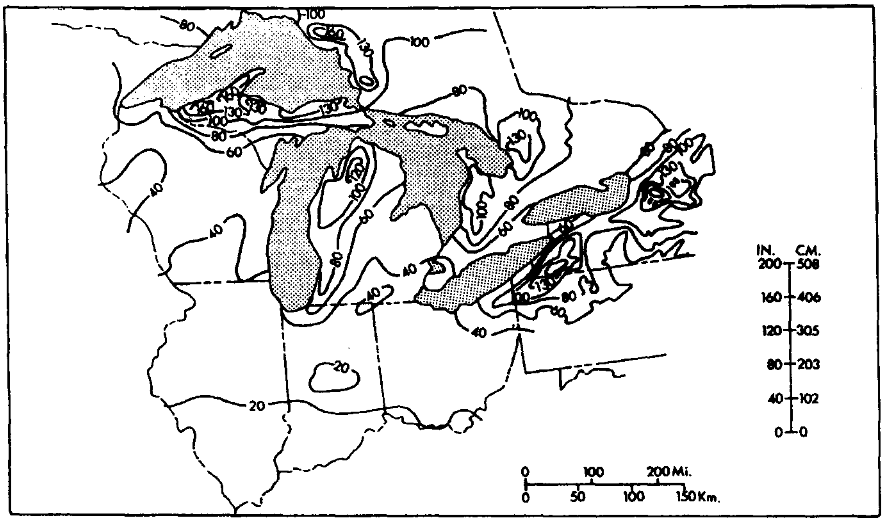


Figure 11 Mean annual snowfall (in.) for the Great Lakes region. (From Niziol et al., 1995; in turn from Eichenlaub, 1979; reprinted by permission of AMS).

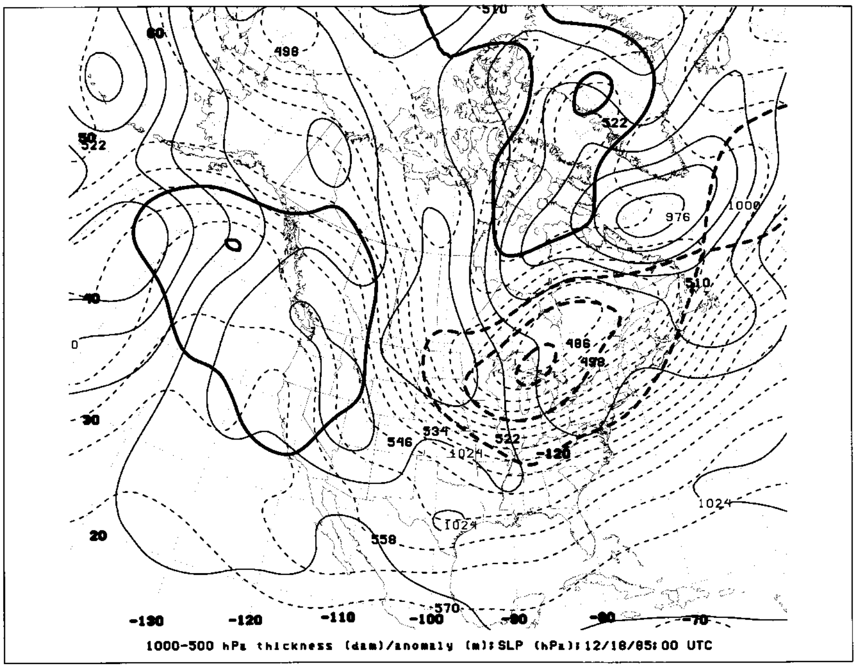


Figure 12 As for Fig. 1, except 0000 UTC, December 18, 1985.

than the climatological average. The tropospheric air was about 12°C colder than the mean in Buffalo. Despite the fact that the most substantial synoptic-scale ascent typically is located to the north and east of a developing surface low, owing to a favorable combination of warm advection and cyclonic vorticity effects, extremely large snowfall rates in excess of 5 cm/h were observed in the Buffalo area. Several factors contributed to such an extreme snowfall rate. These include large-scale ascent associated with a strong upper-level trough. Within this environment, several crucial mesoscale conditions amplify the response to this advection. First, the cold air traveling over the relatively warm waters will hydrostatically destabilize the air. Second, the evaporation of water vapor from the lakes will saturate the air, so that the effective hydrostatic stability will be reduced even further.

Additional physical processes are responsible for enhancing the snow. These include the development of a surface trough (Fig. 12) that provides the larger-scale ascent necessary to trigger moist convection in the presence of instability. The existence of sensible heat transfer from the unfrozen lakes to the atmosphere contributes to ascent. Additionally, the differential roughness between the lakes and the surrounding land will create low-level convergence areas with accompanying ascent. Often, as in the case of the December 17–19, 1985, event, mesoscale bands of heavy snow will develop (Fig. 13), and a key forecast problem is to predict the existence and movement of such band(s).

Freezing rain events can have devastating impacts on economic infrastructure. The synoptic environment of freezing rain is characterized typically by a surface extratropical cyclone advecting warm, moist air above relatively shallow, cold air masses. Areas that experience persistent shallow, cold air are therefore prone to freezing rain. Especially susceptible regions include larger valleys and basins.

Cortinas (2000) has documented the mean synoptic conditions for freezing rain events in the Great Lakes region. Figure 14 illustrates that freezing rain occurs to the northeast of a surface cyclone that advects warm, moist air poleward from either the Gulf of Mexico or the Atlantic Ocean. Typically, the air to the northeast of the low had been associated with a prior cold-air outbreak. The temperature stratification becomes very stable (Fig. 15), as the warm, moist air travels above the relatively cold dense near-surface air. As Figure 15 shows, the lowest layer during a freezing rain event is characterized by an inversion with near-surface air that is less than 0°C, with a deep layer of air aloft that is greater than 0°C. The strong inversion helps to prevent any turbulent mixing of the warm air aloft down to the surface. Furthermore, the stratification decreases markedly above the inversion, to the extent that nearly moist adiabatic conditions in the free atmosphere may exist. The implication for precipitation amounts is that substantial ascent, or even convection, may occur in these upper weakly stratified layers. The lowest-layer air is typically from the east or northeast, and this reinforces the inversion with cold-temperature advection from the down-shear polar air mass. The winds at the top of the inversion are typically warm and moist and blowing from the southwest. Therefore the vertical wind shear reinforces the temperature inversion.

An extreme example of such an event occurred during January 5–9, 1998, when an ice storm deposited greater than 100 mm of predominantly freezing rain in south-

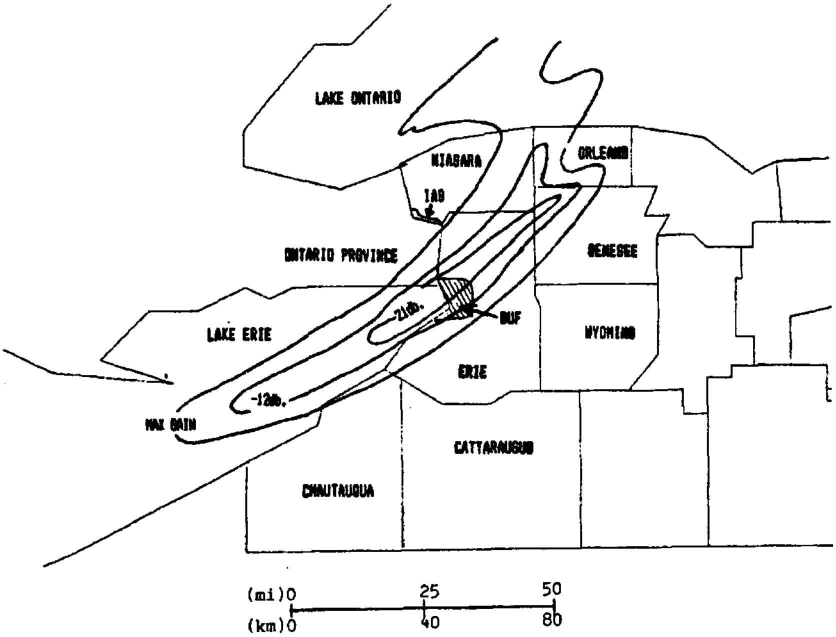


Figure 13 Precipitation echoes from the WSR-57 radar at Buffalo for the Lake Erie snowstorm on 2330 UTC, December 17, 1985. Snowfall intensities are contoured for local use, with linear attenuation, to depict the areas of heaviest snowfall more accurately. (From Niziol, 1987; reprinted by permission of AMS).

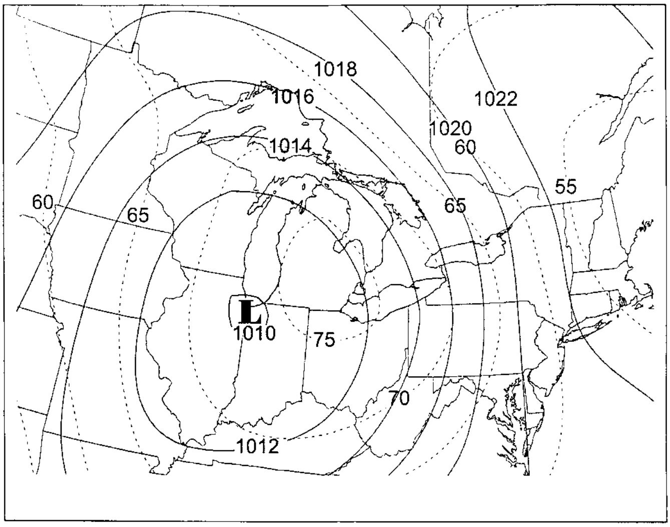


Figure 14 Mean sea-level pressure (hPa, solid line) and 1000–666-hPa relative humidity (%, dashed line) during freezing rain events over the central Great Lakes for the period 1976–1990. (From Cortinas, 2000; reprinted by permission of AMS).

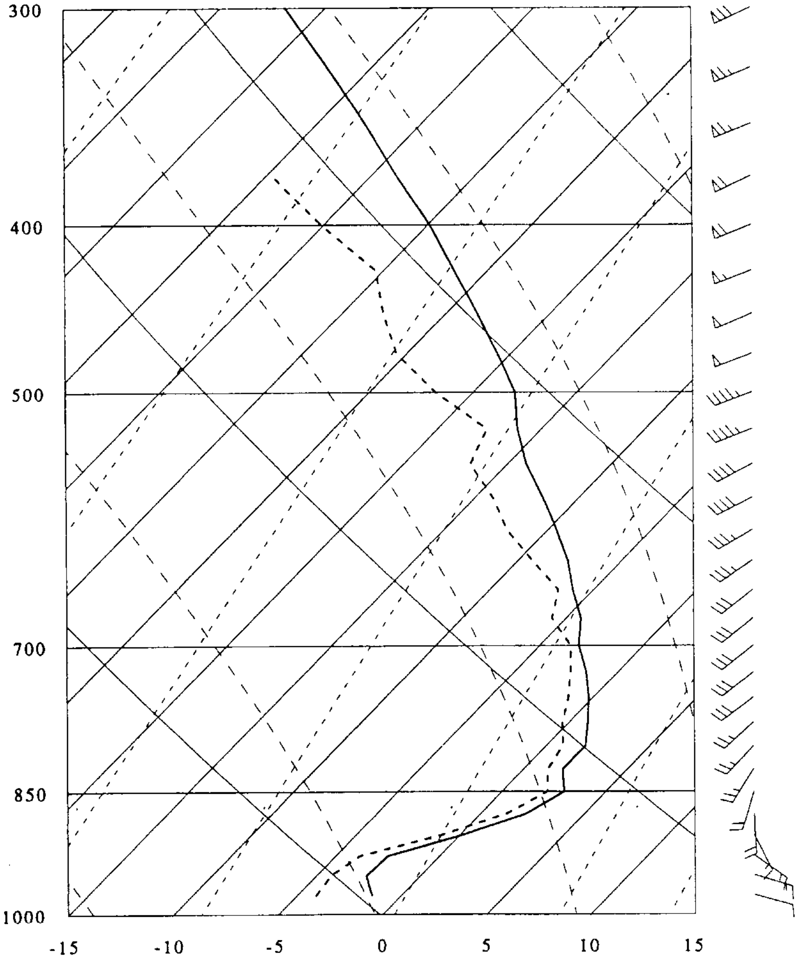


Figure 15 Median values of dry-bulb temperature (°C, dark solid line) and dewpoint temperature (°C, dark dashed line) from a distribution of freezing rain soundings at Flint, Michigan. Data are plotted on a skew *T*–log *p* diagram, with pressure (vertical axis) given in hPa and winds shown by full barb = 5 m/s and half barb = 2.5 m/s. (From Cortinas, 2000; reprinted by permission of AMS).

ern Quebec and upstate New York. Approximately 44 fatalities and damages in excess of $4 billion occurred (NOAA, 2000). The synoptic conditions at 1200 UTC, January, 8, 1998 (Fig. 16), were typical for the event. The freezing rain occurred in a strong geostrophic deformation zone that extended northeast of a surface low from upstate New York into New England. Surface temperatures during the event ranged from −5 to −1°C in extremely large 1000–500-hPa thicknesses of 552 dam or greater. The anomalously strong Bermuda anticyclone assisted

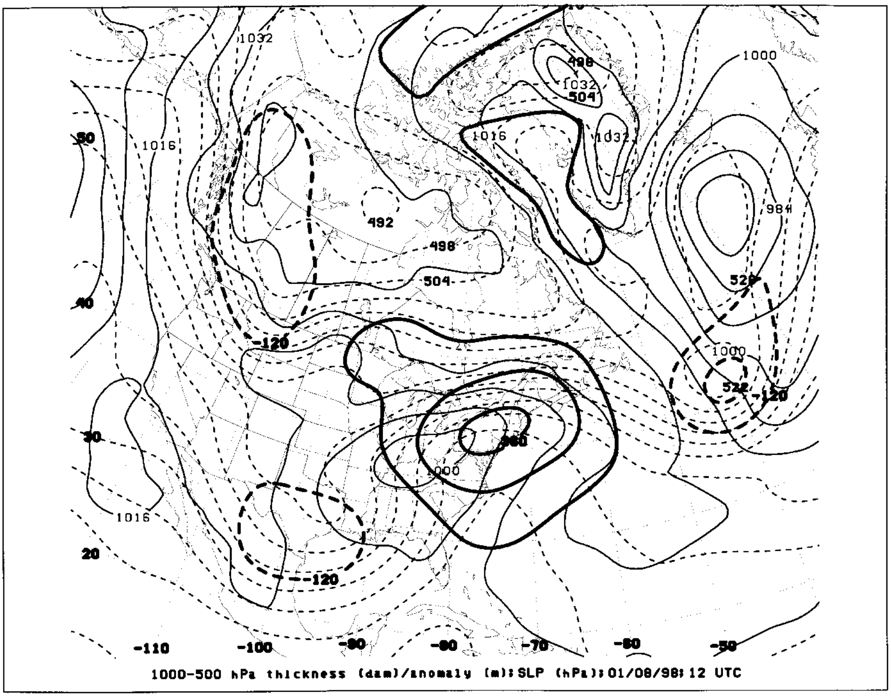


Figure 16 As for Fig. 1, except for 1200 UTC, January 8, 1998.

the tropospheric warming of the affected regions with northward transports of warm and moist air from the Caribbean Sea. Surface winds advected cold air from the northeast along the Saint Lawrence River Valley. These surface winds occurred along the southern flank of a bitterly cold 1044-hPa surface anticyclone centered in the Canadian Northwest Territories. While the instantaneous synoptic-scale features were unusually strong for a freezing rain event, the factor that produced such an extreme event was the persistence for 5 days of this freezing rain–producing pattern.

5 SUMMARY

We have discussed several meteorological phenomena associated with winter weather. These events, associated with extreme cold, wind, and precipitation, have the potential to produce substantial economic impact (NOAA, 2000). We have described these events in terms of the larger-scale meteorological patterns that are conducive to their existence. Additionally, we have seen how these events are produced by the modulation of larger-scale atmospheric patterns by interesting topographic features, such as mountains, lakes, and oceans. Although we have focused our attention on North American phenomena, similar events exist in other regions of the world, where similar interactions exist between the atmospheric flows and topography.

REFERENCES

Cortinas, J., Jr. (2000). A climatology of freezing rain in the Great Lakes region of North America, *Mon. Wea. Rev.* **128**, 3574–3588.

Eichenlaub, V. L. (1979). *Weather and Climate of the Great Lakes Region*, Notre Dame, IN, University of Notre Dame Press.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph (1996). The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteor. Soc.* **77**, 437–471.

Klemp, J. B., and D. K. Lilly (1975). The dynamics of wave-induced downslope winds, *J. Atmos. Sci.* **32**, 320–339.

Kocin, P. J., P. N. Schumacher, R. F. Morales, Jr., and L. W. Uccellini (1995). Overview of the 12–14 March 1993 superstorm, *Bull. Am. Meteor. Soc.* **76**, 165–182.

Lackmann, G. M., and J. R. Gyakum (1999). Heavy cold-season precipitation in the northwestern United States: Synoptic climatology and an analysis of the flood of 17–18 January 1986, *Wea. Forecasting* **14**, 687–700.

NOAA (2000). Billion Dollar US. Weather Disasters, 1980–2000 (available on the World Wide Web at http://www.ncdc.noaa.gov/ol/reports/billionz.html).

Niziol, T. A. (1987). Operational forecasting of lake effect snowfall in western and central New York, *Wea. Forecasting* **2**, 310–321.

Niziol, T. A., W. R. Snyder, and J. S. Waldstreicher (1995). Winter weather forecasting throughout the Eastern United States. Part IV: Lake effect snow, *Wea. Forecasting* **10**, 61–77.

Quiroz, R. S. (1984). The climate of the 1983–84 winter—a season of strong blocking and severe cold in North America, *Mon. Wea. Rev.* **112**, 1894–1912.

CHAPTER 28

TERRAIN-FORCED MESOSCALE CIRCULATIONS

JOHN HOREL

1 INTRODUCTION

The characteristics of the Earth's surface affect climate and weather on all spatial scales (Barry, 1992). However, many weather phenomena that are influenced by surface inhomogeneities in elevation, moisture, temperature, snow cover, vegetation, or roughness are organized on the mesoscale, which spans the range from 2 to 200 km. According to Pielke (1984), weather systems on the mesoscale can be divided into two general categories: those that are forced primarily by instabilities in traveling large-scale disturbances (e.g., squall lines or mesoscale convective complexes) and those that are forced by surface inhomogeneities (e.g., mountain/valley circulations, sea breezes, or urban circulations).

The impact of mesoscale variations in the underlying land surface is evident in an estimate of the average annual precipitation over the United States (Fig. 1). As a general rule, precipitation increases locally as terrain height increases, for example over the Appalachians and the mountain ranges of the western United States. A further general rule is that precipitation is higher near the coasts, where moisture is more abundant, than further inland.

This section emphasizes mesoscale weather phenomena that are strongly modulated by the characteristics of the underlying surface. After discussion of thermally driven and mechanically driven flows, the impact of terrain upon precipitation processes will be presented. The influence of other variations in surface properties

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

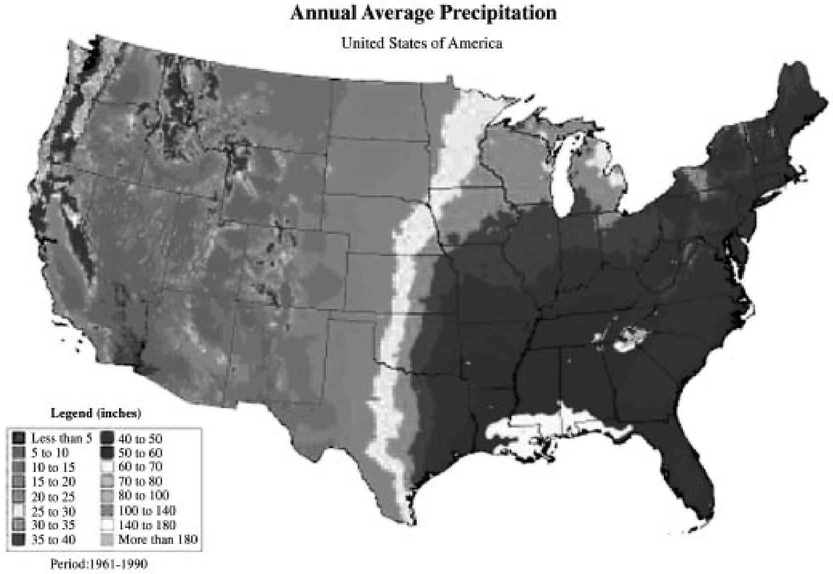


Figure 1 (see color insert) Average annual precipitation in inches, as determined from the Parameter-elevation Regression on Independent Slopes (PRISM) model by Chris Daly, based on 1961–1990 normals from National Oceanic and Atmospheric Administration (NOAA) cooperative stations and Natural Resources Conservation Service (NRCS) SNOwpack TELemetry (SNOTEL) sites. Modeling sponsored by USDA-NRCS Water and Climate Center, Portland, Oregon. Available from George Taylor, Oregon State Climatologist, Oregon Climate Service. See ftp site for color image.

upon mesoscale circulations is followed by discussion of the predictability of terrain-forced mesoscale circulations.

2 IMPACT OF TERRAIN UPON WIND

Thermally Driven Flows

Mesoscale wind circulations are produced frequently by temperature contrasts that develop as a result of terrain variations. As noted by Pielke and Segal (1986), the general nature of diurnal sea and land breezes was well understood in ancient times: “southward goes the wind, then turns to the north; it turns and turns again” (Ecclesiastes 1:6). Sea (or lake) and land breezes are driven by horizontal temperature contrasts that develop between water bodies and adjacent land surfaces (Whiteman, 2000). As shown in Figure 2 for the Great Salt Lake (the largest body of water in the continental United States to the west of the Great Lakes), differences in air temperature develop over the water and land surfaces as a result of the higher heat capacity of the water relative to that of the surrounding land. In the case of the Great Salt Lake during summer, the air over the lake is roughly 1°C warmer than the air over the surrounding land at night and 4°C cooler during the day. These temperature

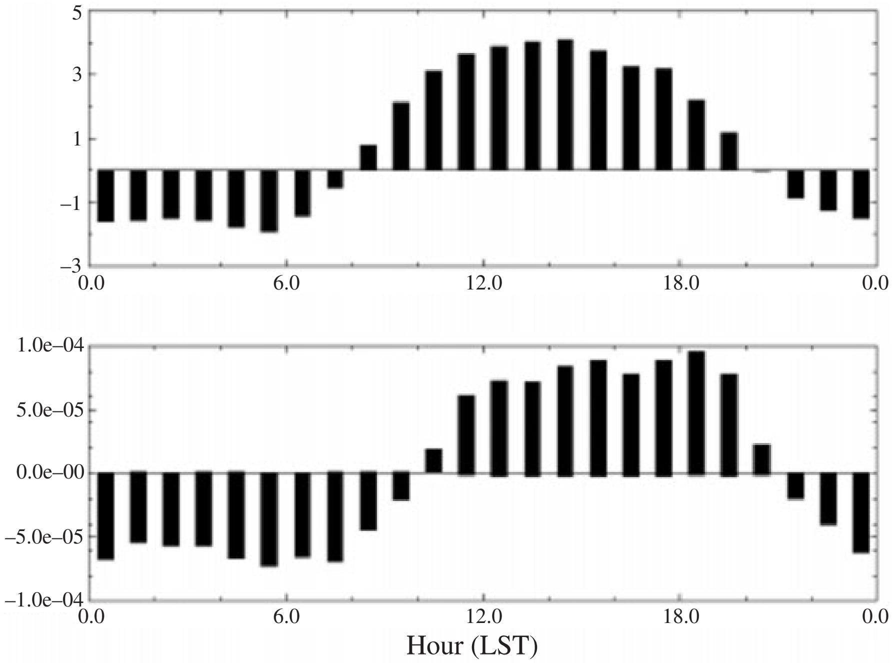


Figure 2 Top panel: Diurnal variation in the difference in air temperature (°C) over the land surrounding the Great Salt Lake, Utah, versus the air over the lake during August 1999. The air over the lake is warmer than the air over the land during the night. Bottom panel: Diurnal variation in surface wind divergence (s^{-1}) over the Great Salt Lake. The air converges over the lake at night and diverges away from the lake during the day (a similar diurnal variation in surface wind divergence is evident over Lake Ontario; Chen, 1977).

differences drive the air from over the water to over the land during the day (lake breeze) and drive the air from over the land to over the water at night (land breeze). The resulting convergence of the surface winds at night over the Great Salt Lake and divergence during the day are evident in Figure 2.

Mountain–valley circulations arise as a result of differential heating between the ground in regions of complex terrain and the free atmosphere at the same elevation. There are two broad categories for mountain–valley circulations: slope flows and along-valley winds (Whiteman, 1990, 2000). Typically, slope flows are driven by horizontal temperature contrasts between the air over the valley sidewalls and the air over the center of the valley. Since a larger diurnal temperature variation occurs near the ground, the higher terrain above the valley serves as a heat source during the day and a heat sink at night. These temperature differences between the slope and free air over the valley drive cold, dense air flowing down the slope at night and warm, light air surging up the slope during the day. Upslope flows tend to be deeper than the shallow nocturnal drainage flows as a result of strong turbulent mixing during the day.

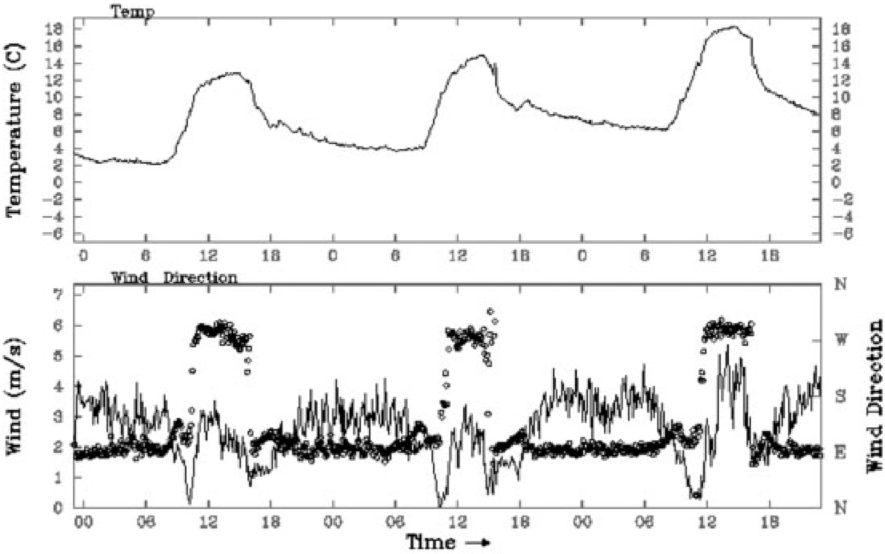


Figure 3 Diurnal mountain–valley circulation in Big Cottonwood canyon in the Wasatch Mountains of northern Utah. Top panel: temperature (°C) from midnight local standard time (LST) October 15 to midnight October 18, 2000. Bottom panel: wind speed (solid line in m/s) and wind direction (circles) for the same period. The weather station is located where the canyon is oriented east–west.

Along-valley winds develop as a result of variations in the strength of slope flows within the valley or differences in temperature between the air in the valley and that found over the adjacent lowlands. As an example of along-valley circulations, Figure 3 shows the evolution of wind and temperature at a station located roughly midway along the length of the Big Cottonwood Canyon in the Wasatch Mountains of northern Utah. During this 3-day period, a warming trend was underway and the skies remained clear; further, the station location within this narrow, steep canyon was shaded during the morning and late afternoon. The diurnal temperature swing of roughly 10°C drove a reversal in the wind from downvalley at night (from the east) to upvalley during the day (from the west). The sudden onsets of the wind reversals were particularly striking during these 3 days.

Dynamically Driven Flows

As summarized by Whiteman (2000), the degree to which a hill or mountain affects the air flow depends upon the characteristics of the terrain feature (e.g., height, width, roughness, orientation relative to flow direction) and the upstream speed and stability of the air. Since the speed and stability of the flow can vary significantly with height, the impact of the terrain upon the flow can change from one atmospheric layer to another (Fig. 4).

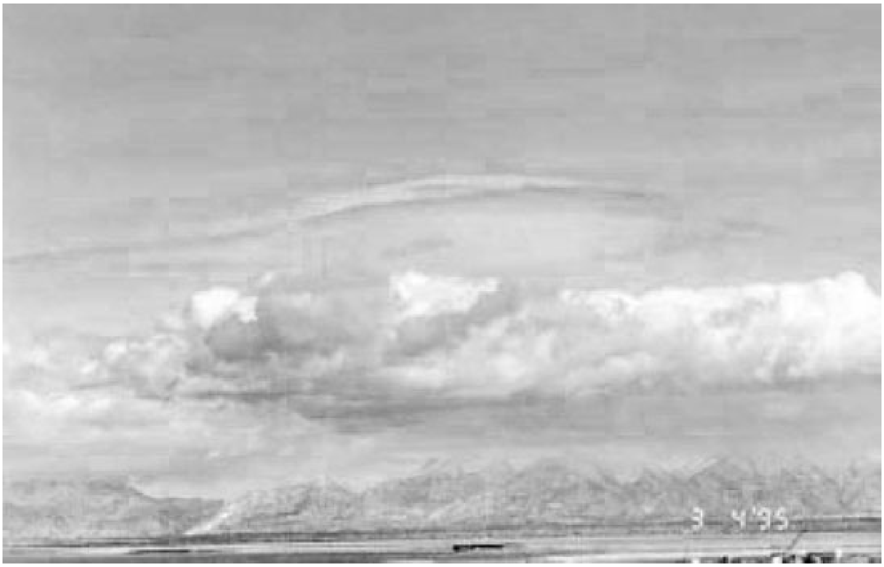


Figure 4 (see color insert) Cumulus clouds, indicative of shallow instability, above the Wasatch Mountains in northern Utah are capped by lenticular wave clouds, indicative of stable air flowing across the mountain barrier (photo by J. Horel). See ftp site for color image.

Whether an obstacle significantly obstructs the flow depends upon whether or not the approaching air has enough kinetic energy to lift the air over it. The nondimensional Froude number (defined as $\mathrm{Fr} = u/Nh$, where u is the speed of the upstream flow, N is the Brunt–Vaisala frequency, a measure of stability, and h is the height of the terrain) provides a conceptual framework to assess the impact of terrain height and flow speed and stability. If the flow approaching a relatively low obstacle has strong winds and weak stability (Froude number greater than 1), then there is sufficient kinetic energy to cross the obstacle. On the other hand, if the flow approaches a relatively high barrier with weak winds and strong stability (Froude number less than 1), then there is not enough kinetic energy to force the air over the obstacle and the flow is either channeled through gaps in the terrain or forced to travel laterally around the obstacle. For example, Figure 5 shows fog spilling through a gap in the coastal mountains of northern California as a result of strong stability at crest level. The flow is more likely to be blocked and required to travel around obstacles if the barrier has a relatively short lateral extent or the flow impinging upon the mountain range is very shallow.

As stable air flows across a mountain barrier (with Froude number less than 1), mountain gravity waves are often created over or in the lee of the barrier (Durran, 1990). The structure of the mountain waves exhibited at any particular time depends upon the charateristics of the barrier (e.g., height, width, multiple ridges) as well as the stability, orientation of the flow relative to the ridge crest, and vertical change of wind with height. If sufficient moisture is present, the uppermost portion of the wave

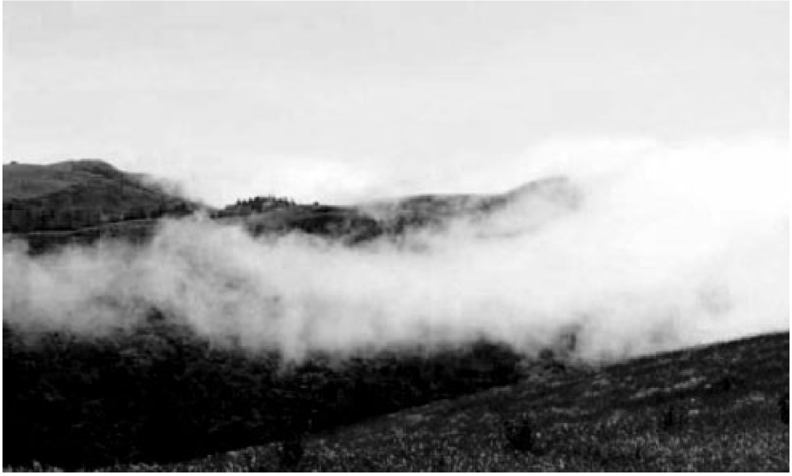


Figure 5 (see color insert) Fog channelled through a gap in the coastal mountains of northern California (photo by J. Horel). See ftp site for color image.

may become visible (e.g., the lenticular clouds in Fig. 4). Viewed from space (Fig. 6), the wavelike nature of the gravity waves is readily apparent. The waves embedded within these flows result from the response to the forced ascent: After the air is carried aloft as high as possible, the lifted air is cooler (and heavier) than the surrounding air and displaced back toward its original level by gravity.

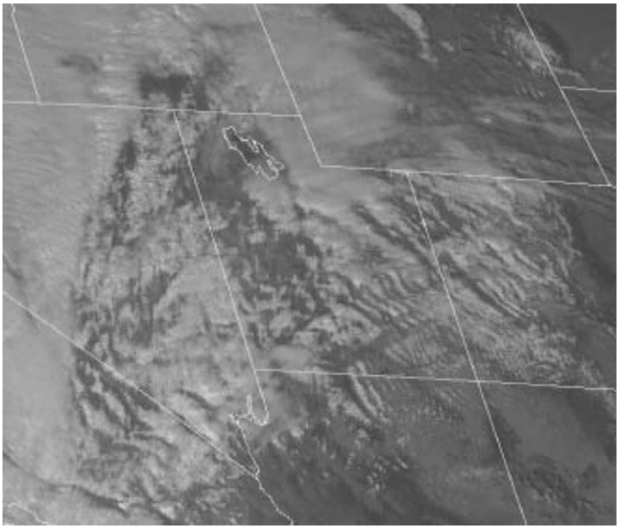


Figure 6 Mountain waves generated by southwesterly flow traversing the mountain ranges of the southwestern United States. This visible satellite image was taken at 2200 Universal Time Coordinated (UTC), February 9, 1999.

When the cross-barrier flow of stable air is sufficiently strong, damaging downslope wind storms in the lee of major mountain barriers are common during the winter season at many locales around the globe (Whiteman, 2000). Local residents refer to these wind storms by many different names (e.g., foehn in the European Alps, bora near the Adriatic Sea, chinook to the east of the Rockies, Santa Ana in southern California, canyon winds in Utah, zonda in Argentina, and oroshi in Japan). While local topography modulates the intensity of these storms, damaging downslope wind storms share one or more of the following common characteristics: pronounced mountain waves, precipitation on the upwind side of the mountain range coupled with adiabatic warming on the downwind side, strong cross-barrier pressure gradient (high pressure upstream, low pressure downstream), inversion (temperature increasing with height) in a layer above ridge crest, or wind reversal above the crest. The wind and temperature structure across the Rocky Mountains during a particularly damaging event is shown in Figures 5 and 6 in Chapter 27. Durran (1990) summarizes how the change in flow characteristics across the barrier (Froude number less than 1 upstream and greater than 1 downstream) may contribute to further acceleration of the wind in the lee of the barrier in a manner analogous to a hydraulic jump.

Barrier jets form when stable, low-level flow impinges upon a mountain range. As the flow piles up in front of the barrier, it is deflected to the left in the Northern Hemisphere as a result of a leftward-directed component of the pressure gradient force. Northward-directed barrier jets have been observed along the Pacific coast when the prevailing wind was from the west (Overland and Bond, 1995). Southward-deflected barrier jets have been observed on the east slopes of the Rockies and Appalachians during cold-air damming episodes (Dunn, 1987; Bell and Bosart, 1988).

3 INFLUENCE OF TERRAIN UPON PRECIPITATION

Orographic Precipitation

Local variations in precipitation as a result of the height, relief, and aspect (i.e., slope direction) of local terrain features can be striking: For example, annual precipitation increases over a distance of 35 km from 16.2 in. (41 cm) at Salt Lake City, Utah, to 58.5 in. (149 cm) at Alta in the nearby Wasatch Mountains. One of the most dramatic mesoscale variations in precipitation in the United States is evident in northwestern Washington where precipitation on the windward side of the Olympic Mountains is over 120 in. (305 cm) yet drops to less than 30 in. (76 cm) in the lee of those mountains (see Fig. 1).

Hills or mountains deflect air near the surface upward or downward, depending on the direction of the air flow relative to the slope of the topography. Banta (1990) notes that mountains have two major roles in forming clouds and precipitation: first, as obstacles to the flow and, second, as high-level heat sources during the day that cause the winds to converge toward the mountain. Clouds are likely to be generated if the ascent caused by either of these mechanisms is strong and there is sufficient moisture present in the air stream. If the ascent carries the cloud water and ice high enough such that the air becomes supersaturated, then the excess water and ice in the cloud may begin to fall and eventually may be deposited at the surface as precipitation.

Figure 7 summarizes the physical processes that control the development of orographic precipitation (Houze, 1993):

- *Upslope Condensation* Stable ascent of saturated air is forced by flow over mountains.
- *Orographic Convection* Lifting induced by terrain leads to convective release of instabilities present in the flow (e.g., Fig. 8). Orographic convection can be further subdivided into the following:
  - *Upslope and Upstream Triggering* Topographically induced motions and upstream blocking trigger convection leading to precipitation on the windward slope of the barrier.
  - *Thermal Triggering* Daytime heating produces an elevated heat source with local convergence near the top of the hill or mountain.
  - *Lee-Side Triggering* Low-Froude-number flow around a hill or isolated mountain leads to convergence in the lee of the obstacle.
  - *Lee-Side Enhancement of Deep Convection* Flow across a mountain range converges with low-level thermally induced upslope flow.

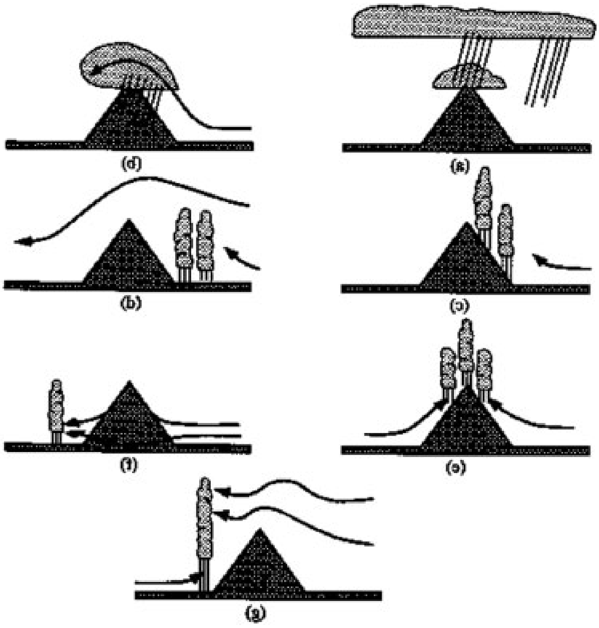
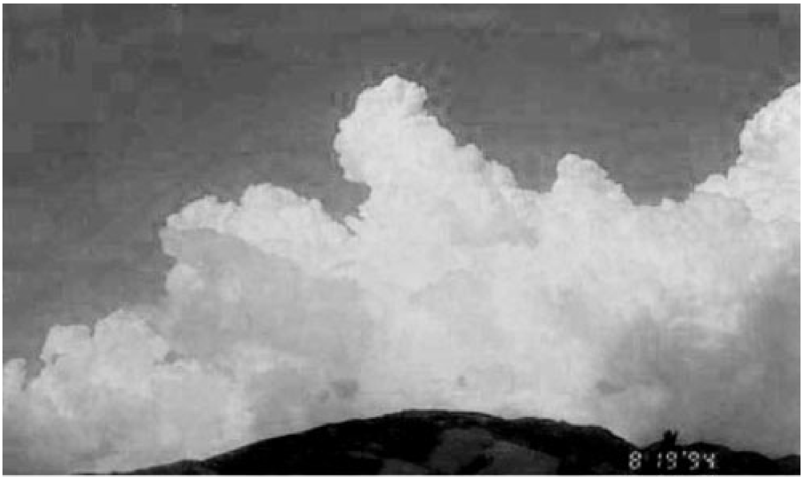


Figure 7 Mechanisms of orographic precipitation (adapted from Houze, 1993); (*a*) seeder-feeder; (*b*) upslope condensation; (*c*) upslope triggering; (*d*) upstream triggering; (*e*) thermal triggering; (*f*) lee-side triggering; (*g*) lee-side enhancement.



8 19'94

Figure 8 Convection developing over the Wasatch Mountains in the afternoon (photo by J. Horel). See ftp site for color image.

- *Seeder-feeder* Convective cells aloft produce cloud water or ice that fall into lower cloud decks; the falling cloud particles grow at the expense of the water content of the lower clouds.

Lake Effect Snow

During the cool season, enhanced precipitation is often observed in the lee of major water bodies such as the Great Lakes (Niziol et al., 1995). As shown in Figure 11 of Chapter 27, the greatest snowfall near the Great Lakes occurs where the prevailing winds blow across the longest fetch of a lake and is enhanced by local orography such as the Tug Hill plateau downwind of Lake Ontario (Niziol et al., 1995). A crippling snowstorm for the Buffalo, New York, area and many other locales downwind of Lakes Erie and Ontario occurred during November 20–23, 2000. Local total snowfall amounts were as large as 31 in. (79 cm) downwind of Lake Erie and 28 in. (71 cm) downwind of Lake Ontrario. Chapter 27 reviews the synoptic-scale and mesoscale weather associated with lake effect snowstorms in the Great Lakes area.

Residents far from the shores of the Great Lakes contend occasionally with narrow lake effect snow bands that extend a considerable distance inland (Niziol et al., 1995). However, the greatest impact of lake effect snowstorms tends to be closer to the shores of the Great Lakes. Ingredients that determine the intensity and structure of lake effect snowstorms include the synoptic setting, the upstream fetch (distance the air travels over water) and degree to which the water body is covered by snow or ice, instability in the boundary layer, wind shear, moisture availability, differences in surface roughness between the water and land surfaces, offshore-directed land breezes, and presence of upstream lakes or downstream orography.

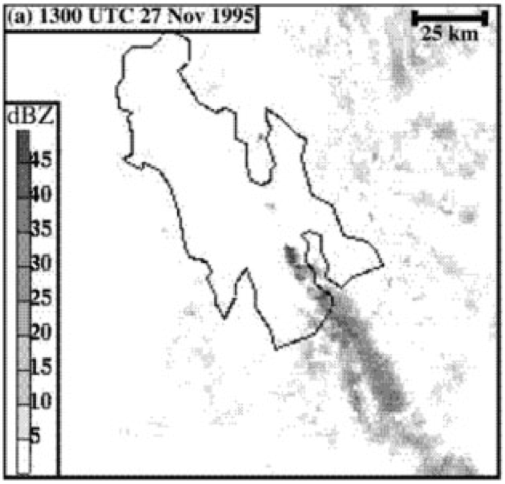


Figure 9 Example of lake effect snow band downwind of the Great Salt Lake, Utah. Lowest-elevation (0.5°) base-reflectivity analysis at 1300 UTC, November 27, 1995. Reflectivity shading based on scale at left. (From Steenburgh et al., 2000.)

Enhanced snowfall occurs as well downwind of smaller water bodies, such as the Great Salt Lake, that can have serious societal impacts (Steenburgh et al., 2000). For example, a wind-parallel band developed on November 27, 1995, over the Great Salt Lake and produced localized snow accumulations of 10 in. (25 cm) downstream (Fig. 9).

4 OTHER IMPACTS OF SURFACE INHOMOGENEITIES

Anthes (1984) and Zeng and Pielke (1995) demonstrate that variations in landscape characteristics can strongly affect the atmospheric boundary layer structure, formation of convective clouds, and fluxes of heat and moisture on the mesoscale. Mesoscale flows can be generated directly by differences in surface temperature and heat and moisture fluxes. In addition, biophysical processes that control moisture availability within the vegetative canopy can influence mesoscale circulations.

Pielke and Segal (1986) summarize the impact of horizontal contrasts in snow cover upon mesoscale circulations. Snow cover increases the albedo, reduces roughness, and alters surface fluxes of heat and moisture relative to nearby snow-free regions.

Urban areas affect the structure of the atmosphere and weather in a variety of ways (Dabberdt et al., 2000). Urban heat islands result from the combined effects of modified thermal and radiative properties of the surface and anthropogenic sources of sensible heat and moisture. The variability of surface roughness in urban areas affects the exchange of heat, mass, and momentum between the surface and the atmosphere. As a result of structures and pavement, hydrological processes are altered substantially.

Residents of urban areas are susceptible to the interactions of mesoscale and convective-scale weather phenomena, as illustrated in Figure 10, the F2 tornado that traversed downtown Salt Lake City on August 11, 1999 (Dunn and Vasiloff, 2001). It has been speculated (e.g., Landsberg, 1981; Bornstein and Lin, 1999) that large urban areas may affect the origin, strength, and movement of convective storms and other weather systems.

5 PREDICTABILITY OF TERRAIN-FORCED MESOSCALE CIRCULATIONS

As mentioned earlier, Pielke (1984) divided mesoscale atmospheric systems into two groups: terrain-induced mesoscale systems and synoptically induced mesoscale systems. He stated that the former are easier to simulate because they are forced by geographically fixed features in the underlying terrain. Paegle et al. (1990) also suggest that terrain-forced circulations may be inherently more predictable than synoptically induced flows, which are more sensitive to errors in the data used to specify the initial conditions of a numerical simulation. However, Paegle et al. (1990) note that the modeling of terrain-induced flows is difficult and susceptible to inaccurate numerical treatment of physical processes such as turbulent mixing,

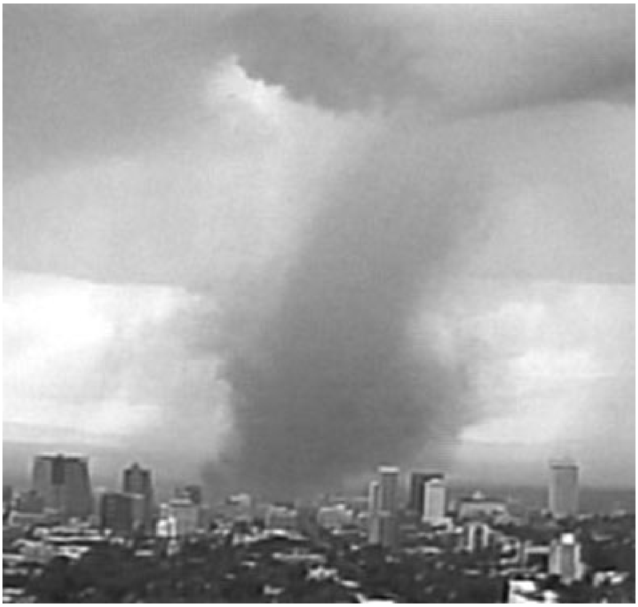


Figure 10 Category F2 tornado of August 11, 1999, in downtown Salt Lake City, UT (photo courtesy of Department of Meteorology, University of Utah, Salt Lake City, UT).

radiative heating, cloud processes, and soil transfer and the numerical errors arising from steep terrain slopes.

From ALPEX (Kuettner, 1981) to CASES (LeMone et al., 2000) and MAP (Bougeault et al., 2001), field programs have been critical to improve the numerical simulation and prediction of mesoscale circulations forced by variations in the underlying surface. Comprehensive datasets are required to provide validation of the model simulations and lead to improved treatment of relevant physical processes in high-resolution numerical weather prediction models.

Considerable debate remains in the mesoscale modeling community regarding the inherent predictability of terrain-induced circulations. As noted by Mass and Kuo (1998), mesoscale predictability in regions of complex terrain may be enhanced due to the relatively deterministic interactions between the synoptic-scale flow and the underlying terrain. Nonetheless, while models are increasingly capable of simulating physically realistic responses to flow over terrain, a corresponding increase in forecast skill has not always been evident (e.g., Colle et al., 2000).

REFERENCES

Anthes, R. A. (1984). Enhancement of convective precipitation by mesoscale variations in vegetative covering in semi-arid regions, *J. Appl. Meteor.* **23**, 541–554.

Banta, R. M. (1990). The role of mountain flows in making clouds, in W. Blumen (Ed.), *Atmospheric Processes over Complex Terrain, Meteor. Monogr.* **23**(45), 229–282.

Barry, R. G. (1992). *Mountain Weather and Climate*, 2nd ed., London, Routledge.

Bell, G. D., and L. F. Bosart (1988). Appalachian cold-air damming, *Mon. Wea. Rev.* **116**, 137–162.

Bornstein, R., and Q. Lin (1999). Urban heat islands and summertime convective thunderstorms in Atlanta, *Atmos. Environ.* **34**, 507–516.

Bougeault, P., P. Binder, A. Buzzi, R. Dirks, R. Houze, J. Kuettner, R. B. Smith, R. Steinacker, and H. Volker (2001). The MAP special observing period, *Bull. Am. Meteor. Soc.* **82**, 433–462.

Chen, W. Y. (1977). Analysis of vorticity and divergence fields and other meteorological parameters over Lake Ontario during IFYGL, *Mon. Wea. Rev.* **105**, 1298–1309.

Colle, B. A., C. F. Mass, and K. J. Westrick (2000). MM5 precipitation verification over the Pacific Northwest during the 1997–99 cool seasons, *Wea. Forecasting* **15**, 730–744.

Dabberdt, W. F., J. Hales, S. Zubrick, A. Crook, W. Krajewski, J. C. Doran, C. Mueller, C. King, R. N. Keener, R. Bornstein, D. Rodenhuis, P. Kocin, M. A. Rossetti, F. Sharrocks, and E. M. Stanley (2000). Forecasting issues in the urban zone: report of the 10th prospectus development team of the U.S. Weather Research Program, *Bull. Am. Meteor. Soc.* **81**, 2047–2064.

Dunn, L. (1987). Cold air damming by the Front Range of the Colorado Rockies and its relationship to locally heavy snows, *Wea. Forecasting* **2**, 177–189.

Dunn, L., and S. Vasiloff (2001). Tornadogenesis and operational considerations of the 11 August 1999 Salt Lake City tornado as seen from two different doppler radars, *Wea. Forecasting* **16**, 377–398.

Durran, D. R. (1990). Mountain waves and downslope winds, in W. Blumen (Ed.), *Atmospheric Processes over Complex Terrain, Meteor. Monogr.* **23**(45), 59–82.

Houze, R. A., Jr. (1993). *Cloud Dynamics*, San Diego, CA, Academic.

Kuettner, J. P. (1981). ALPEX: the GARP mountain subprogram, *Bull. Am. Meteor. Soc.* **62**, 793–805.

Landsberg, H. E. (1981). *Urban Climate*, New York, Academic.

LeMone, M., R. Grossman, R. Coulter, M. Wesley, G. Klazura, G. Poulos, W. Blumen, J. Lundquist, R. Cuenca, S. Kelly, E. Brandes, S. Oncley, R. McMillen, and B. Hicks (2000). Land-atmosphere interaction research, early results, and opportunities in the Walnut River Watershed in southeast Kansas: CASES and ABLE, *Bull. Am. Meteor. Soc.* **81**, 757–779.

Mass, C. F., and Y.-H. Kuo (1998). Regional real-time numerical weather prediction: Current status and future potential, *Bull. Am. Meteor. Soc.* **79**, 253–263.

Niziol, T. A., W. R. Snyder, and J. S. Waldstreicher (1995). Winter weather forecasting throughout the eastern United States. Part IV: Lake effect snow, *Wea. Forecasting* **10**, 61–77.

Overland, J. E., and N. A. Bond (1995). Observations and scale analysis of coastal wind jets, *Mon. Wea. Rev.* **123**, 2934–2941.

Paegle, J., R. Pielke, G. Dalu, W. Miller, J. Garratt, T. Vukicevic, G. Berri, and M. Nicolini (1990). Predictability of flows over complex terrain, in W. Blumen (Ed.), *Atmospheric Processes over Complex Terrain, Meteor. Monogr.* **23**(45), 285–299.

Pielke, R. A. (1984). *Mesoscale Meteorological Modeling*, San Diego, CA, Academic.

Pielke, R. A., and M. Segal (1986). Mesoscale circulations forced by differential terrain heating, in P. S. Ray (Ed.), *Mesoscale Meteorology and Forecasting*, Boston, American Meteorological Society, pp. 516–548.

Steenburgh, W. J., S. F. Halvorson, and D. J. Onton (2000). Climatology of lake-effect snowstorms of the Great Salt Lake, *Mon. Wea. Rev.* **128**, 709–727.

Whiteman, C. D. (1990). Observations of thermally developed wind systems in mountainous terrain, in W. Blumen (Ed.), *Atmospheric Processes over Complex Terrain, Meteor. Monogr.* **23**(45), 5–42.

Whiteman, C. D. (2000). *Mountain Meteorology: Fundamentals and Applications*, New York, Oxford University Press.

Zeng, X., and R. A. Pielke (1995). Landscape-induced atmospheric flow and its parameterization in large-scale numerical models, *J. Clim.* **8**, 1156–1177.

CHAPTER 29

SEVERE THUNDERSTORMS AND TORNADOES

H. BROOKS, C. DOSWELL III, D. DOWELL, R. HOLLE, B. JOHNS, D. JORGENSON, D. SCHULTZ, D. STENSRUD, S. WEISS, L. WICKER, AND D. ZARAS

1 INTRODUCTION

Severe thunderstorms and tornadoes are phenomena that can occur at almost any place on the planet. Unlike hurricanes and synoptic-scale cyclones, these local storms affect areas of 10 to 100 km^2 (e.g., the size of a typical U.S. city) and last a few minutes to several hours. Nevertheless, these storms can produce devastating damage that rivals any other atmospheric storm on earth. Tornadoes kill nearly three-dozen people in the United States per year, and recent tornadoes such as the May 3, 1999, Oklahoma City tornado can cause more than $1 billion in damage. However, the real killers from severe storms are actually flash floods and lightning, as the fatality rate from these events is more than 200 people every year. Therefore, timely forecasts and warnings of severe weather are crucial for mitigating damage and protecting the public.

2 CLIMATOLOGY OF SEVERE THUNDERSTORMS

Severe weather associated with thunderstorms affects almost all of the planet and represents a significant threat to life and property in many locations. The definition of what is considered "severe" depends on operational forecasting considerations that vary from country to country but typically includes phenomena

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

such as tornadoes, large hail (usually of diameter at least approximately 2 cm), strong convective wind gusts (usually approximately 90 km/h or more), and extremely heavy precipitation associated with flash floods (frequently 50 mm/h at a single location). Criteria associated with heavy precipitation are the most variable from country to country and, in some places, even within one country. For instance, in the United States, flash flooding is not considered a severe thunderstorm event, and in Canada the objective definition of heavy precipitation is different in different geographical regions.

Tornadoes

Tornadoes have been observed on every continent except Antarctica, although they are most common in North America, particularly the Great Plains of the United States. Increased efforts to collect information about tornadoes in North America have led to an increase in the number of reports, with an average of about 1200 tornadoes reported annually in the United States in recent years, compared to only 600 just 50 years ago. The increase has been particularly apparent in the number of weak tornadoes (classified F0 or F1 on the Fujita damage scale that goes from F0 to F5). Similar efforts in other countries have also led to large increases in the reported number of tornadoes there, such as in Germany where the average prior to 1950 was about 2 per decade, but in the 1990s was 7 per decade, with more than 20 reported in the year 2000 alone. Climatologies of tornado occurrence in the United States have identified the temporal and spatial structure of the threat. The strongest tornadoes (F2 to F5 on the Fujita scale) are most often found in the Great Plains of the United States (Fig. 1). This is a result of the frequent production of favorable environments with warm, moist air near the ground, dry, relatively cool air aloft, and strong vertical speed and directional shear of the horizontal winds. The Gulf of Mexico acts as a source region for warm, moist low-level air flowing north, and the Rocky Mountains act as a source region for the dry, relatively cool air aloft flowing eastward toward the Plains. The presence of these two fixed geographic features appears to be the dominant reason for the frequency of tornadoes in the Plains.

The rarity of tornadoes in other regions of the world does not mean that, when events occur there, the effects are small. Landfalling tropical cyclones often produce tornadoes. Historically, devastating tornadoes have struck Europe approximately once every 20 years. Since 1984, individual tornadoes with hundreds of fatalities have occurred in Russia, northeast of Moscow, and in Bangladesh. While it seems that strong and violent tornadoes are much less common in other parts of the world compared to the United States, it also appears likely that tornadoes are vastly underreported in the rest of the world. The prime evidence for this is that the majority of reported tornadoes in many parts of the world are fatality-producing events or are especially newsworthy (such as the 1998 tornado in Umtata, South Africa, while the South African president was visiting the town). This situation is similar to that in the United States in the middle part of nineteenth century, when only approximately 25 tornadoes per year were reported. Recent studies have indicated that probability of a

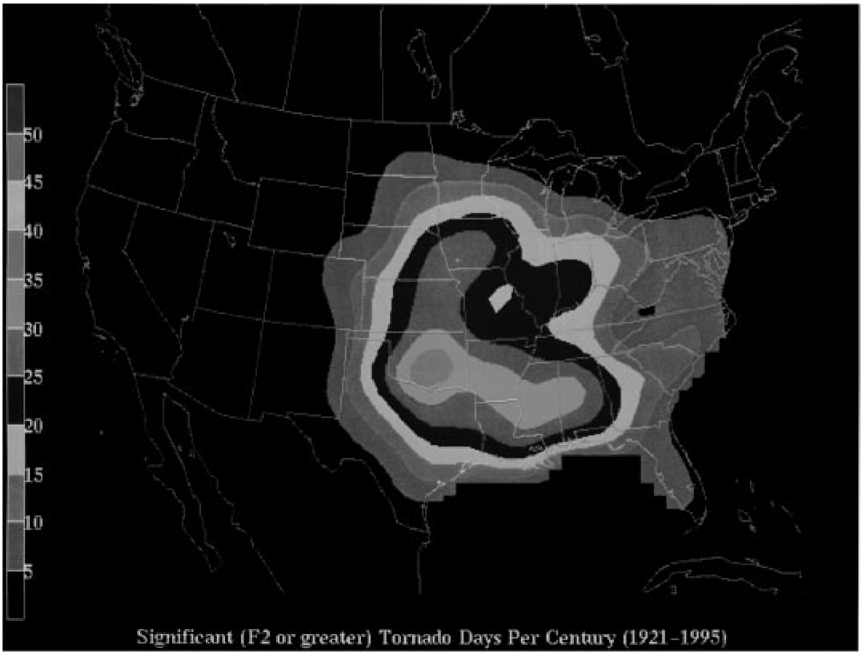

50
45
40
35
30
25
20
15
10
5
Significant (F2 or greater) Tornado Days Per Century (1921–1995)

Figure 1 (see color insert) Number of days per century an F2 or greater strength tornado might occur within 25 miles of a point. For example, in southcentral Oklahoma one would expect an F2 tornado within 25 miles about once every 3 years. Calculations based on data from 1921 to 1995. See ftp site for color image.

particular reported tornado being strong or violent is approximately the same over most of the world (Fig. 2).

Hail

The definition of exactly what is severe hail is troublesome. For some agricultural interests during some times of the year, even 1 cm in diameter hail may be devastating. For urban areas, it may take much larger hail, say 4 cm in diameter, to cause problems. The distribution of regions prone to these levels of threat is very different. The smaller limit occurs in much of the temperate world during the warm season. Larger hail is typically limited to the central part of North America and regions near major mountain ranges in the rest of the world (e.g., the Himalayas and Alps). It has been suggested that extremely large hail is much more likely in supercell thunderstorms than in "ordinary" thunderstorms. This is consistent with the observed distribution of tornadoes, presumably associated with supercells, in the central part of the United States. The lack of a relationship when hail of any size is considered has been pointed out for China by showing that the frequency of hail is maximized in the high plateau regions of western China while tornadoes are more common in the eastern part of the country, particularly the Yangtze River valley.

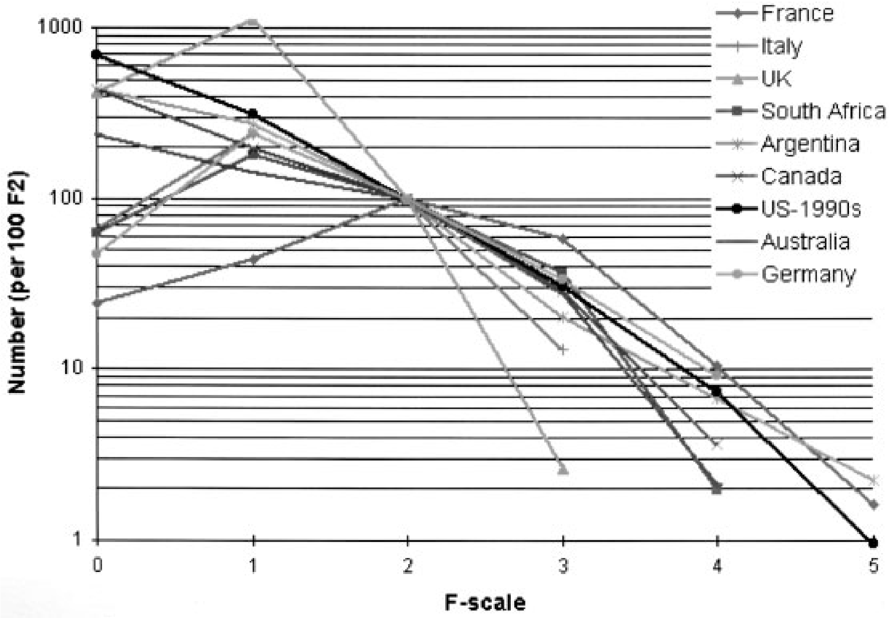


Figure 2 Distribution of F-scale rating for tornadoes in various countries is roughly the same, particularly for the violent (F4 to F5) tornado. See ftp site for color image.

The high plateaus and other regions downwind of mountains may produce large hail because of the steep tropospheric lapse rates that develop as air comes over mountains.

Unfortunately, observations of hail have not been consistent through the years. In the United States, the number of reports of severe hail (approximately 2 cm or larger) have increased by an order of magnitude in the past 30 years. Most of the increase has been in the smaller end of the severe range. As a result, attempts to develop a climatology based on the reports face significant challenges. Researchers are faced with the dilemma of having small sample sizes or an inhomogeneous record. Efforts to use insurance losses are complicated by the issue of what causes the losses (agricultural versus urban interests) and the temporal inhomogeneity of the insured base. Nevertheless, extremely large losses (greater than $500 million) have been associated with hailstorms in areas such as Munich (Germany), Denver, Colorado, and the Dallas–Fort Worth, Texas, region in the last 15 years.

Damaging Convective Wind Gusts

Strong winds associated with thunderstorms are a common feature. Little has been done to document their climatological occurrence until recently, although they are almost certainly the most common severe weather event. Damaging straight-line thunderstorm wind gusts are usually associated with cold air outflow as the downdraft of the storm reaches the ground, producing what has been termed a *downburst.* Factors that influence the generation of damaging wind gusts at the surface include: negative buoyancy enhanced by evaporative cooling within unsaturated air, precipitation loading within the downdraft, and downward transfer of horizontal momentum by the downdraft. Again, aspects of these processes are dependent on storm-scale microphysics including drop size distribution and liquid water content per unit volume, neither of which can be determined from standard observing systems.

Strong winds can occur in a variety of situations. They can be associated with small, short-lived downdrafts and even when they are relatively weak (say less than 25 m/s), they can be a significant hazard to aviation. Considerable effort has been expended in the last 20 years to decrease commercial aircraft accidents due to thunderstorm downdrafts. Radar detection and education of the aviation industry about the threats seems to have limited the number of accidents in the last decade, after several occurred from the early 1970s through the mid-1980s.

Larger areas can be affected by high winds when organized systems of thunderstorms occur. In the United States, widespread convective wind events are sometimes referred to as *derechos.* They occur in association with mesoscale convective systems, which are composed of a number of individual thunderstorms. Often, they are arranged as a squall line, producing a wide area of high winds with new convective cells initiated on the leading edge of the outflow from earlier cells. The system may maintain itself for many hours, provided sufficient low-level moisture and midlevel unstable air can be found as the system moves along.

Flash Floods

Flash floods are the most widespread severe local storm phenomena associated with large loss of life. They occur all over the world, especially in regions of complex terrain. They are the most difficult to forecast, in part because they involve both meteorological and hydrological aspects. Determining their effects is complicated further by interactions with people and buildings. If a flash flood occurs in a location where it does not impact life or property, it is unlikely to be reported. On the other hand, relatively minor precipitation events may produce significant flooding if antecedent conditions exacerbate the flooding as occurred in the Shadyside, Ohio, flood of 1990 with saturated soils or in the Buffalo Creek, Colorado, flood of 1996 when a forest fire cleared vegetation from the area a couple of months before the rain event.

Great loss of life has been associated with flash flooding, even in developed nations. Recently, a campground in Biescas in the Spanish Pyrenees was flooded with more than 80 deaths. In 1998, 11 hikers were killed by a flash flood in a "slot canyon" in northern Arizona when rain-generated runoff from a storm tens of kilometers away was funneled into the canyon. (This storm, by the way, was accurately located by NWS radar in southern Utah, and its potential impacts relayed to Park Service personnel. Unfortunately those who died chose to ignore the warning.) The three biggest convective-weather death toll single events in the United States (with the exception of aircraft crashes) since 1970 have all been flash floods. Death tolls in developing countries are frequently difficult to estimate.

Flash floods are distinguished from main-stem river floods by the extremely rapid rate of rise of water levels. While main-stem river floods may have water stages rising by tens of centimeters per day, flash floods are associated with water stages rising by tens of centimeters per hour or, in extreme cases, per minute. Small streams may carry 100 times their normal capacity and, often, it is very small basins that produce flash floods. (In operational practice, even in the United States, this can cause problems, since these small basins may not be mapped as well as larger basins, particularly for comparison to radar estimates of precipitation. As a result, forecasters may be unaware of the nature of the threat even if accurate estimates of rainfall are available.)

The threat from flash floods has increased in some regions because of the increased use of mountainous regions for recreation. Excellent examples include the Big Thompson (Colorado) and Biescas (Spain) floods. Public response to flash flood forecasts is frequently poorer than for other weather hazards forecasts, such as for tornadoes. Most people have experienced heavy rain events and fail to realize until it is too late that the particular event underway is more dangerous. Further, heavy rain often washes out roads and bridges, which makes escape and rescue difficult.

Lightning

The frequency and location of cloud-to-ground lightning in the United States are now known quite well. The deployment and operation during the last decade of automatic real-time lightning detection sensors have made this possible. An average of about 25 million cloud-to-ground flashes are detected by the National Lightning Detection Network (NLDN) in the United States every year.

The map of network-detected flashes in Figure 3 shows that much of peninsular Florida has the greatest frequency of flashes per area over a year. Flash density decreases to the north and west from there. Important local variations occur along the coast of the Gulf of Mexico, where sea breezes and urban areas enhance lightning frequency. Other maxima and minima are located in and around the western United States where there are mountains and large slopes in terrain.

Lightning is most common in summer (Fig. 4*a*); about two-thirds of the flashes occur in June, July, and August. In the southeastern states, lightning is more common throughout the year, since there often is a significant amount of moisture in the lower and middle levels of the atmosphere. Air needs to be lifted strongly to form lightning; coastlines and large mountains provide persistent updrafts that result in lightning-producing thunderstorms.

During the course of the day, lightning is most common in the afternoon (Fig. 4*b*). Nearly half of all lightning occurs from 1500 to 1800 Local Standard Time (LST). Flashes are most common in the afternoon because the updrafts needed for thunderstorm formation are strongest during the hours of the day when surface temperatures typically are the highest, which results in the greatest vertical instability.

The primary information for lightning deaths and injuries in the United States is the National Oceanic and Atmospheric Administration (NOAA) publication *Storm*

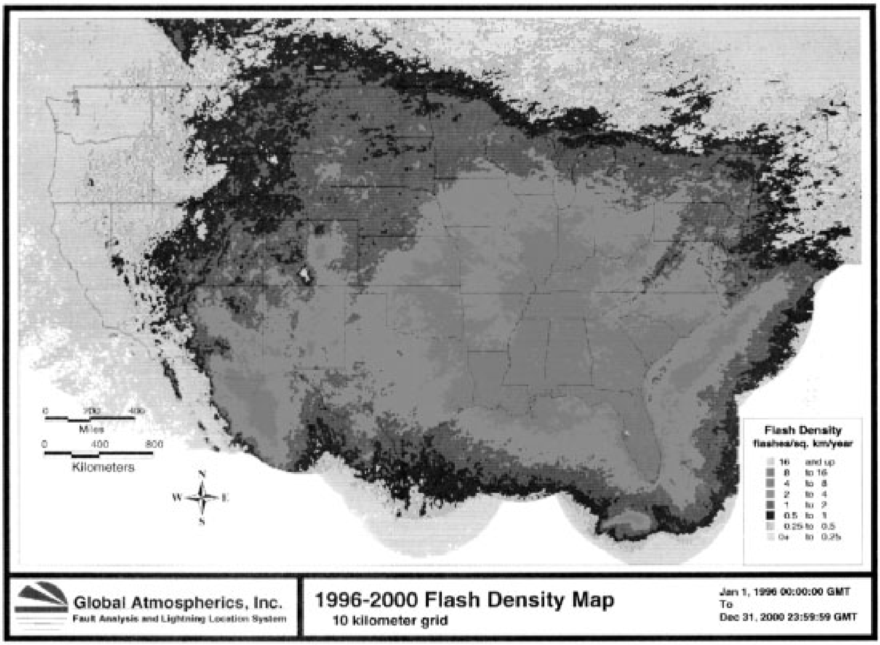


Figure 3 (see color insert) Number of cloud-to-ground lightning flashes per year for the United States based on data from 1996 to 2000. (Courtesy Global Atmospherics, Inc.) See ftp site for color image.

Data. Since most lightning casualties occur to one person at a time, and are more dispersed in time and space than other severe weather phenomena, lightning casualties are underreported.

The spatial distribution of deaths and injuries in Figure 5 shows the largest absolute number to be in Florida, whose total is twice that of any other state. Most of the other states with large numbers of casualties are among the most populous in the country. However, when population is taken into account, the highest rates of lightning casualties per million people are found in the Rocky Mountain states, Florida, and other states in the southeast. The time distributions of lightning deaths and injuries by month and day in Figure 4 show a close resemblance to the distribution of the actual lightning flashes in the same figure.

Over the last 100 years, the rate of lightning deaths has dropped significantly. The decrease parallels a major shift from rural to urban settings for much of the U.S. population and is apparent in similar records in Europe and Australia. The most common activity of lightning casualty victims has also changed during this period from agricultural to recreational.

Around the world, data from lightning sensors on satellites show that lightning occurs most often over land in the tropical and subtropical areas of Africa, South America, and Southeast Asia. The annual rates of lightning per area often exceed those in Figure 3 for Florida.

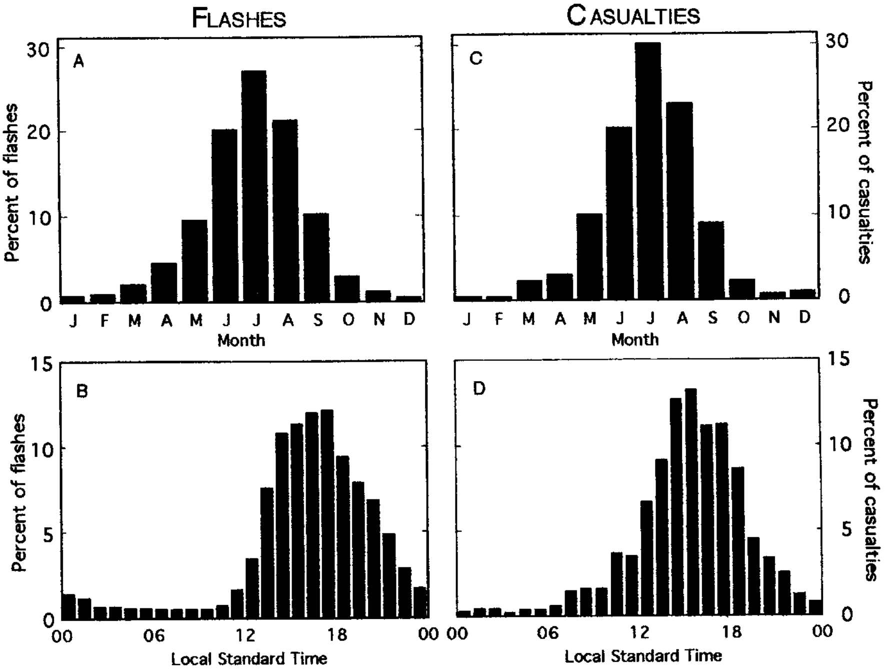


Figure 4 Lightning histograms showing flash rates by (*a*) month of year and (*b*) time of day. (Courtesy of National Severe Storms Laboratory.)

About 100 years ago, the annual U.S. death rate from lightning was 6 per million people, which is an order of magnitude higher than now. The United States then was a more agriculturally oriented society, living and working in ungrounded and less substantial buildings than are now common. This rate may continue to be appropriate for populous tropical and subtropical areas where lightning is frequent.

There are now about 100 lightning deaths per year in the United States, but this total could be around 1000 if the population was still rural, practiced labor-intensive agriculture, and lived in less substantial dwellings. So, the worldwide death total could be expected to be at least 10,000 per year, since many people continue to live in such situations. A ratio of 10 injuries per death gives a global total of 100,000 injuries a year from lightning.

3 CLOUD PROCESSES AND STORM MORPHOLOGY

Ingredients of Deep Moist Convection

Deep moist convection requires three ingredients to occur: instability, lift, and moisture. Interesting weather can occur in the absence of these ingredients, but it will not be deep moist convection. For example, in the absence of instability, forced

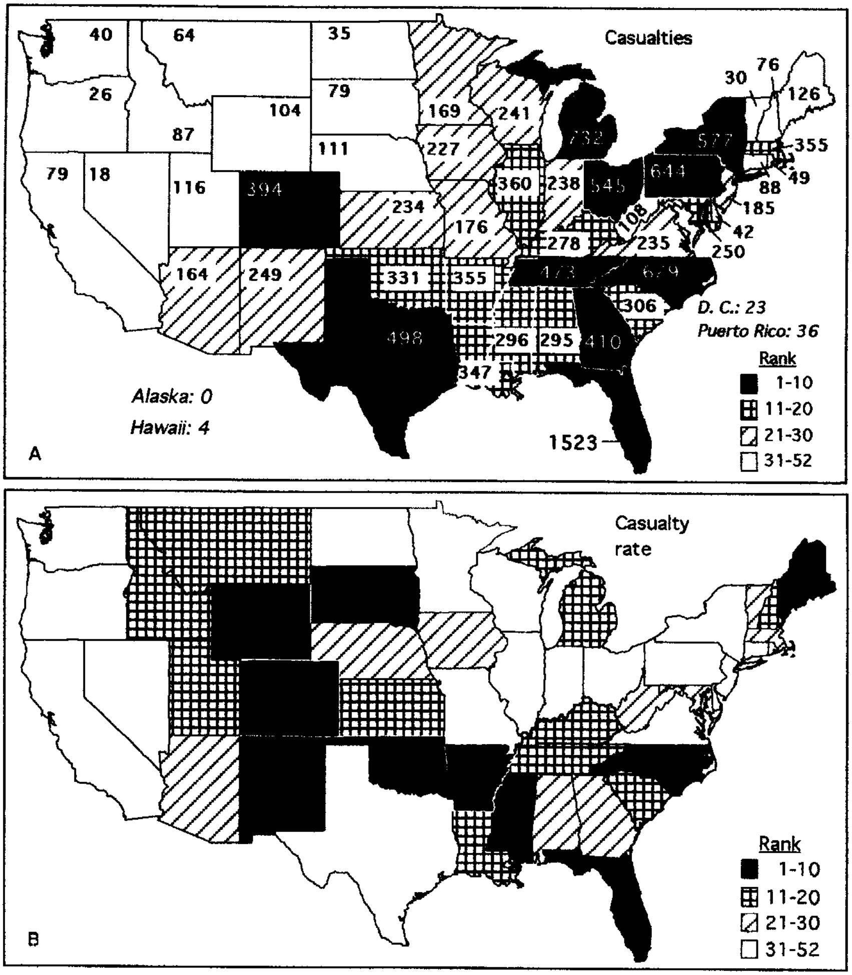


Figure 5 Spatial distribution of deaths and injuries by U.S. state. (Courtesy of National Severe Storms Laboratory.)

ascent of moist air over topography or over a frontal surface can produce heavy precipitation; but, without instability, it is not generally considered convection. Sometimes the term forced convection is applied to this situation. In this section, however, we will focus mainly on free convection, where all three ingredients are present.

Before discussing the ingredients of convection individually, it is necessary to introduce several important concepts. First is the concept of hydrostatic stability. In a hydrostatically stable atmosphere the downward gravitational force associated with the weight of the air is exactly balanced by the upward vertical pressure gradient force (the pressure near the surface of Earth is greater than the pressure aloft, resulting in a pressure gradient force that acts upward). Generally, the atmosphere is very close to hydrostatic stability at most times and locations. Another important concept is that of the lapse rate, a measure of how the temperature in a column of air changes with height. In the troposphere, the temperature almost always cools with increasing height, so a positive lapse rate indicates decreasing temperature with respect to height. As will be explained below, large lapse rates indicate rapidly decreasing temperatures with height, and this is a favorable environment for the development of convection.

Buoyant stability is a measure of the atmosphere's resistance to vertical motions. The primary way meteorologists measure stability is through parcel theory, a pedagogic tool in which idealized bubbles of air called air parcels are employed. An air parcel is a small bubble of air that does not exchange heat, moisture, or mass with its environment, i.e., the thermodynamic process is adiabatic. As air parcels sink in the atmosphere, the increasing pressure they encounter causes them to compress and warm. In fact, this rate of warming is a constant 9.8°C/km and is termed the dry adiabatic lapse rate. Likewise, when air parcels rise in the atmosphere, the decreasing pressure allows them to expand and cool at the dry adiabatic lapse rate. The dry adiabatic cooling rate occurs as long as the parcel's relative humidity is less than 100%. Saturated parcels (i.e., the parcel relative humidity is 100%) cool less as they rise because cooling causes condensation of water vapor. A process that releases latent heat and lessens the rate of cooling with height to a value ranging from 4 to 7°C/km, depending on the temperature and pressure of the parcel. The cooling rate for a saturated parcel is termed the moist adiabatic lapse rate.

Parcel buoyancy is defined by comparing the parcel's temperature to the surrounding environment's temperature. If an air parcel is warmer than its environment, it is said to be positively buoyant and the parcel will accelerate upward. If an air parcel is colder than its environment, it is said to be negatively buoyant and the parcel will accelerate downward. If an air parcel is the same temperature as its environment, it is said to be neutrally buoyant and there is no net force on the parcel. If the parcel buoyancy is large, the accelerations are significant and cause the atmosphere to deviate significantly from hydrostatic balance. These nonhydrostatic forces are often large and an important mechanism for the development and maintenance of severe thunderstorms.

Consider a moist, but unsaturated, air parcel near the surface of Earth. If the actual lapse rate of the atmosphere is between the moist and dry adiabatic lapse rates, then a rising air parcel will cool at the dry adiabatic lapse rate, a rate larger than the environmental lapse rate. Thus, this parcel will be negatively buoyant and will need to be forcibly lifted to continue to rise. As the air parcel cools and expands, it may eventually reach 100% relative humidity, or saturation. The height of this point is called the lifting condensation level (LCL). Further forced lifting will result in the air parcel cooling at the moist adiabatic lapse rate. Eventually, the temperature of the rising air parcel may become warmer than the environmental air at the same height, becoming positively buoyant. The height of this point is called the level of free convection (LFC). Typically, the positive buoyancy continues with height until the air parcel rises near the tropopause, where the stability becomes larger, to its equilibrium level, the point of neutral buoyancy. The integrated positive buoyancy from the LFC to the equilibrium level can be computed and is termed the convective available potential energy (CAPE), which is related to the maximum updraft possible under parcel theory. The integrated negative buoyancy needing to be overcome during forced parcel lifting near the surface is called the convective inhibition (CIN), and the layer over which the CIN occurs is called the cap or lid. When the environmental lapse rate lies between the moist and dry adiabatic lapse rates, conditional instability is said to exist. Conditional instability indicates that the atmosphere will form convection if enough air parcels are lifted to the LFC by some mechanism.

CAPE is an important measure of the integrated instability in the atmosphere. One way to generate a high CAPE environment is to get dry air with large lapse rates on top of warm, moist air. During the spring in the central United States, this so-called loaded-gun sounding frequently occurs when low-level warm moist air from the Gulf of Mexico is overrun by dry air at midlevels from the Mexican Plateau or the Rocky Mountains. This stratification can produce high-CAPE soundings, loaded for a classic Central Plains outbreak of severe weather. A certain degree of CIN is also required so that convection does not break out spontaneously and so that the CAPE can build to large values as the surface temperatures warm during the day.

The second ingredient for convection to occur is that sufficient lift for the air to reach its level of free convection is required. Often this arises due to surface airflow boundaries where convergence occurs, forcing low-level ascent. These surface boundaries can be fronts, drylines, sea-breeze convergence lines, horizontal convective rolls in the boundary layer, or even outflow boundaries from previous convection. These thunderstorms may occur due to buoyant thermals in the warm boundary layer air near the surface of Earth. Sufficiently strong thermals may have enough penetration out of the boundary layer that their LFC may be reached, thus initiating convection.

The final ingredient for convection is an adequate supply of moisture. For most environments, such as the loaded-gun sounding, an increase in low-level moisture results in an increase in CAPE. The dry air aloft is also important since it can quickly evaporate the warm moist thermals bubbling up from the boundary layer. Evidence suggests that it may take several generations of ever-deepening towering cumulus congestus to sufficiently moisten the lower part of the dry layer in the loaded-gun sounding. Without sustained vertical motion and moisture, deep moist convection may never develop, even from a well-primed loaded-gun sounding.

Although not essential for convection to occur, a factor that is important for controlling the type of convective system that may develop is the vertical change in the horizontal wind direction and speed. This is called vertical wind shear. (Readers may be familiar with the term wind shear as it relates to aircraft accidents. In that case, wind shear refers to the horizontal change in the horizontal wind speed and direction. It is important to distinguish between the two different types of wind shear.)

The vertical wind shear is important to deep moist convection for several reasons. Too much wind shear tears apart cloud elements, not allowing updrafts to develop deeply and nearly vertical. Too little wind shear results in the downdrafts suppressing the updrafts, shutting off the flow of warm moist air to the storm. The proper amount of wind shear is important for separating the updrafts from the downdrafts of a thunderstorm, thus permitting the storm to be long lived. More specifically, convection produces cold downdraft air that pools underneath the storm. A gust front, marking the boundary between the cool downdraft air and the warm environmental air, can lift the warm air, sparking new convective storms. In the absence of shear, the downdraft air will spread uniformly in all directions. New cells formed along the gust front will be quickly undercut by the expanding gust front from the parent cell, thus limiting further convective development. As the shear increases, new convective cells will move downshear away from the parent cell faster than the evolving gust front, allowing continual redevelopment of new cells. This effect is particularly important in long-lived multicells and squall lines.

Types of Convective Storms

Ultimately, the role of convection in the atmosphere is to take an unstably stratified sounding and make it more stable by lifting the warm, moist low-level air and bringing cooler, drier air down in the downdrafts. This removes the conditional instability present in the pre-convective atmosphere. It is the interplay between the release of this instability within the storm itself and the environment that produces the panoply of storm types that we see.

Although many of the factors that affect storm structure and evolution are understood, there is still much to be discovered about the relative roles of the large-scale environment and the internal dynamics of the storm itself. This is an important issue because it relates to the limits of predictability of convective storms. If the environment of the storm is a strong factor in storm evolution, then storms are potentially more predictable since the large-scale data in the storm environment is usually well observed. But if the internal dynamics of the storm are the most important factor, then it may be a long time, if ever, until we have measurements within the storm that could be useful for understanding, let alone predicting, the storm evolution. For the purposes of this section, we consider the effect of the environment on the type of convective storm.

At least three environmental factors affect the type of convective storm that forms: wind shear, instability, and synoptic setting (e.g., fronts, drylines, jet streams). The type and direction of wind shear that the convection initiates is important for the morphology (or mode) of convection that results. The amount of instability affects the strength of the convective updrafts. The flow pattern in which the storms develop may also play a role in the mode and strength of the resulting convection. In the following section, the types of convective storms and their characteristics are summarized. More detailed discussion of the individual storm types can be found in later sections of this chapter.

The individual cumulus clouds that constitute so-called *pulse* or *air-mass* thunderstorms are typically a few kilometers in diameter with updrafts on the order of 10 m/s or more. CAPE is usually less than 1000 J/kg and the deep-layer wind shear is weak (less than 10 m/s over 10 km). These storms typically last 30 to 50 min and produce short-lived localized showers with few, if any, reports of severe weather (tornadoes, hail, or damaging winds). Typically, large areas tend to erupt in convective cloud about the same time when the surface temperature reaches the convective temperature, the temperature required to eliminate any low-level CIN. While useful for these situations, the use of the convective temperature for other types of storms is not recommended.

With moderately strong low-level and deep-layer wind shear and moderate to high CAPE, *supercells* may form. The essence of a supercell thunderstorm is a single, nearly steady, rotating updraft. High winds, flooding rains, large hail, and potentially long-lived violent tornadoes can occur with supercells. If the wind shear is constant in direction with height (called a straight-line hodograph), maturing supercells will tend to split into left-moving and right-moving cells. If the wind shear curves counterclockwise with height (in the Northern Hemisphere), right-moving cells will tend to dominate with counterclockwise rotation. On the other hand, if the wind shear curves clockwise with height, left-moving cells will tend to dominate with clockwise rotation in the Northern Hemisphere. Right-moving cells tend to be more common. More will be said about supercells later.

Sometimes, individual convective clouds will join together as gust fronts caused by downdrafts from the parent cells combine to lift unstable environmental air to its LFC, forming secondary convection along the periphery of the parent cells. These storms are called *multicells* and range in organization from poorly organized clusters of individual convective elements to highly organized linear structures including bow echoes and larger-scale squall lines. Individual storms within a multicell usually move to the left (in the Northern Hemisphere) of the mean motion of the multicell itself. Because the warm moist air tends to be ingested on the right side (or equatorward side, in the Northern Hemisphere) of the multicell, new cells form preferentially on this side, forcing older cells toward the left of the multicell storm. Because multicells propagate away from the original convection, a certain amount of low-level shear is needed for gust-front propagation. If storm motion is very slow, local areas may be affected by prolonged periods of rain, making flooding a strong possibility.

Sometimes multicellular convection will organize in lines. Typically this occurs because linear features exist at the surface (such as a cold front) that organizes the convection. Because of the linear forcing, a common location for development of *squall lines* and bow echoes is along fronts. Sometimes these lines move ahead of the front that spawned them; at other times, forcing above the surface will produce squall lines in the warm sectors of extratropical cyclones. In some cases, often when the vertical wind shear is weak, mesoscale convective complexes form.

4 MESOSCALE CONVECTIVE SYSTEMS, MESOSCALE CONVECTIVE COMPLEXES, SQUALL LINES, AND BOW ECHOES

Mesoscale Convective Systems

While isolated thunderstorms are important producers of severe weather, it is just as common for thunderstorms to merge and interact, forming a more complex precipitation system. These more complex thunderstorm systems are called mesoscale convective systems (MCSs)—cloud and precipitation systems that are comprised of a group of thunderstorms that have a contiguous precipitation area of ~ 100 km in at least one direction (Fig. 6). MCSs are particularly important because they are observed worldwide, over both the Tropics and midlatitudes, and over land and water. MCSs also have been shown to produce approximately half of the warm season rainfall in the central United States, indicating that they are important components of the hydrologic cycle. Further, in the form of squall lines and bow echoes, such systems likely account for most, if not all, of the widespread convective windstorms that occur around the world.

Mesoscale Convective Complexes

Observations indicate that MCSs occur in a variety of shapes and sizes. The largest 1% of MCSs are called mesoscale convective complexes (MCCs). They are most commonly identified from infrared satellite imagery as large regions of cold, nearly circular, cloud tops. MCCs typically produce a contiguous cloud shield of 200,000 km^2 and last for over 15 h (Table 1). These cold cloud tops are the collective

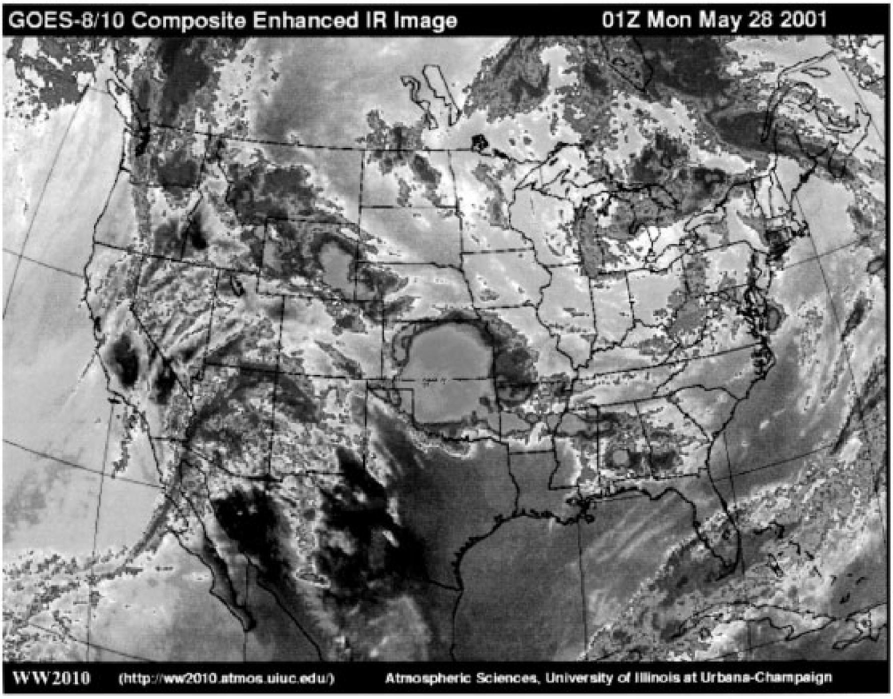

GOES-8/10 Composite Enhanced IR Image
01Z Mon May 28 2001
WW2010 (http://ww2010.atmos.uiuc.edu/)
Atmospheric Sciences, University of Illinois at Urbana-Champaign

Figure 6 (see color insert) Infrared satellite picture of a developing MCS over south-central Kansas and northern Oklahoma at 0100 UTC May 28, 2001. A severe squall line was developing underneath the cloud shield and subsequently moved southeasterly for the next 12 h to the Texas coast. [Image courtesy of University of Illinois WW2010 project (http://ww2010.atmos.uiuc.edu).] See ftp site for color image.

TABLE 1 Mesoscale Convective Complex (MCC) Criteria

Size	A—Contiguous cold cloud shield with infrared (IR) temperatures ≤ 241 K with area $\geq 100{,}000\,\text{km}^2$ B—Interior cold cloud region with IR temperatures ≤ 221 K with area $\geq 50{,}000\,\text{km}^2$
Initiate	Size definitions A and B are first satisfied
Duration	Size definitions A and B must be met for a period of at least 6 h
Maximum extent	Contiguous cold cloud shield (IR temperatures ≤ 241 K) reaches maximum size
Shape	Eccentricity (minor axis/major axis) ≥ 0.7 at time of maximum extent
Terminate	Size definitions A and B no longer satisfied

anvils from interacting thunderstorm cells constituting the MCC. MCCs tend to initiate in the late afternoon, reach their maximum extent around midnight, and terminate early in the morning. They tend to form in situations with large-scale warm advection in flow regimes with strong anticyclonic shear or anticyclonic curvature at the jet-stream level. The processes by which individual cells join to form an MCC are poorly understood. Some MCCs last for several days and can influence large portions of a continent. In addition, when MCCs develop or move over warm ocean waters, they have been observed to evolve into tropical storms.

Squall Lines and Bow Echoes

While MCCs represent only a small portion of all MCSs, commonly observed subsets of MCSs include the squall-line and bow-echo complexes. Squall lines consist of a well-defined line of thunderstorms with an associated stratiform precipitation region and are often identified from radar data. (Fig. 7). The stratiform

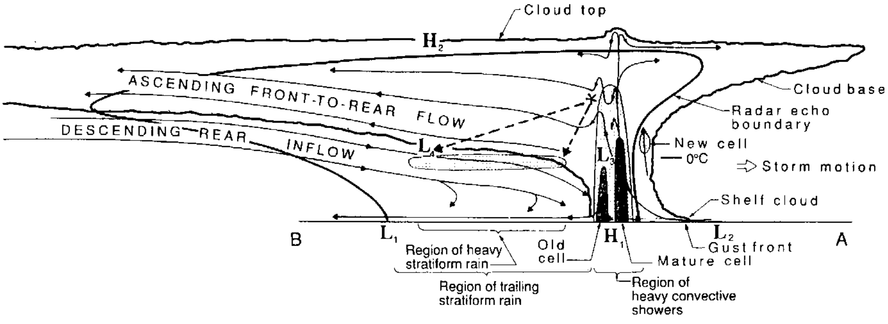


Figure 7 Conceptual model of a squall line with a trailing stratiform area view in a vertical cross section oriented perpendicular to the convective line (i.e., parallel to its motion). (Adapted from Houze et al., 1989, American Meteorological Society.)

precipitation region can be located either in front of, along, or behind the convective line, although it is most common for the stratiform region to be behind the convective line. The convective and stratiform regions often have a symmetric radar depiction early in their life cycle, evolving into a more asymmetric pattern with time (Fig. 8). The distinction between convective and stratiform precipitation is made by comparing the typical vertical air velocities with the terminal fall velocities of ice crystals and snow (~ 1 to 3 m/s). If the vertical air velocities are larger in magnitude than the ice and snow fall velocities, then the precipitation is convective; otherwise the precipitation is stratiform. On radar, the convective regions have much larger values of reflectivity, indicating heavier precipitation or hail. While the typical vertical motions in stratiform precipitation regions are not large, these regions provide ~40% of the total precipitation reaching the ground for many squall lines owing to their large areal extent.

Four typical phases have been identified in the life cycles of MCSs (Fig. 9). The first phase is the formative stage in which the initial thunderstorms are developing and act independently from each other. This is often the time during which the most severe weather occurs. As the thunderstorms grow and merge, the MCS enters the intensifying stage. During this stage the MCS is seen first to have a contiguous precipitation region using radar. The mature stage occurs as a large stratiform rain region is produced, often from older convective cells that weaken and move to the rear of the convective line. This stratiform rain region often grows in size until it is

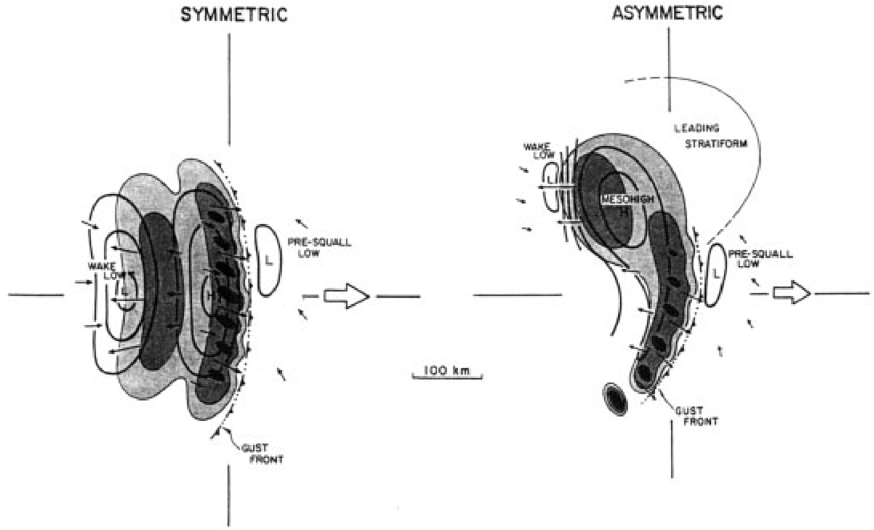


Figure 8 Conceptual model of a midlevel horizontal cross section through (*a*) an approximately two-dimensional squall line, and (*b*) a squall line with a well-defined mesoscale vortex in the stratiform region. In each case, the midlevel storm-relative flow is superimposed on the low-level radar reflectivity. The stippling indicates regions of higher reflectivity. (Adapted from Houze et al., 1989, American Meteorological Society.)

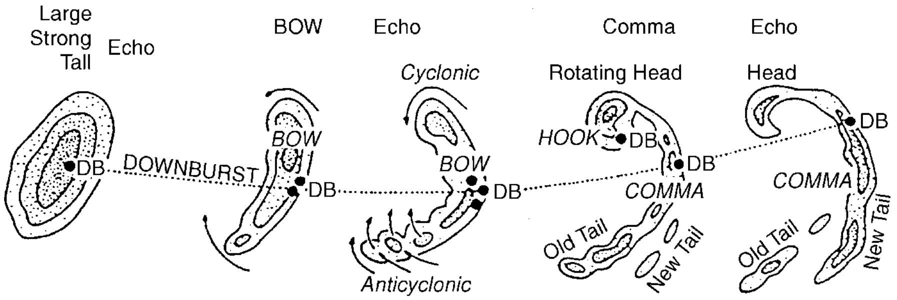


Figure 9 Idealized morphology of an isolated bow echo associated with strong and extensive downbursts. (after Fujita 1978).

an order of magnitude larger in area than the convective region. Significant lightning can occur within the stratiform precipitation region, along with a potential for heavy rainfall and flooding. Positive cloud-to-ground lightning strikes appear to be more common within the stratiform precipitation region than in the convective line. The final, dissipating stage is when the convective portion of the MCS weakens, with fewer and fewer convective cells developing along the convective line. Eventually, all that remains is a weakening region of stratiform precipitation.

One of the most important aspects of MCSs is that the combined effects of the individual thunderstorms and the stratiform precipitation region produce dynamical features that are unique to MCSs and influence both smaller and larger scales of motion. Pools of cooler air are produced near the ground by the evaporation of falling precipitation underneath the thunderstorm cells. In MCSs, these pools merge and influence the development of new convective cells, since warm air approaching these cold pools is lifted up and over the cooler surface air. Thus, MCSs often move in different directions than the individual thunderstorms that constitute the convective line. In the midlevels of the atmosphere, the latent heat released from the change of phase from water vapor to liquid water within the stratiform precipitation region often gives rise to the development of front-to-rear flow in the upper portion of the MCS (Fig. 7). A rear inflow jet also may develop below the front-to-rear flow region of the MCS and approaches the MCS from the rear. Other circulations can develop that produce vortices within the stratiform precipitation regions; these vortices can persist for several days after the initial parent MCS has dissipated. Finally, the upper-level outflows from MCSs alter the mass field near the tropopause and can produce upper-level jet streaks that move away from the MCS and can influence other weather systems downstream. On occasion smaller scale features known as bow echoes (20–200 km in length) form within squall lines or as individual MCSs (Fig. 9). In such situations the severe weather threat, particularly in the form of damaging wind gusts, is enhanced and may continue through much of the life cycle of the MCS. Widespread convection windstorms known as derechos may occur with the longer-lived bow echo dominated MCS events. In such cases it appears the cold pool is enhanced and often moving more rapidly than the mean tropospheric wind. This rapid movement enhances low level storm relative flow and the development of new cells along the gust front. As was mentioned in the last paragraph, predictions of MCS development is complicated. However, in those situations in which bow echo development is dominant, the environmental air above the boundary layer typically displays low relative humidity. This drier air likely enhances the development of the downdraft and cold pool strength as well as the risk of damaging winds at the surface. Owing to the myriad interactions that occur within MCSs, it is probably not surprising that the prediction of MCS development and evolution has proven to be a challenging problem.

5 SUPERCELLS

Definition

The idea that some severe thunderstorms have a markedly different character from other types of thunderstorms owes its origins to the development of radar as an observing tool. Radar gave meteorologists the ability to see the distribution and time evolution of precipitation in thunderstorms. Keith Browning and Frank Ludlam, in England, were pioneers in the interpretation of radar. They participated in field observation campaigns in the United States during the early 1960s after having observed a particularly severe and long-lived hailstorm that struck in and close to the town of Wokingham, England. After observing a number of severe storms with radars, Browning made use of the fact that the distribution of precipitation and its changes with time provides indirect evidence of a thunderstorm's up- and downdrafts. A careful examination of that radar-depicted structure and evolution gave meteorologists a detailed look at what was going on inside a thunderstorm.

Browning and Ludlam noticed that a few severe thunderstorms exhibited a particular set of characteristics: (1) they produced extreme severe weather events (tornadoes, giant hail, and violent wind gusts), (2) they tended to move to the right of the winds, whereas most storms moved more or less with the winds, (3) they exhibited a columnar region of reduced radar echo originally called a vault (later renamed the bounded weak echo region, or BWER), and (4) they had extended lifetimes in comparison to most thunderstorms, persisting for many hours on occasion. At first, these were referred to as severe, right-moving (or SR) storms. In an informal report published in 1962, Browning and Ludlam first used the term *supercell* to refer to these storms. As more and more storms were subjected to scrutiny by radar, the so-called hook echo became associated with supercells (Fig. 10).

The development in the early 1970s of Doppler radar, which can observe the motion of precipitation echoes and, therefore, can be used to infer the airflow, provided conclusive evidence of what Browning and his collaborators had surmised from non-Doppler radar observations: Supercells are characterized by rotation on the scale of the thunderstorm itself. Although from a formal viewpoint, the rotation can be either cyclonic or anticyclonic, the SR storms that Browning called supercells are

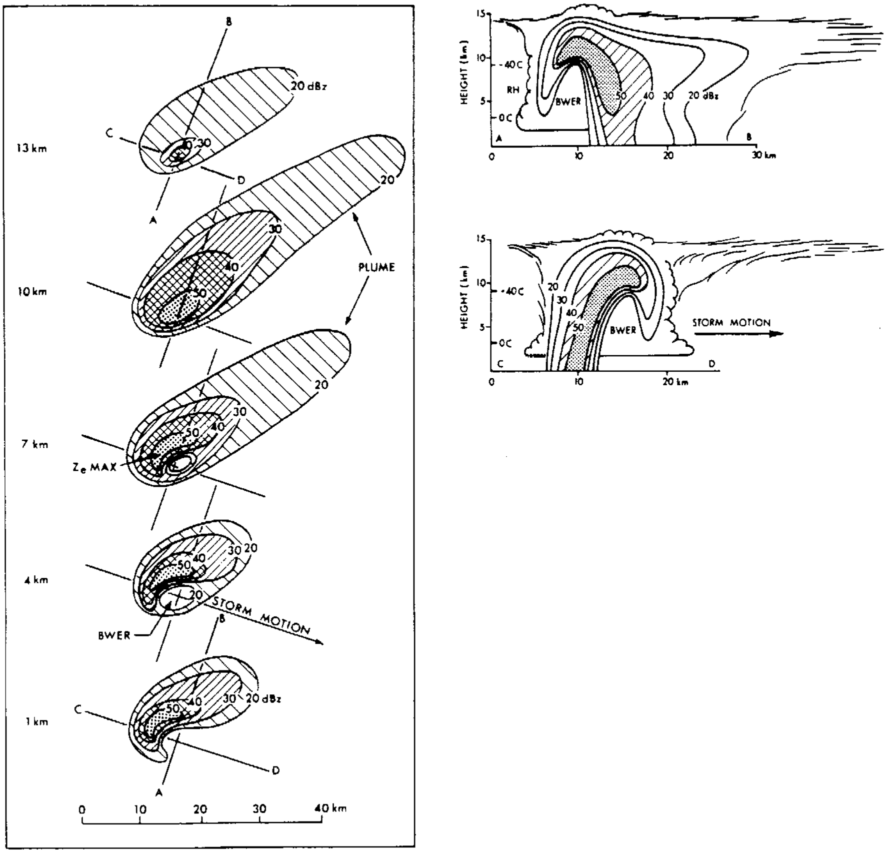


Figure 10 Plan view of supercell showing hook echo, bounded weak echo region (BWER), and forward flank precipitation regions. (Courtesy of National Severe Storms Laboratory.)

characterized by cyclonic rotation. The center of that cyclonic rotation came to be called the mesocyclone. It is the presence of this rotation that distinguishes supercells from other types of thunderstorms. Therefore, it is generally agreed that supercells are thunderstorms that have a deep, persistent mesocyclone. It is the mesocyclone that is deemed responsible for the high probability of severe weather and a supercell's characteristic features.

Supercell Structure and Evolution

Figure 10 illustrates the primary features of a supercell thunderstorm during its mature phase. The fact that supercells can persist for many hours suggests that their structure evolves relatively slowly. However, every storm must have a beginning and end, so no storm is ever truly steady. Some supercells are more nearly steady than others, but all exhibit a basic evolutionary pattern. Thunderstorms begin as pure updrafts, creating towers of cloud called cumulus congestus, which then develop precipitation aloft. The production of precipitation triggers the development of downdrafts, both from the drag effect of having precipitation particles and from the evaporation of some of those particles. Thus, whereas the thunderstorm is initially dominated by updrafts in its mature phase, the thunderstorm has both updrafts and downdrafts. In dissipation, a thunderstorm's updraft weakens and eventually ceases, leaving only the weakening downdraft and precipitation.

Supercells also follow this evolution although the mature phase is prolonged. During the early stages of a supercell, the updraft begins to rotate cyclonically, typically several kilometers above the surface. Thus, the rotation is initially on the same axis as the updraft. With the development of precipitation and the descent of downdrafts, the mesocyclone structure is transformed. Rather than being centered on the updraft, the cyclonic rotation extends into the downdraft, such that the mesocyclone comes to be centered near the interface between updraft and downdraft. The updraft changes its shape from being nearly circular as seen at a given horizontal level, to become elongated and crescent-shaped, with the crescent aligned with the low-level boundary between updraft and downdraft (the so-called gust front). With time, the downdraft and outflow at low levels goes through an evolution called occlusion, whereby the gust front undercuts the updraft, which then dissipates.

Most supercells exhibit a cyclic behavior during their extended mature phase, with new updrafts and mesocyclones forming on the leading edge of the outflow boundary, even as the previous updraft is in the process of dissipation. This process can go on many times, so the mature phase of supercells undergoes a fluctuation on a time scale of 20 to 30 min or so (although the time between "pulsations" is by no means constant). Eventually, of course, the storm can no longer be maintained and the last of updrafts and mesocyclones in the series finishes its life cycle and the storm dissipates.

Origins of Supercell Structure

Browning and his collaborators noticed right from the beginning that supercell storms formed in environments having enhanced vertical wind shear. That is, the winds typically changed both direction and speed with height. Poleward low-level winds became more westerly and increased in speed rapidly in the vertical. Vertical wind shear in the atmosphere is associated with regions of enhanced horizontal temperature differences (or fronts) and also means that the airflow possesses a property called vorticity or spin, which is created when regions of air move past each other at different speeds and/or different directions. When this occurs in association with vertical wind shear, this means the wind profile has horizontal vorticity. This is illustrated in Figure 11, where it can be seen that the change of wind in the vertical implies a potential for rotation about a horizontal axis.

When thunderstorm updrafts develop in regions of strong vertical wind shear, they cause the horizontal vorticity in their surroundings to be tilted upward into the vertical, and it is this uptilted vorticity that gives rise to the rotation of the updraft

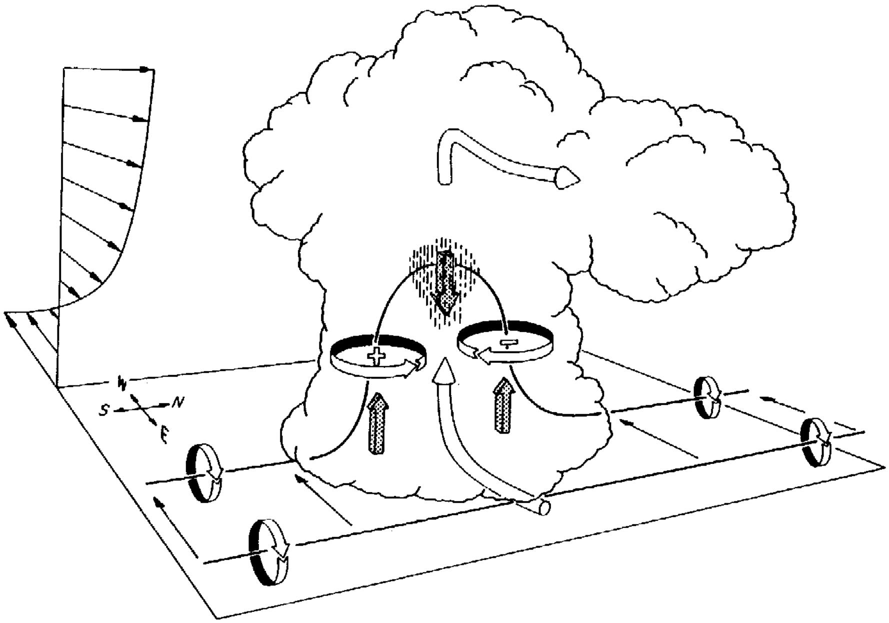


Figure 11 Tilting of horizontal vorticity into the vertical via a thunderstorm updraft. (Courtesy of National Severe Storms Laboratory.)

about a vertical axis during its development. This has been demonstrated in computer simulation models and is clearly the source for mesocyclonic rotation in thunderstorms.

Available evidence indicates, however, that the creation of rotation near the surface is associated with a more complex process that involves a supercell's downdrafts. In some supercells, the mesocyclone aloft and that developing near the surface can interact strongly, creating a deep column of intense rotation. Such storms are the primary producers of long-lasting, strong tornadoes and often result in families of tornadoes, with each tornado being a reflection of another pulsation in a cyclic supercell. Other supercells fail to develop a strong interaction between the low-level mesocyclone and the mesocyclone aloft. Such storms can produce other forms of severe weather, notably giant hail, but tornadoes are relatively infrequent and tend to be brief and weak if they do form.

Hazardous Weather Associated with Supercells

It appears that the development of a mesocyclone has a strong influence on the storm, and that influence is such that the likelihood of severe weather in all forms increases if a storm becomes supercellular. Supercell storms constitute only a small fraction of the total number of severe thunderstorms, but they account for a disproportionate share of the most intense forms of hazardous weather.

Tornadoes are arguably the most hazardous weather event associated with convective storms. Although tornadoes are by no means limited to supercells, those tornadoes associated with supercells have a much greater likelihood to be intense and long-lived than those produced by nonsupercell thunderstorms. In turn, this means that such tornadoes have the highest potential for damage and casualties. This is exemplified by events on May 3, 1999, when an outbreak of 69 tornadoes across Oklahoma and Kansas was produced by only 10 supercell storms. One tornado from the first supercell of the day caused $1 billion in damage and 36 fatalities as it tracked first through rural areas southwest of Oklahoma City, Oklahoma, and then on into the metropolitan area.

As with tornadoes, the hail produced by supercells is much more likely to exceed 2 in. (5 cm) in diameter than in nonsupercell storms. There is strong scientific evidence to believe that updrafts are enhanced by the presence of a mesocyclone, and giant hail requires an intense updraft for its formation. Some supercells are prolific hail producers, creating long swaths of hail up to 10 cm in diameter. When such storms interact with populated areas, they can cause damage on the order of $100 million or more from broken glass, dented motor vehicles, roof destruction, and vegetation damage. Occasionally, people are seriously injured or even killed by being caught outdoors during a fall of giant hail.

Windstorms of a nontornadic nature in supercells can also reach extreme proportions. A few times per year in North America, supercells produce swaths of wind up to 20 km wide and more than 100 km long, within which winds can exceed 25 m/s for more than 30 min, with peak gusts approaching 50 m/s. In forested areas, such events produce vast blowdowns of trees. Interactions with populated areas are rare, but the potential for destruction is enormous. In such events, hail can accompany the strong winds, adding to their destructive potential.

Finally, a few supercells are responsible for prodigious rainfall production. It appears that most supercells are not very efficient at producing rainfall because they often are associated with processes promoting evaporation of precipitation. Moreover, supercells are mostly isolated storms that move over a given location relatively quickly. Nevertheless, their intense updrafts can process a lot of water vapor into precipitation, however inefficiently it is accomplished. Thus, supercells have produced bursts of precipitation exceeding 200 mm/h, even if only for a short time. Such intense rainfall rates can create flash floods, particularly in urban areas where runoff is so high owing to the lack of permeability of urban environments composed mostly of concrete, buildings, and other hard surfaces.

Variations on the Theme

Supercells are not all the same, but they have certain common features. The presence of a deep, persistent mesocyclone is the defining feature of a supercell, so it is possible to look only at storms meeting that criterion. Using this broad definition, we should not be surprised to learn that variations on the supercell theme exist.

The most widely used classification scheme for supercells is based loosely on the notion that mesocyclones vary in the amount of precipitation they contain. The prototypes are called the low-precipitation, classic, and high-precipitation supercell.

Low-precipitation (LP) supercells characteristically have little or no precipitation within their mesocyclones. The absence of heavy precipitation means they typically do not produce strong downdrafts and outflow, nor are they likely to be tornadic. However, they can produce falls of giant hail and, like other supercells, tend to do so in relatively long swaths. They typically are not very large storms and tend to be observed mostly in the transitional environments between arid and moist regions (e.g., the western Great Plains of North America). Figure 12 shows a schematic illustration of the appearance of an LP supercell, both visually and on radar.

Classic (CL) supercells most clearly resemble those described by Browning and his collaborators. They often exhibit most, if not all, of the traditional radar echo morphology associated with them in the literature. They produce all forms of severe weather, including the most extreme examples, although they are not likely to be flash flood producers. Figure 13 illustrates the archetypical appearance of a CL supercell.

Finally, high-precipitation (HP) supercells have mesocyclones deeply embedded in precipitation, although there may be a narrow corridor that, at low levels, remains free of precipitation (sometimes referred to as an inflow notch). Figure 14 shows the typical appearance of HP supercells. With heavy precipitation falling into the downdrafts on a supercell's rear flank, the potential for strong winds is quite high. Clearly, HP supercells also have a potential to be heavy rainfall producers, as well as giant hail. HP supercells stand somewhere between CL and LP supercells in their tornado potential. It is rare for a violent, long-track tornado to be associated with storms displaying a predominantly HP structure.

6 TORNADOES

Tornadoes are rotating columns of air that extend from the surface to the interior of a thunderstorm cloud (or, more correctly, a cloud associated with deep moist convection, with or without thunder). The distinction between tornadoes and other vortices within thunderstorms is not always clear, so tornadoes are sometimes defined as having wind speeds near the surface that have the potential to cause damage. Many tornadoes produce cylindrical or conical clouds of condensed water (Fig. 15*a*), but a funnel cloud is not always present. Some tornadoes are made visible only by lofted debris (Fig. 15*b*).

The number of reported tornadoes in the United States now typically exceeds 1000 per year. Annual totals of fatalities and property damage caused by tornadoes vary considerably from year to year. The death toll, as a fraction of the total population of the United States, has been decreasing on average since the mid-1920s. Improved forecasting, detection, and warning of tornadoes; improved communication of warnings; increased public awareness of safety precautions; urbanization; and changes in building standards may be some of the important factors in explaining

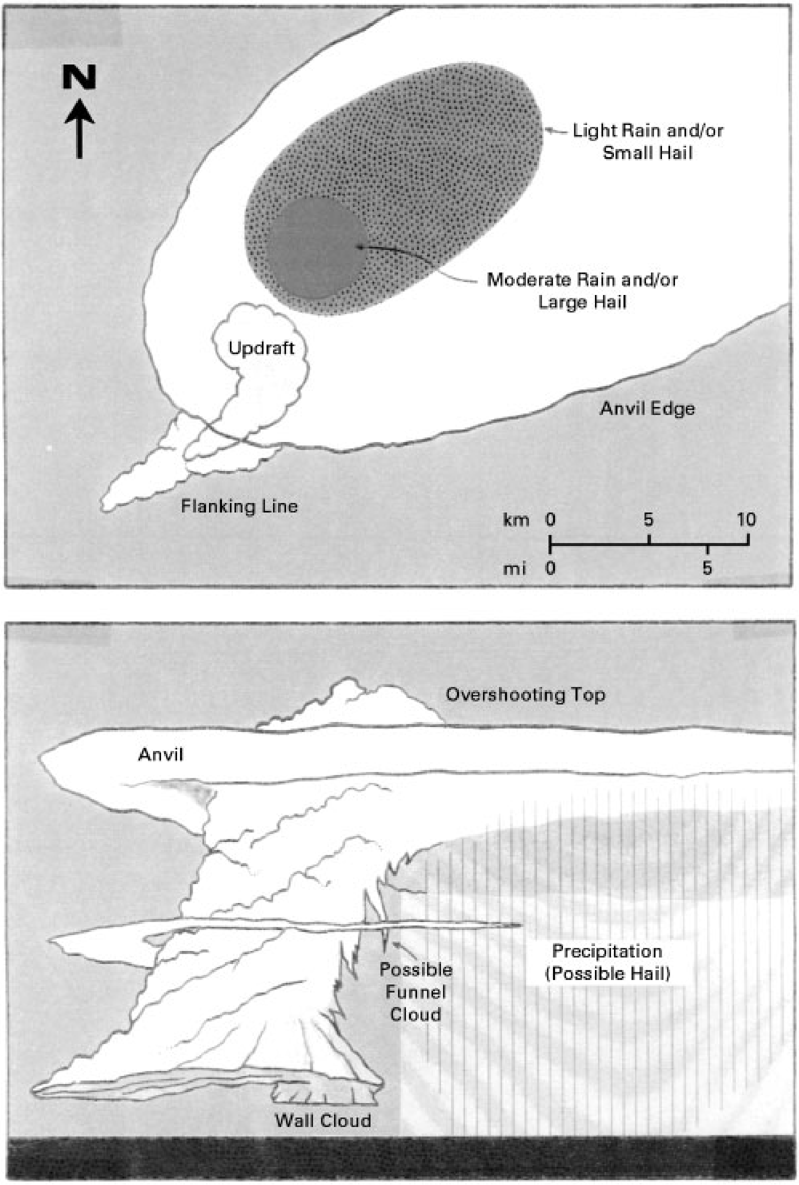


Figure 12 Schematic of a low-precipitation supercell. (Courtesy of National Severe Storms Laboratory.) See ftp site for color image.

this trend. Tornadoes typically last less than 10 min but have been known to persist for over an hour. Wind speeds may exceed 120 m/s (270 mph) in particularly strong tornadoes but are less than 60 m/s (135 mph) in most cases. In some tornadoes, the distribution of strong winds around the vortex core is relatively uniform; in others,

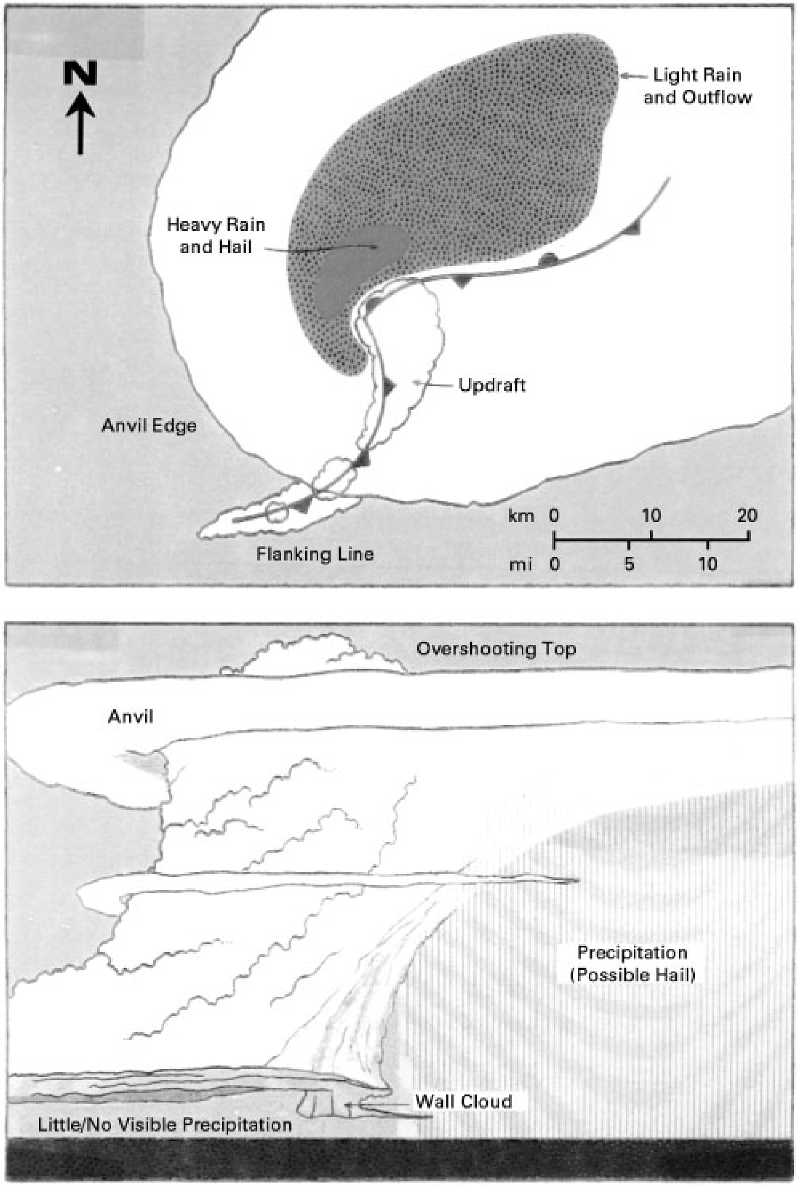


Figure 13 Schematic of a classic supercell. (Courtesy of National Severe Storms Laboratory.) See ftp site for color image.

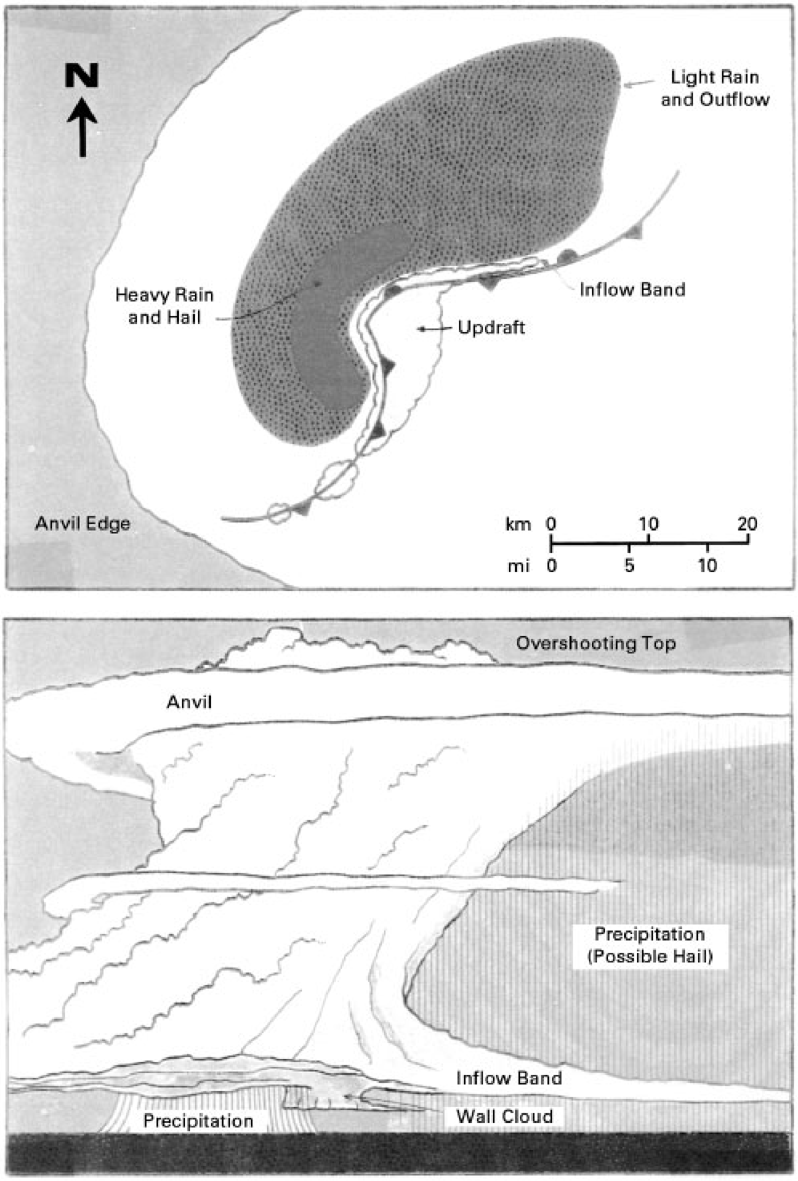


Figure 14 Schematic of a high-precipitation supercell. (Courtesy of National Severe Storms Laboratory.) See ftp site for color image.

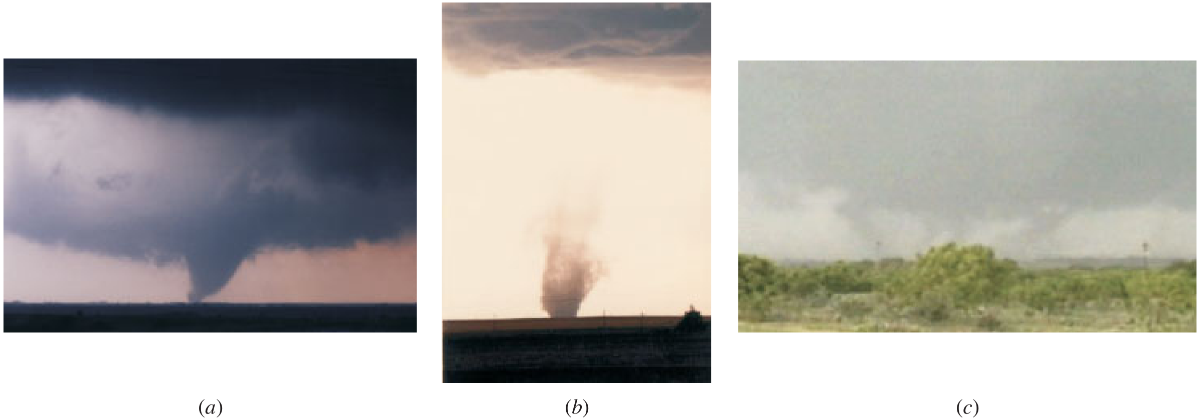

(*a*) (*b*) (*c*)

Figure 15 Tornado photographs: (*a*) Tornado with a cone-shaped condesation funnel near Stockton, Kansas, on May 15, 1999 (copyright 1999 by David Dowell). (*b*) Tornado with no condensation funnel near Denver, Colorado, on July 2, 1987 (copyright 1987 by Bill Gallus). (*c*) Tornado with multiple subvortices near Elbert, Texas, on April 30, 2000 (copyright 2000 by David Dowell). See ftp site for color image.

the strongest winds may occur within relatively small subvortices (Fig. 15*c*). Although tornadoes may extend to over 10 km above ground within the cloud, the strongest horizontal winds in mature tornadoes occur within 100 m of the surface. Near the radius of strongest horizontal winds, air rises at speeds comparable to those of the horizontal winds; the extreme vertical velocities may loft debris to great heights. Near the center of the tornado, there may be descending motion. On rare occasions, the large-scale atmospheric pattern may be favorable for the formation of tornadic thunderstorms over a broad region covering multiple states. The "super outbreak" of April 3–4, 1974, is an extreme example. During this outbreak, at least 148 tornadoes occurred within 14 states from the Midwest to the Southeast. The majority of tornadoes, however, are isolated occurrences. Most thunderstorms form in environments that are not favorable for tornadoes, or when the environment would support the development of tornadoes, storms do not always form.

Scientists continue to be challenged to explain the details of how tornadoes form (and to explain why tornadoes do not always form on days when they are anticipated). Organized field programs to study tornadoes began in the early 1970s and continue today. Scientists use a number of mobile devices (Doppler radar, instrumented automobiles, weather balloons, etc.) to collect measurements of wind, temperature, pressure, humidity, and precipitation in and near tornadoes (e.g., Fig. 16). To assess how a tornado forms in a particular thunderstorm, scientists

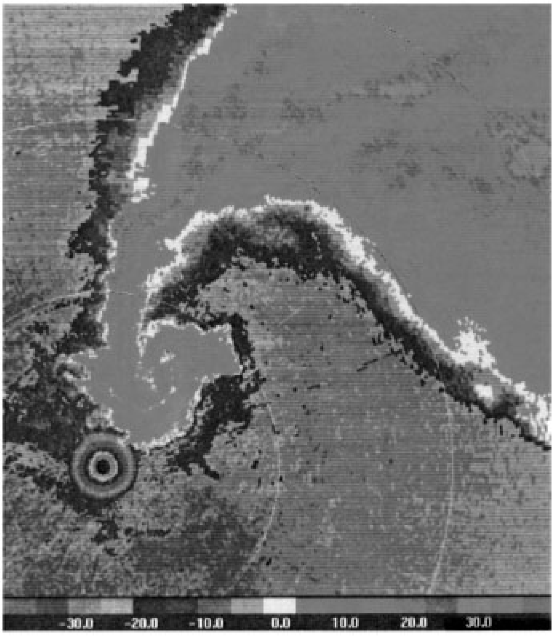


Figure 16 (see color insert) High-resolution measurements of the reflectivity factor (an indicator of precipitation size and number concentration) by the Doppler on Wheels mobile radar in a tornadic storm with a hook echo near Rolla, Kansas, on May 31, 1996. The thin blue rings indicate range from the radar at intervals of 5 km. See ftp site for color image.

require accurate measurements at high temporal and spatial resolution; the need for such complete observations continues to challenge observational capabilities.

Significant advances in our understanding of tornadoes and their parent storms have also come from numerical simulations with high-speed computers. In a numerical model, the thunderstorm and its environment are represented on a three-dimensional grid. During each time step of the simulation, the dynamical equations governing the changes of wind, temperature, pressure, humidity, liquid water content, and ice content are solved at each grid point. Numerical simulations (e.g., Fig. 17) have reproduced storm features analogous to those in observed storms.

The prevailing hypotheses for how tornadoes form all involve horizontal convergence of low-level air that has significant angular momentum. As air retains its angular momentum with respect to the center of the region of convergence while spiraling inward, the rate of rotation increases. The sequence of events leading up to the formation of a tornado is not identical in all cases. If the air near the surface in the environment of a thunderstorm already had significant rotation before the thunderstorm formed, then the mechanism of tornado formation could be relatively simple (Fig. 18). For example, a low-level boundary, along which there is a shift in the wind direction and speed, may be present in the environment. The wind shift may be organized into circulations that are initially a few hundred meters to a few kilometers wide (Fig. 18). If one of these circulations coincides with a growing

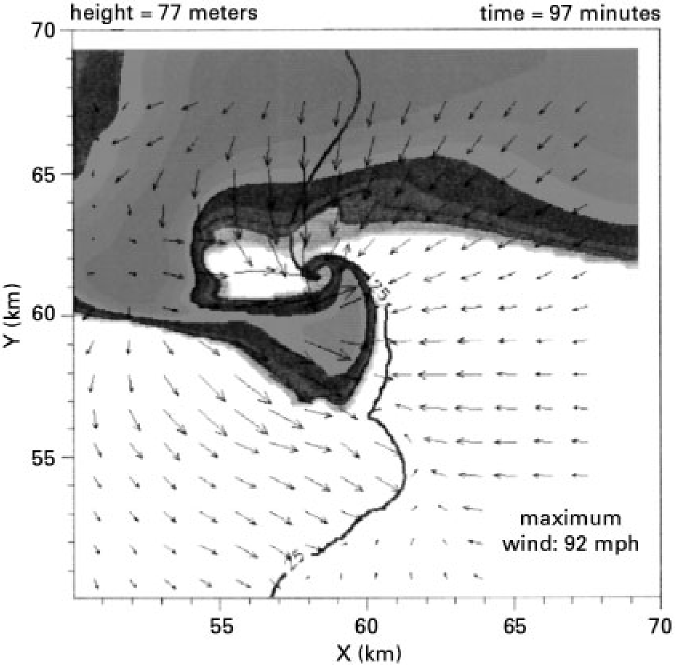


Figure 17 Simulated low-level reflectivity factor (shading) and horizontal wind (vectors) in a numerical simulation of a tornadic storm. (Courtesy of Matt Gilmore of Cooperative Institute of Mesoscale Meteorological Studies.) See ftp site for color image.

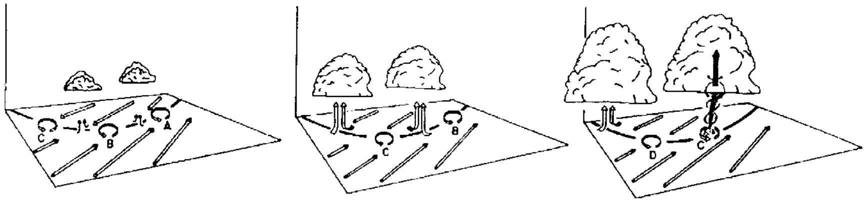


Figure 18 Schematic of nonsupercell tornado formation (by R. Wakimoto and J. Wilson; copyright 1989 by the American Meteorological Society). The arrows indicate the wind direction. Numbered loops represent circulations along a low-level boundary, which is indicated by a black line.

thunderstorm updraft, then a tornado may form as the circulating air at the base of the updraft converges to smaller and smaller radii. This mechanism involving the interaction of an updraft with a preexisting low-level circulation may apply to tornadoes that form within otherwise nonsevere deep moist convection over water ("waterspouts") but may also explain the formation of some tornadoes over land.

The formation of most of the tornadoes in supercell thunderstorms appears to be more complicated. Supercells develop in environments in which there is strong vertical shear (i.e., a change in direction and/or speed with height) of the horizontal wind; such shear is associated with rotation about a horizontal axis. In addition, rotation about a horizontal axis may develop within a mature thunderstorm when there are significant horizontal variations of air density in the region where rain and hail are falling. Updrafts and downdrafts within the supercell thunderstorm may tilt the orientation of the rotation such that a component of it becomes rotation about a vertical axis. If the air that is rotating about a vertical axis is drawn into the region of horizontal convergence at the base of the thunderstorm updraft, then a mesocyclone (a region of rotation a few kilometers wide) and perhaps tornado (a narrower, more intense vortex) may form. A major current research problem is to determine why tornadoes form in some supercell thunderstorms but not in others.

7 RADAR CHARACTERISTICS OF SEVERE STORMS

Use of Radar in Observing Severe Storms

Radar has been used to identify and track severe local storms since its invention in the 1940s. Weather radar detects hydrometeors within storms with a series of pulses of electromagnetic energy, directed by an antenna that is mechanically rotated in azimuth and elevation. It alternates transmitting and receiving those pulses and measures the distance to the target by the time delay between a pulse transmission and echo arrival. Figure 19 illustrates the typical radar beam geometry. Because of the hazards within severe storms to aircraft penetrations, radar has been the primary tool that has been used by meteorologists to probe the inner structure and circulations of storms to better understand the physics of their behavior.

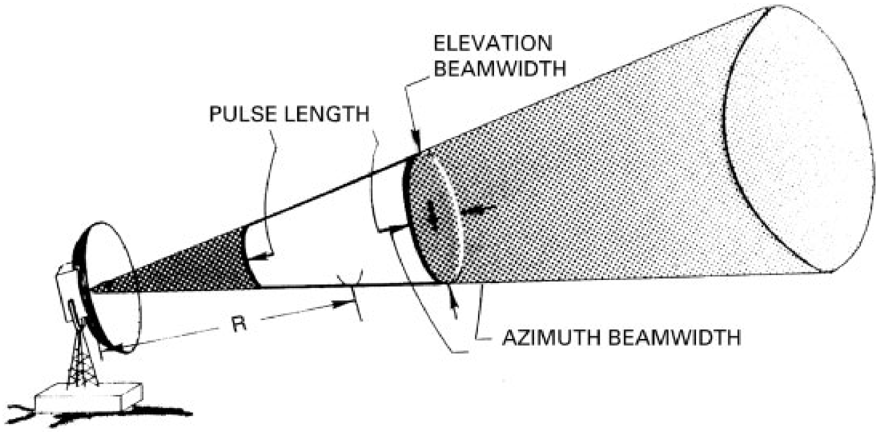


Figure 19 Radar beam geometry. The size of the beam is a function of antenna size and radar wavelength.

Early techniques utilized "incoherent" technology of simply displaying returned power normalized for range on a horizontal display device (termed a *plan position indicator*, or PPI). Severe storms, sometimes producing hazardous weather such as large hail, high winds, and tornadoes, very often exhibit a characteristic structure that could easily be tracked on a PPI display, and suitable warnings could be made for regions in the storm's path by simple extrapolation of storm motion. One of the key developments in radar technology that made radar valuable for distinguishing between tornado-producing storms and other less severe events such as hailstorms was the development of Doppler radar techniques. Doppler radars not only detect and measure the mean power received from a target but also its relative motion. Thus regions of a storm can be seen to approach or recede from a radar site and inferences about rotation within the storm (e.g., tornado mesocyclones that are the parent circulation of tornadoes) can easily be made. This motion detection capability greatly assists weather forecasters to better distinguish between storms that produce tornadoes from ones that do not.

The U.S. National Weather Service (NWS) was quick to recognize the value of radar in providing timely public warnings of severe weather. In the 1950s a network of incoherent radars was deployed (called the WSR-57). With the realization by the 1980s that Doppler technology could offer significant improvement in tornado warning lead time, a new generation of operational weather radars (termed at the time *NEXRAD*, now called the WSR-88D radar) was deployed to meet the warning missions of the NWS, U.S. Air Force, and the Federal Aviation Administration. The network consists of over 130 radars providing nearly complete coverage of the continental United States (Fig. 20). The WSR-88D has greatly increased the tornado warning lead time from nearly zero before the deployment of the WSR-88D network to nearly 10 min today.

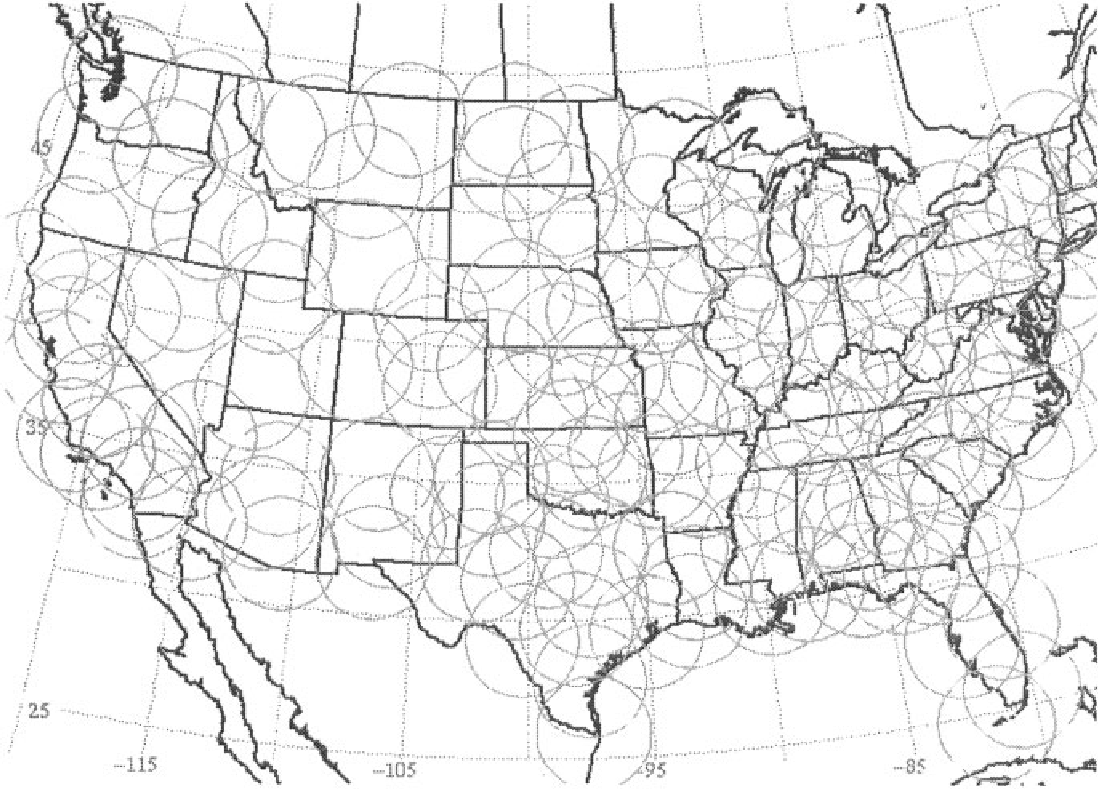


Figure 20 The 230-km range of each WSR-88D radar site.

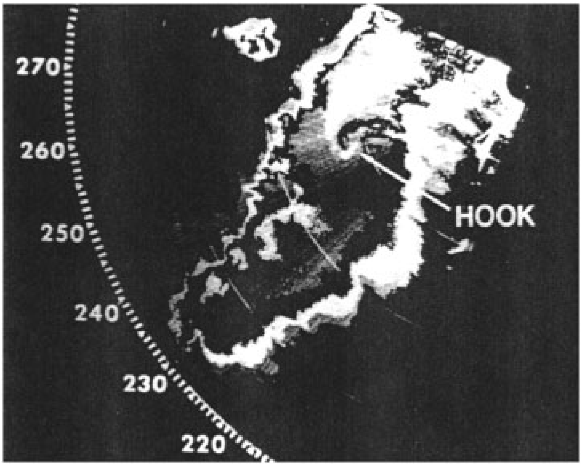


Figure 21 Plan Position Indicator (PPI) display of contoured echo power of the "Tabler" storm on June 6, 1974, in central Oklahoma (from Brandes, 1977). The region labeled "HOOK" is an indication of a likely region for a tornado. Radar location is in the upper right. Range marks (arcs) are spaced 20 km. Echoes are contoured in a gray–white–black pattern starting at 10 dBZ in 10 dBZ steps.

Storm Structure Revealed by Radar Observations

Precipitation echoes from mature thunderstorms usually are easily recognizable even by incoherent radars. On a PPI display they are characterized by high reflectivities (reflectivity is proportional to the sixth power of all the hydrometeor diameters within the pulse volume with units of mm^6/m^3 and is expressed in logarithmic units) and sharp gradients of intensity (Fig. 21). The presence of "hook" echoes, or echoes with notches, on their southern sides has been correlated with tornadic potential. The value of Doppler radar in identifying tornadoes was first shown in the early 1970s when velocities were measured from inside the Union City, Oklahoma, thunderstorm. The pattern of large wind shear (termed a *tornado vortex signature*) between adjacent beams on either side of the tornado location seen by ground observers persisted for over 40 min throughout a deep depth of the thunderstorm. A significant finding for possible tornado warning was the presence of this shear pattern aloft for about 20 min before the tornado was on the ground. Often, however, most Doppler radars do not actually "see" a tornado because of its small size relative to the beam size. It is the much larger parent circulation (termed a *mesocyclone*), often high up in the storm, which is first detected by the Doppler radar (Fig. 22).

While tornado detection by Doppler radar is an important component of the NWS warning responsibility, single Doppler radars also identify many other severe weather hazards. For example, strong downdrafts (often termed *downbursts*) that can affect the safety of airplanes can be identified by the pattern of Doppler velocity

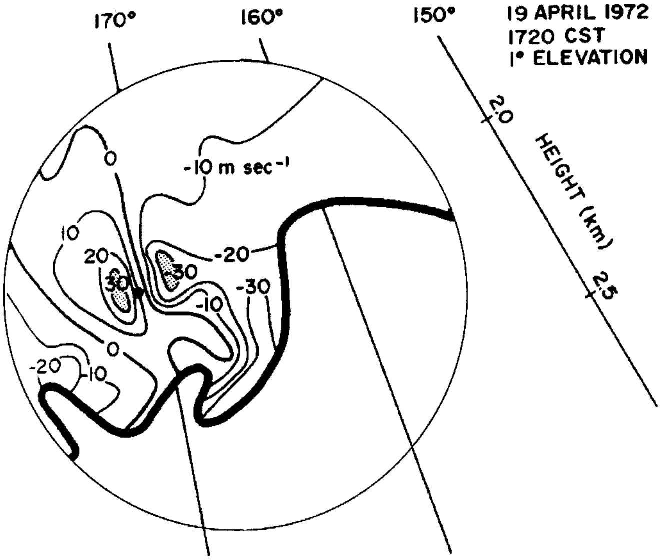


Figure 22 Single Doppler radar signature of a strong mesocyclone indicative of rotation. Dot indicates location of tornado. Negative velocities are approaching the radar. Straight lines labeled in degrees indicate azimuth radials from the radar (from Burgess and Lemon, 1990).

divergence near the ground. Storms that produce large hail are recognized by their elevated reflectivity structure aloft.

Although important for recognizing the presence of severe weather through patterns of velocity, single Doppler radars provide only limited information about the internal circulations of storms, particularly updraft and downdraft strengths that are critical for precipitation formation processes. A single Doppler radar can only provide information about the component of the flow directed either toward or away from the radar. To derive actual wind fields at least two Doppler radars are required. Thus, to understand why storms move the way they do and produce tornadoes, hail, and sometimes damaging straight-line winds, researchers have utilized multiple Doppler radars, and even Doppler radars mounted on airplanes, to provide simultaneous observations that can be combined to produce a three-dimensional description of the flow within severe storms. Figure 23 is an example of combining the radar data from two radars to produce an analysis of the updraft through a strong hailstorm observed on May 26, 1985. In this case the two radars were the Cimarron Doppler radar operated by the National Severe Storms Laboratory and a Doppler radar mounted on the WP-3D aircraft from the National Oceanic and Atmospheric Administration. By flying the aircraft close to the storm, the beams from the aircraft's radar, which was scanning in a vertical orientation normal to the aircraft's flight track, can be combined with the PPI scans from the Cimarron radar.

Progress toward understanding the complex dynamic and microphysical forces acting to control the behavior of severe storms has certainly been aided by multiple

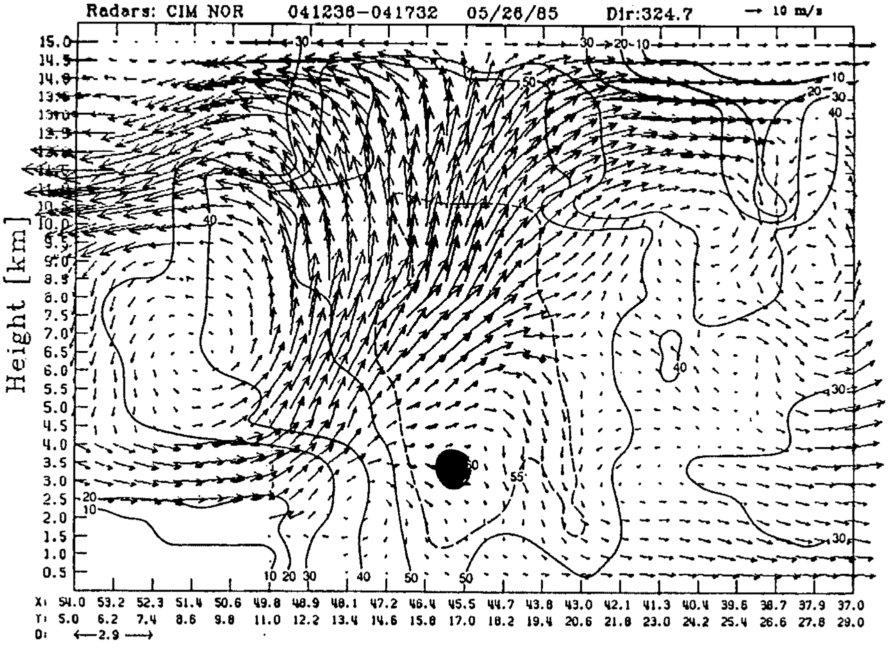


Figure 23 Vertical cross section through the core of a hailstorm simultaneously observed by the P-3 aircraft and the Cimarron Doppler radar on May 26, 1985. Solid lines are reflectivity (dBZ). The 10 m/s wind vector is shown in the upper right.

Doppler case studies. Doppler radar data, however, suffers from a number of limitations, the most severe of which is its restriction to areas where there are hydrometeors that scatter the electromagnetic radiation. Thus clear air regions are usually devoid of observations unless the storm is very close to the radar and the clear air contains some other particle scatterer such as insects. Even in those circumstances the amount of coverage is generally slight. One of the most powerful approaches has been to combine Doppler radar with numerical simulations of convective storms. The simulations use observed environmental conditions to initialize the grid domain and the Doppler observations to validate the computed solutions. Numerical simulations, even using state-of-the-art approaches, have significant limitations. Even so, useful insights have been derived from these simulations in the case of the strongest and longest-lived convective storm, the supercell thunderstorm, as evidenced by the closeness of simulations and Doppler observations where they overlap, such as the high reflectivity region of the storm core where tornadogenesis occurs.

The hypothesis that emerges of tornadogenesis from the analysis of Doppler winds and numerical simulations is that the vertical vorticity (or "spin") develops initially from the tilting of environmental vorticity set up by the low-level wind shear into the vertical by the principal storm updraft. Once initiated, the vorticity is

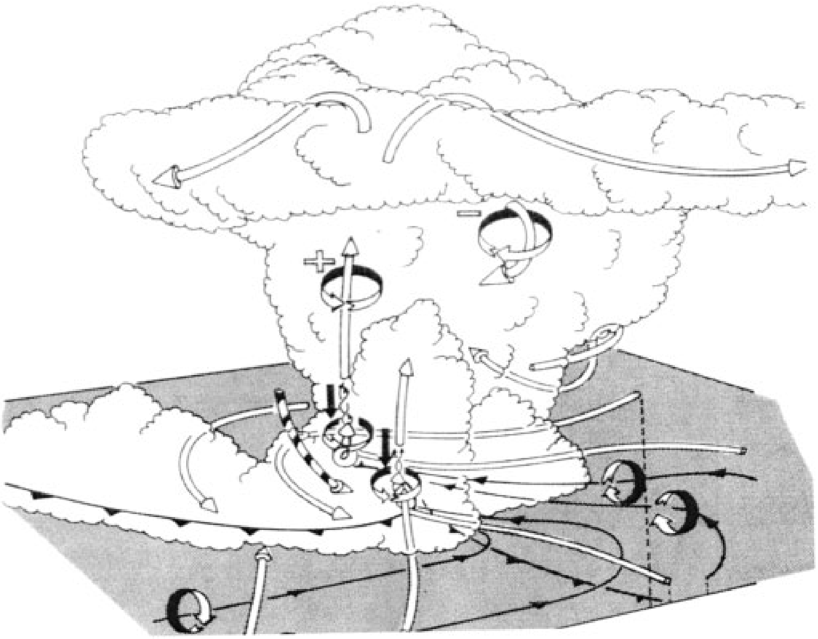


Figure 24 Three-dimensional schematic representation of the processes leading to tornado formation based on Doppler radar and numerical simulations. Cylindrical arrows depict flow in and around the storm. Thin lines show the low-level vortex lines, with vector direction by arrows along the lines and sense of direction also by circular-ribbon arrows. (Adapted by Klemp, 1987.)

increased by concentration of spin by convergence of air similar to the way in which a figure skater increase his or her spin by contracting their arms (called *conservation of angular momentum*). This process is illustrated schematically in Figure 24. Numerical simulations reveal the pressure gradient forces implied in the Doppler winds. Both observations and numerical models have shown that strong low-level rotation promotes a downdraft that spreads cold air at low levels around the storm and helps create convergence along the resultant “gust front.” These features have also been seen visually by storm intercept teams.

Future Advances

Doppler radars have enabled considerable leaps in our understanding and improvements in warning lead-time of severe storms. The establishment of a national network of Doppler radars (Fig. 20) has provided nearly complete nationwide coverage. Some hazards, however, are not optimally detected by the WSR-88D network (e.g., aviation hazards such as downbursts and sudden near-surface wind shifts caused by gust frontal passage) and require specialized radars for each major airport called the Terminal Doppler Weather Radar (TDWR) used by the Federal Aviation

Administration (FAA). These specialized radars can also be used to augment the WSR-88D coverage and improve the NWS warning lead time.

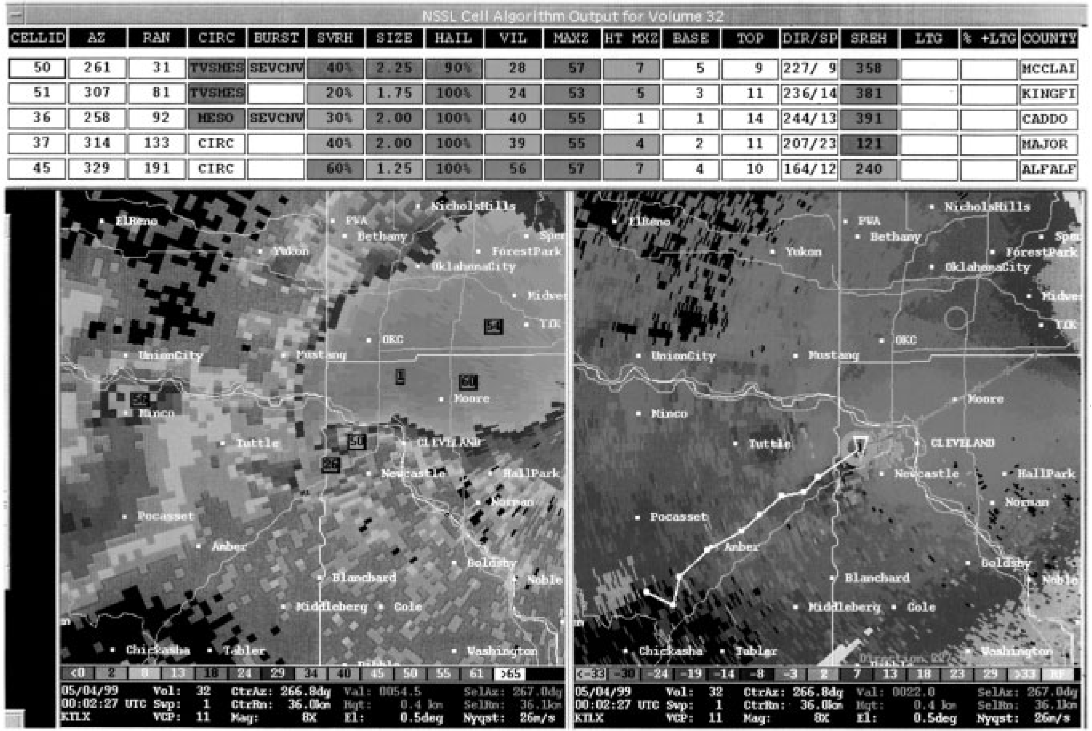
Tied to the national network is the ability to rapidly process and interpret the single Doppler radar data for the detection of a variety of hazards through pattern recognition. This “marriage” of the computer and radar greatly accelerates the detection of patterns that can be very subtle and often are embedded in large areas of high reflectivity. WSR-88D radar data from the May 3, 1999, Oklahoma City tornado outbreak is shown integrated with the computerized warning decision process in Figure 25. The Warning Decision Support System (WDSS) can independently track many storm cells simultaneously and provide forecast guidance as to rain severity, hail size, tornadic potential, and probably accurate tornado location and tracking without requiring a dedicated radar scientist.

Other technological advances in development may help identify and warn of severe storm hazards. For example, improved hail detection and rainfall estimation may be possible with polarization of the radar beam. In this approach, two pulses are alternatively transmitted with orthogonal polarizations (e.g., horizontal and vertical). The polarization of each received pulse is measured and various products (such as the ratio of the horizontal-to-vertical polarizations) can be computed. According to electromagnetic scattering theory, if a particle were not circular, it would scatter preferentially in its long axis, i.e., a pancake-shaped raindrop would scatter more horizontally polarized radiation than vertical polarization. This information can be used to improve the accuracy of rainfall estimates, discriminate hail from heavy rain, and reduce uncertainties about the drop-size distribution that is being sampled. The application of polarimetric observation to weather forecasting is still an active area of research. Even so, the utility of polarimetric data in improving weather forecasting is already recognized and the NWS is already planning to upgrade the U.S. WSR-88D radar network to have this capability.

One of the most severe limitations to multiple Doppler radar analysis is the rather large uncertainty in vertical air motion estimates. If fundamental problems in convective dynamics are to be addressed, these uncertainties need to be reduced so that a more complete picture of the mass, momentum, pressure, vorticity, thermodynamic, electrical, and water substance interactions can be examined. One of the ways to reduce the uncertainties is to observe the phenomena at higher spatial and temporal density since it is known that convective elements possess large kinetic energy on spatial scales of 1 to 2 km and exhibit significant evolution over 1 to 3 min. Therefore “rapid-scan” radars that can sample a storm volume within 1 min at data densities of 200 to 300 m are needed. Such data would permit adjoint analysis methods with cloud resolving numerical models to be implemented. This type of radar is currently used in military applications and possibly could be adapted to examine severe weather.

8 SEVERE STORM FORECASTING

In the United States, the National Weather Service’s Storm Prediction Center (SPC) in Norman, OK, is responsible for forecasting thunderstorm occurrence as well as



NSSL Cell Algorithm Output for Volume 32

CELLID	AZ	RAN	CIRC	BURST	SVRH	SIZE	HAIL	VIL	MAXZ	HT MXZ	BASE	TOP	DIR/SP	SREH	LTG	% +LTG	COUNTY
50	261	31	TVSMES	SEVCNV	40%	2.25	90%	28	57	7	5	9	227/ 9	358			MCCLAI
51	307	81	TVSMES		20%	1.75	100%	24	53	5	3	11	236/14	381			KINGFI
36	258	92	MESO	SEVCNV	30%	2.00	100%	40	55	1	1	14	244/13	391			CADDO
37	314	133	CIRC		40%	2.00	100%	39	55	4	2	11	207/23	121			MAJOR
45	329	191	CIRC		60%	1.25	100%	56	57	7	4	10	164/12	240			ALFALF

```
05/04/99          Vol:  32   CtrAz: 266.8dg  Val: 0054.5       SelAz: 267.0dg
00:02:27 UTC Swp:    1   CtrRn:  36.0km  Hgt:     0.4 km   SelRn:  36.1km
KTLX              VCP:  11   Mag:        8X  El:      0.5deg   Nyqst:  26m/s
```

```
05/04/99          Vol:  32   CtrAz: 266.8dg  Val: 0022.0       SelAz: 267.0dg
00:02:27 UTC Swp:    1   CtrRn:  36.0km  Hgt:     0.4 km   SelRn:  36.1km
KTLX              VCP:  11   Mag:        8X  El:      0.5deg   Nyqst:  26m/s
```

Figure 25 (see color insert) Screen output from the National Severe Storms Laboratory Warnings Decision Support System of the May 3, 1999, Oklahoma City tornado. The table at the top shows the system is tracking 5 different storms with a variety of algorithm-determined characteristics such as direction and speed of motion, presence of hail and its maximum size, whether a circulation or mesocyclone is present, and its echo top. Left panel is radar reflectivity with cities and county boundaries as background. Boxes with numbers correspond to locations of tracked storms. Right panel is radial velocity. White line is storm #50 track with future positions shown as the purple line. Yellow circle shows the presence of the tornado vortex signature. (Courtesy of Greg Stumpf of the National Severe Storms Laboratory.) See ftp site for color image.

most hazardous weather associated with thunderstorms. This includes tornadoes, damaging straight-line winds, large hail, and heavy rainfall that can result in flash floods. The SPC's primary focus is on the risk of tornadoes, damaging straight-line winds (58 mph or greater), and large hail ($\frac{3}{4}$ inch or greater in diameter). In this section we will explain how the SPC makes forecasts for thunderstorms and these three hazards.

Severe thunderstorm forecasting began in earnest in the mid-twentieth century. The U.S. Air Force began making rudimentary internal forecasts for severe weather in the late 1940s. By the early 1950s, the National Weather Service formed a national unit of specialists to forecast severe thunderstorms and tornadoes for the 48 contiguous states. Called the National Severe Storms Forecast Center, this unit was based in Kansas City, Missouri, for more than 40 years before relocating to Norman, Oklahoma, and being renamed the Storm Prediction Center in 1997 (Fig. 26). Currently, 20 specialized meteorologists work in teams of four to monitor weather conditions around the clock, 7 days a week, all year long. An "outlook" forecaster makes forecasts for severe weather out to 3 days ahead (Fig. 27). The other three specialists issue short-term forecasts (1 to 7 h ahead) that include tornado and severe thunderstorm watches as well as other products. Typically, watches are parallelogram in shape covering an area averaging about 25,000 square miles (about the size of the state of Iowa; Fig. 28). A severe thunderstorm watch is issued when there is a significant and concentrated threat of damaging straight-line winds and/or

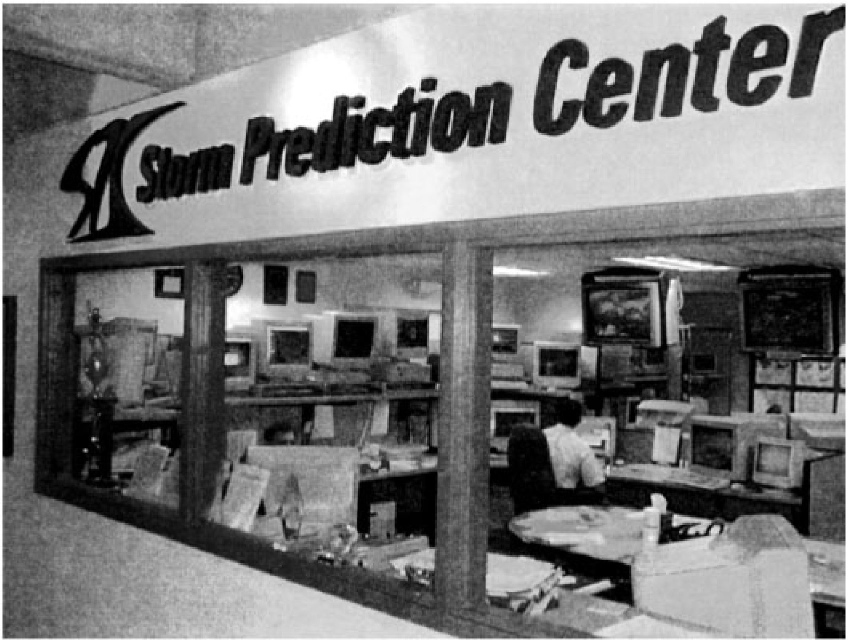


Figure 26 Storm Prediction Center operations area. Forecaster Jeff Peters studies data on one of several high-speed workstations. See ftp site for color image.

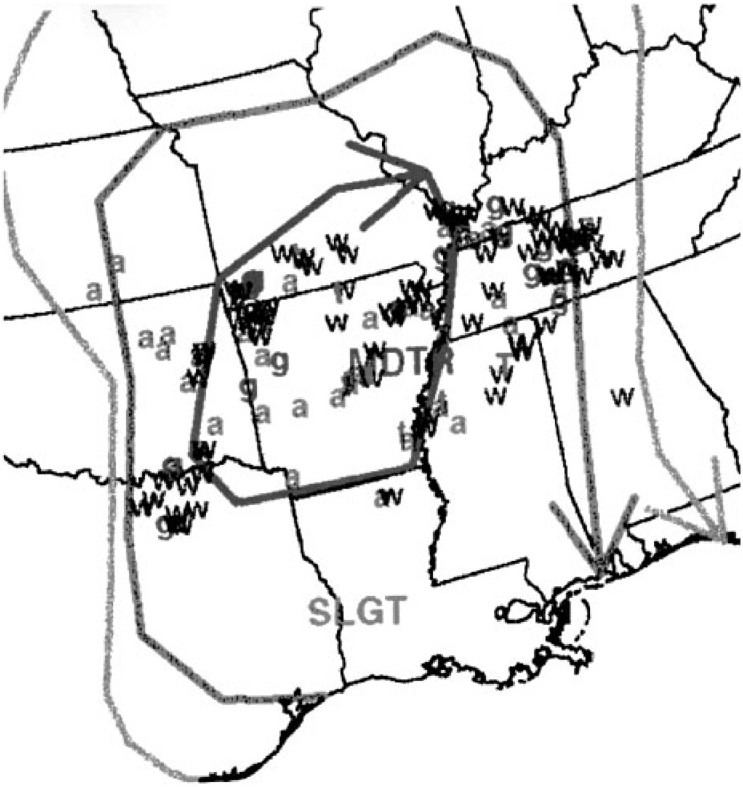


Figure 27 Example of SPC outlook and watch area. See ftp site for color image.

large hail. Tornado watches are issued when there is a threat of tornadoes. There may also be a threat for damaging winds and large hail within a tornado watch. Significant and/or concentrated severe weather typically results from "organized" severe thunderstorms, those that are of the supercell, bow echo, or strong multicell modes.

Generally, forecasting severe thunderstorms involves three concepts: climatology, pattern recognition, and parameter evaluation. Climatology is used by SPC forecasters to know what time of day, season, and area they should expect a higher likelihood of severe weather development. For example, in the Great Plains region of the United States, spring tornadoes are most likely to occur in the late afternoon and evening hours. Given a typical severe weather situation for the region, forecasters anticipate an enhanced risk at that time. In the winter, however, the highest risk of tornadoes in Florida and the coastal regions of the southeastern states is during late night and morning hours, nearly the diurnal opposite of the Plains states in spring. Therefore, given the typical severe weather situation in the southeastern United States in the winter, an SPC forecaster's anticipation of tornado development based on climatology is enhanced for the period from midnight until noon.

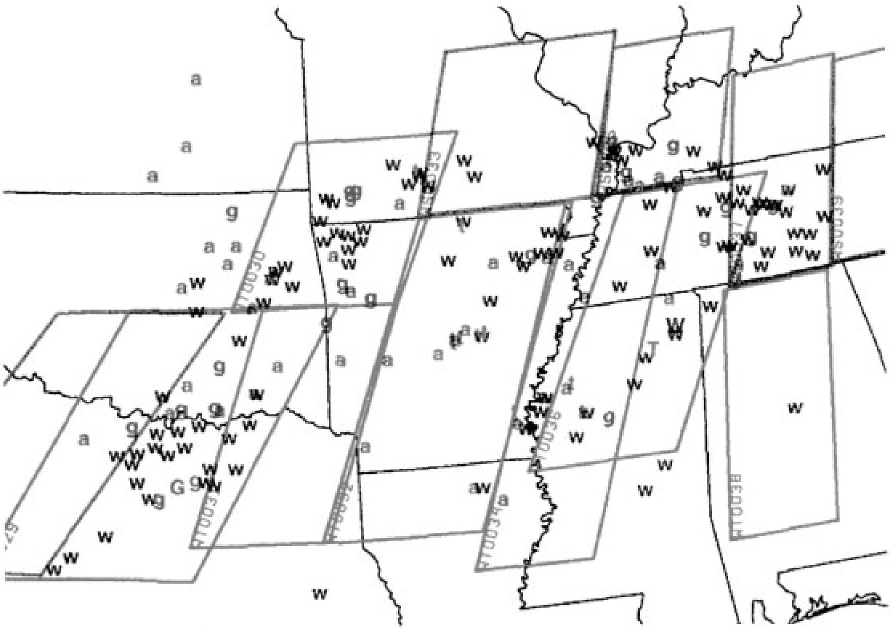


Figure 28 Composite forecast map used by SPC to overlay various surface and upper-level features to help poinpoint potential severe weather threats. See ftp site for color image.

Pattern recognition is used by SPC forecasters as a first approximation concerning severe thunderstorm development. Features such as the upper-level jet stream (20,000 ft or higher above ground level), the low-level jet stream (from 2000 to 5000 ft above ground level), fronts, thermal and moisture axes, etc. are examined. The intensity, orientation, juxtaposition, and movement of these features aid the forecaster in estimating the probability of occurrence, area affected, timing, and severity of potential severe weather episodes. As an example, most large tornado outbreaks are associated with a pattern that includes dual upper jet streams that diverge over the outbreak area and a strong southerly low-level jet stream that transports warm and moist air into the area. The strength and movement of these jet streams play a major role in the severity, area affected, and timing of the outbreak.

Finally, parameter evaluation (also called *ingredients-based forecasting*) has become increasingly important in forecasting severe thunderstorms in recent years. Atmospheric scientists have learned much about thunderstorm development and evolution over the five decades since the first severe weather forecasts were made. This knowledge is derived from observations, theoretical studies, and numerical modeling experiments. Application of this knowledge has resulted in forecast techniques that relate values of meteorological parameters to the type of storms that develops, their evolution, and the types of severe weather (large hail, damaging winds, and/or tornadoes) associated with them. As an example, "isolated" supercells can generally be associated with varying combinations of the amount of vertical wind shear and the degree of buoyancy for rising parcels of air in the troposphere. Given that thunderstorms will develop, SPC forecasters assess the values of these two parameters from both "real-time" observational data and operational numerical forecasts when deciding whether or not to forecast "isolated" supercells. If supercells are predicted, the forecaster then looks at other meteorological parameter values to assess the risk of tornadoes, damaging winds, and/or large hail with the expected supercells.

Making forecasts for severe thunderstorms requires detailed analysis of both real-time observations and operational numerical model forecast data and trends. Since the 1980s, computer workstations have been used to process and display ever-increasing quantities of meteorological data, and their use has helped forecasters gain a better understanding of processes important for severe storm development (Fig. 29). Outlook forecasters rely primarily on analysis of numerical model forecast data for forecasts that extend out as far as 3 days ahead. Composite forecasts of model data are typically constructed for key times within the forecast period (Fig. 30). Adjusting for known model biases and limitations, the forecaster then uses the patterns and positions of features on the composite forecast charts to determine timing and area covered by potential severe storms. Model forecasts of vertical profiles of wind, temperature, and moisture are then examined within the forecast area to help estimate storm mode, intensity, evolution, and severe weather type, if any.

Figure 29 Example of mesoscale analysis used by SPC to determine regions where the threat of severe weather exists. See ftp site for color image.

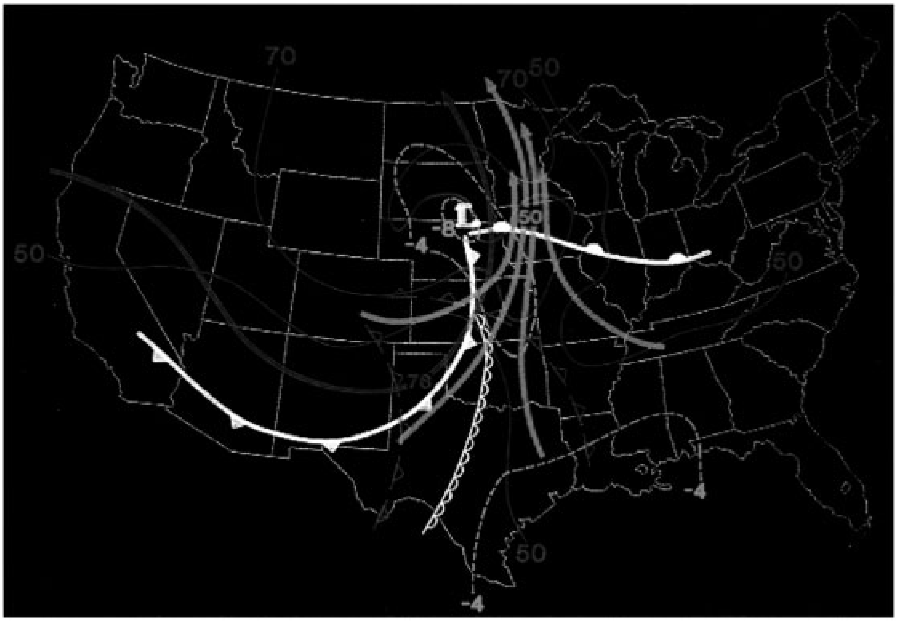


Figure 30 Example of a composite forecast for April 26, 1991. See ftp site for color image.

For short-term forecasts (1 to 7 h ahead), the most recent day 1 outlook forecast (current time to 24 h ahead) is used to focus attention on specific areas. Although short-term operational numerical model trends are noted, the primary emphasis for short-term forecasts is on analyzing real-time data and trends. Real-time data includes both observations (e.g., radar reflectivity and wind data, satellite imagery, lightning strike data, aircraft wind and temperature data, and upper air and surface observations of temperature, moisture, air pressure, and wind direction and speed.) and derived fields computed from this data (e.g., surface pressure changes over time). For timely and accurate short-term products, continuous attention to details and trends is very important because the atmosphere is constantly changing. For example, regional subjective analysis of surface observations (the densest operational data network available) is necessary to assess the short-term severe weather threat (Fig. 31). So, such analysis is typically done each hour. As significant small-scale patterns, parameter values, and trends are diagnosed, mesoscale discussion products consisting of one or two paragraphs are issued. These messages describe the situation and how storms are expected to develop and/or evolve. A severe thunderstorm or tornado watch is issued if the analysis reveals that there is a significant threat of "organized" severe thunderstorm development that will last 4 or more hours and affect more than a localized area (generally greater than 8000 square miles).

Although the ingredients-based approach to severe thunderstorm forecasting (parameter evaluation) has allowed more precision in severe weather forecasting

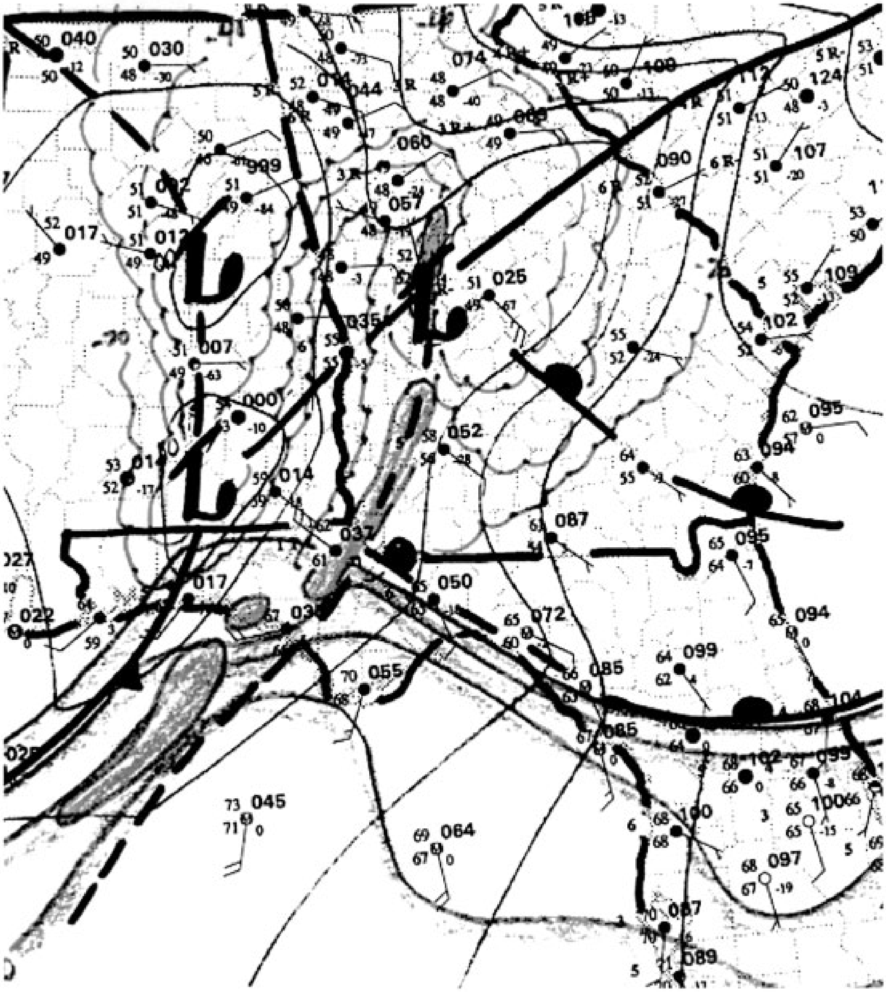


Figure 31 Regional subjective analysis of surface observations. See ftp site for color image.

during recent years, it appears that the climatological and pattern recognition aspects of forecasting will continue to be utilized in the foreseeable future. There has been a rapid increase in knowledge about storm processes in the past several years, but there is still much to learn. Further, despite the increase in the amount of real-time data available to forecasters, the operational data network is still not dense enough to diagnose all parameter values and other features in enough detail to take advantage of some newly understood storm-scale processes.

To complete the tornado warning process, local NWS Warning and Forecast Offices, which have responsibility for much smaller areas, issue very short-range warnings for events that are either occurring or imminent.

FOR FURTHER READING

Bluestein, H. B. (1999). *Tornado Alley. Monster Storms of the Great Plains*, New York, NY: Oxford University Press.

Grazulis, T. P. (2001). *The Tornado: Nature's Ultimate Wind Storm*, Norman, OK: University of Oklahoma Press.

Rinehart, R. E. (1997). *Radar for Meteorologists*, 3rd ed, Columbia, MO: Rinehart.

Vasquez, T. (2000). *Weather Forecasting Handbook*, Garland, TX: WeatherGraphics Technologies.

CHAPTER 30

TROPICAL PRECIPITATING SYSTEMS

EDWARD J. ZIPSER

Precipitation is influenced by phenomena on all scales of motion. In the tropics and subtropics, it is rare to find continuous precipitation on horizontal scales larger than mesoscale, whether related to a larger-scale disturbance or not. The reason is straightforward: The stratification of temperature and moisture is such that the equivalent potential temperature (θ_e) decreases with height in most of the low to midtroposphere. That is, the atmosphere is often both conditionally and convectively unstable, such that large-scale lifting will inevitably result not in slow steady ascent and light precipitation but in convective clouds and possibly heavy precipitation. Adjacent regions normally have subsidence without precipitation, *even within regions of large-scale ascent or disturbances*. This chapter surveys current knowledge of tropical convective and mesoscale precipitation and its organization. We focus first on the physical nature of the precipitation systems themselves, and only later examine the reasons for how those systems are forced, or organized. The organizing systems are then arranged in order of scale, beginning with small-scale orography and land–sea breezes, progressing to large and planetary-scale forcing.

1 DIFFERENCES BETWEEN TROPICAL AND MIDLATITUDE CONVECTION

There are varying perceptions about tropical convection, not always rooted in reality. It is important to examine the basis for statements about tropical phenomena. The old

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

climatological school of thought spoke of the "daily thunderstorm to which one could set one's clock." This belief has a grain of truth in some former British colonial outposts of Malaysia or Africa, some of the time. Knowledge of the oceanic tropics expanded during World War II, leading to descriptions not of boring unchanging climate but of synoptic-scale disturbances such as easterly and equatorial waves, which sometimes intensified into tropical cyclones. Synoptic models of these waves describe useful relationships between phases of the waves and weather. This represented an extension of synoptic meteorology thinking into the tropics, which was helpful in some regions, was irrelevant or misleading in regions where synoptic-scale systems do not control daily events, and mostly ignored mesoscale phenomena.

Knowledge depends upon observations *of appropriate scale*. Motivated by a series of devastating hurricanes in the 1950s, research aircraft penetrations of hurricanes provided such data, leading to major advances in description, understanding, and prediction of tropical cyclones (Marks, F. D. Jr., Chapter 32). In the meantime, quantitative radar and mesoscale data in the severe storm regions of the United States led to analogous knowledge of storms bearing hail and tornadoes (Brooks, H. et al., Chapter 30). It would be nearly 20 years before such tools would be used for "ordinary" tropical weather. The motivating factor was not weather forecasting but the increasing realization that global models of weather and climate required sound treatment of "subgrid-scale" phenomena in the tropics. The conceptual framework for the Global Atmospheric Research Program (GARP) and its Atlantic Tropical Experiment (GATE) was created in the 1960s, under the leadership of Jule Charney, Verner Suomi, and Joseph Smagorinsky. Following a series of smaller field experiments in the late 1960s, the GATE was carried out in the eastern Atlantic in 1974, ending forever any lingering thoughts that large-scale and small-scale phenomena can be treated independently.

The landmark "hot tower" study of Riehl and Malkus (1958) had already conditioned meteorologists to the belief that tropical cumulonimbus clouds were not just decorations, responsible for local showers and the heavy rains that constitute many tropical climates. These convective towers were shown to be a critical link in the general circulation, transporting heat, moisture, and moist static energy from the low to high troposphere for subsequent export to higher latitudes. They also have an essential role in hurricane formation and maintenance. Thus, it becomes easier to accept the truth that tropical convection must be parameterized if global models were to be successful. It perhaps was natural to believe that these hot towers of the deep tropics were also some of the biggest and most powerful storms on the planet, but observations demonstrated otherwise.

Research aircraft equipped to derive vertical velocity made thousands of penetrations of convective clouds, both isolated and embedded in mesoscale convective systems (MCSs, more about these below). Beginning in GATE, but over a 20-year period in other field programs, including the Bay of Bengal, offshore Borneo, offshore northern Australia, offshore Taiwan, the warm pool of the equatorial west Pacific, and tropical cyclones in several oceans, the results were remarkably similar.

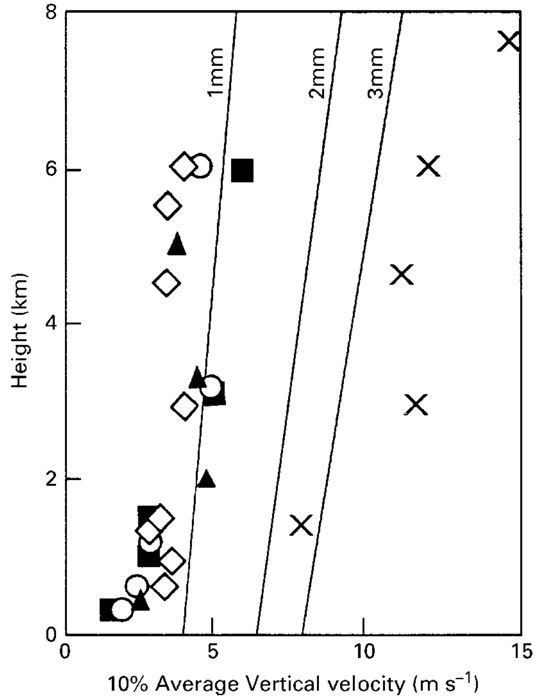


Figure 1 Average vertical velocity in the strongest 10% of updraft cores over tropical oceans from measurements in three different regions (triangles, circles, diamonds) and over land (crosses, Thunderstorm Project). The lines show terminal fall speeds of raindrops as a function of raindrops diameter and height. (after Zipser and Lutz, 1994).

Most tropical convective clouds were weak. Typical maximum updrafts were a factor of 2 to 3 lower than those in ordinary cumulonimbus clouds over land (Fig. 1). Updraft and downdraft velocity, diameter, and mass flux are approximately lognormally distributed.

But it would be wrong to assume that weak convection characterizes the entire tropics. The cited observations are entirely from the *oceanic* tropics. Few data exist from continental tropical regions, not coincidentally because local knowledge of intense convection over land would quickly discourage pilots from attempting penetrations. The true hot towers, therefore, are mainly confined to certain regions, e.g., India–Bangladesh, Argentina, southern United States, and Africa. Recent satellite data show the distribution of intense storms more quantitatively (see Section 7). The common oceanic cumulonimbi still perform their major role of energy transport and precipitation; the vertical speed is simply smaller than previously imagined.

2 MESOSCALE CONVECTIVE SYSTEMS (MCSs)

There is no precise definition of an MCS. The essence of an organized mesoscale convective system is that it is a *recognizable entity of a distinctly larger temporal and spatial scale than its constituent convective clouds*. That is, the MCS lives for several hours and covers a horizontal scale of the order of 100 km in at least one dimension, while the ordinary cumulonimbus cloud may live for one hour and cover the order of 10 km. Many MCSs are unmistakably well organized, such as the squall line or mesoscale convective complex (MCC; Section 4). Some are not.

Mesoscale convective systems are widely distributed throughout the world. The fundamental processes by which groups of cumulonimbus clouds become organized into MCSs, and which govern the structure and evolution of MCSs, are independent of latitude. (The sole exception is the Coriolis effect, which influences the asymmetry and inertial stability of particularly large, long-lived MCSs.) Given the important role of MCSs in the United States, there is some irony in the fact that so much of what we know about them comes from field programs over tropical oceans, transferred to midlatitudes later. The first field program explicitly targeting MCSs over the United States was carried out in Oklahoma and Kansas in 1985.

The literature on MCSs is dominated by studies of squall lines. The reason is the great simplification inherent in a quasi-two-dimensional system. Concentrating on squall lines results in little loss of generality, provided one remembers that more complex or disorganized systems are often more common.

3 SQUALL LINE AND MCS STRUCTURE

The best-known type of squall line has *a leading convective region and a trailing stratiform precipitation region*. It is fairly common in almost all regions, the easiest to study, and therefore the best-known MCS. The convective region is usually 10 to 30 km in width, followed by a region of stratiform precipitation that often exceeds 100 km in width. When such a system passes over a given location, the weather experienced includes a few heavy convective showers (often >25 mm/h) followed by several hours of steady precipitation (often 3 to 5 mm/h). Rainfall from the convective region depends upon details of individual cells, their intensity, and their relation to the observer; from the stratiform region it depends mainly on duration of the moderate rain.

Convective updrafts within MCSs are basically similar to those in more isolated storms but are concentrated in space and time. Updraft speeds rarely exceed 3 to 10 m/s over oceans and 10 to 25 m/s over land. Diameters rarely exceed 5 km except in supercells. The *mesoscale updrafts* occupy the upper half of the troposphere in the stratiform precipitation region, with updraft speeds in the 10 to 100 cm/s range over an area that may extend for 100 km.

The *convective downdrafts* in the convective region form a cold pool at low levels that quickly spreads out to cover a large area. This cold pool can be thought of as one of the exhaust products of the system, analogous to the more easily visible exhaust product of the convective updrafts: the anvil system, which spreads out near and below the level of neutral buoyancy (equilibrium level). The cold pool has the important function of helping to continuously generate new convective cells, usually along its advancing edge (e.g. Houze 1993, Chapters 8 and 9). The interior of the cold pool is usually sufficiently stabilized that new deep convective cell growth is impossible. Even over warm tropical oceans, while heat and moisture transfer can restore the boundary layer properties and make new convection potentially possible, another element of the MCS usually prevents it: the mesoscale downdraft.

The *mesoscale downdraft* is universally found below the melting level in the stratiform precipitation region, beneath the mesoscale updraft. The evaporation of precipitation fails to keep the mesoscale downdraft saturated, and relative humidity is often below 80% and can be as low as 50% in the 1 to 2 km levels. Therefore, in spite of thick anvil-type clouds above, steady rain, and occasionally lightning, there are usually no clouds at all below the melting layer. All precipitation in this region was initially frozen before falling through the melting layer. The mesoscale downdraft cannot penetrate through the shallow cold pool to the surface, so can be detected by aircraft or soundings (Fig. 2). The result is a characteristic onion-shaped sounding (Fig. 3). Under most circumstances, the combination of a shallow, capped cold pool and a warm, dry unsaturated downdraft dictates that the air in the stratiform region of an MCS cannot generate new deep convection for at least 12 to 24 h.

Mesoscale downdrafts have been attributed to three causal mechanisms, and all three mechanisms act together in most MCSs. The latent heat of fusion cools the air when ice melts. Although the magnitude is only 13% of the latent heat of evaporation, the cooling takes place in a concentrated layer only a few hundred meters thick, having the effect of lowering the 0°C isotherm on the mesoscale, generating or accelerating descent. The evaporation of falling precipitation is a powerful cooling process, assuming subsaturation of the ambient air. Once descent begins for any reason, however, evaporation can be effective. A final reason is that the cold pool near the surface always tends to diverge, so air above the cold pool must descend.

In both tropics and midlatitudes, some 40% of rainfall from MCSs falls in the stratiform precipitation region. The question arises: Which process is most important, the condensation/sublimation growth of ice particles in the strong ascent in the mesoscale updraft, or horizontal transfer of convective debris (ice) from the convective region? The answer is that both are required for substantial precipitation rates and that both are present in MCSs. While this result comes from simulation experiments, the remarkable fact is that *no documented case exists of an MCS cloud structure without prior existence of deep convection*. Some earlier descriptions of rainfall systems have mistakenly ascribed the existence of widespread light precipitation to midlevel convergence in subtropical cyclones or monsoon depressions. This is very doubtful; the misconception arises when synoptic-scale reasoning is applied to situations where mesoscale processes are responsible for the specific cloud and precipitation features.

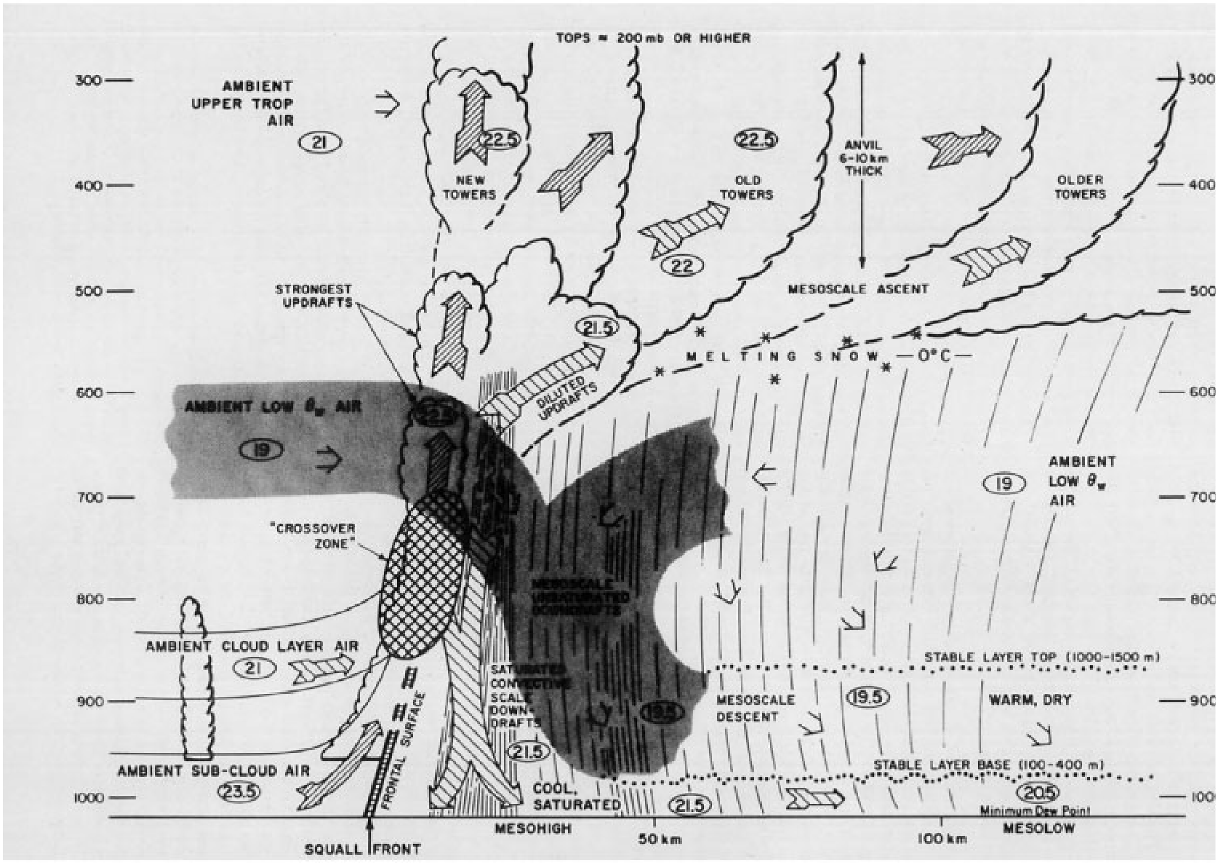


Figure 2 Schematic cross section through a typical squall line system. All flow is relative to the squall line, which is moving from right to left. Circled numbers are typical values of wet-bulb potential temperature, directly related to equivalent potential temperature (in °C). The convective region occupies the first 30 km behind the leading edge, and stratiform precipitation the next 100 km. The air descending in the mesoscale downdraft remains above the cooler air, which descended in the convective downdrafts and spread out in the lowest layers (after Zipser, 1977).

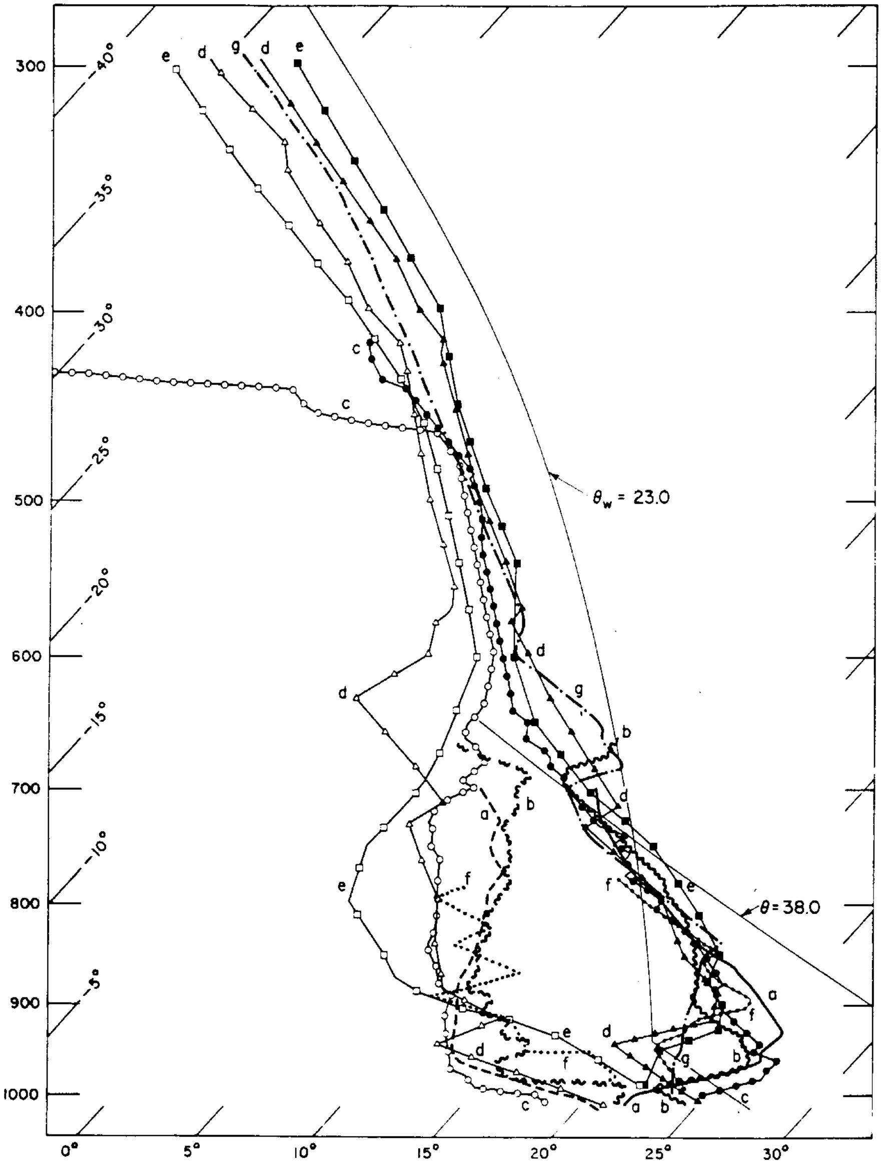


Figure 3 Soundings shown are taken behind squall lines, mostly within or toward the rear of the stratiform precipitation region. The temperature and dew-point curves are far apart in the mesoscale downdraft air below the melting level and near the surface, signifying low relative humidity even in the rain area (after Zipser, 1977).

4 SQUALL LINES, MCCs, OR DISORGANIZED MCSs

What are the environmental conditions that favor MCSs organizing in the form of squall lines? Empirically and theoretically (and confirmed in simulations) the determining factor is the low-level wind shear (LeMone et al., 1998). Substantial low-level wind shear usually results in convection rapidly organizing into lines perpendicular to the shear, moving downshear while leaning upshear and transporting line-perpendicular momentum opposite to the direction of motion (upshear).

What are the environmental conditions that favor MCCs (mesoscale convective complexes)? These are the largest and longest-lived of the MCS family. The question is what favors extremely large and persistent relative inflow of high θ_e air into the system at low levels? Local or purely orographic effects are unable to provide such inflow, which helps to explain why MCCs are often associated with a low-level jet stream. Also, there is often a region of forced ascent for the low-level jet, either a synoptic-scale front, orography, or large cold pool left behind by previous MCSs/MCCs (Laing and Fritsch, 2000). Squall lines and MCCs are not mutually exclusive. Only the criterion of approximately circular shape prevents many large squall lines from satisfying the other MCC threshold criteria.

5 EVOLUTION OF MCSs

All MCSs that have been documented in case studies evolve in much the same manner (Houze, 1993). A group of cumulonimbus clouds forms fairly close to one another for a variety of reasons. The convective rain volume often peaks rapidly but occasionally can persist at a high level for many hours (the typical MCC precursor). Stratiform precipitation usually rises in phase with the convective precipitation but displaced a few hours later in time. If the system is weak or short-lived, there may be little overlap in time between convective and stratiform regions—one may appear to evolve into the other. For strong, long-lived systems, including squall lines and MCCs, stratiform and convective regions may coexist in relative steady state for many hours. The end stages of MCSs inevitably consist of decaying stratiform-only regions.

6 CANDIDATE SYSTEMS THAT ORGANIZE RAINFALL AND MAIN CONTRIBUTORS TO RAINFALL IN SELECTED REGIONS

Organization of precipitation systems is governed by disturbances with a variety of horizontal length scales. The main requirement is that they include regions where ascending motion is organized. The *frequency and amount* of precipitation in a region is often determined by the nature of the disturbances, while the *character and intensity* of the rainfall system (e.g., strength of convection, MCS or not) is often determined by the local wind shear and thermodynamic profiles. There is too much variety to cover the entire range of tropical environments, so selected illustrative examples are given.

As a general rule, there tends to be an inverse correlation between total monthly or annual rainfall, and the intensity of the rainfall systems. For example, the heaviest monsoon rains over land, and the heavy rains of the oceanic intertropical convergence zones (ITCZ), often come without lightning and with poorly organized MCSs. Where are the most violent storms in the tropics? This may seem counterintuitive, but they are often on the fringes of deserts or during the premonsoon seasons or "break monsoon" periods. Some examples are given below.

Orographic Forcing (1) Strong flows of moist low-level air capped by warm dry air. The classic example is the Hawaiian Islands, where persistent, steady trade winds are forced to ascend the windward slopes. Annual rainfall in the adjacent oceans is probably <300 mm, yet parts of the windward slopes of most islands experience >300 mm per month. In the extreme case of Mt. Waileale on Kauai, annual rainfall exceeds 10 m. These orographic rains are not at all steady but consist of shallow convective clouds that produce periodic heavy showers. Other regions with similar orographic enhancement include Jamaica and Puerto Rico, January–May.

Orographic Forcing (2) Weak flow, with the moist layer weakly capped. This distinctly different kind of orographic enhancement is triggered by the heating of elevated terrain, which, in turn, forces a diurnal upslope flow. The organized mesoscale flow often triggers thunderstorms, which may be closely spaced and in turn organize into MCSs or MCCs, occasionally propagating off the mountains as organized squall lines. The Sierra Madre Occidental of western Mexico is the site of a classic example. These mesoscale rain events extend northward into southeastern Arizona during July and August, sometimes referred to as a monsoon. Rainfall can exceed 500 mm monthly in Mexico and 200 mm monthly in southeastern Arizona during this regime. Many other parts of the world experience similar regimes, including the tropical Andes, parts of east Africa, mountainous islands in the Indonesian region, and Puerto Rico in summer.

Land–Sea Breeze Circulation Systems Conditions for effective rain production from sea breeze circulations are similar to that for orographic forcing (2) above. Florida in summer is a classic example. Adjacent oceans have little rain except for tropical disturbances, while peninsular Florida generates some 200 mm rain and 20 to 25 thunderstorms per month, some of which can be locally severe, and/or organize into MCSs, including squall lines. It is obvious from satellite and surface data that these are generated by the ascent along sea-breeze convergence zones from either coastline, and numerous case studies of these storms have been undertaken.

Synoptic-Scale Waves The best known is the easterly wave. These may be the most regular synoptic waves on Earth, generated in Africa, coasting across the Atlantic into the eastern Pacific. They are less common and regular in the Pacific. They have a period of 4 to 5 days, a wavelength about 2500 km, propagating westward at about 6 to 8 m/s. Some 10% of these waves encounter favorable conditions for intensification into tropical cyclones. They modulate convection and MCSs strongly, with the wave troughs wetter than the wave ridges (Reed et al., 1977); in this way they are analogs of traveling waves in the higher-latitude westerlies.

The northern half of an easterly wave usually has a strong midlevel easterly jet that dictates squall line formation, while the southern half of the same waves may generate heavily raining but poorly organized MCSs in the light wind shear regime. Also of importance is the relative motion of the MCS and the system that may have generated it. The squall line moves at roughly twice the wave speed; therefore, it forms preferentially near the wave trough but tends to dissipate as it approaches the wave ridge. This resembles the case of MCSs in the United States, where it is common for the system to move eastward more rapidly than its larger-scale forcing mechanism. If it moves into an unfavorable area, it may dissipate rapidly. If it can regenerate convection on its right rear flank, it may not only live for a long time but also move slowly enough to be a serious flash flood threat (Chappell, 1986). Excessive rain events in the tropics are rarely attributable to a simple wave passage but to unusual interactions among systems of different scales.

Many other regions of the tropics are affected by westward traveling disturbances. Most of these do not fit the description of easterly waves but can be Rossby waves or Rossby-gravity waves. Their existence is well-documented by satellite data analyses, but case studies are rarely undertaken due to almost complete lack of appropriate in situ data.

Cyclones Other than tropical cyclones (Marks, F. D., Jr., Chapter 32), there are a variety of other cyclones that can organize tropical rainstorms. Once again, their existence is well known but definitive case studies are quite rare. They are often categorized by whether the maximum circulation (vorticity) is found in the upper, mid, or low troposphere.

The tropical upper tropospheric trough (TUTT) is a semipermanent feature of the summer hemisphere in the north Atlantic and north Pacific. It is a subtropical feature and generally exists over dry midtropospheric air masses. Upon occasion, some of the common low-pressure centers along this trough become strong enough to organize deep convection and MCSs, usually in the subtropical west Atlantic and west Pacific. It is possible that such activity requires the low to extend downward from its usual upper tropospheric location.

Midtropospheric cyclones may exist in the vicinity of Hawaii, where they are known as Kona lows, and they can generate heavy rainfall events, especially when they move slowly (Barnes, 2001). Similar cyclones have been documented in the Indian Ocean. It is not clear how often these cyclones propagate downward into the low levels to form monsoon depressions, as they are called, or whether the latter are generated independently in low levels. In any event, some of the heaviest rainfalls of the Indian subcontinent are produced in association with the passage of these depressions. They are most common in the Bay of Bengal and frequently move over northern India from there. They are not to be confused with tropical cyclones, which are unable to survive in the strong wind shear regime that dominates south Asia during the heart of the summer monsoon (strong low-level westerlies *and* strong upper-level easterlies).

Intertropical Convergence Zone: Oceans It has been known, literally for centuries, that the trade-wind flows of the Northern and Southern Hemispheres converge along rather narrow zones in the oceans, known as the ITCZ, along which synoptic-scale ascent must take place on average (not at all times). Rain systems are frequent along the oceanic ITCZ, although without great convective intensity. Typical rainfall totals are about 300 mm monthly; where the ITCZ stays over a given location for many months of the year, annual totals may exceed 3 m. There is a large-scale convergence zone extending southeastward from New Guinea toward Tahiti known as the South Pacific convergence zone, within which the rainfall may have similar properties to the ITCZ.

Intertropical Convergence Zone: Land, Monsoons The off-equatorial heating of the continents forces much stronger cross-equatorial flows well into the summer hemisphere. At very low levels, the ITCZ as defined by wind convergence often moves far from the equator, and the low-level flow of moist air attempts to follow suit. However, the deep convection and heavy rainfall does not usually reach the location of this ITCZ but is distributed over a large region.

A prime example is north Africa in August. The wind convergence marking the ITCZ at the surface appears to reach the central Sahara Desert, while the heaviest rainfall remains well to the south, along about 10°N. This is a region of strong temperature gradient, connected with the midlevel easterly jet and the generation of easterly waves. Very strong thunderstorms and squall lines form in the desert margins near 20°N, modulated by the waves, but rare in any one location. The Sahel zone near 15°N is in the strong rainfall gradient, with heavy rainfall during the July–September monsoon season when the large-scale flow brings in moist air from the southwest at low levels and the waves and squall lines are strongest (Zipser, 1994).

Short monsoon seasons are also experienced in western North America, noted above, and northern Australia in December to March. The latter has been the subject of several field programs near Darwin and the Tiwi Islands just to the north. The thunderstorms over these islands are so common in season that Darwinites call them by name (Hector). As in west Africa, the surface convergence zone extends well beyond the area of heaviest rain (near 10°S), into interior Australia. Near Darwin, research has established that the main modulation of these storms and rain systems is not by synoptic-scale waves but by longer period oscillations in the flow (also known as the MJO, see below). During monsoonal low-level westerly wind periods, rainfall is heavy, oceanic in character, and tropical cyclones may form. During the "break" periods, westerlies weaken or become easterlies, and rainfall decreases, but is concentrated into strong thunderstorms and squall lines moving from the continent (Rutledge et al., 1992).

The classic Asian monsoon affects billions of people and requires far more research before its rain systems have been properly described and understood, since few field programs have been undertaken with observations of appropriate scale. It is well known that the southwest monsoon flow is from ocean to continent during northern summer over India, China, and all of Southeast Asia. There is great variety of local and regional details, well beyond our scope here to describe. Over India, there is a characteristic alternation between rainy and break periods on a scale of weeks. Where persistent moist flow is forced to ascend rapidly, as over the hills of Assam in northeast India, world record rainfalls have been recorded.

Madden–Julian Oscillation (MJO) First discovered by Madden and Julian (1972, 1994), there is often a strong variability in tropical winds and rainfall on this intraseasonal time scale of 30 to 60 days. While many wave modes may come into play, the most significant one appears to be an equatorially trapped Kelvin wave (Meehl, G., Chapter 5). It is strongest within 10° of the equator (but its effects extend beyond), and from the western Indian Ocean eastward through the central Pacific Ocean (but its effects extend beyond). In these regions, it is considered normal for rainfall to be excessive for a few weeks and deficient for the next few weeks. The break period has anomalous easterlies at low levels and westerlies at high levels, while the rainy period has westerly wind anomalies at low levels and easterly anomalies at upper levels. During the rainy west wind periods near the equator, tropical cyclones may form on either side of the equator, sometimes on both sides (double vortex). Periods with especially strong west winds are known as westerly wind bursts. These phenomena have been extensively studied during a major field program in 1992–1993 (Godfrey et al., 1998).

7 DISTRIBUTION OF TROPICAL/SUBTROPICAL RAINFALL AND MESOSCALE RAIN SYSTEMS

Maps of annual rainfall totals in the tropics are often accuracy challenged but show the same major features. The “doldrums” or ITCZ regions near the meteorological equator are very rainy; the subtropical high-pressure regions and trade wind regions are generally dry except where orographic lifting occurs. Interestingly, the regions of heaviest rainfall, and the regions with numerous large, strong MCSs, are often quite different.

Total Rainfall Distribution

Figure 4 can be considered a “best estimate” of total rainfall in the tropics, using a combination of satellite and rain gauge data. The rainfall maps include the region covered by the *Tropical Rain Measuring Mission* (*TRMM*) satellite (36°N–36°S) and include one month from each season in 1998–1999. During the first 4 months of 1998 one of the strongest warm episodes of El Niño–Southern Oscillation (ENSO) occurred, and for the next 12 months one of the strongest cold episodes. The rainfall anomalies associated with these events are clearly seen in the expected areas: equatorial central and east Pacific (rainy in January and April 1998, dry in January and

April 1999); Indonesia (rainier in January and April 1999 than in the same months of 1998).

Despite these unusually strong anomalies, these rainfall maps clearly show the "normal" seasonal cycle of rainfall over the tropics and subtropics. Asia is dry in winter, wet in the summer monsoon, with similar extent in both years (despite different amounts in some regions). African rainfall migrates north and south of the equator with season. The oceanic ITCZ migrates very little in the tropical Atlantic and Pacific with the seasons.

MCC Distributions Using Satellite Infrared (IR) Measurements

Laing and Fritsch (1997) have systematically mapped out the regions subject to MCCs throughout the globe (Fig. 5). There are concentrations of MCCs in the central United States, Panama–Columbia, South America east of the Andes from 15° to 35°S, Africa between 0° and 15°N, northeast India–Bangladesh, and lesser concentrations in China and northern Australia. Laing and Fritsch (2000) show how the environments of MCCs in these regions vary, but the similarities are striking: a low-level jet bringing high θ_e air, and an approaching disturbance (which can be weak) providing ascent, triggering intense convection downwind of elevated terrain.

Some tropical regions are notable for high rainfall, but a near absence of MCCs. These include the entire oceanic ITCZ belt, equatorial South America, Southeast Asia, and the entire Maritime continent (Indonesian Archipelago). These regions are characterized by persistent deep and high cloudiness, and by low values of *average* outgoing longwave radiation (Laing and Fritsch, 1997). The contrast between the Amazon and Congo basins is especially striking. McCollum et al., (2000) have recently noted that the satellite estimates of rainfall are biased low in equatorial Africa but are approximately correct in the Amazon, for reasons yet to be determined, but which they speculate are related to differing properties of MCSs and their microphysics.

MCS Distributions Using Passive Microwave Measurements

Passive microwave radiances at 37 and 85 GHz can be used to identify, categorize, and map MCSs. The principle used to diagnose deep precipitating convection is that ice particles are efficient scatterers but poor absorbers and emitters of radiation at frequencies in this range. Therefore, regions with large depth of precipitation-sized ice particles, especially graupel, have low brightness temperatures compared with their surroundings. This approach has the advantage that precipitation particles are sensed directly, compared with IR techniques that sense only the cloud tops, even if they consist mainly of nonprecipitating cirrus. Data are available at microwave frequencies from a series of SSM/I Special Sensor Microwave Imager (SSM/I) instruments on DMSP satellites, since 1987, and on *TRMM* since 1997. The main disadvantage compared to IR techniques is that they are not yet available at high temporal frequencies from geosynchronous orbits.

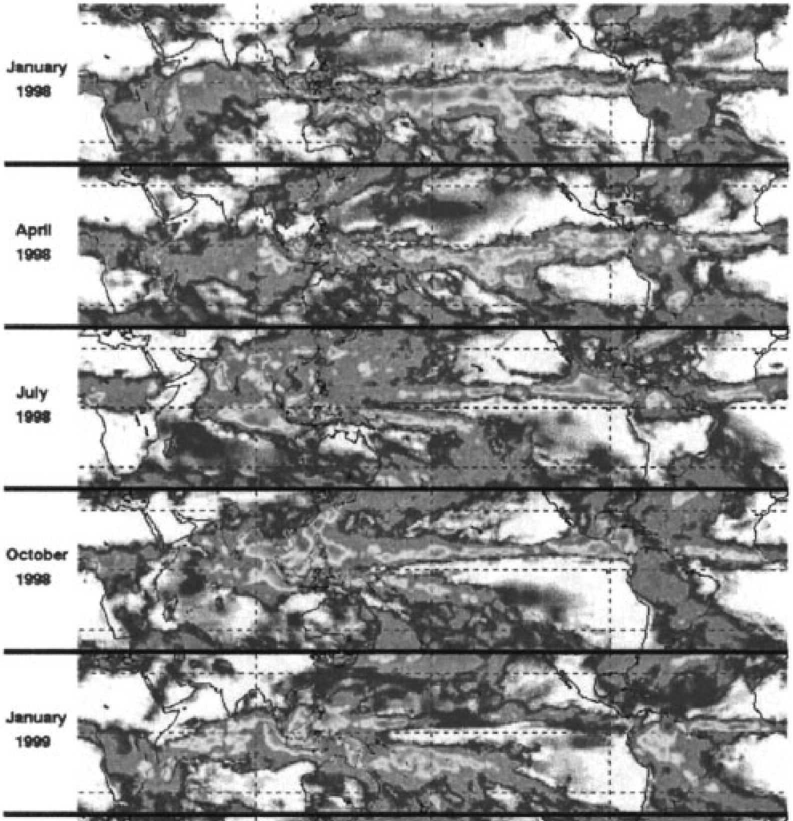


Figure 4 Average precipitation (mm/day) for selected months of 1998–1999 for the *TRMM*-merged analysis (Adler et al., 2000).

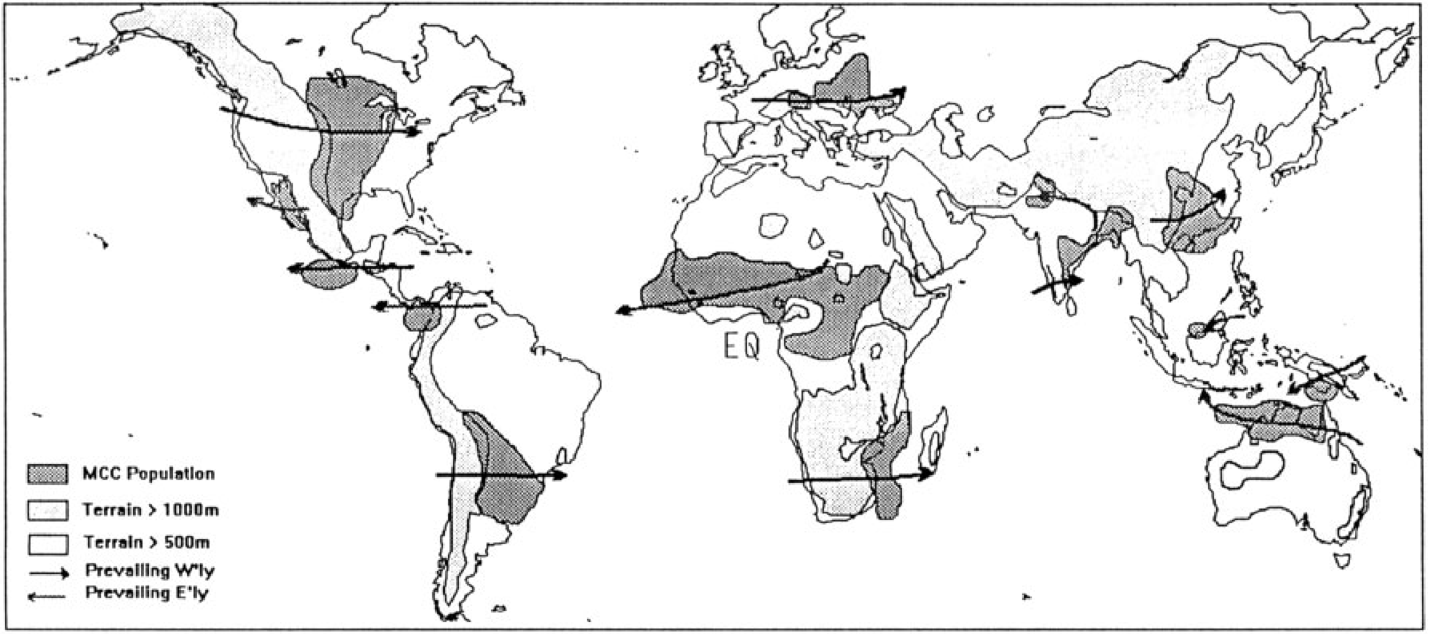


Figure 5 Relationship among mesoscale convective complex (MCC) population centers, elevated terrain, and prevailing midlevel flow (Laing and Fritsch, 1997).

Mohr and Zipser (1996*a*,*b*) mapped MCSs by size and by “intensity,” defined by the minimum brightness temperature (i.e., by the pixel with largest optical depth of ice in a given MCS). They are mapped by season (Fig. 6) and by sunrise vs sunset observation time (not shown). The resulting maps give some additional insights into the properties of precipitating systems in different parts of the world. Unlike MCCs,

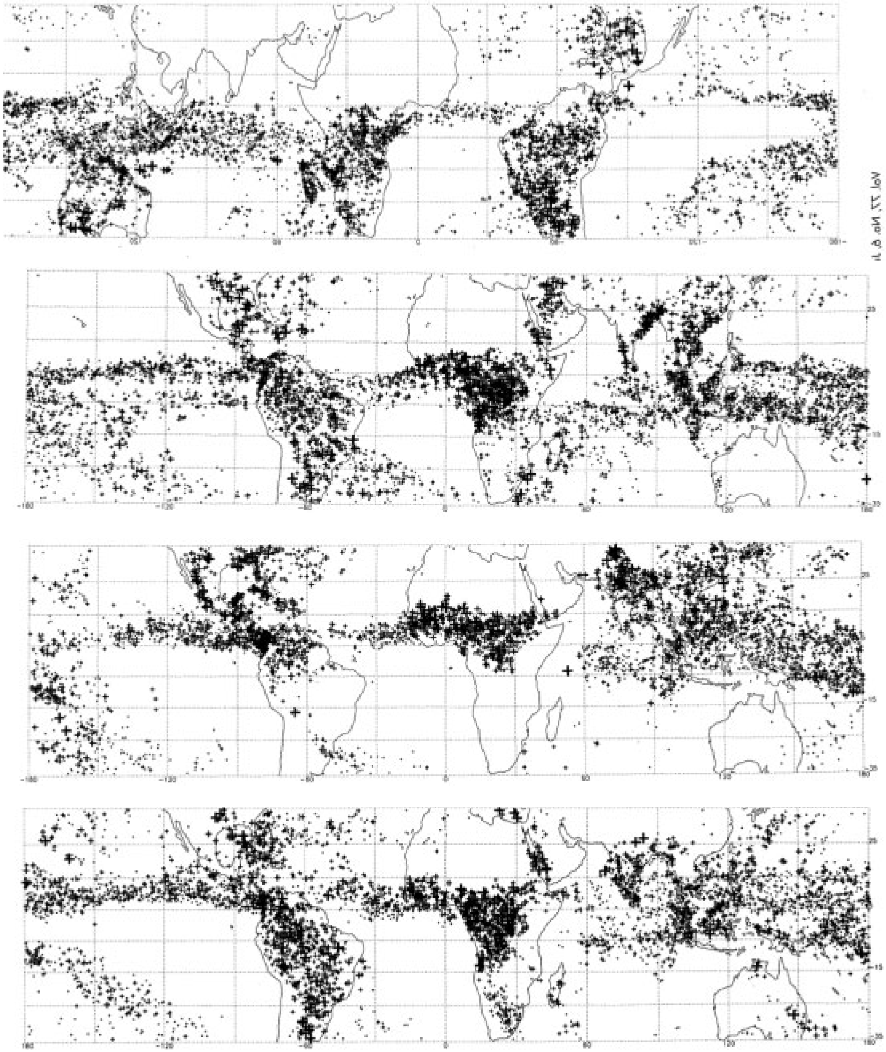


Figure 6 January, April, July, and October, 1993, top to bottom distribution of MCSs according to minimum (polarization-corrected) brightness temperatures. Symbols increase in size and thickness with decreasing temperature, which represents more intense convective pixels. The coldest range, <120 K, has the largest and thickest cross (after Mohr and Zipser, 1996a).

which are nearly absent in tropical oceans, the high rainfall regions of the oceans have abundant (ice-scattering) MCSs, including numerous MCSs with large areas. The MCSs with large (ice-scattering) intensity, however, favor land areas, especially at sunset, a time when growth of MCSs from intense thunderstorms is common. Oceanic regions have more MCSs at sunrise than at sunset.

Using Lightning Measurements

Satellite-derived maps of lightning discharges have been used increasingly in studies of global distributions of thunderstorms. Since 1995 such data have been available from a near-polar orbit, and since 1997 on *TRMM*. Research using these data is relatively new, and comparative studies of lightning and other databases only beginning. The distribution of lightning is even more skewed than the distribution of rainfall, with a relatively small number of storms producing a disproportionate fraction of total lightning. Lightning mapping clearly demonstrates a scarcity of lightning over tropical oceans, resembling the MCC distributions to some extent (Fig. 7). Comparing the lightning, MCS, and MCC distributions, one could hypothesize that the lightning distribution faithfully represents the distribution of strong MCSs and MCCs, which are common over continents and rare over oceans. Recent research that attempts to test this hypothesis indicates that it fails. That is, MCSs over land and ocean with similar brightness temperatures and

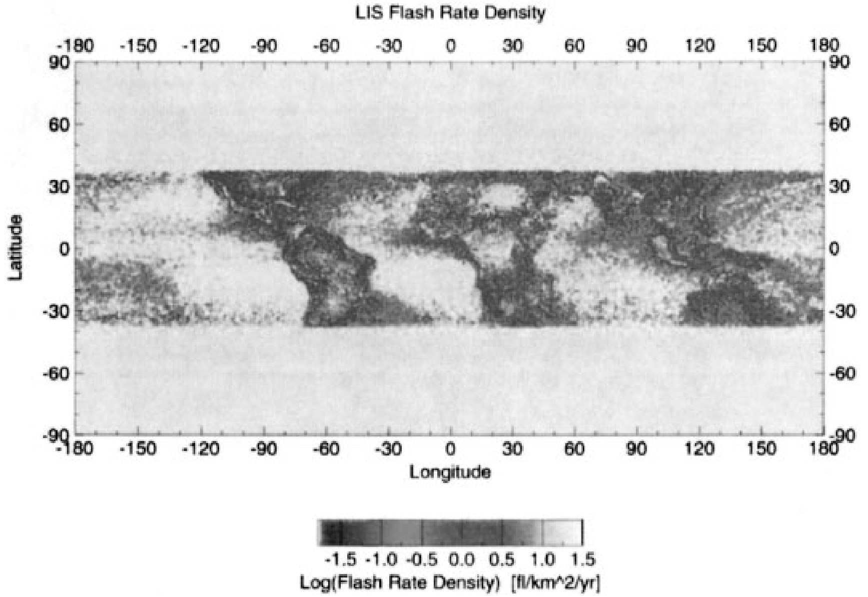


Figure 7 Climatological lightning imaging sensor (LIS instrument on *TRMM*) flash rate density, Dec. 1997–Nov. 1999. Data have been normalized by sensor view time and an assumed detection efficiency of 75% (after Boccippio et al., 2000).

similar radar echoes (from *TRMM*) still have higher lightning frequency over land, for reasons that are under investigation but must involve details of the microphysics.

REFERENCES

Adler, R. F., G. J. Huffman, D. V. Bolvin, S. Curtis, and E. J. Nelkin (2000). Tropical rainfall distributions determined using TRMM combined with other satellite and rain gauge information, *J. Appl. Meteor.* **39**, 2007–2023.

Barnes, G. M. (2001). Severe weather in the tropics, in *Severe Convective Storms*, *Meteor. Monographs*, **28**(50), C. Doswell (Ed.)., Am. Meteor. Soc., Boston MA.

Boccippio, D. J., S. J. Goodman, and S. Heckman (2000). Regional differences in tropical lightning distributions, *J. Appl. Meteor.* **39**, 2231–2248.

Chappell, C. F. (1986). Quasi-stationary convective events. Mesoscale analysis and forecasting, P. S. Ray (Ed.), *Am. Meteor. Soc.* 289–310.

Godfrey, J. S., R. A. Houze, Jr., R. H. Johnson, R. Lukas, J.-L. Redelsberger, A. Sumi, and R. Weller (1998). Coupled Ocean–Atmosphere Response Experiment: An interim report, *J. Geophys. Res.* **103**(C7), 14,395–14,450.

Houze, R. A., Jr. (1993). *Cloud Dynamics*, Academic Press, San Diego, CA.

Laing, A. G., and J. M. Fritsch (1997). The global population of mesoscale convective complexes, *Quart. J. Roy. Meteor. Soc.* **123**, 389–405.

Laing, A. G., and J. M. Fritsch (2000). The large-scale environments of the global populations of mesoscale convective complexes, *Mon. Wea. Rev.* **128**, 2756–2776.

LeMone, M. A., E. J. Zipser, and S. B. Trier (1998). The role of environmental shear and CAPE in determining the structure and evolution of mesoscale convective systems during TOGA COARE, *J. Atmos. Sci.* **55**, 3493–3518.

Madden, R. A., and P. R. Julian (1972). Description of global-scale circulation cells in the Tropics with a 40–50 day period, *J. Atmos. Sci.* **29**, 1109–1123.

Madden, R. A., and P. R. Julian (1994). Observations of the 40–50 day tropical oscillation—A review. *Mon. Wea. Rev.,* **122**, 814–837.

McCollum, J. R., A. Gruber, and M. B. Ba (2000). Discrepancy between gauges and satellite estimates of rainfall in equatorial Africa, *J. Appl. Meteor.* **39**, 666–679.

Mohr, K. I., and E. J. Zipser (1996*a*). Defining mesoscale convective systems by their ice scattering signature, *Bull. Am. Meteor. Soc.* **77**, 1179–1189.

Mohr, K. I., and E. J. Zipser (1996*b*). Mesoscale convective systems defined by their 85 GHz ice scattering signature: Size and intensity comparison over tropical oceans and continents, *Mon. Wea. Rev.* **124**, 2417–2437.

Reed, R. J., D. C. Norquist, and E. E. Recker (1977). The Structure and Properties of African Wave Disturbances as Observed During Phase III of GATE, *Mon. Wea. Rev.* **105**, 317–333.

Riehl, H., and J. S. Malkus (1958). On the heat balance in the equatorial trough zone, *Geophysica* **6**, 503–538.

Rutledge, S. A., E. R. Williams, and T. D. Keenan (1992). The Down Under Doppler and Electricity Experiment (DUNDEE): Overview and preliminary results, *Bull. Am. Meteor. Soc.* **73**, 3–16.

Zipser, E. J. (1977). Mesoscale and convective-scale downdrafts as distinct components of squall-line structure, *Mon. Wea. Rev.* **105**, 1568–1589.

Zipser, E. J. (1994). Deep cumulonimbus cloud systems in the tropics with and without lightning, *Mon. Wea. Rev.* **122**, 1837–1851.

Zipser, E. J., and K. Lutz (1994). The vertical profile of radar reflectivity of convective cells: A strong indicator of storm intensity and lightning probability? *Mon. Wea. Rev.* **122**, 1751–1759.

CHAPTER 31

HURRICANES

FRANK D. MARKS, Jr.

1 INTRODUCTION

The term *hurricane* is used in the Western Hemisphere for the general class of strong tropical cyclones, including western Pacific typhoons and similar systems, known simply as cyclones in the Indian and southern Pacific Oceans. A tropical cyclone is a low-pressure system that derives its energy primarily from evaporation from the sea in the presence of 1-min sustained surface wind speeds $>17\,\mathrm{m/s}$ and the associated condensation in convective clouds concentrated near its center. In contrast, midlatitude storms (low-pressure systems with associated fronts) primarily get their energy from the horizontal temperature gradients that exist in the atmosphere. Structurally, the strongest winds in tropical cyclones are near Earth's surface (a consequence of being "warm core" in the troposphere), while the strongest winds in midlatitude storms are near the tropopause (a consequence of being "warm core" in the stratosphere and "cold core" in the troposphere). Warm core refers to being warmer than the environment at the same pressure surface.

A tropical cyclone with the highest sustained wind speeds between 17 and 32 m/s is referred to as a tropical storm, whereas a tropical cyclone with sustained wind speeds $\geq 33\,\mathrm{m/s}$ is referred to as a hurricane or typhoon. Once a tropical cyclone has sustained winds $\geq 50\,\mathrm{m/s}$, it is referred to as a major hurricane or supertyphoon. In the Atlantic and eastern Pacific Oceans hurricanes are also classified by the damage they can cause using the Saffir–Simpson scale (Table 1).

The Saffir–Simpson scale categorizes hurricanes on a scale from 1 to 5, where category 1 hurricanes are the weakest, and category 5 the most intense. Major hurricanes correspond to categories 3 and higher. The reasons that some disturbances intensify to a hurricane, while others do not, are not well understood. Neither

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

TABLE 1 Saffir–Simpson Scale of Hurricane Intensity

Category	Pressure (hPa)	Wind (m/s)	Storm Surge (m)	Damage
1	>980	33–42	1.0–1.7	Minimal
2	979–965	43–49	1.8–2.6	Moderate
3	964–945	50–58	2.7–3.8	Extensive
4	944–920	59–69	3.9–5.6	Extreme
5	<920	≥ 70	≥ 5.7	Catastrophic

is it clear why some tropical cyclones become major hurricanes, while others do not. Major hurricanes produce 80 to 90% of the U.S. hurricane-caused damages despite accounting for only one-fifth of all landfalling tropical cyclones. Only three category 5 hurricanes have made landfall on the U.S. mainland (Florida Keys, 1935, Camille, 1969, and Andrew, 1992). Recent major hurricanes to make landfall on the United States were Hurricanes Bonnie and Georges in 1998, and Bret and Floyd in 1999.

As with large-scale extratropical weather systems, the structure and evolution of a tropical cyclone is dominated by the fundamental contradiction that while the airflow within a tropical cyclone represents an approximate balance among forces affecting each air parcel, slight departures from balance are essential for vertical motions and resulting clouds and precipitation, as well as changes in tropical cyclone intensity. As in extratropical weather systems, the basic vertical balance of forces in a tropical cyclone is hydrostatic except in the eyewall, where convection is superimposed on the hydrostatic motions. However, unlike in extratropical weather systems, the basic horizontal balance in a tropical cyclone above the boundary layer is between the Coriolis *force* (defined as the horizontal velocity, v, times the Coriolis parameter,* f), the centrifugal force (defined as v^2 divided by the radius from the center, r), and the horizontal pressure gradient force. This balance is referred to as gradient balance, where the Coriolis and centrifugal force are both proportional to the wind speed. Centrifugal force is an apparent force that pushes objects away from the center of a circle. The centrifugal force alters the original two-force geostrophic balance and creates a nongeostrophic gradient wind.

The inner region of the tropical storm, termed the cyclone *core*, contains the spiral bands of precipitation, the eyewall, and the eye that characterize tropical cyclones in radar and satellite imagery (Fig. 1). The primary circulation—the tangential or swirling wind—in the core becomes strongly axisymmetric as the cyclone matures. The strong winds in the core, which occupies only 1 to 4% of the cyclone's area, threaten human activities and make the cyclone's dynamics unique. In the core, the local Rossby number[†] is always >1 and may be as high

*f is the Coriolis parameter ($f = 2\Omega \sin \phi$), where Ω is the angular velocity of Earth (7.292×10^{-5} s^{-1}) and ϕ is latitude. The Coriolis parameter is zero at the equator and 2Ω at the pole.

[†] The Rossby number indicates the relative magnitude of centrifugal (v^2) and Coriolis (fv) accelerations, $\mathrm{Ro} = V/fr$, where V is the axial wind velocity, r the radius from the storm center, and f the Coriolis parameter. An approximate breakdown of regimes is: $\mathrm{Ro} < 1$ geostrophic flow; $\mathrm{Ro} > 1$ gradient flow; and $\mathrm{Ro} > 50$ cyclostrophic flow.

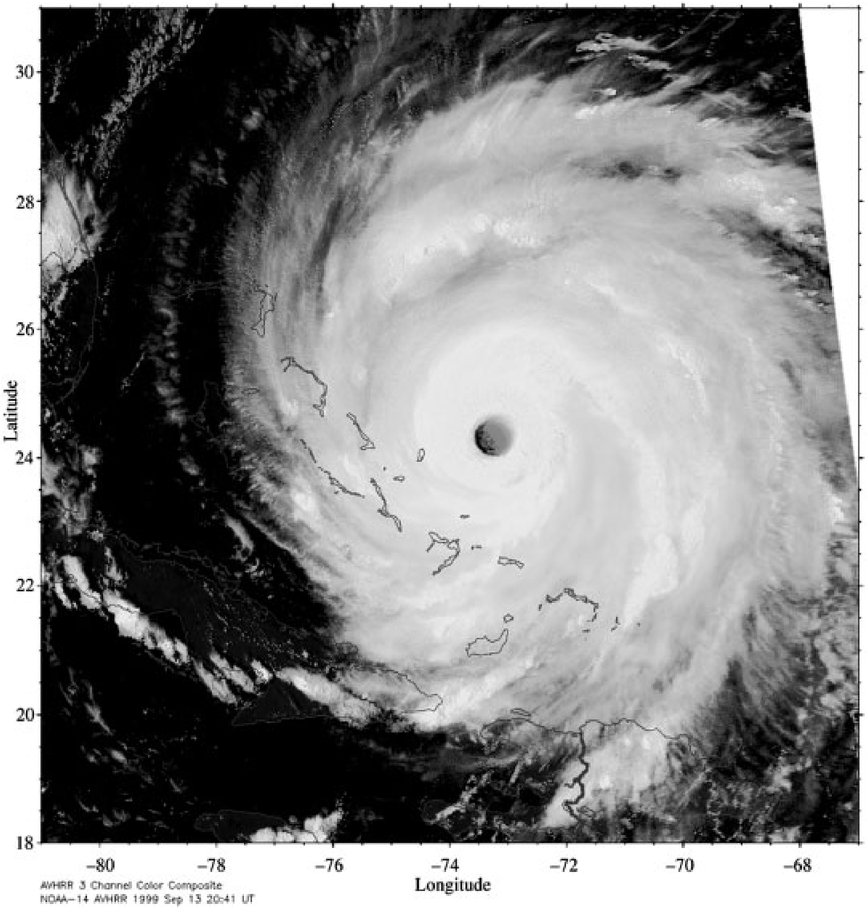

30
28
26
24
22
20
18
Latitude
−80
−78
−76
−74
−72
−70
−68
Longitude
AVHRR 3 Channel Color Composite
NOAA−14 AVHRR 1999 Sep 13 20:41 UT

Figure 1 (see color insert) NOAA-14 AVHRR multispectral false color image of Hurricane Floyd at 2041 UTC, September 13 about 800 km east of south Florida. (Photo courtesy of NOAA Operationally Significant Event Imagery website: *http://www.osei.noaa.gov/.*) See ftp site for color image.

as 100. When the Rossby number significantly exceeds unity, the balance in the core becomes more cyclostrophic, where the pressure gradient force is almost completely balanced by the centrifugal force. The time scales are such that air swirling around the center completes an orbit in much less than a pendulum day (defined as $1/f$).

When the atmosphere is in approximate horizontal and vertical balance, the wind and mass fields are tightly interconnected. The distribution of a single mass or momentum variable may be used as a starting point to infer the distribution of all other such variables. One such variable is potential vorticity (PV), approximately equal to the vorticity times the thermal stratification, which is related to the three-dimensional mass and momentum fields through an inverse second-order Laplacian-like operator. The benefit of such a relationship is that PV variations in a single location are diagnostically related to variations in mass and wind fields at a distance. Areas of high PV correspond locally to low mass, or cyclones, while areas of low PV correspond to anticyclones.

Typical extratropical weather systems contain high PV values around $0.5 \times 10^{-6} m^2/s$ K/kg (0.5 PVU) to 5 PVU, whereas, typical values in the tropical cyclone core are ≥ 10 PVU. Figure 2 shows the wind and mass fields associated with an idealized axially symmetric tropical cyclone PV anomaly with the PV concentrated near the surface rather than in a vertical column. The cyclonic anomaly (positive in the Northern Hemisphere) is associated with a cyclonic circulation that is strongest at the level of the PV anomaly near the surface, and decreases upward. Temperatures are anomalously warm above the PV anomaly (isentropic surfaces are deflected downward). While consistent with the simple PV distribution, the wind and mass fields are also in horizontal and vertical balance. The tropical cyclone being a warm-core vortex, the PV inversion dictates that the winds that swirl about the center decrease with increasing height, but they typically fill the depth of

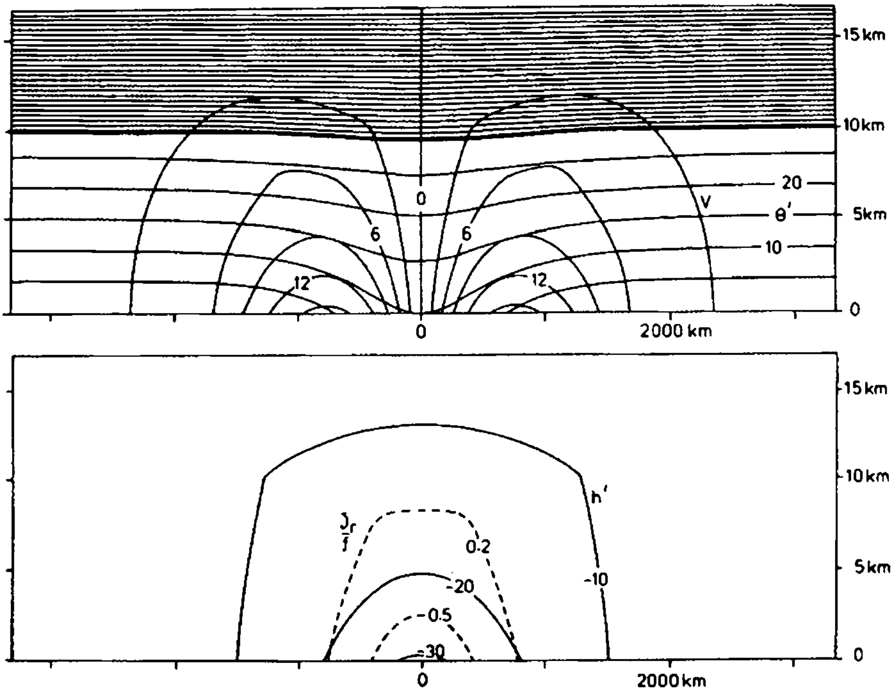


Figure 2 Gradient wind v (m/s) and perturbation potential temperature θ' (K, top panel); and geopotential/height perturbation h' (dm) and relative vorticity scale by Coriolis parameter ζ/f (bottom panel) for a warm core, lower cyclone. The tropopause location is denoted by the bold solid line, and the label 0 on the horizontal axis indicates the core (and axis of symmetry) of the disturbance. The equivalent pressure deviation at the surface in the center of the vortex is -31 hPa. [From A. J. Thorpe, *Mon. Wea. Rev.* **114**, 1384–1389 (1986). Copyright owned by the American Meteorological Society.]

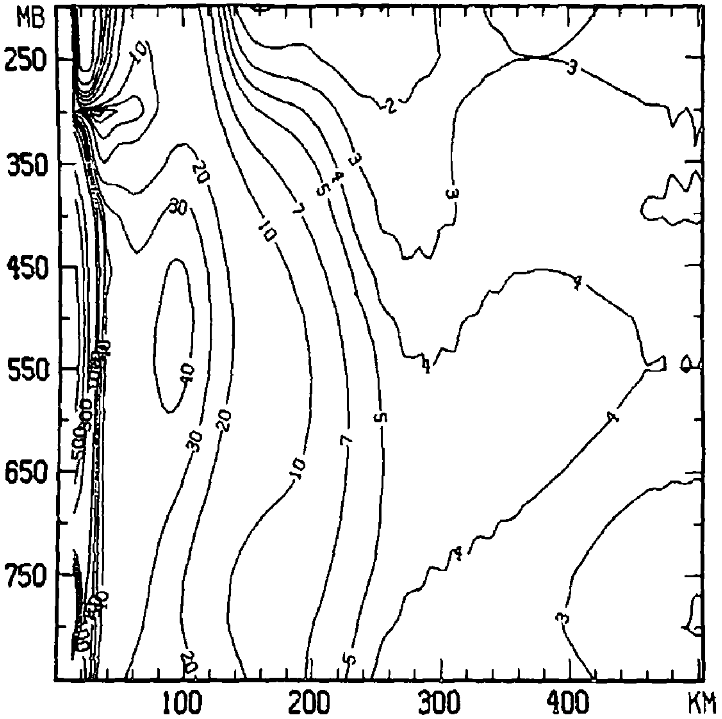


Figure 3 Radial-height cross section of symmetric PV for Hurricane Gloria, September 24, 1985. Contours are 0.1 PVU. Values in data sparse region, within 13 km of vortex center, are not displayed. [From L. J. Shapiro, and J. L. Franklin, *Mon. Wea. Rev.* **123**, 1465–1475, (1995). Copyright owned by American Meteorological Society.]

the troposphere. If the PV reaches values ≥ 10 PVU, the inner region winds can become intense as in Hurricane Gloria in Figure 3. Gloria had PV values exceeding 60 PVU just inside the radius of maximum winds of 15 km where the axisymmetric mean tangential winds exceeded 65 m/s.

Many features in the core, however, persist with little change for (pendulum) days (mean life span of a tropical cyclone is about 5 to 10 days). Because these long lifetimes represent tens or hundreds of orbital periods ($\sim$1 h), the flow is nearly balanced. Moreover, at winds >35 m/s, the local Rossby radius of deformation* is reduced from its normal $\sim 10^3$ km to a value comparable with the eye radius. In very intense tropical cyclones, the eye radius may approach the depth of the troposphere (15 km), making the aspect ratio unity. Thus, the dynamics near the center of a tropical cyclone are so exotic that conditions in the core differ from Earth's day-to-day weather as much as the atmosphere of another planet does.

*The Rossby radius of deformation is the ratio of speed of the relevant gravity wave mode and the local vorticity, or, equivalently, the ratios of the Brunt–Vaisala and inertial frequencies. This scale indicates the amount of energy that goes into gravity waves compared to inertial acceleration of the wind.

2 CLIMATOLOGY

There are 80 to 90 tropical cyclones worldwide per year, with the Northern Hemisphere having more tropical cyclones than the Southern Hemisphere. Table 2 shows that of the 80 to 90 tropical cyclones, 45 to 50 reach hurricane or typhoon strength and 20 reach major hurricane or super typhoon strength. The western North Pacific (27 tropical cyclones), eastern North Pacific (17 tropical cyclones), Southwest Indian Ocean (10 tropical cyclones), Australia/southwest Pacific (10 tropical cyclones), and North Atlantic (10 tropical cyclones) are the major tropical cyclone regions. There are also regional differences in the tropical cyclone activity by month with the majority of the activity in the summer season for each basin. Hence, in the Pacific, Atlantic, and North Indian Ocean, the maximum numbers of tropical cyclones occur in August through October, while in the South Pacific and Australia regions, the maxima are in February and March. In the South Indian Ocean, the peak activity occurs in June. In the western North Pacific, Bay of Bengal, and South Indian Ocean regions tropical cyclones may occur in any month, while in the other regions at least one tropical cyclone-free month occurs per year. For example, in the North Atlantic, there has never been tropical cyclone activity in January.

Some general conclusions can be drawn from the global distribution of tropical cyclone locations (Fig. 4*a*). Tropical cyclone formation is confined to a region approximately 30°N and 30°S, with 87% of them located within 20° of the equator. There is a lack of tropical cyclones near the equator, as well as in the eastern South Pacific and South Atlantic basins. From these observations there appears to be at least five necessary conditions for tropical cyclone development.

- Warm sea surface temperature (SST) and large mixed layer depth: Numerous studies suggest a minimum SST criterion of 26°C for development The warm water must also have sufficient depth (i.e., 50 m). Comparison of Figure 4*a* and 4*b* the annual mean global SST, shows the strong correlation between regions with SST > 26°C and annual tropical cyclone activity. SST > 26°C is sufficient but not necessary for tropical cyclone activity, evidenced by the regions with tropical cyclone activity when SST < 26°C. Some of the discrepancy exists because storms that form over regions where SST > 26°C are advected poleward during their life cycle. However, tropical cyclones are observed to originate over regions where SST < 26°C. These occurrences are not many, but the fact that they exist suggests that other factors are important.
- Background earth vorticity: Tropical cyclones do not form within 3° of the equator. The Coriolis parameter vanishes at the equator and increases to extremes at the poles. Hence, a threshold value of Earth vorticity (f) must exist for a tropical cyclone to form. However, the likelihood of formation does not increase with increasing f. Thus, nonzero Earth vorticity is necessary but not sufficient to produce tropical cyclones.
- Low vertical shear of the horizontal wind: In order for tropical cyclones to develop, the latent heat generated by the convection must be kept near the

TABLE 2 Mean Annual Frequency, Standard Deviation (σ), and Percent of Global Total of Number of Tropical Storms (Winds $\geq$17 m/s), Hurricane-Force Tropical Cyclone (Winds $\geq$33 m/s), and Major Hurricane-Force Tropical Cyclone (winds $\geq$50 m/s).

Tropical Cyclone Basin[a]	Tropical Storm Annual Frequency σ	Percent of Total	Hurricane Annual Frequency (σ)	Percent of Total	Major Hurricane Annual Frequency (σ)	Percent of Total
Atlantic (1944–2000)	**9.8** (3.0)	11.4	**5.7** (2.2)	12.1	**2.2** (1.5)	10.9
Northeast Pacific (1970–2000)	**17.0** (4.4)	19.7	**9.8** (3.1)	20.7	**4.6** (2.5)	22.9
Northwest Pacific (1970–2000)	**26.9** (4.1)	32.1	**16.8** (3.6)	35.5	**8.3** (3.2)	41.3
North Indian (1970–2000)	**5.4** (2.2)	6.3	**2.2** (1.8)	4.6	**0.3** (0.5)	1.5
Southwest Indian (30–100°E) (1969–2000)	**10.3** (2.9)	12.0	**4.9** (2.4)	10.4	**1.8** (1.9)	9.0
Australian/S.E. Indian (100–142°E) (1969–2000)	**6.5** (2.6)	7.5	**3.3** (1.9)	7.0	**1.2** (1.4)	6.0
Australian/S.W. Pacific (142°E) (1969–2000)	**10.2** (3.1)	11.8	**4.6** (2.4)	9.7	**1.7** (1.9)	8.5
Global (1970–2000)	**86.1** (8.0)		**47.3** (6.5)		**20.1** (5.7)	

[a] Dates in parentheses provide the nominal years for which accurate records are currently available.

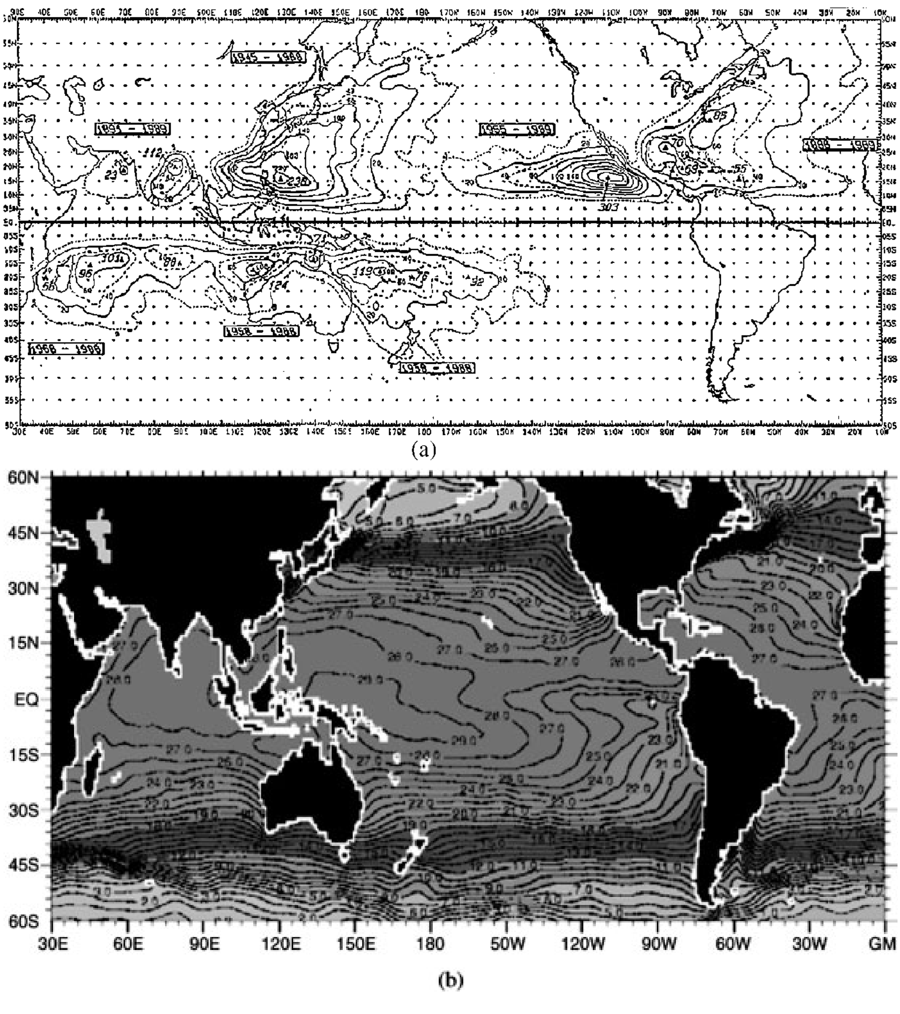


Figure 4 (see color insert) (*a*) Frequency of tropical cyclones per 100 years within 140 km of any point. Solid triangles indicate maxima, with values shown. Period of record is shown in boxes for each basin. (*b*) Annual SST distribution (°C). See ftp site for color image.

center of the storm. Historically, shear was thought to "ventilate" the core of the cyclone by advecting the warm anomaly away. The ventilation argument suggests that if the storm travels at nearly the same speed as the environmental flow in which it is embedded, its heating remains over the disturbance center. However, if it is moving slower than the mean wind at upper levels, the heating in the upper troposphere is carried away by the mean flow. Recent analysis suggests that the effect of shear is to force the convection into an asymmetric pattern such that the convective latent heat release forces flow asymmetry and irregular motion rather than intensification of the symmetric vortex. Thus, if the vertical shear is too strong ($>16\,\mathrm{m/s}$) existing tropical cyclones are ripped apart and new ones cannot form.

- Low atmospheric static stability: The troposphere must be potentially unstable to sustain convection for an extended period of time. Typically measured as the difference between the equivalent potential temperature (θ_e) at the surface and 500 hPa, instability must typically be $>10\,\mathrm{K}$ for convection to occur. This value is usually satisfied over tropical oceans.
- Tropospheric humidity: The higher the midlevel humidity, the longer a parcel of air can remain saturated as it entrains the surrounding air during its ascent. Vigorous convection occurs if the parcel remains saturated throughout its ascent. A relative humidity of 50 to 60% at lower to midlevels (700 to 500 hPa) is often sufficient to keep a parcel saturated during ascent. This condition is regularly evident over tropical oceans.

These conditions are usually satisfied in the summer and fall seasons for each tropical cyclone basin. However, even when all of the above conditions are favorable, tropical cyclones do not necessarily form. In fact, there is growing evidence for significant interannual variability in tropical cyclone activity, where numerous tropical cyclones form in a given basin over a week to 10-day period, followed by 2 to 3 weeks with little or no tropical cyclone activity. Figure 5 shows just such an active period in the Atlantic basin in mid-September 1999, where two hurricanes (Floyd and Gert), both major, and an unnamed tropical depression formed within a few days of each other. During these active phases, almost every disturbance makes at least tropical storm strength, whereas in the inactive phase, practically none of the distur-

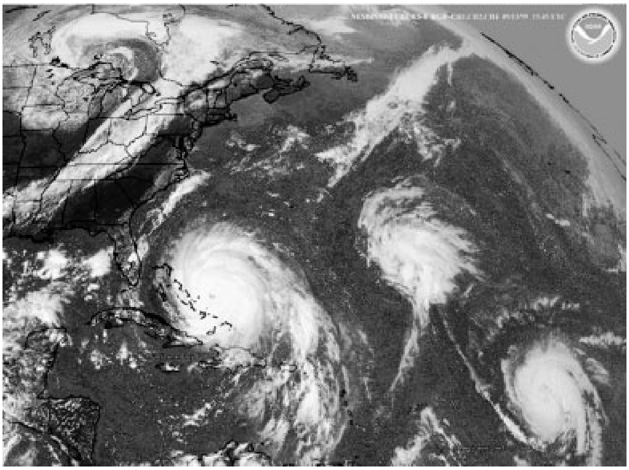


Figure 5 (*a*) *GOES* multispectral false color image of Hurricanes Floyd and Gert and an unnamed tropical depression at 1935 UTC, September 13. 1999. (Photo courtesy of NOAA Operationally Significant Event Imagery website: *http://www.osei.noaa.gov/*.) See ftp site for color image.

bances intensify. The two hurricanes and unnamed depression in Figure 5 represented the second 10-day active period during the summer of 1999. An earlier period in mid-August also resulted in the development of three hurricanes (Brett, Cindy, and Dennis), two of which were major, as well as a tropical storm (Emily). There is speculation that the variability is related to the propagation of a global wave. Because the SST, static stability, and Earth vorticity do not vary that much during the season, the interannual variability is most likely related to variations in tropospheric relative humidity and vertical wind shear.

It has long been recognized that the number of tropical cyclones in a given region varies from year to year. The exact causes of this remain largely speculative. The large-scale global variations in atmospheric phenomena such as the El Niño and Southern Oscillation (ENSO) and the Quasi-Biennial Oscillation (QBO) appear to be related to annual changes in the frequency of tropical cyclone formation, particularly in the Atlantic Ocean. The ENSO phenomenon is characterized by warmer SSTs in the eastern South Pacific and anomalous winds over much of the equatorial Pacific. It influences tropical cyclone formation in the western North Pacific, South Pacific, and even the North Atlantic.

During the peak phase of the ENSO, often referred to as El Niño (usually occurs during the months of July to October), anomalous westerly winds near the equator extend to the date line in the western North Pacific acting to enhance the intertropical convergence zone (ITCZ) in this area, making it more favorable for formation of tropical cyclones. Another effect of the El Niño circulation is warmer SST in the eastern South Pacific. During such years, tropical cyclones form closer to the equator and farther east. Regions, such as French Polynesia, which are typically unfavorable for tropical cyclones due to a strong upper-level trough, experienced numerous tropical cyclones. The eastern North Pacific is also affected by the El Niño through a displacement of the ITCZ south to near 5°N. Additionally, the warm ocean anomaly of El Niño extends to near 20°N, which enhances the possibility of tropical cyclone formation. The result is an average increase of two tropical cyclones during El Niño years. Cyclones also develop closer to the equator and farther west than during a normal year.

The QBO is a roughly 2-year oscillation of the equatorial stratosphere (30 to 50 hPa) winds from easterly to westerly and back. The phase and magnitude of QBO are associated with the frequency of tropical cyclones in the Atlantic. Hurricane activity is more frequent when the 30-hPa stratospheric winds are westerly. The exact mechanism by which the QBO affects tropical cyclones in the troposphere is not clear; however, there are more North Atlantic tropical cyclones when the QBO is in the westerly phase than in the easterly phase.

3 TROPICAL CYCLOGENESIS

Tropical cyclones form because of thermodynamic disequilibrium between the warm near-surface waters of the tropical ocean and the tropospheric column. If one adjusts an air column to equilibrium with the SST of the summertime tropical ocean, the resulting change in surface pressure is commensurate with that found in tropical cyclones. Thus, much of the tropical oceans contain enough moist enthalpy to support a major hurricane.

Throughout most of the trade-wind regions, gradual subsidence causes an inversion that traps water vapor in the lowest kilometer. Sporadic convection (often in squall lines) that breaks through the inversion exhausts the moist enthalpy stored in the near-surface boundary layer quickly, leaving a wake of cool, relatively dry air. This air comes from just above the inversion and is brought to the surface by downdrafts driven by the weight of hydrometeors and cooling due to their evaporation. The squall line has to keep moving or it quickly runs out of energy. A day, or even several days, may pass before normal fair-weather evaporation can restore the preexistent moist enthalpy behind the squall. The situation is like a poorly designed control loop that oscillates around its equilibrium point. The reasons why squall line convection generally fails to produce hurricanes lie in the limited amount of enthalpy that can be stored in the subinversion layer and the slow rate of evaporation under normal wind speeds in the trades.

To make a tropical cyclone, one needs to speed up evaporation and raise the equilibrium enthalpy at the sea surface temperature by lowering the surface pressure. Tropical cyclones are thus finite-amplitude phenomena. They do not grow by some linear process from infinitesimal amplitude. The normal paradigm of searching for the most rapidly growing unstable linear mode used to study midlatitude cyclogenesis through baroclinic instability fails here. The surface wind has to exceed roughly 20 m/s before evaporation can prevail against downdraft cooling.

How then do tropical cyclones reach the required finite amplitude? The answer seems to lie in the structure of tropical convection. As explained previously, behind a squall line the lower troposphere (below the 0°C isotherm at $\sim$5 km) is dominated by precipitation-driven downdrafts which lie under the "anvil" of nimbostratus and cirrostratus that spreads behind the active convection. Above 5 km, condensation in the anvil forces rising motion. This updrafts-over-downdrafts arrangement requires horizontal convergence centered near 5 km altitude to maintain mass continuity. The important kinematic consequence is formation of patchy shallow vortices near the altitude of the 0°C isotherm. The typical horizontal scales of these "mesovortices" are tens to hundreds of kilometers. If they were at the surface or if their influence could be extended downward to the surface, they would be the means to get the system to the required finite amplitude.

The foregoing reasoning defines the important unanswered questions: (1) How do the midlevel mesovortices extend their influence to the surface? (2) What are the detailed thermodynamics at the air–sea interface during this process? Leading hypotheses for (1) are related to processes that can increase the surface vorticity through changes in static stability and momentum mixing, both horizontally and vertically. However, the answers to these questions await new measurements that are just becoming available through improved observational tools.

4 BASIC STRUCTURE

Primary and Secondary Circulations

Inner-core dynamics received a lot of attention over the last 40 years through aircraft observations of the inner-core structure. These observations show that the tropical cyclone inner-core dynamics are dominated by interactions between “primary” (horizontal axisymmetric), “secondary” (radial and vertical) circulations, and a wave number one asymmetry caused by the storm motion. The primary circulation is so strong in the cyclone core that it is possible to consider axisymmetric motions separately, if account is taken of forcing by the asymmetric motions. The primary circulation is in near-gradient balance and evolves when heat and angular momentum sources (often due to asymmetric motions) force secondary circulations, which in turn redistribute heat and angular momentum.

Figure 6 shows that the primary circulation is sustained by the secondary circulation, which consists of frictional inflow that loses angular momentum* to the sea as it gains moist enthalpy. The inflow picks up latent heat through evaporation and exchanges sensible heat with the underlying ocean, as it spirals into lower levels of the storm under influence of friction. The evaporation of sea spray adds moisture to the air, while at the same time cooling it. This process is important in determining the intensity of a tropical cyclone. Near the vortex center, the inflow turns upward and brings the latent heat it acquires in the boundary layer into the free atmosphere. Across the top of the boundary layer, turbulent eddies causes significant downward

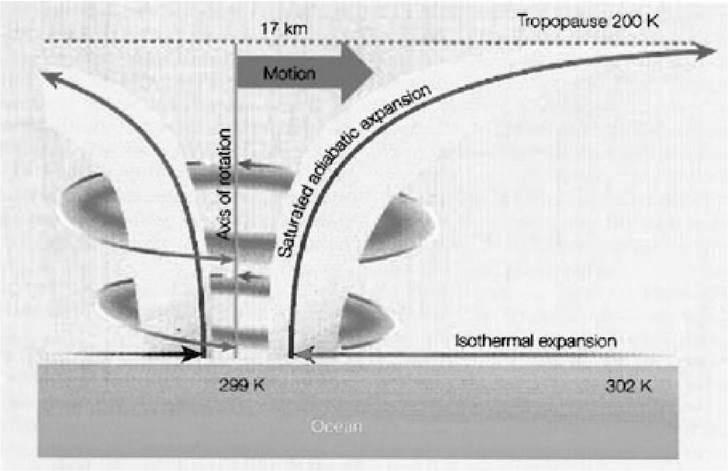


Figure 6 Schematic of the secondary circulation thermodynamics. [From H. E. Willoughby, *Nature* **401**, 649–650 (1999). Copyright owned by Macmillan Magazines, Ltd.] See ftp site for color image.

*Angular momentum, $M = Vr + fr^2/2$, where V is the tangential wind velocity, f is the Coriolis parameter, and r the radius from the storm center.

flux of sensible heat from the free atmosphere to the boundary layer. The energy source for the turbulent eddies is mechanical mixing caused by the strong winds. The eddies are also responsible for downward mixing of angular momentum. Hence, these turbulent eddy fluxes fuel the storm.

As the air converges toward the eye and is lifted in convective clouds that surround the clear eye, it ascends to the tropopause (the top of the troposphere, where temperature stops decreasing with height). As shown in Figure 6 the convective updrafts in the eyewall turn the latent heat into sensible heat through the latent heat of condensation to provide the buoyancy needed to loft air from the surface to tropopause level. The updraft entrains midlevel air promoting mass and angular momentum convergence into the core. It is the midlevel inflow that supplies the excess angular momentum needed to spin up the vortex. The net energy realized in the whole process is proportional to the difference in temperature between the ocean ($\sim$300 K) and the upper troposphere ($\sim$200 K). Storm-induced upwelling of cooler water reduces ocean SST by a few degrees, which has a considerable effect on the storm's intensity.

As shown in Figure 7 the secondary circulation also controls the distribution of hydrometeors and radar reflectivity. It is much weaker than the primary circulation except in the anticyclonic outflow, where the vortex is also much more asymmetric. Precipitation-driven convective updrafts form as hydrometeors fall from the outward sloping updraft. Condensation in the anvil causes a mesoscale updraft above the 0°C isotherm and precipitation loading by snow falling from the overhanging anvil causes a mesoscale downdraft below 0°C isotherm. The melting level itself marks the height of maximum mass convergence. Inside the eye, dynamically driven descent and momentum mixing leads to substantial pressure falls.

In order for the primary circulation to intensify, the flow cannot be in exact balance. Vertical gradients of angular momentum due to vertical shears of the primary circulation causes updrafts to pass through the convective heat sources because the path of least resistance for the warmed air lies along constant angular momentum surfaces. Similarly, horizontal temperature gradients due to vertical shears cause the horizontal flow to pass through momentum sources because the path of least resistance lies along isentropes (potential temperature or θ surfaces). Although the flow lies generally along the angular momentum or isentropic surfaces, it has a small component across them. The advection by this component not the direct forcing, is the mechanism by which the primary circulation evolves.

Some of the most intense tropical cyclones exhibit "concentric" eyewalls, two or more eyewall structures centered at the circulation center of the storm. In much the same way as the inner eyewall forms, convection surrounding the eyewall can become organized into distinct rings. Eventually, the inner eye begins to feel the effects of the subsidence resulting from the outer eyewall, and the inner eyewall weakens, to be replaced by the outer eyewall. The pressure rises due to the destruction of the inner eyewall are usually more rapid than the pressure falls due to the intensification of the outer eyewall, and the cyclone itself weakens for a short period of time. This mechanism, referred to as the *eyewall replacement* cycle, often accom-

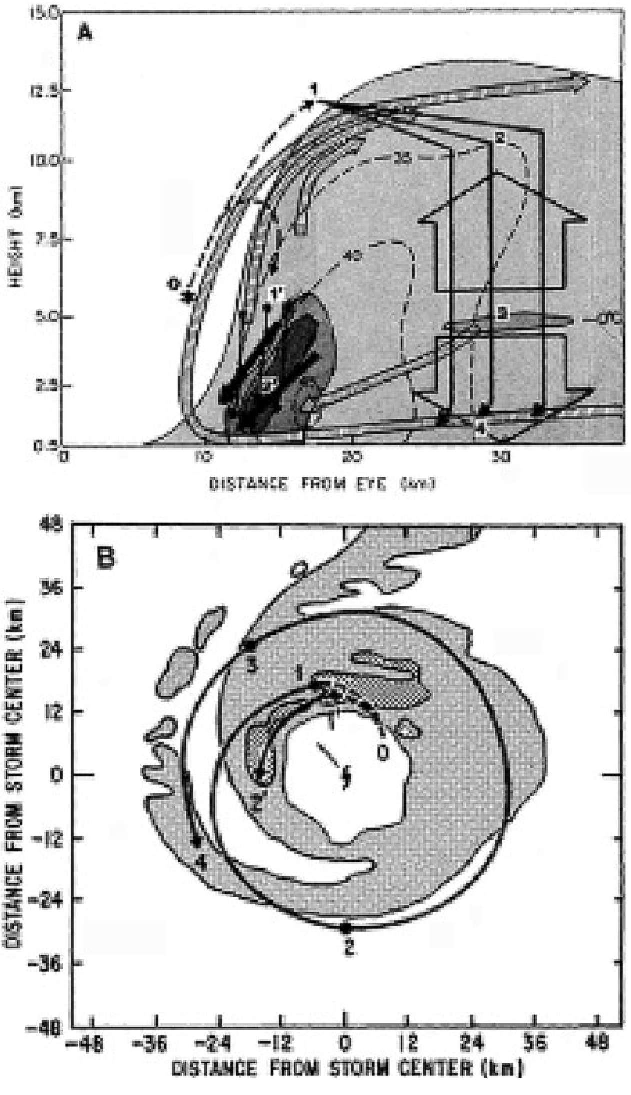


Figure 7 (*a*) Schematic of the radius–height circulation of the inner core of Hurricane Alicia. Shading depicts the reflectivity field, with contours of 5, 30, and 35 dBZ. The primary circulation (m/s) is depicted by dashed lines and the secondary circulation by the wide hatched streamlines. The convective downdrafts are denoted by the thick solid arrows, while the mesoscale up- and downdrafts are shown by the broad arrows. (*b*) Schematic plan view of the low-level reflectivity field in the inner core of Hurricane Alicia superimposed with the middle of the three hydrometeor trajectories in (*a*). Reflectivity contours in (*b*) are 20 and 35 dBZ. The storm center and direction are also shown. In (*a*) and (*b*) the hydrometeor trajectories are denoted by dashed and solid lines labeled 0–1–2–3–4 and 0′–1′–2′ [From F. D. Marks, and R. A. Houze, *J. Atmos. Sci.* **44**, 1296–1317 (1987). Copyright owned by American Meteorological Society.]

panies dramatic changes in storm intensity. The intensity changes are often associated with the development of secondary wind maxima outside the storm core.

A good example of contracting rings of convection effecting the intensification of a hurricane is shown in Figure 8 for Hurricane Gilbert on September 14, 1988. On September 14, two convective rings, denoted by intense radar reflectivity, are evident in Figure 8*a*. The outer ring is located near 80 to 90 km radius and the inner one at 10 to 12 km radius. Figure 8*b* shows that both are associated with maxima in tangential wind and vorticity. Figure 9 shows that in the ensuing 12 to 24 h the storm filled dramatically. However, it is not clear how much of the filling was caused by the storm moving over land and how much was due to the contracting outer ring and decaying inner ring of convective activity.

A process has been proposed whereby (i) nonlinear balanced adjustment of the vortex to the eddy heat and angular momentum sources associated with an upper trough in the storm's periphery produces an enhanced secondary circulation, (ii) a secondary wind maximum develops in response when the forcing reaches the inner radii, and (iii) the wind maximum contracts as a result of differential adiabatic warming associated with the convective diabatic heating in the presence of a inward radial gradient of inertial stability. Under these circumstances, understanding the intensification of the tropical cyclone reduces to determining how the secondary wind maximum develops.

Inner Core—Eyewall and Eye

The most recognizable feature found within a hurricane is the "eye" (Fig. 10). It is found at the center and is typically between 20 and 50 km in diameter. The eye is the focus of the hurricane, the point about which the primary circulation rotates and where the lowest surface pressures are found in the storm. The eye is a roughly circular area of comparatively light winds and fair weather found at the center of strong tropical cyclones. Although the winds are calm at the axis of rotation, strong winds may extend well into the eye. As seen in Figure 10 there is little or no precipitation and sometimes blue sky or stars can be seen. The eye is the region of warmest temperatures aloft—the eye temperature may be $\geq$10°C warmer at an altitude of 12 km than the surrounding environment, but only 0 to 2°C warmer at the surface.

The eye is surrounded by the eyewall, the roughly circular area of deep convection that is associated with the up-branch of the secondary circulation and the highest surface winds. The eye is composed of air that is slowly sinking, and the eyewall has a net upward flow because of many moderate—occasionally strong—updrafts and downdrafts. The eye's warm temperatures are due to warming by compression of the subsiding air. Most soundings taken within the eye are similar to that for Hurricane Hugo in Figure 11. They show a low-level layer that is relatively moist, with an inversion above, suggesting that the sinking in the eye typically does not reach the ocean surface, but instead only gets within 1 to 3 km of the surface. An eye is usually only present in hurricane-strength tropical cyclones.

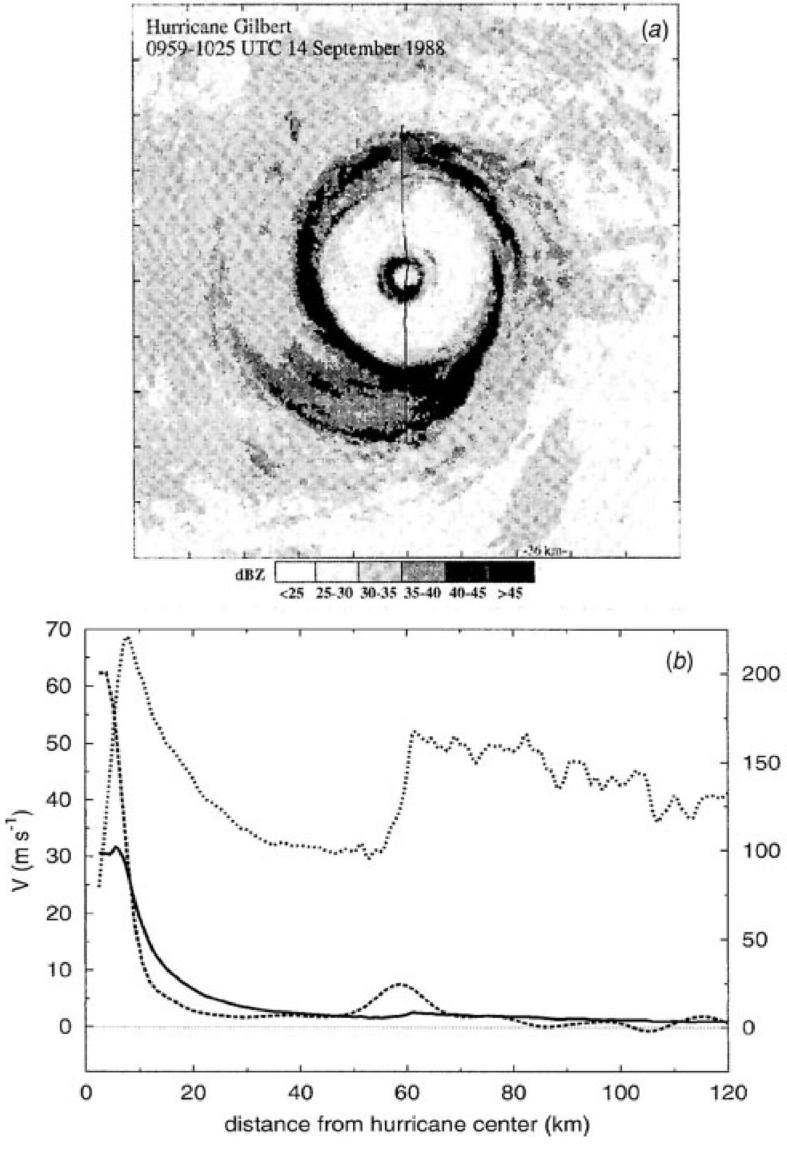


Figure 8 (*a*) Composite horizontal radar reflectivity of Hurricane Gilbert for 0959–1025 UTC September 14, 1988; the domain is 360 × 360 km, with tick marks every 36 km. The line through the center is the WP-3D aircraft flight track. (*b*) Profiles of flight-level angular velocity (solid), tangential wind (short dash), and smoothed relative vorticity (long dash) along the southern leg of the flight track shown in (*a*). [From J. P. Kossin, W. H. Schubert, and M. T. Montgomery, *Atmos. Sci.* **57**, 3893–3917 (2000). Copyright owned by American Meteorological Society.]

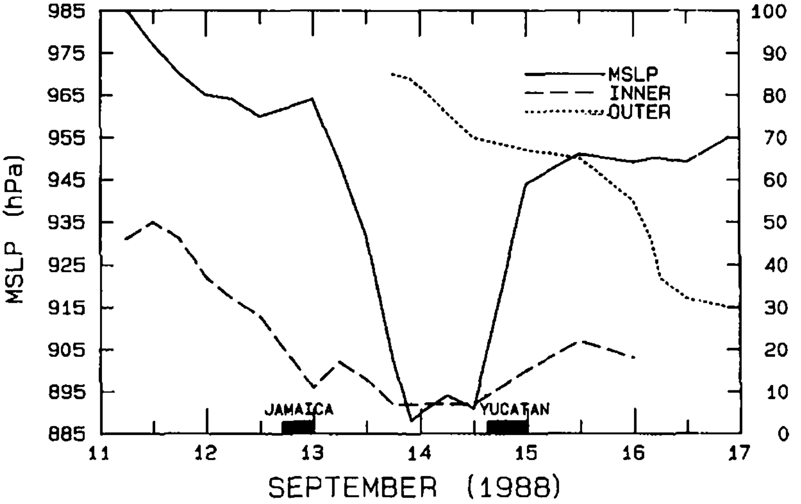


Figure 9 Hurricane Gilbert's minimum sea level pressure and radii of the inner and outer eyewalls as a function of time, September 1988. Solid blocks at bottom indicate times over land. [From M. L. Black, and H. E. Willoughby, *Mon. Wea. Rev.* **120**, 947–957 (1992). Copyright owned by American Meteorological Society.]

The general mechanisms by which the eye and eyewall are formed are not fully understood, although observations shed some light on the problem. The calm eye of the tropical cyclone shares many qualitative characteristics with other vortical systems such as tornadoes, waterspouts, dust devils, and whirlpools. Given that

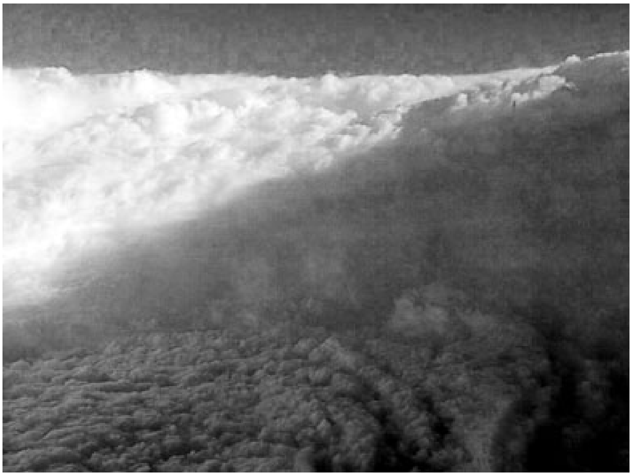


Figure 10 Eyewall of Hurricane Georges 1945 UTC, September 19, 1998. (Photo courtesy of M. Black, NOAA/OAR/AOML Hurricane Research Division.) See ftp site for color image.

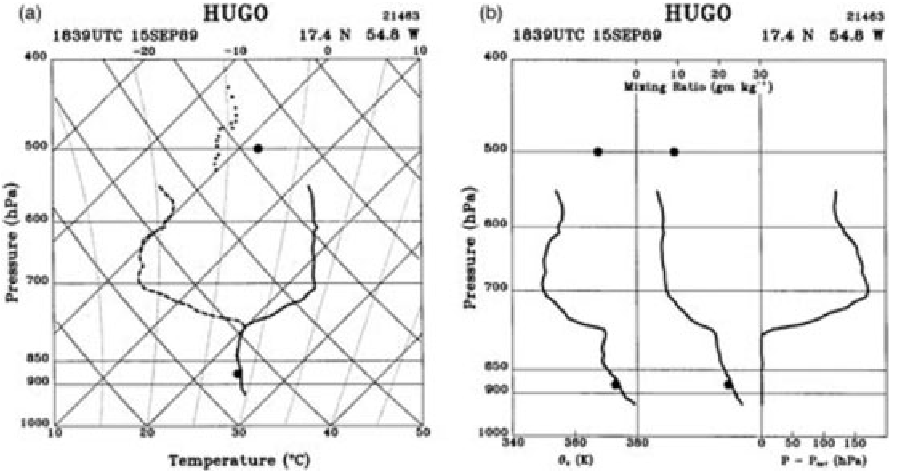


Figure 11 (a) Skew $T \log p$ diagram of the eye sounding in Hurricane Hugo at 1839 UTC on September 15, 1989. Isotherms slope upward to the right; dry adiabats slope upward to the left; moist adiabats are nearly vertical curving to the left. Solid and dashed curves denote temperature and dew point, respectively. The smaller dots denote saturation points computed for the dry air above the inversion, and the two larger dots temperature observed at the innermost saturated point as the aircraft passed through the eyewall. (*b*) θ_e, water vapor mixing ratio, and saturation pressure difference, P–P_{SAT}, as functions of pressure at 2123 UTC. [From H. E. Willoughby, *Mon. Wea. Rev.* **126**, 3189–3211 (1998). Copyright owned by American Meteorological society.]

many of these lack a change of phase of water (i.e., no clouds and diabatic heating involved), it may be that the eye feature is a fundamental component to all rotating fluids. It has been hypothesized that supergradient wind flow (i.e., swirling winds generate stronger centrifugal force than the local pressure gradient can support) present near the radius of maximum winds causes air to be centrifuged out of the eye into the eyewall, thus accounting for the subsidence in the eye. However, others found that the swirling winds within several tropical cyclones were within 1 to 4% of gradient balance. It may be though that the amount of supergradient flow needed to cause such centrifuging of air is only on the order of a couple percent and thus difficult to measure.

Another feature of tropical cyclones that probably plays a role in forming and maintaining the eye is the eyewall convection. As shown in Figure 12, convection in developing tropical cyclones is organized into long, narrow rain bands that are oriented in the same direction as the horizontal wind. Because these bands seem to spiral into the center of a tropical cyclone, they are sometimes called spiral bands. The earliest radar observations of tropical cyclones detected these bands, which are typically 5 to 50 km wide and 100 to 300 km long. Along these bands, low-level convergence is a maximum, and therefore, upper-level divergence is most pronounced. A direct circulation develops in which warm, moist air converges at the surface, ascends through these bands, diverges aloft, and descends on both sides

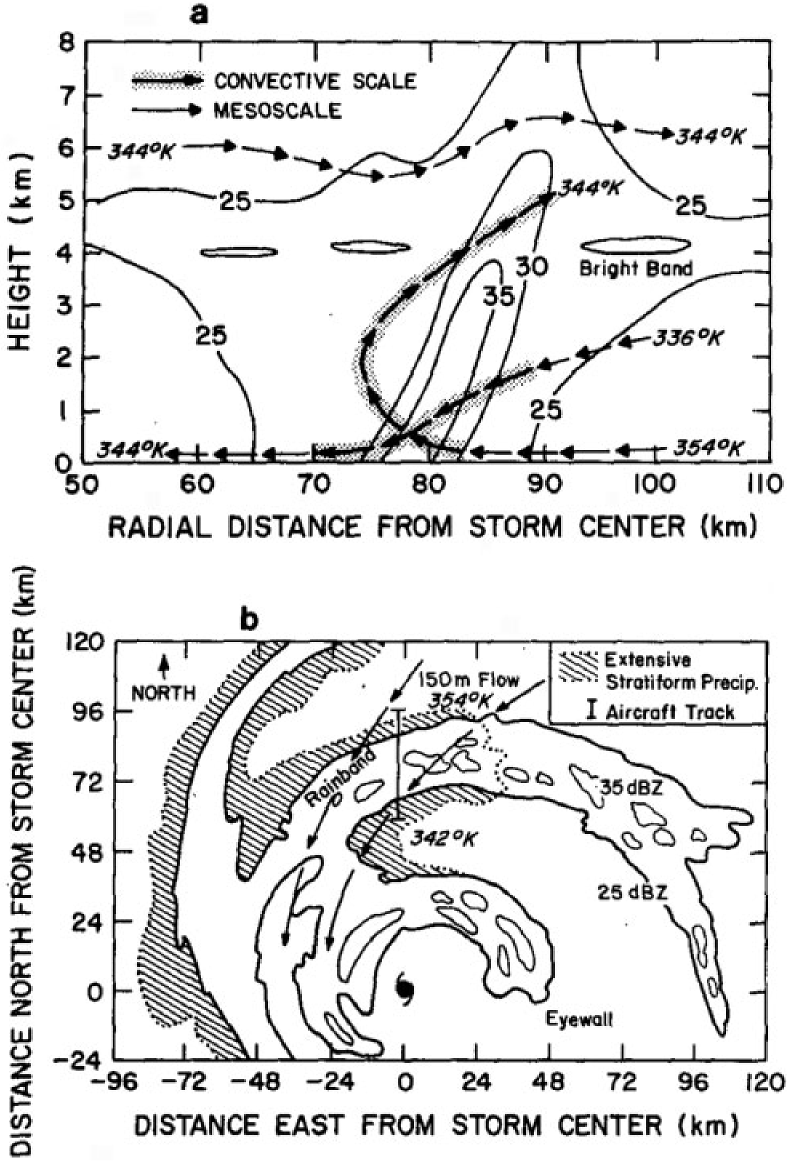


Figure 12 (*a*) Schematic of the rain band in radius–height coordinates. Reflectivity, θ_e, mesoscale (arrows), and convective-scale motions are shown. (*b*) Plan view. Aircraft track, reflectivities, cells, stratiform precipitation, 150-m flow, and θ_e values are shown. [From G. M. Barnes, E. J. Zipser, D. P. Jorgensen, and F. D. Marks, *J. Atmos. Sci.* **40**, 2125–2137 (1983). Copyright owned by American Meteorological Society.]

of the bands. Subsidence is distributed over a wide area outside of the rain band but is concentrated in the small inside area. As the air subsides, adiabatic warming takes place, and the air dries. Because subsidence is often concentrated on the inside of the band, the adiabatic warming is stronger inward from the band causing a sharp contrast in pressure falls across the band since warm air is lighter than cold air. Because of the pressure falls on the inside, the tangential winds around the tropical cyclone increase due to increased pressure gradient. Eventually, the band moves toward the center and encircles it and the eye and eyewall form.

The circulation in the eye is comparatively weak and, at least in the mature stage, thermally indirect (warm air descending), so it cannot play a *direct* role in the storm energy production. On the other hand, the temperature in the eye of many hurricanes exceeds that which can be attained by any conceivable moist adiabatic ascent from the sea surface, even accounting for the additional entropy (positive potential temperature, θ, anomaly) owing to the low surface pressure in the eye (the lower the pressure, the higher the θ at a given altitude and temperature). Thus, the observed low central pressure of the storm is not consistent with that calculated hydrostatically from the temperature distribution created when a sample of air is lifted from a state of saturation at sea surface temperature and pressure. The thermal wind balance restricts the amount of warming that can take place. In essence, the rotation of the eye at each level is imparted by the eyewall, and the pressure drop from the outer to the inner edge of the eye is simply that required by gradient balance.

Because the eyewall azimuthal velocity decreases with height, the radial pressure drop decreases with altitude, requiring, through the hydrostatic equation, a temperature maximum at the storm center. Thus, given the swirling velocity of the eyewall, the steady-state eye structure is largely determined. The central pressure, which is estimated by integrating the gradient balance equation inward from the radius of maximum winds, depends on the assumed radial profile of azimuthal wind in the eye.

In contrast, the eyewall is a region of rapid variation of thermodynamic variables. As shown in Figure 13, the transition from the eyewall cloud to the nearly cloud-free eye is often so abrupt that it has been described as a form of atmospheric front. Early studies were the first to recognize that the flow under the eyewall cloud is inherently frontogenetic. The eyewall is the upward branch of the secondary circulation and a region of rapid ascent that, together with slantwise convection, leads to the congruence of angular momentum and moist entropy (θ_c) surfaces. Hence, the three-dimensional vorticity vectors lie on θ_e surfaces, so that the moist PV vanishes. As the air is saturated, this in turn implies, through the invertibility principle applied to flow in gradient and hydrostatic balance, that the entire primary circulation may be deduced from the radial distribution of θ_e in the boundary layer and the distribution of vorticity at the tropopause.

In the classic semigeostrophic theory of deformation-induced frontogenesis, the background geostrophic deformation flow provides the advection of temperature across surfaces of absolute momentum that drives the frontogenesis whereas, in the hurricane eyewall, surface friction provides the radial advection of entropy across angular momentum surfaces. Also note that the hurricane eyewall is not

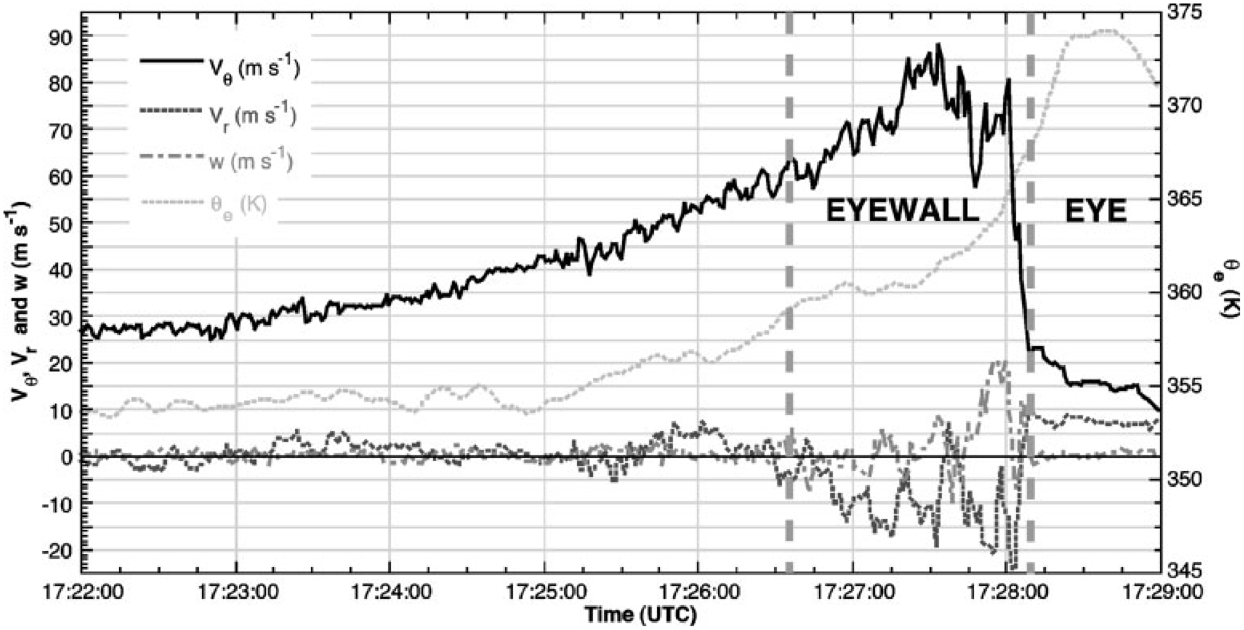


Figure 13 Time series plots of tangential wind (V_θ), radial wind (V_r), vertical velocity (w), and θ_e in Hurricane Hugo for 1721–1730 UTC, September 15, 1989. The aircraft flight track was at 450 m. Thick dashed vertical lines denote the width of the eyewall reflectivity maximum at low levels. See ftp site for color image.

necessarily a front in surface temperature, but instead involves the θ_e distribution, which is directly related to density in saturated air.

There is likely a two-stage process to eye formation. The amplification of the primary circulation is strongly frontogenetic and results, in a comparatively short time, in frontal collapse at the inner edge of the eyewall. The frontal collapse leads to a dramatic transition in the storm dynamics. While the tropical cyclone inner core is dominated by axisymmetric motions, hydrodynamic instabilities are potential sources of asymmetric motions within the core. In intense tropical cyclones the wind profile inside the eye is often "U-shaped" in the sense that the wind increases outward more rapidly than linearly with radius (Fig. 13). The strong cyclonic shear just inside the eyewall may result in a local maximum of absolute vorticity or angular momentum, so that the profile may actually become barotropically unstable. This instability leads to frontal collapse as a result of radial diffusion of momentum into the eye and also may explain the "polygonal eyewalls" where the eyewall appear on radar to be made up of a series of line segments rather than as a circle. It may also explain intense mesoscale vortices observed in the eyewalls of hurricanes Hugo of 1989 and Andrew of 1992.

Once the radial turbulent diffusion of momentum driven by the instability of the primary circulation becomes important, it results in a mechanically induced, thermally indirect (warm air sinking) component of the secondary circulation in the eye and eyewall. Such a circulation raises the vertically averaged temperature of the eye beyond its value in the eyewall and allows for an amplification of the entropy distribution. Feedbacks with the surface fluxes then allow the boundary layer entropy to increase and result in a more rapid intensification of the swirling wind. Thus, the frontal collapse of the eyewall is an essential process in the evolution of tropical cyclones. Without it, amplification of the temperature distribution relies on external influences, and intensification of the wind field is slow. Once it has taken place, the mechanical spinup of the eye allows the temperature distribution to amplify without external influences and, through positive feedback with surface fluxes, allows the entropy field to amplify and the swirling velocity to increase somewhat more rapidly.

Outer Structure and Rain Bands

The axisymmetric core is characteristically surrounded by a less symmetric outer vortex that diminishes into the synoptic "environment." In the lower troposphere, the cyclonic circulation may extend more than 1000 km from the center. As evident in Figure 14 the boundary between cyclonic and anticyclonic circulation slopes inward with increasing height, so that the circulation in the upper troposphere is primarily anticyclonic, except near the core. In the outer vortex, there are no scale separations between the primary and the secondary circulations, the asymmetric motions, or the vortex translation as they are all of the same rough magnitude. The asymmetric flows in this region control the vortex motion and sustain an eddy convergence of angular momentum and moisture toward the center. Interactions between the symmetric motions of the inner core with the more asymmetric

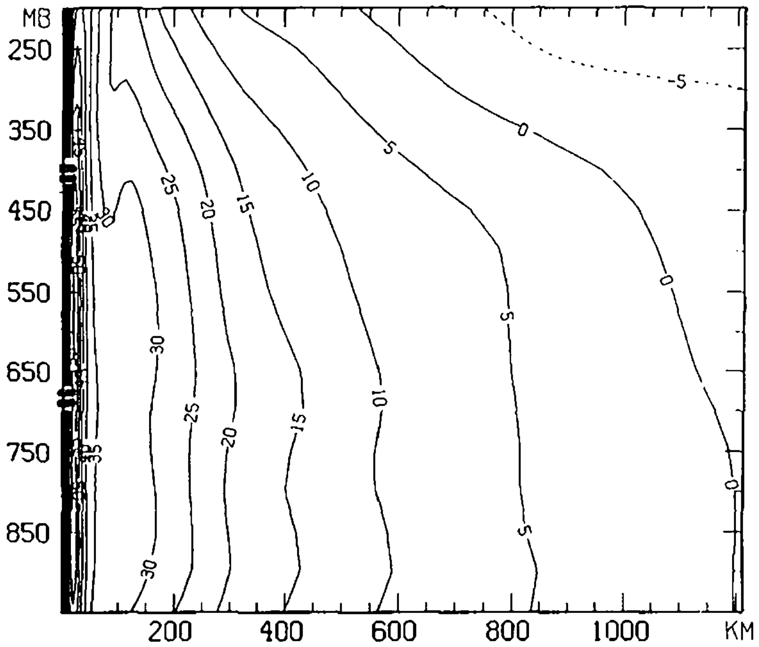


Figure 14 Vertical cross section of the azimuthal mean tangential wind for Hurricane Gloria on September 24, 1985. Anticyclonic contours are dashed. [From J. L. Franklin, S. J. Lord, S. E. Feuer, and F. D. Marks, *Mon. Wea. Rev.* **121**, 2433–2451 (1985). Copyright owned by American Meteorological Society.]

motions in the outer portion of the storm are the key to improved forecasts of tropical cyclone track and intensity.

Spiral bands of precipitation characterize radar and satellite images in this region of the storm (Figs. 1 and 15). As seen in Figures 8, 12, and 15, radar reflectivity patterns in tropical cyclones provide a good means for flow visualization although they represent precipitation, not winds. Descending motion occupies precipitation-free areas, such as the eye. The axis of the cyclone's rotation lies near the center of the eye. The eyewall surrounds the eye. In intense hurricanes, it may contain reflectivities as high as 50 dB(Z),* equivalent to rainfall rates of 74 mm/h. Less extreme reflectivities, 40 dB(Z) (13 mm/h), characterize most convective rainfall in the eyewall and spiral bands. The vertical velocities (both up and downdrafts) in convection with highest reflectivity may reach 25 m/s, but typical vertical velocities are <5 m/s. Such intense convection occupies $<10\%$ of the tropical cyclone's area. Outside convection, reflectivities are still weaker, 30 dB(Z), equivalent to a 2.4 mm/h rain rate. This "stratiform rain," denoted by a distinct reflectivity maximum or "bright band" at the altitude of the 0°C isotherm, falls out of the anvil cloud

*10 $\log_{10}$ (Z), where Z is equivalent radar reflectivity factor (mm^6/m^3).

that grows from the convection. The spiral bands tend to lie along the friction layer wind that spirals inward toward the eyewall (Fig. 12).

Many aspects of rain band formation, dynamics, and interaction with the symmetric vortex are still unresolved. The trailing-spiral shape of bands and lanes arises because the angular velocity of the vortex increases inward and deforms them into equiangular spirals. In the vortex core, air remains in the circulation for many orbits of the center; while outside the core, the air passes through the circulation in less than the time required for a single orbit. As the tropical cyclone becomes more intense, the inward ends of the bands approach the center less steeply approximating arcs of circles. Some bands appear to move outward, while others maintain a fixed location relative to the translating center.

As shown in Figure 15, motion of the vortex through its surroundings may cause one stationary band, called the "principal" band, to lay along a convergent streamline asymptote that spirals into the core. A tropical cyclone advected by middle-level steering with westerly shear moves eastward through surrounding air at low

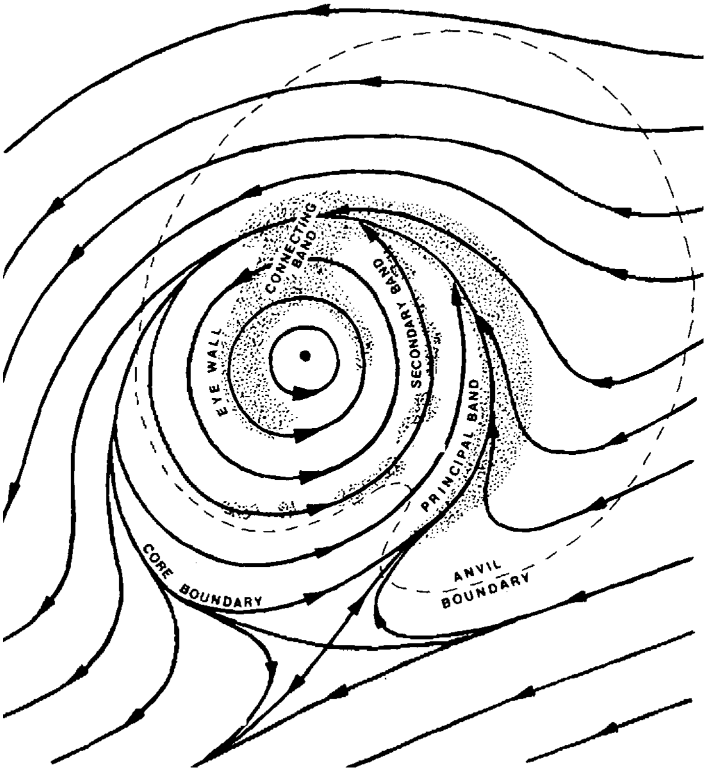


Figure 15 Schematic representation of the stationary band complex, the entities that compose it, and the flow in which it is embedded. [From H. E. Willoughby, F. D. Marks, and R. J. Feinberg, *J. Atmos. Sci.* **41**, 3189–3211 (1984). Copyright owned by American Meteorological Society.]

levels. Thus, the principal band may be a "bow wave" due to displacement of the environmental air on the eastern side of the vortex. Its predominant azimuthal wave number is one.

Moving bands, and other convective features, are frequently associated with cycloidal motion of the tropical cyclone center, and intense asymmetric outbursts of convection are observed to displace the tropical cyclone center by tens of kilometers. The bands observed by radar are often considered manifestations of internal gravity waves, but these waves can exist only in a band of Doppler-shifted frequencies between the local inertia frequency (defined as the sum of the vertical component of Earth's inertial frequency f and the local angular velocity of the circulation, V/r) and the Brunt–Vaisala frequency.* Only two classes of trailing-spiral, gravity wave solutions lie within this frequency band: (i) waves with any tangential wave number that move faster than the swirling wind, and (ii) waves with tangential wave number ≥ 2 that move slower than the swirling wind.

Bands moving faster than the swirling wind with outward phase propagation are observed by radar. They are more like squall lines than linear gravity waves. Waves moving slower than the swirling wind propagate wave energy and anticyclonic angular momentum inward, grow at the expense of the mean-flow kinetic energy, and reach appreciable amplitude if they are excited at the periphery of the tropical cyclone. Alternate explanations for these inward-propagating bands involve filamentation of vorticity from the tropical cyclone environment, asymmetries in the radially shearing flow of the vortex, and high-order vortex Rossby waves. Detailed observations of the vortex-scale rain band structure and wind field are necessary to determine which mechanisms play a role in rain band development and maintenance.

While the evolution of the inner core is dominated by interactions between the primary, secondary, and track-induced wave number one circulation, there is some indication that the local convective circulations in the rain bands may impact on intensity change. Although precipitation in some bands is largely stratiform, condensation in most bands tends to be concentrated in convective cells rather than spread over wide mesoscale areas. As shown in Figure 12, convective elements form, move through the bands, and dissipate as they move downwind. Doppler radar observations indicate that the roots of the updrafts lay in convergence between the low-level radial inflow and gust fronts that are produced by convective downdrafts. This convergence may occur on either side of the band. A 20 K decrease in low-level θ_e was observed in a rain band downdraft and suggested that the draft acts as a barrier to inflow. This reduction in boundary layer energy may be advected near the center, inhibit convection, and thereby alter storm intensity.

5 MOTION

Tropical cyclone motion is the result of a complex interaction between a number of internal and external influences. Environmental steering is typically the most prominent external influence on a tropical cyclone, accounting for as much as 70 to 90%

*Natural gravity wave frequency, the square root of the static stability defined as $(g/\theta)\,\partial\theta/\partial z$.

of the motion. Theoretical studies show that in the absence of environmental steering, tropical cyclones move poleward and westward due to internal influences.

Accurate determination of tropical cyclone motion requires accurate representation of interactions that occur throughout the depth of the troposphere on a variety of scales. Observations spurred improved understanding of how tropical cyclones move using simple barotropic and more complex baroclinic models. To first order, the storm moves with some layer average of the lower tropospheric environmental flow: The translation of the vortex is roughly equal to the speed and direction of the basic "steering" current. However, the observations show that tropical cyclone tracks deviate from this simple steering flow concept in a subtle and important manner. Several physical processes may cause such deviations. The approach in theoretical and modeling of tropical cyclones has been to isolate each process in a systematic manner to understand the magnitude and direction of the track deviation caused by each effect. The β effect,* due to the differential advection of Earth's vorticity (f), alone can produce asymmetric circulations and propagation. Models that are more complete describe not only the movement of the vortex but also the accompanying wave number one asymmetries. It was also discovered that the role of meridional and zonal gradients of the environmental flow could greatly add to the complexity even in the barotropic evolution of a vortex. Hence, the evolution of the movement depends not only on the relative vorticity gradient and on shear of the environment but also the structure of the vortex itself.

Generally, the propagation vector of these model baroclinic vortices is very close to that expected from a barotropic model initialized with the vertically integrated environmental wind. An essential feature in baroclinic systems is the relative vorticity advection through the storm center where the vertical structure of the tropical cyclone produces a tendency for the low-level vortex to move slower than the simple propagation of the vortex due to β. Vertical shear plays an important factor in determining the relative flow; however, there is no unique relation between the shear and storm motion. Diabatic heating effects also alter this flow and change the propagation velocity. Thus, tropical cyclone motion is primarily governed by the dynamics of the low-level cyclonic circulation, however, the addition of observations of the upper-level structure may alter this finding.

6 INTERACTION WITH ATMOSPHERIC ENVIRONMENT

A consensus exists that small vertical shear of the environmental wind and lateral eddy imports of angular momentum are favorable to tropical cyclone intensification. The inhibiting effect of vertical shear in the environment on tropical cyclone intensification is well known from climatology and forecasting experience. The appearance of the low-level circulation outside the tropical cyclones central dense overcast in satellite imagery is universally recognized as a symptom of shear and as an

*Asymmetric vorticity advection around the vortex caused by the latitudinal gradient of$f$, $\beta = 2\Omega \cos \phi$. β has a maximum value at the equator (i.e., $2.289 \times 10^{-1}\text{s}^{-1}$) and becomes zero at the pole.

indication of nonintensification or weakening. Nevertheless, the detailed dynamics of a vortex in shear has been the topic of surprisingly little study, probably because, while the effect is a reliable basis for practical forecasting, it is difficult to measure and model.

In contrast, the positive effect of eddy momentum imports at upper levels has received extensive study. Modeling studies with composite initial conditions show that eddy momentum fluxes can intensify a tropical cyclone even when other conditions are neutral or unfavorable. It has been shown theoretically that momentum imports can form a tropical cyclone in an atmosphere with no buoyancy. Statistical analysis of tropical cyclones reveals a clear relationship between angular momentum convergence and intensification, but only after the effects of shear and SST variations are accounted for. Such interactions occur frequently (35% of the time, defined by eddy angular momentum flux convergence exceeding 10 m/s day), and likely represent the more common upper boundary interaction for tropical cyclones. Frequently, they are accompanied by eyewall cycles and dramatic intensity changes.

The environmental flows that favor intensification, and presumably inward eddy momentum fluxes, usually involve interaction with a synoptic-scale cyclonic feature, such as a midlatitude upper-level trough or PV anomaly. Given the interaction of an upper-level trough and the tropical cyclone, the exact mechanism for intensification is still uncertain. The secondary circulation response to momentum and heat sources is very different. Upper-tropospheric momentum sources can influence the core directly. As can be seen in Figure 16, large inertial stability* in the lower troposphere protects the mature tropical cyclone core from direct influence by momentum sources; however, the inertial stability in the upper troposphere is smaller and a momentum source can induce an outflow jet with large radial extent just below the tropopause. If the eyewall updraft links to the direct circulation at the entrance region of the jet, as shown in Figure 17*c*, and 17*d*, the exhaust outflow is unrestricted. The important difference between heat and momentum sources is that the roots of the diabatically induced updraft must be in the inertially stiff lower troposphere, but the outflow jet due to a momentum flux convergence can be confined to the inertially *labile* upper troposphere. Momentum forcing does not spin the vortex up directly. It makes the exhaust flow stronger and reduces local compensating subsidence in the core, thus cooling the upper troposphere and destabilizing the sounding. The cooler upper troposphere leads to less thermal wind shear and a weaker upper anticyclone.

A two-dimensional balanced approach provides reasonable insight into the nature of the tropical cyclone intensification as a trough approaches. Isentropic PV analysis (Fig. 17), which express the problem in terms of a quasi-conserved variable in three dimensions, are used to describe various processes in idealized tropical cyclones with considerable success. The eddy heat and angular momentum fluxes are related to changes in the isentropic PV through their contribution to the eddy flux of PV.

*A measure of the resistance to horizontal displacements, based on the conservation of angular momentum for a vortex in gradient balance, and is defined as $(f + \zeta)(f + V^2/r)$, where ζ is the relative vorticity, V is the axial wind velocity, f is the Coriolis parameter, and r radius from the storm center.

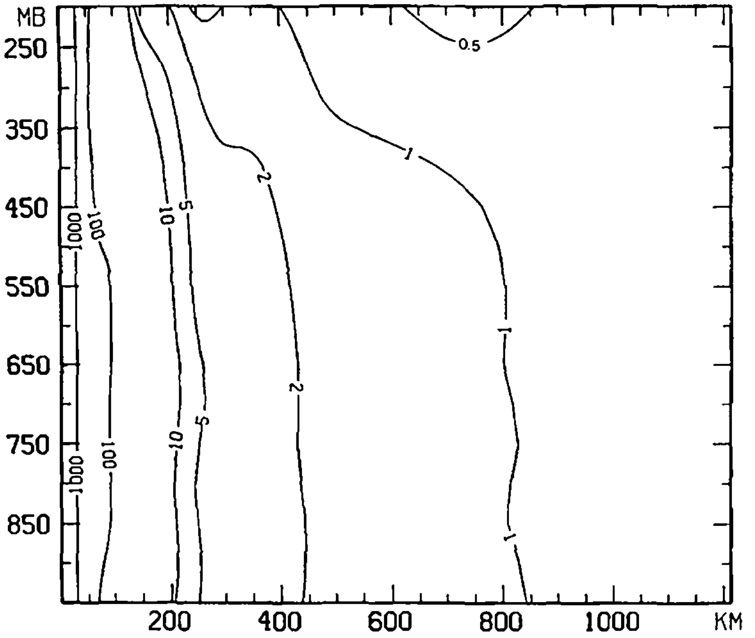


Figure 16 Axisymmetric mean inertial stability for Hurricane Gloria on September 24, 1985. Contours are shown as multiples of f^2 at the latitude of Gloria's center. [From J. L. Franklin, S. J. Lord, S. B. Feuer, and F. D. Marks, *Mon. Wea. Rev.* **121**, 2433–2451 (1993). Copyright owned by American Meteorological Society.]

It has been suggested that outflow-layer asymmetries, as in Figure 17, and their associated circulations could create a mid- or lower-tropospheric PV maximum outside the storm core, either by creating breaking PV waves on the midtropospheric radial PV gradient (Fig. 3) or by diabatic heating. It has been shown that filamentation of any such PV maximum in the "surf zone" outside the tropical cyclone core (the sharp radial PV gradient near 100 km radius) produces a feature much like a secondary wind maximum, which was apparent in the PV fields of hurricane Gloria in Figure 3. These studies thus provide mechanisms by which outflow-layer asymmetries could bring about a secondary wind maximum.

An alternative argument has been proposed for storm reintensification as a "constructive interference without phase locking," as shown in Figure 18. As the PV anomalies come within the Rossby radius of deformation, the pressure and wind perturbations associated with the combined anomalies are greater than when the anomalies are apart, even though the PV magnitudes are unchanged. The perturbation energy comes from the basic-state shear that brought the anomalies together. However, constructive interference without some additional diabatic component cannot account for intensification. It is possible that intensification represents a baroclinic initiation of a wind-induced surface heat exchange. By this mechanism,

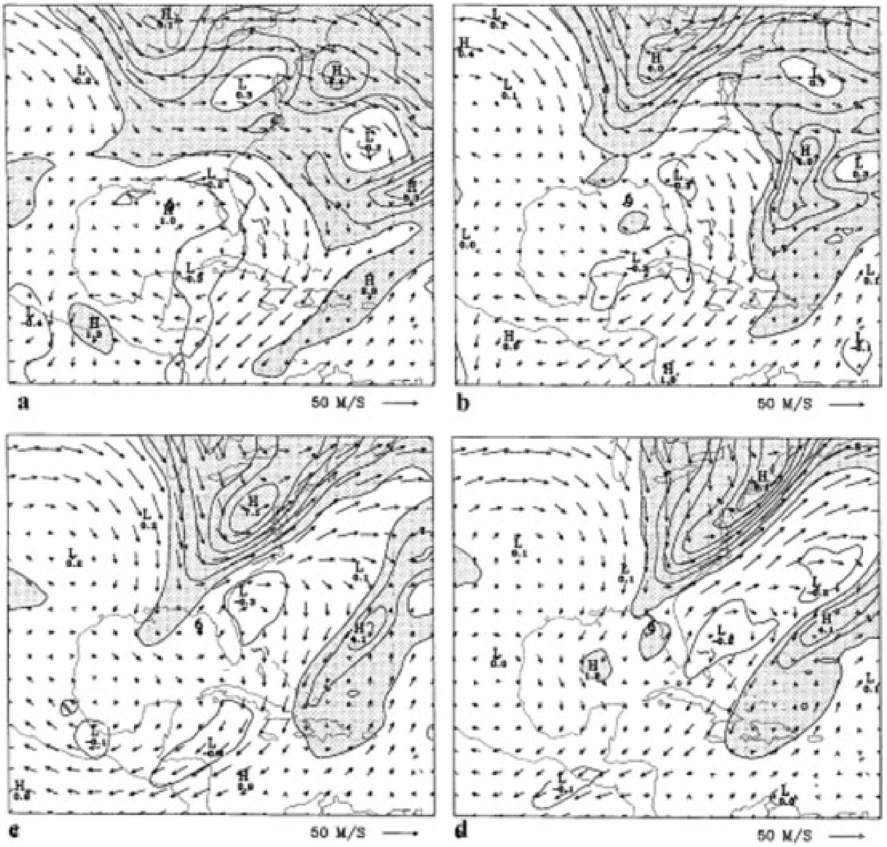


Figure 17 Wind vectors and PV on the 345 K isentropic surface at (*a*) 1200 UTC August 30; (*b*) 0000 UTC August 31; (*c*) 1200 UTC August 31, and (*d*) 0000 UTC September 1, 1985. PV increments are 1 PVU and values >1 PVU are shaded. Wind vectors are plotted at 2.25° intervals. The 345 K surface is approximately 200 hPa in the hurricane environment and ranges from 240 to 280 hPa at the storm center. The hurricane symbol denotes the location of Hurricane Elena. [From J. Molinari, S. Skubis, and D. Vollaro, *J. Atmos. Sci.* **52**, 3593–3606 (1995). Copyright owned by American Meteorological Society.]

the constructive interference induces stronger surface wind anomalies, which produce larger surface moisture fluxes and thus higher surface moist enthalpy. This feeds back through the associated convective heating to produce a stronger secondary circulation and thus stronger surface winds. The small effective static stability in the saturated, nearly moist neutral storm core ensures a deep response so that even a rather narrow upper trough can initiate this feedback process. The key to this mechanism is the direct influence of the constructive interference on the surface wind field as that controls the surface flux of moist enthalpy.

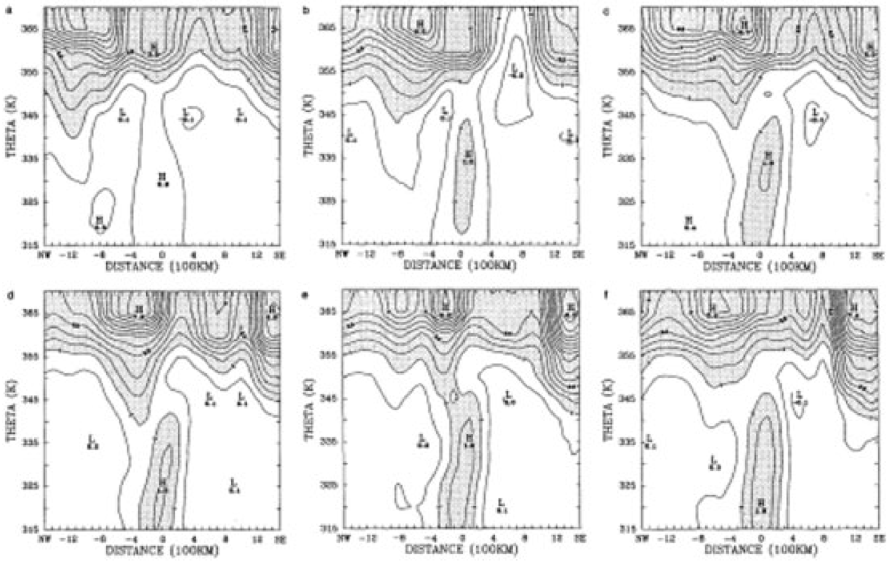


Figure 18 Cross sections of PV from northwest (left) to southeast (right) through the observed as center of Hurricane Elena for the same times as in Fig. 17, plus (*a*) 0000 UTC August 30 and (*f*) 1200 UTC September 1. Increment is 0.5 PVU and shading above 1 PVU. [From J. Molinari, S. Skubis, and D. Vollaro, *J. Atmos. Sci.* **52**, 3593–3606 (1995). Copyright owned by American Meteorological Society.]

7 INTERACTION WITH THE OCEAN

As pointed out in Section 2, preexisting SSTs >26°C are a necessary but insufficient condition for tropical cyclogenesis. Once the tropical cyclone develops and translates over the tropical oceans, statistical models suggest that warm SSTs describe a large fraction of the variance (40 to 70%) associated with the intensification phase of the storm. However, these predictive models do not account either for the oceanic mixed layers having temperatures of 0.5 to 1°C cooler than the relatively thin SST over the upper meter of the ocean or horizontal advective tendencies by basic-state ocean currents such as the Gulf Stream and warm core eddies. Thin layers of warm SST are well mixed with the underlying cooler mixed layer water well in advance of the storm where winds are a few meters per second, reducing SST as the storm approaches. However, strong oceanic baroclinic features advecting deep, warm oceanic mixed layers represent moving reservoirs of high-heat content water available for the continued development and intensification phases of the tropical cyclone. Beyond a first-order description of the lower boundary providing the heat and moisture fluxes derived from low-level convergence, little is known about the complex boundary layer interactions between the two geophysical fluids.

One of the more apparent aspects of the atmospheric-oceanic interactions during tropical cyclone passage is the upper ocean cooling as manifested in the SST (and

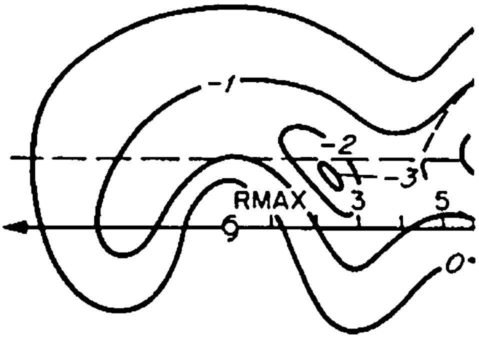


Figure 19 Schematic SST change (°C) induced by a hurricane. The distance scale is indicated in multiples of the radius of maximum wind. Storm motion is to the left. Horizontal dashed line is at 1.5 times the radius of maximum wind. [From P. G. Black, R. L. Elsberry, and L. K. Shay, *Adv. Underwater Tech., Ocean Sci. Offshore Eng.* **16**, 51–58 (1998). Copyright owned by Society for Underwater Technology (Graham & Trotman).]

mixed layer temperature) decrease starting just in back of the eye. As seen in Figure 19, ocean mixed layer temperature profiles acquired during the passage of several tropical cyclones revealed a crescent-shaped pattern of upper ocean cooling and mixed layer depth changes, which indicated a rightward bias in the mixed layer temperature response with cooling by 1 to 5°C extending from the right-rear quadrant of the storm into the wake regime. These SST decreases are observed through satellite-derived SST images, such as the one shown in Figure 20 of the post–Hurricane Bonnie SST, which are indicative of mixed layer depth changes due to stress-induced turbulent mixing in the front of the storm and shear-induced mixing between the mixed layer and thermocline in the rear half of the storm. The mixed layer cooling represents the thermodynamic and dynamic response to the strong wind that typically accounts for 75 to 85% of the ocean heat loss, compared to the 15 to 25% caused by surface latent and sensible heat fluxes from the ocean to the atmosphere. Thus, the upper ocean's heat content for tropical cyclones is not governed solely by SST, rather it is the mixed layer depths and temperatures that are significantly affected along the lower boundary by the basic state and transient currents.

Recent observational data has shown that the horizontal advection of temperature gradients by basic-state currents in a warm core ring affected the mixed layer heat and mass balance, suggesting the importance of these warm oceanic baroclinic features. In addition to enhanced air–sea fluxes, warm temperatures (>26°C) may extend to 80 to 100 m in warm core rings, significantly impacting the mixed layer heat and momentum balance. That is, strong current regimes (1 to 2 m/s) advecting deep, warm upper ocean layers not only represent deep reservoirs of high heat content water with an upward heat flux but transport heat from the tropics to the subtropical and polar regions as part of the annual cycle. Thus, the basic state of the mixed layer and the subsequent response represents an evolving three-dimensional

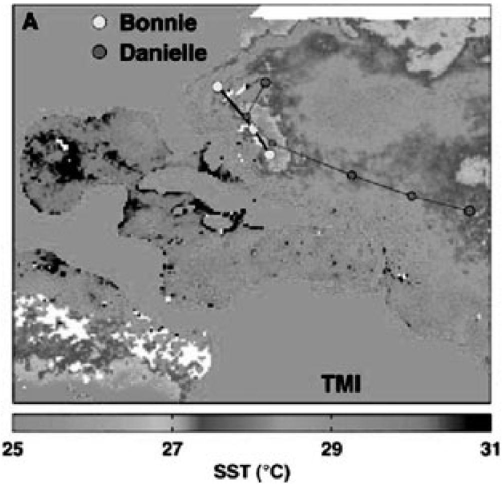


Figure 20 (see color insert) Cold wake produced by Hurricane Bonnie for August 24–26, 1998, as seen by the NASA TRMM satellite *Microwave Imager* (TMI). Small white patches are areas of persistent rain over the 3-day period. White dots show Hurricane Bonnie's daily position from August 24 to 26. Gray dots show the later passage of Hurricane Danielle from August 27 to September 1. Danielle crossed Bonnie's wake on August 29 and its intensity drops. [From F. J. Wentz, C. Gentemann, D. Smith, and D. Chelton, *Science* **288**, 847–850 (2000). Copyright owned by the American Geophysical Union. (*http:/www.sciencemag.org*)]. See ftp site for color image.

process with surface fluxes, vertical shear across the entrainment zone and horizontal advection. Simultaneous observations in both fluids are lacking over these baroclinic features prior, during, and subsequent to tropical cyclone passage and are crucially needed to improve our understanding of the role of lower boundary in intensity and structural changes to intensity change.

In addition, wave height measurements and current profiles revealed the highest waves and largest fetches to the right side of the storm where the maximum mixed layer changes occurred. Mean wave-induced currents were in the same direction as the steady mixed layer currents, modulating vertical current shears and mixed layer turbulence. These processes feed back to the atmospheric boundary layer by altering the surface roughness and hence the drag coefficient. However, little is known about the role of strong surface waves on the mixed layer dynamics, and their feedback to the atmospheric boundary layer under tropical cyclone force winds by altering the drag coefficient.

8 TROPICAL CYCLONE RAINFALL

Precipitation in tropical cyclones can be separated into either convective or stratiform regimes. Convective precipitation occurs primarily in the eyewall and rain bands, producing rains >25 mm/h over small areas. However, observations suggest that only 10% of the total rain area is comprised of these convective rain cores. The average core is 4 km in radius (area of 50 km^2) and has a relatively short lifetime, with only 10% lasting longer than 8 min (roughly the time a 1-mm-diameter raindrop takes to fall from the mean height of the 0°C isotherm at terminal velocity). The short life cycle of the cores and the strong horizontal advection produce a well-mixed and less asymmetric precipitation pattern in time and space. Thus, over 24 h the inner core of a tropical cyclone as a whole produces 1 to 2 cm of precipitation over a relatively large area and 10 to 20 cm in the core. After landfall, orographic forcing can anchor heavy precipitation to a local area for an extended time. Additionally, midlatitude interaction with a front or upper level trough can enhance precipitation, producing a distortion of the typical azimuthally uniform precipitation distribution.

9 ENERGETICS

Energetically, a tropical cyclone can be thought of, to a first approximation, as a heat engine; obtaining its heat input from the warm, humid air over the tropical ocean, and releasing this heat through the condensation of water vapor into water droplets in deep thunderstorms of the eyewall and rain bands, then giving off a cold exhaust in the upper levels of the troposphere (~12 km up). One can look at the energetics of a tropical cyclone in two ways: (1) the total amount of energy released by the condensation of water droplets or (2) the amount of kinetic energy generated to maintain the strong swirling winds of the hurricane. It turns out that the vast majority of the heat released in the condensation process is used to cause rising motions in the convection and only a small portion drives the storm's horizontal winds.

Using the first approach we assume an average tropical cyclone produces 1.5 cm/day of rain inside a circle of radius 665 km. Converting this to a volume of rain gives 2.1×10^{16} cm/day (a cm^3 of rain weighs 1 g). The energy released through the latent heat of condensation to produce this amount of rain is 5.2×10^{19} J/day or 6.0×10^{14} W, which is equivalent to 200 times the worldwide electrical generating capacity.

Under the second approach we assume that for a mature hurricane, the amount of kinetic energy generated is equal to that being dissipated due to friction. The dissipation rate per unit area is air density times the drag coefficient times the wind speed cubed. Assuming an average wind speed for the inner core of the hurricane of 40 m/s winds over a 60 km radius, the wind dissipation rate (wind generation rate) would be 1.5×10^{12} W. This is equivalent to about half the worldwide electrical generating capacity.

Either method suggests hurricanes generate an enormous amount of energy. However, they also imply that only about 2.5% of the energy released in a hurricane by latent heat released in clouds actually goes to maintaining the hurricane's spiraling winds.

10 TROPICAL CYCLONE–RELATED HAZARDS

In the coastal zone, extensive damage and loss of life are caused by the storm surge (a rapid, local rise in sea level associated with storm landfall), heavy rains, strong winds, and tropical cyclone-spawned severe weather (e.g., tornadoes). The continental United States currently averages nearly $5 billion (in 1998 dollars) annually in tropical cyclone–caused damage, and this is increasing, owing to growing population and wealth in the vulnerable coastal zones.

Before 1970, large loss of life stemmed mostly from storm surges. The height of storm surges varies from 1 to 2 m in weak systems to more than 6 m in major hurricanes that strike coastlines with shallow water offshore. The storm surge associated with Hurricane Andrew (1992) reached a height of about 5 m, the highest level recorded in southeast Florida. Hurricane Hugo's (1989) surge reached a peak height of nearly 6 m about 20 miles northeast of Charleston, South Carolina, and exceeded 3 m over a length of nearly 180 km of coastline. In recent decades, large loss of life due to storm surges in the United States has become less frequent because of improved forecasts, fast and reliable communications, timely evacuations, a better educated public, and a close working relationship between the National Hurricane Center (NHC), local weather forecast offices, emergency managers, and the media. Luck has also played a role, as there have been comparatively few landfalls of intense storms in populous regions in the last few decades. The rapid growth of coastal populations and the complexity of evacuation raises concerns that another large storm surge disaster might occur along the eastern or Gulf Coast shores of the United States.

In regions with effectively enforced building codes designed for hurricane conditions, wind damage is typically not so lethal as the storm surge, but it affects a much larger area and can lead to large economic loss. For instance, Hurricane Andrew's winds produced over $25 billion in damage over southern Florida and Louisiana. Tornadoes, although they occur in many hurricanes that strike the United States, generally account for only a small part of the total storm damage.

While tropical cyclones are most hazardous in coastal regions, the weakening, moisture-laden circulation can produce extensive, damaging floods hundreds of miles inland long after the winds have subsided below hurricane strength. In recent decades, many more fatalities in North America have occurred from tropical cyclone–induced inland flash flooding than from the combination of storm surge and wind. For example, although the deaths from storm surge and wind along the Florida coast from hurricane Agnes in 1972 were minimal, inland flash flooding caused more than 100 deaths over the northeastern United States. More recently, rains from Hurricane Mitch (1998) killed at least 10,000 people in Central America, the majority after the storm had weakened to tropical storm strength. An essential difference in the threat from flooding rains, compared to that from wind and surge, is that the rain amount is not tied to the strength of the storm's winds. Hence, any tropical disturbance, from depression to major hurricane, is a major rain threat.

11 SUMMARY

The eye of the storm is a metaphor for calm within chaos. The core of a tropical cyclone, encompassing the eye and the inner 100 to 200 km of the cyclone's 1000 to 1500 km radial extent, is hardly tranquil. However, the rotational inertia of the swirling wind makes it a region of orderly, but intense, motion. It is dominated by a cyclonic primary circulation in balance with a nearly axisymmetric, warm core low-pressure anomaly. Superimposed on the primary circulation are weaker asymmetric motions and an axisymmetric secondary circulation. The asymmetries modulate precipitation and cloud into trailing spirals. Because of their semibalanced dynamics, the primary and secondary circulations are relatively simple and well understood. These dynamics are not valid in the upper troposphere where the outflow is comparable to the swirling flow, nor do they apply to the asymmetric motions. Since the synoptic-scale environment appears to interact with the vortex core in the upper troposphere by means of the asymmetric motions, future research should emphasize this aspect of the tropical cyclone dynamics and their influence on the track and intensity of the storm.

Improved track forecasts, particularly the location and time when a tropical cyclone crosses the coast are achievable with more accurate specification of the initial conditions of the large-scale environment and the tropical cyclone wind fields. Unfortunately, observations are sparse in the upper troposphere, atmospheric boundary layer, and upper ocean, limiting knowledge of environmental interactions, angular momentum imports, boundary layer stress, and air–sea interactions. In addition to the track, an accurate forecast of the storm intensity is needed since it is the primary determinant of localized wind damage, severe weather, storm surge, ocean wave runup, and even precipitation during landfall. A successful intensity forecast requires knowledge of the mechanisms that modulate tropical cyclone intensity through the relative impact and interactions of three major components: (1) the structure of the upper ocean circulations that control the mixed layer heat content, (2) the storm's inner core dynamics, and (3) the structure of the synoptic-scale upper-tropospheric environment. Even if we could make a good forecast of the landfall position and intensity, our knowledge of tropical cyclone structure changes as it makes landfall is in its infancy because little hard data survives the harsh conditions. To improve forecasts, developments to improve our understanding through observations, theory, and modeling need to be advanced together.

SUGGESTED READING

Emanuel, K. A. (1986). An air–sea interaction theory for tropical cyclones. Part I: Steady-state maintenance, *J. Atmos. Sci.* **43**, 585–604.

Ooyama, K. V. (1982). Conceptual evolution of the theory and modeling of the tropical cyclone, *J. Meteor. Soc. Jpn.* **60**, 369–380.

WMO (1995). *Global Perspectives of Tropical Cyclones*, Russell Elsberry (Ed.), World Meteorological Organization Report TCP-38, Geneva, Switzerland.

CHAPTER 32

MODERN WEATHER FORECASTING

LAWRENCE B. DUNN

1 INTRODUCTION

It has often been said that weather forecasting is part science and part art. Essentially, weather forecasting is the application of science to a very practical problem that affects the lives and livelihoods of people and nations. How the science is applied varies somewhat from forecaster to forecaster, and this subjectivity is the aspect of weather forecasting that is still a bit of an art form. Many meteorologists and atmospheric scientists are uncomfortable with the subjectivity of weather forecasting since it inevitably leads to inconsistency, and there is a never-ending attempt to provide forecasters with tools to make the process as objective as possible. Given the preceding discussion, it should be clear that, like forecasting itself, any description of modern weather forecasting will vary, at least slightly, according to the experiences of the author.

Modern weather forecasting is applied to a broad spectrum of scales and applications. Forecasting is done by government and private meteorologists for a variety of users. Forecasts of general conditions for each of the next 7 days are widely distributed and are familiar to virtually everyone. In the United States, forecasts and warnings to protect life and property are provided by government meteorologists, although private-sector forecasters will often add value to these products, especially through dissemination of the products. Specific forecasts for commercial interests are done by private-sector meteorological firms. Forecast applications are numerous. Examples include weather predictions for aviation, the marine community, management of wildfires and prescribed burns, flood control, reservoir management, utility power demand, road surface conditions, agriculture, snowmaking at ski resorts, the retail industry, commodity traders and brokers, the film industry, special outdoor

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

sporting events, and many others. The increased skill of modern weather forecasts has resulted in more and more decision makers in all fields basing high resource commitments on weather predictions.

Early in the twentieth century weather forecasting was based on the collection of surface observations of wind, temperature, pressure, dew-point temperature, clouds, and precipitation. The observations were plotted and analyzed by hand. By the 1930s rawinsonde observations of the upper atmosphere began to supplement the surface observation network that was the backbone of operational meteorology. By mid-century the frontal theory of the time, often referred to as the Norwegian Frontal Model because of its origins, was being applied as a unifying concept to interpret the observations and to guide analysis. Numerous empirical rules existed that related clouds, precipitation, and frontal evolution to the observations. To a great extent these rules described how to prognosticate weather systems through extrapolation.

The nonlinear nature of atmospheric processes limits the utility of extrapolation techniques, and the problem of cyclone development and decay were addressed in the 1940s and 1950s by application of theoretical "development" equations that related upper-level vorticity and thermal advection to divergence and ultimately to changes in surface pressure. The quasi-geostrophic (QG) approximation, basically stating that the wind velocity remains in approximate balance with the pressure gradient, was invoked as part of these development equations. To apply these concepts to weather forecasting, vorticity charts were constructed via manual graphical methods that were quite laborious. The goal of all these efforts was to predict the future position of cyclone centers and the attendant fronts as accurately as possible and then to derive the sensible weather that was characteristically associated with each weather system.

By the late 1950s and 1960s numerical weather prediction using recently developed computers became skillful enough that humans were no longer preparing prognostic charts by hand, and forecasting instead focused on interpreting the output from the various computer models of the atmosphere. This continues to this day and will be discussed in detail later in this chapter.

The 1950s and 1960s were also the time when remote sensing of the atmosphere via radar and satellite began to have an impact on weather forecasting. In particular, these types of observations became primary tools along with the conventional surface observations to make forecasts and warnings of weather phenomena on a short temporal scale (0 to 6 h) and a spatial scale of approximately towns and counties. Remote-sensing technology continued to evolve through the remainder of the twentieth century, and short-term warnings and forecasts of severe and significant weather have become one of the primary activities of operational weather forecasters.

Modern weather forecasting has evolved as a result of changes in observational, computational, and communications technology. Increases in computational resources have made a significant impact on the complexity of numerical weather prediction, but, as importantly, also on display capabilities where forecasters view model output and observational data. Perhaps more important to the evolution of weather forecasting than the changes in technology are the application of new knowledge from research results that examine the fundamental structures and processes of atmospheric phenomena on all scales. Research has led to the development and application of numerous conceptual models of atmospheric phenomena that provide a framework for forecasters as they attempt to interpret the large volume of model output and observational data now available.

The rest of this chapter will be divided into three parts. A discussion of the forecast process will be followed by sections on long-term forecasting and short-term forecasting. The distinction between short-term and long-term will be made based on the time scale when numerical weather prediction begins to become the dominant tool used by the forecaster. At the time of this writing, this transition takes place between 6 and 12 h. For the purposes of this chapter, long-term forecasting covers the period from 12 h through 7 days. Forecasting at ranges longer than 7 days is beyond the scope of this chapter.

2 FORECAST PROCESS

A prediction of the future state of the atmosphere requires that a forecaster must understand what is taking place in the present. This fundamental first step is required regardless of the spatial or temporal scale of the prediction. This step is required even if the basis for the prediction is output from a numerical weather prediction model for reasons that will be described in the long-term section below.

As part of the process of understanding the present state of the atmosphere, a forecaster must pose and then answer a series of questions. Examples of some of the more obvious questions might be: Where is it raining? Where is it not raining? Why is it raining in one location but not another? How are the clouds distributed horizontally and vertically? How has this been changing with time? Where are the centers of high and low pressure? Where are the fronts? How are the troughs and ridges aloft distributed and how have they been evolving? Where are the temperature gradients at the surface and aloft and how are they changing? A quick look at an animated sequence of satellite or radar images will bring these and dozens of other questions quickly to the forefront of the problem to be solved by the forecaster.

This part of the forecast process is called analysis. During analysis, a forecaster will examine both direct and remotely sensed observations and attempt to answer the many questions that, in the aggregate, lead to an understanding of the present state of the atmosphere. Analysis of directly observed parameters is often done objectively by software, with graphical output displayed on a computer. Analysis is also done by hand, with a forecaster carefully scrutinizing a plot of observations with a pencil in hand, drawing contours and making annotations of significant features. Skillful hand analysis can provide a forecaster with insight that is not likely to be brought out by objective analysis routines, but the utility of hand analysis is dependent on the skill of the analyst.

Analysis of remotely sensed data is also performed to ascertain the current state of the atmosphere. The most common example of this type of analysis is the examination of satellite imagery. The three most commonly available satellite data types are infrared (IR), visible (VIS), and water vapor (WV) imagery. Other types of satellite imagery exist that detect precipitable water, rainfall rates, ozone, and other parameters, and these are seeing increased usage in forecasting applications. Animation of WV imagery quickly shows a forecaster the positions of mid and upper-level circulation features such as troughs, ridges, closed cyclones, and anticyclones, as well as areas of ascent and subsidence aloft. Animation of IR imagery shows development and decay of cloud systems by changes in cloud top temperature with time. Visible imagery, which is available only during sunlit hours of the day, shows regions of low and mid-level clouds, as well as regions of fog and snow cover. Visible imagery can also be used to identify areas of convection, especially at times of low sun angle, when convective clouds cast shadows. Forecasters use animation of the different types of satellite imagery to see movement and evolution of features.

Similar analysis techniques of radar data can quickly show a forecaster the location and intensity of both stratiform and convective precipitation. The evolution and vertical structure of individual convective cells can be discerned from careful analysis of radar data. Doppler radar offers forecasters the opportunity to remotely sense the wind speed and direction on a continuous basis at a variety of levels in the vertical if there are adequate reflectors. Wind profilers, which are a type of vertically pointing Doppler radar, also provide remotely sensed information about the three-dimensional wind field.

The various directly and remotely sensed observations described above, along with other parameters not included in this discussion, are often best analyzed in combination. Computational resources are commonly available that allow for the combination of the various observed data sets. To a great extent, this was not possible prior to the 1990s, and many of these data were often analyzed separately from each other. A graphical overlay of rawinsonde and surface observations on a satellite image can make it very easy to answer a fundamental question such as "where is the front at the surface and aloft," but this was not always the case and is still quite difficult over oceanic regions.

To make sense of the observations, forecasters will often attempt to fit a conceptual model to the observed data. A conceptual model is a mental picture of an atmospheric structure or process. They are usually the result of research where intensive observations have been made on a temporal and spatial scale that is far greater in resolution than the operational observational network. The Norwegian Frontal Model is one of the best-known examples of a conceptual model. The structure and evolution of a so-called supercell thunderstorm is another example of a conceptual model. There are hundreds, and perhaps thousands, of different conceptual models in use to describe virtually every type of phenomena on virtually every spatial and temporal scale. Conceptual models are convenient structures for organizing the huge volume of data with which forecasters are faced; however, they are inevitably simplifications of actual atmospheric phenomena and processes.

Research results lead to constant refinement of existing conceptual models, proposals of new models, and even the abandonment of some conceptual models. Models of phenomena are often applicable in some geographic regions, but not others. The Norwegian Frontal Model is a very old paradigm, and it has been revised many times, and its complete abandonment has been recommended on numerous occasions. However, a satisfactory replacement for this conceptual model has not been found, so at least parts of this model continue to be used by forecasters. Most conceptual models have similar life cycles.

Another method employed by forecasters to organize the forecast process is to start their analysis of the data on the largest scale and then move progressively downscale. This method has been called the "forecast funnel" in that a forecaster starts at the top, or wide portion, of the funnel, which corresponds to the planetary scale, and then descends into the synoptic-scale in the middle of the funnel, and finally the narrow portion of the funnel focuses on the forecast problem of a specific place and time, which is determined by conditions on the mesoscale. The forecast funnel is itself a conceptual model in that it is based on the idea that larger scale atmospheric conditions drive smaller scale phenomena, and that it will be difficult or impossible to accurately predict conditions on the smallest scales without taking into account conditions on the next larger scale, and so on.

A simple example of scale interaction and the application of the forecast funnel might be to consider the potential for a snowstorm along the mid-Atlantic states. On the planetary scale, a longwave trough position must be present over eastern North America for there to be a chance of cyclogenesis, and the track of the storm will, to some extent, be determined by the planetary-scale conditions. A migratory short-wave trough(s) on the synoptic-scale is required to initiate cyclogenesis. The interaction of the synoptic-scale troughs and ridges with the local topography will determine whether cold air remains trapped on the east side of the Appalachian Mountains. This synoptic-scale interaction with the terrain drives the mesoscale conditions and modulates the precipitation type and intensity.

Like all conceptual models, the forecast funnel is a simplification. In the example above, one could also argue that the preexisting conditions on the mesoscale are just as important, or more so, for the creation of a snowstorm. Where is the cold air already present? Are there any preexisting temperature gradients or old frontal boundaries? Is the air in the warm sector of the storm unstable enough to produce widespread precipitation and convection, which could lead to further intensification of the cyclone. Forecasters generally find that they must also ascend the forecast funnel at times to take into account "upscale" interactions.

3 FORECASTING 12 h TO 7 DAYS

Modern weather forecasting of conditions beyond 12 h in the future is primarily based on forecaster interpretation of output from numerical weather prediction models. Forecasters have access to many different models. Through the Internet, forecasters can access models from the national centers of various countries, military national centers, and universities around the world. Depending on their employer's computational resources, forecasters may even have access to models run locally. At the time of this writing, model output is typically used in a deterministic sense. This means that given realistic initial conditions and a model with adequate resolution and accurate representations of the pertinent physical process, the model prediction of the future state of the atmosphere can be used to explicitly derive the sensible weather elements at Earth's surface (or aloft for aviation forecasting). By the end of the twentieth century considerable effort was being expended on so-called ensemble forecasting, where output from an ensemble of models is used either by a forecaster or directly through objective techniques to produce probabilistic forecasts of sensible weather, with a goal of including information about uncertainty with each forecast element. A very brief discussion of ensemble techniques is included at the end of this section, but most of what follows is based on a deterministic approach, which is still the most widely used method for most forecasting tasks.

Before using any model, forecasters will go through the process of analysis in an attempt to understand current conditions. This must be done on all spatial scales. The analysis step is the cornerstone of proper use of deterministic model output. It is important for at least two reasons. First, the forecaster must decide if the model initial conditions and the early hours of the forecast capture the essential detail of what is happening in the real atmosphere. Second, through analysis of current conditions, the forecaster must grasp the physical processes that are important. Different models represent the physical processes of the atmosphere in different ways, and if the forecaster expects certain processes to be particularly important, then this information will be a factor in how the forecaster uses the output from the various models.

A critique of the initial conditions of models has become much easier with the use of modern graphic/image workstations. It is a simple task to overlay various model fields with animating satellite imagery. Comparison of model initial fields with plotted observations and satellite imagery allows the forecaster to quickly determine if troughs, ridges, vorticity centers, baroclinic zones, jet stream axes, moisture plumes, and other features are reasonably well depicted in the model's initial conditions. This critique is particularly important over data-sparse regions such as the oceans but is also important over relatively data-rich areas such as North America since there are times when relatively small errors in initial conditions can lead to significant errors in the forecast.

When the initial conditions of one model are superior to all the others, a forecaster may decide to give the solution from this model greater weight as the forecast is developed. Often the situation is not clear as to which model is best initialized. Differences are often subtle, and even with current satellite capabilities, there is considerable uncertainty about the exact locations in the horizontal and vertical of specific gradients and features over oceanic regions. There are times when all of the available models have a poor grasp of the real atmosphere. In these situations, a forecaster will have to fall back to some of the empirical rules of the pre-NWP (Numerical Weather Prediction) era to mentally modify model output in the direction most likely to minimize the error. Most often in this situation the forecaster will have low confidence in what the ultimate conditions will be, and this uncertainty will be expressed quantitatively and qualitatively within the allowable format of the forecast product.

Given a reasonable outcome from the initial condition critique, the forecaster will often next move onto consideration of the different model characteristics. Typically, resolution is greater in limited domain models, while lower resolution models have larger and often global domains. High-resolution models will have more realistic representations of Earth's topography, including mountains, valleys, lakes, and coastlines. Larger domain models will often perform best in regimes with strong westerly flow and fast-moving storm systems.

Similar differences typically exist between models with respect to vertical resolution, particularly in the boundary layer. Greater vertical resolution generally leads to more accurate representation of temperature inversions, unstable layers, and the vertical distribution of moisture. There can be significant differences between models in their treatment of moist processes such as convection and stratiform precipitation due to differences in their vertical resolution. It also can drastically affect how a model predicts temperature and winds near the surface. Models also differ in their vertical coordinate system and whether the basic equations assume hydrostatic balance or not. Some vertical coordinate systems are terrain-following; others are not. Some dynamically place the greatest resolution in regions of large potential temperature gradient; others are fixed regardless of the situation. The vertical coordinate can lead to different model characteristics for phenomena such as mountain-induced windstorms or cyclogenesis that takes place in the lee of orographic barriers.

Different methods of representing physical processes are used in the various models available to forecasters. Processes that cannot be resolved explicitly within the model's formulation of the primitive equations are represented through what are called parameterization schemes. These are basically subroutines that are invoked throughout the model run that use "state" variables that are explicitly predicted, such as wind, temperature, pressure, and some form of moisture, to handle such phenomena and conditions as soil moisture, moist and dry convection, shortwave and longwave radiation, evapotranspiration from vegetation, turbulence, and so forth. The parameterization schemes are designed to feedback into the model and alter the state variables. Although many of these schemes are quite complex, an understanding of some of the simpler schemes can be used effectively by forecasters in their subjective use of model output.

Model output is now widely available in gridded form. Prior to the 1990s, only a very limited set of graphics at specific levels were typically available. These were, to a great extent, the same set of graphics that were produced via manual methods prior to the NWP era. Gridded model output allows forecasters to examine the model solution in great detail, and in particular it allows forecasters to gain insight into the three-dimensional structures of the model's simulated atmosphere. Software and computational resources exist that allow the output to be viewed via three-dimensional renderings, but at the time of this writing these resources are not widely in use yet, and most model output is still viewed via two-dimensional graphics, such as plan views, cross sections, time–height sections, and forecast soundings.

Forecasters will apply their knowledge of the important physical processes gleaned from the analysis step to examination of the model output. Knowledge of local conditions plays a role in the forecaster's decision about which processes are most important in each situation, and this becomes a significant factor since the strengths and weaknesses of different models are considered. If flow interaction with terrain is a key consideration for a given event, such as in cases of orographic precipitation, then a model with high horizontal resolution might be favored. If lake effect clouds and precipitation are significant forecast problems, then a model that correctly initializes the areal extent of the lake and its water temperature might be favored. If two models are forecasting a trough to move through a mean longwave ridge position with the same timing, but with different depths, the larger domain model that has a better representation of the planetary-scale waves might be preferred. If the forecaster knows that precipitation has recently occurred and near-surface conditions are moist, then the model that has most correctly predicted precipitation in recent model runs might be expected to have the best representation of soil moisture and boundary layer conditions. Each situation must be evaluated and forecasters must apply their knowledge of initial conditions and model characteristics to decide, when possible, which model solution offers the greatest skill and utility.

The final step in the deterministic use of model output for forecasting is the derivation of sensible weather. This includes parameters such as precipitation amount, onset, end, type, and areal distribution, potential for severe convection, lightning activity, sustained wind speed and gusts at the surface, wind shift timing, cloud cover, cloud layers, visibility, areas of turbulence, icing, blowing snow, blowing dust, temperature, wind chill, heat index, wave and swell height, sea-ice accumulation on ships, and so forth, depending on the type of forecasting that is being done. Statistical postprocessing of model output exists that attempts to derive sensible weather elements directly from model output. In the United States, some of these statistical methods are called model output statistics (MOS). The MOS regression equations are derived by comparing past model forecasts with subsequent observations at individual locations. This method has the advantage that it can correct for model biases. It has the disadvantage that in the modern era operational models are undergoing nearly constant change, and the developmental data set for MOS techniques may not be representative of the current incarnation of the model. There are other objective algorithms that convert model output into sensible weather elements and more are always being developed, but a complete discussion of this topic is beyond the scope of this chapter. A popular display of this type of algorithm output is known as a meteogram, where a group of time series of derived sensible weather is displayed graphically in the same way that a graphic of observations might look. In general, experienced forecasters are quite expert at deriving the sensible weather from model output, once they have applied the knowledge and techniques described above to determine which model output is most appropriate for a given situation.

In deterministic forecasting, a forecaster will consider a number of different model solutions during the forecast process. In essence the forecaster is examining a small ensemble of forecasts. The number is typically five or fewer. The forecaster may be looking for model-to-model consistency and/or run-to-run consistency from a given model as an indication of the uncertainty associated with a given forecast. Ensemble forecasting techniques are based on the idea that there is inherent uncertainty in the initial conditions and perhaps similar uncertainty about which is the optimal model configuration for a given situation. The goal of ensemble forecast methods is to quantify the uncertainty by running many versions of NWP models with either perturbed initial conditions and/or versions of models with different parameterization schemes, different basic configurations, or different analysis schemes. The method of perturbing the initial conditions, or the basis for deciding the makeup of the ensemble members, is beyond the scope of this discussion, but two key premises are that each member has an equal probability of being correct and that the more ensemble members included the more skillful the probability distribution of the output.

Ensemble techniques are being developed, both as an alternative and as an adjunct to the deterministic forecast method. Although experienced forecasters are very skillful at deriving sensible weather from model output, the task of determining which model output is best in a given situation, and estimating the magnitude and shape of the model's most likely error is quite difficult. Some would argue that this task is essentially impossible. This last statement is a topic of active debate. Another reason for ensemble techniques is that an increasing number of users of forecast information, especially those who must make large resource commitments based on the weather forecast, require a quantification of the uncertainty associated with the forecast. While deterministic forecasts often include uncertainty, such as probability of precipitation statements, and narrative wording that indicates alternative scenarios, the method does not lend itself to quantification of uncertainty of all parameters in time and space, and certainly not with consistency from forecaster to forecaster. A strength of ensemble techniques is that probabilistic output falls out quite naturally, offering consistent quantitative assessment of uncertainty. Of course, skillful algorithms that convert model output to sensible weather elements must exist, and this is still problematic in many situations.

Ensemble forecasting is a rapidly emerging technique. It could become the primary method of weather forecasting in the future. A description of modern weather forecasting in 2010 may look back on deterministic forecasting methods as something from a bygone era, much as today we look back on the pre-NWP era with nostalgia. More likely, some forecast applications will lend themselves to ensemble techniques, especially where a sophisticated user community exists, while other applications will be best accomplished by deterministic methods for the foreseeable future.

4 FORECASTING 0 TO 12 h

Output from numerical prediction models is of limited use in short-term forecasting for a number of reasons. The most obvious reason is that the collection of observations, quality control of the observations, objective analysis, the execution of the numerical model, and the distribution of the model output takes time. Usually these tasks take on the order of 1 to 3 h. Forecasts for this time frame will have to be made without the benefit of new model guidance, although output from a previous model run may still be applicable. Another reason NWP is of limited use in the short-term is that the initial conditions of numerical models do not yet include ongoing precipitation processes. This is an area of research and development and progress is being made, but, presently, there is a so-called spin-up period, typically 3 to 6 h, before a modern numerical model can produce realistic precipitation. Perhaps the most important reason that NWP is of limited use for short-term forecasting is that the atmospheric phenomena of interest are often not explicitly resolved by current NWP models. Predictions of whether an individual thunderstorm will produce large hail or not and where such a storm is likely to track in the next hour are examples of short-term forecasting for which current NWP offers little or no guidance, although research efforts are ongoing in this area.

The short-term forecast process begins much like the longer-term forecast process, the forecaster must understand the current state of the atmosphere. Analysis is even more important for short-term, small-scale forecasting. Objective and manual analyses are useful, and particular emphasis must be placed on careful examination of individual observations. Because specific observations may be smoothed by objective analysis schemes, or even eliminated by quality control algorithms, manual analysis is usually considered a routine part of short-term forecasting, particularly during the warm season when gradients are weak and important features may be subtle.

Most short-term forecasting deals with phenomena on the mesoscale. However, as part of the analysis step, forecasters must consider scale interaction. Although it is debatable whether planetary-scale considerations are important for short-term forecasting, there is no doubt that synoptic-scale conditions are critical to the evolution of mesoscale phenomena. The location and evolution of synoptic-scale features such as frontal boundaries, troughs and ridges aloft, regions of warm and cold temperature advection, high- and low-level jets, and the distribution of moisture in the horizontal and vertical is necessary knowledge for any forecaster attempting to predict mesoscale phenomena.

Because short-term forecasting does not have the benefit of primitive equation model output, it is absolutely critical that the forecaster understand the current atmospheric processes and phenomena. A prediction of conditions 1 to 3 h into the future will generally fail if the current conditions are misdiagnosed. The forecaster must decide whether conditions are evolving linearly or via nonlinear processes. An example of a linear forecast would be extrapolation. A line of rain showers is moving east at 10 mph and is expected to continue east with no change in speed, direction, or intensity. Again, modern graphic/image workstations allow tracking and extrapolation of features on radar and satellite imagery to make even this simple extrapolation forecast more accurate. However, if the forecaster fails to note that this line of rain showers is moving into an environment of greater instability and favorable vertical wind shear, then the nonlinear transformation of the line of showers into a severe squall line could be completely missed in the short-term forecast.

Conceptual models play a very important role in short-term forecasting, perhaps even more so than in longer-term situations because of the lack of NWP output. A good example of applying a conceptual model to a short-term forecast problem can be illustrated for the prediction of supercell thunderstorms. Once thunderstorms have started to develop over an area, the short-term forecast for that area can be very different depending on the character of the convection. Ordinary thunderstorms will have a life cycle of only an hour or so, they will generally not produce large hail or tornadoes, and they will tend to move in a straight line. In contrast, supercell thunderstorm life cycles can be many hours, they are likely to produce severe weather, including large hail and possible tornadoes, and their movement will likely be to the right or left of the movement of the ordinary cells. Clearly, the short-term forecast will be quite different if supercell thunderstorms are expected as opposed to ordinary thunderstorms.

Considerable research has shown that supercell thunderstorms develop in regions with at least 20 m/s of vertical wind shear in the 0 to 6 km depth above ground, and with adequate instability to support deep convection. The basic conceptual model of supercells is that the combination of buoyancy and shear allow a thunderstorm structure to become organized such that the updraft and downdraft do not interfere with each other, as is typically the case in ordinary thunderstorms. The result is a long-lived storm with very strong updrafts and downdrafts, and pressure perturbations that lead to rotation within the storm and propagation rather than just advection of the thunderstorm.

A forecaster will first examine the environment to determine if supercell thunderstorms are possible. Output from NWP can provide guidance about buoyancy and shear. However, detailed analysis of direct and remotely sensed observations is required to identify pertinent features such as old frontal zones, outflow boundaries, gradients in surface characteristics, and diabatic heating gradients due to cloud cover that may modulate the shear and instability on the mesoscale. Upon completion of this analysis the forecaster will know whether supercell thunderstorms are possible and also if some areas are more favorable than others for their development. The second step in this process is to use Doppler radar data to carefully examine the three-dimensional structure of thunderstorms as they develop. The forecaster will look for features such as tilted reflectivity cores, weak-echo regions, bounded weak-echo regions, rotation in the velocity data in the region of the storm updraft, and anomalous storm movement. These are structures that have been documented in research results and are the basis of current conceptual models of supercells. Once identified, appropriate forecasts and warnings can be issued for these storms.

A similar forecast process is used for virtually all short-term forecasting that attempts to predict the evolution of mesoscale phenomena. A precipitation band in a cool-season situation can be caused by many different conditions. A short-term forecast of the behavior of the band can only be made if the forecaster understands the structure and processes responsible for its existence. The short-term forecast will be very different depending upon whether the band is due to conditional symmetric instability, or an internal gravity wave, or postfrontal lake effect, or orographic lift, or low-level convergence.

Ensemble techniques may eventually play a role in the 0- to 12-h forecast process. Often, the differences between environmental conditions that can support certain mesoscale processes and those that cannot are quite subtle. An ensemble of models may be able to offer forecasters insight into which situations are borderline and which are not. Again, the value of the ensemble technique is the quantification of uncertainty.

SECTION 5

MEASUREMENTS

Contributing Editor: Thomas J. Lockhart

CHAPTER 33

ATMOSPHERIC MEASUREMENTS

THOMAS J. LOCKHART

The technology applied to sensing of atmospheric conditions and content improves constantly and at an ever-increasing rate as newly available technologies are applied to old, familiar questions. The increasing capability and decreasing price of personal computers is both a case in point and a cause of expanding technological applications. The best venues for describing and discussing advanced instrumentation are the technical conferences and peer-reviewed journals. The thrust of this section will be toward common applications and common concerns with respect to the representativeness and uncertainty of measurements. The contributors to the measurements section have practical experience dealing with such questions. The goal of this chapter is to describe the problems, compromises, and pitfalls with measurement programs of many kinds. This section will also include some examples of program failures, which seldom reach the readers of technical journals.

In centuries past, instrument design, siting, and data application began with individuals. Curious individuals used measurements to document their observations and to help find explanations for what was happening. This led to a need to duplicate instruments and introduced a commercial opportunity. As communication technology developed, the ability to assemble synchronous observations became possible. Synoptic meteorology generating forecasts and warnings followed. National weather services were born and skills in measurement technology with the necessary set of specifications became a specialty.

It is a fundamental principle that the application of data rests on an understanding of the measurement process and the size of the uncertainty or error bars. It is therefore incumbent on the organization taking the data to retain expertise in these disciplines. Those with such expertise usually are thinking about improvements and investigating new technology. This has led to the various national

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

laboratories and research organizations formed to pursue technological and application progress.

The atomic age introduced a need for radioactive fallout predictions. In recent decades (since the 1960s) federal regulations have driven many of the applications of special measurement programs in the atmospheric surface layer. First, in response to the U.S. Department of Defense (DOD) Army chemical warfare research programs, a need for better understanding of dispersion or diffusion processes was recognized. The Nuclear Regulatory Commission (NRC) required accurate measurements and a quality program to assure compliance with the requirements stated in the NRC Safety Guide 23. A group of volunteers working within the auspices of the American Nuclear Society (ANS) composed a standard (ANS 2.5) that defined the details necessary to meet the ANS requirements for measurement accuracy and the methods for quality assurance.

The U.S. Environmental Protection Agency (EPA) provided similar guidelines in response to the needs of the 1967 Clean Air Act. The EPA requirements were quite similar to those of the NRC. Clearly, a vacuum of measurement standards existed. No one could cite consensus standards to justify claims of data compliance. These two lists of requirements, with the force of the federal government behind them, brought the measurement community to action. It was immediately clear that standard definitions and test methods were necessary to determine if the requirements were being met by the measurement systems used.

The American Meteorological Society's (AMS) Committee on Atmospheric Measurements (CAM) considered the need to write standard methods to determine the various characteristics that were being specified. After due consideration and conversations between the AMS executive secretary and his counterpart in the American Society for Testing and Materials (ASTM), it was decided to form a subcommittee within ASTM D22, Sampling and Analysis of Atmospheres. Subcommittee D22.11, Meteorology, was formed in 1972. The work of volunteer members of D22.11 provided some of the early standards, but the process is open ended. ASTM has a requirement of a 5-year review and reapproval process. New standards are generated as the need, individual willingness to contribute, and expertise come together. This national program is now contributing to a similar international program through ISO (International Standards Organization) TC 146 (Technical Committee 146, Air Quality) SC 5 (Subcommittee 5, Meteorology).

It is generally recognized that approaching perfection in the definitions and performance of instruments is reducing the least important contributor to measurement uncertainty. Take wind speed as an example. A perfect anemometer will only suggest what the wind speed would be if the calibration facility were a true analog to the turbulent environment. Its performance has been tested in the wind tunnel. Its response to a speed change (sensitivity to inertia in a mechanical design) is defined in terms of its distance constant, defined by a standard test using a speed step from zero to the wind tunnel speed. The time constant, $1 - 1/e$, at that speed is measured between the 30 and 74% points (of maximum value) during its recovery to avoid the stall condition at release. The test is run at two or more speeds, which documents that the response is constant with distance rather than time.

It has been recognized that the distance constant is different for a step change from a higher speed to a lower speed than for a step change from lower to higher speeds. The ratio of the two distance constants will predict the "overspeeding" of a cup anemometer. This overspeeding is the inertial effect that causes the anemometer to speed up faster than it slows down, creating an average that is slightly larger in turbulent flow than in laminar flow. The word *overspeeding* has also been used to describe the difference between the scalar average and the resultant vector average of samples taken over a period of time.

There is further confusion when the effect of off-axis or nonhorizontal flow is considered. Cup anemometers will usually turn a little faster when the angle of the flow is not quite horizontal. This is an aerodynamic or lift-drag effect that can be simulated in the wind tunnel. When the distance constant model, based on the wind tunnel test, is used to estimate the underreporting of average speed in turbulent flow, it fails. The step change in speed does not describe the speed changes in atmospheric flow. While the model might predict a 20% error, the measurements in the atmosphere show little or no effect.

There are effects, however, that do bias the measurements. Mounting hardware and/or supporting structures may influence the speed. Nearby trees or buildings will alter the flow. While the anemometer may faithfully report the speed where it is mounted, it may not represent an area as large as the application might require. The overall subject of representativeness will usually contain much more uncertainty than the uncertainty in the transfer function (relationship between wind tunnel speed and measured rate of rotation) of the anemometer. However, there is a general rule of thumb that suggests one should not ignore those parts of the uncertainty analysis that can be understood simply because there are other parts that are larger and more difficult to define.

There are a broad variety of atmospheric variables that can be measured and an equally broad variety of applications that need these measurements, each with their own unique requirements. The contributors to this section will touch on many of these. It remains the responsibility of the person who uses atmospheric measurements to understand, to whatever degree possible or practical, the process by which the measurements were made. This knowledge should range from the engineering specifications of the instrument to the siting biases that might contribute to representativeness of samples.

CHAPTER 34

CHALLENGES OF MEASUREMENTS

THOMAS J. LOCKHART

1 ETHICS

If there is one absolute in the ethics of measurement, it is that the data must speak for themselves. But what can numbers say? If one applies standard statistical processes to a series of numbers, the answer is reproducible. Of course, there is some control in selecting a subset of a series of numbers (see "All That Is Labeled Data Is Not Gold" below). Even if a time series is continuous, the beginning point and the ending point can influence the outcome. There is nothing unethical in presenting a continuous subset of data and listing their statistical parameters, if the fact that it was a selected subset is disclosed. When the parameters are used to make a political (or societal) point, the ice is thinner.

A classic example of mixing true science with "political" science comes from comments made by Stephen Schneider (1987), a proponent of the theory that cholrofluorocarbons (CFCs) are depleting the ozone layer. He said: "We have to offer up scary scenarios, make simplified, dramatic statements, and make little mention of any doubts we may have. Each of us has to decide what the right balance is between being effective and being honest." Those of us who subscribe to the ethics of measurement will have no problem with that decision. Science, including its subset measurement, requires honesty. There is no type-A or type-B honesty. If any consideration influences how numbers are gathered or generalized, that consideration must be defendable without subjectivity.

The measurement community often expresses a genuine skepticism when considering the results of model simulations. The assumptions used for the model simplifications and for the "data" inputs deserve scrutiny. When simulations go beyond the data in time, they become a forecast. There is real value in general circulation models, and

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

there has been real improvement in most aspects of weather forecasting as a result of model outputs. However, it will be some time, if ever, before models will contain the intelligence exhibited by local meteorologists (or local farmers) in local short-term forecasts.

When assumptions go beyond the data in range, it is a guess. If anemometers are calibrated to 50 m/s and report speeds of 70 m/s, it is impossible to know the uncertainty of the measurement (unless some postanalysis is conducted). Extrapolation beyond experience is often the only course available, but a footnote is required to warn the user of that fact. Extrapolation between measurement points may not agree with new measurements specifically located to test the extrapolation. The measurement deserves the presumption of accuracy. There are instrument tests that can confirm the instrument performance. The smoothed model value should not be presumed to be correct.

It is true that caveats may adversely affect the literary quality of pronouncements. Perhaps there should be a sharp division between the statements of "scientists" on political issues and scientists reporting the results of their fundamental research. The public is not capable of sorting the "results" from a reputable scientist that are biased for advocacy from the results from a reputable scientist that contain all the dull but critical caveats that qualify the data supporting the results. It used to be that ethics took care of the problem. If ethics no longer apply, a method needs to be found to warn the public of the advocacy role of "political" scientists.

All That Is Labeled Data Is Not Gold

> Most of us are quite comfortable with the idea of evolution in the animal kingdom or in climate, by which small departures from the norm can spread and eventually change the entire picture. We are not as ready to accept a similar evolution in classical data sets, the gold standard of climatology.
>
> Recently, I have been using a small subset of the data generated by the GEOSECS (Geochemical Ocean Sections) program (1973–1974), in particular, the ocean profiles of oxygen-18 and deuterium measured by Harmon Craig at Scripps Institution of Oceanography in La Jolla, Calif. Until recently, I was secure in the knowledge that the data I was using, gathered from the National Climate Data Center repository, were complete and unadulterated. Earlier this year, while presenting my results and using these data for comparison, a member of the audience asked why I was not using the complete data set. Confused, I defended myself vigorously. Later, it appeared that although I had correctly presented the published data, there was an underground version of unpublished data that was known only to a small circle of initiates. My hosts were kind enough to include me among the cognoscenti and it was at this point that things started to become interesting.
>
> Examining the new version, I started to find many inconsistencies in the two data sets: values different in one than the other, stations appearing in one or the other but not both, different measurement depths, and even different positions for the stations. Puzzled, I went further afield in search of information that could resolve the conflicts.

Looking at the published tables, I worryingly found that the mistakes in station positions seemed to be due to typographical errors.

Soon, though, I found someone else who had a subset of the unpublished data. This surely would clear things up. Unfortunately, it made things even worse. Where this data set purported to give the same data, it was different from both previous versions in seemingly random ways.

In despair, I tried to track down the source of the unpublished data. Finally, I found someone who had an old photocopy of an old-style computer printout whose provenance was claimed to be Harmon Craig's original measurements (including repeats). At last, truth!

It became clear very quickly that this printout was the original version of the unpublished data, and it was equally clear that the data had at some point been typed in by hand. The errors were of the sort caused by inadvertent line skipping, missing decimal points and minus signs, and incorrect averaging of repeat measurements. However, it was also evident that this was the source for the published tables. The published data though, had been subjected to some quality control; outliers had been thrown out, repeat samples with too much dispersion ignored, and possibly, further repeats had been performed. The measurement depths had clearly been refined using more accurate information.

The origin of the third data set remained mysterious until I was informed that it had been traced from a graph showing the data in an old article and then correlated to the original stations. That explained the small but random differences. Only with all this scientific detective work and the discovery of the ancestral data was I now able to amalgamate the published and unpublished results into a consistent data set.

These data are only 25 or so years old and yet there are at least three (maybe more) different and mutually inconsistent versions floating around. The lesson we should take from this is that, if unchecked, small mutations will occur during transcriptions. Unless we are vigilant, "data sets" will evolve and data that have been so painstakingly collected and saved will be distorted and twisted beyond all usefulness. Our care in using data should be a force analogous to natural selection, keeping data sets fit for the purposes intended. If we do not take care, the gold standards of climatology may in time turn to lead.

Gavin Schmidt, NASA Goddard Institute for Space Studies, New York; E-mail: gschmidt@giss.nasa.gov.

2 GENERATION AND IMPLEMENTATION OF A CONSENSUS ON STANDARDS FOR SURFACE WIND OBSERVATIONS

Generation

Robert L. Carnahan, the retired federal coordinator from the Office of the Federal Coordinator for Meteorology (OFCM), contacted the author in 1992 regarding a wind measurement question. Contact was arranged with David Rodenhuis, chairman of the Ad Hoc Group and the Interdepartmental Committee for Meteorological Services and Supporting Research (ICMSSR) on wind observing standards. The author was asked to assist Rodenhuis in organizing a workshop and presiding at the workshop; a plan was devised.

The purpose of the workshop was to seek a consensus standard method of characterizing surface wind measurements that would serve all applications of wind measurement. Two preparatory tasks were essential. First, a group of knowledgeable attendees to represent all applications had to be identified. Second, a syllabus needed to be prepared to set the background for the two-day workshop.

A list of 79 experts was assembled, and, of those, 40 attended; the 39 who could not attend were sent all the materials and reports and had the opportunity to participate by mail. The group included specialists from such applications as agriculture, aviation, climatology, forecasting and warnings—National Weather Service (NWS), manufacturers of instruments and systems, measurements, military, oceanography, buoy operation, and hurricane research, insurance, transport and diffusion, air quality, network operations, and quality control, and wind energy. The names, affiliations, and specialties of all participants are available from the OFCM.

The two-day workshop produced a group consensus, of which the most important elements are listed below.

- The metadata (information about the measurement system and the siting of the sensors) must be available where data are archived. Included will be the following: station name and identification number, station location in longitude and latitude or equivalent, sensor type, first day of continuous operation, sensor height, surface roughness analysis by sector and date, site photographs with 5-year updates, tower size and distance of sensors from centerline, size and bearing to nearby obstructions to flow, measurement system description with model and serial numbers, date and results of calibrations and audits, date and description of repairs and upgrades, data flowchart with sampling rates and averaging methods, statement of exceptions to standard requirements, and software documentation of all generated statistics.
- The basic period of time for surface wind observations should be 10 min. This period provides reasonable time resolution for continuous process analysis (air quality, wind energy, and research) while providing the building blocks to assemble longer averaging periods. Many current boundary layer models require hourly data. However, there is a growing world consensus for 10-min data periods. The periods must be synchronous with the Universal Time Clock (UTC) and labeled with the ending time. If a different time label is used, it must be stated and attached to the data.
- The operating range is either sensitive or ruggedized. The threshold and maximum speed of the sensitive range is 0.5 and 50 m/s. The threshold and maximum speed of the ruggedized range is 1.0 and 90 m/s.
- The dynamic-response characteristics for either type are as follows. The anemometer will have a distance constant of less than 5 m. The wind vane

will have a damping ratio greater than 0.3 and a damped natural wavelength of less than 10 m.

- The measurement uncertainty for wind speed is ±0.5 m/s below 10 m/s and 5% of the reading above 10 m/s. For wind direction, the accuracy to true north is ±5°.
- The measurement resolution is 0.1 m/s for both average speed and the standard deviation of the wind speed. Resolution for average direction is 1° and for the standard deviation of wind direction is 0.1°.
- The sampling rate will be from 1 to 3 s. Sampling for wind speed must be capable of achieving a credible 3-s average.
- The standard data output for archives will include the following:

  Ten-minute scalar-averaged wind speed

  Ten-minute unit vector or scalar-averaged wind direction

  Fastest 3-s gust during the 10-min period

  Time of the fastest 3-s gust

  Fastest 1-min scalar-averaged wind speed during the period

  Average wind direction for the fastest 1-min speed

  Standard deviation of the wind speed samples about the 10-min mean speed

  Standard deviation of the wind direction samples about the 10-min mean direction.

- Special guidance is offered for survivability to assure the support and power will not fail before the sensor itself fails. A suggestion is offered for preserving high-resolution data when the wind is above 20 m/s.
- Quality assurance is required with documentation in the site log.
- A specific method, based on Wieringa (1992), was described.

The consensus was reached with caveats from some members, generally preferring tighter requirements. However, this method was designed as a minimum that all applications could use. Wind energy, for example, needs greater accuracy. There is no reason why special networks or stations could not meet this specification and still provide greater accuracy. Uncertainty is largely a result of calibration. The hope of all the attendees was that the NWS would adopt the consensus and implement it in the new ASOS (Automated Surface Observing System) deployment. Implementation of new technology provides opportunities to improve the value of measurements. Digital data systems can preprocess samples. There is value in the potential richness of data that this consensus could provide as compared to manual observations. Technology has also provided capability to archive the richer data. The National Climatic Data Center (NCDC) representative joined the consensus and agreed that the data could be archived. All that remained was implementation.

Implementation

There was immediate implementation by some members of the workshop. The Oklahoma Mesonet used a version of the consensus in its new (at the time) statewide network of automatic stations. Campbell Scientific Inc. designed an instruction for its data loggers that would generate the required outputs each 10 min. The American Society of Civil Engineers (ASCE) adopted the 3-s average speed at 10 m as its basic wind speed (since the fastest mile has been discontinued). It has been published in its Minimum Design Loads for Buildings and Other Structures ANSI/ASCE 7-95.

Since there now was no expectation that the consensus would be used by the NWS, its essence was introduced to the D22.11, Meteorology, subcommittee of the American Society for Testing and Materials (ASTM). After the normal open process of consideration, modification, and balloting, the Standard Practice for Characterizing Surface Wind Using a Wind Vane and Rotating Anemometer D 5741-96 was adopted. As of 2001, this standard had not been adopted by the OFCM.

One element of the consensus is the definition of peak speed or gust. The ASOS provides a peak 5-s average. This clock-driven averaging period provides a smaller speed than the F420 anemometer attached to a dial or galvanometer recorder has provided during the past several decades. This human observation was considered "instantaneous" since there was no intentional averaging between the generator in the sensor and the dial or recorder in the office. The dial had to be seen at the moment of maximum value. The strip chart records the speed, but the galvanometer does have a frequency response. The damping by the recorder has been estimated as approximately equivalent with a 1- to 2-s average.

It became clear that, from a climate continuity standpoint, the change from the conventional peak speed to the ASOS peak speed decreased the value reported. This could have ramifications for storm or wind warnings. It is also clear that changing from a clock 5-s average to a running 3-s average would move the peak speed value closer to the conventional measurements. The most compelling reason for adopting the running 3-s average for peak speed is the fact that, without standards, the concept of peak speed has no meaning.

In the past, technology was such that whatever one could get was acceptable. Any measurement was better than no measurement. The totalizing anemometer had to be read where the anemometer was mounted. Gears turned a scale not unlike an electric meter. An interim step was the fastest-mile anemometer, which could drive an event pen on a strip chart in the office, thereby recording the time of each mile (so many revolutions of the cup) of air that passed. Time per mile could be converted to miles per hour by using an overlay with different line spacings labeled as speed or by counting the miles that passed in an hour for an hourly average.

With the advent of the cup anemometer driving an electric generator, the measurement could be observed in real time in the office, first on a dial and then on a dial and a chart recorder. When automatic weather stations came into use, the questions of digital sampling and averaging had to be answered. The Federal Aviation Administration (FAA) funded the National Oceanic and Atmospheric Administration (NOAA) and the NWS to make recommendations appropriate for aircraft operations for the ASOS design. The 5-s average for peak speed came from this study. Many other countries also considered the same questions. In the World Meteorological Organization (WMO), the Commission for Instruments and Methods of Observation (CIMO) adopted the 3-s average as the standard for peak wind speed.

It is hoped that our national network of surface weather stations will eventually adopt the consensus standard. In the interim, a campaign is underway to change the peak speed from the clock 5-s average to a running 3-s average on the basis of climate continuity. The American Association of State Climatologists (AASC) voted in their August, 1998, meeting to adopt the running 3-s average for climatological purposes.

3 UNCERTAINTY, ACCURACY, PRECISION, AND BIAS

A critical component of a chapter of this kind is a careful consideration of the meaning of the language used. There are two levels of language in this situation. One is the precise and unambiguous words used by the professionals in the standards field. The other is the routine communication used in the application of technology in a regulated society. A convergence of these two is the goal of this chapter.

Accuracy is a word commonly used in the specification of regulations, procurement documents, and manufacturer's data sheets. According to ISO (1993), measurement accuracy is simply the closeness of the agreement between the result of a measurement and a true value of the measurand. Accuracy is therefore a qualitative concept since the true value must remain indeterminate.

The concept of uncertainty, as carefully defined by standards professionals, provides a method to quantitatively characterize all doubt about the validity of the value of a measurement. There are two types of evaluation methods. Type A is an evaluation of a statistical analysis of a series of observations of the measurand. Type B is an evaluation of other information about the measurement. These two types of evaluations provide what is called the "combined standard uncertainty."

This chapter includes an introduction to these concepts by listing a number of definitions found in the 1993 report by the International Organization for Standardization (ISO). Points that always need reinforcement are also discussed. These include the practices of rounding and the display of significant figures. All of these subjects will be discussed with examples drawn from the measurement of air temperature at 1.5 m above the ground surface.

Background

An effort to find an international consensus on the expression of uncertainty in measurements began in 1978. The world's highest authority in metrology (measurement science, not atmospheric science), the Comité International des Poids et Mesures (CIPM) asked the Bureau International des Poids et Mesures (BIPM) to make recommendations. After consulting the national standards laboratories of 21 countries, a working group was established. This working group assigned the responsibility to the ISO Technical Advisory Group on Metrology (TAG 4). It, in turn, established Working Group 3 (ISO/TAG 4/WG 3) composed of experts nominated by BIPM IEC (International Electrotechnical Commission), ISO, OIML (International Organization of Legal Metrology), and appointed by the chairman of TAG 4.

Working Group 3 was assigned the following terms of reference:

> To develop a guidance document based upon the recommendations of the BIPM Working Group on the Statement of Uncertainties which provide rules on the expression of measurement uncertainty for use within standardization, calibration, laboratory accreditation, and metrology services:
>
> The purpose of such guidance is
>
> - to promote full information on how uncertainty statements are arrived at;
> - to provide a basis for the international comparison of measurement results.

ISO (1993) is the current guide from WG 3.

Definitions

The first paragraph of the introduction to the *Guide* (ISO, 1993) is reproduced to explain the purpose of the *Guide.*

> When reporting the result of a measurement of a physical quantity, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its reliability. Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a specification or standard. It is therefore necessary that there be readily implemented, easily understood, and generally accepted procedures for characterizing the quality of a result of a measurement, that is, for evaluating and expressing its *uncertainty.*

In the scope of the *Guide* (ISO, 1993) is the following:

> This document, hereafter called the *Guide*, establishes general rules for evaluating and expressing uncertainty in physical measurement that can be followed at various levels of accuracy and in many fields—from the shop floor to fundamental research. Therefore, the principles of this *Guide* are intended to be applicable to a broad spectrum of measurements, including those required for:

- maintaining quality control and quality assurance in production;
- complying with and enforcing laws and regulations;
- conducting basic research, and applied research and development, in science and engineering;
- calibrating standards and instruments and performing tests throughout a national measurement system in order to achieve traceability to national standards; and

- developing, maintaining, and comparing international and national physical reference standards, including reference materials.

Annex B of the *Guide* (ISO, 1993) contains definitions of general metrological terms that will be repeated here, with air temperature examples where appropriate.

B.1 (measurable) **quantity**
attribute of a phenomenon, body or substance that may be distinguished qualitatively and determined quantitatively

EXAMPLE: the motion of air molecules interpreted as temperature.

B.2 **value** (of a quantity)
magnitude of a specific quantity generally expressed as a unit of measurement multiplied by a number

EXAMPLE: the temperature of the boiling point of water (at 1 atmosphere) is 100 °C.

B.3 **true value** (of a quantity)
value perfectly consistent with the definition of a given specific quantity

EXAMPLE: the theoretical absolute zero of 0 Kelvins.

B.4 **conventional true value** (of a quantity)

EXAMPLE: the temperature of air when there is no molecular motion is -273.15 ± 0.01°C.

B.5 **measurement**
set of operations having the object of determining a value of a quantity

EXAMPLE: the function of a temperature measurement system or an observer with a thermometer taking a reading.

B.6 **principle of measurement**
scientific basis of a method of measurement

EXAMPLE: the thermoelectric effect applied to the measurement of temperature.

B.7 **method of measurement**
logical sequence of operations, in generic terms, used in the performance of measurements according to a given principle

EXAMPLE: the resistance of a wire, exposed in an aspirated radiation shield, determined by a substitution method.

B.8 **measurement procedure**
set of operations, in specific terms, used in the performance of particular measurements according to a given method

EXAMPLE: a standard operating procedure used by an operator to obtain a measurement or manage an automatic measurement system.

B.9 **measurement process**
all the information, equipment and operations relevant to a given measurement.

(NOTE—This concept embraces all aspects relating to the performance and quality of the measurement; it includes for example the principle, method, procedure, values of the influence quantities, and the measurement standards.)

B.10 **measurand**
specific quantity subject to measurement

EXAMPLE: The temperature of the air at a specific location, often called a variable.

B.11 **influence quantity**
quantity that is not included in the specification of the measurand but that nonetheless affects the result of the measurement

EXAMPLE: the effect of solar or long-wave radiation, wind speed, water (in any phase), insects, environmental influence on measurement circuits, and the condition of any moving parts and connections.

B.12 **result of a measurement**
value attributed to a measurand, obtained by measurement

(NOTES—1. When the term "result of a measurement" is used, it should be made clear whether it refers to: the indication; the uncorrected result; the corrected result; and whether several values are averaged. 2. A complete statement of the result of a measurement includes information about the uncertainty of measurement.)

B.13 **uncorrected result**
result of a measurement before correction for the assumed systematic error

B.14 **corrected result**
result of a measurement after correction for the assumed systematic error

B.15 **accuracy of measurement**
closeness of the agreement between the result of a measurement and a true value of the measurand

(NOTES—1. "Accuracy" is a qualitative concept. 2. The term "precision" should not be used for "accuracy.")

B.16 **repeatability** (of results of measurement)
closeness of the agreement between the results of successive measurements of the same measurand carried out subject to all the following conditions:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time

B.17 **reproducibility** (of results of measurement)
closeness of the agreement between the results of successive measurements of the same measurand, where the measurements are carried out under changed conditions such as:

- principle or method of measurement
- observer
- measuring instrument
- location
- conditions of use
- time

B.18 **uncertainty** (of measurement)
a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand

> (NOTES—1. The parameter may be, for example, a standard deviation (or a given multiple of it), or the half-width of an interval having a stated level of confidence. 2. Uncertainty of measurement comprises, in general, many components. Some of these components may be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations. The other components, which can also be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information. 3. It is understood that all components of uncertainty, including those arising from systematic effects, such as components associated with corrections and reference standards, contribute to the dispersion.)

B.19 **error** (of measurement)
result of a measurement minus a true value of the measurand

B.20 **relative error** (of measurement)
error of measurement divided by a true value of the measurand

B.21 **random error**
result of a measurement minus the mean result of a large number of repeated measurement of the same measurand

B.22 **systematic error**
mean result of a large number of repeated measurements of the same measurand minus a true value of the measurand.

B.23 **correction**
value that, added algebraically to the uncorrected result of a measurement, compensates for an assumed systematic error

B.24 **correction factor**
numerical factor by which the uncorrected result of a measurement is multiplied to compensate for an assumed systematic error

The annex to the *Guide* (ISO, 1993) provides additional examples and comments. Annex C (not discussed here) provides the basic statistical terms and concepts for making a type A evaluation of uncertainty.

Discussion

The traditional use of the term *accuracy*, among applied meteorologists, is understood to include both a (constant) bias term and a (variable) precision term. These can only be estimated from comparisons with other measurements, calibration tests at a laboratory with accepted authority, or reliance on the manufacturer's data sheet claims. These estimations do NOT include comparisons with the true value nor do they include the influence quantity.

Using the 1.5-m air temperature example, the bias term might be $+0.2\,^{\circ}\mathrm{C}$, based on either a calibration made by the National Institute of Standards and Technology

(NIST) or a manufacturer's calibration sheet that came with the sensor. More commonly, however, the sensor calibration is a variable with calibration temperature points. It should be considered a conditional bias with the bias correction dependent upon the temperature measured.

The precision term might be quite small, perhaps ±0.02 °C based on the standard deviation of a series of measurements with the sensor at equilibrium in a constant temperature thermal mass. This is really the precision of the measurement of resistance with the resistance kept constant by the thermal mass. The size of this precision measurement may be an indicator of the quality of the thermal mass.

A much larger precision will be found if the test uses a collocated transfer standard in the atmosphere. Functional precision (ASTM, 1990) is the root mean square of the distribution of differences in measurements made with two identical systems. The precision will be larger because the two sensors are not in the same thermal mass. It will probably be on the order of a few tenths of a degree Celsius.

The uncertainty will contain the influence quantity. This is like saying that the accuracy estimate is not complete until all of the siting and exposure errors are considered. These are the large errors when deciding how good the measurements are with respect to a specific application.

The influence quantities for the example of air temperature at 1.5 m above the ground include the following:

- The heat gained by the sensor from solar radiation arriving directly or indirectly by conduction, convection or reflection.
- The heat lost by the sensor to the night sky by radiation or conduction.
- Wind speed and, to a lesser extent, wind direction, as a modifying condition to the heat loss of the shield. Possible effect on the forced aspiration rate.
- Cloud cover and time of day as a modifying condition to radiative heat transfer. Also as a contributor to reflected radiation.
- Water droplets in the air or on the shield as a modifier to the air sample being measured. The relative humidity is a secondary effect to evaporation rate.
- Surface condition, including snow cover, as a radiation reflector. Parallax error for manually read liquid-in-glass thermometers.

The list can go on and on. The influence quantities will be different for different measurands or variables.

Measurement instrument output values can be reported to any resolution with modern digital systems. A sample measurement value should not be reported with a resolution finer than the uncertainty of the system. A sample air temperature value should be reported to whole degrees. An average of many samples should be reported to a tenth of a degree. Digital systems should carry sufficient resolution to avoid rounding errors. Maximum or minimum values should be reported to a tenth of a degree in order to preserve relative information. The uncertainty of a measurement will be much smaller in the context of a relative inquiry as opposed to an absolute value. Local measurement networks will have a relative uncertainty of value to the network application.

A recent study (Lockhart, 1996) showed that modern automatic data systems still fail to consider the need to maintain resolution for output values. In Lockhart (1979) it was shown that fastest-mile measurements were taken in integer knots and reported for climatological applications in miles per hour. The unit conversion was made with an integer table where the value of 19 mph could not exist, along with many other values. The modern ASOS (Automated Surface Observing System) takes measurements in knots but supplies daily summary data in miles per hour. The unit conversion method does not have the resolution to allow every integer to appear. There is still no 19 mph.

4 VALUE OF COMMON SENSE

During the process of writing a revision to the *EPA Quality Assurance Handbook for Air Pollution Measurement Systems, Volume IV, Meteorological Measurements* (EPA600/4-90-003), several subjects arose that could be resolved only by appeal to common sense. In the more than 30 workshops that have followed the publication of Volume IV, one appeal that is always made is to use common sense. When all else fails, common sense is a good standard.

For example, one always finds the requirement that calibrations be "traceable" to National Bureau of Standards (NBS) [now National Institute of Standards and Technology (NIST)]. What does this mean and how does one document the compliance with the requirement? The common interpretation is that there needs to be some anemometer calibration at NIST that starts the transfer standard path. It might go to another wind tunnel where the NIST calibration is transferred to the new wind tunnel. A calibration in the new wind tunnel might be further transferred to a third wind tunnel. A series of documents could be in the file that records all these calibrations. The operator of the third wind tunnel might provide calibrations traceable to NIST.

A paper trail is not enough when there is a possibility of error in the process. Two cases come to mind. The calibration at NIST has some uncertainty (NIST). Occasionally there is an outlier. The prudent summarization of a NIST calibration is a linear regression analysis, looking critically at differences at each point. Most mechanical anemometers are linear once the nonlinear starting speeds are passed. Outliers or problems with the calibration can usually be seen with this analysis. One wind tunnel operator transferred each point of the NIST calibration to a new anemometer, even the one obvious outlier in the NIST data. This was defendable on paper but failed the common sense test.

Technology is always improving. Cup anemometers block some of the flow in a wind tunnel. How much is not well known. The amount varies with the design of the anemometer but also with the size of the wind tunnel test section. Common sense says that a cup anemometer calibrated in the large test section at NIST (3.25 m^2) will have a small blockage error, probably 1% or less. When the calibrated anemometer is used in a smaller wind tunnel with a test section of $0.4\,\text{m}^2$, there will be more blockage. If the calibration method involves transferring the NIST wind speeds to the wind tunnel fan rpm and then transferring the wind tunnel fan rpm to another identical anemometer, the blockage cancels out. Common sense says you can ignore the blockage effect. If, on the other hand, the wind tunnel fan rpm is used to "calibrate" a different kind of anemometer, the relative blockage of each anemometer in the small test section must be known. Since this effect is difficult to quantify, it is common sense to use NIST transfer standards for each type of anemometer of interest.

When, in the past, wet-bulb and dry-bulb temperatures were measured with a sling psychrometer, it was a new technology. Other methods for measuring dew-point temperature and relative humidity allowed for difference comparisons to be made. Then it became clear that siting bias was a big problem for the sling psychrometers. Even if the thermometer is moving rapidly, any incident solar radiation will heat the thermometer. If shade is found, there may still be a problem with reflected radiation. Body heat and humidity can bias the reading by modifying the air if the thermometer is down wind of the operator. Understanding the measurement process and the application of common sense will minimize these errors.

A very accurate air pressure transducer can be calibrated in the laboratory but when it is installed in the atmosphere the wind effect must be considered. Static ports are now available to minimize wind pressure effects, but two or more decades ago the exposure of the pressure transducer was not considered. The transducer would be mounted in a weatherproof box, but the inside box pressure was assumed to be the same as atmospheric pressure. When the wind was blowing, there was a bias, conditional on the speed and direction of the wind, which could be several times the calibration uncertainty. The common sense in this example must be uncommon until the knowledge of such effects reaches the operational meteorologist who must make decisions about the design of the instrument systems.

There is an "official" data archive for the United States at the National Climatic Data Center (NCDC) in Asheville, North Carolina. If one needs a copy of the "official" data for some purpose, often related to some lawsuit, one can get it from NCDC. It comes designated as official and, no doubt, judges and juries are duly impressed. It is true that the copy is certified to be correct by the head of NCDC, but this does not say anything about the accuracy of the data. The National Weather Service is responsible for the accuracy of the measurements it makes, while NCDC is responsible only for faithfully recording the NWS numbers and copying them when required (although NCDC does perform certain quality control checks for continuity, etc.)

EPA and National Research Council (NRC) list the performance specifications required for measurement systems used on projects under their authority. Performance audits are usually required on a periodic schedule to verify that the systems meet those specifications. The auditor may use auditing methods designed to document conformance. For wind speed there are two tests. One is starting threshold, measuring the starting torque of the shaft and bearings with the cup wheel or propeller removed. Bearings will degrade over time. The time is a function of the exposure environment. There is a starting torque that has been shown to be equivalent of 0.5 m/s. The audit will document the starting torque expressed as a speed. If the result is 0.6 m/s and the requirement is 0.5 m/s, are all of the speed data rejected? If the last audit was 6 months ago, perhaps only the last 30 days of data will be rejected since the bearings will degrade with time. The application of common sense by the auditor, operator, and regulator will result in keeping all the data. The bearings would be changed, but the performance of the anemometer was probably acceptable. The wind in the atmosphere does not start at steady slow speeds as the wind tunnel test requires. The simulation of starting speed by starting torque has some uncertainty.

If the audit showed a starting speed of 2 m/s, the answer becomes more difficult. The operator should examine the speed data for the 6 months or year since the last audit where the starting speed was shown to be 0.4 m/s. Perhaps the record will show a period where something happened to the anemometer. If the site is windy, there may not be many periods with winds less than 2 m/s, in which case a higher starting speed would not degrade the data. Common sense and critical examination will suggest an answer to which the regulator and operator will agree. The last answer to accept is the auditor rejecting all the speed data because of the starting torque test results.

The other auditing method for anemometers is imposing a series of known rates of rotation to the anemometer shaft. This challenges the ability of the measurement system to sense the rate of rotation of the shaft and express this rate in terms of wind speed at the output of the system. When the sensor output is a frequency and the data logger is digital, the test will always pass. The only thing being challenged is the ability of the system to count pulses and apply an algorithm to express frequency as speed. There is nothing in this test that confirms the algorithm of wind speed to frequency. This takes a wind tunnel test or the acceptance of the manufacturer's claim that the generic transfer function for the product is correct.

When dealing with regulators, operators, and consultants, a common sense discussion will usually result in an acceptable solution. What is even more important is that it will result in the exchange of information that leaves all parties more experienced and better prepared for the next question.

REFERENCES

ASTM (1990). Standard Practice for Determining the Operational Comparability of Meteorological Measurements, ASTM, West Conshohocken, PA.

ISO (1993). Guide to the Expression of Uncertainty in Measurement (International Organization for Standardization, Geneva, Switzerland), ISO/TAG 4 N 70 Rev. January.

Lockhart, T. J. (1979). Climate without 19 mph, *Bull. Am. Meteor. Soc.*, **60**, 660–661.

Lockhart, T. J. (1996). Wind climate data continuity study—II.

National Institute of Standards and Technology (1994). Guidelines for Evaluating and Expressing the Uncertainity of NIST Measurement Results, NIST Technical Note 1297, Barry N. Taylor and Chris E. Kuyattt (Ed.), 12th International Conference on IIPS for Meteorology, Oceanography, and Hydrology, Atlanta, GA , Jan. 28–Feb. 2.

Schneider, S. (1987). *Discover* October, p. 47.

Wieringa, J. (1992). *J. Wind Engr. Ind. Aerodyn.* **41**, 357–368.

CHAPTER 35

MEASUREMENT IN THE ATMOSPHERE

JOHN HALLETT

The atmosphere is a turbulent medium. Any measurement is to be interpreted in the context of a set of measurements in a time and spatial dimension having a variance related to the scale considered. Such limits may be considered with respect to a fully developed turbulent field and limited by physical constraints—a solid boundary as at the surface, a variable boundary as with motions limited by the stability constraint as at an inversion top, or temporal as the diurnal or annual cycle. Any measurement is made with an instrument with a given time and spatial characteristic—whether it be a thermometer in sampling air flow or a satellite remotely measuring emissivity from a surface. The instrument design determines the lower limit of temporal and spatial measurement scales through the response time and the size of the instrument. Subsequent analysis determines how individual measurements are to be combined—as a mean and variance time series at a given location or a spatial field at a given time as a conventional synoptic map, with the lower scale necessarily limited through the initial instrument design. Such combinations of observations have been a central theme of understanding the different processes in the atmosphere and key in the initial development of meteorology (Fitz Roy, 1863; Brunt, 1917) and have been long realized (Middleton and Spilhaus, 1953); (*Handbook*, 1956).

We make measurements in the atmosphere for a variety of reasons. On a local scale, a measurement provides information for specific decisions, whether it be to wear a sweater, to begin harvesting a crop or to cancel an aircraft take off. On a broader scale, measurements are put together to provide information for a forecast for similar decisions at remote places. Such synthesis is required for initialization of a model of atmospheric processes whether it be to extrapolate frontal motion or to

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

solve numerically the equations of motion leading to such progression. Further synthesis may occur as such measurements are combined to give data to be used for climatology as a design for living, in terms of means and departures therefrom to estimate criteria for building design and for levels of investment in public utilities. Crudely, the local measurement requires a scale of meters and a time of minutes; the synoptic scale the dimensions of Earth and times of days; the climatological use requires a similar spatial scale but a time scale of a year to a decade or longer. Thus our instruments must be sufficiently small to be used locally, yet sufficiently robust and capable of calibration such that observations can be combined over extended scales of time and space.

Instrument response may be of first or second order depending on the nature of the design. Thus, a standard thermometer responds to an environmental change by approaching the new value almost exponentially and cannot overshoot. A wind vane, on the other hand, that approaches a new direction can be (and usually is) designed to overshoot so it can then approach the new direction more quickly than could be achieved if it were designed not to overshoot—by, for example, having a highly viscous damping system. Thus first-order instruments are to be designed having a time constant (lag) determined primarily by their size; a second-order instrument is designed to have a time constant determined by its size but also a period of oscillation determined separately by the damping characteristics, which also are influenced by size. The design must meanwhile ensure that any particular measurement responds solely to the quantity of interest; for example, rainfall should not influence the measurement of air temperature. The science and art of instrument design juggles these parameters to meet the needs of the way in which the data is to be used.

A further consideration lies in the spatial distribution of instruments and the frequency at which observations are made. While economic considerations may provide an upper boundary for the total number of instruments, the scale of the phenomenon of interest and its velocity of propagation should determine the optimum space distribution and frequency of measurement.

1 COMBINATION OF MEASUREMENTS

Many quantities of meteorological interest are to be derived by combination of independent measurements from separate instruments (Stankov, 1998) of the same region of interest or from regions which are to be expected on physical grounds to be highly correlated. Thus satellite measurements of ground emissivity at different wavelengths over the same area should be comparable, as should radiative properties of a given volume of particulates for identical viewing geometries. Under some circumstances instruments combine to give fluxes, as with a correlation between a temperature fluctuation (as measured by a thermocouple) and an air velocity fluctuation (as measured by a hot wire). Clearly the two instruments in the latter case are never quite collocated, and problems arise from fluctuations on the scale of the differences of location. Yet further complications arise should different instruments be of different order response (as with a wind vane and a thermocouple) and of different time lags. The question arises concerning the comparability of instruments located at a field site or on a moving platform as an aircraft. In the former case, distances of order meters may be involved, to be combined with differences in local airflow because of slightly different location geometry. In the case of aircraft, competition for space may lead to quite different locations for different instruments. Thus, air intakes for aerosol and gas measurements tend to be located at different sites on the fuselage, whereas particle measurement probes tend to be located on wing tip pylons; with different locations on the airplane, instruments are subject to different airflow geometries.

For measurement of properties of rare particles, in concentrations occurring, say, once per second in the sampling volume, it is clear that some hundred particles need to be sampled—or some hundred seconds of data need to be collected to obtain numbers meaningful, from a Poisson statistics viewpoint, at a 10% level. Thus a scale of 10 km is a practical limit of resolution. Relating this to meaningful measurements required to give insight into specific cloud processes—important, for example, for aircraft icing or the structure of cirrus—may require measurements on a scale well below 100 m (Liu et al., 2002). Hence the design of an instrument of adequate volume sample, of adequate time response and, for derived quantities, of appropriate collocation is no trivial undertaking. It is clear that instruments designed to measure *simultaneously* several properties of a given air particle sample will give more meaningful data to investigate the inhomogeneities shown by microwave studies (Mace et al., 1998).

2 PARTICLE MEASUREMENT

Particles in the atmosphere range from the smallest of aerosol some tens of nanometers in diameter in concentration in excess of $10^8\,\text{cm}^{-3}$ to large hailstones of 10 cm in diameter, in concentration less than one per $10\,\text{m}^3$. Such particles comprise precipitation, cloud, and the atmospheric aerosol itself. Measurement may characterize concentration, size distribution, and shape distribution. As an example, we examine ice crystal characteristics—spectra of habit, density, size, and concentration necessary for derivation of radiative fluxes in appropriate wavelengths and precipitation fluxes under differing dynamical conditions. A "point" measurement is idealized, but it is clear that a sample time, which differs for different size and other characteristics, needs to be specified. It is of interest to start with the pioneering work of cirrus crystal forms carried out from an open cockpit using a hand-held varnish-coated slide to give replicas for subsequent microscopic observation (Weickmann, 1947). This technique clearly showed the presence of three-dimensional bullet rosettes. Yet even in this work it is clear that an assessment of the *relative* occurrence of various crystal forms in the atmosphere—both as pristine and as complex crystal shapes—is a major undertaking because of the intrinsic variability of their nucleation and growth conditions. The early surface measurements (Bentley and Humphries, 1962; Nakaya, 1954) tended to select ideal forms as being of greater aesthetic value for sketching or photography, and later observations have similarly selected regions where crystal forms are relatively uniform. The occurrence of multiple habits and multipeaked size spectra at a point measurement is well documented in aircraft measurement and simplistically may be attributed initially to different nucleation and growth processes (Korolev et al., 2000; Bailey and Hallett, 2002), and subsequently to different fall speeds as well as the effects of mixing in lateral shear as in Kelvin–Helmholtz instability at an inversion top.

A more fundamental question needing to be addressed is the meaning of any data set of crystal shape, size, and habit distribution in a given measurement. The question of time and spatial scale of the sample is of major importance, and the detail of the averaging process is crucial as to how the data may be used. From a fundamental viewpoint, we may be interested in the nucleation and growth processes in a given volume of air that retains some coherence over growth times of interest—say some hundreds of seconds, with some hope of characterizing individual crystals over such a period. A Lagrangian observation strategy is therefore attractive, if not easily accomplished. From an applied viewpoint, crystals need to be characterized over, say, a volume of a lidar pulse some 1 m^3; the volume of a radar pulse some $10^6\,\mathrm{m}^3$, the volume of a satellite footprint some 100 km^3. To assess precipitation from a frontal system over its precipitation history, a volume of air some $10^8\,\mathrm{km}^3$ is more realistic; a precipitation of 1 cm over 1 km^2 requires some 10^{16} individual crystals. The realities of individual crystal measurement cannot compete, and the question of what is necessary for a meaningful sample arises. Surface collection and microscopy obviously gives a remarkably small sample and cannot provide a realistic sample for such a use. Electro-optical systems (PMS 2DC; PMS 2DP) give greater ease of data collection and are subject to some degree of automation, yet still provide a meager sample in relation to the above numbers. A similar consideration applies to more recent systems (Lawson et al., 1998). Some idea on variability in cirrus can be obtained on a broad scale from microwave radar (Mace et al., 1998), and it is clear that a cellular structure of order at least 100 m exists, as can readily be seen from a cursory visual inspection of any field of cirrus.

One may resort to a broader approach by assuming that in a sufficiently large volume of space, particle concentration, or indeed any other characteristic results from a combination of random events such that Shannon's maximum entropy principle applies. In this case, a Weibull distribution results (Liu and Hallett, 1998), implying that any spectrum measurement is but one of a family, and a sufficient data set may be specified to provide the "best" (most probable) distribution. It is necessary to specify a time or spatial boundary for such measurements; it may be convenient to do this on physical grounds. For example, limiting time by a well-determined effect (sea breeze/convection life time; Rossby wave transition time in Eulerian frame; a field of wave clouds, etc). Any individual measurement necessarily departs from this ideal. The reality of any particle measurement lies in the statistics of the numbers in each size bin. In general, there are fewer larger particles and an upper limit is set for realism by Poisson statistics for the large, rare particles. Recall that radar scattering relates to ΣNr^6, mass vertical flux to $\Sigma Nr^{5,4,3.5}$ depending on fall regime, mass to ΣNr^3, optical effects to ΣNr^2, particle diffusion growth rate to ΣNr, and nucleation processes to ΣN (Σ = sum over all particles). Uncertainties arise in derived quantities at different sizes depending where the size cut off in the measurements occurs for the selected sampling time and spatial average. A further consideration lies in direct measurement of properties of individual particles—such as impurity content and mean density, or density related to radius. The cloudscope class of instruments is a candidate for these measurements (Hallett et al., 1998).

From the viewpoint of an aircraft measurement, necessarily a long thin ribbon along its flight path, a longer path (space or time) to improve the statistics necessarily implies the likelihood of leaving the area (also defined in space or in time) where particles are occurring—meters or some tens of kilometers at most (arguably) for a cirrus regime or hundreds of kilometers for a field of convective storms (Fig. 1). Hence some formal definition of the *spatial and volume* scale and geometry of

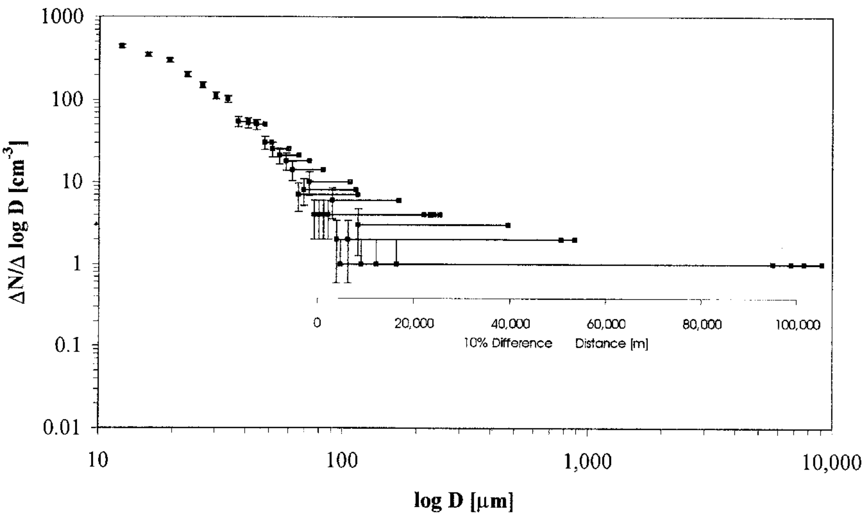


Figure 1 Analysis routine for an ice particle distribution collected by an airborne cloudscope (an instrument which images particles collected on a forward facing optical flat), sample volume about 5 liters per second. The protocol is set initially by the selection of the number of size bins on a logarithmic scale. It is further set by the level of uncertainty which can be tolerated—for example ±10%. The uncertainty in the number of particles actually counted in each bin (N) is given by Poisson statistics as $N^{1/2}$, represented by the **vertical** error bars on each point. The uncertainty is obviously large for the small concentrations of large particles. The **horizontal** line from each point represents a flight distance necessary to sample 100 ± 10 particles (10% uncertainty) at the concentrations observed using the instrument of designated sample volume. The physical domain over which any set of measurements must be analyzed is then selectable. This could be (as this case) 10 km of a hurricane outflow for a specific size, but would be something like 1000 km for marine stratus or the whole Northern Pacific Ocean for frontal systems for larger rarer particles. In order to reduce the sampling uncertainty, it would be necessary to use an instrument with a greater sampling volume or sample for a longer time/distance in regions of interest delineated by other considerations—for example energy dissipation rate or radiance at a given wavelength. (Data collected on the NASA DC-8 in outflow of hurricane Earl approaching the Louisiana coast, September 2, 1998.)

measurement becomes imperative for synthesizing data from any set of observations, whether it be area of cloud cover or a particle distribution having the moments discussed above. It may be desirable under some circumstances to average particles by updraft location—updrafts in the upshear of convective clouds have quite different microphysical structure from the downshear, and combine the two averages for remote-sensing comparison. Other combinations may be required to achieve a sufficient approach to reality.

From the viewpoint of idealizing ice particle formation and linking the physics of the individual particle nucleation and growth with ambient conditions, specific situations may be identified. A mountain wave lenticular cloud, formed under conditions of high stability has the merit of having low turbulence levels with traceable air trajectories from conservative variables. There is a well-defined rate of change of relative humidity ahead of the cloud and rate of availability of water vapor in the cloud itself, enabling specification of growth conditions. This situation also occurs for orographic clouds in upslope flow, which can lead to continuous production of uniform crystal type and size under quasi-steady conditions as in a lenticular cloud. Conditions change along such trajectories over times of order several minutes as can be judged from observation of the location of the cloud leading edge. Shallower clouds formed in gravity waves on frontal surfaces give somewhat longer growth times. Of current interest is the ability to produce relatively low ice supersaturations that persist over a long time, giving low nucleation rates of Cloud Condensation Nuclei (CCN), which freeze homogeneously as they dilute at temperatures somewhat below -40°C. Small droplets freeze as single crystals and grow slowly under ambient conditions in the form of crystals with uniform flat facets to give well-defined and bright optical effects. The stability is very important since it maintains conditions so that a conceptional model may be put together (Hallett et al., 1999). Spectacular optical effects are reported extensively in polar regions and are known to be associated with such pristine crystals—both plates and columns. Less well reported are spectacular midlatitude summer displays at low temperatures and high levels. A way of producing a low supersaturation is a rapid overrunning of a cold layer by a warmer moister layer followed by a slow diffusion of properties from one to the other. Such effects can be idealized by a time-dependent solution of the heat/vapor diffusion equation in one dimension, with initial boundary conditions being uniform temperature and mixing ratio with an initial sharp discontinuity at the interface (Carslaw and Jaeger, 1959). Heat and vapor diffuse almost together (vapor is a little faster) and since the saturation vapor pressure of ice is near exponential with respect to $1/T$ (temperature in Kelvin), a pulse of supersaturation spreads into the cold air with time constant:

$$\frac{X^2}{D}, \quad \frac{X^2}{\kappa}$$

where X is the distance from the initial interface, D is water vapor diffusivity in air, and κ the thermal diffusivity of air. For distances of tens of centimeters, the times are of order 100 s; for meters of order a few hours. Shear can influence the boundary conditions and change these times, but maintaining stability is important to the process. This behavior results from the assumption of infinite lateral extent; in reality the discontinuity may be linear or circular, giving a decreasing supersaturation as the disturbance propagates. To maintain a crystal in growth conditions would require a special combination of progressive gravity wave having such a diffusion supersaturation field but moving with the growing crystal. This is necessarily an infrequent occurrence (the sky is not universally covered with spectacular optical events) but may be sufficiently frequent to explain the occasional occurrence of spectacular optical displays under both high cirrus and boundary layer diamond dust conditions, required for the local production of uniform crystal sizes and shapes. The detailed structure of inversions is far from well known and can be extremely sharp—with differences of 10 K sometimes extending over a meter or less. Such sharp discontinuities would not appear in any standard sounding analysis. The spatial distribution of supersaturation under these conditions is an interesting topic for further study.

The trail of defining and specifying cirrus cloud follows the process of nucleation, growth, and evaporation of individual crystals. It requires a knowledge of the dynamic and radiation conditions under which they evolve and ultimately specification on *physical grounds* of a spatial and temporal averaging domain. This domain is related to the time and spatial resolution of the instruments both for in situ and remote sensing. For multiple instruments, measurements of the *same* volume and geometry of air by the different techniques is crucial to a combination of observations. All need to be considered in any numerical comparison, a difficult but not impossible undertaking.

3 INSTRUMENTATION AND MODELING

Major advances in the atmospheric sciences often come from the development and use of specific instrumentation. The barometer and measurement of pressure led to the realization that weather systems existed. The development of telegraphy and radio communication enabled the construction of weather maps with only a few hours delay. The radiosonde, developed in the 1940s, opened our understanding of the upper troposphere, the basis for description of the weather on a global scale, and forecasting as a practical application. With the advent of major computing power (dependent on the production of fast processing chips and even smaller but more powerful memory devices by the semiconductor industry), opportunity to access very large databases has emerged, together with the ability to perform rapid calculations in the use of complex numerical models of a variety of atmospheric phenomena.

Current models of a variety of processes in Earth's atmosphere are an outcome of this technology. Such models are necessarily limited in application by our ability (or lack thereof) to provide measurements of initial and subsequent conditions over a sufficiently small grid length, both of the thermodynamic variables themselves (e.g., temperature, dew point, mixing ratio, radiative flux) and of what may be called structure-sensitive variables (cloud droplet nuclei, CCN, ice nuclei, trace NH_3, particle defect structure). The implication in the use of such models is that justification of their prediction comes from a scheme that compares their output with some set of measurements. This is called *verification* or perhaps *validation* (Hallett, 1996).

This sounds very grand, but when we look at things more closely, we find that it may be nothing of the sort. What it *is* is a check for consistency of the model against a set of measurements. It should be realized that any agreement may be coincidental and requires justification as the progenitor of the model assesses the sensitivity to both the physical assumptions together with the variability and rational for any parameterization of the measurement. This is the case for models on all scales in the atmosphere and also for the weather forecast itself.

The atmosphere is not a controlled laboratory; it is a turbulent, highly variable environment where parameterizations, sometimes overgenerous in their believability (e.g., size distributions of aerosol or precipitation for the reasons discussed above) may be no more than a figment of the modeler's imagination. Hopefully, it represents a space and time average that smoothes local effects—this is the nature of statistics. But how much larger a scale—how long a time? The conclusion is that physical and chemical concepts (still models in the generic terminology but based on our knowledge of the real world) need to guide closely what is real and what is not real in the numerical model. This may lead to quite specific measurements—not to validate the model (impossible) but to give some justification for the physical/chemical ideas on which it is based. The concept of parameterization of any atmospheric process needs to be offset against the reality of an appropriate space–time scale, whether it be mixing at the top of a cumulus or transfer of some property (heat, momentum) in a boundary layer or gravity wave.

The availability of suitable instruments is a necessary prerequisite to progress, here defined as insight into useful ideas that can be used in a predictive sense. In a completely unrealizable utopia, measurement ultimately provides a one-to-one comparison between a model and a measured reality. From a theoretical basis, self-consistency and logic necessarily have their place, but the reality is so complex that the assumptions and approximations dominate. The ideas of a theme song of an early radar conference "measure everything, everywhere, all the time" provide as much of a hindrance to progress as does the model in the absence of a prescribed set of measurements for its possible justification.

This brings us to the development of instruments for a specific purpose. A wide variety of instruments can be bought "off the shelf" for investigation of atmospheric processes—providing a sufficient market has existed to encourage their development. Thus, instrumentation is readily available, maybe for attacking yesterday's problem. But, what of today's problems? What of tomorrow's problems? This is where we need to be perceptive in development of instrumentation to optimize the future development of our subject. We need to train students not only in instrument design, construction, and use but in critical evaluation of such measurements. From a practical consideration we have a challenge to develop instruments and techniques that give us an edge for the evaluation of specific models and to enable a judgment to be made on their utility. There is an important caveat: The models themselves must be designed to be amenable to such an approach through the predication of specific instrumentation.

Thus, there is an important aspect of this concept in instrumentation evolution. If a class of instrument is to be developed with a specific output in mind, it requires input from the user whose ultimate aim is to justify an approach to a problem through physicochemical insight; it requires input from the engineer who makes the final design; and it requires input from the design of the numerical model such that there are very tight criteria for arguing that the model output makes sense in the overall scheme of things. This concept applies to the choice of sensors and their combination (as for turbulent flux measurement), it applies to the choice of wavelengths for assessing radiative flux budget both in solar and thermal infrared, and it applies to characterizing shapes and sizes of nonspherical particles for interaction with radiation fields of different wavelengths. The measurements made are to be justified in terms of the whole approach, as also is the model that is being tested for consistency.

Thus, success in new sensor technology lies in elegance of design for specific purposes by combining outputs to give meaningful quantities directly and measured *simultaneously* in the same volume of air. This is critical and predicates an integrated approach to instrumentation and model design at an early stage of any project in investigation of specific atmospheric problems. An observed consistency between model and measurement can then be used to give a better rationale both for the combination of observations and in the predictive use of the model output.

4 COMPLEXITY AND DIVERSITY

Diversity of biological species may result from the spread of ecological niches associated from a spread of microclimates. Measurements at widely separated locations miss such phenomena; remote sensing may miss such phenomena because of lack of resolution. The complexity of natural phenomena as they occur in the atmosphere (Rind, 1999) may remain hidden if it is not realized that extremes of environment may be quite unrealistically represented; similar considerations may apply elsewhere in our society (Arthur, 1999). Tails of distributions of such phenomena may be quite unsuitable for extrapolation, assuming standard distributions, as has been argued for years by the statistical community. This is certainly true for many physical processes—extremes of temperature may lead to species elimination both hot and cold. A few large drops lead to rainfall by coalescence and may or may not be present depending on the presence or absence of distinctive physical processes. Measurement of such phenomena must therefore proceed from a prior knowledge of such mechanisms, necessary to suggest the nature of the instruments to be built and used to verify (in its true sense) the physical processes responsible.

REFERENCES

Arthur, W. B. (1999). Complexity and the economy, *Science*, **284**, 107–109.

Bailey, M. and J. Hallett (2002). Nucleation effects on the habit of vapour grown ice crystals from −18 to −42°C, *Q. J. R. Meteorol. Soc.* **128**, 1461–1483.

Bentley, W. A., and W. J. Humphreys (1962). *Snow Crystals*, New York Dover (reprint of 1932 edition).

Brunt, D. (1917). *The Combination of Observations*, Cambridge, England, Cambridge University Press.

Carslaw, H. S., and J. C. Jaeger (1959). *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford, UK.

Fitz Roy (Rear Admiral) (1863). *The Weather Book: A Manual of Practical Meteorology*, London, Longman, Green, Longman, Roberts, & Green.

Hallett, J. (1996). Instrumentation and modeling in atmospheric science, *Bull. Am. Meteorol. Soc.* **77**, 567–568.

Hallett, J., W. P. Arnott, R. Purcell, and C. Schmidt (1998). *A Technique for Characterizing Aerosol and Cloud Particles by Real Time Processing*, PM2.5: A Fine Particle Standard Proceedings of an International Specialty Conference, Sponsored by EPA Air & Waste Management Association, J. Chow, and P. Koutrakis, Eds., Vol. 1, pp. 318–325.

Hallett, J., M. P. Bailey, W. P. Arnott, and J. T. Hallett (2002). *Ice Crystals in Cirrus*, Ch. 3, pp. 41–77, New York, Cirrus, Oxford University Press.

Handbook (1956). *Handbook of Meteorological Instruments, Part 1, Instruments for Surface Observations*, Meteorological Office, Air Ministry, London, Her Majesty's Stationery Office, M.O. 577.

Korolev, A., G. A. Isaac, and J. Hallett (2000). Ice particle habits in stratiform clouds, *Q. J. R. Meteorol. Soc.* **126**, 2873–2902.

Lawson, P. R., A. V. Korolev, S. G. Cober, T. Huang, J. W. Strapp, and G. A. Isaac (1998). Improved measurements of the droplet size distribution of a freezing drizzle event, *Atmos. Res.* **47**, 181–191.

Liu, Y., and J. Hallett (1998). On size distributions of cloud droplets growing by condensation: A new conceptual model, *J. Atmos. Sci.* **55**, 527–536.

Liu, Y., P. H. Daum, and J. Hallett (2002). A generalized systems theory for the effect of varying fluctuations on cloud droplet size distributions, *Am. Meteorol. Soc.* **??**, 2279–.

Mace, G. G., K. Sassen, S. Kinne, and T. P. Ackerman (1998). An examination of cirrus cloud characteristics using data from millimeter wave radar and lidar: The 24 April SUCCESS case study, *G. Res. Lett.* **25**, 1133–1136.

Middleton, K. W. E., and A. F. Spilhaus (1953). *Meteorological Instruments*, 3rd ed., Toronto, University of Toronto Press.

Nakaya, U. (1954). *Snow Crystals, Natural and Artificial*, Cambridge, Harvard University Press.

Rind, D. (1999). Complexity and climate, *Science* **284**, 105–107.

Stankov, B. B. (1998). Multisensor retrieval of atmospheric properties, *Bull. Am. Meteorol. Soc.* **79**, 1835–1839.

Weickmann, H. K. (1947). Die Eisphase in der Atmosphare, Reports and translations, Ministry of Supply (A) Vbkenrode (H. M. Stationery Office, London, 1947, Volkrende R&T).

CHAPTER 36

INSTRUMENT DEVELOPMENT IN THE NATIONAL WEATHER SERVICE

JOSEPH W. SCHIESL

1 INTRODUCTION

Meteorological instrumentation is a very challenging arena. The first difficulty in the design of these instruments is finding a common set of requirements upon which the various users can agree. For example, hydrologists are generally happy with a precipitation resolution of 2.5 mm (0.1 in.) while meteorological modelers prefer a resolution of 0.25 mm (0.01 in.) for forecast verification.

Instruments must survive the very elements that they are measuring. There are the temperature and humidity extremes, high winds, heavy rains, ice accretion, sand and dust, solar insolation, ultraviolet radiation, salt spray, and altitude variations. Also, there are the nonmeteorological elements like corrosion, fungus, insects, birds, radio frequency interference, power line surges, pollen, and jet fuel. Many laboratory measurement technologies do not survive the passage to the outdoors with its diverse conditions.

There are many hundreds of instruments that have been used by the National Weather Service (NWS) and its predecessor organizations since 1870. Some were short lived while others that existed before the NWS are still in use. The instruments that emerged with the more interesting histories are those that measure the meteorological parameters in which the public is most interested, i.e., temperature, precipitation, and wind. The public wants to know if they have to dress for hot or cold, do they need an umbrella or boots, or will the wind blow their hats off or tip the trashcans?

Not all the instruments will be addressed, but those discussed will capture the flavor of the challenge of the times. There are many books available for those

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

wishing to know a complete history or the detailed theory of operation of meteorological instruments.

2 TEMPERATURE AND HUMIDITY

Air temperature readings have and still are made by liquid-in-glass thermometers. These simple devices are based on the principle that materials expand or contract with temperature changes. The liquids are either mercury or alcohol, depending on the range of temperatures to be measured. Mercury can only be used to $-38.9°C$ where it reaches its freezing point.

Regardless of the sensor used, the temperature of the ambient air cannot be measured properly if the sensor is not adequately exposed. The sensor must be protected from direct and indirect solar radiation while still being exposed to a free flow of ambient air. To provide this protection, shelters or screens are constructed of wood and painted white with louvered sides for ventilation. The roof is double layered to provide air insulation. The door to the shelter faces north, in the Northern Hemisphere, to minimize the possibility of direct radiation when the door is opened. See Figures 1 and 2. Normally, temperatures measured in these shelters are representative of the ambient air. However, when the solar radiation is high and the winds are below 5 knots, the shelter temperature has been shown to be up to 3°C (5.4°F) higher than the ambient temperature. To overcome most of this bias, some shelters are aspirated by a fan rather than depending on natural aspiration.

Liquid-in-glass thermometers do not provide continuous temperature records, so the possibility of capturing the maximum and minimum temperature is very remote. A maximum thermometer is made by placing a constriction in the base of the bore

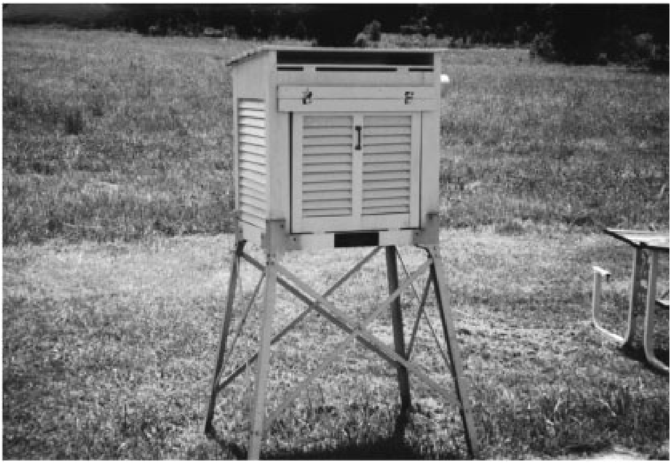


Figure 1 Cotton Region temperature shelter. See ftp site for color image.

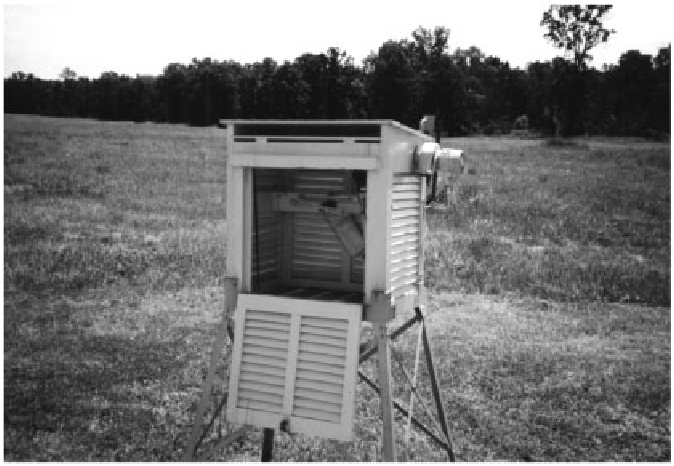


Figure 2 Cotton Region temperature shelter with door open. See ftp site for color image.

just above the bulb. A temperature increase causes the liquid to rise, but the force of gravity is insufficient to allow the liquid to subside. After the maximum temperature is read, the mercury is shaken down below the constriction. See Figure 3.

In the minimum thermometer, a glass barbell-shaped object is placed in the alcohol column. The thermometer is placed in a near horizontal position with the bulb side down. When the alcohol contracts with decreasing temperatures, the meniscus pushes the barbell down. The top of the barbell stays at the minimum temperature position when the alcohol expands. After the minimum temperature is read, the barbell is tipped back to the meniscus. The minimum thermometer can also

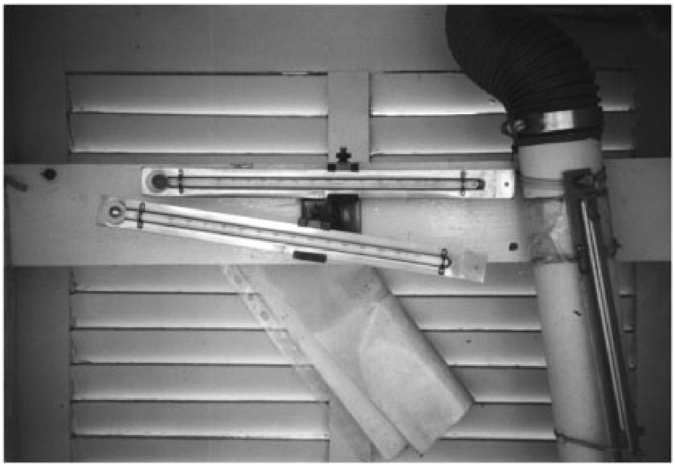


Figure 3 Maximum (bottom) and minimum (top) thermometers in Townsend support. See ftp site for color image.

be used to obtain the current temperature without disturbing its primary function. See Figure 3.

Other methods were used to record temperatures. One consisted of two bonded strips of metal with differing coefficients of expansion. Temperature changes caused the bonded strip to deform because of the different expansion coefficients. This motion was transformed by levers to make an ink trace on a chart called a thermograph.

In the nineteenth century, scientists were aware that the resistance of a material was a function of its temperature. This relationship was used in reverse to determine temperature. Electrical thermometers in which the resistance of the element increases with the temperature are known as resistance temperature devices. Those in which the resistance of the element decreases with increasing temperature are called thermistors. The small size of the electrical thermometer elements, usually less than an inch in length, allows them to have quick response times.

Measuring the amount of water vapor in the air accurately, over a wide range of temperatures outdoors, is one of the most challenging of meteorological parameters. Some terms used to express this measurement are wet-bulb temperature, dew-point temperature, vapor pressure, specific humidity, and relative humidity. The instruments used to measure water-vapor content are called hygrometers.

Many materials with which the public is familiar are coarse versions of hygrometers, although the specific values of moisture content are not apparent with them. Many materials expand with increasing humidity, and this is exhibited by sticking doors and windows. Certain fibers expand significantly with humidity, which caused problems in the non-air-conditioned buildings of the garment industry at the turn of the century. As the sewing threads expanded in high humidity, a point was reached where the thread could not pass through the sewing machine needle eye easily, and the thread snagged and broke. Humidity readings were taken to determine the proper needle eye size for that condition. This expansion property was used in one of the first hygrometers to measure humidity. The material used was human hair. As strands of hair changed in length with humidity changes, their movement was exaggerated with levers causing an inked pen arm to move. This record of humidity on a clock-driven drum was called a hygrograph. The problem with this instrument was that each hygrograph had to be calibrated individually and often, depending on the type of hair. Corrections also had to be made to separate the expansion caused by temperature versus that caused by the humidity.

One of the more common manual instruments used to measure the amount of water vapor in the air is the sling psychrometer. It consists of two liquid-in-glass thermometers mounted on a metal strip. One of the thermometer bulbs is wrapped in muslin. The muslin is dipped in distilled water and both thermometers on the strip are whirled, which evaporates water from the wetted bulb. The temperature decreases to a reading that will remain constant, even with further evaporation. This temperature is called the wet-bulb temperature. The amount of heat given up by the bulb is a function of the water vapor in the ambient air. If the dry-bulb and the wet-bulb temperature are the same after whirling, the air is saturated and the humidity is 100%. The difference in temperatures is called the wet-bulb depression. With psychrometric tables, the dry-bulb and wet-bulb temperatures can be used to determine the dew-point temperature, the humidity, and the amount of water vapor in the air per unit volume. Variations of this instrument were used, where the aspiration was provided by a motor-driven fan rather than by hand slinging.

The NWS, in addition to using the sling psychrometer to determine temperature and wet bulb, uses another instrument to measure both the temperature and the dew point. It is called the hygrothermometer and is in use at hundreds of airports across the United States. See Figure 4. The data are used in support of aircraft operations at airports and by meteorologists for forecasting. These systems were first deployed in the 1960s.

The dry-bulb temperature sensor is a resistance temperature device. The dew point is determined by a resistance temperature device imbedded in a mirror. A light-emitting diode provides a light source, and the dew point is determined by the ratio of the light received by two phototransistors. One measures the direct reflectivity from the source light, while the other measures the indirect. The heating and cooling of the mirror is effected by a Peltier device. See Figure 5. As the mirror is cooled, the formation of dew/frost on the mirror will cause an increase in the indirect sensor level. Cooling continues until the indirect sensor level is about 80% of the direct sensor level and the servo-loop equilibrium is achieved. This is the dew-point temperature. The device electronically compensates for a loss of mirror reflectivity as contaminates accumulate on the mirror surface. Another feature of this system is a circuit used to detect the reduction of flow through the instrument. This reduction can be caused by blockage (leaves and insects) or complete or partial failure of the fan. Specific mirror-cleaning procedures are critical to its proper operation. These procedures involve the use of isopropyl alcohol, lacquer thinner, and distilled water. The criticality of following specified procedures was emphasized, when a nonprescribed brand of cotton-tipped swabs was used to apply these

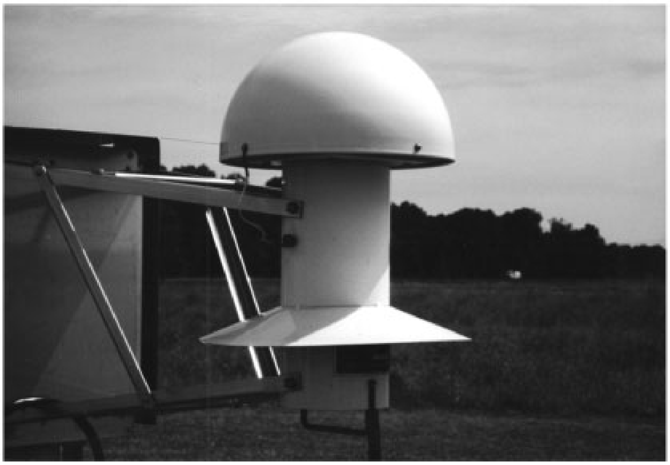


Figure 4 Hygrothermometer enclosure. See ftp site for color image.

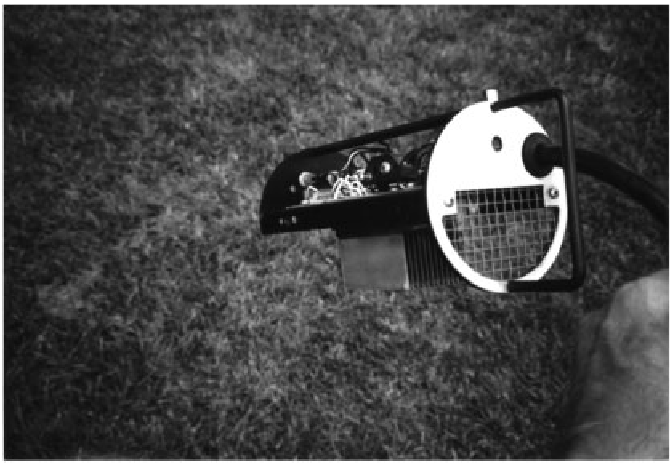


Figure 5 Hygrothermometer interior showing intake air screen. Finned object on bottom is Peltier device. See ftp site for color image.

chemicals. The adhesive used in the substitute brand was dissolved by the lacquer thinner, which in turn contaminated the mirror.

The hygrothermometer sensors are housed in an enclosure to protect them from elements like precipitation, solar radiation, dirt, and insects. Any of these could cause false readings. The shelter also includes a fan that ensures a free flow of air over the sensor. Even with these precautions, the shelter is subject to small variations in temperature caused by solar radiation that varies from day to day and from place to place. Varying wind speeds also affect the amount of airflow through the shelter, which can cause temperature changes. In addition, the electronic components can cause a small calibration drift. Because of these factors affecting the temperature readings, the hygrothermometers are checked weekly using liquid-in-glass thermometers.

Even though these instruments at airports are not meant to meet climatological requirements (there are separate networks for those), these data are used as an index of change for the local area. In addition, airports are notoriously bad for climatological measurements because of the heating effects of runways, roads, buildings, and moving aircraft.

3 TEMPERATURE SITING STANDARDS

Temperature sensors should be located over terrain (earth or sod) that is typical of the area around the station. Unfortunately, some thermometers are located on rooftops for security purposes. Rooftops are not desirable locations for measuring air temperature. The siting problem is the main source of the bank thermometer phenomenon, whereby surface air temperatures are frequently reported too warm.

Because the sensing element of a thermometer absorbs more radiation when exposed to the sun, than the air itself, it must be shielded. The shielding also protects the sensing element from precipitation, dirt, and insects, any of which could cause a false reading.

The preferred temperature sensor height is about 2 m (5 ft) above the ground (eye level). This height needs to be adjusted in areas that accumulate deep snow cover. Sensors located too close to the ground will read too low for minimum temperatures and, conversely, maximum temperatures will read too high during the day. Sensors mounted on towers may be unrepresentative because temperatures can vary greatly in the lowest levels of the atmosphere near the ground. They can be especially unrepresentative during clear, calm mornings with inversions, where temperatures can vary 8°C (14.4°F) or more in the first 60 m (200 ft) above the ground.

Temperature sensors should be installed in a position to ensure that measurements are representative of the free air circulating in the locality and not influenced by artificial conditions such as large buildings, cooling towers, or expanses of concrete and tarmac.

4 PRECIPITATION

Many centuries ago, the people of the Middle East measured precipitation with buckets and measuring sticks. The data were used to levy taxes since precipitation was associated with agricultural output. Although the number of measuring techniques has increased, the simple bucket and measuring stick prevails in numbers over any other methodology used throughout the world.

The NWS uses the 20.3-cm (8-in.) (size of the orifice) precipitation gauge for manual observations. See Figure 6. The precipitation falling through the orifice is

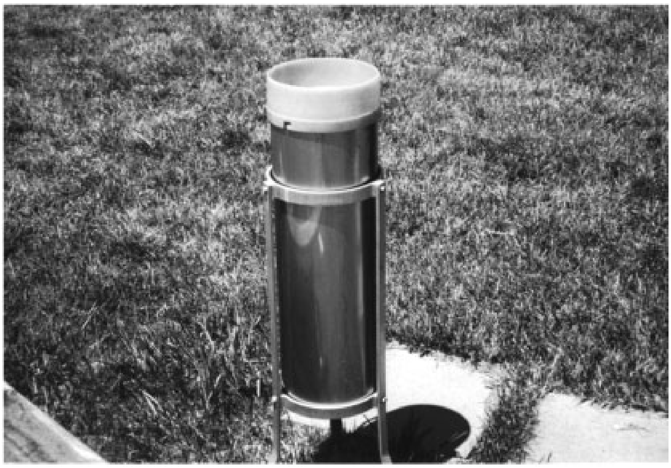


Figure 6 Manual precipitation gauge, 20.3-cm diameter. See ftp site for color image.

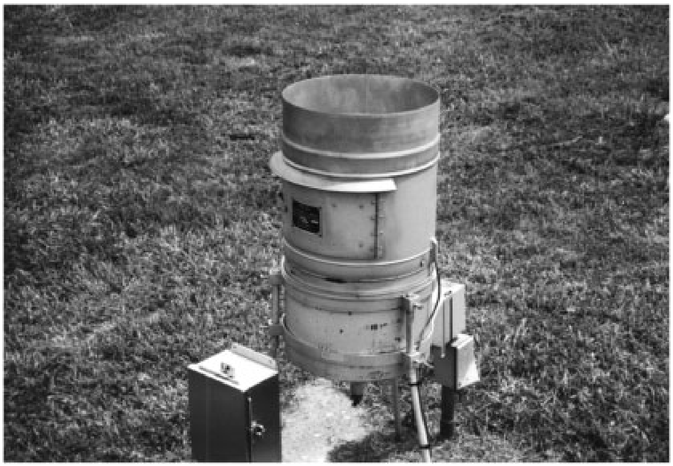


Figure 7 Tipping bucket rain gauge. See ftp site for color image.

funneled into an inner tube whose cross-sectional area is one tenth that of the outer can. This ratio provides the magnification of the measuring resolution by a factor of 10. If the precipitation exceeds the total capacity of the measuring tube (2 in.), it spills over into the overflow can and must be poured into the receiving tube to be measured. For solid precipitation, the collector funnel and receiving tube are removed and the precipitation is caught in the overflow can. The catch is then melted to determine the water equivalent of the precipitation. This gauge also serves as the standard to which other NWS gauges are compared.

Unfortunately, such simplicity could not serve all of the NWS's requirements. In the latter part of the nineteenth century, the tipping bucket gauge was developed to record rainfall rates and amounts. See Figure 7. Rainfall was funneled from a 12-in.-diameter orifice through a small spout to a tipping bucket mechanism. This mechanism consisted of two buckets. The particular bucket under the spout filled to capacity, lost its balance, and tipped. As it tipped, it activated a mechanical counter to record the total rainfall. See Figure 8. In later years, the tip caused the closure of a mercury switch, which sent an electronic impulse to an electronic counter. With the concern over the potential hazard of mercury, the mercury switch is being replaced with a reed switch. The number of tips recorded over a period of time determined the rainfall rate. The rainfall amount was recorded in a collector tube beneath the gauge. The need to separate the means of measurement of rate and total amount is caused by splashing at the tipping mechanism during higher rainfall rates.

Another limitation of the tipping bucket gauge is that it does not work accurately with solid precipitation. A heated version was tried, but the heat caused further inaccuracies by causing evaporation and, in some cases, a thermal plume that prevented all the snowfall from falling into the gauge.

To catch all types of precipitation, the first of a number of weighing gauges was developed. See Figures 9 and 10. This "universal" gauge caught precipitation falling

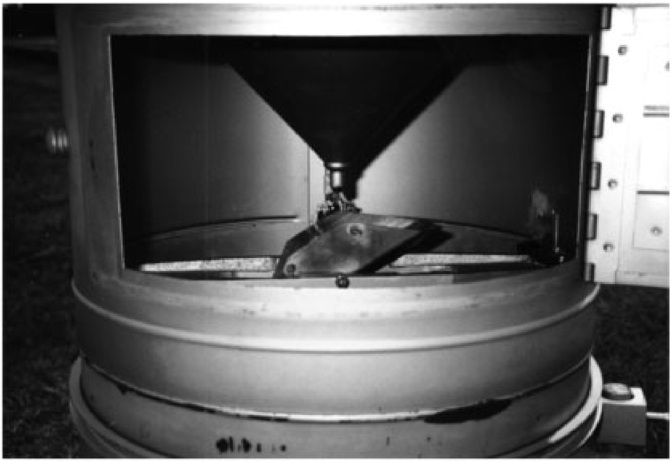


Figure 8 Tipping bucket mechanism. See ftp site for color image.

through a 20.3-cm (8-in.) orifice and was directed into a galvanized bucket on the platform of a spring-scale weighing mechanism. The weight of the precipitation depressed a scale and through a mechanical linkage deflected a pen arm across a clock-driven rotating-paper chart. The most commonly used version of this gauge has a capacity of 12 in. of liquid precipitation and a 24-h recording period. For increased resolution, not accuracy, the recording is expanded across a dual traverse of the pen arm. During the winter season when snow, ice, and freezing temperatures are likely, the gauge is winterized with an antifreezing solution to prevent damage to the measuring bucket. Other precipitation gauges have been developed that measure

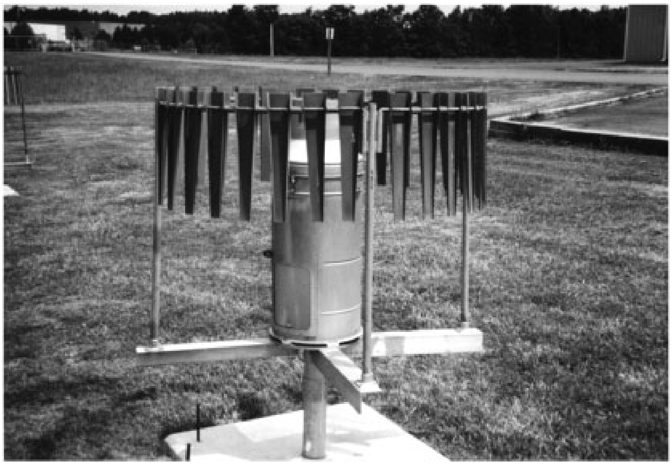


Figure 9 Universal precipitation gauge with wind shield. See ftp site for color image.

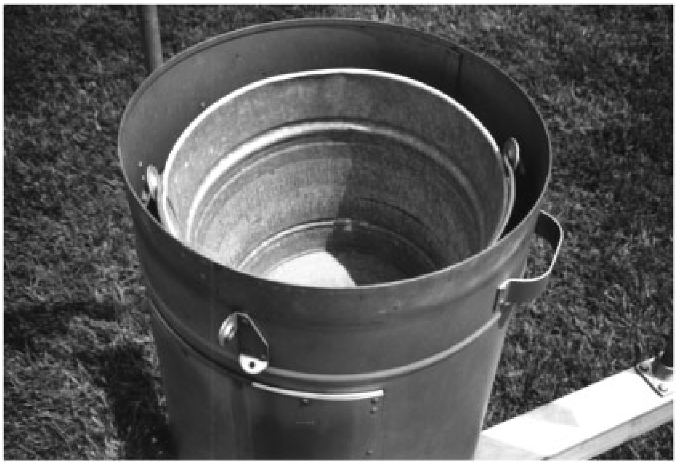


Figure 10 Weighing bucket in universal gauge. See ftp site for color image.

the amount of precipitation collected in a container. The amount is determined by technologies such as strain gauges, shaft encoders, and load cells.

5 PRECIPITATION STANDARDS

The gauge should be mounted so the orifice is in a horizontal plane. The height of the orifice should be as close to the ground as practicable. In determining the height of the orifice, consideration must be given to keeping the orifice above accumulated/ drifting snow and minimizing the potential for splashing into the orifice. The immediate surrounding ground can be covered with short grass or be of gravel composition, but a hard flat surface, such as concrete, gives rise to splashing and should be avoided.

The catch of a precipitation gauge should represent the precipitation falling at that point. This is not always attainable because of the effect of wind, and thus care should be exercised in the selection of a precipitation gauge site to minimize the wind effects. Towers and rooftops are poor locations for precipitation gauges.

Precipitation gauges should be located on level ground at a distance from any object of a minimum of two, and preferably four, times the height of the object above the top of the gauge. An object is considered an obstruction if the included lateral angle from the sensor to the ends of the object is 10° or more. Beyond this range, objects that individually, or in groups, reduce the prevailing wind speed, the turbulence, and eddy currents in the vicinity of the gauge may provide a more accurate catch. Thus, the best exposures are often found in orchards, openings in groves of trees, bushes and shrubbery, or where fences and other objects together form effective windbreaks. Rain gauges should never be installed on roofs or towers because of increasing winds with height and the presence of eddy currents.

In order to reduce losses caused by wind, a windshield is recommended to be installed on gauges where 20% or more of the annual average precipitation (water equivalent) falls as snow.

6 WIND

From earliest times, attempts have been made to measure the effect of the speed of the wind. The wind force was referenced to the effect it had on objects. Without realizing it, people of old used a form of the Beaufort wind scale without the Beaufort number or the speed of the wind. The wind direction was something more obvious even without instruments. References to wind direction can be found in the early books of the Bible.

Somewhere in time, an early scientist must have tried to associate the revolutions of objects like a windmill, in a given amount of time, as a measure of wind speed. In the seventeenth century, seamen estimated the wind by the angular movement, from the vertical, of a flat plate that rotated on a horizontal bar. This was followed by the mounting of objects on rotating wheels to catch the wind. Some of the wind catchers were made of sail cloth, and different shaped wooden and metal disks.

By the time the federal government got into the weather business, the rotating cup anemometer was the most prominent method to measure wind speed. For many decades, meteorologists experimented with various anemometers. It was then known that the rotational speed of the cups was a function of wind speed and the density, viscosity, and turbulence of the air. What was also important was the diameter of the cups, cup shape, the number of cups, arm length, the moment of inertia of the system, and the effect of precipitation. By the late 1920s the four-cup, 10-cm (4-in.) diameter, hemispherical-shaped cups were replaced by the three-cup, 12.5-cm (5-in.) diameter, hemispherical-shaped cups. A few years later, the hemispherical-shaped cups were replaced by semiconical cups. See Figure 11. The semiconical cups did not overestimate gusty winds as much as the hemispherical. Throughout these transitions, the NWS made correction tables available so the speeds of the different systems could be compared. Other than for the methods of recording wind speed and direction, this system has remained essentially the same.

Early wind recorders were composed of worm and toothed gears that indicated the passage of a mile of wind. Subsequent to this, electrical contacts were used to measure miles of wind. These miles of wind were indicated by buzzers, blinking lights, or marks on a chart. Wind speed could then be determined by the amount of time it took between consecutive contacts. To improve the resolution of wind speed measurement, contacts were eventually made for every 1/60th of a mile.

This system was eventually replaced by a direct reading anemometer. See Figure 12. It contained a magneto or small electric generator using a permanent magnet. A spindle is connected to the armature of the magneto. The revolutions per minute of the spindle, caused by the rotating cups, determine the amount of electrical current generated by the magneto. See Figure 13.

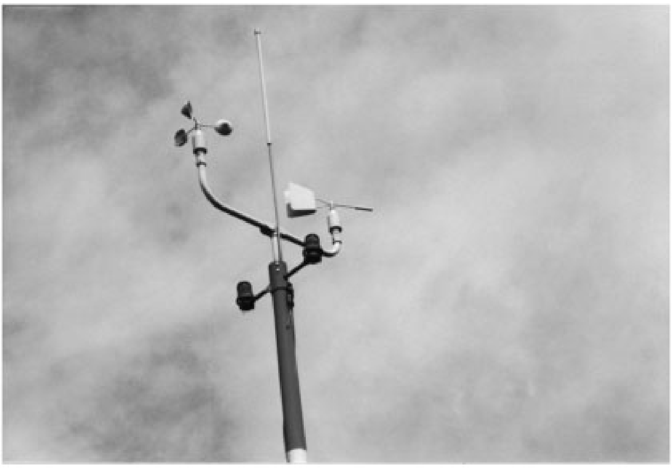


Figure 11 Rotating cup anemometer and wind direction sensor atop 10-m tower. See ftp site for color image.

In the 1980s, the internal electronics were changed. Light from a light-emitting diode was pulsed by slits in a disk that rotated around the spindle in the anemometer. This "light chopper" replaced the magneto and provided improvements. One such improvement was reducing the starting torque with the removal of the magneto. Another improvement was the elimination of questions regarding time constants with the generator system. Time constants with the former system had to be estimated because the dial indicator and gust recorder have time constants determined by inertia. The output of the light chopper is simply the number of light counts per second.

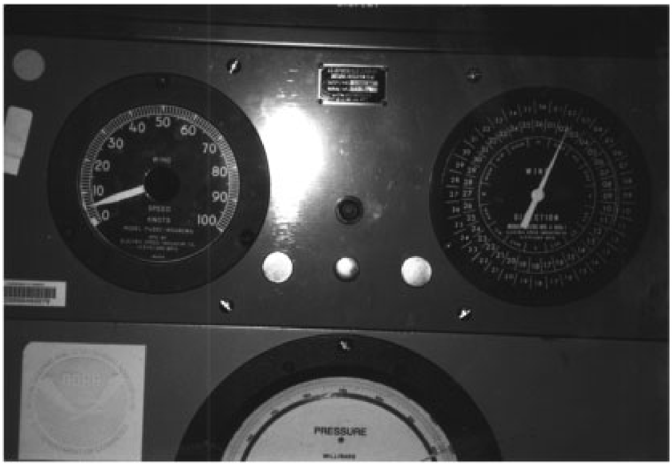


Figure 12 Wind speed and wind direction indicators. See ftp site for color image.

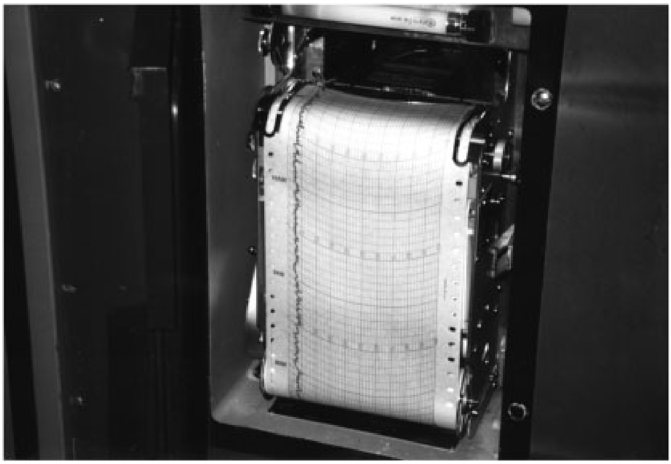


Figure 13 Wind gust recorder. See ftp site for color image.

Wind sensors should be oriented to true north. The site should be relatively level, but small gradual slopes are acceptable. The sensors should be mounted on a freely exposed tower, 10 m (30 to 33 ft) above the average ground height within a radius of 150 m (500 ft).

7 PRESSURE

In the seventeenth century, Torricelli balanced the pressure exerted by the atmosphere against the weight of a column of mercury with a vacuum above the mercury column. The height of this column was approximately 760 mm (30 in.). This is basically the same technology still being used today. These terms of height are still being substituted for pressure units in certain applications, such as aviation. Synoptic pressure analyses use the hectopascal, which is a recent standard change from the millibar.

The NWS has been phasing out the use of the Fortin tube to determine atmospheric pressure. In this device, a level of mercury in a cistern is raised by turning a thumbscrew to a zero scale. The zero scale is an ivory tip. Just before the top of the mercury in the cistern touches the ivory point, the image of the tip can be seen in the surface of the mercury. When the tip and its image just touch, or the mercury is "dimpled," the proper level has been reached. The height of the mercury is then measured by a vernier scale. Corrections are made for gravity and the expansion of glass and mercury with temperature.

With the advent of aviation, pressure readings were required more often. The reading of the mercury column was not something that could be done quickly, so the aneroid barometer was used. With the aneroid, the pressure of the atmosphere is balanced against the force of springs in an evacuated chamber made of metal. Later

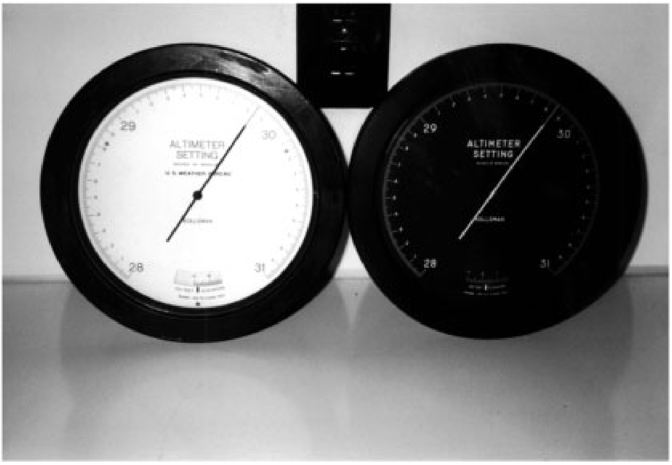


Figure 14 Altimeter setting indicators. See ftp site for color image.

on, the construction of the metal bellows chamber itself provided the balancing force. The motion of these chambers were exaggerated by levers to a dial. See Figure 14.

Permanent graphic records of the pressure values were provided by using a pen arm that made a trace on a clock-driven drum. This recording device was called a barograph. Aneroids had to be calibrated against the mercury barometer periodically and differences were posted on the aneroid as a correction factor.

With the concern about the dangers of exposure to mercury from a broken mercurial barometer, these barometers are being replaced by pressure transducers. In these instruments, the atmospheric pressure is still balanced against a bellows, but the pressure is measured by the force on a resonator. The frequency of oscillation of a crystal quartz resonator varies with the atmospheric pressure. Corrections are made for pressure changes caused by temperature. The NWS has found that these instruments are very stable and are being used to replace the former aneroids and the mercury reference barometers.

Pressure-sensor-derived values are of critical importance to aviation safety and operations. This is one of the parameters that cannot be sensed by humans other than by the rapid rise or fall of pressure. This is usually experienced by “ear popping” or sinus pain.

Care should be taken to ensure that pressure sensor siting is suitable and accurate. The elevations of the sensors shall be determined to the nearest foot. Pressure sensors are usually located indoors. Sensors should not be exposed to direct sunlight or in drafts from heating and cooling. With pressure tight buildings, each pressure sensor should be individually vented to the outside to avoid pressure variations due to “pumping.” These pumping or pressure variations occur with the cycling of heating and air-conditioning systems.

8 CLOUD HEIGHT EQUIPMENT

The height of clouds is most important for aviation. Earliest measurement methods were by the use of ceiling balloons. Balloons of various weights, typically 10 g with a nozzle lift of 45 g or a 30-g balloon with a nozzle lift of 139 g, were inflated with helium. A watch was used to time the interval between the release of the balloon and its entry into the base of the clouds. The point of entry for layers aloft was considered as midway between the time the balloon began to fade until the time the balloon completely disappeared. With surface-based clouds, the time interval ended when the balloon completely disappeared. During the day, red balloons were used with thin clouds and black balloons were used with thicker clouds. At night, a battery-powered light was attached.

In the 1930s, the beam from a ceiling light was projected at a 45° elevation angle into the sky. The projector was rotated about the vertical axis until the light beam hit the lowest cloud. The observer then paced off the distance from the projector to where he was directly under the cloud hit. With the geometry of this scheme, the paced distance equaled the height of the cloud. It soon became obvious that it would be less time consuming to project the light vertically into the air. See Figure 15. At a specified baseline away, a clinometer was used to measure the angle between the ground and the spot of light on the cloud. Knowing the baseline length and elevation angle in this right-triangle situation made it easy to determine the height with a lookup table. Clinometers were used to determine elevation angles. This instrument had wire cross hairs at one end and a narrowed neck at the siting end. This shape gave it the name "beer bottle."

The human detector end of this scheme was later automated by the use of a photocell that scanned the vertical path until the spot of light on the cloud was observed. The projector light was modulated so the photocell would only pick up

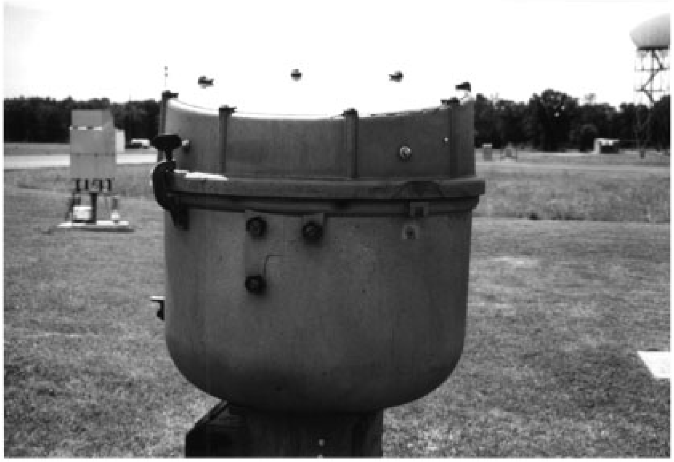


Figure 15 Fixed-beam ceilometer. See ftp site for color image.

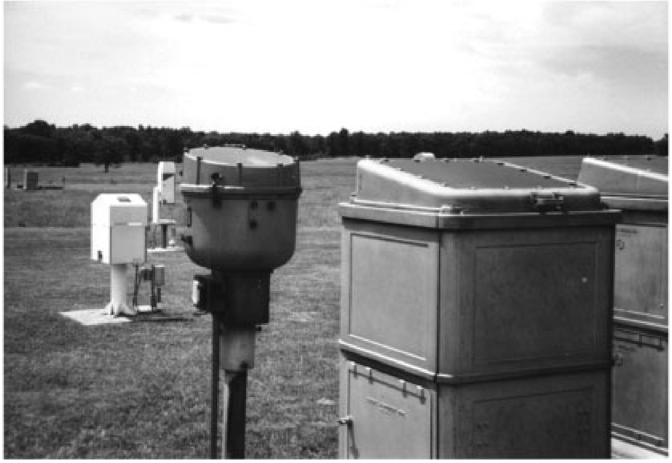


Figure 16 Fixed-beam ceilometer (cylindrical shaped) and rotating-beam ceilometer detector on right. See ftp site for color image.

this type of light during the daytime. The angle of the cloud hits were displayed automatically at the observers console.

The next version of cloud height indicators was the rotating beam ceilometer. As the name implies, the beam of light rotated, and the vertically looking detector measured any cloud hits directly overhead. See Figure 16. The angle of the cloud hits was displayed either on a scope or on a recorder chart. Height measurements were limited to heights no greater than 10 times the baseline. Above this ratio, the value of the tangent function increased too quickly to ensure the accuracy of a measurement.

The latest cloud height indicator is one that was put into use in 1985. This sensor sends laser pulses vertically into the atmosphere. The pulse rate varies with the temperature to maintain a constant power output. The time interval between the pulse transmission and the reflected reception determines the height of the clouds. The reporting limit of this instrument had been 3800 m (12,000 ft). Indicators are now available with a range of 7600 m (25,000 ft). See Figures 17 and 18.

9 VISIBILITY

In the 1940s, attempts were made to automate visibility observations for aircraft carriers. Because of motion problems, these attempts were not fully successful. Work continued in this area and, in 1952 a system became operational at the Newark, New Jersey, airport. This airport was located in the meadowlands adjacent to Newark Bay and experienced dense fogs. An interesting aside was the effect of the fog on the adjoining New Jersey Turnpike. Large propellers were installed at the top of poles by local authorities along the turnpike to disperse the fog, but the attempts were unsuccessful.

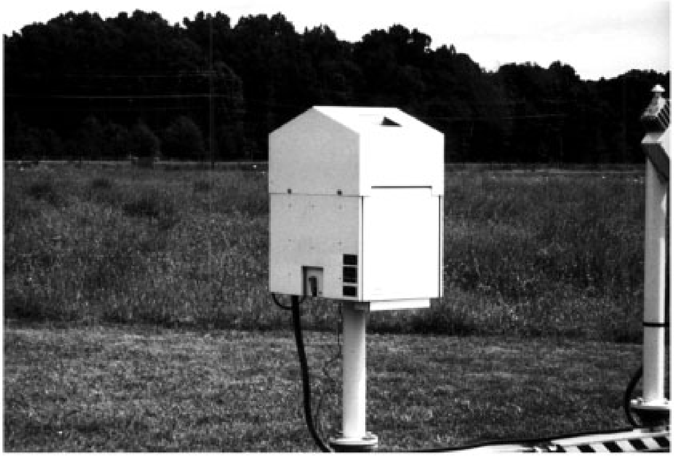


Figure 17 Laser beam ceilometer. See ftp site for color image.

Runway visibility measurements were important to aviation for landings and takeoffs. A fixed intensity visible beam of light was transmitted horizontally toward a receiver along a baseline of 500 ft. The height of the sensors was at 5 m (14 ft) above the centerline of the runway, the average height of an aircraft cockpit of that time period. The particulates in the air attenuated the light beam. The photons arriving at the receiver photodiode generated a small current, and an extinction coefficient was determined. This system was called a transmissometer. See Figure 19. The transmissometer readings were converted to meteorological visibility at this time. In 1955, Runway Visual Range (RVR) was implemented. In addition to the transmissometer reading, this system took into account the light setting or

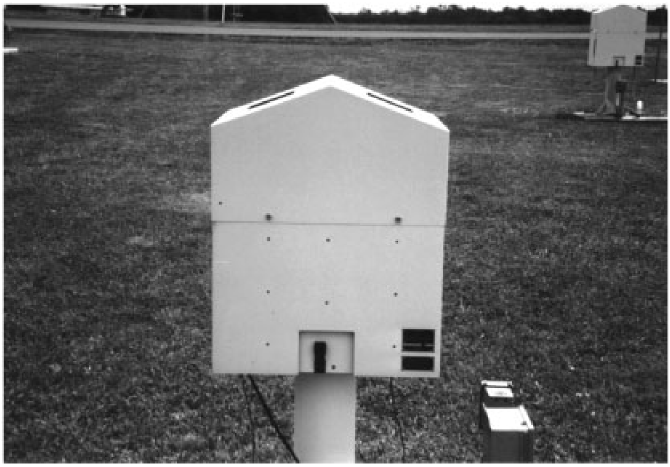


Figure 18 Laser beam ceilometer showing transmitter and receptor windows. See ftp site for color image.

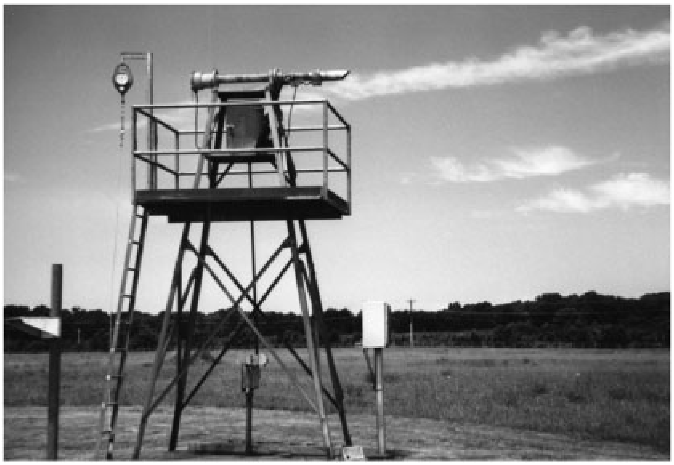


Figure 19 Transmissometer. See ftp site for color image.

intensity of the runway edge lights and whether it was day or night by the use of a photometer. Daytime readings were based on the attenuation of contrast, while nighttime readings were based on the attenuation of flux density. The derived visibility was expressed in increments of 200 ft below 4000 ft and increments of 500 ft above 4000 ft. The initial range of values was from 1000 ft to 6000+ ft. Transmissometer baselines were lowered to 250 ft at certain airports to permit RVR readings down to 600 ft.

A forward scatter visibility sensor was developed in the 1970s. A beam of light is emitted from a projector and is scattered forward by any particulates in the air to a receiver. See Figure 20. The axis of the projector and the receiver are placed a few

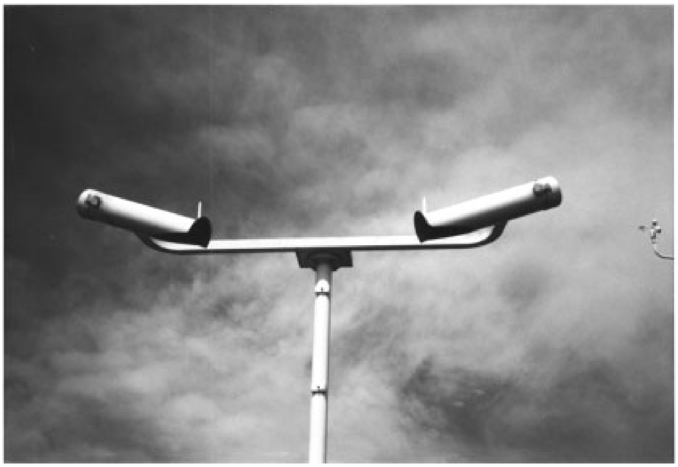


Figure 20 Forward scatter visibility sensor showing projector and receiver. See ftp site for color image.

degrees apart to prevent any direct detection of the beam of light. The light source is visible xenon light pulsed twice per second. The sampling volume is 0.75 ft^3. An extinction coefficient is determined and a sensor equivalent visibility is determined with day/night input from a photometer. Experiments were conducted using infrared light, but it was found to be inferior to visible light, especially with certain atmospheric particles such as haze.

10 CALIBRATION, MAINTENANCE, AND QUALITY CONTROL OF INSTRUMENTS

Obtaining quality meteorological sensors and locating them according to siting and monitoring standards are only the first steps in establishing a quality data acquisition network. Provisions need to be made for maintenance, calibration, and quality control.

Sensors are usually calibrated initially at the factory but sometimes need to be recalibrated at the site because of shipping and site peculiarities. The sensors also need to be recalibrated from time to time because of mechanical and electrical drifting. Recalibration intervals vary with each type of sensor. Data network managers need to know what initial calibration was done, when the sensors need periodic calibration, and have procedures in place to ensure proper calibration.

Sensors fail and maintenance philosophies vary. They range from on-site repair to replacing the failed sensor, to abandonment. Another consideration is the availability and location of maintenance, the required response time and the cost.

Sensors produce data, but procedures need to be in place to determine if the data produced are within the sensor specifications. Users will usually detect grossly bad data. However, a lack of knowledge about local climatology or subtle changes occurring with time or particular weather events, where a number of different weather conditions are possible, may result in bad data going undetected. A monitor(s) needs to quality control the data from a network to ensure its quality. If the equipment is malfunctioning, this person needs to notify the appropriate maintenance personnel.

When a network is new, the above issues may not seem important, but with the passage of time, they become the most important issues threatening data quality, continuity, and availability.

11 TELEMETRY AND AUTOMATION

The sensors described previously are part of the surface and hydrologic data networks. The most familiar parts of this network are located at airports throughout the country. Observations from these networks were usually transmitted by teletypewriter. These observations from a large-scale network were essential for forecasting and making specific point observations. They fell short, however, in providing sufficient data for the hydrologic, marine, agriculture, climatic, and fire weather programs. In the early years of the NWS, data from these off-airport networks were taken manually and relayed to a weather office either by telephone or by radio. Sometimes data took hours to reach a forecast center. There was the time interval between the observation and the phone call. More time delays were caused by busy phones at a collection center and the time needed to relay the report to a forecast center by phone, radio, or teletypewriter. In every step, there was a chance for human error.

In 1945, a three-station experimental network was deployed south of Miami to report pressure, wind speed, and direction every 3 h. The data were transmitted by a radio powered by batteries that were charged by a wind generator. Data were deciphered by the position of an information pulse between two reference pulses.

During these early years of automation, the emphasis in systems development was on telemetering hydrometeorological parameters. Systems were designed that would permit data to be transmitted from a remote sensor to a collecting office by hardwire, telephone line, or radio. See Figure 21. Sensors were designed with cams, coded discs, weighing mechanisms, potentiometers, shaft encoders, and other apparatus that would transform analog data into an electrical signal. In general, these signals had to be manually decoded at a collection center by counting beeps.

In the 1960s, the transistor era, plans were made to automate at least part of this network for three reasons. The first was to make it possible to obtain data from sites where observers were unavailable. Second, data were needed to be available at any time to make forecasts and warnings timelier. The third reason was to reduce the workload at weather offices. During weather emergencies, many hours were required to manually collect data by telephone.

An automated hydrometeorological observing system was developed using solid-state electronics. Data were collected via telephone from remote instruments connected to a collection device and powered by battery and solar cells. See Figure 22. The remote system transmitted a fixed-format, variable-length message coded in American Standard Code for Information Interchange (ASCII) format. In this same time period, the Geostationary Operational Environmental Satellites (GOES) with communications relay facilities became available. Now data could be made accessible from areas where no or unreliable telephone service was available.

With time, further improvements were made. Automatic event reporting sensors replaced sensors reporting only on a fixed schedule to improve sampling for small areas and short events. These data are also being used to provide ground truth for satellite imagery and Doppler radar. This was a big step from when the beam from the old WSR-57 radars was stopped above a rain gauge equipped with a transponder. The amount of precipitation in the gauge was displayed as blips on the radar screen.

Almost any analog or digital sensor can now be interfaced to a microprocessor-based data collection system. The first level of processing converts raw sensor data to engineering units. The second level is used to determine such things as maximums, minimums, means, variances, standard deviations, wind roses, histograms, and hydrographs. Simple or sophisticated algorithms allow transmissions to be sent following an out-of-limits reading on a specific sensor.

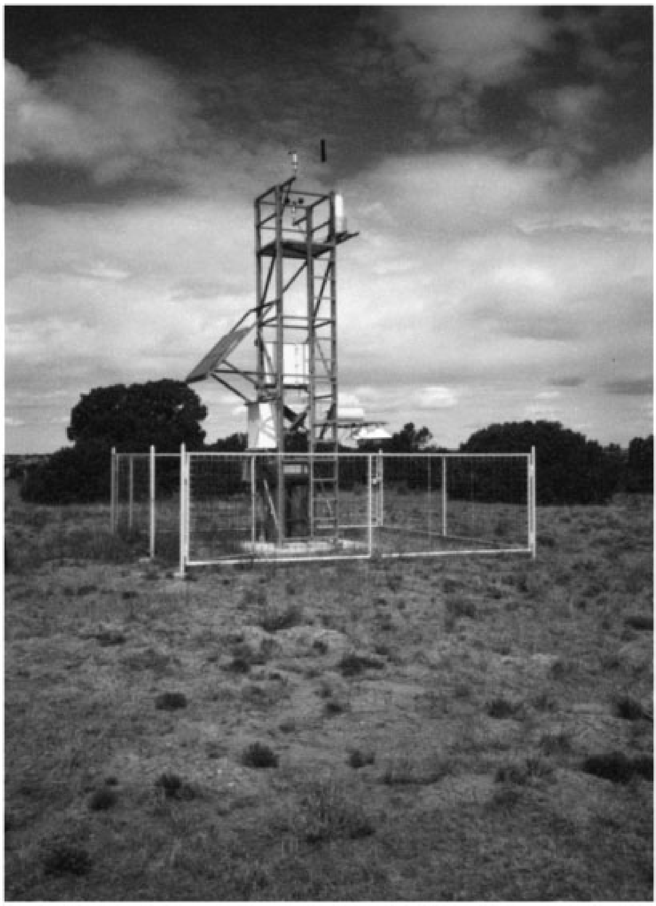


Figure 21 Early remote automated meteorological observing system (RAMOS), late 1960s. See ftp site for color image.

Systems such as the GOES Data Collection System also afford effective data sharing among agencies rather than what agencies would encounter using their own independent systems. The data from the approximately 14,000 remote sites in this coordinated network, not only provide data for forecasting and warning needs but are also cost-effective to the taxpayer when interagency duplication is avoided. See Figure 23.

12 IMPORTANCE OF STANDARDIZATION AND CONTINUITY

Users of meteorological data can only validly compare those data that have been collected by standard methods. This conformity makes it possible to determine meteorological patterns and how they are changing with time. Good standards

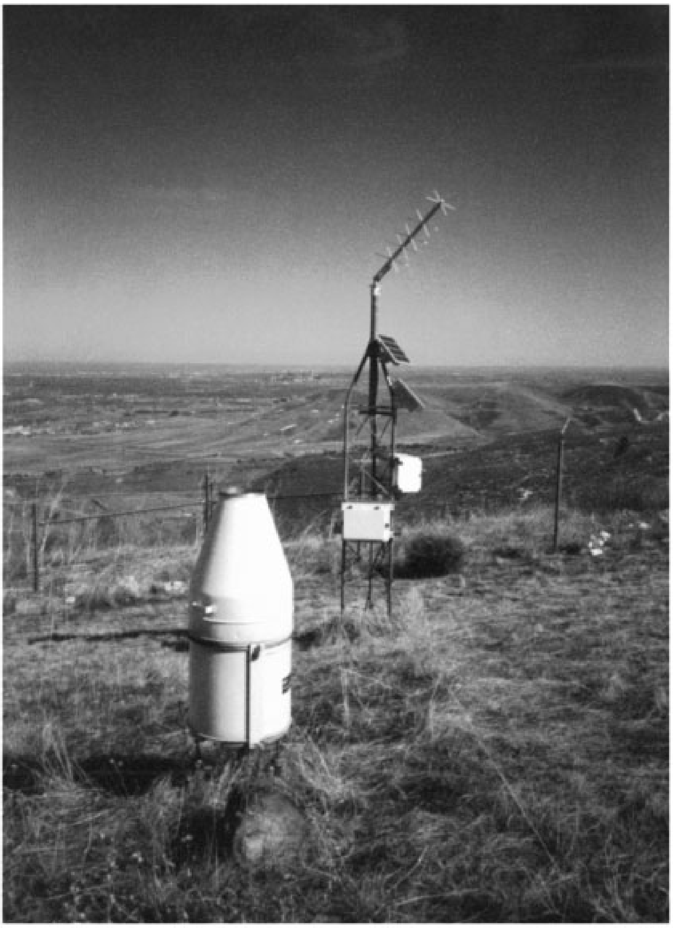


Figure 22 Automated hydrologic observing system (AHOS) with GOES radio transmitter and antenna in background. See ftp site for color image.

evolve from a process that involves many of the users with their various interests. When the needs of participants such as academia, government, and the private sector are considered, the results will be a consensus and consequently accepted and used.

Changes in instrument types, modifications, and relocations are inevitable because of economic and scientific reasons or for changes in individual user missions. These changes need to be documented or data discontinuities will occur. Users need to know the resolution, accuracy, and the temporal and spatial specifics of each data site. This is becoming more important with the many studies being conducted on the state of the environment and determining changes in it. There is increasing cooperation among data providers within certain disciplines, but generally there is a lack of understanding across these disciplines with regard to data standards, quality, and continuity. See Figures 24 and 25. Without this cooperation, we

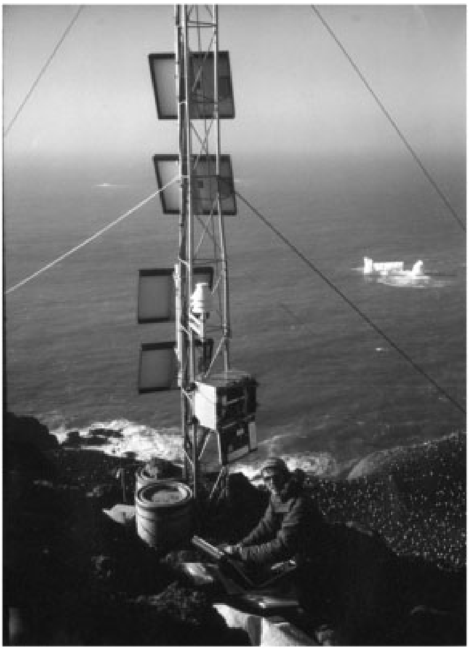


Figure 23 Interagency automatic remote collector (ARC) using satellite communications on Macaroni Ridge in Antarctica. Weather data are used in penguin breeding studies. Hundreds of penguins in lower right background. See ftp site for color image.

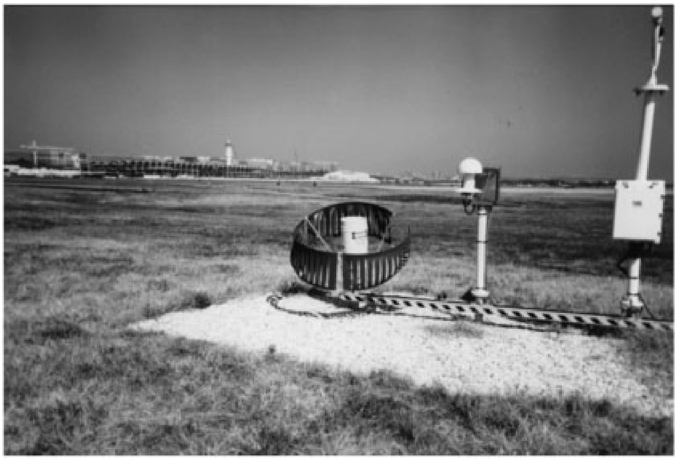


Figure 24 Sensors located too close to a jet takeoff area. Windscreen around the rain gauge is damaged from jet blast. Blast also causes temperature to rise and wind gusts to increase. See ftp site for color image.

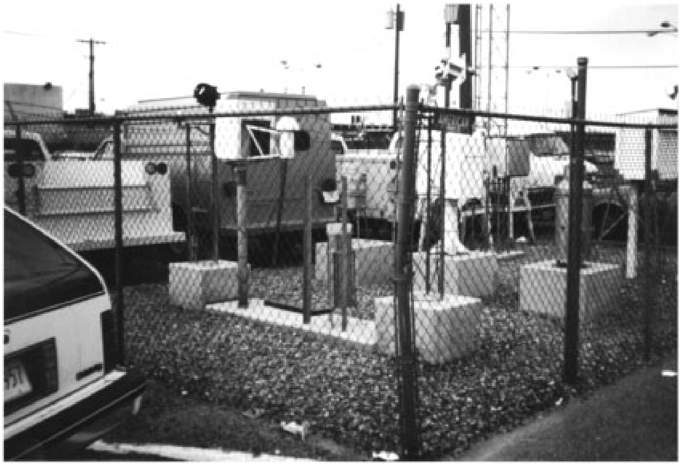


Figure 25 Sensors poorly exposed in a parking lot. See ftp site for color image.

will not be able to determine whether we are measuring changes in the environment or just variations in the data system.

BIBLIOGRAPHY

Cornick, J. C., and T. B. McKee (1993). A Comparison of Ceiling and Visibility Observations for NWS Manned Observation Sites and ASOS Sites, Atmospheric Science Paper #529, Colorado State University, Fort Collins, CO.

Federal Aviation Agency and Department of Commerce (1965). Aviation Weather for Pilots and Flight Operations Personnel, Washington, DC, U.S. Government Printing Office.

Flanders, A. F., and J. W. Schiesl (1972). Hydrologic data collection via geostationary satellite, *IEEE Trans. Geosci. Electro.* **GE-10**(1).

Kettering, C. A. (1965). Automatic Meteorological Observing System, Weather Bureau Technical Note 12-METENG-16.

National Weather Service (1995). *Federal Meteorological Handbook No. 1, Surface Observations*, FCMH-1, Washington, DC, U.S. Government Printing Office.

Schiesl, J. W. (1958). Measurement of Dynamic Strain, Needle Research Report, Singer Manufacturing Company, Elizabethport, NJ.

Schiesl, J. W. (1976). Automatic Hydrologic Observing System, Paper presented at the International Seminar on Organization and Operation of Hydrologic Services, July 1976. Ottawa.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1994). Federal Standard for Siting Meteorological Sensors at Airports, FCM-S4-1994.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1997). National Geostationary Operational Environmental Satellite (GOES) Data Collection System (DCS) Operations Plan, FCM-P28-1997.

U.S. Department of Commerce, Weather Bureau (1963a). *Manual of Barometry*, Washington, DC, U.S. Government Printing Office.

U.S. Department of Commerce, Weather Bureau (1963b). *History of Weather Bureau Wind Measurements*, Washington, DC, U.S. Government Printing Office.

CHAPTER 37

CONSEQUENCE OF INSTRUMENT AND SITING CHANGES

JOSEPH W. SCHIESL AND THOMAS B. McKEE

1 INTRODUCTION

In 1984, the National Weather Service (NWS) began the deployment of the H083 as the standard operational hygrothermometer in use at approximately 400 NWS and Federal Aviation Administration (FAA) field stations across the United States. The data are used in support of aircraft operations at airports and by NWS meteorologists for forecasting.

The purpose of the H083 was to measure temperatures at airports where the required accuracy was ±1°C (1.8°F) at temperatures between −50 and 50°C (−58 and 122°F). This was also the aviation standard of the World Meteorological Organization (WMO). Even though these instruments at airports were not meant to meet climatological requirements (there are separate networks for those), some people use these data as an index of change for the local area. In addition, airports are notoriously poor for climatological measurements because of the solar absorption characteristics of runways and roads plus the heating effects of buildings and moving aircraft. See Figure 1. It is important and even critical to determine these heating factors since they are essential in computing aircraft takeoff distance.

2 EARLY MODIFICATIONS

Improvements were made to the H083 since its initial deployment to improve its accuracy, reliability, maintainability, and to correct several deficiencies that were identified after long-term exposure in a field environment.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2 © 2003 John Wiley & Sons, Inc.

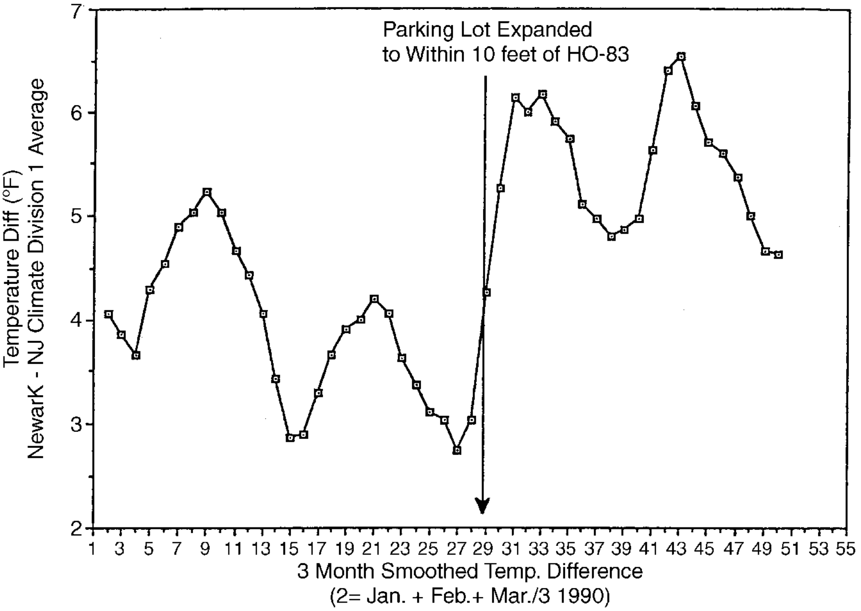


Figure 1 Temperature difference between Newark, New Jersey, airport, and surrounding climate stations. Note the change coincident with the parking lot expansion.

The improvements were concentrated in four areas: (1) an optical block, (2) board/connector corrosion resistance, (3) a new mirror, and (4) aspirator housing and aspirator improvements. See Figures 2 and 3.

The sensor assembly was constructed with an optical block, which was devised to simplify factory alignment and eliminate field misalignment of the electro-optical components during mirror cleaning. The electro-optical components of the dew-point sensor were previously soldered to a circuit board and were aligned at the factory by bending the component leads into position. These components are now rigidly mounted in an optical block assembly, which ensures precise component alignment relative to the reflective mirror surface. However, it is important that the optical loop calibration procedures be followed correctly. Electronic readings must be allowed to settle. This is easy to say but difficult to do in subzero temperatures or other inclement weather conditions.

The redesign of the new sensor assembly and the use of new fabrication techniques eliminated the formation of corrosion on the sensor board. Corrosion on the sensor assembly, especially in the vicinity of the connector, can result in electrical leakage paths on the printed circuit boards in high-humidity conditions. This leakage can lead to erroneous calibration and simulated temperature diagnostics, incorrect temperature readings, and, eventually, complete failure of the sensor. The new design featured a hand-wired, potted connector whose pigtails are soldered to plated-through holes on the sensor board. Printed wiring circuits on the boards were

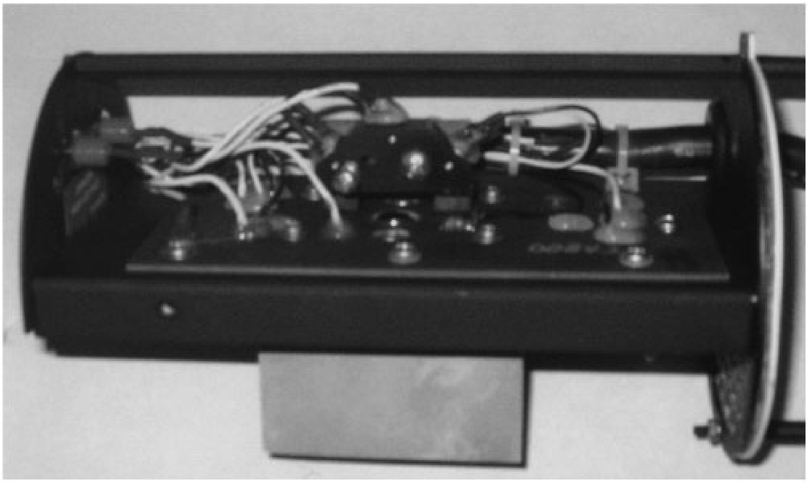


Figure 2 Hygrothermometer sensor board. Pyramid shaped object is the optical block. A round mirror is beneath the optical block. Peltier device is below the mirror on the opposite side of the sensor block (See Fig. 3).

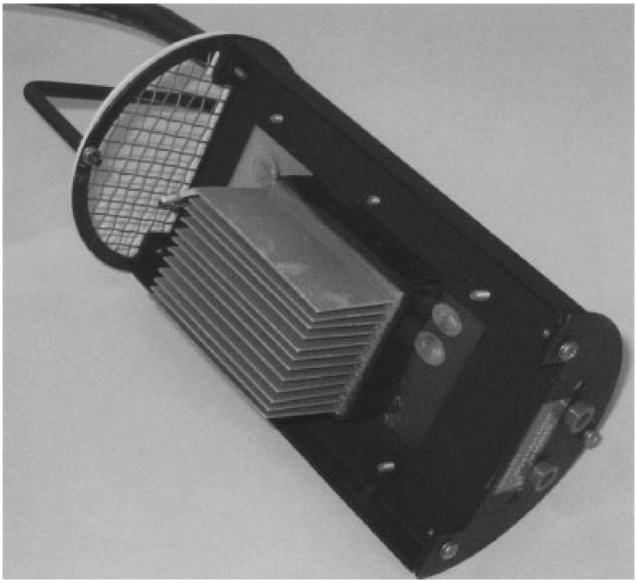


Figure 3 Peltier device on back side of sensor board. The aspirator would be located at the end of the sensor board toward the upper left in the figure.

virtually eliminated and the board was manufactured with a protective solder mask coating.

The new dew-point mirror was designed to provide proper operation with the optical block and to extend the life of the mirror surface. The mirror life was extended by improving fabrication techniques, including thorough cleaning prior to plating and using improved precision plating technology.

Consistency is critical for the components in the optical loop. Variations in the mirror, light-emitting diodes, and phototransistors cause calibration errors especially at high and low temperatures. Combinations of inconsistency, corrosion, excessive contaminants, improper cleaning, and calibration result in mirror icing.

The H083 aspirator/housing modifications were designed to minimize the effects of solar heating of the aspirator. The first modification was a “top hat” for the hemispherical part of the aspirator, which provided for an inch of expanded polystyrene insulation and additional shading of the stem. The second modification provided increased airflow through the aspirator by decreasing the airflow restrictions on the temperature side of the aspirator.

3 NEW CONCERNS

While plans were being made to do validation testing of the redesigned sensor and aspirator before deployment in the field, another concern surfaced. In July, 1989, meteorologists at the University of Arizona expressed concern about anomalous high-temperature readings at Tucson Airport. The Tucson Airport was breaking numerous records. Scientists at the National Climatic Data Center confirmed a warm bias at Tucson but also stated that they did not see this bias in other parts of the country.

At this point, NWS management established a Hygrothermometer Working Group (HWG) to investigate this problem and come up with an appropriate solution. The HWG was composed of meteorologists and engineers from the government, academia, and the manufacturer. To determine the extent of the problem, a survey of the weekly temperature comparisons around the country was conducted. The results showed that only Tucson was out of specifications, but there was definitely a warm bias in the Southwest on days with low winds and high solar radiation.

Tests conducted in the spring and summer of 1990 by NWS’s Equipment Test and Evaluation Branch in Sterling, Virginia, indicated that the H083 was susceptible to solar-radiation-induced errors that could approach 1.7°C (3.0°F) under certain low wind speed conditions. While errors of this magnitude were acceptable in the H083 temperature specification as it existed at that time, steps were taken to improve the accuracy of the hygrothermometer.

The next task of the HWG was to clarify the H083 temperature specification. The original H083 requirements were for an accuracy of 0.5°C (0.9°F) root-mean-squared error (rmse) based on monthly computations, a maximum error of ±1.0°C (1.8°F), and 95% of the monthly data should be less than the specified errors. The revised H083 temperature specification was also adopted for the model 1088 hygrothermometer, also known as the Automated Surface Observing System (ASOS) hygrothermometer. The new specification, requires an accuracy of 0.5°C (0.9°F) rmse (computations based on 7 days of data) and an absolute maximum error of ±1.0°C (1.8°F). These requirements were for temperatures between −50 and 50°C (−58 and 122°F). The rmse and absolute errors were doubled for temperatures outside this range down to −62°C (−80°F) and up to 54°C (130°F). The 95% requirement was deleted, which tightened the accuracy requirements significantly.

Through the preliminary testing, all hygrothermometer testing had been done at Sterling. The HWG decided that concurrent testing should be conducted at additional field sites to determine whether the solar-radiation-induced errors seen at Sterling were representative of those that would be expected under worst-case conditions. It was decided to conduct the initial field testing at Tucson, Arizona. First, there was an interest in the hygrothermometer readings at Tucson. Second, Tucson is in a climatological area where we expected meteorological conditions to be most conducive to maximum solar-radiation-induced errors—high solar radiation with low winds that occur during portions of spring and again in late summer.

Besides Tucson, the group decided to test the modifications in the north central (Sault Ste. Marie, Michigan) and the southeastern (Daytona Beach, Florida) part of the country for the extremes of temperature, humidity, and solar radiation, both direct and indirect. Soon after this, another test site was established at the Ocean City, Maryland, airport because of its proximity to the Atlantic Ocean. This site proved valuable for accelerated corrosion testing.

The H083s that incorporated the early modifications were field tested at Sterling in late 1990 and early 1991 and were found to meet the improved H083 accuracy requirements. Based on these data, the HWG decided to validate the modifications in Tucson. A field test bed was installed in April, 1991, but it was soon evident that solar loading/wind conditions in Tucson contributed to errors that exceeded the accuracy specification.

The group then undertook the task of working to reduce the amount of solar-induced temperature measurement error. Several system modifications were made to insulate the temperature probe from solar heating and improve aspiration through the sensor assembly. Preliminary results showed a dramatic improvement in sensor performance.

4 INITIAL TUCSON TEST RESULTS

A prototype of a modified H083 sensor (mod 1) was installed in the Tucson, Arizona, test site on June 24, 1991. Mod 1 had a larger fan with approximately three times the airflow capacity of the earlier version. In addition, the flow direction was reversed. The reversal was made to minimize the heating of the ambient air by the sensor housing itself. Data, during days of high solar loading, showed the average solar-induced temperature rise to be approximately 0.6°C (1.0°F) versus the 1.1 to 1.7°C (2.0 to 3.0°F) for the early double domed insulated hat, and 1.7

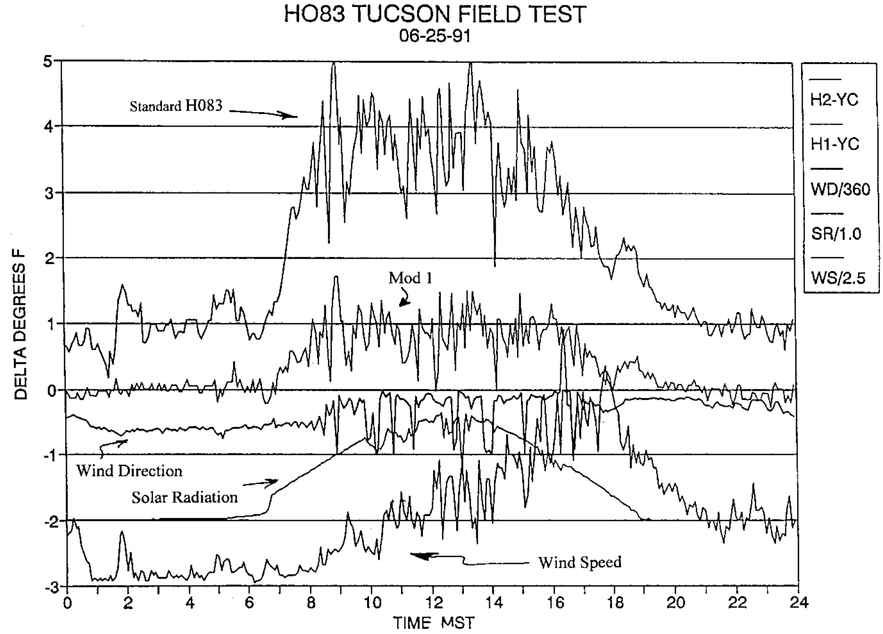


Figure 4 Comparison between standard H083 and Mod 1 at Tucson, Arizona. Mod 1 has a larger fan with approximately three times the air flow of the standard H083 and the flow direction has been reversed. Temperature deltas (°F) are in comparison to the Young reference sensor.

to 2.8°C (3.0 to 5.0°F) for the standard H083. See Figure 4. These deltas are all in comparison to the Young aspirated precision reference thermometer.

Even though good progress was made in solving the solar loading problem in the Southwest, the HWG was not in a position to firmly state that the problem was solved. The final modification had to be tested through the four seasons and in the different climatological regimes that were chosen. Since the temperature sensor was now located at the bottom of the instrument at the aspirator inlet, the group had to be certain the sensor was protected from any albedo effect either off snow or a natural reflective soil typified at Tucson, Arizona. Also, the HWG had to make sure the minimum temperatures were not affected or more importantly the dew point. Tucson, Arizona, with its temperature–dew-point depressions reaching 47°C (85°F), represents only one end of the humidity spectrum.

5 OVERALL TEST RESULTS

While the HWG was evaluating the solar loading modifications, it became apparent that there were a number of other problems with the H083. During temperature bath calibration tests, calibration errors were found that varied from −0.8 at 54°C to 1.2 at −51°C (−1.5 at 130°F to 2.1 at −60°F). It was also found during calibration stability tests that the sensor calibration drifted excessively with temperature, which would affect the actual sensor accuracy in the operational mode. These performance characteristics add a large degree of uncertainty to interpretation of H083 field performance test data. The problems were serious enough to suspend field testing during November and December of 1991.

Based on testing conducted in late December, 1991, and early January, 1992, the HWG determined that the latest modified sensor now had more stable electronics and deployed these modifications at the field test sites.

Both the standard and modified systems were installed at Tucson, Sault Ste. Marie, Daytona Beach, Ocean City, and Sterling. The HWG was able to measure not only the instrumental biases at these diverse climatological sites but also could determine why the biases occurred. This was possible because pyrheliometers, wind sensors, and reference temperature sensors were installed and carefully calibrated at each site. The effects of a partial solar eclipse were captured on both systems, which clearly demonstrated the solar loading problem. See Figure 5. These test beds were set up to determine the temperature biases between the different sensors under different wind and solar loading conditions. The resolution of the data was in minutes, rather than by the hour or the daily maximums and minimums presented in previous studies. Having a calibrated standard at each site also allowed the determination of how

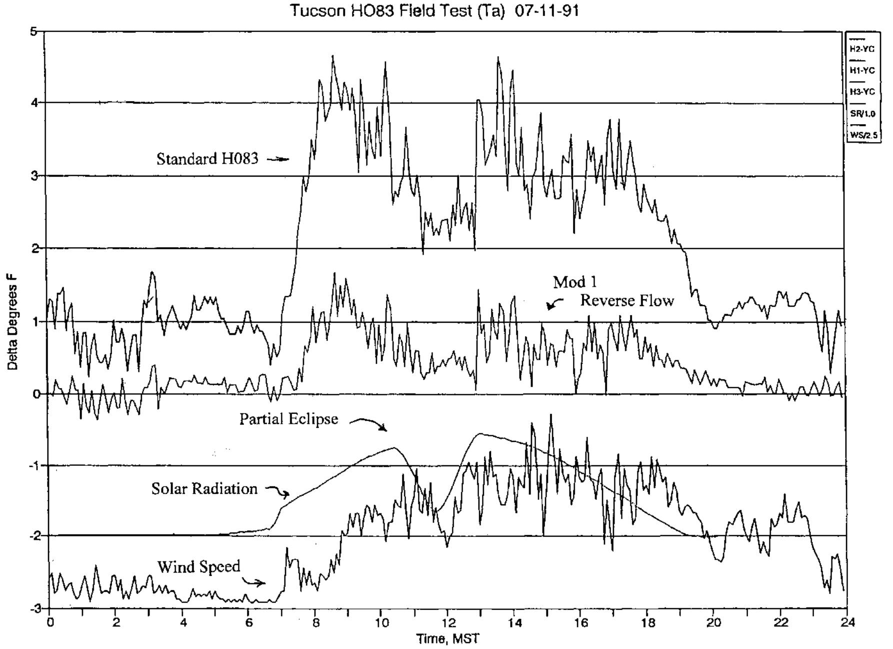


Figure 5 Similar to Fig. 4, but with a partial solar eclipse.

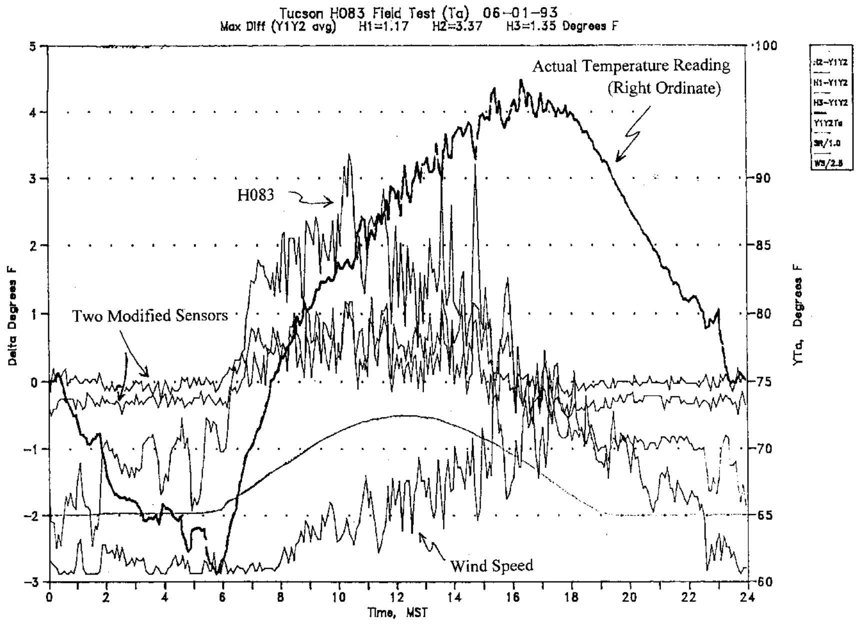


Figure 6 Similar to Fig. 4, but with two modified sensors.

much each sensor differed from an established reference, rather than just stating the difference between the standard and modified systems. See Figure 6.

6 AN OVERCONSERVATIVE DATA QUALITY CHECK

Of all the ASOS sensors, the hygrothermometer generated the most data quality failures. These failures generally indicate that the data coming from the sensor are suspicious. This suspicion can be caused by an internal self-test or the data quality algorithms. One of the processes initiated by this failure is the setting of the maintenance character ($) in all future observations that can only be cleared via the technician interface. In the case of the hygrothermometer, the data quality algorithm is the main cause of quality failures.

The ASOS hygrothermometer continually measures the ambient temperature and provides sample values about six times per minute. Processing algorithms in the hygrothermometer use these samples to determine a 1-min average temperature valid for a 60-s period ending at M+00 (minute + 00 seconds). Once each minute the 5-min average temperature is calculated from the 1-min average observations. These 5-min averages are rounded to the nearest degree Fahrenheit, converted to the nearest 0.1 degree Celsius, and reported once each minute as the current 5-min average temperature.

A number of data quality checks are performed by an acquisition control unit, including a rate of change check. Originally, if the current 1-min temperature differed from the last respective, nonmissing, 1-min reading within the previous 2 min by more than 3.3°C (6.0°F), it is marked as missing. If there are less than four valid 1-min average temperatures within the past 5 min, then the current 5-min average temperature is not computed. In this case, ASOS will use the most recent 5-min average calculated temperature within the last 15 min. If no valid 5-min average temperature is available within the last 15 min, a sensor failure is indicated. The 15-min delay allows for a once-a-day calibration heat cycle to occur without causing a data quality flag.

This initial rate turned out to be too conservative for the way the atmosphere was behaving. Many meteorologists were naturally skeptical of large temperature changes over short time periods for a number of reasons. Forty to 50 years ago, with the limited data available, it was envisioned that temperature fluctuations of more than two or three degrees Fahrenheit per minute were very rare events. High-resolution temperature data in both space and time from mesonetworks were not generally available. Thermographs were available, but their limitations will be discussed later.

Experience reinforced this conservative view of temperature changes. Most times, forecasters only saw the hourly weather reports where larger temperature rates were smoothed out. Observations in between the hourlies (specials) that were generated by significant changes or the onset or cessation of other meteorological elements did not even require the recording of temperature and dew point.

Those who examined thermograph traces saw more drastic changes over time, but these too were muted. In our efforts to protect the sensor from the elements as well as the sun or other radiating surfaces, we used thermometer screens. These lengthened the response time.

This conservative view was in contrast to some records that were available. For example, the following extraordinary events are noteworthy. There was the temporal change in the surface temperature on January 22, 1943, at Spearfish, South Dakota, from −20 to 7°C (−4 to 45°F) in 2 min, which was caused primarily by a Chinook wind. Then there were the spatial differences caused by cold air collecting in the hollows on low-wind nights. This was dramatized by the car ride by Middleton and Millar through Toronto in 1936 that showed differences of 14°C (26°F) over a mile.

Then there was the way thermometers were read. Temperatures and wet bulbs at airport stations were read to tenths of degrees primarily for the computation of the dew point and humidity. At low temperatures, a difference of only a few tenths in wet-bulb depressions may mean differences of relative humidity of about 10%. Temperatures can be measured electronically to a resolution of a thousandth of a degree; however, in meteorological applications this is meaningless.

In the late 1960s, with the advent of automated and telemetered data systems like the Automatic Hydrologic Observing System (AHOS), data were taken at 1-min intervals. More abrupt changes in hydrometeorological data became evident. Subsequent development of automated systems that also transmit data based on rate and threshold algorithms, such as the Automatic Remote Collector (ARC) in

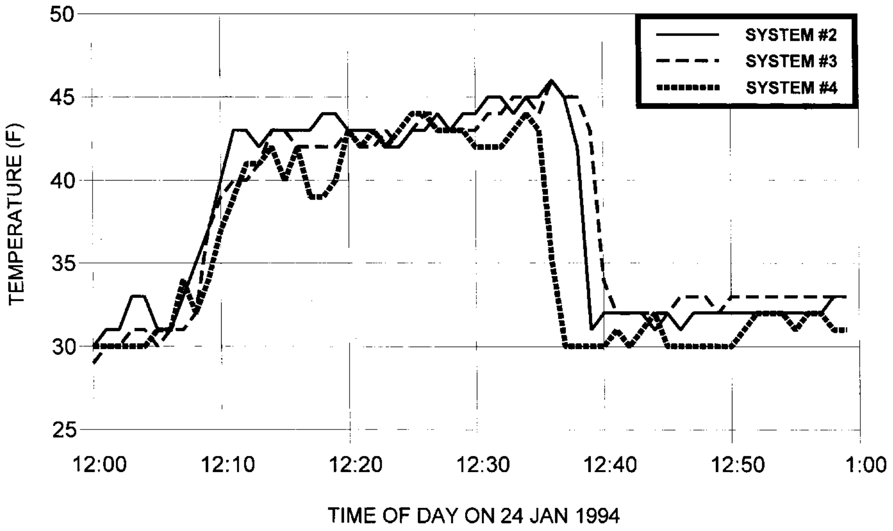


Figure 7 One-minute temperature data showing cold front passage from three noncollocated ASOS hygrothermometers at Sterling, Virginia.

the late 1970s, brought these changes to light even more. Today, these systems and ASOS provide us with very detailed data.

Figure 7 shows the passage of a cold front as recorded by three noncollocated hygrothermometers at Sterling, Virginia. Note the drastic changes in temperature over a very short period of time. All three systems logged temperature data quality errors and the sensors were flagged inoperational as a result of this temperature drop.

Based on temperature rate data from the HWG's studies over 3 years, the HWG recommended that the temperature and dew-point data quality algorithm be changed. The change was implemented in a subsequent firmware load in ASOS. Now, the quality failure does not occur until the last reading from the temperature or dew-point sensor varies from the previous reading by more than 5.6°C (10°F).

7 RELOCATION BIAS

In addition to the instrumental biases discussed so far, another factor was to affect temperature readings at most ASOS sites. ASOS sensors are typically located near the touchdown zone of the primary designated instrument runway and on occasion at a center field location. These exposures for temperatures at airports were generally better where the H083s were not already remotely located from the airport office. Some earlier station temperatures were taken on roofs or near parking lots. See Figure 8. Even where the H083 was already remotely located, the ASOS sensor was sometimes located a mile or two away from the previous location. There were

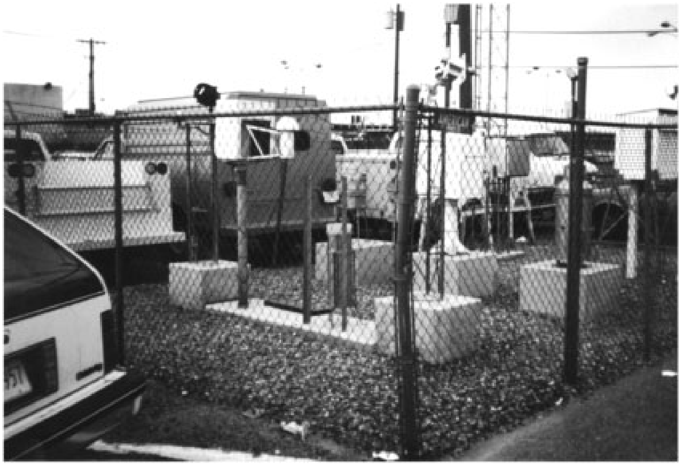


Figure 8 Sensors poorly exposed in a parking lot. See ftp site for color image.

sometimes significant exposure differences over these distances. Runways are flat but the areas around them are not always so. Suitable exposures for meteorological instruments are not always available because of the proximity to taxiways and instrument intrusion into safety regulated airspace. See Figure 9. As a result, sensors are sometimes located in swales where the temperatures are subject to cold air drainage. Regardless of the reason for the change in location, the differences needed to be accounted for.

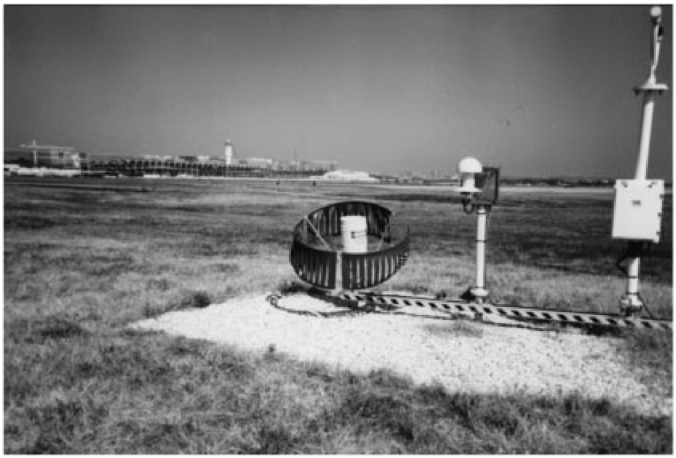


Figure 9 Sensors located too close to a jet takeoff area. Windscreen around the rain gauge is damaged from jet blast. Blast also causes temperature to rise and wind gusts to increase. See ftp site for color image.

8 DATA CONTINUITY STUDIES

The basic dilemma was that even if the 1088 were perfect, it was now known that the H083 was not. Thus, at all sites, an account must be made for that bias given by the uncertain mixture of H083 bias and removal factors. This could only be done by leaving the H083 in place and comparing it to the modified and much improved 1088. It was obvious that the simple sling psychrometer would not suffice. If the sling is perfect and the 1088 is perfect and they are collocated, nothing is proved. For true data continuity, there was a need to develop for each station its unique bias signature.

Unfortunately, because of resource limitation and schedules, only a subset of the ASOS sites could be studied. The H083s had to be returned to the manufacturer and modified so they would be the sensor equivalent of the improved 1088. Only a limited number of the original H083s could be left in place with the ASOS 1088 version. With this restriction, the sites selected for a climate data continuity project took into account as many different climatological regimes as practical. This research was conducted by the Department of Atmospheric Science at Colorado State University in Fort Collins, Colorado, and was reported by McKee et al. (1996, 1997). Studies of data continuity began with the 15 sites given in Table 1 for data collected from June, 1994, through August, 1995. All sites had received all of the modifications described above by June, 1994. These 15 locations retained the H083 during this period so the maximum and minimum daily temperatures were available.

TABLE 1 Climate Data Continuity Study (CDCP) Comparison Sites for Daily Maximum and Minimum Temperatures

Number	Site ID	Station Name
1	AMA	Amarillo, TX
2	AST	Astoria, OR
3	BRO	Brownsville, TX
4	BTR	Baton Rouge, LA
5	COS	Colorado Springs, CO
6	DDC	Dodge City, KS
7	GLD	Goodland, KS
8	GRI	Grand Island, NE
9	ICT	Wichita, KS
10	LNK	Lincoln, NE
11	OKC	Oklahoma City, OK
12	PWM[a]	Portland, ME
13	SYR	Syracuse, NY
14	TOP	Topeka, KS
15	TUL	Tulsa, OK

[a] Station commissioned in August, 1994.

The goal of the analysis was to understand the physical causes of the observed temperature difference between the two hygrothermometers. An initial step was to confirm the absolute accuracy of the ASOS hygrothermometer. The NWS examined several of the ASOS instruments at Sterling, Virginia, and found a range of ±0.20°F. Three additional comparisons were made in the field at Colorado Springs, Colorado (COS), Oklahoma City, Oklahoma (OKC), and Tulsa, Oklahoma (TUL), which showed a range of ±0.30°F relative to a field standard (RM Young), which had been calibrated against a secondary standard at Sterling, Virginia. Thus, ASOS does not have a temperature bias. The model used to assess the temperature differences had the analytic form:

$$\Delta T = \Delta T_i + \Delta T_\lambda + \Delta T_s \tag{1}$$

where ΔT is the observed temperature difference of ASOS–H083, and the subscripts are i (instruments), λ (local effect of location), and s (solar heating effect). All of the separations between the hygrothermometers were less than 1 mile and the hygrothermometers were collocated at four sites, which would have no local effect, by definition. Local effects due to site location could be different for day and night, various weather systems, and seasons. Solar heating was included since it was known that the H083 had a solar heating problem.

Results of the analyses are given in Table 2 for the 15-month period. Observed differences in the maximum and minimum temperatures (labeled M_x and M_n) show ASOS to be cooler by 1.17°F (M_x) and 0.86°F (M_n) averaged over all sites. Two methods were used to determine the instrument bias of the H083 (ΔT_i) assuming ASOS had no bias. The first was to examine comparisons when wind speeds were high enough to reduce local effects resulting in a narrow frequency distribution of ΔT. A speed of nearly 15 miles per hour was required, but there were not enough observations remaining to be meaningful at all sites. The second approach was to have an ASOS reported overcast sky at night, which meant cloud base was 12,000 ft or lower. Downward infrared radiation from an overcast low cloud would reduce horizontal temperature differences. In fact the frequency distribution of observed temperature differences was very narrow, which indicated the local effects were reduced. The instrument bias determined from this analysis is given in Table 2 in the column marked ΔT_i. The average ΔT_i was 0.57°F, and the range was quite large from 0.16 to 1.06°F. The ASOS at Lincoln, Nebraska, was moved midway in our data collection period. Two LNK sites are thus included. LNK-1was the first location and the ASOS instrument was moved to LNK-2, which was essentially collocated with the H083. The cloud analysis yielded the same estimate of ΔT_i from the two locations. Next the H083 bias was subtracted from the M_x and M_n observations, and the remainder for the M_n is ΔT_λ at night and for the M_x is a combination of ΔT_λ in the daytime plus the solar effect. Notice the resulting ΔT_λ from the M_x has a wide range of 0.56 to -1.10°F. This shows that local effects can be quite important for distances less than 1 mile. For the daytime local effect plus solar effects, several of the sites showed a magnitude of 1°F or larger indicating a likely strong solar effect, which was previously established as a problem for the H083. As a consequence of the

TABLE 2 ASOS–CONV (°F) June, 1994 Through August, 1995

				Bias Removed		
Station	M_x	M_n	ΔT_i	M_x $(\Delta T_s + \Delta T_\lambda)$	M_n (ΔT_λ)	ABOS–CONV Diurnal Range
AMA	−0.76	−0.59	−0.35	−0.41	−0.24	−0.17
AST	−0.59	0.03	−0.28	−0.31	0.31	−0.62
BRO	−0.93	−0.32	−0.45	−0.48	0.13	−0.61
BTR	−1.96	−1.19	−0.89	−1.07	−0.30	−0.77
COS[a]	−1.38	−0.41	−0.16	−1.22	−0.25	−0.97
DDC	−0.48	−0.91	−0.61	0.13	−0.30	0.43
GLD	−1.16	−1.19	−0.21	−0.95	−0.98	0.03
GRI	−1.25	−0.69	−0.67	−0.58	−0.02	−0.56
ICT[a]	−0.79	−0.27	−0.29	−0.50	0.02	−0.52
LNK-1	−2.14	−2.00	−0.96	−1.18	−1.04	−0.14
LNK-2[a]	−2.38	−1.03	−0.96	−1.42	−0.07	−1.35
OKC	−0.64	−2.03	−0.93	0.29	−1.10	1.39
PWM	−0.78	−1.18	−0.47	−0.31	−0.71	0.40
SYR[a]	−0.80	−0.42	−0.29	−0.51	−0.13	−0.38
TOP	−0.48	0.06	−0.50	0.02	0.56	−0.54
TUL	−2.17	−1.55	−1.06	−1.11	−0.49	−0.62
Average	−1.17	−0.86	−0.57[b]	−0.60	−0.29	−0.31

[a] Co-located sites.

[b] Value has LNK twice.

differences for M_x and M_n, the ASOS observed diurnal range also decreased in comparison to the H083 observations. A summary of this data continuity analysis is that ASOS is a better instrument than the H083 with much less bias, less solar influence, and a new recognition that local effects are important.

9 INTERFACE WITH OUTSIDE GROUPS

The ongoing work on temperature measurement was done in coordination with groups such as the National Academy of Science, the National Center for Atmospheric Research, a few universities, and a number of state climatologists. What became apparent was that the measurement of temperature is more complex than originally thought. For instance, one reason electronic thermometers report higher temperatures is because of their quicker response time. Liquid-in-glass thermometers miss quick temperature fluctuations because of their slower response. Cotton Region temperature shelters have an even larger bias than the H083 under the low-wind, high-solar-load conditions. The climatological community is looking at these problems, and the NWS will be working with them on documenting the biases among the various sensors, climatological regimes, sensor locations, and by certain weather parameters.

Without this cooperation, unexplainable temperature shifts would be a murky blend of a number of biases caused by different factors, and certain data would go down in the books with the comment: “Data quality/continuity uncertain.”

REFERENCES

Federal Aviation Administration and Department of Commerce (1965). *Aviation Weather for Pilots and Flight Operations Personnel*, Washington, DC, U.S. Government Printing Office.

Flanders, A. F., and J. W. Schiesl (1972). Hydrologic data collection via geostationary satellite, *IEEE Trans. Geosci. Electron.* **GE-10**(1).

McKee, T. B., N. J. Doesken, and J. Kleist (1996). Climate Data Continuity with ASOS (Report for the period September 1994–March 1996), Climatology Report No. 96-1, Colorado Climate Center, Atmospheric Science Department, Fort Collins, CO, Colorado State University, March.

McKee, T. B., N. J. Doesken, J. Kleist, and N. L. Canfield (1997). Climate Data Continuity with ASOS—Temperature, Preprints, 13th International Conference on IIPS, AMS, 2–7 February, Long Beach, CA, pp. 70–73.

National Weather Service (1992). ASOS (Automated Surface Observing System) User’s Guide. National Oceanic and Atmospheric Administration.

Schiesl, J. W. (1976). Automatic Hydrologic Observing System. Paper presented at the International Seminar on Organization and Operation of Hydrologic Services, July, 1976. Ottawa.

Schrumpf, A. D., and T. B. McKee (1996). Temperature Data Continuity with the Automated Surface Observing System, Atmospheric Science Paper No. 616, Climatology Report No. 96-2, Colorado State University, Fort Collins, CO.

U.S. Army Quartermaster Research and Development Command (1957). Environmental Protection Research Division, Weather Extremes around the World, Natick, MA, May.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1994). Federal Standard for Siting Meteorological Sensors at Airports, FCM-S4-1994.

CHAPTER 38

COMMERCIAL RESPONSE TO MEASUREMENT NEEDS: DEVELOPMENT OF THE WIND MONITOR SERIES OF WIND SENSORS

ROBERT YOUNG

The Wind Monitor wind sensor is an example of a commercial product development in response to customer measurement needs. Requirements of a single government agency led to the eventual development of an entire series of wind speed and direction sensors. The current models are the result of input from a multitude of agencies and end users whose application requirements vary dramatically—from sensitivity and responsiveness to survival in high winds, snow, and ice as well as desert heat and blowing sand—from ease of mounting and maintenance to durability and long-term performance—from simple analog output signals to polled digital serial data streams.

1 BACKGROUND

The Wind Monitor was created in 1979 as a small development project to try to satisfy the requirements of the National Data Buoy Center (NDBC) at Stennis Space Center, Mississippi. Now, approximately 23 years later, seven standard models plus several additional special models of the Wind Monitor are being manufactured to serve different customer requirements. Over 50,000 units have been produced (as of February, 2002) and are in worldwide service in more than 85 countries as well as

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

the Arctic and Antarctic regions and ships and buoys at sea. New applications are still being discovered.

From our company's early days we were advocates of the use of helicoid propellers for measurement of wind speed. Several products, which originally were categorized as "sensitive wind instruments" combined a helicoid propeller with a wind vane to provide measurements of both wind speed and wind direction in a single instrument. We manufactured propellers from very lightweight polystyrene foam for sensitivity and also from injection-molded polypropylene for greater durability. The typical wind sensor matched one of these propellers with a wind vane of comparable sensitivity. Typical fabrication materials were aluminum castings, machined aluminum and stainless steel, standard metal and plastic component parts, with stainless steel fasteners used for assembly. The wind speed transducer was typically a miniature tachometer generator, which required slip rings and brushes, and the typical wind direction transducer was a wirewound or conductive plastic potentiometer.

2 DEVELOPMENT

In the fall of 1979 we entered into a development contract with the National Data Buoy Center (NDBC; then referred to as the National Data Buoy Office) to design a propeller-vane-type sensor that would address its future needs for deployment on offshore buoys and remote stations. At the time the operational wind sensors for these applications were the J-Tec, which measured wind speed by means of a vortex shedding technique (utilizing a detector that counted the number of vortices shed from an obstruction in the tail assembly), and the Bendix Aerovane, a rugged sensor that utilized a propeller and vane combination. The main problem encountered with these sensors was the difficulty in mounting at sea due to their size and weight plus the need to install mounting bolts between the base and the mounting fixture on the buoy. Many of the buoys were a long distance offshore, which sometimes meant servicing the sensors in difficult weather conditions.

The design goals of the development contract were: small size (maximum 36 cm overall height; 46 cm overall length), light weight (2.5 kg maximum; 1.5 kg desired), simple mounting allowing one-hand replacement, corrosion resistant, no electrical slip rings, three- or four-blade helicoid propeller (maximum diameter 20 cm), working range 0 to 60 m/s and 0 to 360°, cost reduction, and reliability (18-month MTBF). The contract called for delivery of three prototype sensors for testing and evaluation.

The first prototype, which was delivered to NDBC in November, 1980, was machined aluminum with a detachable tail assembly (Fig. 1). It was intentionally designed for testing flexibility. Many different size and shape tail fins were tested to determine the optimum trade-off between overall size and dynamic performance. The wind speed sensor was an injection-molded polypropylene four-blade propeller, 18-cm diameter by 30-cm pitch, which had been previously developed. The wind speed transducer was a magnetically operated Hall effect switch, actuated by a two-

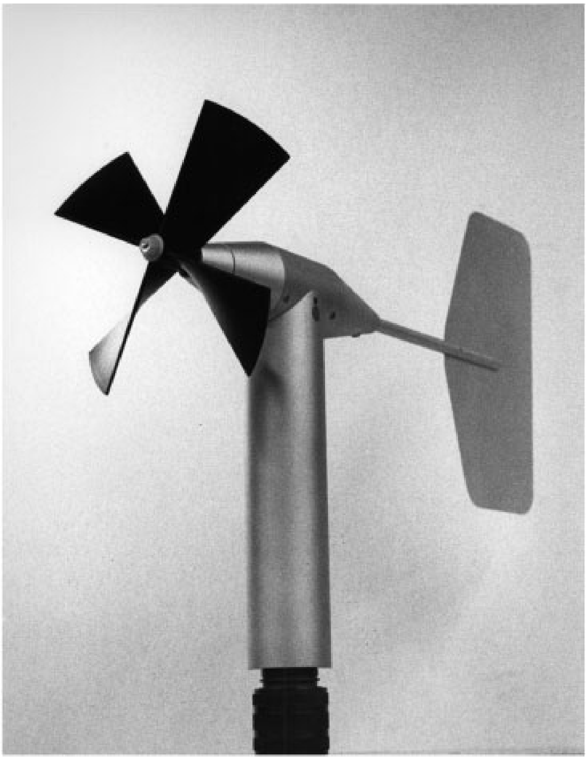


Figure 1 First prototype wind sensor, November, 1980.

pole magnet on the propeller shaft, which provided an output of one pulse per propeller revolution. The wind direction transducer was a conductive plastic 1 kΩ, 352° function angle, potentiometer which was mounted concentrically in the sensor housing and coupled to the movable vane above. Replaceable ball bearings and sleeve bearings were provided to conduct torque and speed tests as well as to assess long-term reliability. Early wind tunnel testing showed that sleeve bearings would overheat at the higher wind speeds and would not be suitable. The most difficult part of the design was actuation of the wind speed transducer that, to eliminate the need for slip rings and brushes, had to be mounted on the nonrotating inner part of the main body. Since the potentiometer was also located concentrically in the same area and being actuated by a coupling from the vane assembly, it was necessary to mount the Hall switch slightly off-center to clear the coupling. To accomplish this, the Hall switch was sandwiched between two soft iron circular plates. The circular magnet attached to the propeller shaft would then always be the same distance from the edge of the plates and the magnetic flux would be directed through the Hall switch as the magnet rotated.

3 PLASTIC FABRICATION

Two identical units of a second prototype design were delivered in March, 1981 (Fig. 2). The main purpose of these prototypes was to develop a streamlined and economical fabrication method since cost reduction was one of the primary design goals. The main housing, nose cone, and tail assembly of these units were constructed of thermoformed black vinyl plastic. Wood molds were fabricated for forming the main housing and tail in two identical halves, which were then cemented together. These molds could be easily modified to incorporate any desired changes. The tail assembly was filled with polyurethane foam. To achieve the desired combination of rigidity and low weight several different thickness and color vinyl materials were tried as well as different foam densities. The wind speed and wind direction transducers were the same as those used in the first prototype. A total of six sensors of this design were fabricated for evaluation by a couple of different organizations.

Work on a third prototype design was already underway while prototype II units were being field tested. We acquired a small injection molding machine (previous injection-molded propellers were made by an outside supplier) and began to redesign the nose cone assembly and other internal parts for injection-molding. To minimize outdoor ultraviolet (UV) deterioration, we had selected a black vinyl material for

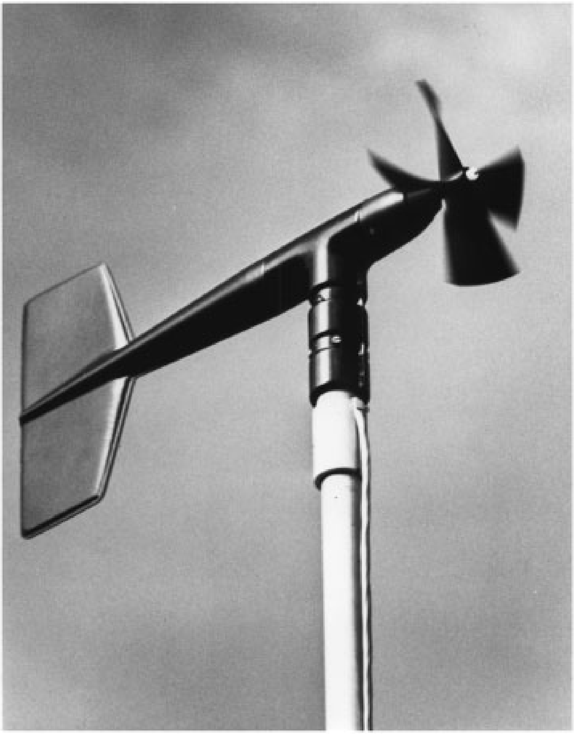


Figure 2 Second prototype wind sensor, March, 1981.

thermoforming the main housing and tail assembly. The wind speed and wind direction transducers remained the same. A machined aluminum mounting post with set screws provided for mounting on standard 1-in. schedule 40 pipe. This design was designated as Model 05101 Wind Monitor. Approximately 25 units of this design were produced in January, 1982. Field testing of these and the previous prototype units revealed a problem with the black vinyl softening in the hot sun and residual expansion of the internal foam causing irregular bumps in the surface of the tail assembly.

We changed to a superior material for thermoforming the main housing and tail assembly. The new material was a white ABS (Acrylonitrile-Butadiene-Styrene) plastic with UV inhibitors, which had been developed and extensively tested specifically for long-term outdoor exposure. Some changes were made in the shape of the main housing, and additional mating parts were designed for injection molding. An aluminum orientation ring was also supplied. The orientation ring, which was installed directly below the mounting post, had an index pin that allowed the sensor to be removed and replaced for service while maintaining its original orientation on the pipe support. These changes, which were included in the next production lot of 60 units produced in April, 1982 (still designated Model 05101), made a dramatic improvement in the performance of the sensor, especially in tropical conditions (Fig. 3).

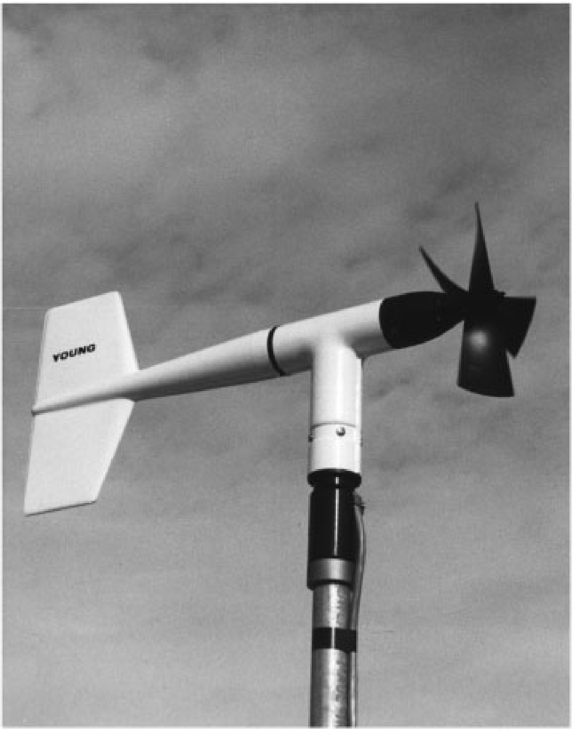


Figure 3 Model 05101 Wind Monitor, April, 1982.

4 CUSTOMER INPUT

Meanwhile a number of changes in design continued to be made in response to customer requests. The most significant change was the wind speed transducer. The Hall switch was replaced with a coil that was wound on a bobbin with a central hole, which allowed space for the potentiometer coupling to pass through. The wind speed signal then became an alternating current (ac) sine wave (in place of the 5-V square wave from the Hall switch) with the frequency directly proportional to wind speed. The main advantages of the coil were greater reliability and the elimination of the magnetic attraction of the soft iron flux plates required for the Hall switch that caused a noticeable "jogging" of the propeller at threshold wind speeds. A separate sensor interface circuit was developed to convert the wind speed signal to a square-wave pulse output (similar to the Hall switch) as well as a calibrated analog voltage output (0 to 1 V $=$ 0 to 100 mph). A small metal junction box with a terminal strip was added to the sensor mounting post to provide an easy cable connection point. Designated Model 05102 Wind Monitor, approximately 340 units of this design were produced between October, 1982, and July, 1984.

Early in the fall of 1983 we had begun to incorporate additional changes desired by NDBC and other customers. The potentiometer was changed from 1 kΩ, 352° function angle to 10 kΩ with a 355° function angle. NDBC also wanted a reduction in the number of external fasteners, which were difficult to deal with in a marine environment. Meanwhile Climatronics Corporation in Bohemia, New York, had been awarded a contract by the Federal Aviation Administration (FAA) to instrument 51 major U.S. airports with Low Level Wind Shear Alert Systems (LLWAS). The Climatronics proposal included the Model 05102 Wind Monitor as the system wind sensor. This contract resulted in a blanket order for 450 Wind Monitors that, added to the requirements of other organizations, was cause for a major redesign effort to incorporate numerous improvements that had been suggested by several different customers. One of the most significant changes was the elimination of nearly all external screws and set screws. The main mounting post and orientation ring were changed from machined aluminum, fastened with set screws, to injection molded plastic with stainless-steel band clamps for tightening on the standard 1-in. pipe. The entire nose cone assembly and the main body of the sensor were now being injection molded while the tail assembly continued to be fabricated by vacuum forming the white ABS plastic and filling with foam. A simple internal spring latch was designed to secure the main housing to the rotor of the mounting post assembly. The latch was accessed by removing the nose cone assembly that was now threaded with an O-ring seal. Previously, the nose cone was secured to the main housing with four screws. An injection-molded junction box with a slide cover was added to the mounting post. Special packaging was designed with die-cut foam fitted to a custom carton to provide optimum protection during shipment. This latest design was introduced in August, 1984, and designated Model 05103 Wind Monitor (Fig. 6). A sensor interface circuit and a 4- to 20-mA line driver circuit both with a choice of wind speed scaling in meters per second, miles per hour, or knots and a choice of 360° or 540° azimuth range were also made available to provide calibrated outputs for direct input into most recorders and data loggers, and also for long cable runs.

5 ICING PROBLEMS

The FAA was very concerned about the performance and survivability of the LLWAS Wind Monitor in icing conditions. A number of different ideas were tried for heating the sensor, and several test programs were carried out both in a controlled laboratory environment and in field conditions. A variety of commercially available wind sensors were tested simultaneously. It was apparent that wet snow, clear ice, and rime ice affected the performance and calibration of all mechanical wind sensors to varying degrees. Snow had a greater effect on cup anemometers than on propeller anemometers while both clear ice and rime ice caused degradation of performance of both types. The physically larger anemometers were able to operate longer in icing conditions before performance was badly deteriorated. Heating of rotating joints allowed some sensors to operate longer before complete freeze-up; however, because of loss of calibration with ice build up, the benefit of heating only rotating joints was doubtful. Climatronics developed a heating scheme that utilized a heat element on the sensor mounting pipe and an aluminum mounting post (in place of plastic) to conduct some of the heat to the sensor bearings. It was felt that this provided some benefit in helping the sensor to recover from freeze-up and was therefore adopted for the LLWAS program. Our conclusion was that only heating the entire sensor assembly with an array of heat lamps was sufficiently effective to warrant the cost. The heat lamp array worked satisfactorily for low wind freezing rain events when temperatures did not drop very far below freezing but was not useful for situations where higher wind speeds and colder temperatures made the heating ineffective.

During the winter of 1983–1984 a Wind Monitor was installed on Mt. Washington in New Hampshire to check on the effects of severe rime icing. During a prolonged rime ice event with sustained winds, the sensor continued to operate reasonably well due to the high rotation rate of the propeller, which prevented the rime ice from adhering, and the continuous action of the vane, which prevented the rime ice from bridging the gap between the vane skirt and the mounting post. After several hours of high winds and rime ice buildup, the sensor failed when it was impacted by an ice chunk that had dislodged from a structure up wind. From this test we concluded that the Wind Monitor could operate effectively in light rime ice situations with continuous wind but would probably not be suitable for heavy rime ice conditions.

6 WIDESPREAD ACCEPTANCE

Several organizations had been evaluating the Wind Monitor for some time. In 1984 Campbell Scientific in Logan, Utah, added it to its standard line of wind sensors and began to place multiple orders. In 1985 The U.S. Army Atmospheric Sciences Laboratory (ASL) at White Sands Missile Range began operational use of the Wind Monitor and began purchasing multiple units through New Mexico State University. Also in 1985 the National Oceanic and Atmospheric Administration (NOAA)/Pacific Marine Environmental Laboratory (PMEL) began operational use of the Wind Monitor for marine applications. Late in 1985 the National Data Buoy

Center began operational use of a special model of the Wind Monitor that was modified for its application. About 2 years later further modifications, including a machined aluminum mounting post, were made to provide NDBC a "ruggedized" version that could better withstand the rigors of buoy deployment in rough seas. Units previously shipped to NDBC were also retrofitted with these modifications.

In the fall of 1985 a new model of the Wind Monitor was developed to meet the requirements of the U.S. Environmental Protection Agency and the Nuclear Regulatory Commission for air pollution applications. This model utilized an expanded polystyrene foam tail for greater sensitivity (improved damping ratio and lower threshold). The polypropylene four-blade propeller was normally supplied; however, an optional larger and lighter propeller was offered at additional cost. This optional propeller was hand fabricated of carbon fiber and proved to be very popular. We imported these from a supplier in France; however, the cost of hand fabrication plus transportation and duty made them disproportionately expensive. The calibration of these propellers was also more variable than desired. Eventually we developed a much less expensive and more uniform injection-molded carbon fiber thermoplastic propeller, which we manufacture in our own plant and supply as standard. This second standard version of the instrument is designated Model 05305 Wind Monitor-AQ (air quality model) (Fig. 4).

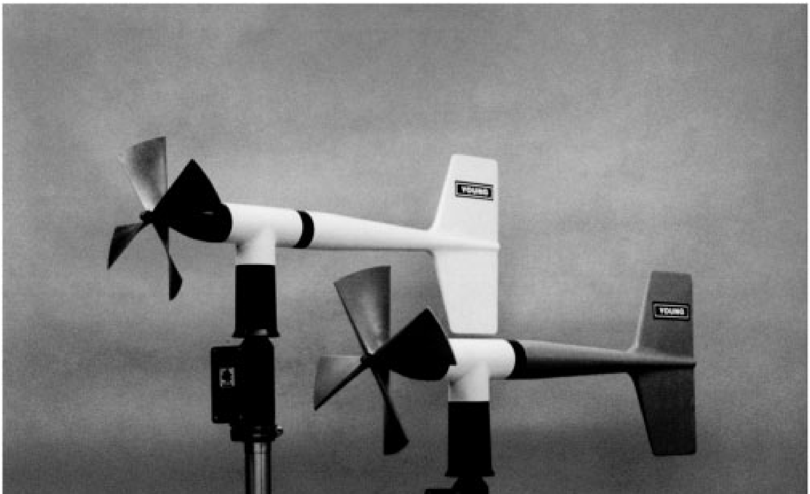


Figure 4 (*Top*) Model 05305 Wind Monitor-AQ (air quality model), February 1986. (*Bottom*) Model 05701 Wind Monitor-RE (research model), July, 1987. See ftp site for color image.

7 SPECIALIZED MODELS

In early 1985 we received another development contract from the National Data Buoy Center to design a special model of the Wind Monitor with a tachometer generator for the wind speed transducer. This model required the use of slip rings. We calibrated the output of the tachometer generator so that it exactly matched the output of the Bendix/Belfort sensors, which were the operational sensors at many NDBC sites. With the addition of a special cable connector and special mounting adapter fabricated at the NDBC facility, this special model Wind Monitor could directly replace the other sensors in the field without any changes to the data collection package. When we began making deliveries in December, 1985, we designated this special-purpose Wind Monitor as Model 05203-1. It was only manufactured for NDBC plus a very few other customers for special applications. At a later date we developed a low-power circuit that converted the frequency output of the standard Wind Monitor into an analog voltage with the same calibration as the tachometer generator version. The combination of the special circuit box plus a standard Wind Monitor achieved the same result and was more reliable than the special Model 05203-1.

8 DESIGN IMPROVEMENTS

The magnet and coil wind speed transducer had proved to be very reliable; however, the signal level was very low when rotating slowly near threshold speeds providing a marginal signal-to-noise ratio. In 1986 we changed the wire size in the coil from 26AWG to 4OAWG, allowing more than twice as many turns and a corresponding increase in signal level. We were concerned about the durability of the much finer wire, but field testing indicated that it was quite satisfactory. We invested in a special coil winding machine that was able to properly control the tension and layering of the finer wire on the special injection-molded bobbins. Also in 1986 we changed the wind direction transducer to a custom-designed very high quality conductive plastic potentiometer, custom manufactured to our own specifications. Previously we had used potentiometers made by several different well-known manufacturers. These were all standard-type conductive plastic potentiometers with modifications to meet our requirements. While they all worked fairly well, the failure rate seemed to be unacceptably high with many warranty replacements. The new custom potentiometer, while much higher in initial cost, proved to be far more reliable and therefore less costly due to reduced repair costs and greatly improved customer satisfaction.

A fairly large number of Wind Monitors were deployed by Pacific Gas and Electric Company (PG&E) in a study of the dynamic thermal rating of power transmission lines. In 1986 PG&E reported a problem with the wind speed signal from an instrument located near a substation. The electromagnetic pickup from the 60-Hz power lines was causing a wind speed reading of about 6 mph when the actual wind speed was at zero. This problem was solved by adding an RC low-pass filter circuit that attenuated the 60-Hz signal below signal conditioning trigger levels. The circuit was designed to pass the normal ac wind speed signal at the very low frequencies near threshold. It also passes the 60-Hz real wind speed signal, which has a higher amplitude than the induced power line signal. This circuit has been incorporated on all subsequent models.

In July, 1987, we introduced a third standard model of the Wind Monitor, which was designed for maximum sensitivity for research and special applications requiring measurement of very low wind speeds. The tail assembly is similar in construction to the air quality model but with a much lower foam density, resulting in an optimum damping ratio of 0.65. The very sensitive helicoid propeller is made from the same expanded polystyrene foam as the tail, resulting in a very low threshold and a short distance constant. This sensor, designated Model 05701 Wind Monitor-RE (research model), is sold only in limited quantities (Fig. 4).

At the same time the different versions of the Wind Monitor (as well as other products) were being developed, we were gaining valuable experience in injection molding. We had acquired new CNC (Computer Numerical Control) machinery for mold making and also new and larger injection-molding machines with better controls and greater capability. In mid-1987 we completed a fairly sophisticated mold to injection mold the entire tail assembly of the standard Wind Monitor. This resulted in a much more uniform and durable tail assembly, which we could produce at a significantly reduced cost.

By now many Wind Monitors, especially the Wind Monitor-AQ, were being used in critical applications that required wind tunnel certification and field auditing. From the first production runs all units were serial numbered; however, since the propellers were interchangeable between sensors, in March, 1988, we began to serial number all three types of production propellers. Previously, serial numbers were added only to propellers that had been individually tested in our wind tunnel.

9 JUNIOR MODEL

In April, 1988, the National Data Buoy Center asked us to design a new reduced size model of the Wind Monitor. Its buoys were becoming progressively smaller and it was felt that a smaller sensor would be required in the future. All NDBC buoys use duplicate wind sensors for redundancy. A smaller sensor would allow the two sensors to be mounted far enough apart to avoid mutual interference while keeping the sensors close enough to avoid damage during buoy deployment. The new smaller Wind Monitor turned out to be about two-thirds the size of the standard model. Prototypes were delivered in October, 1989, for field testing; however, this model has not yet been deployed operationally by NDBC. This fourth standard model, which was designated Model 04101 Wind Monitor-JR (junior model) (Fig. 5), was not advertised until November, 1993. Eventually it proved useful for several special applications with a number of different customers and several hundred have been produced to date.

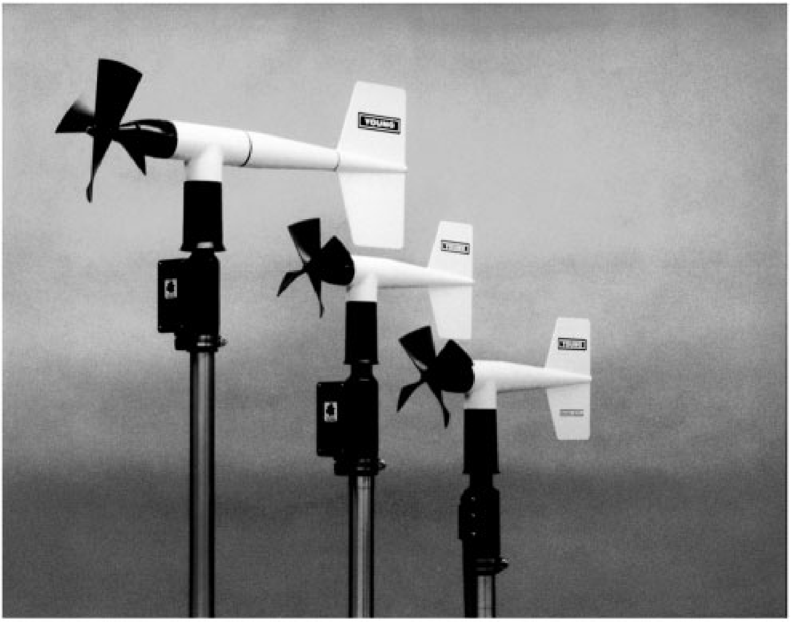


Figure 5 (*Top*) Model 05103 Wind Monitor (standard model, shown for relative size). (*Middle*) Model 04101 Wind Monitor-JR (junior model), November, 1993. (*Bottom*) Model 04106 Wind Monitor-JR/MA (junior marine model), October, 1994. See ftp site for color image.

10 CONTINUING DESIGN MODIFICATIONS

Until the spring of 1988 all Model 05103 Wind Monitors used ball bearings that had light-contacting plastic seals and were lubricated with instrument oil. The Model 05305 Wind Monitor-AQ and Model 05701 Wind Monitor-RE use the same size ball bearings but with noncontacting metal shields and instrument oil lubrication. This combination provides lower threshold as required by agency specifications, but the bearings are more easily contaminated and require routine maintenance. To achieve a longer field life, the bearing lubrication in Model 05103 was changed from oil to a special, wide-temperature-range grease. This resulted in a slight sacrifice in threshold but provided a dramatic increase in service life, which was extremely valuable to those customers with remote installations.

We had tested these sensors in our own wind tunnel and in the NDBC wind tunnel and had established a gust survival specification of 80 m/s. We had also done several sustained high-speed tests with the sensor mounted on a vehicle on an extended highway trip. We were beginning to get inquiries regarding survival at even higher wind speeds, especially for monitoring hurricane- and typhoon-prone areas. In August, 1988, we were able to make arrangements to test a Model 05103 Wind Monitor in the NASA/AMES Research Center wind tunnel at Moffett Field, California. In this test the sensor was installed in their 7 × 10 ft Wind Tunnel No. 1 and subjected to wind speeds that were stepped from 0 to 250 mph (110 m/s) and back to 0 (in 12 steps). The wind tunnel was held for 2 min at each speed. The Wind Monitor performed well throughout the entire speed range. With these data we increased our gust survival specification to 100 m/s, and we began to recommend the sensor for applications that required higher wind exposure.

Woods Hole Oceanographic Institution (WHOI) had tested and deployed many Wind Monitors beginning with the early models. Beginning in March, 1989, WHOI began to incorporate the Model 05103 Wind Monitor into its IMET (Improved Meteorological Measurements from Buoys and Ships) "intelligent" sensor program. To satisfy its requirements, we had to adapt the upper portion of the sensor to a specially designed base assembly that WHOI supplied to us. The potentiometer, normally supplied for the wind direction transducer, was omitted and an elongated coupling was provided for connection to an optical encoder located in the special base assembly along with the signal conditioning and data logger circuits. Several of these units have been built and successfully field tested on buoys and research vessels by WHOI in an ongoing program. This same type of sensor package is now being proposed for the new NOAA/Voluntary Observing Ship (VOS) program, which may eventually include 300 to 400 vessels.

During the next several months additional design changes were made to improve performance and to overcome some reported field problems. The junction box, which contained the terminal strip for cable connections, was enlarged to accommodate a printed circuit board with transient protection devices. The circuit board also mounted an improved device for cable connections. There had been a number of potentiometer failures in the field that occurred during thunderstorm activity and also similar failures associated with extremely dry and windy sites such as the Antarctic. After extensive testing in our engineering department, we discovered that the terminations of the conductive plastic element in the potentiometer were susceptible to failure when a static charge would build up on the housing. Two changes were made to combat this problem. A ground wire was added to the housing of the potentiometer and connected directly to the earth ground terminal in the junction box. In addition the molded plastic parts surrounding both the coil and the potentiometer were changed to a conductive-type plastic. The mounting post was also changed to conductive plastic. With these changes and a properly grounded installation the static discharge failures were almost completely eliminated.

11 TESTING & CERTIFICATION

The World Meteorological Organization (WMO) had scheduled an intercomparison study of wind instruments to take place between July, 1992, and October, 1993. The Swiss Meteorological Institute (SMI) was already familiar with the Wind Monitor and recommended that we include it in the study. We submitted an application through the U.S. National Weather Service (NWS) to include the Model 05103 Wind Monitor, which was accepted. The WMO Wind Instrument Intercomparison was conducted jointly by METEO-FRANCE and the Swiss Meteorological Institute. The test site was the Mont Aigoual Observatory in the Cevennes Mountains in southern France. The observatory is at an altitude of 1567 m and normally experiences strong winds and heavy rime ice conditions during winter. The Wind Monitor was installed and operated satisfactorily without maintenance during the entire study, including winter, except during several periods of heavy rime ice. During the ice events the sensor became frozen and ceased to function; however, upon thawing it recovered and again functioned within specifications. The final report was released by METEO-FRANCE in October, 1997. We were surprised and pleased to learn that, after careful data analysis, the Wind Monitor had been selected as the reference sensor for the study. As the results of the intercomparison became known, there was a marked increase in acceptance of the Wind Monitor in Europe.

Late in 1992, to update our wind tunnel transfer standards, we arranged to have a Model 05103 Wind Monitor (with a four-blade polypropylene propeller) and a Model 05305 Wind Monitor-AQ (with a four-blade carbon fiber thermoplastic propeller) tested at the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland. Both models were run in the NIST 1.5 × 2.1 m (5 × 7 ft) Dual Test Section Wind Tunnel (DTSWT) from 0 to 46 m/s (103 mph) as well as the NIST Low Velocity Facility (LVF) from 0 to 10 m/s (22 mph). Based upon the data from these tests we found that the calibration of the four-blade polypropylene propeller differed from our published calibration by approximately 1%. The calibration of this propeller can be shifted slightly during manufacture by altering injection-molding temperature and pressure. Therefore we were able to correct the slight discrepancy in calibration. Unfortunately, the four-blade carbon fiber thermoplastic propeller used on the Model 05305 Wind Monitor-AQ exhibited a calibration discrepancy of 4.5%. The material used in the manufacture of this propeller does not allow any adjustment during the molding process. It was therefore necessary to publish a new calibration for this model and advise our customers of the revision. Fortunately many customers were able to apply a correction factor to their data to improve its quality.

12 MARINE MODEL

The number of Wind Monitors being used in marine applications was steadily increasing and in April, 1993, we introduced a standard model that incorporated many of the special modifications that we had done for the National Data Buoy Center. The sealed ball bearings are lubricated with a special waterproof grease. The junction box, normally attached to the mounting post, is omitted and a 3-m (10 ft) heavy-duty cable is supplied with a waterproof cable gland. The surge suppression components are contained in a potted module with the cable connections inside the mounting post. This fifth standard sensor is designated Model 05106 Wind Monitor-MA (marine model) (Fig. 6). The following year we introduced a marine version of the Wind Monitor-JR that incorporated the same special modifications. The Model

04106 Wind Monitor-JR/MA is the sixth standard model in the Wind Monitor line (Fig. 5).

Late in 1995 we had to do some extensive testing of all products, including all models of the Wind Monitor and associated signal conditioning and display modules, to determine conformance with the European requirements regarding electromagnetic emissions and susceptibility to electromagnetic interference known as the "EMC Directive" (Electro-Magnetic-Compatability). Most products met the requirements without modification, but a few circuits required some modification, and in most installations the use of shielded cable is required. After it was determined that the sensors met these requirements, we were required to affix a "CE" label (from the French 'Conformité Européan') to the sensor and also add a "Declaration of Conformity" to the instruction manual for each shipment destined for Europe.

13 SERIAL OUTPUT

After many months of development yet another standard model of the Wind Monitor was added to the product line. This new model incorporates an absolute optical encoder in place of the potentiometer for the wind direction transducer. Because of the very limited space available within the body of the Wind Monitor, it was necessary to design and fabricate a custom encoder that utilizes a unique scanning technique to determine azimuth angle to better than one degree. The operation of the encoder and the signal conditioning are controlled by a microprocessor in the junction box. There are three selectable output protocols, which are changed by moving a small jumper on the microprocessor circuit card. One output is specially developed by the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, for its portable weather stations. A second output is a standard used in the marine industry developed by the National Marine Electronics Association (NMEA). The third output is our own, which we refer to as the "RMY" (R. M. Young) protocol. The outputs are a serial data stream, which can either be polled or continuous. In addition there are calibrated analog voltage outputs. This seventh standard model was introduced in December, 1995, as Model 09101 Wind Monitor-SE (serial output) (Fig. 6).

Icing continues to be a major limitation on Wind Monitor operation. The Wind Monitor is currently being deployed in many remote areas where there are occasional heavy rime ice conditions and limited power for any effective form of heating. We believe that the most promising potential solution is some form of passive surface modification. We have been investigating a couple of different surface treatments that we are hopeful will reduce the ability of clear and rime ice to adhere to the propeller and other outer surfaces of the sensor. Several researchers have reported varying degrees of success with a spray-on material called Vellox, which was developed specifically to reduce ice adhesion. This material seems to work quite well for a period of time following its application, but it rubs off easily and requires reapplication at regular intervals to be fully effective. Another form of permanent hydropho-

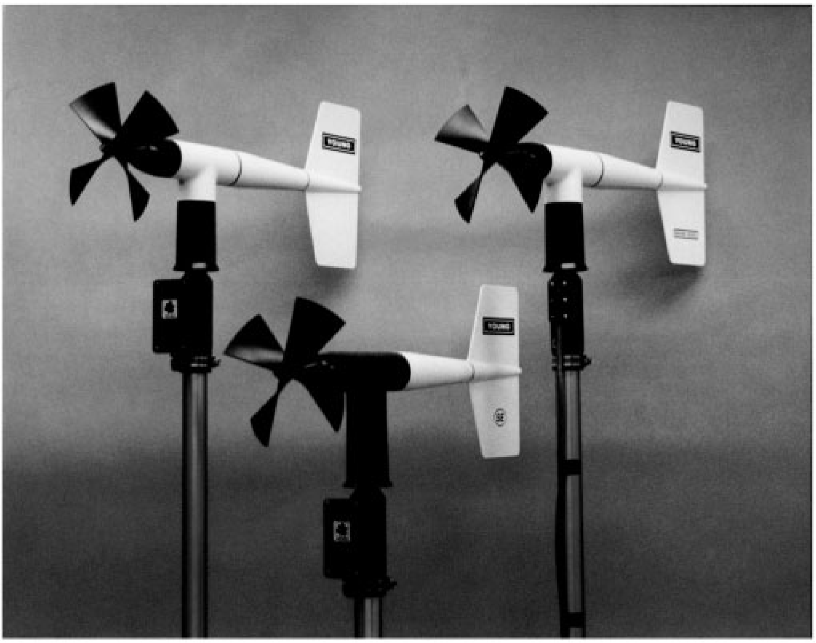


Figure 6 (*Top left*) Model 05103 Wind Monitor (standard model), August, 1984. (*Bottom middle*) Model 09101 Wind Monitor-SE (serial output), December, 1995. (*Top right*) Model 05106 Wind Monitor-MA (marine model), April, 1993. See ftp site for color image.

bic surface treatment is currently being investigated but results have so far been inconclusive and further testing is still needed. We are also investigating a new epoxy paint that is specially formulated to prohibit ice adhesion. If we are able to find a surface treatment that is practical and effective, it may result in the introduction of yet another model of the Wind Monitor.

The continually evolving needs of customers have dictated many changes as well as the ongoing development of a series of similar but quite different instruments. By this process products are continuously being improved and adapted to new demands.

BIBLIOGRAPHY

Anemometer Comparison Project (1986). A Report to the Atmospheric Environment Service of Canada, RVI/50.09, Resource Ventures Incorporated, Charlottetown, Prince Edward Island C1A 3W2.

Crawford, K. C., F. V. Brock, R. L. Elliott, G. W. Cuperous, S. J. Stadler, H. L. Johnson, and C. A. Doswell, The Oklahoma Mesonetwork—A 2 Pt Century Project, Preprints, Eighth International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, Atlanta, American Meteorological Society, pp. 27–33.

Frietag, H. P., M. J. McPhaden, and A. J. Shepard (1989). Comparison of equitorial winds as measured by cup and propeller anemometers, *J. Atmos. Oceanic Technol.* **6**, 327–332.

Hayes, S. P., L. J. Magnum, J. Picaut, A. Sumi, and K. Takeuchi (1991). TOGA-TAO: A moored array for real-time measurements in the tropical Pacific Ocean, *Bull. Am. Meteor. Soc.*, **72**(3), 339–347.

Jones, D. W., D. B. Hatton, D. A. Jenkins, and A. P. Scott (1992). A Field Comparison of Some Wind Sensors, Report No. 49, WMO Technical Conference on Instruments and Methods of Observation (TECO-92), Vienna, pp. 347–350.

Michelena, E., and J. Holmes (1983). A Rugged, Sensitive, and Lightweight Anemometer Used by NDBC for Marine Meteorology, Preprints, Fifth Symposium on Meteorological Observations and Instrumentation, Toronto, American Meteorological Society.

Michelena, E., and J. Holmes, The Meteorological and Oceanographic Sensors Used by the National Data Buoy Center, Proceedings MDS 9186 Marine Data Systems International Symposium, New Orleans, Marine Technology Society, pp. 596–601.

Payne, R. E. (1988). The MR, a meteorological sensing, recording and telemetering package for use on moored buoys, *J. Atmos. Oceanic Technol.* **5**, 286–297.

Phillips, C., D. Bumham, L. Jacobs, and D. Hazen (1992). 1991 LLWAS Anemometer Test Program, Final Report DOT/FAA/W/92-1, DOT-VNTSC-FAA-92-6, Research and Special Programs Administration, John A. Volpe National Transportation Systems Center, Cambridge, MA 02142-1093, September. (Document available through N.T.I.S., Springfield, VA 22161).

Qualid, G., P. Gregoire, M. Gilet, B. Hoegger, and A. Heimo (1993). WMO Wind Instrument Intercomparison, Preprints, Eighth Symposium on Meteorological Observations and Instrumentation, Anaheim, American Meteorological Society, pp. 274–278.

Weller, R. A., and D. S. Hosom (1989). Improved Meteorological Measurements from Buoys and Ships for the World Ocean Circulation Experiment, Proceedings Oceans 1989, Seattle, IEEE, pp. 1410–1415.

Weller, R. A., D. L. Rudnick, R. E. Payne, J. P. Dean, N. J. Pennington, and R. P. Trask (1990). Measuring near-surface meteorology over the ocean from an array of surface moorings in the subtropical convergence zone, *J. Atmos. Oceanic Technol.* **7**, 85–103.

Weller, R. A., M. A. Donelan, M. G. Briscoe, and N. E. Huang (1991). Riding the crest: A tale of two wave experiments, *Bull. Am. Meteor. Soc.* **72**(2), 163–183.

WMO (1997). Wind Instrument Intercomparison 1992–1993, Meteo-France/Swiss Meteorological Institute, World Meteorological Organization, Commission for Instruments and Methods of Observation, Final Report, October.

CHAPTER 39

COMMERCIAL RESPONSE TO MEASUREMENT SYSTEM DESIGN

ALAN L. HINCKLEY

1 INTRODUCTION

The commercial opportunity to produce multiple measurement systems began when measurement instrumentation was first used to document observations. While the following contains examples specific to Campbell Scientific's experience in measurement system development, the information presented reflects the direction taken by the measurement industry as a whole.

As a commercial producer of more than 10 different measurement systems, a great deal of practical experience has been gained in their design and in the representativeness and uncertainty of their measurements. Over 75,000 systems have been sold since the first battery-powered datalogger, the CR5, was shipped in August, 1975. Although this commercial success has given growth and economic stability, the greatest gain has been what we and our customers have learned about improving unattended environmental measurements.

Experience has taught us that measurement system design is optimized when it is based on theory and observation. The degree of confidence one can have in a system design is the degree to which the theory and observation agree. A measurement system may be designed and operational, but without some theoretical basis sources of uncertainty cannot be investigated (Campbell, 1991).

New designs should include a method by which the machine can detect and respond to its most probable failures—by observing and recording the errors and failures of current systems, better ones may be produced. Experience with similar designs that have been produced and supported in quantity is a most valuable asset in the design of new systems. Two principles that lead to good system design are (1)

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

making sure that the needs of the customer are met and (2) recognizing that there is an optimum high level of quality at which the total costs of designing, producing, selling, and servicing a product are minimized.

From manually measured mercury thermometers and staff gages, to chart recorders with spring-powered clocks, to microprocessor-based dataloggers, the tools for making good measurements have evolved. These new tools are more accurate, operate longer in harsher environments, and provide many more data retrieval options.

While the modern measurement system eliminates many of the uncertainties associated with manual measurements, others remain and new ones have been introduced. Using the measurement of temperature as an example, the modern systems have solved, within budget limits, the problem of insufficient human resources to measure the temperature in remote areas over extended periods of time. Most errors associated with manual data transcription and data transcription delays have been eliminated. Remaining are siting and exposure errors associated with the new temperature sensors and radiation shields. Inaccuracies in the liquid-in-glass thermometers are replaced by inaccuracies in the electronic temperature sensor (thermistor, resistance temperature detector (RTD), or thermocouple). The eye-sight-induced parallax error associated with manual measurement of a liquid-in-glass thermometer has been replaced by the datalogger's measurement error. Introduced are the effects of temperature changes on the measurement system's performance. While they do occur in the human observer, memory failure, run-down power supplies, and lightning damage are generally considered new uncertainties. Also introduced but often unresolved are a new set of "phantom" events like the rain caused by geckos (a tropical lizard) playing "teeter-totter" on the tipping bucket of a rain gage or the instantaneous half-meter of snow caused by an elk that lies down under the acoustic snow depth sensor.

The rest of this chapter will discuss some important features designed into the modern measurement system that improve its performance and reduce the uncertainties associated with unattended measurements. Where possible, the uncertainties will be defined and quantified. This information will facilitate the selection of good equipment and provided needed metadata (characterizing information) about the new measurement systems. Where they are helpful, anecdotes illustrate problems that have been overcome and failures that have led to greater understanding.

2 MEASUREMENT SYSTEM

Today's microprocessor-based measurement systems range from dedicated units with fixed sensor configurations, measurement rates, and reporting intervals to programmable systems with flexible measurement electronics. Dedicated systems have advantages in certain applications but, in general, the ability to accommodate different sensor types without additional signal conditioning as well as the ability to perform on-site calculations offers advantages for all but the most routine applications. Only electronic digital systems that are capable of operation without alternating current (ac) power are considered in this chapter.

All digital measurement systems must include (1) the measurement electronics to convert sensor signals to digital values, (2) either electronic storage media to collect the data on-site or telecommunications hardware to transmit the digital values, or both, (3) sensors that provide an electronic signal, and (4) mounting hardware, protective enclosures for the electronics, and power supplies (Tanner, 1990). In many of today's systems, a stand-alone environmental datalogger performs the measurement, processing, and on-site data storage functions as shown in Figure 1.

The datalogger, the cornerstone of a data acquisition system, has specific design requirements determined by the application. Our first fully digital datalogger, the CR21, was designed to meet the weather station needs of the agricultural and environmental studies disciplines (Schimmelpfennig et al., 1979). The datalogger design features required by these applications include:

- *Sensor Compatibility* Direct connection and measurement of sensor signals without external signal conditioning circuitry reduces cost, complexity, measurement error, and power requirements. The ability to resolve signals to the required precision affects the choice of sensor.
- *On-Site Processing* Field processing reduces data storage requirements, scales sensor signals to engineering units, and provides logic decisions for control applications.
- *Field Observation of Measurements* Field verification of sensors requires the ability to continuously observe instantaneous measurements, in engineering units, from a display or printer.
- *Input Transient Protection* Environmental datalogging is vulnerable to major hardware damage caused by large lightning-induced transients entering the system on sensor leads. Protection hardware such as transorbs and spark gaps, and proper grounding procedures are required to minimize damage.
- *Hardware Microprocessor Reset* Unattended, processor-based instrumentation should reset the processor, restoring normal execution in the event it is altered due to input transients or intermittent component failure. User-entered programs should exist in write-protected memory, minimizing the possibility that the processor can overwrite the program should an abnormal execution state occur.

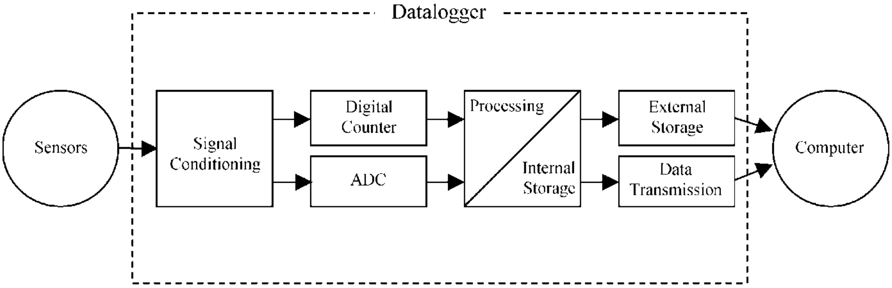


Figure 1 Generalized data acquisition sequence.

- *Low Power Consumption* Operation from direct current (dc) power supplies (batteries) and low current drain are required. An average power consumption of 50 mW (12 V at 4 mA) allows 2.5 months of operation on eight, alkaline D cells (7.5 Ah).
- *Operation in Adverse Environmental Conditions* Operation at high temperature and humidity are the main concerns in agricultural applications (at least until snow algae becomes the next food craze), but environmental dataloggers must operate and maintain measurement specifications over a minimum range of −20°C to +55°C. Solar heating can raise datalogger enclosure temperatures 20°C above air temperatures. Tight enclosures and desiccant provides the simplest and most cost-effective means of preventing water from condensing on the electronics.
- *Remote Communication Capability* Telephone was the only remote communication method available for the CR21. Since that time, data telemetry has become increasingly important. Models requiring climatic data are run daily to predict water use, crop development, disease and pest growth, etc. Synoptic models utilizing 15-min data are run several times a day. These applications require the timely transfer of data via radio, telephone, or satellite.
- *Logic and Control* The ability to compare values or time, with programmable limits and make decisions, provides powerful control capability. Samplers can be switched on or off, sensors with high current drain can be powered only during the measurement.

Good system design requires that the sensors be properly sited and mounted and that the electronics be protected and grounded. The accuracy and response times of the sensor needs to match those of the datalogger and the needs of the investigator. The data exchange between the measurement electronics and the data retrieval/storage hardware needs to be easy, reliable, and accurate.

Sales of our seventh measurement system, the Basic Data Recorder (BDR), did not meet expectations. Only 2000 BDRs were sold compared to sales of 35,000 of the datalogger built just before it. The BDR met the specific design requirements spelled out in a government bid, but it lacked the measurement versatility, telecommunication options, and software support required by our “typical” customer. Had the datalogger been built to fill both requirements, it would have been more successful. In spite of the failure, valuable lessons were learned. A new “table-based” data storage method was developed. Tighter government radio frequency (RF) noise requirements led to improved RF testing, shielding, and eventually to European market “CE Compliance.” Both have been invaluable in the development of today’s dataloggers.

3 DATALOGGER

A datalogger is made of pieces and processes that must fit and function together to make accurate measurements on low power under extreme environmental conditions. Seemingly simple design details like the style of a connector or the circuit board cleaning process are critical to the long-term performance of the unit. For example, the switch from a rosin-core solder flux to a water-soluble flux midway in the life of the first datalogger flooded the service department with repairs, threatening the existence of the new company. The dataloggers were failing because the flux remaining on the boards would adsorb water under humid conditions, become conductive, and cause shorts. Water-soluble fluxes were not as good as they are today and our procedure for completely cleaning it from the circuit boards had not yet been perfected.

Most modern datalogger uncertainties can be separated into two categories, failures or routine inaccuracies. Most failures result in the loss of part or all of the digital data. Only a few compromise the quality of the recorded data. Routine inaccuracies are the day-to-day errors in the datalogger's measurements that affect the quality of the recorded digital data.

Measurement system failures can be caused by:

- Failure of one of the components or circuit boards within the datalogger.
- Environmental conditions such as lightning, humidity, heat/cold, dust, floods, fire, wind, hail, ice, ultraviolet (UV) degradation, humans, animals, birds, insects, mold, etc.
- Human error in the wiring or programming of the datalogger.
- Failure of one of the power system components (battery, regulator, solar panel, etc.). Sometimes a datalogger component fails in a mode where it draws excessive current running down the battery prematurely.
- Failure of a sensor or sensor cable.
- Failure of the data storage/telemetry system.
- Failure of the protective enclosure or mounting hardware.

Inaccuracies of the measurement system that are routine in nature are caused by:

- Inaccuracies inherent in the voltage measurement, pulse measurement, or clock circuits over the temperature range of the datalogger. These uncertainties are quantified by the accuracy specification given by the manufacturer and should include the temperature range over which it is valid.
- Inaccuracies in a sensor's measured parameter induced by another environmental variable (e.g., temperature-induced inaccuracies in a relative humidity measurement).
- Changes over time of components in the voltage measurement, pulse measurement, and clock circuits–recalibration issue.
- Changes over time in the sensor's response to the parameter being measured—recalibration issue.

Fortunately, well-designed dataloggers rarely fail and their routine inaccuracies are relatively small and measurable. Many of the above-mentioned failures and inaccuracies are prevented or minimized by good datalogger design and manufacturing, system maintenance, and user training. Some are minimized by paying more for units with greater accuracy or reliability. Other failures, polar bears, for example, are simply Murphy's law at work. While it is not comprehensive, the following discussion details some of the more important features designed into a good environmental datalogger.

Low-Power Microprocessors

The microprocessor and quartz crystal clock are the brain and the heart of the datalogger. The processor does the computations and controls the operational sequences while the clock controls the timing and to some extent the speed. As evidenced by the cooling fans built on top of the Pentium processors in today's personal computers (PCs), most processors use a large amount of power. In a datalogger, which needs to run for 6 months on the power in eight alkaline D cells, it is important to use processors that require very little power.

In the early 1970s, a new logic circuit technology called CMOS (complementary metal-oxide semiconductor) was introduced. The first CMOS chips released were simple AND, OR, and NOR gates. About 2 years later, RCA introduced the 1802, the first CMOS-based microprocessor. The power consumed by a CMOS processor is low while changing states (processing), but it drops to less than 1 mA when it is inactive or "quiescent." Low-power system design requires that power consumption be minimized by placing CMOS or equivalent processors in the quiescent mode in between events.

The opportunities presented by this new microprocessor technology were recognized with great excitement by Eric Campbell, one of the founders of the company. His enthusiasm was such that he and his brother Evan built their younger brother Wayne a Christmas present, a CMOS-based "alarm clock" with speakers and light-emitting diodes (LEDs) that went off at 6 a.m. Christmas morning. The CMOS logic chips were first utilized in the CA9 path averaging laser anemometer and in the first datalogger, the CR5. Shortly after its release, the 1802 was added to the CR5 datalogger making it the first, or one of the first, microprocessor-based dataloggers. Three subsequent dataloggers—the CR21, CR7, and 21X—also utilized the low-power processing of the 1802. Quick recognition of the benefits of the 1802 processor placed the company right in the middle of an adventurous revolution in measurement system design.

Customer requirements and technology are constantly changing. Requirements for increased measurement and computational speed led to the use of two different versions of Hitachi's 6303 CMOS microprocessor in the next four dataloggers. The increased capability of this processor enabled subsecond measurement rates on a small number of channels or one-second rates on all channels. This speed is sufficient to meet or exceed most measurement needs in the weather, climate, and water fields. New microprocessors are introduced into new or upgraded dataloggers to meet changing customer requirements or because evolving technology has made the older parts unavailable. However, each microprocessor has its own quirks. Testing on the 6303 processor showed that its internal “overrun timer” was not reliable, so other methods had to be developed to work around the bug. Good system design requires extensive testing of a new processor to become familiar with its strengths and weaknesses.

“Watch Dogs”

Microprocessors specifically and digital circuits in general are quite susceptible to static electricity. Quite often, however, the static will “bomb” the processor or latch up a chip or change an address without causing permanent damage. Because lightning and other forms of static electricity are a very real part of the datalogger’s environment, good system design requires that a “watch-dog” circuit be added to the system to reboot it when it crashes. A watch dog is a hardware count-down timer circuit that reboots the processor when the processor bombs for more than a preset time. It is like an automated “Ctrl–Alt–Del.” The watch dog may also include a set of software checks that reboot the system if its system parameters are not within set limits. Once a watch-dog reset occurs, the processor reboots and special software is executed that attempts to restore the system to proper operation without the loss of data. The watch dog also assists in the development of datalogger software because it detects when a software bug causes the system to crash. The watch dog is a critical design requirement for all environmental dataloggers.

Clock

The clock needs to be inexpensive and low power but accurate enough that it can be left unattended for up to a year in places like Antarctica without a significant shift in time. The datalogger’s clock is a quartz crystal that ideally oscillates at a constant frequency. In low-power less expensive crystals the frequency varies slightly as a function of temperature. The raw quartz crystal’s accuracy of ±4 min per month is improved in the datalogger to ±1 min per month by measuring the clock’s temperature and correcting the clock based on an algorithm permanently stored in the datalogger. While satellite applications require even greater clock accuracy, clocks that stay accurate to within 1 min per month meet most requirements.

Self-Contained Measurement Circuitry

Most of the sensors used in the fields of weather, climate, and water output a voltage or change resistance in response to the parameter they are sensing. The rest either output pulses or digital data via a serial protocol. Well-designed environmental dataloggers measure multiple voltage, pulse, and serial signals without the need for additional external signal conditioning, thus reducing the cost, complexity, errors, and power requirements associated with the extra circuitry.

Voltage Measurements

While environmental dataloggers can have analog inputs capable of measuring either voltage or current signals directly, Campbell Scientific dataloggers have always measured voltages and use an external shunt resistor to measure current signals. Voltage measurements are made by switching the signal from one of the analog channels into the analog-to-digital (A/D) circuitry. After waiting for the signal to stabilize at its correct value, the signal voltage is allowed to charge up a capacitor for a fixed amount of "integration" time. The voltage on the capacitor is then held for the A/D conversion. The A/D conversion is made using a successive approximation technique. For example, to measure a signal of +1000 mV on a ± 2730-mV range the microprocessor first compares the signal to 0.0 mV to determine the sign (+ or −) of the signal. Since +1000 mV is greater than 0.0 mV, a value would be stored indicating a positive signal and the 12-bit digital-to-analog converter (DAC) would then output +1365 mV, half of the remaining range. The signal is now less than the DAC output, so a 0 would be stored for the first bit and the DAC would next output 682.5 mV. The process repeats until all the bits have been set, each time closing in on the sensor's voltage by half of the remaining range.

A datalogger typically has one DAC that outputs voltages across the largest full-scale range (FSR). For a datalogger with a 12-bit DAC and a ±2500-mV FSR to measure a thermocouple with a 2.5-mV full-scale output, it must measure the small signal by switching in a gain circuit that electronically multiplies the voltage by a factor of 1000. Thus a 2.5-mV thermocouple signal is really measured on the same 2500-mV range with the same 12-bit resolution DAC as the largest full-scale voltage. Having multiple FSRs with voltage multipliers preserves the resolution and accuracy of the measurements. Good system design requires that the datalogger have multiple FSRs that have been carefully chosen to cover the voltages produced by the sensors it is to measure.

The resolution of a voltage measurement is the smallest incremental change detectable in the signal and therefore represents the fineness of the measurement. The greater the number of bits for a given full-scale voltage range, the finer the resolution of the measurement. An uncertainty of 1 bit must be assumed in the measurement. Continuing the previous voltage measurement example out to 12 bits would yield a signal of +1000.089 with a resolution of ±0.666 mV. A 13-bit and a 16-bit measurement would yield +999.925 ± 0.333 mV and +999.985 ± 0.042 mV, respectively. The number of bits a measurement is divided into is determined by the number of bits in the DAC plus a couple of tricks. A 12-bit DAC ($2^{12} - 1 = 4095$) is one that can change its output voltage in increments of 1/4095 of its full-scale output. An extra bit of resolution is gained by determining if the signal is positive or negative before the successive approximation starts. A second extra bit is gained when a measurement is the average of two integrations (e.g., a differential measurement includes two integrations). To convert the resolution of a measurement range in millivolts into engineering units, multiply the resolution of the appropriate FSR by the ratio of the engineering unit range to the signal range. For example, given a resolution of ±0.333 mV on the ±2500 mV range, and a −40 to +60°C temperature signal represented by a 0- to 1000-mV signal, the resolution would be

$$\frac{\pm 0.333\,\text{mV} \times [60 - (-40)^\circ\text{C}]}{(1000 - 0\,\text{mV})} = \pm 0.0333^\circ\text{C}$$

When the resolution approaches the magnitude of the input noise, the latter determines the measurement uncertainty. Input noise is statistical and is specified in terms of the root-mean-square (rms) value. Numerical averaging of N measurements reduces the input noise by a factor of $N^{-1/2}$. One method of reducing the effect of noise on the measurement of a signal is to integrate the signal over a longer time period. Integration time is the time over which the signal is being electronically averaged. The integration time can be long for greater averaging, which smoothes the noise-induced peaks and valleys, short if more frequent measurements are required, or it can be broken into two integrations, one half of an ac cycle apart to remove either 60- or 50-Hz noise. The use of shielded sensor leads with the shield connected to system ground helps reduce high-frequency external noise.

Accuracy is often expressed in terms of the FSR, e.g., 0.1% FSR, and should be stated for a specific temperature range because of the temperature dependency of most errors. The accuracy of voltage measurements depends upon the accuracy of an internal voltage reference diode and the self-calibration ability of the datalogger over time and temperature. Input offset voltages also affect accuracy, but a good microprocessor-based system corrects for this error by shorting the input to ground and measuring the input offset voltage as part of the signal measurement sequence. The offset error is then removed from the result numerically.

The accuracy of each datalogger is carefully measured over several hot/cold temperature cycles in automated precision environmental test chambers using regularly calibrated National Institute of Standards and Technology (NIST) traceable voltage sources. The dataloggers are tested at the environmental extremes quoted in the specifications and then tested at points 10°C beyond those extremes to ensure their reliability and accuracy. Even in the "startup" days of the company when money for test equipment was limited, environmental testing was so important that dataloggers were manually calibrated while still cold from "soaking" in a chest freezer with dry ice and then again after soaking in a "hot box" made from an old refrigerator and a heat lamp.

In the field, the accuracy of the internal voltage reference is maintained over time and temperature by an automated self-calibration procedure that utilizes a precision reference diode to calibrate the rest of the system. To avoid interfering with the routine measurements, the calibration is done in the background in small segments throughout the 3-min self-calibration cycle. The calibration results are then averaged across a 15-min period, which keeps the logger calibrated across most naturally occurring environmental temperature changes.

"Single-ended" and "differential" amplifiers are commonly used for voltage measurements. With single-ended measurements the signal is measured with respect to instrumentation ground. Input connections therefore have one active side and one ground connection. With differential measurements, one side of the signal is measured with respect to the other side, requiring that both inputs be active. Often these connections are labeled as “high” and “low” or + and −. In some designs either the high or low side of the differential channel can be used to make two single-ended measurements. Most sensors require that the voltage channels have high impedance (use very little current).

Some signals can be measured either single-ended or differentially. Others, such as resistive full bridges, must be made differentially because the signal is referenced to one side of the bridge and not to ground. Single-ended measurements are adequate for high-level voltages, thermistors, potentiometers and similar signals. Differential measurements provide better rejection of noise common to both sensor leads, and for this reason they should be used for sensors having low-level signals (e.g., thermocouples). In addition, the accuracy of a differential measurement is improved by internally reversing the inputs and making a second measurement. The second measurement with the inputs reversed cancels out thermal or noise-induced offset errors in the following manner: Suppose a small change in the datalogger’s temperature since the last self-calibration is making the high input read 1 mV high so with a 10-mV signal, the high input reads 11 mV, the low reads 0 mV, and the first differential voltage measured is 11 mV. When the inputs are reversed, the high input reads 1 mV, the low input reads 10 mV, so the second differential voltage measured is −9 mV. When the sign of the second is reversed and the two measurements are averaged, the 1-mV offset cancels yielding a differential voltage of 10 mV. Good system design requires careful attention to even the smallest measurement details.

Thermocouples

One of the common sensors used to measure temperature is the thermocouple. In addition to the need for an accurate measurement of a voltage signal usually less than 5 mV, thermocouple measurements require a reference junction temperature and polynomial equations to convert the voltage to temperature.

Precision Excitation

Resistive sensors, those whose resistance change in response to the parameter they are sensing, require an excitation voltage to output a voltage that corresponds to the changing parameter. The excitation voltage needs to be selectable so the output voltage fills one of the full-scale ranges. It also needs to have accuracy and resolution characteristics that match those of the input voltage range. Power consumption is reduced by shutting off the excitation after the sensors have been measured.

Most resistive sensors can be measured ratiometrically, which means that the measured voltage is divided by the excitation voltage. Because both the measurement of the signal voltage and the generation of the excitation voltage utilize the same voltage reference, dividing one by the other cancels out errors in the voltage reference. As a result, the accuracy of the measurement increases from 0.1% FSR across the −25 to +50°C range to 0.02% across the same temperature range.

Some of the sensors included in this group are thermistors, RTDs, potentiometers as used in wind vanes, weighing rain/snow gages, pressure transducers, and load cells.

Pulse Measurements

Several important sensors output low-level ac signals, square-wave pulses, or even just a switch closure. The datalogger needs a voltage supply and a bounce elimination circuit to measure switch closures. Amplifiers are required for the low-level ac signals. Dedicated counters that accumulate while the logger is sleeping in between measurement intervals keep the current drain low. The processor must be able to detect occasions when it is too busy to measure and reset the pulse accumulators at the designated time. The pulses from the excessively long interval must be discarded if the signal is a rate or accumulated if it is not. Vibrating wire transducers, commonly used for long-term monitoring of water pressure or deformation, require special circuitry and software to determine the frequency of their limited-duration signal.

Serial Inputs and Digital I/O Ports

In an effort to increase their accuracy or to even be able to make some of the more difficult measurements, several sensors are now microprocessor based. These sensors make the required measurement, digitize the value, and then transmit the value in digital form to a datalogger via a serial protocol such as SDI-12, RS232, RS485, etc. In the case of the SDI-12 protocol, the datalogger tells SDI-12-based sensors when to make a measurement and then requests the data after the wait time specified by the sensor. Some smart sensors simply transmit their data after the measurements are made on their programmed time interval. Most serial signals are transmitted and received using the digital input/output (I/O) port.

Digital ports are needed to input status signals or to output control signals. This capability allows the datalogger to measure, control, or alarm based on time or measured values.

Programming Versatility

A well-built datalogger is not much good if it cannot be easily told what needs to be done. A less versatile datalogger programming language works fine in markets where there is a lot of repetition. A more versatile programming language is needed in diverse markets, especially where research is the primary objective.

A powerful datalogger language consisting of canned measurement instructions and canned mathematical processing instructions take the work out of doing good complicated measurements. If the versatile programming language is backed up with good PC support software to facilitate the program development, routine programs become easy and complex programs become possible.

On-Site Data Processing

The advantage of processing measured values to obtain more efficient data storage has been mentioned. Data compression is particularly important in remote applications where site visitations are costly. Processed results such as averages, standard deviations, extremes, and values recorded conditionally all reduce data storage and handling logistics.

Linear calibration constants are entered into the datalogger to convert measurements into engineering units immediately. The datalogger displays the sensor signal in engineering units that enable the user to verify the correctness of the signal and its conversion. Field calibration of sensors is possible. User-entered polynomial coefficients are used to linearize nonlinear measurements. Linearization and the scaling of sensor outputs into engineering units make it possible to correct sensor readings on-site (e.g., correcting a piezometer reading for barometric pressure yields pore pressure). Nonlinear signals converted to engineering units can then be averaged, but in their nonlinear form they cannot. The ability to convert sensor signals into correct engineering units is extremely useful when verifying the performance of the sensors and the datalogger in the field.

The ability to compare values or time with programmable limits and make decisions provides useful control functions such as sampling at a faster rate, measuring a selected sensor, or initiating a radio or phone communication for an alarm message.

Internal Data Storage

The last 25 years of computer hardware innovation has greatly increased the amount of memory stored per chip. On-board data storage that used to be in increments of 2 kilobytes now comes in increments of 1 and 4 megabytes. To keep up with the rapid advances in memory storage technology, dataloggers are designed and built with the ability to add the next generation of memory chips as soon as they became available or cheap enough to justify their use.

The limited on-board memory of the past required one or more of the following: compression of the data into hourly or daily summaries, immediate data transfer to an external data storage device and even those had limited capability, or immediate data transfer to a computer via one of the telecommunication methods. Today's extensive memory often eliminates the need for on-site external data storage devices, allows for more frequent recording of the data, and permits data retrieval on less frequent intervals.

Today's on-board data storage is not only more extensive but it is also more reliable. Part of the data, the system program, and the clock are maintained by an internal lithium battery that takes over the instant the system voltage drops below the operational level of 9.6 V. The extended memory is stored in FLASH, a memory chip that only erases or records data when power is present.

Reliability

At first glance, reliability and low cost work against each other. For example, military-grade (mil-spec) parts have higher reliability than industrial-grade components but they cost much more. Only when the full costs of a datalogger, including warranty repair, support, and marketing (ever try to sell a lemon?) are understood, can one properly balance cost and reliability. For example, there are ways to significantly reduce the cost of a datalogger without sacrificing reliability. Industrial-grade parts can be tested upon arrival. Where possible, specific parts that have proven reliable in previous dataloggers can be designed into new dataloggers. Rigorous environmental testing of the completed units result in mil-spec or better reliability at a fraction of the cost.

Production and repair personnel keep careful records of the parts that fail, hoping to detect a bad batch of parts before many of the dataloggers ship. This method catches most of the bad parts but, unfortunately, there are those rare occasions when a part only fails after time or after exposure to humidity or temperature variations in the field. This happened to the very popular CR10 datalogger. In the spring of 1998, a sharp increase in capacitor failures in CR10s returned for warranty repair was noted. Examination of the carefully kept internal records finally linked the cap and datalogger failures to a change in the cap manufacturing process early in 1997. The cap manufacturing change caused an internal crack in the capacitor that eventually led to its failure. Although nearly 2500 dataloggers were shipped with the faulty parts, the carefully kept records helped resolve the problem. New capacitors from a new vendor were installed in units that began to ship in the fall of 1998. A recall was issued in the fall of 1999 when the rising failure rate indicated the seriousness of the problem.

Good system design requires good personnel in the test and repair groups. Feedback from these groups on the problems and parts that fail must make its way back to the system design group so that problems can be solved and better parts used in revisions of current dataloggers or in the design of new dataloggers. A monthly meeting of a quality control group made up of production, repair, testing, and engineering personnel helps maintain high quality. Reliability is a direct result of quality. High reliability requires that quality be an issue of continuous concern.

In the field, dataloggers are frequently exposed to voltage surges and RF noise. Dataloggers are best protected from lightning by providing it a low-resistance path to a *single* good earth ground. Datalogger lightning protection design has improved with each new datalogger. The most recent dataloggers are built with separate grounds for power while the rest of the circuit board is covered with a ground plane, which serves as the ground for analog signals and sensor shields. Up to 2 A of current can be run through the ground plane without inducing voltage gradients, which affect analog measurements. A good beefy connection from the ground plane to an earth ground gives lightning a good low-resistance path to travel. Spark gaps protect the analog input channels.

The datalogger both creates and receives RF noise. The amount of RF noise is greatly reduced by the metal package surrounding the electronics. Tighter RF standards and better test equipment have lead to much better RF shielding.

Inexpensive rugged packaging that offers portability, protection from humidity, dust, and temperature extremes helps increase system reliability. The ability to power the datalogger from a variety of power supplies facilitates its use in many environments.

In the competitive datalogger market, where a company's livelihood is earned by small sales to private and publicly funded researchers, most on limited budgets, success requires good technical support, word-of-mouth advertising, a reasonable price, and great reliability.

4 DATA RETRIEVAL

The reliability and convenience with which measurements are transferred from the field to the computer are important design goals in any measurement system. Most of the datalogger design requirements, especially low power and environmental ruggedness, apply to the data retrieval components. Whether the data are stored on-site and retrieved during site visits or retrieved remotely using various telecommunication options, the manufacturer should provide the necessary hardware and software tools to accomplish this task. The challenge has been to keep hardware and software current given the rapid changes in the fields of electronics, telecommunications, and personal computers. The following sections briefly describe most of the data retrieval options currently available. Examples of some specific benefits resulting from new or enhanced data retrieval methods are included.

On-Site Data Storage

An important part of almost all data retrieval systems is the storage of data on-site. On-site storage holds data between site visits or data transmissions. Should the telecommunication system fail, data may be retransmitted once the system is repaired. Printer and magnetic tape methods of on-site data storage have been replaced by data storage in memory either inside the datalogger or in an external memory module.

Our first fully digital datalogger, the CR21, stored data internally in random-access memory (RAM) for telecommunication and externally on either thermal printer paper or a magnetic cassette tape for data retrieval during a site visit. The cost of memory limited internal data storage to only 600 processed data values. The printer paper held 35,000 data values, but it had to be manually transcribed into the computer. Cassette tapes held 180,000 values on one side of a 60-min tape. A tape reader transferred the data to a computer. While the printer stuck in high humidity and the tape recorder failed below -5°C, they were the only cost-effective on-site data storage options available at the time.

In 1983 Sequoia National Park requested a solid-state data storage module so it could gather winter precipitation data at high elevations in the Sierra Nevada. The desire to meet this customer's need combined with the falling cost of RAM led to the development of a RAM module powered by a small lithium battery. Due to its portability and reliability in cold and hot conditions, the memory module and its successors have become the preferred method of external on-site data storage and site visit data retrieval. Memory module development led to convenient data storage under cold winter conditions.

The increased portability and ruggedness of laptop computers and personal digital assistants (PDAs) allows them to be used to retrieve internally or externally stored data. Expansion of the dataloggers internal memory has contributed to this trend by reducing the frequency of site visits.

Data Retrieval via Telecommunication

Data can be transferred from remote field stations to a computer through wire, radio waves, or a combination of both. Variations on these options have greatly increased in number and quality in recent years. The new methods have increased data transfer speed and reliability. Cost, reliability, operating distances, and data rates are important in determining the usefulness of a telecommunication system for a particular application. Remote data retrieval provides timely reporting, early detection of equipment malfunctions, and, for more isolated sites, may be the only practical means of data recovery. The expense of manual data collection, particularly in larger networks, often justifies automated techniques even when near-real-time reporting is not required. Telecommunication improvements have changed data transmission methods from teletype or voice into a fully automatic digital transfer from the datalogger in the field to the computer running the synoptic forecast model or generating the reports.

Where only one-way communication is supported, such as with satellites or certain RF-based designs, redundant transmissions are often used to obtain high data capture rates. Interrogated networks based on two-way communication allow error checking and retransmission of improperly received data. Two-way communication also allows the user to remotely change the datalogger's program, clock, and control ports. The following information details advantages and disadvantages of most currently available telecommunication options.

Dedicated Cable

Short-haul or multidrop modems operate up to distances of several kilometers, the expense of the cable and its installation being the more practical distance limitation. Short-haul modems require a PC COM port and an independent cable (two twisted pairs) for each remote site accessed by the base station. A local area network (LAN) links multiple remote sites to a single PC COM port by a single two-conductor cable (typically COAX). Individual stations within the network are accessed through addresses set in the LAN modems. Direct cable links are useful in more permanent operations and for combining several local sites into a network (e.g., on a watershed), which can be accessed remotely by a single telephone or RF link. The communication rate is dependent upon both the cable characteristics and its length. Model SRM-5A short-haul modem manufactured by RAD Data Communications, Inc. specifies operation at 1200 and 9600 bit per second up to distances of 10.5 and 8 km, respectively, for 19 AWG wire. A network of 10 of Campbell Scientific's LAN modems (Model MD9) operates at 9600-bit per second up to a distance of 2 km assuming the loss characteristic of the coaxial cable is 0.02 dB/m.

Switched Telephone

Where available, standard voice-grade telephone lines provide a simple, reliable method for retrieving data. A typical PC and modem calls and retrieves data from each remote site. System support software facilitates the scheduling and setup of the call parameters. The modem at the measurement site needs to run on minimal dc power and function under hot and cold field conditions. Communication rates of 300, 1200, 4800, and 9600 bits per second are standard. Standard phone company procedures for transient suppression should be followed at the remote site. Suppression devices included in the modem may be augmented by external suppression devices.

Campbell Scientific's (CSI's) first rugged, low-power 300-bit per second phone modem (DC103A) combined nicely with the CR21-based automatic weather station (AWS). These telephone-linked weather stations were quickly utilized to form AWS networks in New Mexico, Nebraska, and California. In the CIMIS project in the Sacramento Valley of California, a computer polled each of 40 sites daily, providing data for evapo-transpiration (ET) models. The introduction of telephone telemetry to the AWS benefited farmers with daily crop water usage information.

Cellular Telephone

The cellular phone systems established in many countries provide vast networks of cellular repeaters in most urban areas and along major highways in rural areas. A cellular phone provides the radio link from a remote site to the nearest cellular repeater where the message is converted to telephone signals and passed down a standard phone line to the PC. Stationary sites able to utilize a cellular yagi antenna can communicate across distances of 20 km if there is line-of-sight to the cellular repeater site. The cellular link eliminates the installation costs of a standard phone line. In addition, the monthly "air time" fee for a cellular site accessed once a day is typically less than the monthly service fee for a standard phone line. The current used when transmitting is typically 1.8 A. The stand-by current drain of 170 mA requires either a larger power supply or that the cellular phone be turned on only a few hours during the day. Data throughput is limited by the bandwidth of the frequency to 4800-bit per second. Though the first cell phone systems were analog, overlapping digital networks are rapidly being installed.

Internet

Where Internet-grade T1 phone lines are available, an Internet modem is used to access data from one datalogger or a radio-linked network of dataloggers.

Single-Frequency RF

Single-frequency ultrahigh frequency (UHF) or very high frequency (VHF) networks with individual stations accessed by unique addresses are in general less expensive and logistically simpler than older style, multifrequency networks. Designs based on low-power transceivers (2 to 5 W) are available. Operating distances are limited to line-of-sight, typically 25 to 35 km, but this range is extended where individual stations may be used as repeaters to access more distant stations. Access of an RF network through telephone lines further extends the data collection range. Average power consumption of approximately 80 mA dc (1 A when transmitting) is typical for many low-powered radios. In standard UHF and VHF applications, a Federal Communications Commission (FCC) license is required. The bandwidth is regulated, limiting the RF bit per second rate. Data "through-put" rates depend upon the communication protocol, the overhead times required to establish the link, and the length of the transmitted data block. For example, data through-put rates of around 30 data values per second are reasonable.

At many sites, the cost of installing and maintaining a phone line are prohibitive. To obtain data from remote mountain tops economically, a radio modem and VHF/UHF radios were added to meet the needs of customers such as the military who were doing tests across the 8300 km^2 White Sands Missile Range. Ski resorts doing avalanche control with remote, mountain-top data also benefited from the radio links.

In 1992 a faster polling scheme was developed for radios. The increased speed of the new method allows data to be retrieved every 5 min from a 35-station network monitoring meteorological conditions across the 1400-km^2 research complex at Idaho National Engineering and Environmental Laboratory (INEEL). Kennecott Utah Copper smelter used the same setup to obtain air quality data from pollution monitors at 20 ambient sites and 3 stack sites every 2 min.

The same technology was taken one step further in the Oklahoma Mesonet Project (Brock et al., 1995). In that project, 108 weather stations uniformly distributed across the 180,000 km^2 of the state measure air temperature, humidity, barometric pressure, wind speed and direction, rainfall, solar radiation, and soil temperatures. Each station is polled every 15 min for its 5-min data. From a typical site, the data is transmitted several kilometers by VHF radio to the nearest police station or sheriff's office. From there it is relayed by the Oklahoma Law Enforcement Telecommunications System (OLETS) on a 9600-bit per second synchronous dedicated line to the central computer at the Oklahoma Climatological Survey office in Norman, Oklahoma. The central site ingests the data, runs quality assurance tests, archives the data, and disseminates weather information and warnings in real time to a broad community of users, primarily through a computerized bulletin board system.

In 2001, St. John's water management district in south Florida installed an even faster radio telemetry network of 85 sites with multiple repeaters. The 85 sites are polled every 15 min. The data are used to monitor levels and control gates and pumps thus controlling surface water levels within an acceptable range. The benefit of increased telemetry speed is real-time monitoring and control.

Spread-Spectrum RF

Spread-spectrum radio modems look and act like a wireless RS232 port. Constrained by the FCC to less than 1 W effective radiated power, an FCC license is not required to operate them. Transmitting and receiving on multiple frequencies in one of several bands (902 to 928 MHz was the first), they work well in industrial areas where line-of-sight is not available but distances are less than 5 km. With line-of-sight and yagi antennas, distances up to 30 km are possible. Spread-spectrum radios utilize one of two methods to transmit their data: frequency hopping or direct sequencing. The frequency hopping method is preferred for stationary sites. It also allows the frequency to be shared by more sites. Direct sequencing is preferred for mobile sites. Current spread-spectrum radios require 100 mA in the receive mode and 600 mA in the transmit mode. A spread-spectrum radio under design at CSI should require less than 75 mA in the transmit mode.

Satellite

Satellite transmitters provide data transfer from very remote sites and networks with wide spatial coverage. Most satellite links only provide one-way data transfer. Data transfer via satellite is generally more costly due to the required infrastructure. At the end of 1998, approximately 1300 satellites orbited Earth. Estimates put their number at 2400 by the year 2008. Satellites are grouped into three main classes based on height and type of orbit. They are: geosynchronous or geostationary Earth orbiting satellites (GEOs), middle Earth orbit satellites (MEOs), and Low Earth orbiting satellites (LEOs). The height of their orbits also determines the maximum geographical area they can cover.

Geosynchronous or Geostationary Earth Orbiting Satellites GEOs, positioned 22,300 miles (36,000 km) above the equator, appear to be stationary in the sky as they turn synchronously with Earth (hence the name geosynchronous or geostationary). This allows a ground station antenna to point at one place in the sky to send and receive signals. GEOs are the farthest from Earth and cover the largest surface area. Three GEOs are needed to cover the entire planet below 70° latitude.

The Geostationary Operational Environmental Satellite (GOES) system was installed and is maintained by the U.S. government. GOES channels (frequency and time slice) are reserved for use by federal, state, and local U.S. governments and some foreign governments. The recently introduced high data rate (HDR) GOES data throughput is limited to:

- 300- or 1200-bit per second data rate
- 20-s transmission window
- Transmissions intervals of 1, 3, or 4 h

The highest elevation weather station in the Americas was installed at 6542 m on the summit of Nevado Sajama in Bolivia in 1996 (Hardy et al., 1998). Installed to better understand tropical ice core variations and global climatic change, the station utilizes a GOES transmitter for near-real-time data and memory modules for on-site storage and manual data retrieval. The remote, high-elevation location precludes unscheduled service or repair visits. Snow and ice buildup on the antenna reduced near-real-time data recovery to 80% during the wettest month, but 100% of the on-site data was recovered during the annual site visit. In addition to permitting the start of the data analysis, the near-real-time data indicated that the failure of the lower sensors midyear was due to their burial by an El Niño induced larger than expected snowfall.

Middle Earth Orbit Satellites MEOs, positioned between GEOs and LEOs, orbit from 1000 to 22,300 miles. Depending on the altitude, as many as 10 or 12 are needed to cover the entire planet. In addition to voice and data transmission, MEOs are often used by surface navigation systems such as the Global Positioning System (GPS).

Many environmental measurement applications benefit from being able to record GPS position data. In addition, the GOES system now requires that transmitters use GPS clock data to keep their transmissions accurately timed.

Low Earth Orbiting Satellites LEOs orbit at altitudes from 100 to 1000 miles. Approximately 46 to 66 LEOs are needed to cover the entire planet. A LEO orbiting satellite is above the local horizon for less than 20 min. Messages must be short or LEO systems must support satellite-to-satellite hand-off to maintain communications.

One of the LEO systems called ARGOS, utilize the National Oceanic and Atmospheric Administration's (NOAA's) polar orbiting satellites. These are frequently used either in oceanography because station position can be tracked, or at latitudes greater than 70° where GOES satellites cannot be seen. The ARGOS platform provides good data transfer at the poles because the satellites cross the poles 24 times per day, which is three times more often than at the equator. ARGOS data throughput is limited by:

- The number of times per day the satellites pass overhead
- One satellite pass overhead lasts from 10 to 14 min

- Redundant 32-byte messages are transmitted every 90 or 200 s to ensure data integrity
- 400-bit per second data rate

Satellite communication is changing rapidly. Several projects currently underway will be launched in the near future. The satellite business is not easy due to the numerous governments involved and the fight for frequencies. Many projects have merged together so that the merged company has a better chance of seeing its project to completion. The following is an incomplete list of websites that provide links to companies that offer data transmission services via one of the GEO, MEO, or LEO satellites:

GEO Systems	Website
GOES	http://NOAA.WFF.NASA.GOV
Inmarsat-C	http://217.204.152.210/index.cfm
Qualcomm	http://www.qualcomm.com

MEO Systems	Website
ICO	http://www.ico.com
ORBLINK	http://www.orbital.com

LEO Systems	Website
Iridium	http://www.iridium.com
Orbcomm	http://www.orbcomm.com
Globalstar	http://www.globalstar.com
Argos	http://www.argosinc.com

5 SUMMARY

Commercial measurement system design requires careful attention to all details. Only the most important details were presented here. The experience gained through the design of similar measurement systems that have been produced and supported in quantity is valuable in the design of a new system. It is important to quickly utilize the technological advances in electrical components, sensors, and telecommunication methods. There is an optimum high level of quality at which the total costs of designing, producing, selling, and servicing a product are minimized. Careful, high-quality records help maintain high-quality standards. Low turnover of key personnel is important to keep expertise and continuity. Finally, the personal joy and satisfaction that comes from service to customers with legitimate needs is the basis for our existence as an organization (Campbell, 1987).

REFERENCES

Brock, F. V., and K. C. Crawford (1995). The Oklahoma mesonet: A technical overview, *J. Atmos. Oceanic Tech.* **12**, 5–19.

Campbell, E. C. (1987). Overview. "Good Morning, This is Campbell Scientific . . ." (in-house newsletter), March, 1.

Campbell, E. C. (1991). Overview, Independent Verification. The Campbell Update, A Newsletter for the Customers of Campbell Scientific, Inc., **2**(1), 2.

Hardy, D. R., M. Vuille, C. Braun, F. Keimig, and R. S. Bradley (1998). Annual and daily meteorological cycles at high altitude on a tropical mountain, *Bull. Am. Meteorolog. Soc.* **79**, 1899–1913.

Schimmelpfennig, H. G., B. D. Tanner, and E. C. Campbell (1979). Applications of a Minature, Low Power, Computing Datalogger in Environmental Investigations. *Fourteenth Conference on Agriculture & Forest Meteorology and Fourth Conference on Biometeorology, April 2–6, 1979, Minneapolis, MN*, American Meteorological Society, Preprint Volume, pp 154–155.

Tanner, B. D. (1990). Automated weather stations, *Remote Sensing Rev.* **5**(1), 73–98.

CHAPTER 40

DESIGN, CALIBRATION, AND QUALITY ASSURANCE NEEDS OF NETWORKS

SCOTT J. RICHARDSON AND FRED V. BROCK

1 INTRODUCTION

The calibration and quality assurance needs of meteorological networks vary significantly depending on factors such as end-user needs, data accuracy and resolution requirements, site maintenance schedule, climatology, and, of course, the budget of the project. This chapter describes general principles that can be used to ensure quality data in meteorological networks.

The design of a meteorological observation network depends on many factors, perhaps most importantly what the system is to measure and with what accuracy. However, design of a measurement system is also powerfully affected by other considerations such as choice of sensor and data logger. Selection of the measurement platform, data communication system, and type of power system has a profound affect on overall system design. Communication system limitations may dictate the location of remote sites forcing compromises in site location. Power limitations may prohibit the use of certain types of sensors.

A very important and sometimes overlooked aspect of system design is future data requirements and network upgrades. A system designed to meet today's requirements will not necessarily satisfy tomorrow's needs and/or may not work well with new technology. A data collection system should be designed with this in mind so that upgrades can be made without the need to replace the entire infrastructure of the network. Note that once a data collection system and/or vendor is chosen, it may be necessary to use this same system for the duration of the project including upgrades or modifications. This is because, for example, once a data logger is chosen it may

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

not be possible or feasible to switch to a different vendor or system because of compatibility problems. This is not necessarily bad but should be considered when choosing a manufacturer. For example, one should consider if the manufacturer will be in business when network upgrades are required and, if not, how much of the system will require replacement.

The text draws from the experiences of the Oklahoma Mesonet (Brock et al., 1995) as well as the Atmospheric Radiation Measurement (ARM) Program (Stokes and Schwartz, 1994; DOE, 1990); a brief description of both will be given to put the subsequent writing in context. More details on network design and instrument performance can be found in *Meteorological Measurement Systems* by Brock and Richardson (1991). *An Introduction to Meteorological Instrumentation and Measurement* by T. P. DeFelice (1998) is also a good reference.

The Oklahoma Mesonet

The Oklahoma Mesonet is a system of 115 remote surface observing stations across the state of Oklahoma and is a joint effort between Oklahoma State University and the University of Oklahoma. Parameters measured at all 115 sites include pressure, wind speed and direction at 10 m above ground level (agl), air temperature and relative humidity at 1.5 m agl, rainfall, global solar radiation, and ground temperature at 10 cm under bare soil and native ground cover. Additional parameters measured at about half the sites include wind speed at 2 m agl, air temperature at 9 m agl, soil moisture at 5, 25, 60, and 75 cm below ground, and additional soil temperature measurements down to 30 cm. Surface flux measurement capabilities are being added to the Oklahoma Mesonet through the OASIS (Oklahoma Atmospheric Surface-layer Instrumentation System) Project (Brotzge et al., 1999). Once completed in 2000, Oklahoma Mesonet sites will measure net radiation, ground heat flux, and sensible flux, and latent heat flux using a profile technique (at 90 sites) and eddy correlation techniques (at 9 sites).

Five-minute averaged data are collected from all 115 stations every 15 min and transmitted through the Oklahoma Law Enforcement Telecommunication System (OLETS) to an archival system at the University of Oklahoma. All stations are solar-powered with battery backup; site maintenance is typically performed 3 to 4 times per year. The system was commissioned in 1994 and each year more than 99.9% of possible observations (approximately 15,537,000 of a possible 15,547,000 observations) are archived. The use of OLETS was important to the success of the Oklahoma Mesonet, since it and allowed data collection and two-way communication between the base station and all remote sites without the need for expensive phone lines.

Atmospheric Radiation Measurement Program

The ARM Program is a multilaboratory, interagency program that was created in 1989 with funding from the U.S. Department of Energy (DOE). The ARM Program is part of the DOE effort to resolve scientific uncertainties about global climate change with a specific focus on improving the performance of general circulation models (GCMs) used for climate research and prediction. These improved models will help scientists better understand the influences of human activities on Earth's climate.

In pursuit of its goal, the ARM Program establishes and operates field research sites, called Cloud and Radiation Testbeds (CARTs), in several climatically significant locations (the north slope of Alaska, the tropical western Pacific, and the U. S. Southern Great Plains). Data are collected over extended periods of time (years) from large arrays of instruments (both state-of-the-science and conventional instrumentation are used) to study the effects and interactions of sunlight, radiant energy, and clouds on temperatures, weather, and climate. Specifically, ARM focuses on cloud–radiation interactions.

The ARM Program has taken advantage of the advances in communications and the World Wide Web. Data are ingested and available in near real time via the Web, trouble reports are submitted via the Web, complete site and instrument history information is archived and available via the Web, etc.

This chapter will emphasize design, calibration, and quality assurance needs of smaller networks such as the Oklahoma Mesonet or smaller.

2 SYSTEM DESIGN

Sensors are typically mounted on a stationary platform (a simple mast or tall tower) or on a moving platform (balloons, planes, ships, etc.). Ideally, data are communicated in real time from the measurement site or platform to a central archiving facility. In some cases, real-time communication is not possible but, instead, data are manually collected at periodic intervals, usually in some electronic form. Availability of electrical power, or the lack of it, may seriously affect the system design.

Instrument Platforms

It is not surprising that virtually every type of instrument platform is used in meteorology because the atmosphere is so extensive and because most of it is quite inaccessible. These platforms include masts, instrument shelters, tall towers, balloons, kites, cars, ships, buoys, airplanes, rockets, and satellites. Synoptic data platforms include balloons and satellites supplemented by buoys and ships over the ocean. In addition, aircraft are used for hurricane observation and some data are collected from commercial flights to fill in gaps in the observation networks. Aircraft are extensively used for research investigations around thunderstorms or wherever high-density upper air data are needed.

When selecting a platform, consideration should be given to where the measurement is to be made and whether the platform can be permanently fixed or is moving, cost, and exposure. To some extent, any platform, even a simple tower for surface measurements, interacts with the atmosphere and affects instrument exposure. A simple 10-m tower, shown in Figure 1, has a wind sensor at 10 m and temperature and a relative humidity (T&RH) sensor at 1.5 m in addition to a radio antenna for data transmission, a solar panel and battery for power, a barometer, and a data logger. These sensors must be mounted with due consideration for exposure to prevailing winds to minimize tower effects.

Communication Systems

A communications network is a vital part of almost every meteorological measurement system at all scales. Historically, meteorological communications have relied primarily on land-line and radio links. More recently, polar orbiting and geostationary satellites are used for data communications in macroscale or synoptic measurement systems and even in many mesoscale systems. Commercial satellites are used to broadcast data from central points, with sophisticated uplinks, to users equipped with fairly simple antennas and receivers (inexpensive downlinks).

The ideal communications system would reliably transmit data from the remote instrument platform to a central facility and in the reverse direction with little or no time delay, without limiting the volume of data to be transmitted. Communication both to and from the remote site is used to synchronize local clocks in the data loggers, to load operating programs into the data loggers, and to make special data requests, to name a few. Two-way communications is not essential but highly desirable.

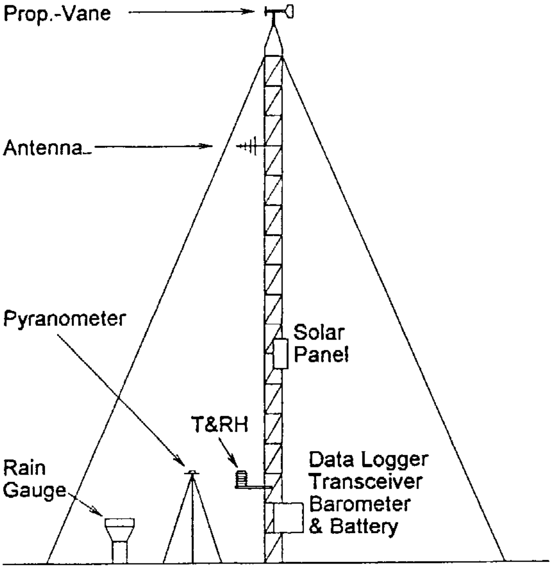


Figure 1 A 10-m tower for surface measurements. At the top of the tower is a propeller-vane anemometer that measures wind speed and direction. Below the anemometer is an antenna for two-way data communications, and power is provided by a solar panel. The pyranometer is mounted to the south where tower and guy-wire shadows will not affect the data.

Telephone Commercial telephone systems provide adequate signal bandwidth, are generally reliable, and cover most land areas. The cost is prohibitive if one must pay for running lines to each station, especially for a short-term project. Even for long-term projects like the Oklahoma Mesonet and the ARM Program, phone lines were either avoided entirely or used very sparingly due to the expense involved.

Recent advances in cellular telephone technology coupled with decreasing airtime charges means cellular data communications has become a viable alternative to traditional phone lines.

Direct Radio Direct radio links from the remote stations to a central base station are desirable because they offer flexibility but Earth curvature limits line-of-sight links. Figure 2 shows the maximum line-of-sight distance between two stations if the remote station antenna is at a height of 10 m. For a base station or repeater antenna height of 200 m, the line-of-sight link is only a little more than 60 km. Direct radio, even when augmented with repeaters, severely limits the size of a network and causes immense difficulties in complex terrain. For example, if the path of the signal from a remote station to the repeater or to the base station is too close to the ground, the signal could be trapped in an inversion layer and ducted away from the intended destination, thereby losing the connection between the sensor and the base station.

Satellite The first communications satellite that permitted an inexpensive uplink (low-power transmitter and simple antenna) was the *Geostationary Operational Environmental Satellite* (*GOES*). An inexpensive uplink is essential when a large

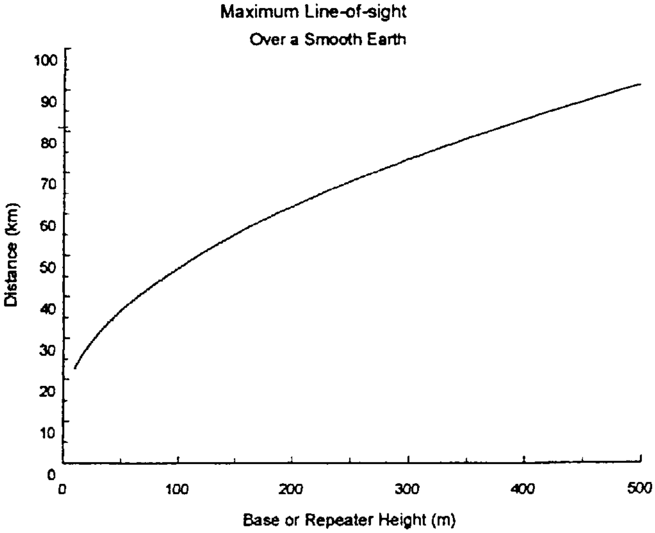


Figure 2 Line-of-sight distance over a smooth earth as a function of height of one end of the link when the other end is fixed at 10 m.

number of remote stations are involved. In addition, stations may be powered by batteries and solar panels, thereby requiring low-power radio transmitters. As satellite communications technology evolves, communication restraints will be eased, which will lead to vastly improved meteorological networks, especially for the mesoscale.

Power Source Electrical power consumption of a measurement system is often a vital consideration; the primary concern is cost. Where commercial power is available, cost is not usually a problem. However, many systems are required to be portable or to operate in locations where commercial power is not available. In these cases, the power source is usually batteries, perhaps supplemented with solar panels. Such systems must operate on a severely limited power budget and that constraint affects the selection of components and the overall system design. Battery-powered systems are constrained to select sensors with low power consumption and/or to switch the sensors on only as needed to conserve power. Heaters generally cannot be used and local computational capability may be severely limited. Therefore, all components must be rated for operation over the expected temperature range.

Unique methods are sometimes used to power remote meteorological stations and are often used out of necessity. The following is an example of what can be done when no traditional power source is available.

The year-long SHEBA (Surface Heat Budget of the Arctic Ocean) field experiment used PAMII ~1 (the third-generation Portable Automated Mesonet) remote meteorological stations and was located well above the Arctic Circle. PAMIII stations require 1.5 to 2.5 A at 12 V dc for continuous operation and, for most deployments, the power system consists of photovoltaically charged deep-cycle storage batteries with nominal capacity of 3 to 6 days of operation in the absence of sunshine. Obviously, modifications were required for SHEBA due to the lack of sunshine during the winter months and the extreme cold. Three separate power sources were eventually used for differing reasons and at different times of the year.

The primary source of power for each station was a small propane-fueled 12 V dc electro-thermal generator (TEG) and a single battery. To mitigate the effects of extreme cold, the generator, 2 propane bottles, the battery, and PAIII electronics were all placed inside an insulated container mounted on a sled for transportation between sites. Two 33-lb propane bottles allowed for 15 to 20 days of unattended operation. Heat from the TEG kept the electronics and more importantly the propane and battery warm when outside temperatures dropped below −15 °C. In the spring two photovoltaic panels were installed at each station and paralleled with the TEG to augment system power. In addition, two small wind generators were added in spring to the least accessible remote stations. The wind generators operated at 12 V dc that allowed parallel operation with both the TEG and photovoltaics through the PAIvIIIiI battery charging and monitoring system.

Some problems were encountered with the TEG, caused by poor fuel quality. However, the system described above was in general successful and was a unique solution to a difficult situation.

3 CALIBRATION

The calibration needs of a network depend on, among other factors, the desired accuracy of the field measurements. Sufficiently accurate laboratory standards are readily available for most meteorological sensors, but acquiring this same level of accuracy in the field is often very difficult if not impossible. For example, it is relatively easy to calibrate temperature sensors in the lab to an accuracy of $\pm 0.1°\mathrm{C}$ or even $\pm 0.01°\mathrm{C}$. However, measuring atmospheric air temperature with this same accuracy requires great care and knowledge of the measurement system. Therefore, extremely accurate laboratory calibrations may not be required in some situations. Nevertheless, the total error associated with an instrument in the field is the square root of the sum squared errors (assuming the errors are independent) and, therefore, the total measurement error can be reduced if a careful laboratory calibration is performed.

Calibration standards are maintained by standards laboratories in each country such as the National Institute of Science and Technology (NIST) in the United States. Standards for temperature, humidity, pressure, wind speed, and for many other variables are maintained. The accuracy of these standards is more than sufficient for meteorological purposes. Every organization attempting to maintain one or more measurement stations should have some facilities for laboratory calibration including transfer standards. These are standards used for local calibrations that can be sent to a national laboratory for comparison with the primary standards. This is what is meant by traceability of sensor calibration to NIST standards. Ideally the calibration of all sensors can be traced back to such a standards laboratory.

Initial calibration of all instruments should include a full range test if at all possible. Even when instrumentation is new from the manufacturer, *all instruments should be tested prior to use in the field*. Testing each instrument can be time consuming and expensive but to not do so is a mistake that will inevitably cost many times more. Even seemingly simple instruments require a laboratory check before field use. For example, a tipping bucket rain guage has an output that is a function of rain rate and not the same for all instruments; therefore, a different calibration curve may have to be developed for each sensor. In addition, slight variations in bucket volumes can result in differences in measured rainfall and can be accounted for. As each rain guage is "calibrated," sensor deficiencies (e.g., faulty bearings or closure switches) may be detected prior to field deployment where such problems could go undetected.

The importance of checking, if not calibrating, each sensor prior to field use cannot be overemphasized. Instrument manufacturers, no matter how reputable, may not be able to maintain the desired standards. Even if sensor specifications meet or exceed the required accuracy, instruments can drift or be damaged during shipment and storage or electronic components can fail. Therefore, to maintain the highest level of confidence in a network's data, sensor performance should be verified prior to use in the field.

Fortunately, in many cases it is not necessary to purchase expensive calibration chambers to achieve the level of accuracy required for meteorological measurements. Simple yet effective “calibration” methods that require a minimal investment can be developed for many meteorological instruments (calibration is put in quotes to indicate that true calibrations, which require a sensor with an order of magnitude more accuracy, may not be occurring). For example, rain guage calibration can be performed with a reasonably accurate electronic scale with RS-232 or analog output, a simple water reservoir, a simple flow valve, and a data collection system like that used in the network. The scale output is connected to a computer (or datalogger) and the rain guage is connected to the data collection system. The water reservoir is placed on the scale and water is siphoned off into the rain guage. The scale output (weight of water remaining in the reservoir) is easily related to the water input to the rain guage and the siphon rate can be adjusted to simulate different rain rates. This calibration does not account for many factors that can affect the rain guage accuracy but does provide a simple yet effective check of the instrument performance.

An equally simple setup can be developed to test temperature sensors (Richardson, 1995). Required equipment includes a temperature transfer standard with errors less than about 0.030°C, a temperature chamber (e.g., a small freezer), a bath, and a stirrer. The use of a temperature standard means it is not necessary to maintain precise chamber temperatures nor is it necessary to hold the chamber temperature constant; it is only necessary to control the rate of change of temperature. The rate must be low enough that errors induced by spatial gradients between the temperature standard and the test sensors and the errors induced by the time response of the sensors are small compared to the acceptable calibration error.

Dynamic errors are fairly easy to detect if the test sensors lag the reference sensor during increasing temperature, and lead it when the temperature is falling, the rate of change is too high. The response of the sensors will be a function of the kind of bath (air or liquid) and the amount of stirring. The bath also affects spatial gradients. These can be detected by correlating errors with sensor position.

4 QUALITY ASSURANCE

Quality assurance methods and/or final data format may depend on the primary end user of the data, e.g., will they be used for research or by the general public? For example, relative humidity (RH) sensor inaccuracy specifications allow for the RH reported by the sensor to be in excess of 100% (as high as 103%). A sensor reporting an RH of 102% may be operating correctly but could cause problems for under-informed users. For example, modelers ingesting RH may need to be aware that values greater than 100% are possible and do not necessarily indicate supersaturation conditions. In addition, those unaware of instrument inaccuracy specifications may think the data is incorrect if RH is above 100%.

A more complete data quality assurance program can be developed for a network of stations than for a single measurement station, so this discussion will address the issue of quality assurance for a network. A single-station assurance program would be a subset of this. In designing a measurement system, it is useful to consider the impact of automation on data quality.

Schwartz and Doswell (1991) claim that automation sometimes has been accompanied by a decrease in data quality but that this decrease is not inevitable. When a measurement system is automated, some sources of human error are eliminated such as personal bias and transcription mistakes incurred during reading an instrument, to name only two. However, another, more serious form of error is introduced: the isolation of the observer from the measurement process.

One tends to let computers handle the data under the mistaken assumption that errors are not being made or that they are being controlled. Unless programs are specifically designed to test the data, the computer will process, store, transmit, and display erroneous data just as efficiently as valid data. Automatic transmission of data tends to isolate the end user from the people who understand the instrumentation system. Utilization of computers in a measurement system allows data to be collected with finer time and space resolution. Even if the system is designed to let observers monitor the data, they can be overwhelmed by the sheer volume and unable to effectively determine data quality.

Automation, or the use of computers in a measurement system, can have a beneficial impact on data quality if the system is properly designed. Inclusion of data monitoring programs that run in real time with effective graphical displays allow the observer to focus on suspect data and to direct the attention of technicians.

The objective of the data quality assurance (QA) system is to maintain the highest possible data quality in the network. To achieve this goal, data faults must be detected rapidly, corrective action must be initiated in a timely manner, and questionable data must be flagged. The data archive must be designed to include provision for status bits associated with each datum. The QA system should never alter the data but only set status bits to indicate the probable data quality. It is probably inevitable that a QA system will flag data that are actually valid but represent unusual or unexpected meteorological conditions. Flagged data are available to qualified users but may not be available for routine operational use.

The major components of a QA program are the measurement system design, laboratory calibrations, field intercomparisons, real-time monitoring of the data, documentation, independent reviews, and publication of data quality assessment. Laboratory calibrations are required to screen sensors as they are purchased and to evaluate and recalibrate sensors returned from the field at routine intervals or when a problem has been detected. Field intercomparisons are used to verify performance of sensors in the field since laboratory calibrations, while essential, are not always reliable predictors of field performance. QA software is used to monitor data in real time to detect data faults and unusual events. A variety of tests can be used from simple range and rate tests to long-term comparisons of data from adjacent stations. Documentation of site characteristics and sensor calibration coefficients and repair history is needed to answer questions about possible data problems.

Independent reviews and periodic publication of data quality are needed since people close to the project tend to develop selective awareness. These aspects of the overall QA program must be established early and enforced by project leaders. The whole QA program should, ideally, be designed before the project starts to collect data.

Laboratory Calibrations

As discussed previously, laboratory calibration facilities are required to verify instrument calibration and to obtain a new calibration for instruments that have drifted out of calibration or have been repaired. However, a laboratory calibration is not necessarily a good predictor of an instrument's performance in the field. This is because laboratory calibrations never replicate all field conditions. For example, a laboratory calibration of a temperature sensor does not include all the effects of solar and Earth radiation nor is it likely to simulate poor coupling with the atmosphere due to low wind speeds. This type of error is called "exposure error" and is a result of inadequate coupling between the sensor and the environment. Exposure errors can be very large (e.g., radiative heating of temperature sensors can result in air temperature errors in excess of 5°C; Gill, 1983) and are not accounted for in lab calibrations. In addition, the manufacturer does not specify this type of error because it is not under its control. Therefore, knowledge of the measurement system and instrument design is crucial to minimize measurement error.

Field Intercomparisons

Two types of field intercomparisons should be performed to help maintain data quality. First, a field intercomparison station should be established and, second, when technicians visit a station, they should carry portable transfer standards and make routine comparison checks.

The field intercomparison station should be comprised of operational sensors and a set of reference sensors (higher quality sensors). Both should report data to the base station but the reference station data should be permanently flagged and not made available for operational use.

Portable transfer sensors can be used to make reference measurements each time a technician visits a station. This method can detect drift or other sensor failures that could otherwise go undetected. These sensors can include a barometer and an Assmann psychrometer. In addition, technicians can carry a lap-top computer to read current data, make adjustments to calibration coefficients when sensors are changed, set the data logger clock (*if* it cannot be set remotely), and reload the data logger program after a power interruption.

Data Monitoring

Neither laboratory calibration nor routine field intercomparisons will provide a real-time indicator of problems in the field. In a system that collects and reports data in real time, bad data will be transmitted to users until detected and flagged. The volume of data flow will likely be far too great to allow human observers to effectively monitor the data quality. Therefore, a real-time, automatic monitoring system is required.

The monitor program should have two major components: scanning algorithms and diagnostic algorithms. The function of the scanning algorithms is to detect outliers while the diagnostic algorithms are used to infer probable cause. The monitor program can analyze the incoming data stream using statistical techniques adapted from exploratory data analysis, knowledge of the atmosphere, knowledge of the measurement system, and using objective analysis of groups of stations during suitable atmospheric conditions.

Exploratory data analysis techniques are resistant to outliers and are robust, i.e., insensitive to the shape of the data's probability density function. Knowledge of the atmosphere allows us to place constraints on the range of some variables such as relative humidity that would be flagged if it were reported greater than approximately 103% (due to sensor inaccuracy specifications). Knowledge of the measurement system places absolute bounds on the range of each variable. If a variable exceeds these limits, a sensor hardware failure may have occurred.

The monitor program must be tailored for the system and should be developed incrementally. Initially, it could employ simple range tests while more sophisticated tests can be added as they are developed. The QA monitoring program will never be perfect; it will fail to detect some faults and it will label some valid data as potentially faulty.

Therefore, *the monitor program must not delete or in any way change data* but set a flag associated with each datum to indicate probable quality. The monitor program should have a mechanism to alert an operator whenever it detects a probable failure. Some of these alerts will be false alarms, i.e., not resulting from hardware failure, but may in fact indicate interesting meteorological events.

Documentation

There are several kinds of documentation needed: Of individual station characteristics, a station descriptor file, and a sensor database.

Station characteristics can be documented by providing an article describing the station and its instrumentation. In addition, there should be a file of panoramic photographs showing the fetch in all directions and the nature of the land. Aerial photographs can also provide valuable information about a site, as can high-resolution topographical maps.

As part of the system database, a system descriptor file must include the location and elevation of each station and the station type (e.g., standard meteorological station, special-purpose agricultural station, or sensor research station).

It is necessary to maintain a central database of sensors and other major components of the system including component serial number, current location, and status. Some sensors have individual calibration coefficients so there must be a method of accounting for sensors to ensure that the correct calibration coefficients have been entered into the appropriate data logger. It is also necessary to keep records of how long a component has been in service and where it was used so that components that suffer frequent failures can be identified. This would help to determine if the component was seriously flawed or if the defect was characteristic of the component design.

This kind of accounting cannot be left to chance; if it is, sensors will inevitably be matched with the wrong calibration coefficients or put back into service without having been repaired or recalibrated. Some sensors require periodic recalibration, but it is not feasible to recalibrate them all at once. Therefore a formal database system should be set up. All technicians should be required to report maintenance activity including swapping components, and this information should be entered into the database. The database system should be able to generate reports indicating the serial number of every component at a station, the number of components awaiting repair at any given time, the number of spares available, the history of any given sensor or sensor type, etc.

Independent Review

For much the same reason that scientific proposals and papers are reviewed, periodic, independent reviews of a network's performance should be invited. It is always possible for people in constant close proximity to a project to become blind to problems; an independent review would help alleviate this at the root cause.

Publication of Data Quality Assessment

There will be frequent data faults in any network even with the data quality assurance program outlined above. To assist critics in making a realistic assessment, it is desirable to publish, periodically, an honest appraisal of the network performance including all data faults, causes when known, and action taken.

REFERENCES

Brock, F. V., K. C. Crawford, R. L. Elliott, G. W. Cuperus, S. J. Stdaler, H. L. Johnson, and M. D. Eihs (1995). The Oklahoma Mesonet: A technical overview. *J. Atmos. Oceanic Technol.* **12**, 5–19.

Brock, F. V., and S. J. Richardson (2001). Meteorological Measurement Systems, Oxford University Press, Inc. New York.

DeFelice, T. P. (1998). *An Introduction to Meteorological Instrumentation and Measurement*, Prentice Hall, New Jersey.

DOE (1990). U.S. Department of Energy, Atmospheric Radiation Measurement Program Plan, DOE/ER 0441, National Technical Information Service, Springfield VA.

Gill, G. C. (1983). Comparison Testing of Selected Naturally Ventilated Solar Radiation Shields, Report submitted to the NOAA Data Buoy Office, Louis, MS.

Richardson, S. J. (1995). Automated temperature and relative humidity calibrations for the Oklahoma Mesonetwork, *J. Atmos. Oceanic Tech.* **12**, 951–959.

Schwartz, B. E., and C. A. Doswell III (1991). North American rawinsonde observations: Problems, concerns and a call to action, *Bull. AMS* **72**, 1885–1896.

Stokes, G. M., and S. E. Schwartz (1994). The Atmospheric Radiation Measurement (ARM) Program: Programmatic background and design of the cloud and radiation test bed, *Bull. Am. Meteor. Soc.* **75**, 1201–1221.

CHAPTER 41

DATA VALIDITY IN THE NATIONAL ARCHIVE

GRANT W. GOODGE

1 INTRODUCTION

As we enter a new millennium the United States and indeed the whole world is becoming increasingly sensitive to various meteorological and climatological extremes. Given the potential for economic and political disruption as a result of these anomalies, there has been an increased interest in their prediction. Such predictions include frequency, magnitude, areal extent, and location. As of this time one of our best methods of understanding the potential for future climatic anomalies is to look at the past climatic record.

The integrity of any scientific research depends a great deal on the quality of the data used as well as the user's understanding of the data and all relevant information associated with its collection and processing (metadata).

These metadata include, but are not restricted to, instrument type, exposure, automatic, manual, maintenance/calibration, and the nature of data transfer from the point of measurement to archive. The final archive for much of the meteorological data for the United States and its territories resides at the National Climatic Data Center (NCDC) in Asheville, North Carolina. A considerable volume of similar data from other countries and the world's oceans are also archived at NCDC. The NCDC is one of several environmental and geophysical data centers in the United States and is administered by the National Environmental Satellite Data and Information Service (NESDIS), which is one of the major components of the National Oceanic and Atmospheric Administration (NOAA).

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

The NCDC is not directly involved with the observation of meteorological variables, sensor maintenance, or observer training. That portion of data collection is managed by the National Weather Service (NWS), which is also a major component of NOAA. The NCDC does, however, actively engage in providing relevant feedback to NWS managers concerning sensor and/or observer problems that are discovered during the quality assurance processes at NCDC.

Unfortunately, errors in the meteorological databases may be introduced anywhere between the point of observation and its final storage in the digital archive or websites. Several examples are observer transcription errors, radio frequency (RF) interference or lightning damage to electronically driven sensors, transmission signal interruption, keying errors of manually observed data during data entry and/or update during quality assurance, and computer program transfer of those files to the digital archives or websites. Also an increasingly problematic issue in today's electronic world is that of missing data. This can be of particular significance with the transmission of data from the Automated Surface Observing System (ASOS) where data are stored on site for only 12 h. Obviously, any interruption of the link between the sites and NCDC that exceeds this period will result in the permanent loss of data from affected sites until the link is restored. We will return to a more detailed discussion of some of these issues later in this chapter.

Before proceeding further, this writer would like to note that within the constraints of limited budgets and personnel, NCDC has always endeavored to produce high-quality databases and publications. Any discussion of the weaknesses in the observational systems and the quality assurance of the resulting observations are mentioned only to alert the research community to known data problems and encourage a better understanding of the data prior to use in any study. It is *not* intended as an indictment or criticism of those responsible for the collection and archive of these environmental observations.

2 TEMPERATURE

In the recent era there has been much public and scientific debate of the issue of "global climate change," and many of those debates center on the parameter of temperature. Certainly the fact that the magnitude of atmospheric model predicted centennial mean global temperature rises are on the same order (1°F) as the required accuracy (±1°F) of NOAA-sponsored instruments indicate the necessity of quality data. Prior to the 1960s official Weather Bureau thermometers were all compared against a National Bureau of Standard's "certified" thermometer, and any mercurial thermometer having a greater error than 0.3°F (above 0°F) from the certified standard was not accepted for observational use. During the 1960s most primary Weather Bureau Stations transitioned from the use of mercurial thermometers to an HO-60 series temperature sensor that employed the use of a fan-aspirated thermistor exposed in a metal shield with the sensed temperature electrically transmitted to an electromechanically driven dial in the weather office. Unfortunately, some of these instruments did not meet the more relaxed standard of ±1°F as compared to the ±0.3°F standard of the mercurial thermometers they replaced.

In one documented case the mean temperatures, sensed by the HO-62 hygrothermometer, at the Asheville, North Carolina, NWS Airport station, migrated upward more than 4°F (as compared to three surrounding cooperative stations) in less than 4 years (1965–1969). Since that time there have been other migrations and/or shifts of ±2°F. The most recent of these have been as a result of sensor changes and/or relocations at the airport. Certainly this station's record would not be a good candidate for climate trend determinations. Unfortunately, sensor-driven data errors and discontinuities for this station and others like it remain unadjusted in the official databases and archives and in some cases undocumented.

There are several reasons why NCDC does not adjust or correct such errors during the quality assurance (QA) process. The primary reason is that NWS stations (i.e., Chicago, Charlotte, Chattanooga, etc.) are quality assured on a single station basis and as a result are unable to identify instrument "drift" unless that drift were to exceed station monthly climatic extremes or cause conflict with another reported value such as ambient temperature versus dew-point temperature. Second, there are seldom any coincident data that would allow comparison or correction of the station in question. Lastly, the historic digital data storage at NCDC has been sequential in nature, which makes the correction of identified, postprocessing errors very expensive.

It should be noted that a change in a temperature sensor's exposure can make as much difference in the observed readings as a change in the type of sensor. Relocation of meteorological instruments at primary weather stations has been a problem for climatologists for many years. Some have been moved from rooftops of post offices in a downtown (urban) environment to rooftops of terminal buildings at airports and finally from the terminal buildings to airport field level exposures. After moving to the field level, sensors have continued to be moved due to airport expansions or alterations. Even if the sensor has not moved, increased extent of paved surfaces may alter the radiative flux of the area surrounding the sensor.

These relocations of sensors highlight the importance and need for timely and accurate station history information called metadata. Without it, a user or researcher is left to guess as to what may have caused the sudden change in the data. Accurate metadata is just as important in the documentation of observations from the cooperative observer network. Even though these stations only record data on a daily basis, it is still essential that any retrospective user know the type of instruments used, the exposure, and the time of day when the readings are made. Unfortunately, NCDC's receipt of the B-44 (metadata) forms lagged several months behind the date of actual station move or instrument change, and thus the publication and databases improperly indicated the location, instrument type, time of observation, or other critical station information.

In about 1984 the NWS began replacing the liquid-in-glass thermometers at cooperative observing sites with an electronic resistance thermometer (hereafter referred to as the Maximum Minimum Temperature System (MMTS)). The sensing accuracy of these MMTS units are generally compatible with the liquid-in-glass thermometer they replaced, but the introduction of the MMTS brought several new undesirable results. First is its susceptibility to electrical interference whether that be generated by radio frequency (RF) from nearby radio transmitters or induction surges from nearby lightning strikes. Usually, the display inside the observer's home or the thermistor outside will be burned out by a nearby lightning strike and result in data loss until the field manager either makes a site visit to replace the unit or ships the affected part(s) to the observer for their own replacement. Other causes of a data loss from the MMTS resulted from the connecting cable being cut by nearby utility construction or in the western states by various energetic burrowing animals. Electrical power interruptions to the units either by thunderstorm or winter snow and ice have also caused data loss over the years. These data losses can be significant when the power outages are widespread and extended, such as was the case in the New England ice storm of January 1998.

As undesirable as the above-mentioned data losses are, they are at least apparent and recognizable to the user. This, however, is not true when RF, lightning, or the salt air of the marine environment cause small changes in the resistance in the electrical loop between the thermistor and the readout in the observer's home or business. Such a resistance change caused abnormally cool temperatures, some of which were below freezing, to be recorded at a low elevation site in Hawaii. In another case, a lightning strike caused a resistance change equal to about 4.5°F that went undetected by the observer, NWS, and NCDC for 6 months. In these cases the errors were not detected by NCDC's automated QA program but rather by quality assurance specialists.

3 PRECIPITATION

In terms of identifying global climate changes, anthropogenic or natural, quality precipitation data is equally as important as temperature; however, because of the noncontinuous temporal nature and spatial variability of precipitation, its quality assurance is more difficult. As was the case with temperature data, errors in precipitation data can be introduced at any stage from the original observation to the final quality assurance and archive.

Before continuing with the discussion of precipitation data, it should be noted that there are many more uses of meteorological data than just long-term climate analysis. In fact there are thousands of users that depend on high-quality data, and since many of those users are interested in the condition at a single point and time, they cannot afford the perceived luxury expressed by some climatologists that errors in the data sets will likely be random and thus average out over the long run. These "single point and time" uses of the data range from the meteorological conditions at the site of an auto accident, aircraft mishap, or the structural failure of a bridge or building. Even though the original reasons for the collection of meteorological data did not envision such applications, they are none the less vital to our understanding of the effects of extreme meteorological conditions on human activities and their supporting infrastructure.

As in the case of temperature measurements, one of the best assurances of quality precipitation data at the observation site is to have some element of redundancy or reality check. The high costs of even the basic manual precipitation gages employed by the government-sponsored networks proved to be too expensive to install more than one identical gage at most cooperative stations. There are, however, about 850 of those stations that are equipped with two gages, one manual and the other automatic. We will discuss this subject later in the chapter. There also was more than one precipitation gage located at most primary NWS sites that were staffed by trained government or contract weather observers.

Observational procedures at these sites, however, did not require a comparison of the measured amounts of each gage. The high quality of precipitation data from these sites came primarily through the aspect of their reporting of hourly, 6 hourly, and 24 hourly amounts, which provided for cross reference at the station as well as at NCDC when processing the data for archive and publication.

Unfortunately this cross-referencing process was significantly compromised with the advent of the Automated Surface Observing System (ASOS) in the mid 1990s. ASOS is equipped with a heated tipping bucket precipitation sensor that does not always dependably and accurately measure precipitation. Its greatest difficulty occurs during periods of frozen or freezing precipitation. In the years prior to ASOS, human observers monitored conditions so they could resolve such problems through the use of a backup gage or make an estimation of precipitation amounts; however, now with no observers at most ASOS sites to augment the automated system, the weather indicator Light Emitting Diode Weather Identifier (LEDWI) may report snow while the precipitation gage does not report any corresponding meltwater equivalent and, if temperatures warm to near or above freezing, the snow will melt and produce a false time and intensity for the precipitation event.

Even in those cases where observers are in place to augment the ASOS observations, they do not have the ability to alter the one-minute data file from which the hourly precipitation files and publications are produced, the net effect being that no longer does the hourly precipitation data from the ASOS sites always agree with the 6-hourly, daily, or monthly totals. Needless to say such discrepancies seriously undermine the integrity and use of the data, particularly in litigation or insurance adjustment cases.

As was earlier mentioned about 850 of the cooperative weather stations are equipped with both manual and automatic precipitation gages. One of the gages is the "standard 8," which has an 8-in. diameter receiving funnel that directs the precipitation into a measuring tube that is one-tenth the area of the receiving funnel and thus expands the true depth of the rainfall by a factor of 10 and provides for an accurate measurement of the precipitation to one hundredth (0.01) of an inch. Today the vast majority of the 2250 automatic gages are Fischer–Porter type of gages, which also have an 8-in. diameter receiving chute that directs the precipitation into a large basin where the precipitation is weighed. That weighed amount is recorded onto a binary punched paper tape every 15 min to a maximum precipitation resolution of one tenth (0.10) of an inch.

The use of a weighing-type recording precipitation gage in the cooperative network began in the mid-1930s and continued until the mid-1960s when the Fischer–Porter type began to replace it. This earlier weighing gage was known as the Universal and was not only used at about 3000 cooperative stations but also became the primary gage for the primary NWS stations from the early 1960s until the mid-1990s when they were phased out with the commissioning of ASOS.

There were two major advantages of the Universal-type gage over the Fischer–Porter. First the Universal recorded data to the hundredth of an inch (0.01), and secondly it made a constant trace of the precipitation event on a chart that allowed a time resolution of as little as one minute. The accepted world record 1 min rainfall of 1.23 in. (3.12 cm) was extracted from one of the Universal autographic charts.

For quality assurance purposes at NCDC it has been unfortunate that for many years the data from the co-located standard and automated gages at the cooperative stations could not be compared during quality assurance and publication because the receipt and reduction of the punched paper tapes and or autographic charts lagged about a month or more behind that of the manually reported daily data forms.

The resulting inability to compare the data from these co-located gages prevented NCDC from effecting a higher level of QA for those sites and in some cases allowed significantly different precipitation amounts to be published for the same location. The worst known difference was published for a site in Puerto Rico. A heavy tropical rain event moved into the south central part of the island in October, 1985. The heavy rains washed out a major highway bridge during the night hours and allowed 29 motorists to drive to their death, one vehicle at a time.

A $65 million lawsuit was later initiated. Obviously, the true amount of rainfall was critical to the case as well as the knowledge of the true return period rainfall potential for the area. Unfortunately, the official published daily total from the manual 8-in. gage was 8.80 in. greater than the sum of the hourly values later published for the co-located automated Fischer–Porter gage.

In this case the automated gage was equipped with a self-siphoning overflow feature that not only drained the collecting bucket when it reached its capacity, but because of the extremely heavy rain, continued to siphon until the heavy rain ended, and thus the gage never accumulated or recorded the balance of the precipitation during the storm.

Fortunately, a more timely receipt and processing of the Fischer–Porter punched paper tapes now allows NCDC to do a comparison of the co-located station's precipitation records. Results of this QA process have also aided NWS cooperative network field managers to identify and correct gage or observer problems in a more timely fashion.

During the twentieth century various automated precipitation gages have been employed not only to measure and report precipitation in real time but also in a retrospective mode to help identify the frequency and nature of extreme precipitation events throughout the United States and its territories. Any user of these data files should pay particular attention to the significant changes that have occurred in the measurement and data extraction over the years.

The most significant of these changes occurred in 1973 when the derived extreme values changed from that of "excessive precipitation" to "maximum short duration precipitation." Excessive precipitation data began to be summarized in 1896 and continued (with several procedural changes) until 1973 when it was terminated in favor of the simpler maximum short duration precipitation (MSDP), which continues to this day. Both the excessive and MSDP systems used the same time periods (5, 10, 15, 20, 30, 45, 60, 80, 100, 120, 150, and 180 min) for which amounts were determined; however, there were two fundamental differences in their determination.

The first was that in determining values for excessive events there were certain precipitation intensity criteria that had to be met before any values were computed. There were some variations in the criteria over the early years, but most centered around that of $.01 \times t + 0.20$ in. (where $t =$ time in minutes) except in the southeastern third of the United States where $.02 \times t + 0.20$ in. was used until 1949 when these southeastern states reverted back to $.01 \times t + 0.20$, which placed the whole nation on the same criteria. Anytime a precipitation event met the above criteria at a given station, excessive precipitation values were determined for that event, and, if there were 10 excessive events during a station-month, then there would be 10 events documented.

This brings us to the second primary difference between the excessive and MSDP systems. Under MSDP there is no intensity criteria that must be met before computation. Therefore, if the maximum intensity for an event were only 0.01 in. in a station-month, it would be summarized and become the extreme for the month. That is because under MSDP procedures there is one extreme summarized, recorded, and published for each station for each month. As was earlier indicated, the excessive method might have any number (10, 12, 15, or none) excessive events in a station-month, while under MSDP if there were multiple excessive events only the most extreme precipitation of each time period (i.e., 5, 10, 15, ..., 180 min) would be summarized, recorded, and published for that station. Anyone who might combine the data from these two files would truly be mixing meteorological apples and oranges. NCDC has a digital file (TD 9656) of excessive precipitation data for the period of 1962–1972, and maximum short duration precipitation (TD 9649) for the period 1973 to the present. Prior to 1962 all excessive precipitation is available only in manuscript form. The MSDP data is also available in manuscript form in the monthly *Local Climatological Data* (LCD) publications from January 1982 through present.

There are also some similar extreme precipitation values that are published in the back pages of the NCDC monthly publication *Hourly Precipitation Data* (HPD). These data are primarily derived from the automated Fischer–Porter gages in the cooperative observing network but also include data from the primary NWS stations as well. Since the maximum time resolution for the automated gages in the cooperative network is 15 min, the "precipitation maxima" summary pages only report precipitation values for 15, 30, and 45 min as well as 1, 2, 3, 6, 12, and 24 h for each station-month.

Once again caution should be exercised when comparing these values with those recorded/published in either the excessive or MSDP systems. The reason being that all values presented in the precipitation maxima tables of the HPD publication are based on data derived from automated gages that are confined to the fixed 15-min clock times while the excessive and MSDP systems are free to move minute by minute along the duration of an extreme event(s) searching for the most extreme precipitation intensity. A close examination of the precipitation maxima table will also show that there are no 15, 30, or 45-min values published for the primary NWS stations. These time periods were omitted in an effort to avoid confusion with the values presented for the same time periods presented in the MSDP tables in the LCD publication for the same station. There are 1-, 2-, 3-, 6-, 12-, and 24-h values published for primary NWS stations in the precipitation maxima table of the HPD; however, for continuity purposes, the maxima table precipitation values for the primary stations are fixed "clock" hour times like the cooperative stations that constitute the majority of stations in the HPD publication. Therefore, these 1-, 2-, and 3-h values will seldom agree with the 60-, 120-, and 180-min values published in the MSDP table of the LCD for that same primary NWS station.

4 SNOW

Given the numerous problems in measuring snow accurately and uniformly (Doesken and Judson, *The Snow Booklet*, 1997) it is no surprise that the quality of snow data in the climate databases are not always what we would like to see. This is true not only of data from the cooperative network stations but unfortunately also from some of the primary NWS stations where, prior to the installation of ASOS (which does not measure snowfall or snow depth) in the mid-1990s, there were professionally trained personnel making the observations.

Obviously, no after-the-fact assurance review of snow data, or indeed any meteorological data, can know exactly what occurred at each reporting site; however, with other accompanying meteorological data or information from that site one can make some reasonable assessments of the validity of the snow reports. The scope of this chapter does not allow a full discussion of all the observational problems encountered in the measurement and quality assurance of snow, but some of the major difficulties encountered will be listed. These will be grouped by (1) observations from the cooperative network observers and (2) those from the NWS primary station network.

Snow observations from cooperative stations presented the greatest quality assurance challenges not only because of a wide variety in the knowledge and training of each observer, but also because of limited supporting meteorological data from the same station. For example, at least 2000 of the average 6800 published cooperative stations report precipitation only (no temperatures), and if an observer at any one of those 2000 sites reports liquid precipitation but no snowfall or snow depth, it is obviously difficult to know whether to assign a "missing" or a "0" to the database and publication.

Such determinations are not too difficult in North Dakota in January or even in Alabama during the blizzard of March, 1993. Unfortunately, these two cases are not representative of most stations during the transitional months of the snow season or for stations located at the geographic margins of a snowstorm.

For those cooperative stations that do report temperatures in addition to precipitation, a "snow–no snow" determination is greatly improved. Automated checks can easily find precipitation reports with both maximum and minimum temperatures below freezing and flag any that do not also include frozen precipitation.

Once the validity of a frozen precipitation event has been established, then the next challenge is to confirm the magnitude of the reported amount. This thought may come as a surprise to some, when one realizes that a significant portion of the U.S. population at large (which includes weather observers) does not clearly understand the proper placement of decimals in writing a bank check (thus the required worded expression of the numerical value) or reporting precipitation on a weather form. Liquid precipitation and/or melted snow is reported in the United States to inches and hundredths (0.00); snowfall to inches and tenths (0.0), and snow depth to whole inches (0).

A survey (by this writer) of precipitation (including snowfall and snow depth) reports during the March, 1993, snowstorm showed that only about 25% of the reports were correct as received by NCDC. Some had only liquid (meltwater) amounts reported, while most reported snowfall or snow depth with no meltwater. Combine the above reports with snow depth written in the snowfall column or snowfall (inches and tenths) written in the snow depth column (whole inches) and a 24.5-in. snowfall would become a 245-in. snow depth. Several reports had the snowfall reported in feet, which uncorrected would have a 3-ft (3′) snowfall recorded as 0.3 in. in the database and publication. Despite these numerous observational problems, NCDC's quality assurance specialists are quite adept in extracting the correct information from the manuscript reports. Much of that ability came from a knowledge of how each station entered their data in the past as well as an areal station-to-station check for magnitude errors.

The automated checks of cooperative precipitation and snow observations at NCDC is very intensive, and the major reason this could be accomplished was because the observation of each precipitation element (i.e., meltwater, snowfall, and snow depth) is made at the same time each day at any given station. Unfortunately, the same cannot be said for the primary NWS stations because their observing practices require a midnight-to-midnight [local standard time (LST)] period for precipitation/meltwater and snowfall, but a 1200 UTC time (7 a.m. EST, 6 a.m. CST, 5 a.m. MST, and 4 a.m. PST, etc.) for the snow depth observation, and if the snow depth on the ground was ≥ 2 in. at 1800 UTC time for the measurement of the meltwater equivalent of the snow on the ground. These observational time differences caused automated comparison checks to falsely flag more good data than bad, and, as a result, the automated checks were turned off. As one might expect, data errors passed on through to the archives and publications (Schmidlin, 1990).

The presence of these errors (mostly meltwater equivalent of snow on the ground) in the database was not totally due to NCDC oversight. There has been a longstanding verbal understanding between NCDC quality assurance specialists and NWS observers that NCDC would not change any observed values without agreement from the observing site in question. Even when presented with significant errors, some observers refused to agree to any correction. One observer from a northern state admitted to this writer that his station only made meltwater equivalent observations once a week instead of the required once daily time period. Following are two other examples of obvious water equivalent errors that were recorded at two primary NWS stations during the March, 1993, blizzard and still exist in the database and publications to this day: A station in the southeast United States recorded 1.85 in. of liquid precipitation for the 18.2 in. of snowfall that fell during the storm; however, the station also reported 3.8 in. of liquid meltwater equivalent for the 18 in. of snow on the ground. Note there was no snow on the ground prior to the storm. The second station was in the northeast and reported 6.3 in. of liquid meltwater, equivalent for 23.2 in. of new snow depth when most other stations in the area reported 1.5 to 2.0 in. of liquid equivalent for similar depths of new snow. Clearly, these two cases were in error.

5 WIND

Throughout the history of the U.S. Weather Bureau and the National Weather Service, the basic instrumentation used for the determination of wind speed has been a rotating cup anemometer (History of Weather Bureau Wind Measurements 1963). Despite this common design, there have been and continue to be many other factors that influence the values of wind speed that are recorded in the publications and databases. These factors include cup design, time constants/averaging periods, units of measure, and sensor exposure.

The first real cup anemometer (History of Weather Bureau Wind Measurements 1963) was introduced by Robinson in 1846, but it was not until the mid-1870s that the U.S. Army Signal Service (predecessor to the U.S. Weather Bureau, which began in 1890 under the U.S. Department of Agriculture) began using these cup anemometers to record wind speed data. These early anemometers were constructed with four cups that were hemispherical in shape (i.e., shaped like half of a ball) and had smooth edges. Over the early decades of the twentieth century scientists learned that three cups shaped like cones and possessing beaded (rounded) edges on the windward face of the cups dramatically improved the accuracy of cup-derived wind measurements. Just to give the reader some sense of the inaccuracy of the four-cup anemometer speeds, they were about 15% too high at 20 mph, 20% too high at 50 mph, and 25% at 100 mph. In comparison the three-cup speeds were <1% too high at 20 mph, 2% too high at 50 mph, and <5% at 100 mph.

The above information becomes more important when one learns that the vast majority of the wind speeds from the four-cup and early three-cup anemometers were *not* corrected prior to entry on the official observation forms until January 1, 1932. Beginning on that date speeds were corrected prior to entry on the forms. These corrections continued until the 1950s when the accuracy of the official three-cup anemometer (Model F-420C) became sufficiently accurate that corrections were no longer needed.

The three-cup, conical-shaped, beaded-edge design of the F-420C was carried through to the advent of the Automated Surface Observing System (ASOS) in the 1990s. However, the method of measuring the rotational speed of the cups was changed from an analog method to a digital. We will revisit this important issue later.

Despite the fact that electricity was used as far back as the 1870s to assist in the recording of wind speed and direction, the technology of the time still did not provide the instantaneous measurement of wind speed. At that time wind speeds were determined by counting (manually and/or graphically) the number of miles or tenths of miles per unit of time of the anemometer as indicated by the number of marks drawn on an autographic chart or the number of times an indicator light was illuminated and counted by the weather observer. This type of measurement is essentially the same as reading the miles or tenths of miles off the odometer of an automobile. If one measures the time required to travel a mile, he or she will then be able to compute the average speed for that mile. However, there is no way to know what the maximum speed was during the mile. Therefore, there was no way to obtain a maximum "instantaneous speed" without a continuous direct read out of the speed (i.e., a speedometer in the above analogy). It was not until the 1950s that instantaneous "peak gust" wind values began to be recorded by the military and not until the 1960s to 1970s for the U.S. Weather Bureau. These direct reading anemometers used a small electric generator (also known as a magneto) that produced an electric current that is linearly related to the rotational speed of the cups of the anemometer and thus the wind speed. The generated electrical current was displayed on a wind dial or strip chart recorder that was located in the weather office.

Operational requirements of the aviation community were influential in bringing about the measurement of these shorter duration wind speeds, and the resulting recorded values were of a great value to the engineering community for determining building and construction standards as well as the analysis of post-storm-related damage or structural failure. Prior to the advent and use of these direct reading (speedometer) type of anemometers the shortest duration (odometer) wind speed value that was recorded and published was the "fastest mile." The observer usually derived these values at the end of the observational day by finding the two closest mile marks on the single or triple register autographic recorder and then enter that value on the observation form. It has probably already occurred to the reader that a "mile" is a distance and not a duration. Obviously, the time that it takes the wind or a car (to continue our analogy above) to go 1 mile will vary greatly with the speed. (For an average speed of 30 mph, 1 mile will take 2 min; at 60 mph 1 min; and at 120 mph only 30 s.) The extreme wind databases contain not only these fastest mile values but also "fastest one-minute" values. These one-minute values were introduced randomly in time during the 1970s and 1980s at the National Weather Service stations as the old triple-register/multiregister recorders broke down. Whenever these instruments ceased operating, there was no longer any standard method by which the fastest mile could be observed or recorded, therefore the NWS instructed the weather observers to observe the wind dial (speedometer) for 1 min and record the value both for the regular hourly observation as well as the "fastest observed one-minute" for the day. Assuming an observer could accurately integrate the movement of the wind speed dial for a

1-min period, the derived values would NOT be comparable (for climatic or engineering use) to the earlier discussed fastest mile values. Remember that 1 min is a unit of time while 1 mile is a unit of distance and the only speed where the two values are equal is 60 mph. Let's go back to our earlier example of 30, 60, and 120 mph. A true "one-minute" wind of 30 mph is one-half the time of a fastest mile of 30 mph, an equal time at 60 mph, and twice the time of a fastest mile of 120 mph. One other reason why the fastest observed one-minute and fastest mile values are not comparable without statistical adjustment is the fact that the fastest mile value was taken from a recorder chart that would include every mile of wind that passed the anemometer during the day while the fastest observed one-minute value only represented the observed wind speed for one minute out of each hour. Obviously, the true peak one-minute wind would likely not be recorded under these circumstances.

During the 1970s and 1980s the NWS began installing strip chart "gust recorders" that constantly recorded the wind speed and allowed the observer to see the near instantaneous speed of the wind. This made it easier for the observer to make a more accurate estimation of a one-minute wind speed but was still difficult during periods of strong or gusty winds.

However, these gust recorders did provide the accurate measurement of wind "gusts," i.e., short duration wind speeds of 1 to 3 s depending upon the speed and gustiness of the wind. As was previously mentioned, these short duration values were important for aviation, structural design, and forensic use. For instance, the maximum dynamic pressure of the wind upon an average size home can occur in as little as 2 to 3 s. Smaller structures or objects can be affected by even shorter duration gusts.

Starting in 1992 the NWS and later the Federal Aviation Administration (FAA) and U.S. Military Services began installing automated weather observing platforms (ASOS/AWOS). Even though most of the platforms use a three-cup anemometer that is nearly identical to the older F-420 series used by the NWS, the internal workings of the anemometer are different in that a "light-chopper" device sends digital pulses to the computer, which samples the wind speed every second and builds a temporary storage of five (1 s) wind speed values and then computes a 5-s average that is stored for the observational day to be transmitted in real-time reports as well as the now "peak 5-s" wind for the day that is passed on to the climate databases and publications.

A 5-s gust differs significantly from the 1- to 3-s gusts that were recorded on the old analog strip chart gust recorders when you realize that the 5-s digital value is truncated. That is because the 5-s sum and resulting computer average is not recomputed with each new 1-s value. Rather the ASOS wind algorithm ingests five values and computes an average and then repeats the process 5 s later. The net effect of this logic dampens the magnitude of a true 5-s wind gust if computed with a running 1-s update.

Unfortunately, for data users the difference between the old analog gust values and the newer digital ASOS records were further complicated by the discovery that the first 3 or 4 years of ASOS-derived 5-s gust values were actually generated by the manufacturer's "test" algorithm that produced wind speed values that were lower than the correct running 5-s means. The erroneously determined values remain in the official databases and publications without adjustment or documentation as to when the correct algorithm was installed.

In addition to the previously discussed changes in wind speed observations, there is one other significant change that coincided with the implementation of ASOS and that was a doubling of the averaging period from 1 to 2 min for the regular hourly wind speed observations.

The final issue to be discussed here in relation to wind speed observations is not as significant as those previously discussed but nevertheless may show up in the reader's research or analysis. This issue deals with the units of measure.

Historically, wind speed measurements in the United States have been in miles per hour (mph); however, with the increasing influence of the military services on civilian weather observations during and after World War II, the knot (nautical mile of 6076 ft or 1.1508 statute miles) was eventually adopted by the U.S. Weather Bureau as the standard unit for wind speed in the mid-1950s. However, for public consumption, miles per hour were still used both in terms of nonmarine forecasts as well as historic publications like the *Local Climatological Data* (LCD) published by the NCDC. The NCDC digital databases and publications contain wind speed values in units of mph and/or knots.

ASOS wind observations are also recorded and disseminated to the aviation community in knots, but disseminated to the general public and published by NCDC in the daily summaries of the LCD in mph.

The reason for raising this issue is that in the arithmetic conversion (1.1508 × knots) of a whole knot value to a whole mph value causes certain mph values to be omitted. For example, 3 knots × 1.1508 = 3.45 mph when rounded becomes 3 mph while 4 knots × 1.1508 = 4.60 mph rounded becomes 5 mph and thus the value of 4 mph never shows up.

The same is true for the mph values of 11, 19, 27, 34, 42, etc., thus undermining wind speed frequency distributions of 1 mph classes.

6 WEATHER

Changes in equipment and observational practices are not restricted to those discussed above; rather, this is probably just the tip of the proverbial iceberg. The transition from human to automated determination of sky cover and visibility has brought its own set of discontinuities. In terms of sky cover the human observer viewed the whole sky while the automated (ASOS/AWOS) systems use a fixed laser beam and depend on the sky/clouds to move past them to be able to integrate a sky condition. This integration process works well most of the time, but still is limited to a ceiling height of 12,000 ft above ground level (agl) and thus does not indicate the presence of any high-level altocumulus, altostratus, or cirrus clouds. As for visibility, it is determined from an approximately 3-ft baseline at one location and is limited to a maximum reported distance of 10 miles. Therefore, any historic visibility studies that involved values greater than 10 miles would end at the time ASOS was commissioned at each observation site.

One of the other data losses that resulted from ASOS was that of the duration and intensity of thunderstorms. Work continues on the improvement and installation of a lightning sensor from which proximate thunder may be inferred.

This writer would be remiss if he did not mention one other major discontinuity to climatological studies and other retrospective uses of meteorological data. That is the conversion (on July 1, 1996) from the long-standing "airways" observational codes to a modified version of the European "METAR" code. This transition brought about numerous changes, but the scope of this chapter does not allow a full treatment of all these changes. Hopefully, the listing of a few will cause the reader to dig further before using any related data.

1. Ambient and dew-point temperatures are now recorded in Celsius (degrees and tenths) instead of whole degrees Fahrenheit.
2. Sky cover is now reported in oktas (eighths) instead of tenths.
3. Reported weather types of which there used to be about 28 have been reduced in number and altered in their two-letter abbreviations. Following are a few examples of the better known ones:

Weather Type	Airways	Metar
Hail	A	GR
Rain	R	RA
Ice pellets	IP	PE
Snow	S	SN
Drizzle	L	DZ
Fog	F	FG
Thunderstorm	T	TS

4. Rounding of negative temperatures changed from arithmetically absolute (i.e., -1.5° became -2°) to temperature absolute (i.e., -1.5° now becomes -1.0°). This change most often affects the computation of heating degree days (HDD) when the sum of the maximum and minimum temperature are odd and, when divided by 2 to obtain the mean temperature for the day, leave a product of whole units and tenths.

For example, Max Temp of -2° + Min Temp of $-9^\circ = -11^\circ \div 2 = -5.5^\circ$, which used to become -6° but now becomes -5°

7 SUMMARY

It is the strong recommendation of the author that anyone using meteorological and climatological data from any source investigate its background prior to analysis for research, publication, or use in the applied and forensic fields. Not to do so is an abrogation of scientific integrity. Hopefully, the discussion of this chapter will cause the reader to pause before blindly accepting the validity of any data set or value.

BIBLIOGRAPHY

Doesken, N. J., and A. Judson, (1997). *The Snow Booklet*, Colorado State University, 87 pp.

Engelbrecht, H. H., and G. N. Brancato (1959). World record one minute rainfall at Unionville, MD, *Mon. Wea. Rev.*, **87**, (8), 303–306.

Fujita, T. T. (1992). U.S. Department of Commerce, National Oceanic and Atmospheric Administration, NCDC, Asheville, NC, *Storm Data*, September 1992, Vol. 34, No. 9, p. 27.

Karl, T. R., and R. G. Quayle (1988). Climate change in fact and in theory; Are we collecting the facts? *Climate Change* **13**, 15–17.

Karl, T. R. (1995). *Long-Term Climate Monitoring by the Global Climate Observing System*. Kluwer Academic Publishers, pp. 51–56.

Schmidlin, T. W. (1990). A critique of the climate record of water equivalent of snow on the ground in the United States. *J. Appl., Meteor.* **29**, 1136–1141.

Schmidlin, T. W. et al. (1992). Design ground snow loads for Ohio. *J. Appl. Meteor.* **31**, 622–627.

Sherlock, R. H. (1947). Gust factors for the design of buildings, International Assoc. For Bridge and Structural Engineering, Vol. 8, pp. 207–235.

Sissenwine, et al. (1973). Extreme wind speeds, gustiness, and variations with height for MIL-STD 210B. U.S. Air Force, Cambridge Research Labs. TR-73-0560, p. 12.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1985). Storm Data; October 1985, Vol. 27, No. 10. National Climatic Data Center, Asheville, NC, 40 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1992). ASOS Users Guide; National Weather Service, Silver Spring, MD, 98 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1993). Climatological Data, New York; March 1993, Vol. 105, No. 3, National Climatic Data Center, Asheville, NC, 40 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1993). Climatological Data, North Carolina; March 1993, Vol. 98, No. 3, National Climatic Data Center, Asheville, NC, 36 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1994). ASOS Progress Report—ASOS Wind Sensor, Problems Identified and Solved. National Weather Service, Silver Spring, MD, 8 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1994). Natural Disaster Survey Report, “Superstorm of March 1993.” National Weather Service, Silver Spring, MD, 152 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1996). National Weather Service, Observing Handbook No. 7—Surface Weather Observations and Reports, July 1996. National Weather Service, Silver Spring, MD, 461 pp.

U.S. Department of Commerce, Weather Bureau (1957). History of Observational Instructions as Applied to Temperature Recordings. U.S. Government Printing Office, Washington, D.C., 7 pp.

U.S. Department of Commerce, Weather Bureau (1958). Excessive Precipitation Techniques. U.S. Government Printing Office, Washington, D.C., 12 pp.

U.S. Department of Commerce, Weather Bureau (1963). History of Weather Bureau Precipitation Measurements. U.S. Government Printing Office, Washington, D.C., 19 pp.

U.S. Department of Commerce, Weather Bureau (1963). History of Weather Bureau Wind Measurements. U.S. Government Printing Office, Washington, D.C., 68 pp.

U.S. Department of Commerce, Weather Bureau (1968). Final Report—Test and Evaluation of the Fischer and Porter Precipitation Gage. U.S. Government Printing Office, Washington, D.C., 28 pp.

U.S. Department of Commerce, Weather Bureau (1969). Specification No. 450. 1016 for Liquid in Glass Thermometers. U.S. Government Printing Office, Washington, D.C., 9 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1985). Climatological Data, Puerto Rico and Virgin Islands; October 1985, Vol. 31, No. 10. National Climatic Data Center, Asheville, NC, 23 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration (1985). Hourly Precipitation Data, Puerto Rico and Virgin Islands; October 1985, Vol. 15, No. 10. National Climatic Data Center, Asheville, NC, 15 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC continues to publish monthly and annual Local Climatological Data volumes that began in 1948. They are published for about 270 cities in the United States and some of its territories.

CHAPTER 42

DEMANDS OF FORENSIC METEOROLOGY

W. H. HAGGARD

1 SEARCH FOR THE TRUTH

The role of the forensic meteorologist is to assist the court in the search for the truth by presenting the most accurate description possible of the meteorological events pertinent to the case in litigation. To fulfill this role, the forensic meteorologist must utilize the best possible data set in his or her retrospective reconstruction of the weather. The quality of the analysis is dependent on the availability of representative measurements of meteorological parameters.

2 DATA NEEDED

The weather event reconstructions frequently require data and information beyond the standardized meteorological observations collected by governmental agencies. These frequently include eyewitness statements, field investigations, interviews, and data from other fields (e.g., recordings of radar-deduced tracks of aircraft, ship log books, air traffic control voice transcripts, and the like).

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

Reliability

Legal tests of the courtroom admissibility of expert testimony have been devised at various times. In 1923 the *Frye test* required that scientific evidence be of a type that is "generally accepted" in the relevant scientific community. In the late 1990s the *Daubert test* to eliminate "junk science" in the courtroom requires the proffered scientific evidence to be "reliable and based upon scientific methodology."

Hindsight

Several factors impact on the reliability of the results of a retrospective meteorological analysis. There must be data available upon which to rely. The data should be representative of the conditions of the atmosphere at the time and place of measurement. The instruments, sensors, observers, radars, satellites, etc. should have the capability of accurately sensing the requisite data. They should be properly calibrated, and their clocks should be set correctly. Since forensic meteorological analyses for litigation involve retrospective reconstructions, the data utilized come from the archives, and high quality of archived data is vital to proper reconstruction.

The forensic meteorologist enjoys the luxury of hindsight—usually believed to be "better" (or at least easier) to apply correctly than foresight. Like the forecaster, however, he or she is limited in analytical capability by the limitations in availability, representativeness, and quality of the available measurements pertinent to the circumstances.

Reconstructing the bases of the cloud above a 2900-ft mountain ridge between Nome and Koyuk, Alaska,[1] when there were no cloud base data within 100 miles of the crash site is more difficult than reconstructing the cloud bases at Flagstaff, Arizona[2] (where a continuously recording laser beam ceilometer record was available) (see Fig. 1).

Expert Testimony

Meteorology comes into the courtroom through the testimony of expert witnesses who are there because they possess an accumulation of education, knowledge, and experience that qualifies them to provide weather-related testimony beyond the normal capability of the court to obtain by other means. The qualified practitioner of forensic meteorology is a formally educated atmospheric scientist, often with specialized training in applied climatology, hydrometeorology, micrometeorology, aviation meteorology, marine meteorology, and/or disaster preparedness who is relied upon to assist legal counsel in clarifying scientific or technical issues that relate to the weather relevant to specific cases being litigated (Falconer and Haggard, 1990).

Forensic meteorologists do more than a "simple" weather event reconstruction (Haggard, 1980a,c, 1983, 1985, 1989; Haggard and Smutz, 1994). They may:

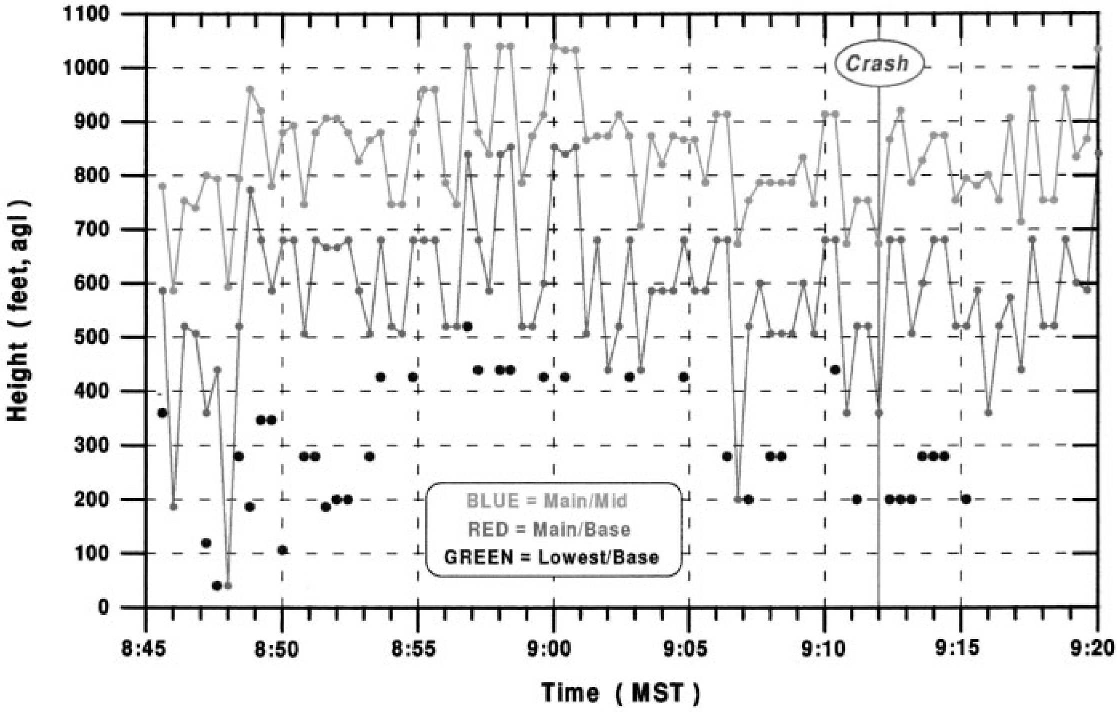


Figure 1 Decoded laser beam ceilometer data, Flagstaff, Arizona, 8:45–9:20 a.m. MST, December 20, 1991, showing values converted to heights above ground (agl) of lowest points of reflectivity (green) above ground; height of the base of main reflectivity (red); and midpoint of recorded main reflectivity (blue). See ftp site for color image.

- Acquire and interpret basic weather records (i.e., observations, operational charts and analyses, archived forecasts, or advisories, etc.).
- Assess the validity, representativeness, pertinence, and adequacy of the available data.
- Advise engaging counsel on the meaning and applicability of these basic weather data to issues pending before the court.
- Perform specialized meteorological analyses based on weather and other records (e.g., topographic surveys, eyewitness reports, damage photos, etc.).
- Prepare a comprehensive report.
- Provide credible expert testimony as to the weather pertinent to the matters under litigation.

High-quality records of measurements are essential to the accomplishment of these functions.

3 CASE EXAMPLES

Many retrospective reconstructions are for a site- and time-specific weather event (e.g., the collision of the *Summit Venture* with the Skyway Bridge, which spans the entrance to Tampa Bay, on May 9, 1980,[3] or the crash of a "TriStar" airliner on approach to Runway 17L at Dallas–Ft. Worth Airport at 6:04 pm CDT on August 2, 1985.[4] Others may involve a moderate span of time and an area (e.g., devastating floods over 16 counties of western North Carolina on November 5–6, 1977,[5] affecting thousands of square miles and several days). A few will be quite dependent on remotely sensed data (e.g., a microburst at Mayaguez, Puerto Rico[6]).

Some may involve a climatological study of decades of data (e.g., a wind study relating to the legality of constructing a N–S runway at Anchorage International Airport as a "crosswind runway"—a study involving detailed analysis of 23 years of hourly wind data[7]). The availability, adequateness, representativeness, pertinence, and quality control of the data impacted the credibility of the analytical results in these exemplar cases.

Visibility

Key questions in the litigation over the *Summit Venture*/Skyway Bridge collision (Haggard, 1985) related to the visibility in the ship channel and speed of the vessel, just prior to the collision with the bridge, far removed from any official weather-reporting site. Rain was so intense the bow of the ship could not be seen from the bridge.

Probably the most pertinent objective weather measurements relative to determination of the visibility were the rainfall rates at the ship as it traversed the channel. The National Weather Service Weather Surveillance Radar (WSR-57) located at Ruskin, Florida, 14 miles from the collision was in a favorable location to provide such data.

The collision occurred at 7:34 a.m. EDT. The radar at Ruskin was struck by lightning at 7:13 a.m. EDT, became inoperable, and did not provide further measurements until 7:43 a.m. EDT. The most potentially valuable and pertinent measurements were not available for 21 min before and 9 min after the event.

The meteorological system producing the intense rain—which reduced visibility to tens of feet—was sufficiently time consistent that is was possible to interpolate the "probable radar image" at the collision time by "bridging the gap" from the series of images prior to and subsequent to the radar outage. The resultant hypothetical image showed the ship and bridge channel spans to be within the area of the strongest reconstructed radar reflectivity, consistent with nonobjective lay eyewitness reports.

Microbursts

The principal litigation resulting from the crash of Delta Air Lines Flight 191 involved a 14-month trial in Federal District Court in Fort Worth Texas. Dr. T. Theodore Fujita (1986) performed an intensive retrospective weather analysis that relied on the combined meteorological measurements made by standard instrumentation, the 7-second LLWAS [low-level wind (shear)–alert system] data at sensors near and on the airport, weather satellite and radar imagery, the parametric data measured by the digital flight data recorder on the accident aircraft, and on-site physical evidence. This artful combination of all available pertinent data permitted a reconstruction of the interaction of aircraft and the atmosphere and led to a detailed reconstruction of the dynamics of the microburst that the aircraft penetrated.

Another microburst-related accident occurred at Mayaguez, on the western coast of Puerto Rico, on June 7, 1992, killing crew and passengers of a commuter plane attempting to land. The detailed weather reconstruction (Fujita et al., 1992) was largely dependent on a sequential analysis of satellite and radar measurements made remotely. The growth and collapse of thunderstorms augmented by the topography and sea breezes of the island were remotely sensed. The results of analyses of those measurements were integrated with multiple witness statements from both airborne and ground-based persons closer to the event. Finally the dynamics of the aircraft and its track were fitted into the meteorological analysis to locate the point of impact of the microburst relative to the trajectory of the plane immediately prior to impact (see Fig. 2). The coordinated use of remotely sensed meteorological measurements, their augmentation by anecdotal local witness statements, and their combination with tracked and timed aircraft positions were essential to the reconstruction of the weather scenario.

Winds

Measurement validity was an element of dispute in the litigation over construction of the "crosswind" north–south runway at Anchorage International Airport (Haggard, 1980b). The Federal Aviation Administration (FAA) conducted a study of winds relative to the existing east–west runway. Though 23 years of wind data existed, they

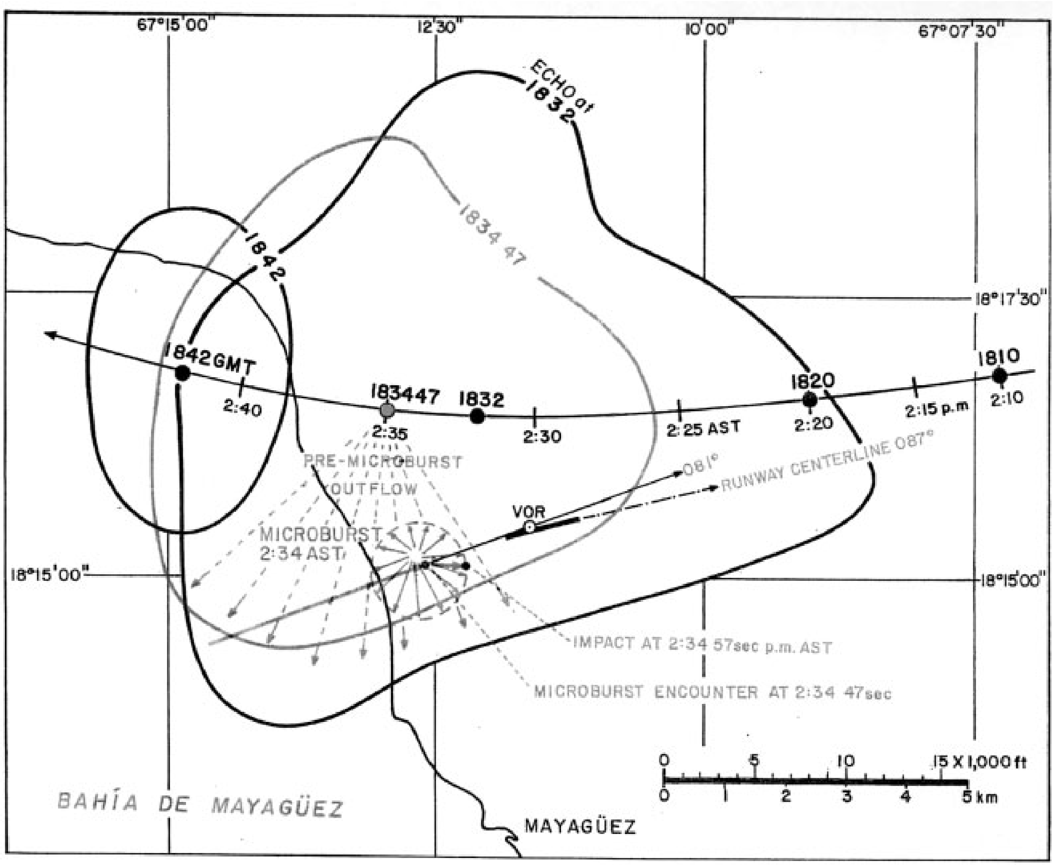


Figure 2 Motion and collapse of San Juan observed radar echo over Mayaguez, Puerto Rico, on June 7, 1992. Curved black line shows center of reflectivity at 1810, 1820, 1832, and 1842 UTC; echo outlines at 1832, and 1842 are shown in black; interpolated echo outline at 1834:47 is shown in blue; aircraft track is shown in red; and probable microburst encounter is depicted in blue (times are shown in Atlantic Standard Time = AST; AST = UTC minus 4 h, such that 1820 UTC = 1420 AST = 2:20 pm AST). See ftp site for color image.

were not compatible. Some were observed at 16 compass points, others at 36 compass directions; some were recorded in miles per hour, others in knots; some were digitized at hourly intervals, others only at 3-h times; the "official wind observations" were made at 12 different consecutive sites over the 23 years of available "on airport" measurements. The life span of the various sites ranged from 11 months to 10 years, averaging 5.2 years.

In its design study, the FAA used the latest 7 years of compatible 36 point, 3-h wind data, which yielded a qualifying (for cross-wind runway construction) cross wind of 15 knots or greater 5.88% of the time (5.0% is required for a cross-wind runway).

An affected property owner had a 17-year wind study made and determined the qualifying cross winds were apparently present less than 5% of the time. A federal judge stopped the $33 million construction project and ordered a wind study utilizing "all available data."

In the undertaking of this complex task, utilizing separate analyses for the differing periods of noncompatible data units, it was discovered that all data from one anemometer location (3/21/61 to 1/17/64) yielded strikingly low cross-wind components compared to prior and subsequent data. Wind speeds were compatible with earlier and later records, but the direction appeared anomalous. A statistical (F) test showed the data from that nearly 3-year sample to be statistically from a different population than the remaining data.

Adjusting the wind data of this sample by 30° (the difference between true and magnetic north at Anchorage) removed the statistical "bias" and made the sample compatible with the prior and subsequent data, suggesting the wind vane directional calibration might have been based on magnetic rather than true north during that sampling period.

Clouds

In aviation meteorology, the *sky condition* is a key element, determining whether aircraft operations may be conducted under visual flight rules (VFR) or instrument flight rules (IFR). While most scheduled commercial aviation flights are conducted under IFR, regardless of the observed conditions, a large percentage of "general aviation" flights are conducted under VFR conditions, and many pilots are not "IFR qualified and equipped." To them the accuracy of the "measurement" of sky condition is vital for a determination of whether they can legally operate their aircraft under visual flight rules.

The measurement of sky condition has historically been made by human observers viewing the sky and recording the percentage (in oktas or tenths) of the sky obscured by clouds. Recently, this task has been increasingly taken over by the Automated Surface Observing System (ASOS), vertically scanning laser ceilometer beams that substitute percentage of time clouds are sensed at various heights in the atmosphere below 12,000 ft above the ground level for the areal estimates formerly made by human observers.

Recent litigation[8] resulting from a six-death crash of a private plane attempting a nighttime VFR landing at a Florida airport hinged on whether the sky condition "measured" by a human observer was "measured 600 foot broken clouds" (covering 6/10ths of the sky with opaque clouds), technically IFR conditions, or whether these clouds had moved away, allowing the field to become VFR.

The case was complicated by a number of conflicting eyewitness postaccident statements regarding the sky condition at the airport at the time of the accident. Two testifying forensic meteorologists offered very differing opinions in testimony before the court during trial. Statements by two IFR-rated pilots who "lost sight of the airport on arrival" and "encountered an unreported cloud layer on departure" near the accident time were significant in the retrospective weather reconstruction.

Conflicting Views

Another aviation litigation case in which the testimony of two eminently qualified forensic meteorologists offered dramatically different opinions on the atmospheric conditions involved a tragic in-flight breakup of a twin-engine aircraft over Anthony, Kansas[9] (Haggard, 1983).

Far removed from any direct measurements, the question was whether:

1. The plane was in level flight between cloud layers (perhaps in light snow) with no wind shear, no turbulence, and no icing when both engines and various control surfaces broke off the aircraft; or
2. The plane flew through a strong wind shear zone with a dramatic weather change into freezing rain and severe turbulence, resulting in an upset and overstresses, which led to the in-flight breakup.

Both experts utilized the same archived weather measurements. One relied upon a computer analysis that indicated strong rotation but minimal shear in the wind field at flight altitude. The other relied on a "hand analysis" (not dependent on computer-programmed analysis procedures) that showed less rotation but a very strong shear zone in the wind field.

Changing Environment

The validity of weather measurements may be degraded by changing environmental circumstances. A recent tragic example was the failure of the LLWAS at Charlotte, North Carolina, to detect and warn of the small but strong microburst associated with the crash of USAir Flight 1016 on the evening of July 2, 1994 (Fujita and Haggard, 1995; Haggard and Fujita, 1994).

Installed in 1981, the six sensors of the system were mounted on poles at heights ranging from 20 to 68 ft on and near the Charlotte, North Carolina, Douglas International Airport in rolling terrain largely covered by early-growth pine. Between 1981 and 1994 many of the pines grew to heights equal to or greater than the heights

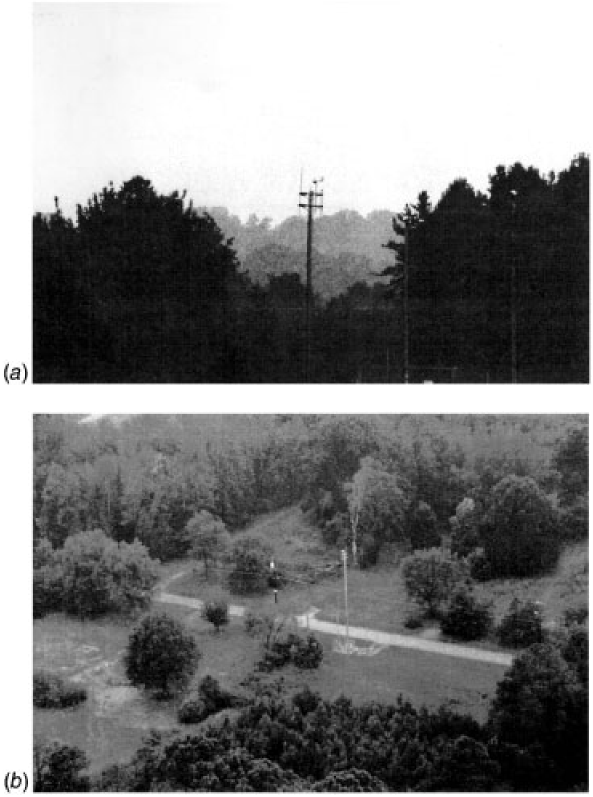


Figure 3 LLWAS anemometers at Charlotte, North Carolina, in 1994. Upper photo is of northwest sensor; lower photo is of southeast sensor; each shows sensors in relation to forest growth. See ftp site for color image.

of the wind sensors (see Fig. 3), degrading their wind measurements (and their ability to sense and warn of significant wind shear).

4 WITNESS CREDIBILITY

In these (and other) examples of weather-related litigation, the court (judge and/or jury) must consider the "testimony and demeanor" of each witness to determine their credibility. Credibility is enhanced by demonstrated reliance on valid meteorological measurements and sound scientific procedures.

The adversarial system of justice in the United States should not impact on the expert witness. The parties and the attorneys are adversaries and advocates. The expert witnesses should be objective scientists utilizing valid measurements and objective techniques. Procedures for bringing scientific facts forth in the courtroom are very different than in a scientific symposium (Bradley, 1989).

Decisions by various courts suggest (Haggard, 1989) that it is essential that the testifying expert:

- Reduce his or her opinion to the simplest terms possible.
- State (and visually illustrate) the opinion(s) clearly and concisely.
- Demonstrate (visually and orally) the factual basis for the opinion.
- Produce authenticated copies of all weather measurements relied upon.
- Demonstrate reliance upon the most accurately observed and adequately quality controlled meteorological measurements available.
- Ensure the bridge from demonstrable facts to the stated opinion is as short and solid as possible (i.e., "let the data speak for themselves").

The meteorological expert must avoid any appearance of advocacy (Saks, 1987) and adhere to the scientific analysis of the highest quality validated meteorological measurements available (Haggard, 1985).

CASE CITATIONS

1. *Kavairlook v. Ryan Air Service*, Superior Court for the State of Alaska (Nome-Koyuk 12/10/94).
2. *Kaufman v. Beech Aircraft*, Superior Court of the State of California for the County of Los Angeles (Flagstaff, AZ, 12/20/91).
3. *M/V Summit Venture*, United States District Court Middle District of Florida Tampa Division (Sunshine Skyway Bridge, Tampa, FL, 5/9/80).
4. *Kathleen Connors et al. v. USA*, United States District Court for the Northern District of Texas, Ft. Worth Division (DL 191, DFW, TX, 8/2/85).
5. *American Enka v. Southern Railroad*, United States District Court for the North Carolina District of Asheville, NC (Enka, NC, 11/7/7).
6. *Leslie et al. v. American Airlines et al.*, United States District Court for the District of Puerto Rico (Mayaguez, PR 6/7/92).
7. *John Overby v. USA*, U.S. District Court for the District of Alaska (N-S runway at Anchorage International Airport).
8. *McNair v. USA*, United States District Court Northern District of Florida Gainesville Division (Gainesville, FL, 6/7/95).
9. *Buzzard v. Piper*, District Court for Oklahoma County, Oklahoma City, OK (Anthony, KS, 2/11/77).

REFERENCES

Bradley, M. D. (1983). The Scientific and Engineer in Court, Am. Geophys. Union Water Resources Monograph 8, Washington, DC.

Falconer, P. D., and W. H. Haggard (1990): *Forensic Meteorology, Forensic Sciences*, New York, Matthew Bender, Chapter 35.

Fujita, T. T. (1986). DFW Microburst on August 2, 1985, SMRP Res. Paper No. 217; Chicago, University of Chicago Press.

Fujita, T. T., W. H. Haggard, and W. A. Bohan (1992). Puerto Rico's Weather of June 7, 1992 Related to the Crash of Executive Air Flight 5456 at Mayaguez, Puerto Rico, submission by Construcciones Aeronauticus, SA [CASA] to the National Transportation Safety Board, Washington, DC, November.

Fujita, T. T., and W. H. Haggard (1995). The Microburst at Charlotte, North Carolina in Relation to the Crash of USAir 1016 on July 2, 1994, A Report to The Air Line Pilots Association ALPA, January.

Haggard, W. H. (1980a). Radar Imagery for Past Analysis of Mountain Valley Flash Floods, Proc. 2nd Conf. on Flash Floods, Atlanta, GA, Boston, American Meteorological Society.

Haggard, W. H. (1980b). What's in a Wind Study? Proc. 2nd Joint Conf. on Industrial Meteorology, New Orleans, LA, Boston, American Meteorological Society.

Haggard, W. H. (1980c). Some Micro-Economic Aspect of Applied Climatology, Proc. Conf. on Climatic Aspects and Social Responses, Milwaukee, WIs, Boston, American Meteorological Society.

Haggard, W. H. (1983). Weather after the Event, Proc. 9th Conf. on Aerospace and Aeronautical Meteorology, Omaha, NE, Boston, American Meteorological Society.

Haggard, W. H. (1985). Meteorologists as Expert Witnesses, Proc. 15th Conf. of Broadcast Meteorology, Honolulu, HI, Boston, American Meteorological Society.

Haggard, W. H. (1989). Weather Testimony in Litigation, Proc. 3rd Int. Conf. on the Aviation Weather System, Anaheim, CA, Boston, American Meteorological Society.

Haggard, W. H., and T. T. Fujita (1994). The LLWAS at Charlotte, North Carolina in Relation to the Microburst of July 2, 1994, A Report to The Air Line Pilots Association ALPA, September.

Haggard, W. H., and S. W. Smutz (1994). Forensic Meteorology; Proc of 7th Aviation Law/Insurance Symposium, Daytona Beach, Fl, Embry Riddle Aeronautical University.

Saks, M. J. (1987). *MIT Tech. Rev.*, Aug/Sept.

CHAPTER 43

SURFACE LAYER IN SITU OR SHORT-PATH MEASUREMENTS FOR ELECTRIC UTILITY OPERATIONS

ROBERT N. SWANSON

1 INTRODUCTION

Electric utility operations, as interpreted in the following discussion, involves day-to-day operations, engineering requirements for planning, designing, and operating a system, applied research, as well as environmental concerns and/or requirements. Several aspects of measurements are discussed elsewhere in this *Handbook* and will not be repeated in this chapter. Meteorological data sets, collected as part of any operational or research program, require detailed knowledge of sensor and data retrieval system characteristics and employ proper quality assurance/quality control (QA/QC) procedures. It also involves sensor siting guidelines to assure maximum information is obtained from the measurement program. Since utilities measurement needs may extend throughout and above Earth's boundary layer, their programs are not necessarily restricted to the use of in situ or short-path sensors.

2 RECENT HISTORY OF METEOROLOGICAL REQUIREMENTS BY ELECTRIC UTILITIES

Until the early 1950s, meteorological information used by electric utilities, if any, were those collected and transmitted by the federal government. Simple adjustments were sometimes made to these data to compensate for local extreme conditions for purposes of power generation or system line loading. Occasionally systems were

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

operated at above rated levels, and the situations were not considered serious until they impacted operations in some manner. Reliability of power delivered to the customer was not always an overriding concern with power outages generally tolerated and/or expected by the user. In the 1950s, when nuclear power plants were first being built on a commercial basis, there was a developing interest in the transport and dispersion of potential releases of nuclear material. This interest or concern resulted in the development of meteorological monitoring programs at both nuclear and conventional power plants. These monitoring programs were developed to provide inputs to existing dispersion models. Early monitoring programs were quite simplistic in that they frequently collected very limited information of questionable quality. For example, at Humbolt Bay Nuclear Plant in northern California, wind measurements at this nuclear plant consisted of a single aerovane mounted on top of a tower near the plant. Calibration and/or maintenance of this sensor was not well established and possibly was never done.

As time passed, new and more complex dispersion and transport models were developed that, in turn, required additional meteorological information. Because of this requirement, as well as because of an increased concern by the public over safety, more sophisticated and better designed monitoring programs were developed. After the belief that nuclear power would be both inexpensive and plentiful was challenged, there was an increased interest in alternate energy development and in energy conservation. Both of these programs had additional meteorological requirements beyond those provided by the National Weather Service (NWS) with the largest uncertainties associated with planned alternate energy projects. This situation was readily apparent because NWS sites were generally located at airports with a lesser number of observation sites in urban areas. Neither of these sites would be representative of wind energy farms that were being evaluated. Until the mid-1970s wind energy and/or photovoltaic power sources were given little or no attention by electric utilities. This lack of interest was a result of many things, including cost, equipment availability, equipment reliability, and, of course, the already existing power-generating network.

In general each of the alternate energy programs, including wind, photovoltaic, hydroelectric, or geothermal power, have their own particular meteorological monitoring requirements. Also increasing concerns over public health issues and conservation of natural resources have created many changes in meteorological monitoring needs after the early 1950s.

3 MEASUREMENT REQUIREMENTS

Meteorological measurement requirements for electric utility operations can normally be met with existing off-the-shelf sensors and digital recording systems. Siting of these sensors to best address specific concerns or needs of a project is likely a more difficult task than is proper sensor selection. It is important to consider that measurements collected for one program may well be integrated into another, possibly unrelated, program. Therefore consideration should be given to this possibility when selecting and/or siting sensors and in data processing and archiving incoming data. This consideration should in no way detract from the original goal of collecting the best possible data set for the specific project to be evaluated. Additional information generated in data processing and archiving should always be considered since it represents only a minimal increase on the overall level of effort and may well provide valuable information to be used with another project.

Measurement programs should be designed to meet as many of the objectives as possible within budget constraints. These designs can become very difficult at times. For example, when monitoring for wind energy or photovoltaic farms in complex terrain, consideration must be given to many factors including terrain elevation and contours, cloud cover changes, temperature differences, and winds. If the site(s) locations are near large water masses, additional issues may exist. Frequently discussions with long-time residents in the area will provide valuable insight into specific conditions that would otherwise take an extended monitoring period to determine.

4 MEASUREMENT GUIDELINES

As mentioned above, most measurement program needs for electric utility operations can be easily satisfied with existing equipment. However, needs arise to examine the atmosphere above the lower boundary layer. When these needs exist in situ and/or short-path sensor measurements will have to be augmented with additional sensors such as long-path remote sensors or acoustic sounders. One example of such a program is cloud seeding for snow pack enhancement in an attempt to increase hydroelectric power. For this effort the atmospheric structure must be well defined through a very deep layer. Siting off the non-in-situ sensors for use in special programs is project specific and therefore cannot be given detailed guidelines except to follow rigid QA/QC policies. In designing a measurement program, the primary consideration is to determine the overall objective(s) of the project and then design a total system that will best address all concerns within the budgetary constraints that are levied on the effort.

Effective measurement programs require adhering to good QA/QC guidelines. Examples of such guidelines are ANSI/ANS-3.11 (ANS, 2000b) and EPA (1989). Both of these documents provide guidelines that will, if followed, assure a reasonably successful monitoring program, but, as with most efforts, deviations from the guidelines may have to occur. One typical deviation is in actual sensor siting where it is frequently impossible to meet all guidelines on distances of the sensor from obstructions. In such cases common sense must prevail and the sensors should be located where the smallest negative impacts will be found.

5 DATA ACQUISITION

Digital data acquisition systems can provide totally adequate data collection, storage, and transmission packages because of their speed, accuracy, and overall reliability.

Backup devices, such as strip-chart recorders, can be included in a monitoring program, but they add a significant level of effort and cost to the program while providing little redeeming information not already available with properly programmed digital systems. All maintenance and calibration efforts must be done using technically competent personnel. Data retrieval, processing, checking, editing and archiving should also be patterned after an accepted guideline, such as ANSI/ANS-3.11, in a manner similar to programs for field measurements. Statistical procedures should be established, using incoming data, to assist in assessing data quality and operating condition of measurement devices. These statistical programs should be established at or near the beginning of a measurement program to capture diurnal and seasonal variations. It is critical that any deterioration of sensor performance be detected and corrected as soon as possible to maintain a valid collection efficiency of 90% or better from all sensors on an annual basis. The definition of "valid" data capture must also be quantified to prevent all incoming values from becoming valid readings regardless of quality.

6 EXAMPLE OF METEOROLOGICAL REQUIREMENTS BY AN ELECTRIC UTILITY

A large electric utility in western United States, in the early 1990s, had a complete weather forecast office in addition to a significant sized group of personnel in a meteorological projects section. Many requests for information and/or studies done within the weather forecast office will not be discussed below as the following tasks represent only the major items requested of the projects section. This list is not intended to be all inclusive but is given only to illustrate the types of programs with which meteorology may become involved. Also it is not intended to reflect the meteorological research needs of other electric utilities since it is unique in size, type, and expanse of service territory and generation mix. Major study efforts by the projects section included:

1. *Electric load management and load research.* These are primarily regulatory issues because electric rates are determined by typical meteorological conditions within specific geographical areas. Routine measurements of temperature and humidity at about 25 locations throughout the utilities service area satisfied the requirements for this project.
2. *Transmission line loading (dynamic thermal rating).* This program consisted of wind and associated temperature measurements along major transmission lines. Critical conditions that may cause transmission line ratings to be exceeded occur with very low ventilation rates and high ambient temperatures. Wind speeds of interest, perpendicular to the power line, are in the 0.25 to 2.0 m/s range so the wind sensors must have a low starting threshold. Temperature measurements within about 1.0° of true are quite sufficient for this type of study. Since this program was implemented in the early 1990s,

new and probably better measurement systems for dynamic thermal rating have entered the market.

3. *Alternate energy system feasibility.* Studies in this program include those for photovoltaic, wind, and geothermal development. Photovoltaic and wind energy farm potentials are very site specific. Photovoltaic farm feasibility is dependent on distribution of cloud cover, evaporative cooling, atmospheric turbidity, and ambient temperature. Wind power is dependent on wind flow distribution in time and space, quality of wind such as wind shear and turbulence, time of day, and time of year. Geothermal power concerns are primarily transport and diffusion of released pollutants and of cooling tower siting and their operating efficiency. Each of these alternate energy programs had its own measurement requirements and each monitoring effort was unique. Wind tunnel modeling, covering a several square kilometer range, was completed for a potential wind energy farm site.
4. *Power line and transformer contamination.* This study centered around transmission lines, transformers, insulators, and substations where contaminant buildup can create equipment malfunction. Costs of washing these pieces of equipment are very high and a realistic washing schedule was desired. Conventional sensors for measuring winds, humidity, net radiation, and surface temperatures of transformers were used.
5. *Environmental impacts from released pollutants.* These studies were primarily done for planned and/or operational fossil fuel, nuclear, geothermal, and hydroelectric power plants. General concepts, techniques, and instrumentation needed for estimating pollutant transport and dispersion are discussed elsewhere in this *Handbook* and so will not be discussed here. However, most of the power plants are located in complex terrain where released pollutant transport and dispersion is modified by the plume's flow around terrain obstacles. Realistic estimates of environmental impacts from released material frequently required additional measurement sites along the plume's trajectory. Three-dimensional wind tunnel modeling of several of the geothermal sites were completed as part of a transport and dispersion study.
6. *Wind, ice, and snow loading on transmission towers and power lines.* Much of the information used in this study was derived from climatological data with a limited number of measurement sites in the field. Since this program was site or area specific, the measurement program was tailored for each area of concern. One major problem area in the implementation of this study was to employ sensors that would provide accurate data at times when icing or snowing conditions occurred. This problem was not adequately solved, but new systems such as use of line tension monitoring sensors, may be of considerable assistance in this type of study.
7. *Cloud seeding to increase snow pack for hydroelectric operations.* This effort involves three-dimensional wind speeds and directions, temperatures, and humidities from the seeding area to heights above clouds having seedable moisture. In situ measurements of wind and precipitation used in this study

must be capable of operating under adverse weather conditions where freezing rain, ice, and snow can be expected. Potential benefits of this study are very large in that, for a small expense, a considerable increase in water runoff from melting snow may occur. This water runoff is then used to generate hydroelectric power.

8. *Long-range transport of pollutants.* This study was initially designed to describe the fate of pollutants from one or more power plants as they were transported to downwind areas far removed from them. This study was co-sponsored by several governmental groups and had the mission of developing models for both air quality and acid deposition over distances of hundreds of kilometers. Meteorological sensors used in this project included many in situ and short-path sensors along with long-path sensors, upper air sensors, and a variety of aircraft measurements.

7 EXAMPLE OF DATA PROCESSING PROCEDURES

As part of the transmission line loading (dynamic thermal rating) effort, a simple prearchiving program was developed for in situ wind speed and direction measurements where inexpensive on-site digital recording systems were used. Data resulting from this program provide basic data sets that may be somewhat applicable to other project studies. In general the program was designed to collect, generate, store, and, on demand, transmit the following data for any desired time period:

1. Peak scalar speed
2. Time of peak speed
3. Unit vector wind direction at time of peak speed
4. Mean scalar wind speed over sampling period
5. Mean unit vector wind direction over sampling (period excludes observations of calm winds)
6. Mean resultant vector wind direction over sampling period
7. Standard deviation of wind speeds over sampling period
8. Standard deviation of unit vector wind directions over sampling period
9. Standard deviation of resultant vector wind directions over sampling period
10. Standard deviation of nonoverlapping 1-min mean speeds over sampling period
11. Standard deviation of nonoverlapping 1-min mean unit vector directions over sampling period

If the desired sampling period was longer than the basic time of 10 min, e.g., 1 h, then values for items 10 and 11 would be repeated for longer term means of 2, 5, or even 10 min. The purpose of generating the many standard deviations from averaged values was to detect the energy-containing frequencies within the lengths of record considered. More extensive or different analyses of incoming raw data can be easily programmed, but cost–benefit relationships as well as storage capacities and data QC must be considered. Again the example is intended only to be an illustration of a practical data processing approach and not a recommended guideline.

8 CONCLUSIONS

The above discussion of the recent history of meteorological support for an electric utility and of programs used within a selected utility are not complete but do serve as an indication of the importance of high-quality measurement programs as well as the need to match measurement requirements with study efforts. Hopefully, it also provides some insight into the need for making as complete as practical measurements for any given study so resulting data packages can be used in other studies. Other electric utilities in different areas of the country, or world, will undoubtedly have a different list of meteorologically related problems or concerns such as lightning protection and cooling water temperatures.

REFERENCES

ANS (2000a). American Nuclear Society, *Determining Meteorological Information at Nuclear Facilities*, La Grange Park, IL.

ANS (2000b). ANS/ANS-3.11-2000, American Nuclear Society, La Grange Park, IL.

U.S. EPA (1989). *Quality Assurance Handbook for Air Pollution Measurement Systems*. Vol. IV, *Meteorological Measurements*, prepared by Thomas J. Lockhart for the U.S. Environmental Protection Agency, Research Triangle Park, NC.

CHAPTER 44

INDEPENDENT AUDITING ASPECTS OF MEASUREMENT PROGRAMS

ROBERT A. BAXTER

There are a number of aspects that need consideration in the design and execution of measurement programs to assure the data collected are of documented quality and meet the program data quality goals. This chapter looks at the independent auditing aspects of monitoring as an integral tool to the overall data collection effort and provides examples of how independent audits contribute to the understanding of the quality of the data collected.

Prior to entering the role of audits, it is helpful to understand the details of monitoring programs and see where audits fit into the overall data collection scheme. The monitoring program and its goals are described in a monitoring plan.

1 MONITORING PLAN

A monitoring plan is a general description of the overall plan to collect data. A monitoring plan includes the monitoring program goals, methods of data collection, locations where data will be collected, internal and external checks of the measurement program, and the overall management and reporting structure. It is typically prepared well in advance of the collection of data so that management and/or regulatory review can identify and resolve any deficiencies in the measurement program.

Once the monitoring plan is agreed upon and ready to be implemented, a quality assurance project plan is prepared that addresses all of the details of the data collection program.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

2 QUALITY ASSURANCE PROJECT PLAN (QAPP)

A quality assurance project plan is a formal document describing in comprehensive detail the necessary activities that must be implemented to ensure that the results of the work performed will satisfy the stated performance criteria.

The QAPP defines all details of the program including the methods of data collection, the specific instrumentation used for collection of each variable, methods of calibration, data storage, backup, validation, and reporting. Also included is the audit plan to verify the implementation of the methods and procedures defined in the QAPP. The U.S. Environmental Protection Agency (EPA) provides guidance for the preparation of QAPPs for environmental monitoring programs (U.S. EPA, 1998). After reading the guidance and recognizing the recommended level of detail in a QAPP, one may ask the question "Are all of the checks and documentation really necessary? After all, I have been collecting data using a variety of measurement methods for many years." The answer is a qualified yes. The qualification steps back to what the data will be used for and how defendable one wants to make it. Backyard weather forecasting requires little in oversight and documentation. Permit applications and compliance on the other hand requires well-documented, defensible data.

One learns through years of experience that even the most competent professional or scientist can make mistakes through the redundant process of setup, data collection, and data validation. These mistakes may be minor but could have a detrimental effect on the quality of the collected data. In some cases, quality and the proper way to operate a system may be compromised because of budget limitations or lack of physical resources. It is extremely important to identify the instances when data has been compromised and then quantify the impact it has on the quality of the data reported. This will provide end users with records and information needed to assess whether the collected data will meet their needs.

A properly designed program will have cross checks built into the overall measurement plan. These checks will be defined in the QAPP in the sections on quality control and quality assurance.

Quality Control (QC) The overall system of technical activities that measures the attributes and performance of a process, item, or service against defined standards to verify that they meet the stated requirements established by the client.

Quality Assurance (QA) An integrated system of management activities involving planning, implementation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the client.

Quality control activities are designed to control the quality of a product so that it meets the user's needs. This includes the routine calibrations, data validation, preventive maintenance, equipment certification, etc. Quality assurance is the process whereby the implementation of the QC program and other activities is checked and verified. It encompasses the various activities needed to assure the QC program is being implemented and is working. QA includes QC as one of the activities needed to ensure that the product meets defined standards of quality. Details on quality assurance in meteorological measurements can be found in U.S. EPA (1995).

A key element in QA is the use of independent audits to review operations and make reports to management on the status of the measurement program. These audits are performed by an individual or group that is independent of those making the measurements. This independence allows the identification of potential problems without any conflict of interest the findings may have with either the technical merit of the data or financial resources needed to correct deficiencies.

Audit A systematic and independent examination to determine whether quality activities and related results comply with planned arrangements and whether these arrangements are implemented effectively and are suitable to achieve objectives.

In performing the audits, there are specific roles of the auditor and auditee.

Auditor A person that is qualified to perform audits.

Auditee Person or organization that is operating a measurement program and is the one being audited.

Ideally, the auditor will be an expert in the field of the measurements being performed. While not always true, the primary prerequisite is that the auditor understand the measurements being performed and be able to identify if the methods followed are consistent with the monitoring and quality assurance plans. The auditor is always a guest at the measurement site and should not perform any of the instrument removal or other activities that are part of the normal operation of the site. In past instances the auditor has helped in the removal of equipment and performance of duties for the operator and inevitably an instrument breaks or becomes inoperable that becomes attributable to the auditor.

3 CASE STUDIES

Given the above overview of the elements in the planning and oversight of a measurement program, it is useful to review some case studies to demonstrate the value in independent audits to support overall data collection and documentation. The case studies below describe field experiences in auditing some aspects of meteorological measurement programs.

In this first case study, a major measurement program was carried out in the western United States that included measurements made by federal, state, and local agencies as well as private contractors. The design of the program included an independent QA contractor responsible for system and performance auditing of the installation and operation of the surface and upper air meteorological systems. The surface measurements included wind measurements using conventional cup-

and-vane or propeller-vane anemometers. Some of the systems were part of existing measurement programs while the majority were installed specifically for this study.

The auditing plan for the study called for initial siting audits that assessed the appropriateness of the selected sites for the measurements. The results of these audits helped program management determine if changes were needed in the selected sites. Once final sites were selected and instruments installed, audits were performed on the surface meteorological systems. While there may have been several specific problems noted for each of the sites, there was a common equipment alignment problem noted with one of the measurement groups. This problem was present at most of their sites. The alignment and orientation of sensors have always presented challenges. While the most common method of orientation is to use the local magnetic declination to correct the alignment to true north, local anomalies in the magnetic field can create errors of 10° or more. With this particular measurements group (called group A), the number of sites set up and operated led them to streamline the alignment process using a hand-held "data scope." The scope allowed the direct entry of the local declination to the magnetic readings making the alignment checks quick and easy. A second group (called group B) used magnetic methods to determine the alignment but did not apply a specific declination. Instead, the magnetic readings were corrected to true north based on a measurement of the sun's azimuth angle and the calculated true angle of the sun for the site's latitude and longitude.

Figure 1 shows the alignment audit results from groups A and B for the surface sensors. In the overall program plan, the data quality objective for the wind direction

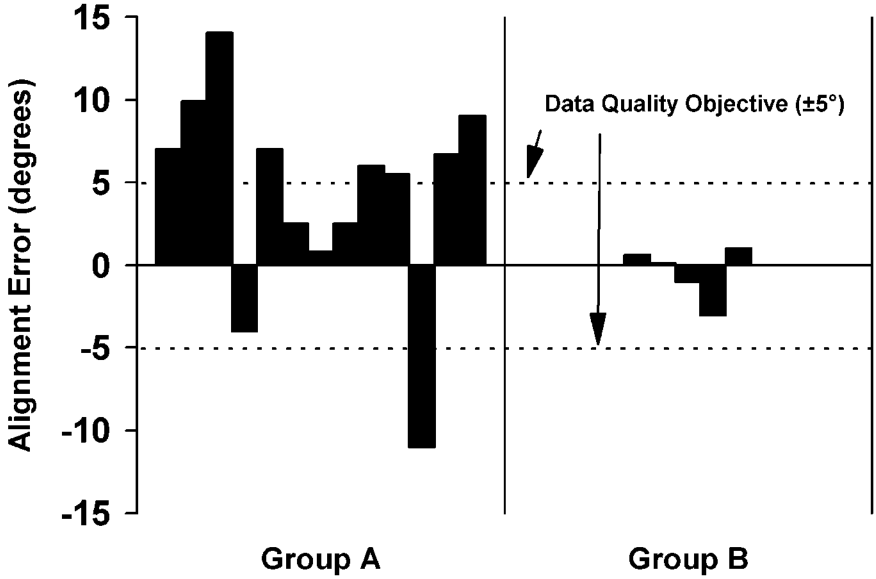


Figure 1 Surface wind direction sensor alignment results from two different measurement groups.

data was established as ±5°, as indicated in the figure. The results of the alignment audits showed group A had much more scatter in the alignment accuracy, which was a direct result of the method used for alignment. The problem was systematic throughout the sites set up by group A. Agreement was reached between the auditee and auditor on appropriate methods for alignment of the systems. The procedures used by the auditee were modified to obtain the needed accuracy. The audit in this case served as a valuable training and teaching exercise for group A performing the measurements and corrected a long-term systematic problem in aligning sensors.

In a second example, a station was audited that collected both air quality and meteorological data. Meteorological sensors were located on a 10-m tower adjacent to the trailer. This particular station demonstrated an all-too-familiar example of how the meteorological sensors generally take second place to the air quality measurements. Due to the relative complexity of the air quality instrumentation, much of the maintenance effort focused on those instruments. The meteorological sensors were set up, operation verified, and then the sensors left alone to collect data without any further checks. This is a common scenario that arises from the ability to visually look at the sensors, see them rotating in the wind and aiming in appropriate directions. This gives the impression that the equipment is operating acceptably and no further checks need to be performed. To the contrary, bearings in the wind sensors may fail, cups or propellers may be damaged, and potentiometers that convert the vane direction to an electrical signal may wear. Other sensors such as those used for temperature may become corroded and produce erroneous signals. The audit in this example looked at wind and temperature measurements and included a system and siting audit to determine the appropriateness of the site for the intended measurements.

Results of the audit showed the following:

1. The entire meteorological system had not been serviced since installation with no calibrations performed.
2. The temperature sensor had corroded into the radiation shield and could not be removed to verify its proper calibration.
3. The wind direction sensor had corroded into the mounting and the connector failed when removed from the sensor. This hampered the performance evaluation of the sensor.
4. The wind speed bearings had corroded to the point where the starting threshold of the sensor was almost 2 m/s.
5. The alignment of the wind direction sensor was incorrect, resulting in a fixed offset in the measurements of nearly 10°.

While the items noted above are critical in the collection of valid data, the biggest problem was in the siting of the sensors. Figure 2 shows the location of the measurements relative to a nearby building and pollution sources. The building was about 8 to 10 m high and located about 20 m from the sensors. The problem was compounded by the proximity to the major air pollution source that was the focus

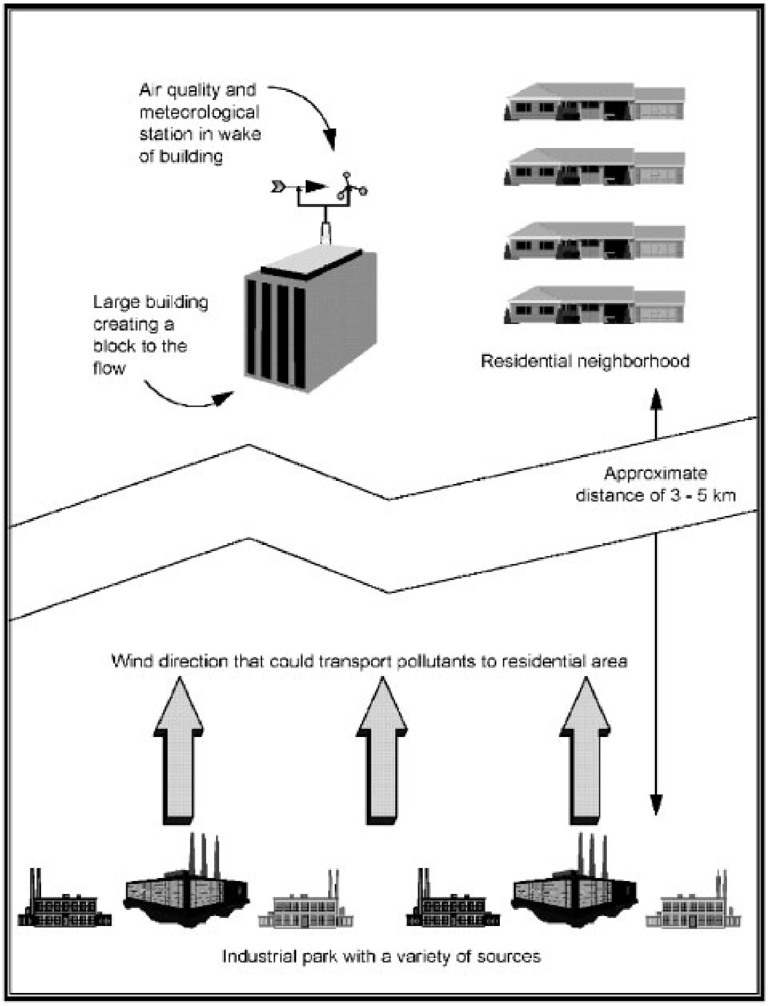

Air quality and meteorological station in wake of building
Large building creating a block to the flow
Residential neighborhood
Approximate distance of 3 - 5 km
Wind direction that could transport pollutants to residential area
Industrial park with a variety of sources

Figure 2 Location of meteorological sensors relative to an adjacent building and nearby pollution sources. See ftp site for color image.

of the overall measurement program. Located to the south was an industrial park with many sources of pollution. This air quality and meteorological station was intended to document the transport of pollutants from the industrial park to the station, which was located adjacent to residential housing. The data was then used in subsequent computer modeling of the sources to determine the impacts. With the building between the station and the source, winds from the south would be significantly altered and not be representative of the area meteorological conditions. The results of the independent audit noted these problems and made recommendations on where an acceptable site could be located and how the sensors should be maintained to collect accurate, defendable data. The five sensor-specific items noted above were highlighted and recommended servicing intervals provided in accordance with EPA guidance. The results of the audit will eventually lead to data that can be used for the intended modeling purposes.

The third case study draws more on the role of quality auditing in the planning and execution of a measurement program in the early stages of setup. As part of a large air quality and meteorological measurement program, the quality assurance contractor held a workshop for all who were making meteorological measurements as well as those who were performing the audits. The purpose of the workshop was to introduce the QA program to all measurement personnel and identify the needs and requirements of the program. This included the audit procedures, criteria used for passing or failing the audits, and the overall data quality objectives. In this manner all personnel making the measurements would understand the common goals of the program and be prepared for the auditing steps that would follow.

All participants in the program had many years of experience in the measurements being performed. One in particular had indicated objections to the workshop and the time it would take to go through the "orientation" process when it still had half of its 35 plus sites to install. This contractor had vast experience in setup of large networks of meteorological instrumentation on the east coast of the United States. Reluctantly, but fortunately, they agreed to participate. As part of the workshop exercise, one of the tasks was for all auditors and measurement contractors to measure the true direction of a distant landmark. A variety of methods were employed by the participants and the results were compared. For reference, the workshop director, who was also the technical director for the meteorological audit program, used his measurement as the "standard" by which all others were compared. For the most part all results were within ±2°. One participant was heard mumbling and groaning, indicating his answer differed by more than 30°. After a brief review of the method used to determine the true pointing direction, it was learned that he had applied the local declination in the wrong direction. This happened to be the contractor with extensive east coast experience where the local declination is applied in the opposite direction. It was also the same contractor that had responsibility for setup of over 35 stations, half of which were already in operation with the improper alignment, an error that was eventually corrected.

The lesson of this story is that the independent audit program implemented at the start of the field study identified and corrected a major problem early. This allowed all participating in the project to understand the requirements, perform the measurements in a consistent manner, learn from each other's expertise, and ultimately collect data of known quality that are defensible.

As a final example, it is valuable to address the traditional challenge of calculating an "average" wind direction. This is not a trivial problem since the wind direction is a circular function that is not amenable to simple analog averages. For years there have been debates on the appropriate method to handle the average, whether it be dealt with on a vector basis or through some algorithms that deal with the north crossing through 360°. The EPA has provided guidance on procedures for calculating wind direction in regulatory driven and other monitoring programs (U.S. EPA, 2000) that have been incorporated in one form or another into commercially available data logging systems. The discussion below summarizes an experience with an audit of a program that identified a flaw in a widely accepted algorithm for calculating the scalar wind direction using a single pass method. More extensive details on the findings can be found in Baxter (1995).

Wind speed and direction data were collected as part of a dust monitoring program at a construction site with the data used to assess the contribution of the construction activities to the downwind particulate matter concentrations. The site was located in the middle of a densely populated urban area with a number of tall buildings surrounding it. This made it virtually impossible to meet the EPA siting criteria for exposure of wind sensors (U.S. EPA, 2000). While subject to building wake turbulence, the measurements were deemed adequate to assess the general wind direction and aid in the evaluation of the dust-producing activities.

As part of the overall program an audit was performed that identified unusual patterns in the collected data. The site was located in southern California, and the location was strongly influenced by the afternoon sea breezes. These late morning to early evening flow patterns produce a very consistent southwest wind. The data collected by the meteorological system, however, showed frequent interruptions of the afternoon southwest flow with winds from a variety of other directions, including those from the northeast. Further investigation into the data logging system identified the logger used a single pass scalar average algorithm generally accepted and described in the EPA guidelines. This algorithm corrects for the rotation through north, allowing the proper interpretation of wind direction when winds vary from northwest (270°–360°) to the northeast (0°–90°).

With the identification of the anomalies in the afternoon wind direction data, several tests were performed to determine whether the problems were in the physical instruments or whether it was in the calculation methodology. The first test involved programming the existing data logger with an additional unit vector algorithm and comparing the two sets of wind direction calculations. The unit vector algorithm is also described in the EPA guidelines. Figure 3 shows the comparison of 20 days of wind direction data when wind speeds were 1 m/s or greater. The comparison shows some agreement, but for a significant number of values there is no obvious relationship.

The second test placed an identical wind sensor near the first one and logged wind direction data on an independent data logger. This second system collected data using the unit vector algorithm. Figure 4 shows unit vector comparison data between the first and second systems when the wind speeds were 1 m/s or greater. It is clear there is good agreement between the two systems when the unit vector algorithm was used.

On the basis of the first two tests it was obvious there were questions about the calculation results of the single pass scalar algorithm. The third test performed used a model to generate test wind data and perform wind direction averaging calculations with a variety of methods. A simulated 1-s interval wind data set was generated and a simple rotation introduced in the middle of the 3600-point hourly data set. This data set was then evaluated using three averaging techniques: simple arithmetic, unit vector, and the single pass scalar algorithm. The data set and results are shown in

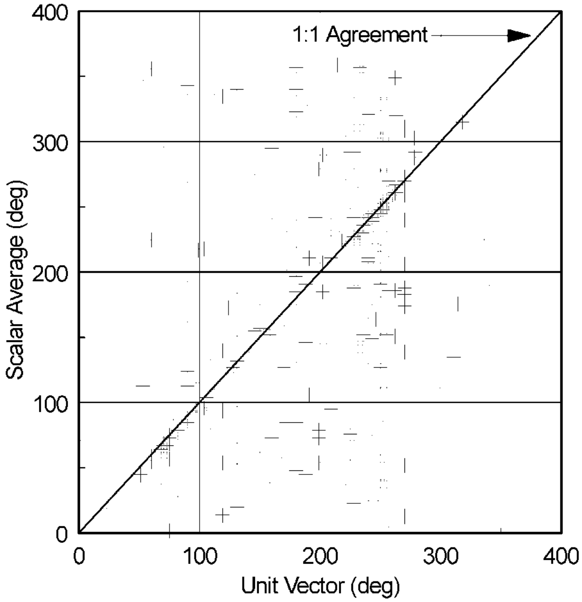


Figure 3 Comparison of 20 days of wind direction data using the single pass scalar and unit vector averaging algorithms.

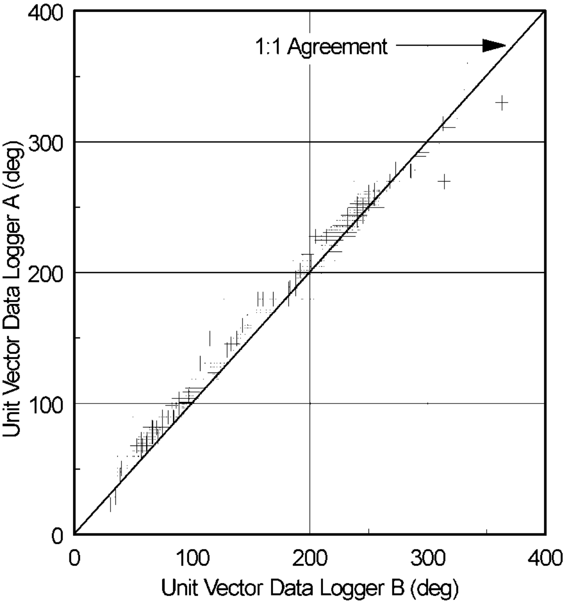


Figure 4 Comparison of 20 days of wind direction data using the unit vector averaging algorithm programmed into two different data loggers.

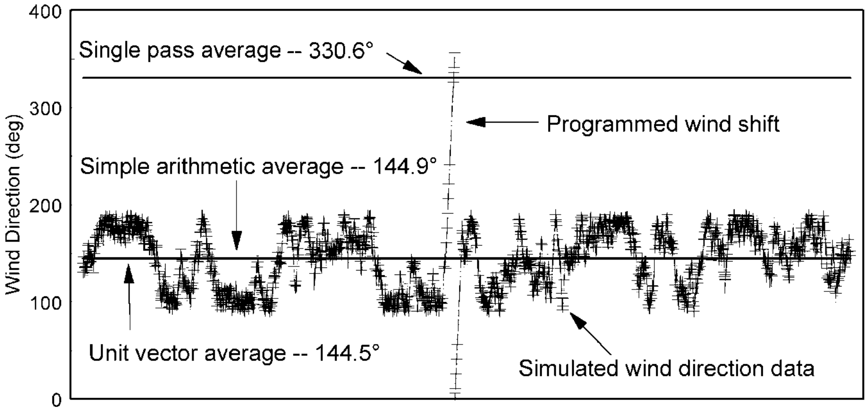


Figure 5 Simulated hourly wind direction data showing a 360° wind shift in the middle of the averaging period. The results of three different averaging techniques are plotted.

Figure 5. It was obvious from the results that there was something wrong with the single pass scalar method technique. Further investigation into the algorithm revealed that a complete circular rotation of the wind direction anytime during the hour would result in erroneous values with the magnitude of the error effectively being random. If there was not a complete rotation during the averaging period, then there was excellent agreement between the methods. However, introducing one or more complete 360° rotations during the averaging period resulted in an unrecoverable error in the reported average.

The results of this audit were extremely important in understanding potential errors in the collected data and resolving the ambiguous values obtained during what should have been very consistent wind directions. Further discussions with individuals who developed the scalar algorithm recognized the problem but also carried the purpose of the algorithm one step further. Part of the derivation of the algorithm was to develop a method that would correctly quantify the variability of the wind direction expressed as the standard deviation or sigma theta. That calculation is performed correctly using the scalar technique, but at the expense of the reported wind direction. Given these results, the optimum technique may be to use a vector method for the wind direction (either a unit vector for non-wind-speed weighted or straight vector for wind-speed weighted) and the scalar method for the standard deviation of the wind direction. Each of these methods are described in the EPA guidelines (U.S. EPA, 2000).

4 CONCLUSION

In summary, independent audits have become an integral part of many measurement programs. The extent of the audit scope and frequency of performance depends on the specific data needs and any applicable requirements. Not only do the audits aid in the improvement of the data quality, but they can in many instances identify potential problems that could invalidate collected data. Additionally, most regulatory driven programs such as Prevention of Significant Deterioration (PSD), National Air Monitoring and State and Local Air Monitoring Stations (NAMS and SLAMS), and Photochemical Assessment Monitoring Stations (PAMS) require independent audits as part of the normal measurement program. So not only are audits good for the data quality, they are also a regulatory requirement that must be fulfilled in order to use the data in modeling or analysis projects.

REFERENCES

Baxter, R. A. (1995). Evaluation of the Wind Data Collected Using Different USEPA Approved Calculation Algorithms. Paper presented at the Ninth Symposium on Meteorological Observations and Instrumentation, March 27–31, Charlotte, North Carolina.

U.S. EPA (2000). United States Environmental Protection Agency, Meteorological Monitoring Guidance for Regulatory Modeling Applications EPA-454/R-99-005, Office of Air Quality Planning and Standards.

U.S. EPA (1995). United States Environmental Protection Agency, *Quality Assurance Handbook for Air Pollution Measurement Systems*, *Vol. IV, Meteorological Measurements*, Document EPA/600/R-94/038d, Atmospheric Research and Exposure Assessment Laboratory.

U.S. EPA (1998). United States Environmental Protection Agency, EPA Guidance for Quality Assurance Project Plans EPA QA/G-5, Document EPA/600/R-98/018, Office of Research and Development.

CHAPTER 45

REGULATORY APPROACHES TO QUALITY ASSURANCE AND QUALITY CONTROL PROGRAMS

PAUL M. FRANSOLI

1 BACKGROUND

Meteorological data are often obtained to support environmental and engineering studies. Such data are evaluated by regulatory bodies for suitability in various risk, compliance, and design analyses. Information used in the regulatory arena is subject to intense scrutiny by all involved parties, including those attempting to challenge the credibility of work performed. To ensure that meteorological data are acceptably correct and complete, and that the regulations are being applied consistently to all applicants, some regulatory groups have promulgated quality assurance and quality control (QA/QC) requirements and guidelines. Meeting the monitoring requirements and following the guidelines helps to ensure acceptance of the data. It is usually preferable to agree on a monitoring plan with the regulatory group prior to proceeding with the monitoring program. It also helps to protect proposed projects from time-consuming demonstrations that the monitoring equipment and methods were satisfactory for the intended purpose.

Site characterization in engineering design or regulatory environmental studies differs from typical weather data collection programs. Site characterization programs seldom require real-time information flow, which is an important element in routine weather observations made for forecasting purposes. Instead, data from remotely operated stations can be stored on-site for long time intervals prior to collection. Some on-site processing is to obtain some statistical properties of the measurements, such as means, extremes, and standard deviations.

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

The primary quality factors appearing in most regulatory measurement programs are accuracy, precision, validity, completeness, and representativeness.

- **Accuracy and precision** are addressed in depth in another chapter in this part. The accuracy and precision requirements contained in regulatory guidance are based on both application needs and capability of the equipment used. Early atmospheric dispersion models had very simple data input requirements; typical synoptic weather observations were adequate model input. The use of site-specific (also known as on-site) data, that is, data intended to be representative of the source and/or receptor areas, helped reduce the uncertainty from the dispersion modeling portion of environmental studies. Leapfrog improvement advances between model and measurement capabilities continue to challenge both the measurement and the modeling fields.
- **Validity** of the data implies compliance with the accuracy and precision requirements and goes beyond to ensure that erroneous measurements have been removed from the validated data set. A data validation protocol often begins with removing data from known periods such as quality control checking and maintenance work that interrupts the measurements and continues with identifying periods of sensor or on-site data processing and recording equipment failures. While modern sensor and data system reliability has increased immensely, all equipment is subject to error. Some sensor failures can start as intermittent problems that may introduce a slight error that is difficult to distinguish from normal variations. Other problems can occur with sensors becoming temporarily incapacitated by external forces such as ice. Many measurement programs have found that a meteorologist knowledgeable of the measurement process and the specific program should be included in the data review team.
- **Completeness** means attaining an adequate amount of data for the intended purpose. Modeling programs intending to identify short-term phenomena that produce unacceptable hazards or risks need to be based on a sufficiently long monitoring program time period to identify the high-risk meteorological episodes. The most typical completeness requirement is 90% data recovery, though 80% applies to some remote measurement stations. Modern equipment reliability makes this requirement attainable, though frequent site checks by knowledgeable operators are important to ensure that sensors have not been damaged.
- **Representativeness** refers to measurements being made in a location that is representative of the area being characterized. The area could be the source itself, or the potential receptors of an airborne effluent, or significant points along a potential airflow pathway. Representativeness of the measurement location is addressed in siting guidelines. In addition to the measurement location being representative of an intended area, siting requirements also include instrument exposure. Wind measurements can be affected by obstacles, including the structure supporting the sensors. Temperature measurements can

be affected by nearby heat sources or sinks, such as parking lots and cooling towers.

2 STANDARDIZATION AND REGULATORY GUIDANCE

Regulatory monitoring requirements and guidance are based on the requirements set by the governing bodies, such as the U.S. Nuclear Regulatory Commission (NRC) and the U.S. Environmental Protection Agency (EPA). Others include the American National Standards Institute (ANSI) working in conjunction with the American Nuclear Society (ANS) or the American Society for Quality Control (ASQC). Of these, the EPA guidelines on prevention of significant deterioration (PSD) monitoring made the most significant recent advances in quality assurance and quality control applied to meteorological monitoring. The primary EPA monitoring guidance document is *Meteorological Monitoring Guidance for Regulatory Modeling Applications* (U.S. EPA, 2000). Volume IV (U.S. EPA, 1994) in a series of quality assurance handbooks for air pollution measurement systems contains some of the most detailed technical guidelines on monitoring techniques. Volume IV became the "how-to" book that explained many of the "why" questions behind the guidelines and guided technicians through the rigorous detail of correctly performing the equipment tests.

Early work in the nuclear power industry to site, license, and operate large nuclear-powered electric-generating stations triggered some guidelines such as *Meteorology and Atomic Energy* (Slade, 1968) and the Nuclear Regulatory Commission Safety Guide 23, which became Regulatory Guide 1.23. Technical guidance in these documents established the equipment and network design specifications for numerous nuclear-power-related projects. The 60-m tall meteorological tower with wind and temperature measurements at the 10- and 60-m levels above the ground became synonymous with RG 1.23 programs. Another important nuclear-related guidance document appeared in 1984 as ANSI/ANS-2.5: Standard for Determining Meteorological Information at Nuclear Power Sites (American Nuclear Society, 1984). This document helped to update the outdated RG 1.23 and was used for many safety plans at nuclear power plants. This standard became less useful by advances in monitoring technology. The replacement for this document was recently approved as ANSI/ANS-3.11: Determining Meteorological Information at Nuclear Facilities (American Nuclear Society, 2000).

Regulatory guidance is increasingly focused on using voluntary consensus standards, rather than having separate regulatory agencies promulgating independent requirements. The American Society for Testing and Materials (ASTM) subcommittee D22.11 (Meteorology) has produced consensus standards and practices related to meteorological monitoring equipment and application methods. The standards describe equipment design and testing considerations; the practices describe techniques to use the equipment effectively in operational programs. Many of the standards and practices address wind measurement, though atmospheric pressure, temperature, and humidity are also covered. These standards and practices are continually being reevaluated, and new material is being developed.

3 PRIMARY ELEMENTS

The basic elements of meteorological monitoring programs that achieve the data quality factors described above are summarized in this section. This material is intended to be a "primer" on the topic; compliance with current requirements and guidance should be based on the material effective during the time of a given program. The primary information needs are input to atmospheric dispersion models, with other applications of on-site measurements such as engineering and hydrology.

Siting

Siting criteria for meteorological monitoring programs include the number of stations, their locations, and the measurements to be made. Simple programs in flat terrain may only require a single station. The primary wind measurements are made at 10 m above ground level (agl), and temperature and atmospheric moisture at 2 m agl. Additional wind measurements may be necessary at higher levels to properly document airflow in complex terrain, near large manmade obstacles or bodies of water, particularly if the source is located much higher than 10 m agl. Such terrain or obstacles may also require more stations to properly document trajectories of airborne material or the locations of extreme wind or temperature conditions.

Siting criteria also include instrument exposure at the station itself. Wind, temperature, moisture, and precipitation sensors are easily influenced by supporting structures, such as towers and poles. Nearby natural and manmade obstacles can also adversely influence a measurement, making it unrepresentative of the area intended.

Equipment Selection and Procurement

Equipment selection should be driven by guidance specifications, availability of vendors to perform calibration services (which could include the manufacturers), product and service reliability, and cost. When initial cost is one of the few factors used in the selection process, the overall cost over the first few years of operation can far exceed initial differences. The vendor exhibits at trade shows and discussions with other program operators are valuable information resources.

Operator Training

Even the best equipment systems can produce unusable data if the station operators and data processing staff are inadequately trained and supervised. Some regulatory groups and private companies offer valuable training programs. As with most quality control and quality assurance activities, it is wise to document the training received.

Installation (Testing, Location Documentation)

Proper equipment installation involves adequate field testing to ensure that the system was not damaged in transit and has been properly installed. Equipment checking is addressed in the next section. Another essential station startup step all too easily overlooked is proper station location documentation. Seemingly obvious local landmarks can easily change in time, particularly when a station is the first step in a major development. The resolution and reference coordinate system should be compatible with the purposes of the data.

Calibrations, Checks, and Corrective Actions

Three important quality control activities are equipment calibrations, simple checks, and corresponding corrective actions. Calibrations imply formal comparisons of equipment responses with known conditions produced by a standard. Standards must have performance traceable to reliable sources, such as the National Institute of Standards and Technology (NIST). Less complex testing and checks can be accomplished in the field to ensure continued reliability of the measurement process. Such checks can be comparative measurements made by the monitoring equipment and a collocated standard, or by placing the equipment in a known operating condition, such as aiming a wind vane toward a known direction. Calibration and check results can either be made to demonstrate operation within the required tolerance limits or to allow for instrument adjustments that bring the monitoring equipment response within specified limits of the known condition.

Equally important to the calibration or check itself is the corresponding corrective action to be applied when the instrument response does not meet the desired tolerance limit. In addition to performing adjustment or maintenance to bring the response within the desired range, the nonconforming response of equipment operating on-line to collect data should be carefully documented and provided to the data validation staff to ensure that the out-of-tolerance data can be removed from a validated data set.

Routine Operations and Maintenance

Continuing the credibility of the data beyond the careful equipment selection, operator training, and checks requires vigilant routine operations and proper maintenance. Procedures specific to the given operating program should be available to (and used by) equipment operators and data validation staff. Site checks should be documented on checklist forms or in logbooks. Many meteorological sensors used in monitoring programs for regulatory purposes are designed for a balance of sensitivity to subtle airflow conditions and reliability of operation. Such equipment can be susceptible to damage from natural hazards, such as ice, blowing dust, birds, and other animals. It is difficult to tell from data acquired by telemetry if one of the anemometer cups is missing.

Maintenance includes preventive and corrective actions. A solid preventive maintenance program can preclude a large portion of equipment failure events. Corrective maintenance should be applied as promptly as possible to minimize downtime, provided that proper documentation is made of the failure symptoms or condition and that the replacement or repaired equipment is properly tested when installed.

Data Processing

Data processing begins during the measurement phase of the program; sensor responses are translated to a numerical value that can be averaged, totaled, stored as an extreme, or used to calculate a variance about the mean. Modern data recording equipment offers the user numerous options, as well as pitfalls. As with other activities, proper testing and documentation of the on-site processing routines is an essential first step in demonstrating that data processing is correctly applied. Most programs are enhanced by including data-identifier information with all data records, so that raw files can be uniquely identified by time period and station location. Once the data are collected from the field station, the data should be traceable throughout the data validation and editing processes.

Audits

An important step in demonstrating data credibility to outside groups is to document the results of independent verifications of compliance with established procedures and tolerance limits. The term *performance audit* implies tests made by knowledgeable staff independent of those performing the routine site operations and checks using independent testing equipment, which is also traceable to standards. The complementary function is a *system audit*, in which independent staff examine the work products and documentation to ensure that activities comply with the established procedures.

Reports

The final link in the chain of demonstrating data quality is to produce credible reports of the monitoring program results. Report content and format should fit the objectives of the program. It is wise to include summaries of the quality control and quality assurance activities in sufficient detail that the results can be accepted.

REFERENCES

American Nuclear Society (1984). ANSI/ANS-2.5: Standard for Determining Meteorological Information at Nuclear Power Sites, American Nuclear Society.

American Nuclear Society (2000). ANSI/ANS-3.11: Determining Meteorological Information at Nuclear Facilities, American Nuclear Society.

Slade, D. H. (Ed.) (1968). *Meteorology and Atomic Energy*, U.S. AEC, July.

U.S. EPA (1994). *Quality Assurance Handbook for Air Pollution Measurement Systems, Vol. IV: Meteorological Measurements*, U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-95-038d, April.

U.S. EPA (2000). *Meteorological Monitoring Guidance for Regulatory Modeling Applications*, U.S. EPA Office of Air Quality Planning and Standards, Research Triangle Park, NC, EPA-454/R-99-005, February.

CHAPTER 46

MEASURING GLOBAL TEMPERATURE

JOHN R. CHRISTY

Every part of the Earth system may be described by the variable of state we call temperature. Fundamentally, temperature is the magnitude of the average molecular kinetic energy of a substance—the higher the molecule's average speed, the greater the temperature. Today, temperature is estimated by several kinds of instruments using a variety of techniques.

In situ measurements require, in some way, direct contact between the instrument and the moving molecules of the targeted substance. The most familiar of these types of instruments is the liquid-in-glass thermometer whose bulb contains a liquid that expands or contracts, allowing the liquid to move through an opening to a narrow, graduated tube. The bulb is inserted into the medium of interest (e.g., air, water, ice, or earth) and the liquid responds according to the amount of molecular motion detected on contact. Another in situ device utilizes the fact that the electrical resistance of a conductor is proportional to its molecular kinetic energy so that the amount of electricity able to flow through the conductor indicates temperature. The velocity of sound through a substance (usually air or water) is also directly related to its temperature and thus is an indirect characteristic that can be used to monitor temperature in situ.

Remote methods, in contrast to in situ, have the advantage of measuring temperature from a distance. The most direct of these methods employs radiometers that measure the intensity of radiation emitted by a substance. The magnitude of the intensity is often proportional to the substance's temperature. Satellite radiometers fall into this category of devices as they monitor the upwelling radiation from the various components of Earth's system. Other relatively new remotely sensed methods estimate temperature by measuring (a) the speed and refraction of radio signals through a material (e.g., Global Positioning System satellites), (b) the physical

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

height of the sea surface (higher altitudes mean expanded or warmer water) and even the thermal "color" of the ocean affected by microorganisms that preferentially appear in waters of certain temperatures.

The distinction between in situ and remote may be rather blurred. One thinks of a bucket dropped over the side of a ship, filled with seawater, hoisted back up on deck into which a thermometer is inserted, and from which a temperature reading is determined after some time as in situ. However, one probably thinks of a satellite radiometer, which measures the actual, unaltered photons representing the exact character of the substance in question and traveling at the speed of light, as a remote measurement. One can see that mere proximity to the intended medium may not assure the most accurate estimate of temperature.

A unique and quasi-direct method is one that measures the temperature at various depths of a very stable borehole (e.g., in bedrock or ice cap) and then estimates what the surface temperature would have been in the past to produce the temperatures observed at each given depth. Simply put, the deeper the temperature reading, the longer in the past it was influenced by the surface temperature.

Temperatures from all of the above devices are referred to as being part of the "instrumental" record and in some sense may be called direct measurements. There are, however, several indirect methods available in which some organism or physical process preserves in its history the character of the environmental factors, including temperature, affecting it. In this category of "proxy" records are tree rings, ice core composition and thickness, isotope ratios in ice, sea floor, and sediment cores, pollen distributions in sediments, erosion rates, coral bands, sea level height, evidence of the extent of mountain glaciation, plant and animal fossil types and distributions, and many more. With these types of proxy and borehole records, some estimates of the climate prior to the instrumental record are possible.

The *global temperature* is a rather ambiguous term since every part of the global system has its own temperature. When used in the context of climate change, it usually means the temperature of Earth's atmosphere about a meter or two above the surface, often termed *near-surface air temperature*. However, one could speak of the temperature of the land itself, or of the sea water at various depths, or of the ice in the ice caps, or of the atmosphere at any of several altitudes. It is even possible to measure the temperature of the "cold space" in which Earth orbits and find a value of about 2.7 K. Because Earth is a system of many interactive components, the temperature of each is truly necessary to document global temperature.

Most data sets of global temperature are in fact not global in extent nor systematic in quantity measured. So, not only are there variations in "how" and "where" temperature is measured, one must be careful to know "what" aspect of the Earth system is being measured. Because we as humans experience weather and sustain our existence on the surface of this planet, the near-surface air temperature is usually the quantity of first importance to us.

For all of the temperature data sets above and to which the term global is applied, there are issues of uncertainty—spatial and temporal homogeneity, calibration (or lack thereof) of sensors and techniques, changes in instrumentation (type, method etc.), degradation of instruments over time, corruption of proxies over time, changes in local environment due to nonclimatic factors, poor information on observational practices, and many others. Each of these present potential problems and are usually difficult to assess so that the total level of uncertainty in any measurement is not completely quantifiable. In the attempt to understand precisely what the climate system is doing in terms of temperature, the unpleasant notion of potential error is never totally absent.

The various measures of global temperature have been described in publications too numerous to list. The four Intergovernmental Panel on Climate Change reports (IPCC, 1990, 1992, 1996, 2001) are probably the best sources of information that document the more prominent data sets.

1 PROXY RECORD OF SURFACE TEMPERATURES

The average of proxy-estimated temperatures of widely scattered sites prior to instrumental record gives a rough idea of what the temperature may have been in the past. One set of these time series is shown in Figure 1 (see also IPCC, 1990, Fig. 7.1). Even in these very low time-resolution diagrams, there is clear evidence of large variability in global temperatures. Fluctuations in volcanism, solar insolation, orbital parameters, atmospheric composition (including changes due to human activities), aerosols, natural chaos, etc. certainly play their roles, sometimes in isolation and at other times in interdependent ways, as probable explanations for the variations. Efforts have been attempted to quantify these causes, usually by a combination of empirical calibration and analytical theory (e.g., Mann et al., 1998; Tett et al., 1997; Wigley and Raper, 1990), but such discussion is not the focus of this chapter.

Figure 1 (top) shows two periods of interglacial (warm) temperatures over the past 150,000 years, the current being the Holocene that began over 10,000 years ago. The mid-Holocene (~6000 years ago) was relatively warm as was the period about 1000 years ago (Fig. 1, middle and lower). There is considerable evidence that the period between 1400 and 1850 was somewhat cool and is commonly called the Little Ice Age, though recent evidence suggests that it was a period of several types of climate fluctuations in various parts of the globe. Solar variability and significant volcanism are among those forcing parameters thought to play roles in causing these "recently" depressed temperatures.

2 INSTRUMENTAL RECORD OF SURFACE TEMPERATURES

For most of the instrumental period, the past 150 years or so, temperature has been determined from the familiar liquid-in-glass thermometers housed in various types of shelters, usually 1.5 m above the ground, around the world's land areas. Over the oceans, the sea surface temperature (or SST, i.e., the temperature of the seawater, not the air) is the preferred quantity because of the ocean's more spatially and temporally coherent temperature field. The manner by which the SSTs are measured has varied considerably through the years and requires careful adjustment factors to account for

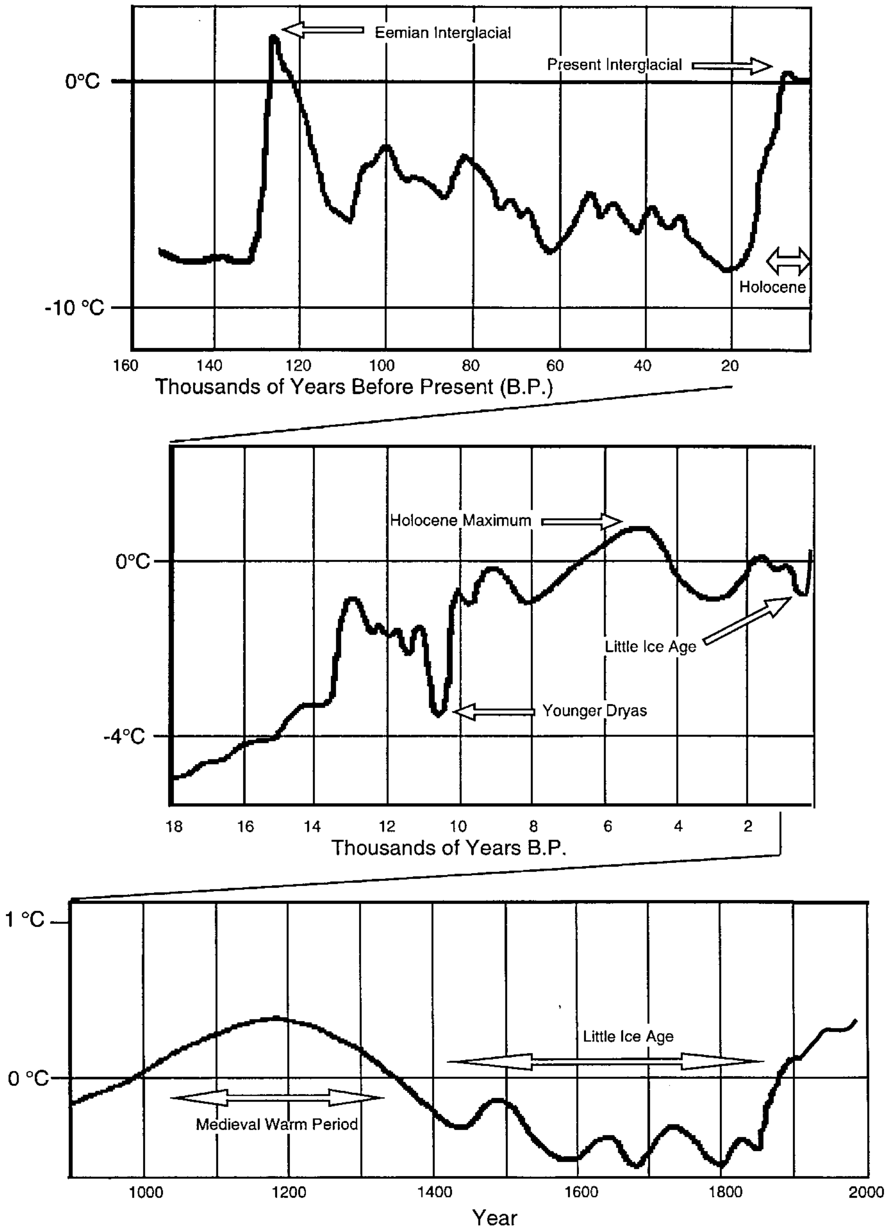


Figure 1 Estimates of global temperature variations over three long periods ending with the present day determined from proxy information (*EarthQuest*, Office of Interdisciplinary Earth Studies, Spring 1991, Vol. 5, p. 1).

biases introduced when, for example, buckets or ship-intake values are used (Folland and Parker, 1995). The "global" temperature most commonly reported is a combination of the near-surface air over land and SST reports over oceans, however geographically sparse they might be.

Only a handful of serious efforts have achieved the goal of collecting as many reports as are presently accessible to document the temperature in several places around the globe over the past 150 years. (Scattered records go back further, one to the midseventeenth century in central England; see Manley, 1974.) The IPCC usually focuses on the data sets produced by (1) the Climate Research Unit of the University of East Anglia (near-surface air temperature over land; Jones and Briffa, 1992), (2) UK Meteorological Office (SSTs; Parker et al., 1995), (3) Goddard Institute for Space Studies, NASA (near-surface air temperature over land; Hansen and Lebedeff, 1988), (4) Russian (near-surface air temperature over land, Vinnikov et al., 1990), and (5) NOAA satellite-based SSTs (Reynolds and Smith, 1995). Two other significant efforts are the Global Historical Climate Network (NOAA/NCDC; Vose et al., 1995) and a database of individual SST reports known as COADS (Woodruff et al., 1987). At present, some of these groups are combining the data sets of temperatures over land with SSTs to produce global coverage (e.g., IPCC, 1996; Hansen et al., 1996).

The time series of global, annual surface temperature anomalies for three of the groups is displayed in Figure 2. (All of the time series produced from the various groups are quite similar, and this is to be expected since the critical broad variations depend on basically the same primary source data for each data set.) These time

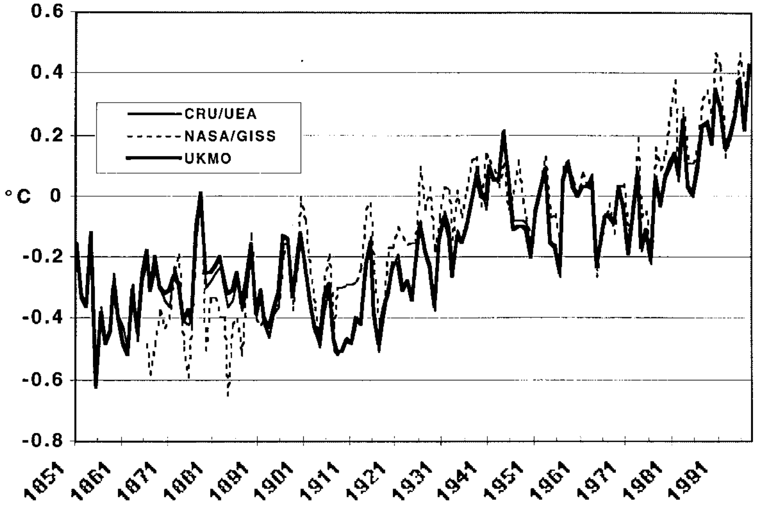


Figure 2 Annual, global anomalies of surface temperature 1851–1997 as a combination of near-surface air temperatures over land and SSTs over oceans (CRU/UEA, UKMO) and land-only (NASA/GISS).

series were described in the IPCC (1996) as showing a 0.3 to 0.6°C rise over the past century, the range being due to the uncertainty factors mentioned above.

The features of this time series are well-known; a fairly significant rise from about 1910 to 1940 (~0.4°C), then somewhat random variations to 1980 and a rise since that time (an additional ~0.2°C). A comparison of the most recent 30 years with those of 1870–1899 reveal an increase in temperature of about 0.4°C while a comparison of the most recent 10 years versus the earliest 10 (1860–1869) shows a rise of about 0.6°C.

3 BOREHOLE TEMPERATURES

An insulated, homogeneous column of material with one end exposed to temperature variations over a long period will reveal temperature variations throughout the column according to the speed at which the fluctuations propagate from the exposed end. An analysis of the temperatures throughout the column at a given point in time, then, may contain enough information to recover those external temperature variations. The general idea is that the greater the depth from the exposed end, the further back in time the temperature at that depth might represent. Of course, as time passes, the perturbations tend to smear, and it becomes difficult to extract meaningful information about what may have happened at the surface.

Certain types of stable, homogeneous bedrock and large ice plateaus lend themselves to temperature recovery through measurements taken in boreholes. The theory and complications regarding the inversion of the temperature profiles into time series are quite complex, but the general results show that many land regions have experienced warming in the past three centuries, though the results show variations among the individual sites (Deming, 1995). Also, evidence from boreholes on the Greenland Plateau suggest that the period around 1000 years ago was warmer than the present century (Cuffey et al., 1994; Dahl-Jensen et al., 1997). Both of these results are consistent with the temperature estimates in Figure 1.

4 UPPER AIR TEMPERATURES

Upper air measurements, which could be compiled into large-scale averages, became possible in the late 1950s due to the expansion of the network of radiosonde stations—sites that release balloon-borne instruments to measure temperature, wind, and humidity to altitudes up to and exceeding 10 km. Sources of large-scale averages of temperatures at various levels and for various layers have been provided by NOAA (Angell, 1988; 63 stations; Oort and Liu, 1993; 800+ stations), UK Meteorological Office (UKMO) (Parker et al., 1997; 350+ stations), and Russian IHMI-WDC (Sterin et al., 1997; 800+ stations).

Temperature time series constructed from individual radiosonde sites often display inhomogeneities most often related to changes in the instrumentation or the algorithms by which the raw data are processed into pressure-level data (Gaffen, 1994). These inhomogeneities can be quite prominent at the highest elevations because errors or changes tend to accumulate as the balloon ascends. Oort and Liu (1993) generate global maps by objective analysis of their world-wide radiosonde dataset. The UKMO product uses selected stations with better records of homogeneity and length and produces a quasi-global analyses with limited interpolation for filling in vacant grids. As with Oort and Liu (1993), the RIHMI-WDC produces global analyses by objective means but also applies a complex quality control algorithm to the data to remove obvious inconsistencies that violate hydrostatic constraints and those of spatial coherency.

Since 1979, nine NOAA polar orbiting satellites have carried an instrument, the microwave sounding unit, or MSU, designed to provide temperature information in both clear and cloudy areas where infrared (IR) methods are ineffective. The sensor measures the intensity of radiation near the 60-GHz oxygen absorption band, which is proportional to atmospheric temperature. Though not intended to provide long-term climate information, data from the nine MSUs, which have orbited since 1979, have been calibrated and merged into a single time series (Christy et al., 1995). With daily global coverage (over 30,000 observations per day) of a very robust quantity (a volume of air over 50,000 km^3 per observation), these data have some advantages over other types of data sets.

Two MSU temperature products are widely used: the lower troposphere (the average temperature of the surface to about 7 km or 1000 to 400 hPa) and the lower stratosphere (17 to 22 km layer or about 120 to 70 hPa). Version D of the data described in Christy et al. (2000) takes into account recently discovered influences on the satellites that affect the observations.

We have found in the several years of constructing the MSU data sets that the issues of greatest impact on the long-term record are (1) the intersatellite biases, (2) the orbital time drift, and (3) the orbit decay (see Wentz and Schabel, 1998). Because the MSU data sets are products we produce, I shall devote a relatively large amount of discussion to them.

Though calibrated to high precision on Earth in thermally controlled vacuum chambers, the MSU, once in the environment of space, may acquire or display unexpected characteristics. In one case (NOAA-12) an electronic gain change occurred after launch so that the anticipated calibration target temperature of cold space (2.7 K) was not correct, being measured at about 6 K. Corrections for this effect were generated by Mo (1995). Also, the MSU, as a cross-track scanner, monitors temperatures to the left and right of the track. In an unexpected result, temperature comparisons of these two sides produce average differences as great as 3°C and as little as 0.05°C in different instruments. This asymmetry is likely due to variations in the antenna beam patterns from instrument to instrument—i.e., the actual location of the "cone" through which the energy upwells to the sensor is not as precisely located as anticipated. This is a systematic effect (it will not change during the instrument's life) but will produce overall biases if not accounted for.

Continuing with nonclimatic effects, there are two spurious consequences of the slow east–west drift of a satellite. One is that the MSU observes Earth at later or earlier local times as it passes over a given region as it drifts. Thus, the natural diurnal cycle of Earth's temperature is aliased through time and can appear as an artificial trend in the data. In version D, we apply an independently determined diurnal trend correction to all satellites based on the differences of the individual footprints across a scan line.

In Christy et al. (1998) we noted a newly discovered effect in which the temperature of the MSU instrument itself influences the temperature of the observations. This effect is manifested as both intraannual and interannual variations, especially for the satellites in the 0200/1400 orbit time node. The temperature of the instrument fluctuates due to the east–west drift, which induces a changing solar shadowing effect on the instrument itself. The net effect of these east–west drifts is to introduce an artificial warming into the time series.

Wentz and Schabel (1998) have discovered yet another effect that for the lower troposphere product has an important influence. During periods of high solar activity, the upper atmosphere expands and exerts greater drag on the satellites, causing them to descend a few kilometers over the 2 or 3 years of increased solar activity from their average altitude of about 850 km. The lower tropospheric temperature utilizes a retrieval that is very sensitive to satellite altitude. The net effect of the "falling" satellites is to introduce an artificial cooling to the lower tropospheric time series.

To summarize, then, we know that the MSU data contain two artificial warming effects and one artificial cooling effect. These have been quantified and removed from the data set, and in fact they are almost exactly offsetting in their net effect. However, we must not presume that this set of issues, combined with what we have discovered previously, constitutes the complete body of knowledge regarding the long-term stability of the MSUs and their peculiarities. These instruments are in space, so we cannot examine them directly for anomalies we suspect may be developing. We are forced to diagnose problems based on the data they transmit, and this is a difficult enterprise. However, this effort has been a largely successful endeavor because we have the ability to examine the MSU data in light of completely independent data from radiosonde observations. The results, in our view, are very interesting.

In Figure 3 we show the annual anomalies of global tropospheric temperature from Angell, UK Met. Office, RIHMI-WDC, and the MSU D (see Christy, 1995). We must note that Angell studies the thickness temperature of the 850- to 300-hPa layer, RIHMI looks at the average temperature of the 850- to 300-hPa layer, and the UKMO at the raob-simulated MSU temperature. Geographical coverage of the radiosonde data sets is limited, while the MSU essentially observes the entire planet. However, the spatial coherence of the troposphere is such that fewer spatial degrees of freedom exist than at the surface; thus fewer sites are necessary to define the average global temperature.

For the period 1979–1997 all data sets indicate that the temperature of the troposphere has declined slightly. Before that period, all show a warming trend, principally a sharp jump during 1976–1980, though the RIHMI anomalies are much less in magnitude (the reasons are probably due to a stiff interpolation method). Many such

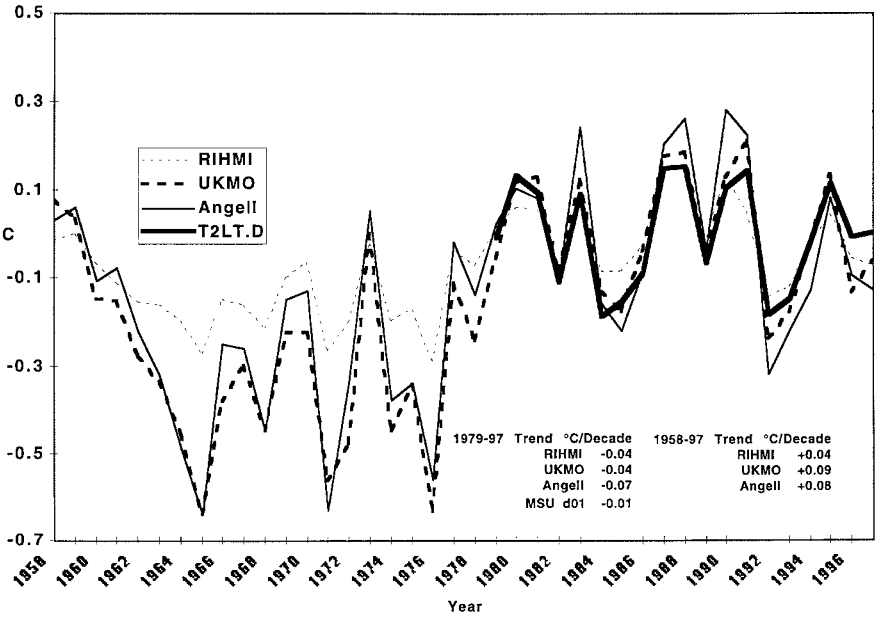


Figure 3 Annual global tropospheric temperature anomalies 1958–1997 determined from radiosonde data sets and the MSUs on polar orbiting satellites. The layer is roughly between 1 and 8 km altitude.

comparisons of multistation averages have been performed, and they indicate excellent agreement between the radiosondes and MSU temperatures.

Some concern has been raised regarding the lack of a tropical warming trend observed independently by radiosondes and the MSU since 1979. It has been proposed that the tropical troposphere should behave in concert with variations in tropical SSTs as indicated by climate model results, which are forced with observed SSTs. The observed atmospheric data seem to indicate the tropical troposphere is somewhat more decoupled from the boundary layer than presumed. The issue was put forth by Hurrell and Trenberth (1997) in which they suggested that when compared with tropically averaged SSTs, there was evidence for two "spurious" downward jumps in the MSU record, one of 0.25°C in late 1981 and of 0.1°C in late 1991. Since these events seemed to be near the time when new satellites were merged into the data set (NOAA-07 and NOAA-12, respectively), they suggested that improper biases might have been applied to those two satellites.

It is possible to test these claims by producing the anomalies of each of the satellites independently. A comparison of the anomalies indicate no such break occurred in 1981 or in 1991, i.e., the anomalies of the time series remained the same with or without the addition of the new satellites. In addition, comparisons with radiosonde anomalies across both of the boundaries of alleged jumps verified the MSU anomalies (Christy et al., 1997, 1998). Speculation cited by Hurrell and

Trenberth (1997) for the causes of the "jumps" (i.e., surface emissivity effects, bias errors, etc.) were investigated thoroughly and shown to be without foundation. Mentioned, but not emphasized, by Hurrell and Trenberth was the possibility that the SSTs and troposphere indeed do show evidence of slight independence in multidecadal trends (a phenomenon already verified in other regions, Ross et al., 1996).

The apparent jumps between SSTs and the troposphere are an intriguing observation. We examined this effect further and found the real source of the shifts appear to be due to a relative warming of the SSTs in the tropical oceans from the Indian eastward through the central Pacific. We compared the tropical SST time series with that of the Night Marine Air Temperatures (NMATs), and in Figure 4 we see that, remarkably, there are shifts between the water and air temperatures at the suspected points in time (between 1981–1982 and 1991–1992). These shifts are evident without any appeal to tropospheric temperatures and shows that the NMATs produce a trend 0.11°C/decade cooler than the SSTs immediately underneath for this very large region (Christy et al., 1998). These shifts are not seen in the tropical Atlantic (Fig. 5). One possible factor being observed here is an apparent slight increase in the instability of the tropical atmosphere on the order of 0.2 to 0.3°C over a vertical extent of about 4 to 5 km (but mostly in the lowest 20 m) since 1979.

An additional possible factor for explaining the difference between surface and tropospheric trends is related to the change in the SST observing system of the tropical Indian and Pacific Oceans. Since the mid-1980s, moored and drifting buoys have supplied a massive increase in observations. In some months in the near-equatorial Pacific, over 90% of the observations now are derived from buoys. It is possible, though not proven as of yet, that the buoy temperatures may be warmer than traditional ship reports because their nominal depth of measurement is 1 m, a

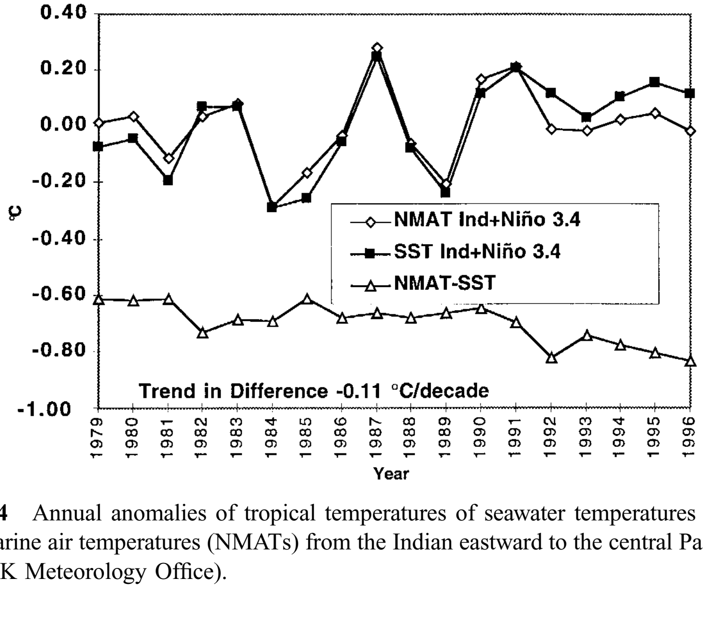


Figure 4 Annual anomalies of tropical temperatures of seawater temperatures (SSTs) and night marine air temperatures (NMATs) from the Indian eastward to the central Pacific oceans (from UK Meteorology Office).

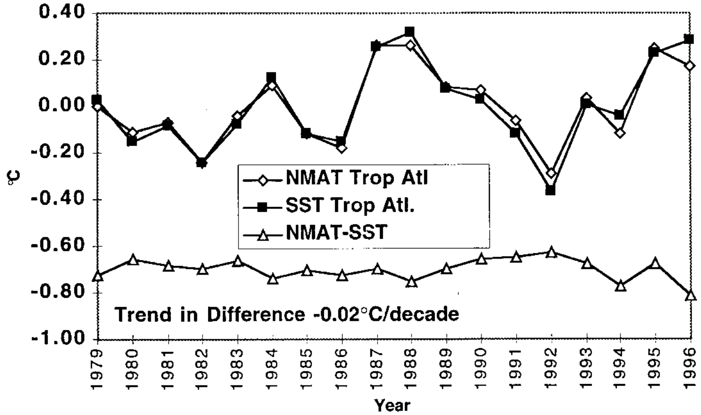


Figure 5 As in Fig. 4 but for the tropical Atlantic Ocean.

shallower and warmer depth than the typical ship-intake depth of 3 to 20 m. The increase in buoy reports may have introduced a slight warming bias since the mid-1980s. This is another issue that will be investigated to assure that the time series of SSTs is as homogeneous as possible (C. Folland, personal communication).

It appears that notions of vertically fixed linkages between the tropical SSTs and troposphere require some refinement in light of the comparisons shown above. In any case, the evidence presented here shows that over the past 20 years (a woefully brief period in terms of climate variations) the global troposphere has not experienced any significant warming or cooling. What may be of most interest in terms of climate is finding the explanation for the apparent difference in warming rates between the tropical surface and the free atmosphere.

The lower-stratospheric temperatures reveal considerable interannual variations (Fig. 6). Warming episodes related to eruptions of Agung (1963), El Chichon (1982), and Mt. Pinatubo (1991) stand out as the largest features. Underlying this punctuated time series is a general downward trend in the global stratospheric temperature, with an apparent acceleration in the last years. The location and degree of decline correlates well with similar decreases in the concentration of stratospheric ozone (Christy and Drouilhet, 1994; McCormack and Hood, 1994). With only a few decades available, however, it is difficult to know the character and extent of natural variations for this layer.

5 CONCLUSION

It should be apparent that all of the data sets mentioned above contain uncertainties in their ability to answer for us “what is the global temperature?” They provide information for different parts of the global system and thus should be viewed as pieces that imperfectly highlight specific components of the climate puzzle.

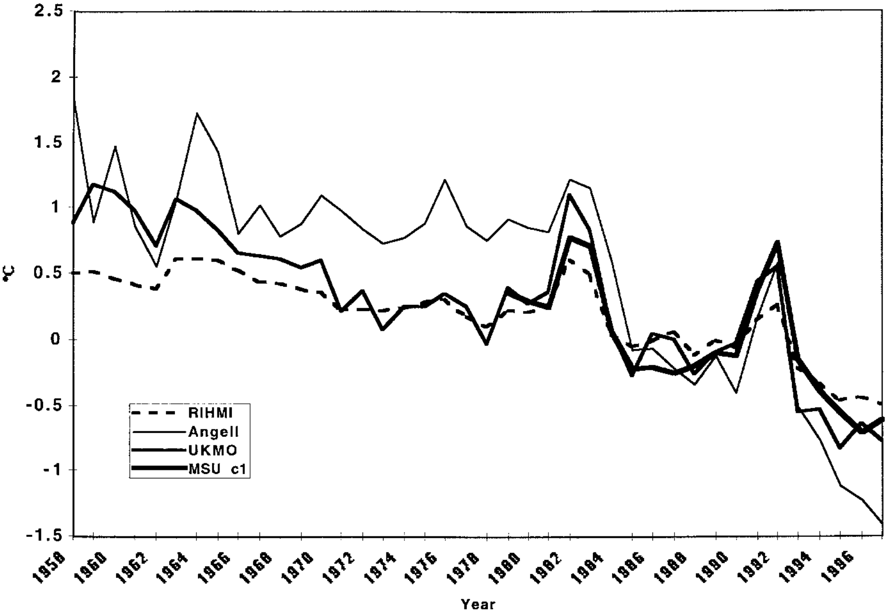


Figure 6 Annual global lower-stratospheric temperature anomalies 1958–1997 determined from radiosonde data sets and the MSUs on polar orbiting satellites. The layer is centered near 20 km altitude.

The global surface air temperature has surely risen in the past 150 years, while between 1400 and 1850 it appears to have been relatively level. In the centuries before 1400, the scant and murky evidence implies a period of relative warmth unrelated to any human factor. The troposphere has shown warming since 1958, which is due to a relatively rapid shift during only 5 years, 1976–1980. Without this shift, there would be no warming observed (which points to the limitations of discussing linear trends determined from a fluctuating time series). The global stratosphere clearly has experienced a decline in temperature, consistent with both ozone depletion and enhanced concentrations of greenhouse gases, though the character of natural variability is largely unknown for this layer.

The presence of so many types of uncertainty in our ability to determine the global temperature should draw attention to the critical needs we currently face. Our global observing system requires a more spatially and temporally complete set of both in situ and remote observations made with instruments that are precisely calibrated and consistent in quality. The current concerns of climate change have brought the present inadequacies to light and hopefully will spur action to increase our global network of observations so we will have the necessary information to make reasonable decisions about possible human impacts on climate.

Author’s note: Much of the material contained in this chapter was developed for the Norwegian Academy of Technological Sciences and appears in *Do We Under-*

stand Global Climate Change? (an international seminar at Oslo at Holmen Fjordhotell, Asker, 11–12 June, 1998) Trondheim, June, 1998.

REFERENCES

Angell, J. K. (1988). Variations and trends in tropospheric and stratospheric global temperatures, 1958–87, *J. Climate* **1**, 1296–1313.

Christy, J. R., and S. J. Drouilhet (1994). Variability in daily, zonal mean lower-stratospheric temperatures, *J. Climate* **7**, 106–120.

Christy, J. R., R. W. Spencer, and R. T. McNider (1995). Reducing noise in the MSU daily lower-tropospheric global temperature dataset, *J. Climate* **8**, 888–896.

Christy, J. R. (1995). Temperature above the surface layer, *Clim. Change* **31**, 455–474.

Christy, J. R., R. W. Spencer, and W. D. Braswell (1997). How accurate are satellite "thermometers"? *Nature* **389**, 342–343.

Christy, J. R., R. W. Spencer, and W. D. Braswell (2000). MSU Tropospheric temperatures: Data set construction and radiosonde comparisons, *J. Atmos. Oceanic Tech.* **17**, 1153–1170.

Christy, J. R., R. W. Spencer, and E. S. Lobl (1998). Analysis of the merging procedure for the MSU daily temperature time series, *J. Climate*, **11**, 2016–2041.

Cuffey, K. M., R. B. Alley, P. M. Grootes, J. F. Bolzan, and S. Anandakrishnan (1994). Calibration of the δ^{18}O isotopic paleothermometer for central Greenland, using borehole temperatures, *J. Glaciology* **40**, 341–349.

Dahl-Jensen, D., N. S. Gundestrup, K. Mosegaard, and G. D. Clow (1997). Reconstruction of the past climate from the GRIP temperature profile by Monte Carlo inversion, *EOS Abstracts*, AGU 1997 Fall Meeting, San Francisco, CA, F6.

Deming, D. (1995). Climatic warming in North America: Analysis of borehole temperatures, *Science* **268**, 1576–1577.

Folland, C. K., and D. E. Parker (1995). Correction of instrumental biases in historical sea surface temperature data, *Q. J. R. Meteorol. Soc.* **121**, 319–367.

Gaffen, D. J. (1994). Temporal inhomogeneities in radiosonde temperature records, *J. Geophys. Res.* **99 D2**, 3667–3676.

Hansen, J., and S. Lebedeff (1988). Global surface temperatures: update through 1987, *Geophys. Res. Lett.* **21**, 2693–2696.

Hansen, J., R. Ruedy, and M. Sato (1996). Global surface air temperature in 1995: Return to pre-Pinatubo level, *Geophys. Res. Lett.* **23**, 1665–1668.

Hurrell, J. W., and K. E. Trenberth (1997). Spurious trends in MSU satellite temperatures due to merging of different satellite records, *Nature* **386**, 164–167.

IPCC, 1990: Climate Change, The IPCC Scientific Assessment, J. T. Houghton, G. J. Jenkins, and J. J. Ephraums, Eds., Cambridge University Press, Cambridge, UK, 365 pp.

IPCC, 1992: Climate Change, 1992: The Supplementary Report to the IPCC Scientific Assessment, J. T. Houghton, B. A. Callander, and S. K. Varney, Eds., Cambridge University Press, Cambridge, UK, 198 pp.

IPCC, 1996: Climate Change 1995, The Science of Climate Change, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate

Change, J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell, Eds., Cambridge University Press, Cambridge, UK, 572 pp.

IPCC 2001: Climate Change 2001, The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, J. T. Houghton et al. Eds., Cambridge University Press, UK, 881 pp.

Jones, P. D., and K. R. Briffa (1992). Global surface air temperature variations over the twentieth century, Part 1: Spatial, temporal and seasonal details, *Holocene* **2**, 165–179.

Jones, P. D., T. J. Osborn, and K. R. Briffa (1997). Estimating sampling errors in large-scale temperature averages, *J. Climate* **10**, 2548–2568.

Manley, G. (1974). Central England temperatures: Monthly means 1659 to 1973, *Q. J. R. Meteorol. Soc.* **100**, 389.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1998). Global scale temperature patterns and climate forcing over the past six centuries, *Nature* **392**, 779–788.

McCormack, J. P., and L. L. Hood (1994). Relationship between ozone and temperature trends in the lower stratosphere: Latitude and seasonal dependencies, *Geophys. Res. Lett.* **21**, 1615–1618.

Mo, T. (1995). A study of the Microwave Sounding Unit on the NOAA-12 satellite, IEEE Trans. *Geosci. and Remote Sensing* **33**, 1141–1152.

Oort, A. H., and H. Liu (1993). Upper-air temperature trends over the globe, 1958–1989, *J. Climate* **6**, 292–307.

Parker, D. E., C. K. Folland, and M. Jackson (1995). Marine surface temperature observed variations and data requirements, *Clim. Change* **31**, 559–600.

Parker, D. E., M. Gordon, D. P. N. Cullum, D. M. H. Sexton, C. K. Folland, and N. Rayner (l997). A new global gridded radiosonde temperature data base and recent temperature trends, *Geophys. Res. Lett.* **24**, 1499–1502.

Reynolds, R. W., and T. M. Smith (1995). A high-resolution global sea surface temperature climatology, *J. Climate* **8**, 1571–1583.

Ross, R. J., J. Otterman, D. O'C. Starr, W. P. Elliot, J. K. Angell, and J. Susskind (1996). Regional trends of surface and tropospheric temperature and evening-morning temperature difference in northern latitudes: 1979–93, *Geophys. Res. Lett.* **23**, 3179–3182.

Sterin, A. M., V. A. Orshekhovskaya, and N. M. Mishina (1997). Comparison of upper-air temperature variations in the past and current decade, derived from the global radiosonde database and from the microwave sounding unit, *Proc. 22nd Ann. Climate Diagnostics and Prediction Workshop, Berkeley, CA, USA.* U.S. Dept. Commerce, Sills Bldg., 5285 Port Royal Road, Springfield VA 22161.

Tett, S. F. B., J. F. B. Mitchell, D. E. Parker, and M. R. Allen (1997). Human influence on the atmospheric vertical temperature structure: detection and observations, *Science* **247**, 1170–1173.

Vinnikov, K. Ya., P. Ya. Groisman, and K. M. Lugina (1990). Empirical data on contemporary global climate changes (temperature and precipitation), *J. Climate* **3**, 662–677.

Vose, R. S., T. C. Peterson, R. L. Schmoyer, and J. E. Eischeid (1995). The Global Historical Climatology Network, a preview of version 2. Ninth Conf. on Applied Climatology, Dallas TX, *Amer. Meteor. Soc.* 59–64.

Wentz, F. J., and M. Schabel (1998). Effects of satellite orbital decay on MSU lower tropospheric temperature trends, *Nature*, **394**, 361–364.

Wigley, T. M. L., and S. C. B. Raper (1990). Natural variability of the climate system and detection of the greenhouse effect, *Nature* **344**, 324–327.

Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer (1987). A Comprehensive Ocean-Atmosphere Dataset, *Bull. Amer. Meteor. Soc.* **68**, 1239–1250.

CHAPTER 47

SATELLITE VERSUS IN SITU MEASUREMENTS AT THE AIR–SEA INTERFACE

KRISTINA B. KATSAROS

In this chapter we explore the trade-offs in selecting surface in situ versus satellite platforms to measure properties near or at the air–sea interface. The most obvious difference between the two observing platforms is sampling coverage in time and space. A surface platform can obtain continuous measurements at a point, while a polar-orbiting satellite instrument samples, at most, twice per day depending on the swath width of the sensor. A geostationary satellite can sample the surface as frequently as every 15 min (once per hour is typical), but the high altitude (38,000 km) limits the resolution that is achievable for some sensors. To focus the discussion, we compare the following two variables commonly measured by both in situ and satellite systems: the sea surface temperature (SST) and surface wind speed, U, or wind vector, $\bar{U}$.

1 SEA SURFACE TEMPERATURE

SSTs were traditionally measured from ships by the insertion of a mercury thermometer into water samples obtained by buckets lowered from deck. Currently, the temperature of the water intake of ships is recorded and reported every 3 h via satellite to the international World Meteorological Organization (WMO) weather telecommunications network (Global Transmission System, or GTS) and distributed to all weather services globally. Moored and drifting buoys have expanded the in situ network to cover larger areas of the global ocean. In the case of moored buoys,

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

continuous time series are obtained. Examples of such networks are the U.S. National Data Buoy Center (NDBC) buoys located around the U.S. coastline, the Tropical Atmosphere Ocean (TAO) network in the tropical Pacific Ocean, operational since the mid-1980s (McPhaden et al., 1998), and the global drifter bouys (Swenson and Niiler, 1996; Bushnell, 1996).

Sea surface temperatures from satellites are obtained by measuring the radiance emitted from the sea surface at electromagnetic frequencies within spectral regions where Earth's atmosphere is only weakly absorbing, so-called atmospheric window regions. The main window for SST measurements is in the infrared spectrum between 8 and 12 μm. Clouds are opaque at infrared wavelengths, however, which is a serious limitation for observing SST in certain regions of the world. Certain wavelengths in the microwave spectrum do penetrate clouds, but the technology does not allow fine spatial resolution at these wavelengths. Table 1 presents the main atmospheric window regions used for SST measurements.

For SST, the in situ and satellite systems are complementary and are used in conjunction for the National Centers for Environmental Prediction (NCEP) SST product, which provides global SST data at 1° latitude by 1° longitude resolution, averaged over one week (Reynolds and Smith, 1994). The in situ SST are used to calibrate the satellite values inferred from infrared signals in the atmospheric window regions. Measurements are obtained at two wavelengths whose absorption values by the intervening atmosphere are different. By differencing the two measurements, the effect of atmospheric absorption can be measured and corrections applied. This technique provides the primary atmospheric corrections (McClain et al., 1985). The in situ SSTs from moored and drifting buoys are combined with the satellite data after this correction for atmospheric transmission in an optimal interpolation scheme (Barton, 1995).

TABLE 1 Spectral Regions Used for Determining SST Satellites[a]

	Frequency/Wavelength Region		
Satellite Type	3.6 μm	10–11 μm	5 cm (6.6 GHz)
Polar orbiting	*TIROS/NOAA* (1960s onward)		*Seasat*, (1978) *Nimbus 7* (1978–1985)
Resolution	4 km	4 km (1 km maximum)	100s of km
Low orbit in subtropical/ tropical latitudes	—		*Tropical Rainfall Measuring Mission* (TRMM, 1997–)
Resolution	—	—	15 km
Geostationary	*GOES Series*		—
Resolution	4 km	4 km	—

[a] The satellites carrying the instruments and the typical surface resolution.

The basic measurement of temperature in situ is not a problem: Thermistor or resistance thermometer units are usually reliable and the calibration remains stable, although drift in the electronics due to seasonal changes in the mean environmental temperature must be accounted for. Similarly, satellite infrared technology, filters, and lenses have had a 35-year history or longer and are also very reliable. The most difficult problem for achieving stable results has been due to injections of aerosols from volcanos (Reynolds, 1993).

The measurement problems associated with SST determination include, for the in situ, heating of the platform by solar radiation on the buoys or ships and, in the latter case, also the ship's engine heating the cooling water and general heat diffusion from the bulk of the ship. For the satellite infrared observations, effects of atmospheric aerosols and clouds in the field of view are persisting problem areas, which have regional variations whose corrections have not been readily available. Table 2 compares in situ and satellite accuracies.

As the much relied-on SST product of NCEP is produced with a fit of the satellite information to a "calibration surface" generated with the available in situ data, inaccuracies develop when the sampling coverage of the in situ data is inadequate. This is the case in large regions of the Southern Hemisphere oceans, where the "fit" may become unreliable and can even *generate* errors. The presence of sea ice can also confuse the correspondence between in situ and satellite observations (Reynolds, 1999, personal communication).

Another difficulty with SST determination is that a satellite instrument senses the radiation from the top 0.5 mm or less of the sea surface (Katsaros, 1980; Robinson et al., 1984), while a buoy or ship measures the temperature from 0.5 to 5 m depth. Since the oceans lose heat to the atmosphere from the sea surface, there is typically a negative gradient in temperature toward the surface mainly confined to the nonturbulent region nearest the interface. We refer to the layer exhibiting the temperature difference, ΔT, across it as the "cool film". Many processes affect the magnitude of the ΔT, as illustrated in Figure 1. The ΔT values of the order of -0.3 to -0.7°C have been observed on the open ocean (Schlüssel et al., 1990; Wick et al., 1996).

TABLE 2 Sea Surface Temperature Uncertainties and Absolute Accuracies

	Intrinsic Errors	
	Random Error	Absolute Error
In situ sensors		
Thermistor	±0.1	Of the order of ±0.5
Resistance	±0.1	
Satellites		
Polar orbiting	±0.5	Of the order of ±0.7
Geostationary	±0.5	

However, on a calm sea during the day, the uppermost layer can be warmer than the subsurface layers by several degrees Celsius due to solar heating. This is not a common occurrence but has been observed, particularly in the tropics, and even from space (Katsaros et al., 1983). It is more likely to occur where fresh water added by heavy rain may stabilize the surface layer (Wijesekera et al., 1999).

Estimates of temperature drop across the "cool" film on the sea surface would be required to correctly relate the buoy in situ SST to the satellite observations. One proposed method would be to apply a simple model for the estimated ΔT across the cool film (Soloviev and Schlüssel, 1994).

The uncertainties in both the satellite and surface SST measurements are of the same order of magnitude as the typical temperature drop across the cool film, but since the cool film is present over most of the ocean most of the time, it could be argued that ignoring it introduces a bias or additional uncertainties. Most heat flux parameterizations and other uses for SST information have been based on the bulk value of SST provided by buoys and ships, so it is wise to consider that the satellite information represents the temperature at the interface.

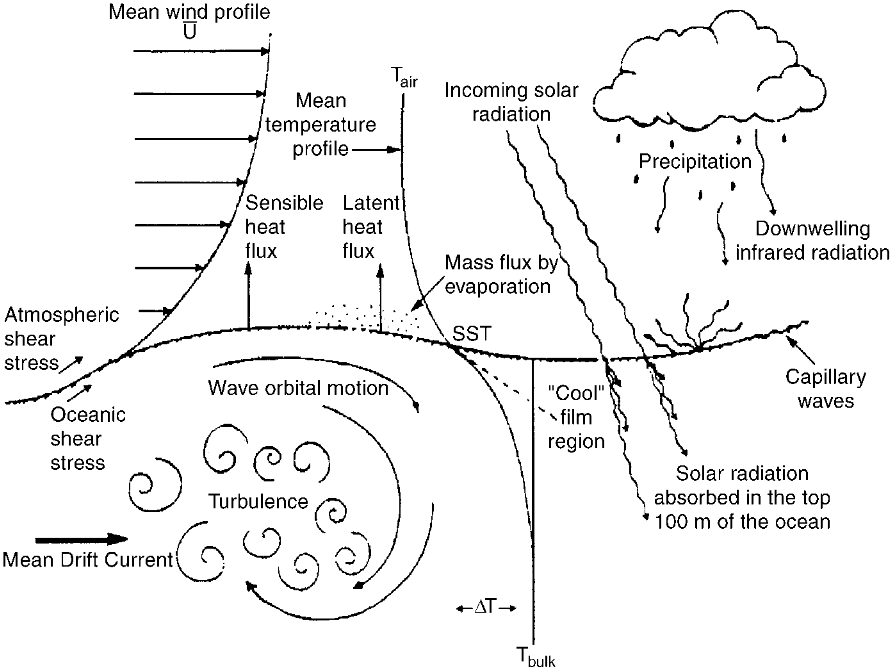


Figure 1 Illustration of the "cool" film and the many processes that affect the temperature difference across it. These processes include variable shear stress on the surface, radiative, sensible and latent heat fluxes, precipitation, wave breaking, and the presence of oil films and slicks (after Katsaros, 1980).

2 SURFACE WIND

Another sea surface variable observable by in situ sensors or from satellites is the surface wind. The classical sea surface wind speed observations were qualitative descriptions by mariners of the appearance of the sea surface as it was roughened by the wind. The observations were reported via the Beaufort scale (Weather Bureau, 1948; Shaw and Simpson, 1906). This scale has since been calibrated versus direct measurements by anemometers. Table 3 gives the Beaufort scale and conversions. The wind direction was determined from the directions in which the shorter gravity waves were traveling with respect to the ship's compass.

The replacement of these qualitative observations by anemometers on ever larger and larger ships has created much confusion and, perhaps, more erroneous data because to provide unaffected exposure of an anemometer on a bulky, moving ship is almost impossible due to severe distortion of the flow both in speed and direction (Yelland et al., 1998). Modern research ships carry two anemometers to provide alternate sensors to choose from, depending on the wind direction relative to the ship's heading.

Anemometers on buoys have the problem of being "pumped" by the rocking of the buoy in the wave field . The cup anemometers rectify the buoy motion, i.e., the anemometer rotation is increased, whether the buoy is in the "to" or "fro" cycle of its rocking, while propeller anemometers can turn in both directions and, therefore, average out the effect of the rocking. Propeller anemometers are, therefore, used almost exclusively on buoys (exceptions are stable buoys, such as the spar buoys, which do not follow the waves). A directional vane also suffers errors due to buoy motion.

The satellite method for measuring the fleeting wind is either by microwave backscatter from the sea surface by an instrument called a scatterometer (Ulaby

TABLE 3 Beaufort Scale for Wind Speed and Its Conversion to Knots and Meters per Second (m/s)

Beaufort Number	International Description	Wind Speed (knots)	Wind Speed (m/s)
1	Light air	1–3	0.5–1.5
2	Light breeze	4–6	2–3
3	Gentle breeze	7–10	4–5
4	Moderate breeze	11–16	6–8
5	Fresh breeze	17–21	9–10
6	Strong breeze	22–27	11–13
7	Moderate or near gale	28–33	14–17
8	Gale or fresh gale	34–40	18–20
9	Strong gale	41–47	21–24
10	Whole gale or storm	48–55	25–28
11	Violent storm	58–63	29–33
12	Hurricane	>64	>33

et al., 1982) or by variations in the emissivity (for microwave radiometers). Scatterometers measure the diffuse backscatter of a radar signal from the sea surface with two to three antennas, or by a rotating antenna, such that the same piece of ocean is observed several times from different look angles, e.g., the European Remote Sensing Satellite, ERS, scatterometer, the NASA scatterometer NSCAT, and QuikSCAT, so named for its short cycle from planning to launch. The difference in the backscatter power at the different look angles allows inference on the wind direction, as well as the wind speed. The *Seasat*, *Nimbus 7*, and SSM/I microwave radiometers only provide the wind speed as a consequence of the increased emissivity and presence of foam as the wind speed increases. These sensors were pioneered in the early satellite days, but saw their main debut with the *Seasat* satellite launched in 1978 (Jones et al., 1979; Lipes et al., 1979). The satellite wind measurement techniques have in common with the qualitative Beaufort scale that it is the effect of the wind on the sea that is being observed. The reason that this works for the satellite radars of wavelengths between a few to tens of centimeters is that the surface waves that interact with the electromagnetic radiation (via the Bragg scattering mechanism) are the capillary-gravity waves, which have a short life time and are closely linked to the wind speed (or more exactly the wind stress).

After a long hiatus of 13 years, a scatterometer was first launched again in 1991 by the European Space Agency and in 1996 by the National Aeronautics and Space Administration in cooperation with the Japan Space Agency. The *Seasat* instruments, as well as the instruments of the 1990s, were calibrated against buoy anemometer measurements (Bentamy, 1998; Graber et al., 1996), as well as against numerical weather prediction analyses with the buoy calibration having a somewhat more realistic wind speed range; atmospheric numerical models tend to smooth the wind variations and, therefore, lose the extremes. However, for very low wind speeds and for wind speeds >25 m/s, serious questions about the representativeness of the buoy observations are also emerging (Large et al., 1995; Bentamy et al., 1996).

The workhorse for satellite surface wind estimates from 1987 to the present has been the microwave radiometer, the Special Sensor Microwave Imager (SSM/I), partly for lack of a scatterometer in space. Microwave radiometry is less well adapted to satellite wind observations due to interference by heavy clouds and rain, exactly the condition of paramount meteorological interest. The active systems, scatterometers [and other radars such as altimeters and synthetic aperture radars (SARs)], have more signal power and are, therefore, better able to penetrate clouds. A new polarimetric microwave radiometer is expected to overcome this difficulty (but has not yet been flown on a satellite).

Again, sampling is the major problem for ship, buoy, and satellite systems. Reliance on one system alone is often unsatisfactory because of the large temporal variations of the wind. If a measurement at one specific place in the ocean is sought, then the data from an anemometer on a moored buoy will suffice, but if horizontal variability of the wind is desired, a combination of satellite and observations with an array of buoys may best serve the purpose.

Another important challenge of wind analysis lies in how to join various wind measurements together. The *surface wind* is typically defined to apply at 10 m height above the sea (this height is an arbitrary choice and others have been in use). Ships and buoys rarely have their anemometers at 10 m height. To make measurements intercomparable, we convert the wind speed to a common height by applying a correction to an individual measurement at height z to translate it vertically to the 10 m or other desired reference height. For that calculation, we must adopt an analytical vertical wind profile that depends on the atmospheric stratification, which is a function of the turbulence, which, in turn, depends on the wind stress and the buoyancy. For a neutrally stratified atmospheric boundary layer (no buoyancy fluxes or very strong wind shear), the profile is logarithmic. For stable and unstable stratification (negative or positive buoyancy fluxes) caused by sensible heat and/or water vapor flux, the profile differs from the logarithmic shape (Kraus and Businger, 1994).

Methods to provide the height correction have been established (Liu et al., 1979; Fairall et al., 1996), mostly derived from simultaneous flux and profile measurements over land. The application of the established profile shapes to the ocean have not yet been fully proven due to the difficulty in measuring profiles from moving or bulky platforms or the effects of the wave field on the lowest level atmospheric measurements for stationary towers.

The satellite measurement of the wind vector and wind speed depends on the roughening of the sea surface by the wind, particularly by the centimeter-scale gravity-capillary waves. These are forced by the wind stress exerted on the surface, and the stress is directly related to the wind profile. Thus, the fields of air–sea interaction, wave generation, and electromagnetic backscatter theory are involved in the remote measurement of surface wind. The enterprise is saved from the whole of this complexity under most circumstances by the footprint size of the radar data (12.5 to 50 km), which assures a certain averaging over the variability and, thereby, a randomization of the errors. However, in regions of large wind speed gradients and variable wave fields, this may not be true.

The measurements of SST and surface wind are now well developed and are only being refined by better sampling techniques, data recordings, and inclusion of auxiliary measurements to improve the corrections. An example is simultaneous measurements on recent (and future) satellites of the aerosol content in the atmosphere, which allows the elimination of the radiation error caused by the aerosol via radiative transfer models. Therefore, we can now embark on producing long, consistent time series of both SST and sea surface wind for climatological studies. The crux is in keeping the long time series consistent with proper intercalibration between sequential satellite sensors and when some inevitable design improvements are made.

REFERENCES

Barton, I. J. (1995). Satellite-derived sea surface temperatures: Current status, *J. Geophys. Res.* **100**, 8777–8790.

Bentamy, A., Y. Quilfen, F. Gohin, N. Grima, M. Lenaour, and J. Servain (1996). Determination and validation of average wind fields from ERS-1 scatterometer measurements, *Global Atmos. Ocean Syst.* **4**(1), 1–29.

Bentamy, A., N. Grima, and Y. Quilfen (1998). Validation of the gridded weekly and monthly wind fields calculated from ERS-1 scatterometer wind observations, *Global Atmos. Ocean Syst.*, **5**, 373–396.

Bushnell, M. H. (1996). Preliminary Results from Global Lagrangian Drifters Using GPS Receivers, WMO/DBCP Technical Doccument No. 10, pp. 23–26.

Fairall, C. W., E. F. Bradley, J. S. Godfrey, G. A. Wick, J. B. Edson, and G. S. Young (1996). Cool-skin and warm-layer effects on sea surface temperature, *J. Geophys. Res.* **101**(1), 1295–1308.

Graber, H. C., N. Ebutchi, and R. Vakkayil (1996). Evaluation of ERS-1 scatterometer winds with wind and wave ocean buoy observations, *Tech. Rep.*, *RSMAS 96-003*, Div. of Applied Marine Physics, Univ. of Miami, FL.

Jones, W. L., P. G. Black, D. M. Boggs, E. N. Bracalente, R. A. Brown, G. Dome, J. A. Ernst, I. N. Alberstan, J. E. Overland, F. Peteherych, W. J. Pierson, F. J. Wentz, P. M. Woicefiyn, and N. J. Wurtele (1979). SEASAT scatterometer: Results of the Gulf of Alaska workshop, *Science* **204**(4400), 1413–1415.

Katsaros, K. B. (1980). The aqueous thermal boundary layer, *Boundary-Layer Meteorol.* **18**, 107–127.

Katsaros, K. B., A. F. G. Fiuza, F. Sousa, and V. Amann (1983). Sea surface temperature patterns and air–sea fluxes in the German Bight during MARSEN 1979, Phase 1, *J. Geophys. Res.* **88**(C14), 9871–9882.

Kraus, E. B., and J. A. Businger (1994). *Atmosphere-Ocean Interaction*, Oxford University Press, New York.

Large, W. G., J. Morzel, and G. B. Crawford (1995). Accounting for surface wave distortion of the marine wind profile in low-level ocean storms wind measurements, *J. Phys. Oceanogr.* **25**(11), 2959–2971.

Lipes, R. G., R. L. Bernstein, V. J. Cardone, K. B. Katsaros, F. G. Njoku, A. L. Riley, D. B. Ross, C. T. Swift, and F. J. Wentz (1979). SEASAT scanning multichannel microwave radiometer: Results of the Gulf of Alaska workshop, *Science* **204**(4400), 1415–1417.

Liu, W. T., K. B. Katsaros, and J. A. Businger (1979). Bulk parameterization of air–sea exchanges of heat and water vapor including the molecular constraints at the interface, *J. Atmos. Sci.* **36**(9), 1722–1735.

McClain, E. P., W. G. Pichel, and C. C. Walton (1985). Comparative performance of AVHRR-based multichannel sea surface temperatures, *J. Geophys. Res.* **90**, 11,587–11,601.

McPhaden, M. J., A. J. Busalacchi, R. Cheney, J.-R. Donguy, K. S. Gage, D. Halpern, M. Ji, P. Julian, G. Meyers, G. T. Mitchum, P. P. Niiler, J. Picaut, R. W. Reynolds, N. Smith, and K. Takeuchi (1998). The Tropical Ocean-Global Atmosphere observing system: A decade of progress, *J. Geophys. Res.* **103**, 14,169–14,240.

Reynolds, R. W. (1993). Impact of Mount Pinatubo aerosols on satellite-derived sea surface temperatures, *J. Climate* **6**(4), 768–776.

Reynolds, R. W., and T. M. Smith (1994). Improved global sea surface temperature analyses using optimum interpolation, *J. Climate* **7**(6), 929–948.

Robinson, I. S., N. C. Wells, and H. Charnock (1984). The sea surface thermal boundary layer and its relevance to the measurement of sea surface temperature by airborne and spaceborne radiometers, *Int. J. Remote Sens.* **5**, 19–45.

Schlüssel, P., W. J. Emery, H. Grassl, and T. Mammen (1990). On the bulk-skin temperature difference and its impact on satellite remote sensing of sea surface temperature, *J. Geophys. Res.* **95**(C8), 13,341–13,356.

Shaw, W. N., and G. C. Simpson (1906). *The Beaufort Scale of Wind Force: Report of the Director of the Meteorological Office upon an Inquiry into the Relation between the Estimates of Wind-Force According to Admiral Beaufort's Scale and the Velocities Recorded by Anemometers Belonging to the Office, with a Report Upon Certain Points in Connection with the Inquiry*, Darling and Son, London.

Soloviev, A. V., and P. Schlüssel (1994). Parameterizations of the cool skin of the ocean and of the air-ocean gas transfer on the basis of modeling surface renewal, *J. Phys. Oceanogr.* **24**(6), 1339–1346.

Swenson, M. S., and P. P. Niiler (1996). Statistical analysis of the surface circulation of the California Current, *J. Geophys. Res.* **101**(C10), 22,631–22,645.

Ulaby, F. T., R. K. Moore, and A. K. Fung (1982). *Microwave Remote Sensing, Active and Passive, Vol. II; Radar Remote Sensing and Surface Scattering and Emission Theory*, Addison-Wesley, Reading, Mass.

Weather Bureau (1948). Beaufort Scale of Wind Force, WB Form 4042A, Washington, DC.

Wick, G. A., W. J. Emery, L. H. Kantha, and P. Schlüssel (1996). The behavior of the bulk-skin sea surface temperature difference under varying wind speed and heat flux, *J. Phys. Oceanogr.* **26**(10), 1969–1988.

Wijesekera, H. W., C. A. Paulson, and A. Huyer (1999). The effect of rainfall on the surface layer during a westerly wind burst in the western equatorial Pacific, *J. Phys. Oceanogr.* **29**(4), 612–632.

Yelland, M. J., B. I. Moat, P. K. Taylor, R. W. Pascal, J. Hutchings, and V. C. Cornell (1998). Wind stress measurements from the open ocean corrected for airflow distortion by the ship, *J. Phys. Oceanogr.* **28**(7), 1511–1526.

CHAPTER 48

RADAR TECHNOLOGIES IN SUPPORT OF FORECASTING AND RESEARCH

JOSHUA WURMAN

1 INTRODUCTION

Forecasters and researchers, and the computer models that support them, require detailed information concerning the three-dimensional state of the atmosphere and how it evolves with time. Some forecasting and research needs overlap; others are particular to those respective tasks. But whether the goal is weather prediction or scientific understanding, no tool can provide more wealth and diversity of information than existing and emerging technologies in weather radar (Fig. 1). Weather radars can sample large volumes of the atmosphere nearly continuously in space and time at fairly high resolution vertically, horizontally, and temporally. Radars can penetrate through clouds and precipitation to measure precipitation intensity, precipitation type, and wind motions in many weather conditions, throughout the depth of the atmosphere. They are a uniquely versatile tool.

A forecaster needs to know where it is raining or snowing, how hard, and whether snow or hail is occurring. He needs to know whether the precipitation is caused by convective systems or more stratiform uplift, whether it is moving toward the forecast area or away, whether it is strengthening or weakening. He needs to know if an airport runway will soon be affected by a gust front or microburst, or if a mesocyclone will move into his forecast area, whether a snow band will move onshore, or whether a sea breeze will initiate storms. For many forecasts of 6 h or less, radar is the crucial tool that provides the most up-to-date and comprehensive information to a forecaster. Mesoscale forecasting models

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

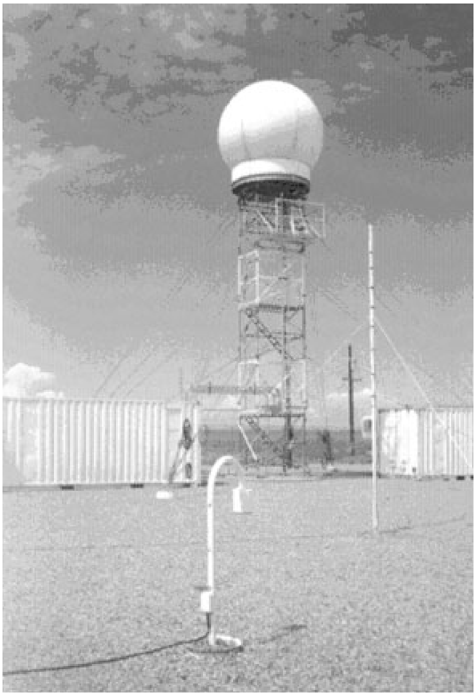


Figure 1 The MIT 5-cm (C-band) research radar deployed in Albuquerque, New Mexico. The nearly spherical radome protects the radar antenna and sits on top of a tower that places the antenna above blocking objects. The transmitter, receiver, and operating consoles are located in the buildings below the tower. (Photo courtesy of MIT Weather Radar Lab, D. Boccippio.) See ftp site for color image.

of the future will become more and more dependent on initializations and nudgings provided by the nation's radar network.

Research needs are similarly diverse. A scientist may want to study large-scale rainfall patterns, not only at weather stations, but in between, or to study much smaller scales, down to suction vortices in tornadoes or rolls in microbursts. He may want to understand the microphysical processes that cause hail, icing, or charging of particles and lightning, or to study the flow along a front or dry line, or the wind field of a hurricane rain band. Researchers use radar to probe the upper atmosphere and the boundary layer, intense and violent weather, and quiescent overturning on clear days. Almost all atmospheric phenomena with scales of 10 m to 100 km have been and are being intensively studied using radar.

2 BASIC RADAR OPERATION

This chapter is not a radar textbook, but it is valuable to briefly review some basics of the theory and operation of radars to understand the nature and quality of the data that they produce, the limitations of these data, and potential uses both now and in the future. The following discussion will not be exhaustive and will cover only some major technologies.

Most weather radars focus pulsed beams of microwaves on meteorological targets and listen for returned signals. (Not all radars operate in this fashion. But the exceptions are primarily fairly exotic research systems.) By measuring the strength, timing, and other parameters of these signals, information about the targets can be obtained. A pulse of focused radiation, typically about 1 to 2 μs in duration, leaves the radar (Fig. 2). At various points along its travel, the pulse encounters precipitation, insects, suspended particles, airplanes, mountains, or density discontinuities in the air and some of the radiation is scattered back toward the radar and elsewhere, while the remainder continues to travel outward. After the pulse leaves the radar, hardware and software listen for returned energy. The delay time between transmission and return uniquely defines the distance to the precipitation that caused the scattering, $R = c\ \Delta t/2$, where R is the distance to the precipitation, c is the speed of light, and Δt is the time between transmission and return of the energy. The returned signals are grouped or sampled at intervals, typically about 1 to 2 μs, resulting in distance resolution, using the above formula, of 150 to 300 m. After approximately 0.5 to 2 ms, another pulse is transmitted and the process is repeated. The strength of

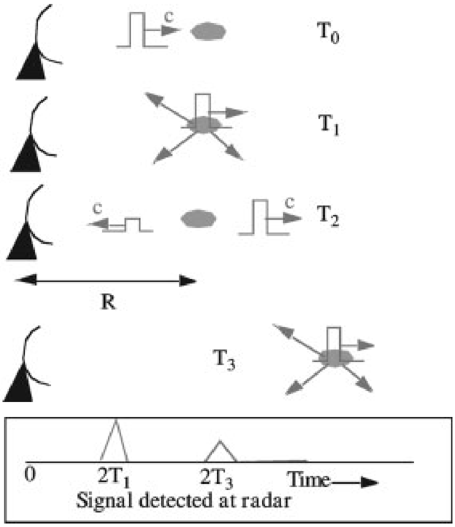


Figure 2 Pulse of microwaves travels from a radar toward a raindrop at T_0. At T_1, the pulse encounters the raindrop, which acts as an antenna and radiates in many directions. At T_2, the transmitted pulse continues outward while some of the microwaves emitted by the raindrop continue back toward the radar. The microwaves travel at the speed of light, c, from the radar to the drop and back again, so the distance to the drop can be calculated from the round-trip time, $2T_1$. Radiation emitted by drops encountered further along the transmitted pulse's travel T_3, will return to the radar at later times. See ftp site for color image.

the returned signals can be used to estimate the intensity of precipitation that causes the scattering, using what is known as the radar equation. The difference in phase of the returned radiation from subsequent pulses is used to calculate the motion of the precipitation. (Most radars do not, in fact, actually measure the Doppler shift of returned radiation.) The transmitting antenna rotates, usually horizontally, surveying a wide area, then inclines and repeats the rotation, surveying a similar area at greater elevation. Some of these processes will be discussed further below.

Choice of Wavelength: Attenuation Versus Antenna Size

A typical weather radar generates microwaves with a wavelength of 3 cm (X-band), 5 cm (C-band), or 10 cm (S-band). Other wavelengths are used for specialized research purposes as will be discussed later. (For comparison, a fluorescent light bulb generates energy at about 500 nm, an FM radio station 300 m, and a cordless phone at 40 cm.)

All other things being equal, it is usually best to use the longest practical wavelengths for a weather radar. This is because when a radar beam of microwaves passes through precipitation, it becomes attenuated due to scattering by the raindrops. Essentially, the beam is consumed by the process of scattering (back toward the radar and in other directions), and some is absorbed and converted to heat. (Not much heat, however; you cannot evaporate a cloud with a normal weather radar.) Shorter wavelength radiation suffers much more from attenuation; the effect is proportional to worse than the inverse square of wavelength $1/\lambda^2$. Precisely calculating how much a beam will be attenuated is difficult since complex scattering effects and absorption must be taken into account, but a rough approximation is possible. The attenuation rate is roughly proportional to the intensity of rain occurring in the cloud. A beam will lose a certain fraction of its intensity each kilometer: attenuation (dB/km) $= CR$ where R is in mm/h and $C = 0.01$ (3 cm), 0.02 (5 cm) or 0.003 (10 cm). So, if the rain rate is 10 mm/h, a 5-cm radar beam will lose $0.02 \times 10 = 0.2$ dB/km, a fairly significant amount since the effect is cumulative during the outward and return travel of the beam. After passing through 50 km of 10 mm/h rain and returning to the radar, the measured signal will be 20 dB weaker than otherwise expected. There are techniques for correcting for this attenuation, but they are prone to large errors. The velocity data from attenuated beams can still be good, but, in intense rain or hail, the beams can become mostly or completely extinguished, complicating or preventing the retrieval of useful data. In heavy rain all data from ranges beyond 20 km can be lost if a 3-cm radar is used. Additional attenuation can be caused by heavy rain forming opaque sheets of water on the radomes that protect antennas from the weather, sun, and wind.

But, all other things are not equal. Due to the physics of diffraction, the ability of an antenna to focus radiation into a narrow beam is proportional to its size. A large antenna can produce a much more focused beam than a small antenna. The focusing power of an antenna is also dependent, inversely, on the wavelength being transmitted, with longer wavelengths being more difficult to concentrate into a narrow beam. Thus, the beamwidth of the transmitted energy is roughly equal to $80\lambda/D$,

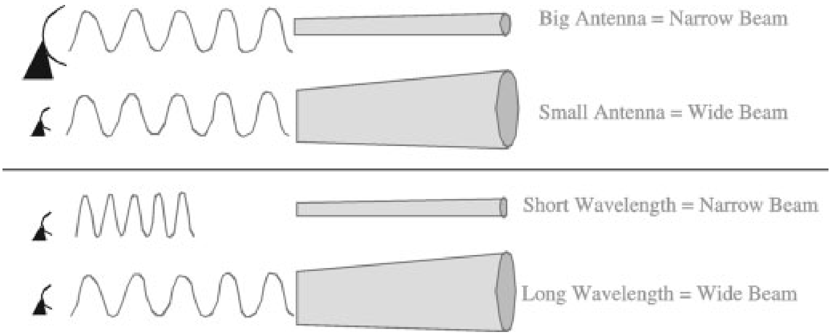

Big Antenna = Narrow Beam
Small Antenna = Wide Beam
Short Wavelength = Narrow Beam
Long Wavelength = Wide Beam

Figure 3 How beamwidth is affected by antenna size and wavelength. See ftp site for color image.

where λ is the radar wavelength and D the antenna diameter. As a result, it requires a 8-m diameter antenna to produce a 1° wide beam of 10-cm radiation, but only a 2.4-m antenna if 3-cm microwaves are used (Fig. 3). The area, weight, wind resistance to rotation, and the power required to turn an antenna are proportional to the square, or worse, of its diameter. The design trade-off is between penetration ability, better at 10 cm, versus low cost and logistical ease, better with smaller antennas transmitting at 3 cm. When the current operational weather radar network in the United States was constructed, a significant investment was required to produce radars with 1° beamwidths using 10-cm wavelengths. Each radar used a 8.5-m diameter antenna inside a 12-m radome, and cost approximately $5 million, requiring significant infrastructure, power, and maintenance. The benefit is that the network of WSR-88D radars can penetrate through many kilometers of intense rain and hail. Many other countries, researchers, and media have chosen lower cost, shorter wavelength options with the limitation of only moderate or poor penetration ability.

In certain specialized cases, other factors enter into the choice of wavelength, notably with mobile and airborne systems as will be discussed later.

Transmitter Type

Weather radars usually use one of two basic types of transmitter, magnetron or klystron, though some are designed with other mechanisms. Klystron transmitters are most expensive but produce very stable and coherent signals. Transmitted radiation varies very little in frequency and the phase of the radiation is synchronous from pulse to pulse. Until very recently, klystron radars were far superior in their ability to measure velocities since the phase of returned radiation is used in these calculations. Cheaper magnetron transmitters produce radiation that exhibits random phase from pulse to pulse, and the frequency drifts rapidly within a narrow band from second to second. In the past, many radars that did not attempt to measure precipitation motion (non-Doppler radars) used magnetron transmitters. Today, however, almost all

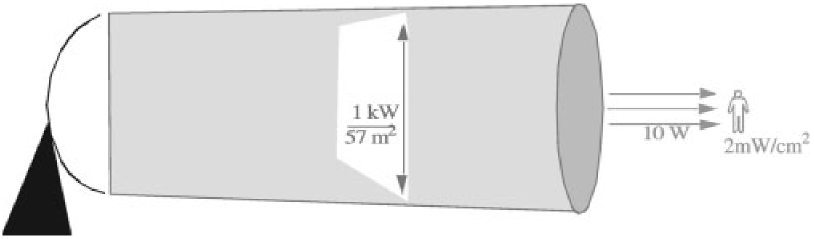


Figure 4 Intensity of microwaves emitted by typical weather radar. A person standing directly in front of a typical weather radar would absorb 10 W of microwave radiation with an exposure of 2 mW/cm^2. See ftp site for color image.

weather radars in the United States, including magnetron systems, measure precipitation motion (Doppler) using modern hardware or software techniques.

A more intense transmitted beam results in more energy scattered back toward the radar, permitting the measurement of the properties of smaller or more tenuous weather targets. More intense beams also penetrate further into precipitation, despite attenuation. As with antenna and frequency choice, there are trade-offs involving transmitter strength. It costs more to construct high-intensity transmitters. It requires more power to transmit large amounts of energy. Also, intense electric fields can be hazardous and can complicate system design requiring pressurization of parts of the system with common or exotic gasses (SF_6) to prevent arcing. Transmitter strength is usually reported in terms of peak power, the power transmitted during the short pulses. Typical weather radars transmit pulses with 40 kW to 1 MW power. Some military systems use much more; some specialized research radars much less. Since the transmitter is off most of the time between pulses (e.g., on 1 μs, then off 999 μs, . . .), the average transmitted power is about 500 to 1000 times less, typically 100 W (like a light bulb) to 1 kW (like a hair drier). U.S. federal safety standards prohibit exposure to more than 10 mW/cm^2 average power (other nations have similar standards, but with different levels of allowable exposure). If a person stood immediately in front of a WSR-88D 1000 W radar with a 8.5-m diameter antenna (area 57 m^2), he or she would be exposed to approximately 2 mW/cm^2, well below the safety limit (Fig. 4). The person (cross-sectional area about 0.5 m^2) would intercept about 1% of the radar beam, or about 10 W total. The intensity at the center of the beam might be a couple of times higher. However, if one were to place a hand over the feedhorn (area < 0.05 m^2), much higher levels would be experienced.

Scattering

When the pulse of radiation impinges on raindrops in its path, scattering occurs. This is a complex electromagnetic process but, with some approximations, some simple statements can be made. When radiation interacts with a raindrop, it excites the water molecules in the drop. The molecules become electrically polarized. The radar beam causes this polarization to change orientation rapidly, at the radar frequency.

Thus, opposite sides of the drops become charged one way, then the next, billions of times per second. The drops become miniature antennas, analogous to the antenna of a walkie talkie, and radiate microwaves outward. Some of this radiation is radiated back toward the transmitting radar. (In drops that are similar in size to the transmitted radiation, e.g., hailstones, this process is considerably more complicated.)

The amount of radiation that a drop emits is proportional to D^6, the sixth power of the drop's diameter, so a 5-mm diameter drop radiates 15,000 times more than a 1-mm drop and 10^8 times more than a 50-μm cloud droplet. This complicates the interpretation of returned data since a single 5-mm drop returns as much energy as 15,000 1-mm drops, but the latter contain 125 times as much mass (Fig. 5). Therefore, it is difficult to know whether a particular strength of radar echo is due to a few large drops or a plethora of smaller ones. A particular amount of water mass scatters more microwaves if the water is contained in a few large drops. Raindrops are more efficient radiators than ice particles, scattering about 5 times as much than equivalent diameter ice particles. The D^6 approximation, called the Rayleigh approximation, applies only to drops that are much smaller than the radar wavelength. So, when short wavelength radars are used or when hail is present, more complicated formulations must be used.

Scattering can occur off other airborne objects, most notably insects and birds. Being mostly composed of water, these animals scatter in a fashion similar to raindrops of similar diameter. Importantly, however, they may not be passively moving with the wind. Bird echoes, moving at a different velocity than the air, can cause significant contamination to measured wind fields. Sometimes the military or researchers will release small strips of aluminized mylar or other objects into the air, called chaff. In the former case this is to confuse enemy radars, in the latter to provide passive scatterers carried by the wind in regions of little natural scattering. Scattering will occur where there is any contrast in the refractive index, usually caused by density changes, in the air. So, where turbulence is mixing cold and warm air parcels, or along precipitation-free dry lines or gust fronts, some energy is reflected back toward the radar by this process, called Bragg scattering. Researchers use these signals to observe the edges of clouds and other phenomena.

Scattering or reflections also occur when radar energy hits land, vegetation, or water surfaces. These echoes, called ground clutter or sea clutter, can be very intense and overwhelm the signals from the raindrops in the air near them. Often, clutter signals are filtered out by software that effectively blocks signals that have very low velocity and are presumed to originate from stationary objects. But this technique

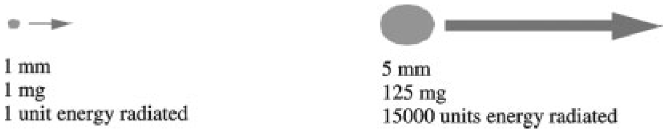


Figure 5 Relative diameter, mass, and scattered energy from small and large drops. Large drops scatter much more energy than small drops. See ftp site for color image.

does not work well with sea clutter contamination. Much of the clutter contamination arises from scattering from stray radiation that is not perfectly focused by the antenna into a narrow beam. This energy, called side-lobe radiation, hits objects in all directions and is scattered back to the radar. Since side-lobe energy scattered back by a very strong radar target, like a mountain or water tower, can be stronger than the weak signals scattered back by raindrops arriving back at the radar at the same time, this can cause significant contamination to weather radar data.

Propagation Paths

The narrow beam produced by a radar spreads with increasing distance from the radar. Even a 1° beam is 160 m wide at 10-km range and 1.6 km wide at 100-km range. At the maximum range measured by typical radars, 200 to 300 km, the beam may have spread to 3 to 5 km in extent. This critically affects the ability of the radar to detect weather, since objects smaller than the beamwidth cannot be resolved and only objects several times larger than a beamwidth can be accurately measured. Thus, microbursts and tornadoes are often difficult to detect at great range. Some very specialized research radars use extremely large antennas or short wavelengths to produce ultra-narrow beams, but this is not practical for most weather radar applications.

To a first approximation, radar beams travel in straight lines. Since Earth is curved, this means that a radar beam aimed at the horizon will soon become significantly raised from the surface and eventually depart into space (Fig. 6). This means that objects behind the horizon cannot be detected, preventing the resolution of near surface weather beyond a limited range. Fortunately, the atmosphere is more dense near the ground, resulting in an index of refraction gradient that bends radar beams partially back toward the curving surface of Earth, permitting some over-the-horizon visibility. The approximate height of the center of a beam aimed 0.5° above the ground, in "average" weather conditions, is 1.5 km at 100-km range and 3.5 km at 200-km range. The bottom of the beam would be approximately 400 m and 1.5 km above the ground at the same ranges.

Sometimes the gradient of atmospheric density is so high that it can bend the radar beams back into the earth. This can occur if very cold dense air lies near the surface. In this case, called anomalous propagation, energy will reflect off the ground or water surface, some back toward the radar.

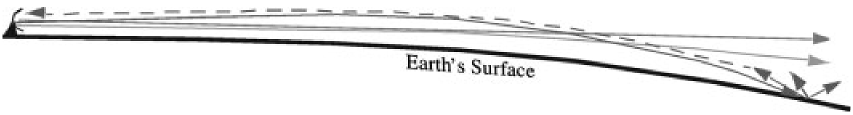


Figure 6 (see color insert) Beam paths assuming straight propagation (red), typical atmospheric density gradient bending beam partially back toward Earth (blue), strong density gradient, possibly temperature inversion, bending beam back into Earth (green) where scattering off surface sends energy back toward radar. See ftp site for color image.

Data Processing

Once the transmitter generates a pulse, it is focused by the antenna, interacts with objects in its path, and scattered energy returns to the radar. The returned signal must be converted into useful meteorological data. This is accomplished by the radar receiver and signal processing system. This is one of the most complex, varied, and rapidly evolving areas of radar technology. The basic concepts are relatively straightforward, however.

Radar Gates Most radars digitize (sample) the received signals. The digitalization rate determines the gate size and is one determiner of the resolution of a radar. If the returned signals are digitized at a rate of 1 MHz, or every 1 μs, the gate size will be 150 m ($c\ \Delta t/2$). Faster sampling will result in shorter gates. However, sampling intervals less than the duration of a pulse of the radar have diminishing added utility since the length of the transmitted pulse effectively blurs the returned signals and is another determining factor in true radar resolution. Frequently, the pulse length and gate length are matched.

Reflectivity The amount of power that returns to the radar from any scattering volume (defined by the beamwidth and sampling interval) is dependent on the amount of energy that impinges on the volume, the nature, number, size, shape, and arrangement of the scattering particles, radar wavelength, and distance to the weather target, attenuation, and other factors. These are related through the radar equation, which appears in many forms, but can be simplified to $P_r = CZ_e/R^2$, where P_r is the returned power, C is called the radar constant and contains all information about the transmitter, pulse length, antenna, wavelength, etc., R is the distance to the target, and Z_e is equivalent radar reflectivity factor, more commonly referred to as Z, or reflectivity. Z is a rather strange parameter; it has units of volume (mm^6/m^3) and it is usually expressed in terms of 10 times its base 10 logarithm, or $\text{dB}Z = 10\ \log_{10} Z$. The amount of Z that would be measured from a raindrop is proportional to D^6, the sixth power of the drop diameter. The Z measured from a volume of drops is thus $\Sigma N_i D^6$, where N_i is the number of drops of each diameter in the volume. Because large drops are much more effective radiators, a certain value of Z can be due to a very small number of large particles or a large number of small particles; it is impossible to tell which by using Z alone.

It is difficult to precisely relate Z values to meteorologically useful quantities like liquid water content or rain rate. This is because it is dependent on the sum of the sixth power of raindrop sizes, not the sum of the masses of the raindrops. Numerous theoretical and empirical relationships, called Z–R relationships, exist to convert between Z, rain rate (R) and other quantities. Very roughly, 15 dBZ corresponds to light rain, 30 dBZ to moderate rain of several mm/h, 45 to 50 dBZ to 50 mm/h, 50 to 57 dBZ to 100 mm/h, and higher dBZ levels, 55 to 70, to hail or rain/hail mixes.

Typically Z is averaged over many pulses, 32 to 256, since it can vary greatly due to constructive and destructive interference from the radiation emitted from each drop in the illuminated volume. It is necessary to obtain several “independent” measurements to calculate an accurate value of *Z*. Independence means that the particles in the illuminated volume have reshuffled so that their arrangement is effectively decorrelated with their arrangement during the passage of the previous pulse. It can require a time spacing of several pulses before independence occurs, so the measurement of *Z* cannot take full advantage of all the 32 to 256 pulses mentioned above. The time to independence is shortest when there is high turbulence and/or short (i.e., X-band) transmissions. It is very difficult to calculate the radar constant, *C*, accurately, and the measurement of *Z* is prone to errors of approximately $\pm 2\,\mathrm{dB}$. This can be very significant since small changes of *Z* can result in large differences in predicted *R*, particularly at high *Z* and *R*, where one cares the most.

The minimum power that a typical weather radar can measure is about 10^{-14} W, which corresponds to about $-5\,\mathrm{dB}Z$ at a range of 50 km. This depends on the wavelength, antenna, quality of electronics, pulse length, number of pulses per average, etc. Though rarely an issue except in very close range research applications, there is a maximum power that can be detected before radar hardware/software saturates and is effectively blinded. This is seldom realized except in heavy rain within a few kilometers of a radar.

Doppler Velocity Most weather radars can measure the component of scatterer motion toward or away from the radar. While these radars are usually called Doppler, most do not directly use the Doppler effect to measure this motion. Typically the radar measures the path length to a raindrop (actually the sum of the path lengths to a volume of raindrops) during subsequent pulses (actually the remainder, noninteger portion) to calculate the motion (Fig. 7). The most common calculation technique is called pulse-pair processing, whereby the phase of the returned energy from each pulse is measured. Another technique called spectral processing, or Fourier processing, can be used also. While an acceptable velocity measurement can be made using just two pulses (in stark contrast to the several independent measurements needed to get an accurate *Z* measurement), typically many pulses are averaged to reduce error.

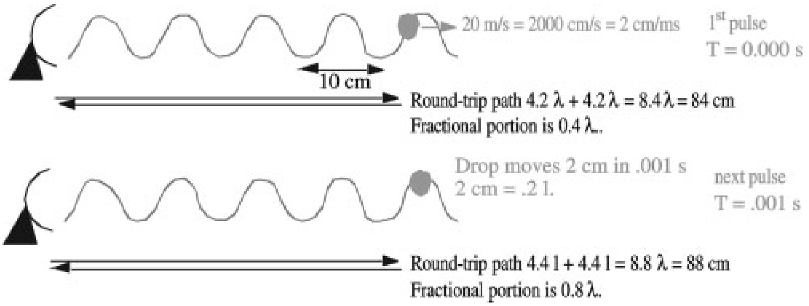

20 m/s = 2000 cm/s = 2 cm/ms
1st pulse
T = 0.000 s
10 cm
Round-trip path 4.2 λ + 4.2 λ = 8.4 λ = 84 cm
Fractional portion is 0.4 λ.
Drop moves 2 cm in .001 s
2 cm = .2 l.
next pulse
T = .001 s
Round-trip path 4.4 l + 4.4 l = 8.8 λ = 88 cm
Fractional portion is 0.8 λ.

Figure 7 Illustration of pathlength changes used to calculate toward/away component of velocity. Signal processing is able to measure the fractional portion of the pathlength change. In reality, this calculation is performed on the energy scattered by many raindrops. See ftp site for color image.

The calculation is conceptually simple when the energy from just one raindrop is considered. However, typical radar volumes contain many raindrops, each moving with the wind, but with some random component, each radiating an amount of energy proportional to D^6, interfering constructively and destructively. If a radar beam is pointed horizontally, it is usually assumed that the drops are moving with the wind, $V_d = V_a$. But, if the radar beam is inclined, the terminal velocity of the drop will enter into the measurement: $V_d = V_a + V_t \sin \theta$. Estimation of V_t is difficult, and must take into account that V_d is the D^6 weighted average. Typically, it is assumed that V_t is a function of Z and atmospheric density, and is about 8 m/s in heavy rain near sea level.

A critical limitation of single radar "wind" measurements is that they can only detect the wind component toward and away from the radar (the radial wind) of the three-dimensional wind field. Even very strong cross-beam wind components cannot be detected with a single normal weather radar (see multiple Doppler and bistatic sections below).

Spectral Width In addition to the radial wind, averaged over many pulses, many radars also calculate the spectral width, which is just the standard deviation of the individual pulse-to-pulse wind measurements or frequency domain calculations. The drops in a radar volume can exhibit different motions for several reasons. They may have different terminal velocities as just discussed. They may be embedded in sub-resolution-volume-scale turbulence. The resolution volume may span a large-scale meteorological feature like a front or mesocyclone, so different portions of the beam illuminate different portions of the phenomena containing different characteristic velocities. There is also always some measurement error.

Range Ambiguity Once a pulse is transmitted from a radar, it will continue indefinitely until it is totally consumed by reflection, absorption, or scattering. Elevated radar beams quickly pass above the troposphere into regions where there are few scatterers other than the moon and planets. But, beams that are oriented almost horizontally can remain in the troposphere for hundreds of kilometers. However, the useful range of a radar is frequently limited by what is called the ambiguous range. The ambiguous range is determined by the maximum range to which a pulse can travel and return before the next pulse is sent. If the pulse repetition time (PRT) is 1 ms, then this range is 150 km. It takes 1 ms for the pulse to travel to 150 km, scatter off raindrops at that range, and return to the radar. Energy emitted by raindrops beyond that range will reach the radar after the next pulse is sent. The radar has no simple way of knowing whether this energy originated from raindrops illuminated by the first pulse beyond 150 km or by the second pulse just a short distance from the radar (Fig. 8). Since the energy returning from both pulses is superimposed, the data is contaminated. The amount of energy that is returned by the raindrop at great range is reduced significantly due to distance ($P_r \sim 1/R^2$), but if the distant weather system is intense, and the nearby weather weak, the data from the nearby weather can be obscured.

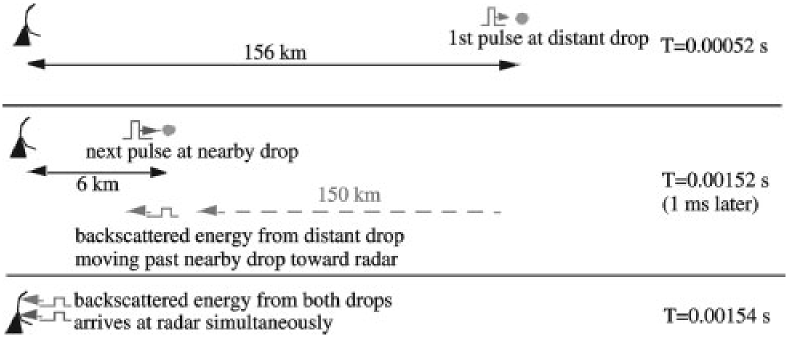

1st pulse at distant drop
156 km
T=0.00052 s
next pulse at nearby drop
6 km
150 km
T=0.00152 s
(1 ms later)
backscattered energy from distant drop
moving past nearby drop toward radar
backscattered energy from both drops
arrives at radar simultaneously
T=0.00154 s

Figure 8 Illustration of range ambiguity phenomena. Energy scattered back from raindrops at 156-km range arrives at the radar simultaneously with energy from the next transmitted pulse scattered back from raindrops at 6-km range. See ftp site for color image.

The ambiguous range can be increased by slowing the PRT. This is frequently done to measure storms at great range. A PRT of 2 ms increases the range to 300 km. But this may complicate velocity processing as discussed in the next section. Newer techniques include the addition of phase offsets to the transmitted pulses so that the true range to the scatterers can be retrieved.

So-called second-trip echoes can be detected by trained observers since they have elongated and unrealistic shapes. In the case of random phase magnetron radars, the Doppler velocities in the second-trip echoes will be incoherent, not smoothly varying as in correctly ranged echoes.

Velocity Ambiguity: The Nyquist Interval For a given transmitted wavelength and PRT, there is a maximum velocity that can be ambiguously measured. This is because Doppler radars do not actually measure the Doppler shift of returned radiation. Referring back to Figure 7, the fractional portion of the path length from radar to target back to the radar is measured. It is usually assumed that this path length changes by less than one full wavelength during the PRT. But, this may not be so. Consider a 10-cm radar with a PRT of 1 ms. If a raindrop is embedded in a strong wind, say 40 m/s away from the radar, then it will move 4 cm during the PRT. This means that the round-trip path length will increase by 8 cm, say from 10^9 to $10^9 + 0.8$ wavelengths. Ambiguity arises because it is difficult to distinguish between a 8 cm increase in path length and a 2-cm decrease since each will result in the same fractional portion difference, namely 0.8 wavelengths [i.e., $\text{frac}(10^9 + 0.8) = \text{frac}(10^9 - 0.2) = 0.8$]. The Nyquist interval is the range of velocities that will produce path length changes between $\pm\frac{1}{2}$ wavelength and can be calculated as Nyquist Interval $= \pm\lambda/(4\ \text{PRT})$. Thus, in the above case, the Nyquist interval would be $0.1\,\text{m}/(4 \times 0.001) = \pm 25\,\text{m/s}$.

The Nyquist interval can be increased by decreasing the PRT, essentially giving the raindrops less time to change the round-trip path length. Note, however, that this would increase data contamination from storms beyond the now shortened ambiguous range. There is a trade-off between maximizing Nyquist interval and maximizing ambiguous range. Some research radars use techniques called dual PRT or staggered PRT whereby alternating (or other patterns) of long and short PRTs are used to gain a least-common-multiple effect and much larger effective Nyquist intervals.

Data from regions exhibiting velocities greater in magnitude than the ambiguous velocity are considered to be folded or aliased and need to be corrected, or unfolded, or dealiased, to produce correct values (Fig. 9). This can be a difficult process and can be conducted either automatically or manually.

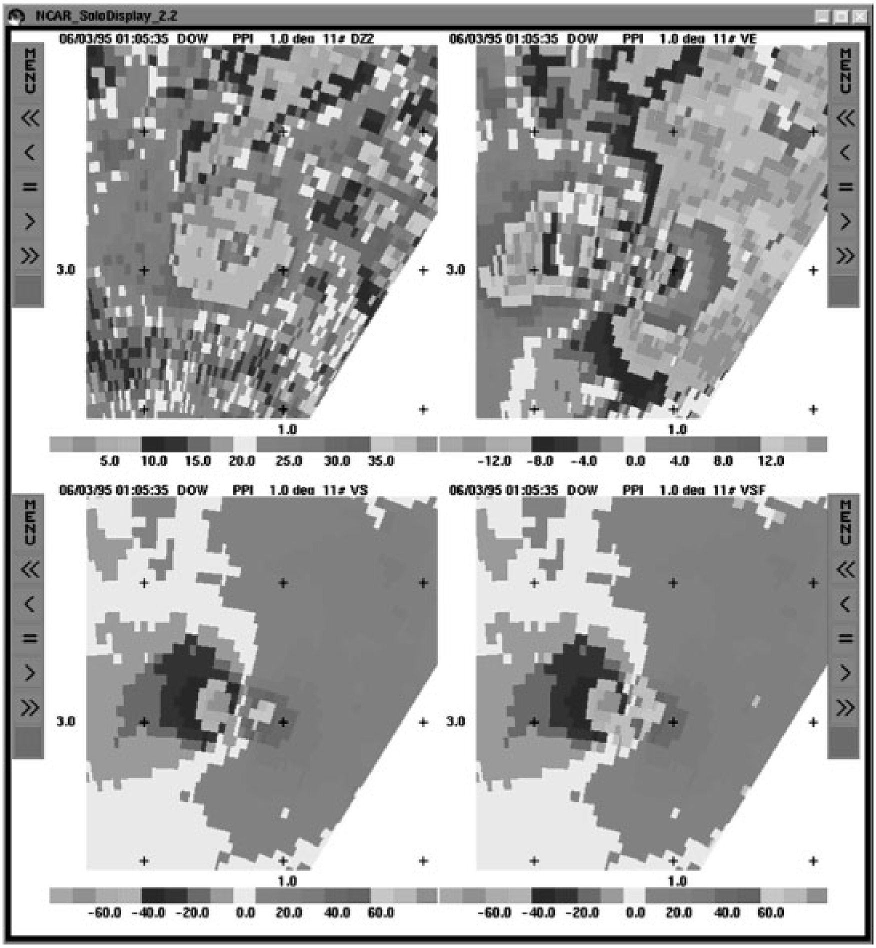


Figure 9 (see color insert) Illustration of dealiasing and cleaning of radar data. (*Top left*) Reflectivity in tornado showing ring debris. (*Top right*) Raw Doppler velocity with aliasing. (*Bottom left*) Velocity after dealiasing. Strong away and strong toward velocities adjacent to each other imply rotation, in this case over 70 m/s. (*Bottom right*) Velocity after values with high spectral width or contaminated by echoes from the ground (ground clutter) have been removed. Data is from DOW mobile radar in the Dimmitt, Texas, Tornado on June 2, 1995, from a range of 3 km. See ftp site for color image.

Scanning Techniques and Displays Most radars scan the sky in a very similar way most of the time. They point the antenna just above the horizon, say 0.5° in elevation, then scan horizontally (in azimuth), until 360° is covered. Then the radar is moved to a higher elevation, say 1.0° or 1.5°, and the process is repeated. This continues for several to many scans. Using this method, several coaxial cones of data are collected. These can be loaded (interpolated) onto three-dimensional grids or displayed as is to observe the low and mid/high levels of the atmosphere. There are infinite variations on this theme, including interleaving scans, fast and slow scans, repeated scans at different PRTs, scanning less than 360° sectors, etc. When a single scan is displayed, it is usually called a plan position indicator (PPI).

Sometimes, mostly during research applications, a radar will keep azimuthal angle constant and move in elevation, taking a vertical cross section through the atmosphere. These can be very useful when observing the vertical structure of thunderstorms, the melting layer, the boundary layer, and other phenomena. These types of scans are called range height indicators (RHI).

Since PPI scans have a polar, conical, geometry, lower near the radar and higher as the beams travel outward, it is often useful to use a computer to load the data from several scans into a Cartesian grid and display data from several different scans as they pass through a roughly constant altitude above Earth's surface, say 1 or 5 km above ground level (agl). These reconstructions are called constant altitude plan position indicators (CAPPIs). Since the data at a given altitude can originate from several scans, there can be ringlike interpolation artifacts in these displays.

Scanning strategy strongly influences the nature of the collected data. Operational radars typically scan fairly slowly through 360°, using many scans, requiring about 5 to 6 min for each rotation. This provides excellent overall coverage, but can miss the rapidly evolving weather such as tornadoes and microbursts. In specialized research applications, much more rapid scanning, through limited regions of the sky, is often employed.

New radar technology is being developed that may someday permit very rapid scanning of the entire sky as discussed below.

3 RADAR OBSERVATIONS OF SELECTED PHENOMENA

There are literally thousands of examples of weather phenomena observed by weather radars. Only a few will be illustrated here in Figure 10 to show some the range of phenomena that can be observed and the typical nature of the data. The interpretation of the data from weather radar is a complete study unto itself and could occupy far more space than is appropriate in this short overview.

4 GOAL: TRUE WIND VECTORS

Radial velocities provide much qualitative information about weather phenomena. However, the true wind field is a three-dimensional wind field comprised of three-

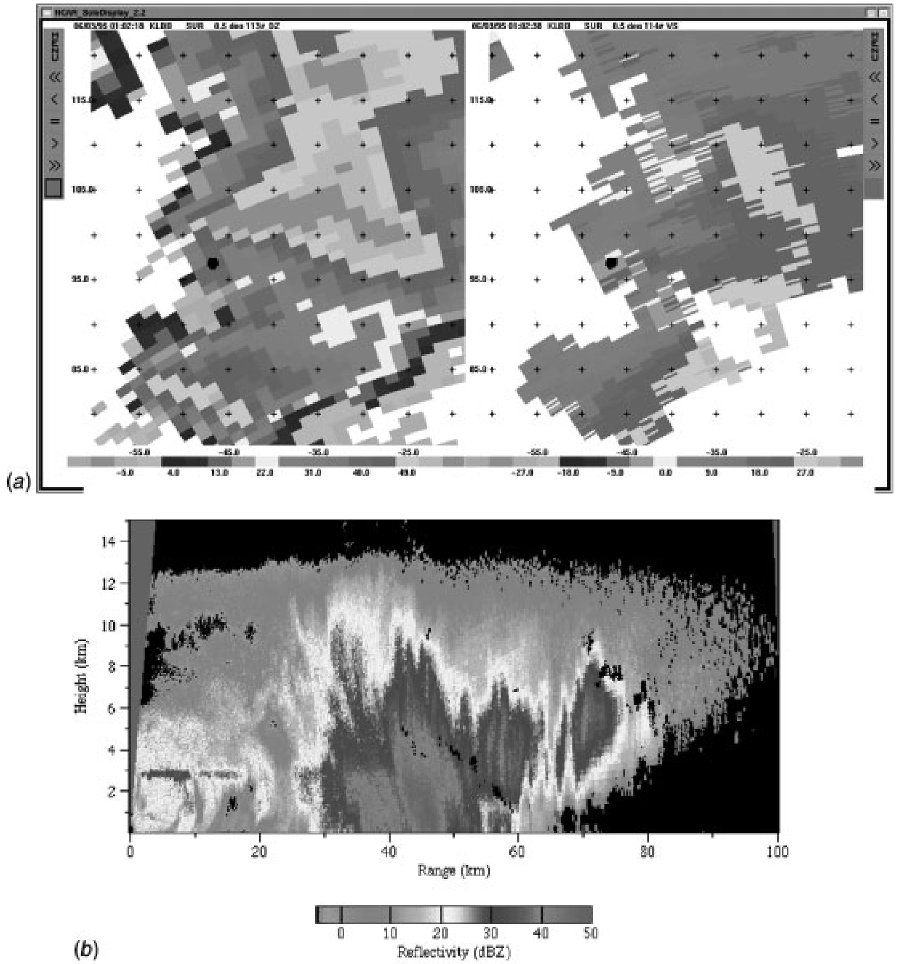


Figure 10 (see color insert) (*a*) A tornadic supercell thunderstorm observed by a WSR-88D operational radar. Reflectivity (*left*) and Doppler velocity (*right*) are shown. Classic hook echo extends from the western side of the supercell. An intense circulation, suggested by the strong away and toward velocities near the hook, is the mesocyclone associated with a tornado that was occurring. (*b, c*) Reflectivity and Doppler velocity in a vertical cross section (RHI) through a portion of a squall line. The high reflectivity core and lower reflectivity extending to 12 km are visible. Strong toward and away Doppler velocities are associated with the up and down drafts of the cell, as indicated by arrows. Data from the MIT radar in Albuqurque, New Mexico. (Courtesy of MIT Wea. Rad. Lab. D. Boccippio.) (*d*) Reflectivity (*lower*) and Doppler velocity (*upper*) in a winter storm. Reflectivity is somewhat amorphous but is enhanced in a ring corresponding to the melting layer. In the melting layer, large, wet slow-moving particles cause high reflectivity. The velocity pattern provides a vertical sounding of the atmosphere. Winds are from the NNW at low levels (near the radar), but from the southwest aloft (away from the radar as the beams diverge from Earth’s surface). Cold advection is implied. Data from the MIT radar in Cambridge, Massachusetts. (Courtesy of MIT Wea. Rad. Lab. D. Boccippio.) See ftp site for color image.

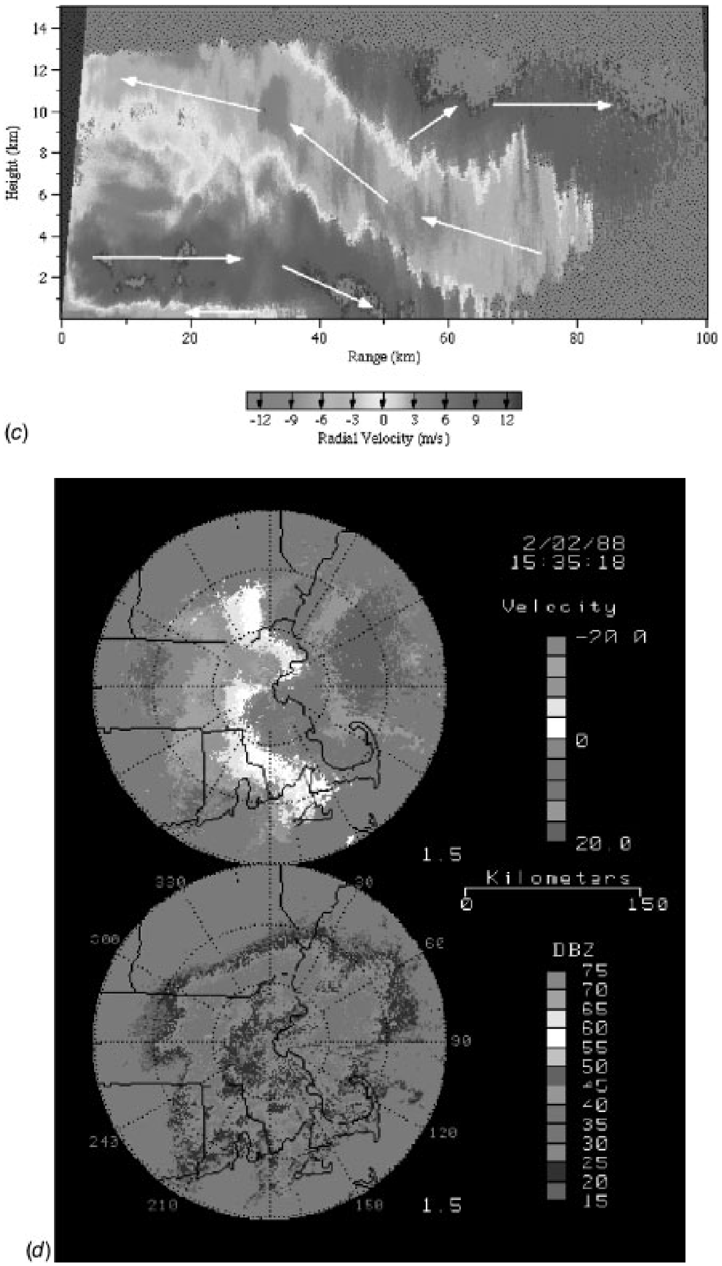


Figure 10 (*continued*)

dimensional vectors, evolving in time. Physical equations used in research and in forecasting models operate on these vector wind fields, which really contain the physics of the phenomena. So there is great value in estimating or measuring the full vector wind field and several techniques have been developed.

Single-Doppler Retrievals

One class of techniques for obtaining the vector wind field is called single-Doppler wind field retrievals. These techniques use various physical assumptions to convert data from a single radar into vector wind fields. One simple method assumes that the reflectivity field is a passive tracer that moves with the wind. In simple terms the Z field is examined at different times, and the wind field necessary to move the Z features from one place to another is calculated. Of course, evolving systems complicate these analyses. Another assumes that the wind field is composed of certain simple mathematical components. These are extensions of what radar meteorologists do visually when they look at the zero line of the Doppler velocity and assume that the wind is moving perpendicularly to it. Several sophisticated and combination methods are in development.

These techniques are very useful since they work with just one radar, but they are limited by the validity of the physical assumptions. Some also only work in cases of high reflectivity or velocity gradients, others only when precipitation covers much of the surveyed volume.

Currently, these techniques are being developed in the research environment, but it is hoped that they will be used to introduce wind fields into operational computer forecasting models in the future.

Dual and Multiple Doppler

In some research experiments, two or more radars can be deployed. Each radar can survey a target region from a different vantage point. A simple mathematical calculation can convert the two or more Doppler velocity measurements into a wind vector (Fig. 11). This technique is very powerful and has been a favorite of research meteorologists. It is not used frequently in operational forecasting because few permanent radars are close enough for dual-Doppler calculations to be useful. The large spacing of the U.S. WSR-88D network makes dual-Doppler calculations, while possible, not very useful due to the large beamwidths at the typical 200-km ranges to weather targets.

Typically, but not always, the horizontal components of the vector wind are calculated directly from the radar measurements. The vertical component is calculated by integrating the equation of mass conservation. Essentially, this says that if there is strong divergence in the boundary layer as in, say, a microburst, the air must have come from above, implying downward vertical motion (Fig. 12). Strong convergence near the ground implies an updraft. Similarly, strong divergence at storm top indicates an updraft from below, etc. Unfortunately, the vertical motions calculated in this manner result from the integration of quantities that are derivatives

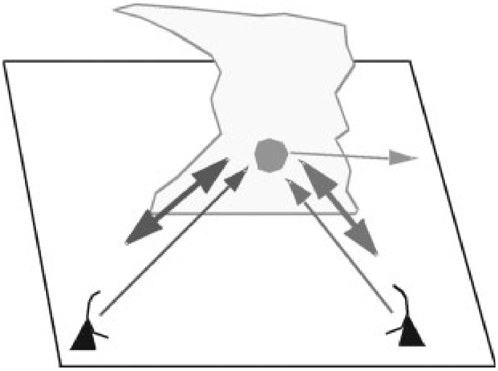


Figure 11 Dual-Doppler network with radars measuring motion of raindrops from different vantage points. The different Doppler radial winds (red and green) are combined mathematically to produce the horizontal projection of the true raindrop motion vector (blue). See ftp site for color image.

of actual measurements. Both the integration and differential processes are prone to errors. The resultant vertical wind estimates can be substantially incorrect. Accurate determination of vertical motions from radar data is probably the largest unsolved problem in radar meteorology today. An example of a dual-Doppler reconstruction of the vector wind field near and in a tornado is shown in Figure 13.

Rarely, three or more radars are in close enough proximity that the vertical component of the raindrop motion vector can be calculated directly. This method is called triple-Doppler and is exclusively a research technique.

There are two major limitations to multiple-Doppler techniques. The first is that radars are very expensive. Multiple-Doppler data is only affordable in a small frac-

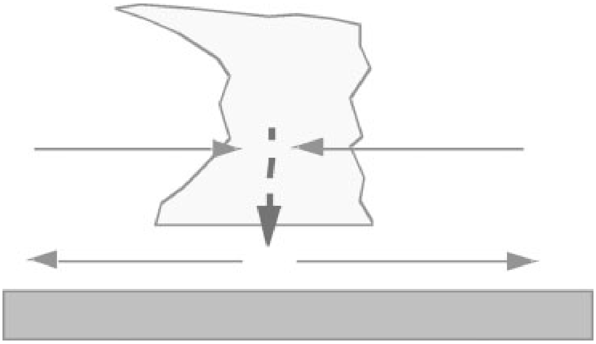


Figure 12 Since the vertical component of motion is rarely measured directly, the equation of mass continuity is usually used (sometimes in very complex formulations) to derive the vertical component of air motion. The physics of the method is very straightforward. If divergence is observed at low levels (*right*), then air must be coming from above to replace the departing air, implying a downdraft. See ftp site for color image.

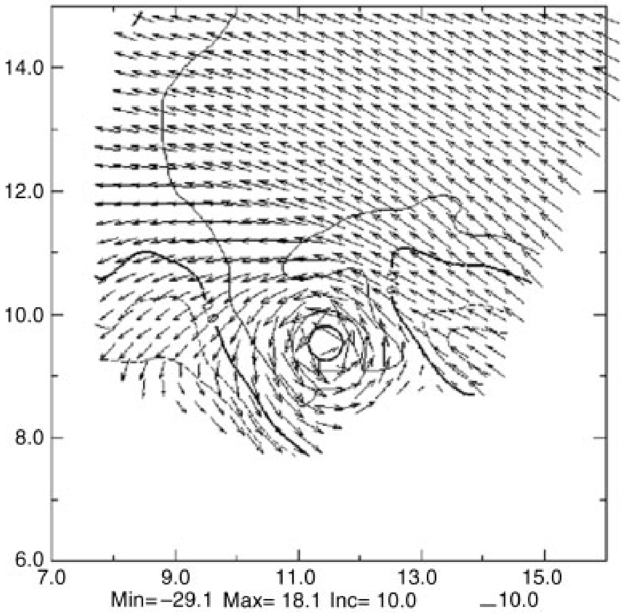


Figure 13 Dual-Doppler analysis showing horizontal component of the vector wind field in a tornado. Contours are *Z* with the small circle representing the low *Z* "eye" of the tornado. The axes are labeled in kilometers. Peak winds are about 60 m/s. (Courtesy Y. Richardson.)

tion of short-term research experiments and rarely in operational applications. The second is that the observations of weather targets by the different radars may occur at different times, sometimes a few minutes apart. Rapid evolution of some of the most interesting phenomena between these observations will contaminate the calculated vector wind fields.

A common expression among multiple-Doppler users is "you always get a vector," meaning that the technique always produces a result, but the result can be quite bogus. Multiple-Doppler and single-Doppler reconstructions should always be viewed with a sceptical eye.

Bistatic Radars

When the transmitted radar beam interacts with raindrops, only some of the reemitted energy travels back toward the transmitter. Most is scattered in other directions. The bistatic radar technique involves placing small passive radar receivers (Fig. 14) at various places to measure this stray radiation. The data from the bistatic receivers is combined with that from the transmitter using a variation on standard multiple-Doppler formulations.

The biggest advantage of the bistatic technique is the comparatively low cost of the bistatic receivers, less than 10% that of a WSR-88D. Several to many receivers

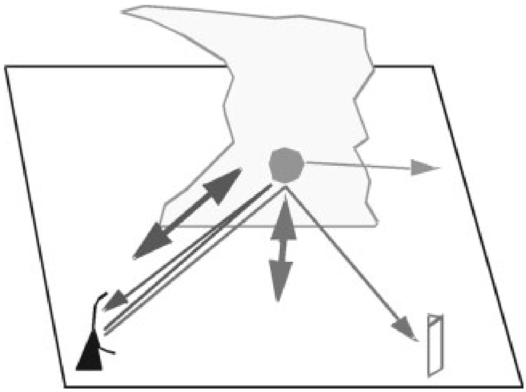


Figure 14 Bistatic dual-Doppler network with receive-only radar (red) measuring a different component (red arrow) of the raindrop motion vector (blue arrow) than the transmitter (green arrow). The components are combined mathematically in a fashion similar to that used in traditional dual-Doppler radar networks to calculate the horizontal projection of the raindrop motion vector (blue arrow). See ftp site for color image.

can result in more accurate data at a low cost. Another advantage is that the observations of individual weather targets are made simultaneously since there is only one source of radar illumination. Rapidly evolving weather can be well resolved. Thus two of the major limitations of traditional dual-Doppler networks are avoided.

Currently there are several research bistatic networks. Data from one are illustrated in Figure 15. It is anticipated that operational networks and operational computerized forecast models will use bistatic data in the future.

5 DATA ASSIMILATION INTO COMPUTER MODELS

Computerized weather forecasting models require accurate initializations to produce meaningful predictions. A major thrust of current modeling research involves how to best introduce radar data into these initializations, particularly into mesoscale simulations. Some models have successfully ingested both reflectivity and single-Doppler-retrieved wind fields, but none are yet used operationally.

6 NEW AND NONCONVENTIONAL TECHNOLOGIES

Dual and Multiple Polarization Radars

Most radars emit microwaves with just one polarization. This means that when they cause charge to move in raindrops as discussed above, the charge moves one direction then the opposite (say left, then right), but the charge distributions in other directions (say up and down) are largely unaffected. The intensity of microwaves

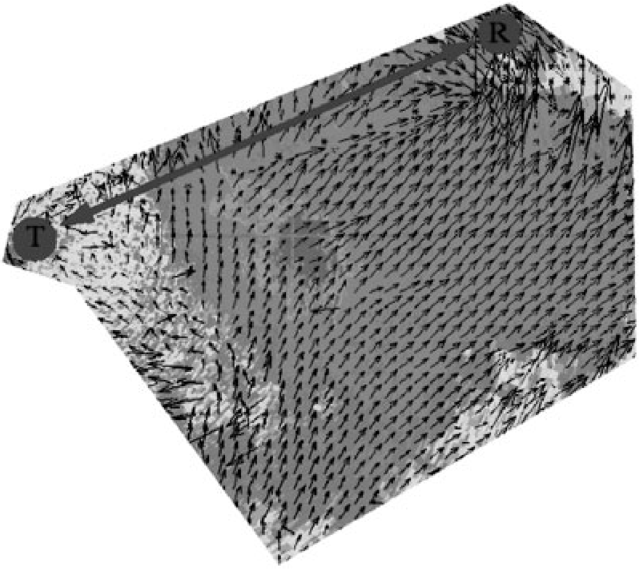


Figure 15 Bistatic dual-Doppler horizontal wind field. The transmitter (T) and passive, receive-only radar (R) are located 35 km distant. *Z* field is shaded. Data from NCAR CASES experiment in Kansas in 1997. See ftp site for color image.

emitted by a drop is proportional to D^6. But, large raindrops are hamburger bun shaped, ice particles have many shapes, and hail and insects can be very irregular. The intensity of the emitted microwaves is proportional to the D^6 in the direction of polarization of the radar.

It is possible to obtain information about the shape of the rain or ice particles by transmitting both horizontally and vertically polarized radiation. (There are other exotic techniques too, like using circularly polarized radiation, or radiation polarized at intermediate values between *H* and *V*.) If the beam hits a large hamburger-shaped drop, more horizontal energy will return than vertical energy (Fig. 16). The sixth power dependence means that small differences in D_h and D_v can cause large

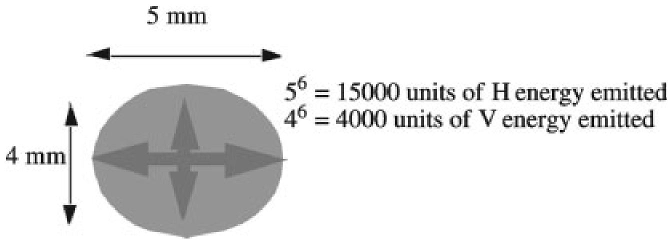

5 mm
5^6 = 15000 units of H energy emitted
4^6 = 4000 units of V energy emitted
4 mm

Figure 16 Oblate, hamburger-shaped raindrops scatter much more horizontally polarized energy than vertically polarized energy. See ftp site for color image.

differences in the emitted energy. This difference is called ZDR and can be as much as several decibels. Even though hailstones are often very irregular, they tumble randomly and the sum of the ZDR returns from thousands of hailstones is very close to zero (Fig. 17). Ice particles, however, can exhibit preferred orientations just like raindrops, and can produce ZDR. A good rule of thumb is that high Z with high ZDR = heavy rain while high Z with low ZDR = hail.

New techniques are in development to make use of other measurements possible with multiple polarization radars, such as the linear depolarization ratio (LDR), which measures how much radiation from the horizontal beam is reemitted from drops with vertical polarization, and specific differential phase, Φ_{dp}, which measures the differences in the phase of the horizontally and vertically polarized returned energy. Some of these hold promise to refine the identification of particle types (rain, hail, large hail, ice, etc.) and rain rate. Φ_{dp} holds particular promise since changes in Φ_{dp} are thought to be proportional to D^3, the volume of water in a resolution volume, and therefore more directly related to rain rate than Z. LDR is a very difficult measurement that requires a very precisely manufactured, therefore expensive, antenna.

Polarization radars are now primarily used in research but will probably be used operationally by forecasters in coming decades.

Mobile Radars

No matter which choices are made for radar wavelength and antenna size, the radar beam will always spread out with distance. Few radars have beams that are much narrower than 1°. This means that at 60-km range the beams are 1 km wide, and at 120 km they are 2 km wide, much too large to resolve small phenomena such as tornadoes, microbursts, etc. The best solution found to this problem so far has been

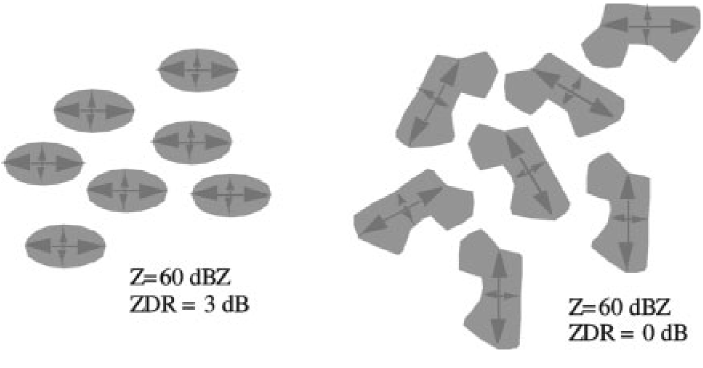


Figure 17 Large raindrops (*left*) are oriented similarly, so the ZDR effect from individual drops adds constructively to produce large ZDR values. Hailstones (*right*) tumble and are therefore oriented randomly so individual ZDR values tend to cancel out producing ZDR = 0. See ftp site for color image.

to put the radars on vehicles: ground based, in the air, and on the ocean. The vehicles are then deployed near the weather to be measured.

These systems tend to be very specialized and, with the exception of the hurricane hunter aircraft, are used mostly for research. Their mobile nature makes design and operation difficult.

Ground-Based Mobile Radars A leading example of the mobile radar concept is the Doppler On Wheels (DOW) research radars (Fig. 18). These radars have obtained unprecedented high-resolution three-dimensional data in tornadoes, hurricane boundary layers, etc., at scales as small as 3 to 60 m. Typically two or more DOWs are deployed in a mobile multiple-Doppler network to retrieve high-resolution vector wind fields. The DOWs use 3-cm radiation and 2.44-m antennas to produce 0.93° beam widths, which are comparable to the WSR-88Ds (but 3-cm radiation suffers more from attenuation in heavy precipitation). Fast scanning and very short pulses and gate lengths are combined for fine spatial and temporal scale observations. Other mobile radars using energy with wavelengths ranging from 3 mm to 10 cm have also been used by researchers to study similar phenomena.

None are currently in use for forecasting. But the idea of using DOW-type radars to augment the stationary radar network, part of a concept called adaptive observations, is being explored. In the future, forecasters who need information in specific areas, say a hurricane landfall, severe weather outbreak, flood, or the Olympics, may be able to request high-resolution tailored multiple-Doppler radar measurements.

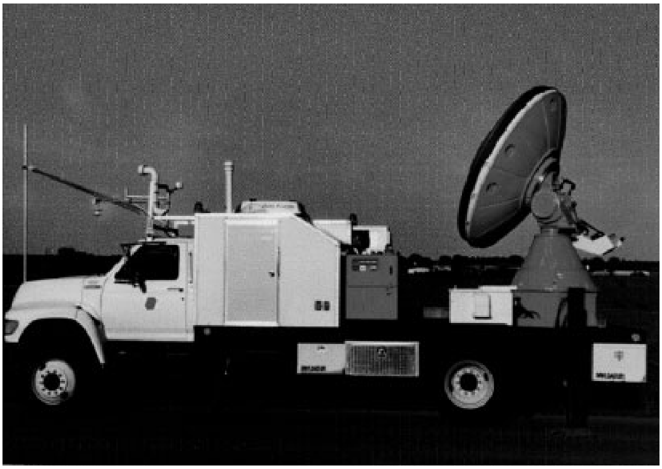


Figure 18 DOW2 mobile radar. The DOW2 uses a 2.44-m antenna transmitting 3-cm radiation with a 0.93° beam. It is used to intercept tornadoes, hurricanes, and is deployed in mountain valleys or wherever an easily movable system is needed. See ftp site for color image.

Airborne and Ship-Based Radars Since many areas are inaccessible to DOW-type trucks, particularly oceanic areas that cover 70% of Earth including hurricane spawning grounds, and because trucks are limited to highways speeds of 30 m/s, limiting deployment ranges, researchers and operational meteorologists use radars mounted on aircraft (Fig. 19) and ships to get closer to the weather of interest. These aircraft regularly fly into hurricanes before they make landfall to aid in predictions. They have been used to study meteorology as diverse as tornadoes in the Midwest and tropical climates in the western Pacific.

Bistatic Radar Networks

These collect stray radiation emitted in many directions by raindrops to measure vector wind fields and were discussed above.

Rapid Scan Radars

Military radars have long been able to move their beams electronically rather than having to mechanically move an antenna. Since the beam can be moved almost instantaneously, these hold the promise of extremely rapid scanning of weather. Instead of requiring 6 min to survey the entire sky, it might be sampled in only 10 to 30 s. A type of antenna called a phased array is used. Very few exist outside military applications, in part due to the high cost of construction. There are efforts being made to adapt this military technology to meteorological use.

A new type of rapid scan radar is under development and holds promise as a research and operational tool primarily due to its low cost, compared to phased array systems. These radars transmit multiple frequencies from an unusual antenna designed to split the various frequencies into simultaneous multiple beams (Fig. 20).

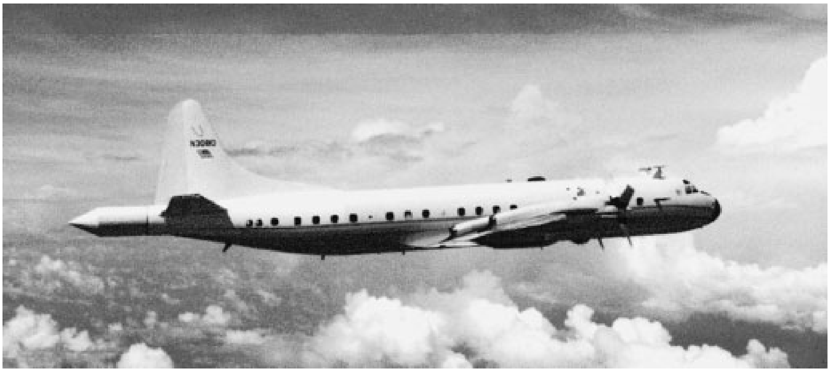


Figure 19 ELDORA airborne radar can be deployed quickly to almost any point in the world, even over oceans. It has a sophisticated radar in the protuberance extending beyond the normal tail. Similar systems are used operationally to intercept hurricanes. (Courtesy NCAR/ATD.) See ftp site for color image.

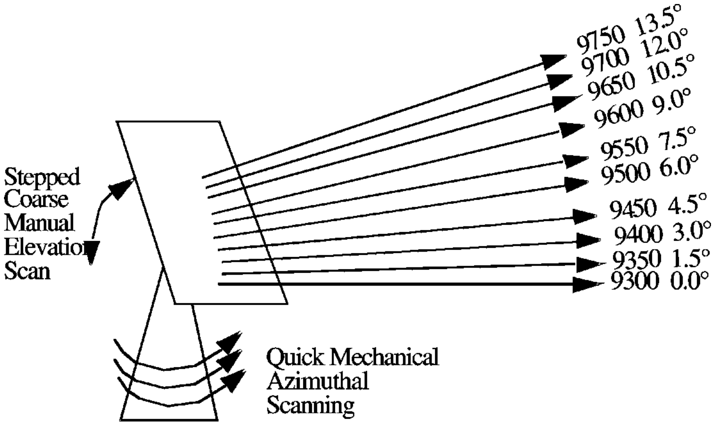


Figure 20 A flat panel, slotted waveguide array with 100 individual slotted waveguide antennas produces beams that emanate at different elevation angles depending on the frequency of the transmitted radiation. Therefore, nearly simultaneous transmissions at multiple frequencies, can be simultaneously received from several elevation angles at once. Ten or more elevation angles can be surveyed in ~10 s, providing truly rapid scanning with a nonphased array system.

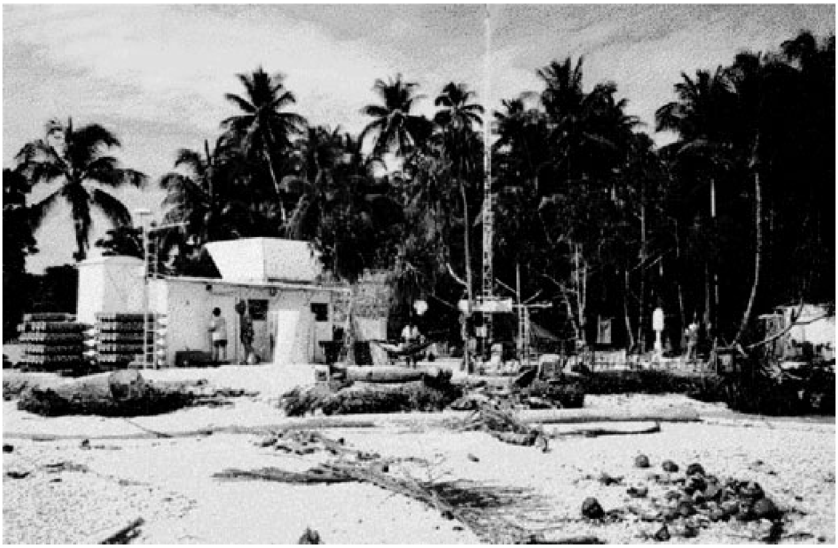


Figure 21 NCAR ISS in Kapingamarangi, Pohnpei State, Federated States of Micronesia. A wind profiler antenna is pointed vertically and shrouded by a clutter fence visible on top of the shelter. Aluminum cylinders shield a portion of a RASS radio acoustic sounding system. See ftp site for color image.

A single mechanical sweep of this antenna produces multiple beams, as many as 10 or more, permitting a typical volumetric scan to occur in just 2 sweeps, in as little as 10 s.

Wind Profilers

There is another class of radars called wind profilers that usually point vertically, do not scan, and sample vertical cross sections of wind and temperature. The principles of operation share some similarities to conventional weather radars, but there are substantial differences. They usually collect just one vertical sounding, averaged over about 30 min. Many can collect wind velocity data in clear air, in the absence of precipitation, through substantial depths of the lower troposphere. There is a network of profilers in the Midwest that augments the balloon sounding network by providing half hourly measurements at intermediate locations. Data from wind profilers is valuable for the forecasting of severe weather outbreaks in the Midwest. Deployable wind profilers are used by researchers to provide vertical soundings of wind between balloon launches (Fig. 21). An instrument called a RASS, or radio acoustic sounding system, uses an ingenious method whereby emitted sound waves disturb the atmosphere in a manner that can be measured by the profiler radar to measure the temperature of the atmosphere in the vertical column.

CHAPTER 49

BASIC RESEARCH FOR MILITARY APPLICATIONS

W. D. BACH, Jr.

> If you know the enemy and know yourself, your victory will never be endangered; if you know Heaven and know Earth, you may make your victory complete
>
> —Sun Tzu, The Art of War X, 31 circa 500 BC

The ancient Chinese general succinctly summarizes necessary ingredients for success in war. Understanding that "Heaven" represents the atmospheric environment, the need to know the weather is deeply rooted in military preparation and tactics. History is replete with examples of commanders using weather as an ally or suffering its bad effects. Even with today's modern technology, the dream of an all-weather military has not become a reality. Thus the military continues to seek ways to use the atmosphere as a "combat multiplier" in the order of battle.

1 MILITARY PERSPECTIVE

The military must understand the atmosphere in which it operates. That understanding will ultimately depend on scientific understanding of the atmospheric processes, an ability to use all of the information available, and an ability to synthesize and display that information in a quickly understandable fashion.

The military's needs for understanding the atmosphere's behavior has changed as a result of geopolitics and advancing technology. On the battlefield, future enemies are more apt to resort to chemical or biological attacks as a preferred method of mass destruction. The atmospheric boundary layer is the pathway of the attack. Increased

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6

reliance on “smart” weapons means that turbulent and turbid atmospheric effects will influence electromagnetic and acoustic signals propagating through the boundary layer. Intelligence preparation of the battlespace, a key to successful strategies, depends on full knowledge of the enemy, weather, and terrain. It requires an ability to estimate atmospheric details at specific locations and future times to maximize strategic advantages that weather presents a commander. It also avoids hazards and strategic disadvantages.

A convergent theme of basic research in the atmosphere is the interdependence of research progress in acoustic and electromagnetic propagation with the measurement and modeling of the dynamical boundary layer. The scattering of propagating energy occurs because of fine structure changes in the index of refraction in the turbulent atmosphere. The connection between the two—the refractive index structure function parameter, C_n^2 —is proportional to the dissipation rate of the turbulent kinetic energy.

Military operations are often in time (seconds to hours) and space (millimeters to tens of kilometers) scales that are not addressed by conventional weather forecasting techniques. Atmospheric information is required for meteorological data-denied areas. Furthermore the real conditions are inhomogeneous at various scales. Significant adverse effects may have short lifetimes at high resolution. Such breadths of scales make accurate quantification of important boundary layer processes very difficult. Carefully planned and executed experiments in the uncontrolled laboratory of the atmosphere are needed for almost every phase of the research. The combined range of atmospheric conditions and potential propagation types and frequencies that are of interest to the military is too broad to condense into a few nomograms. The propagation studies require intensive and appropriate meteorological data to understand the atmospheric effects. Understanding of heterogeneous atmospheric fields arising from inhomogeneous conditions requires measurements achievable only by remote sensors.

Chemical and biological agents constitute major threats to military field units as well as to civil populations. Electromagnetic propagation and scattering are the principal means of remote or in-place detection and identification of agents. Current models of atmospheric transport and diffusion of the agents over stable, neutral, and convective conditions are marginally useful and based on 40-year-old relationships. Significantly improved models of the transport and diffusion are needed to estimate concentrations and fluctuations at various time and space scales for simulated training, actual combat, and for environmental air quality.

2 RESEARCH ISSUES

Basic research for the military in the atmospheric sciences comes under two broad, interdependent scientific efforts:

Atmospheric wave propagation, which is concerned about the effects of the atmosphere, as a propagating medium, on the transmission of electromagnetic (EM) or acoustic energy from a source to a receptor, and

Atmospheric dynamics, which is concerned with quantifying the present and future state of the atmosphere.

Atmospheric Wave Propagation

Electromagnetic and acoustic propagation and scattering is a large subject that addresses problems in both characterizing the battlespace and in detecting targets for engagement and situational awareness. Although there has been much previous research in the general area of wave propagation and scattering, there remains a pressing need for research that specifically address issues related to propagation and scattering in realistic atmospheric environments.

Electro-optical and infrared propagation in the atmosphere is limited by turbulence and by molecular and aerosol extinction by both absorption and scattering. The relative importance of turbulence and extinction is dependent on the wavelength of the radiation and the atmospheric conditions. Modern, high-resolution imaging systems are often turbulence limited, rather than diffraction limited, impairing long-range, high-resolution target acquisition, recognition, and identification. Laser-based systems are limited in their ability to retain spatio-temporal coherence and effectively focus at tactical ranges. Like acoustic signals, electromagnetic signals are primarily useful for target detection or ranging and for remote sensing of the atmosphere itself. Although much research has been done in the area of electromagnetic extinction, and current models are generally good, applications of these models for remote-sensing purposes is an area of active research. A large body of research exists in the area of atmospheric turbulence effects for electromagnetic propagation, but with less conclusive results. Hence understanding of atmospheric turbulence for optical and infrared propagation is still an active area of research.

Understanding of the interaction between propagation and turbulence is rather straightforward in the theoretical sense. Practically, the time scales of the EM propagation interactions with the air are much shorter, $O(10^{-9}\,\mathrm{s})$ along a path, than the capability to locally sample the atmospheric turbulence or density fluctuations, $O(10^{-1}\,\mathrm{s})$ and to characterize the turbulence spectra, $O(10^{+3}\,\mathrm{s})$. Acoustic interactions also occur on short time scales. Second, the turbulent structure of the atmosphere between the source and detector is largely unknown, so the scale of the disturbance is undetermined. Furthermore, with models, grid volume samples are realized at $O(10^{0}\,\mathrm{s})$ in large eddy simulation models and at $O(10^{2}\,\mathrm{s})$ in fine mesoscale models. To make the connection between propagation effects, turbulence characteristics, and atmospheric models, new measurement techniques are needed. In careful field campaigns to relate propagation to atmospheric conditions, the biggest problem is independent ground truth.

Atmospheric Dynamics

The military emphasis on global mobility requires weather forecasting support at all scales from global to engagement scales. In recent years, the explosion in computer capacity enabled highly complex numerical weather prediction models on most of these scales. Model accuracy has improved even as complexities are added. Model results have become more accepted as a representation of the atmospheric environment. As computation power increases, finer and finer scales of motion are represented in smaller grid volumes. Lately, large eddy simulations represent dynamics at grid sizes of a few meters in domains of 5 km^2 by 1 km deep.

As the models go to finer scales, the variability imposed by large-scale synoptic and mesoscale influences the Atmospheric Boundary Layer (ABL) is modified by both the cyclical nature of solar radiation and metamorphosis due to the stochastic behavior of clouds and other natural processes as well as anthropogenic causes. The resulting (stable and unstable) boundary layers are sufficiently different that current models of one state do not adequately capture the essential physics of the other or the transition from one state to the other. Accurate predictions become more difficult, principally because the atmosphere is poorly represented. Data are lacking at the scales of the model resolutions. Parameterizations required to close the set of equations are inadequate. Observations are lacking to describe the four-dimensional fields of the forecast variables at the model resolution. The models are unlikely to improve until better theories of small-scale behavior are implemented and observations capable of testing the theories are available.

For the military, techniques to represent the inhomogeneous boundary layer in all environments—urban, forest, mountains, marine, desert—is absolutely essential for mission performance. In most cases, this must be done with limited meteorological data. Furthermore, the military is more frequently interested in the effects—visibility, trafficability (following rain), ceiling—than in the “weather” itself. Nevertheless, without high-quality, dependable models that represent the real atmosphere in time and space, the effects will not be representative.

Basic research attempts to improve modeling capability by increasing the knowledge base of the processes of the atmosphere. Until new measurement capabilities are developed and tested, our ability to characterize the turbulent environment—affecting propagation and dispersion of materials—will be severely limited.

3 PROTOTYPE MEASUREMENT SYSTEMS

Military basic research has participated in several new techniques to measure winds and turbulence effects in the atmospheric boundary layer. These techniques concentrate on sampling a volume of the atmosphere on the time scales of (at least) the significant energy containing eddies of the atmosphere.

The Turbulent Eddy Profiler, developed at the University of Massachusetts, Amherst, is a 90-element phased-array receiving antenna operating with a 915-MHz 25-kW transmitter. It is designed to measure clear air echoes from refractive index fluctuations. The received signal at each element is saved for postprocessing. These signals are combined to form 40 individual beams simultaneously pointing in different directions every 2 s. In each 30-m range gate of each beam, the intensity of the return, the radial wind speed is computed from the Doppler shift, and the spread of the Doppler spectrum is calculated. These data are displayed to show a four-dimensional evolution of refractive index structures in the boundary layer. Further processing gives estimates of the three components of the wind velocity. Combining the wind vectors with refractive index structures has shown significant horizontal convergence occurring with strong refractive index structures. At 500 m above the ground the volume represented is approximately a 30-m cube.

The Volume Imaging Lidar at the University of Wisconsin measures backscatter from atmospheric aerosols in three dimensions at 7.5-m range resolutions. The evolution of boundary layer structures can be seen in a variety of experiments. Capabilities to estimate horizontal winds at selected altitudes have been developed and demonstrated.

An eye-safe, scanning Doppler lidar operating at 2 μm has been jointly developed with NOAA/ERL/ETL to measure radial wind speeds at 30-m resolution. The error of the measurement is $< 0.2\,\mathrm{m/s}$. Various scanning approaches have shown several different evolutions of the morning transition and breakdown of a low-level jet.

Another lidar system has been developed at the University of Iowa to measure the horizontal wind in the boundary layer at about 5-m increments every minute. Teamed with existing FMCW radars, having comparable vertical resolution, should add to meteorological understanding of layers of high refractive index.

4 CURRENT MEASUREMENT CAPABILITIES

Reliable measures of atmospheric temperature and moisture at high resolution in space and time are still lacking. Some progress has been made but does not yet achieve the resolutions of the wind measurements.

Prototype and existing wind/wind field data do not yet measure the turbulence. Although the Doppler spectrum width is an indicator of the eddy dissipation rate, few researchers or equipment developers report the variable or its spatial variability. To date, none have done so operationally.

5 CONCLUSIONS

Some progress is being made by the military to develop necessary instrumentation to measured atmospheric fields at high time and spatial resolution. This is motivated by the need to improve the theory and subsequently the models of atmospheric motions at small scales to provide the armed services with reliable models on which to gain superiority over the adversary and successfully complete the mission.

CHAPTER 50

CHALLENGE OF SNOW MEASUREMENTS

NOLAN J. DOESKEN

1 INTRODUCTION—CHARACTERISTICS AND IMPORTANCE OF SNOW

Snow remains one of the truly incredible wonders of nature. Its delicate beauty, its pure whiteness, its endless variety and changeability delight children and adults. Its beauty is offset for some by the reality of the cold obstacle it presents to life's daily activities. The older we get, the greater an obstacle it becomes. How tiny and fragile crystals of ice totally transform a landscape even to the point of bringing temporary silence where the clamor of urban life usually prevails is indeed remarkable. The process of snow formation has gradually been explained by generations of ardent scientists (Bergeron, 1934; Nakaya, 1954; Mason, 1971; Hobbs, 1974; Takahashi et al., 1991). Still the reality of trillions of ice crystals forming and efficiently harvesting atmospheric water vapor and tiny cloud droplets as they fall through cloud layers on their path toward bringing moisture to Earth still seems miraculous to those that think and ponder.

The importance of snow in society today cannot be understated. While it is loved by some and hated by others, it greatly affects economic activities in the mid- and high-latitude nations of the world. Each year millions travel long distances to ski, snowboard, snow mobile, or participate in other winter recreation. Millions of others spend their hard-earned money escaping the cold and snow. Hundreds of millions of dollars are now spent annually in the United States alone clearing sidewalks, streets, highways, parking lots, and airport runways so that commerce and transportation are slowed as little as possible (Minsk, 1998). Despite these efforts, hundreds of lives are

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.

lost each winter to snow and ice-related accidents. Thousands more are injured in a variety of types of accidents. The economic cost of closed highways, blocked businesses, canceled flights, and lost time is enormous.

Snow is much more than an impediment to commerce. Just think of all the forts, snowballs, and snowmen made each year. Snow is also a structural material providing practical temporary shelter and protection from extreme cold. When smoothed and compacted, it can make an effective temporary aircraft runway in cold, remote areas. Left uncompacted, snow is an excellent insulation material protecting what lies below it from the extreme cold that may exist in the air immediately above.

The weight of snow is a necessary consideration in the design and construction of buildings. Almost every year some buildings are damaged or destroyed following extreme storms or periods of great or prolonged snow accumulation.

The high albedo (ability to reflect light) of fresh snow has profound impacts on the surface energy budget of the globe, which, in turn, dramatically affects climate both locally and over larger areas. Snow-covered areas are significantly colder than adjacent land areas under most weather conditions. Interest in global snow accumulation patterns has risen greatly during the past two decades due to its great significance in the global climate system (Bamzai and Shukla, 1999; Walsh et al., 1982) and the extent to which snow affects and is affected by large-scale global climate change.

In some mountainous areas, the largest threats posed by snow are avalanches. Subtle changes in crystal structure within the snowpack are always occurring. Certain weather patterns, temperature changes, and snowfall sequences lead to layering within the snowpack where some layers are “weak” and do not adhere well to adjacent layers. With the addition of new snow, these unstable snowpacks can be very prone to avalanches that claim dozens of lives in North America and Europe each year. The study of avalanches is a field of its own, involving hundreds of scientists around the world and requiring some unique measurements of internal snowpack characteristics.

Great improvements have been made during recent decades in weather prediction on the scale of a few hours to a few days. Improvements are also evident in long-range forecasts. One of the biggest challenges in weather prediction today remains the quantitative prediction of precipitation. Predictions of snowstorms and the location of the rain/snow/ice boundaries remain difficult even just hours in advance. Almost every year there are examples of large snowstorms that catch us by surprise, or forecasted storms that never materialize. If forecasts are to continue to improve, adequate observations must be taken on the scale needed to track and model precipitation processes.

Perhaps most important of all, snow is water. The hydrologic consequences of snow are so great that it is imperative to carefully track snow accumulation and its water content. Accumulating snowfall becomes frozen reservoirs that later release water into the soil, into aquifers, into streams and rivers, and into reservoirs. This water source is critical to water availability and hydroelectric power generation throughout the year, especially in mountainous regions and where snow contributes a significant fraction of the year’s precipitation. If snow melts quickly or if heavy rains fall on melting snow or downstream of melting snow, flooding also becomes a possible consequence.

2 MEASUREMENTS OF SNOW

Because of the importance of snow within our natural and socioeconomic environment, it is essential that we measure this remarkable substance and its salient features. We measure so that we can study, learn, explain, and teach others about snow—its properties and its impacts. We also measure so that we can describe and document what has occurred and note changes that occur over time. This allows us to compare, prepare, and predict so that our society can adapt as well as possible to the challenges and benefits derived from snow.

Much of what we know about snow, its spatial distributions and its contribution to the hydrologic cycle, we have learned from very simple observations taken at a large number of locations for a long period of time. Here is a list of common measurements. Details about instrumentation and methodologies are provided later in this chapter.

- *Precipitation Amount* This refers to the water content of snowfall plus any other liquid, freezing or frozen precipitation falling during the same period such as rain, freezing rain, or ice pellets (sleet). Measurements of precipitation amount are traditionally taken with a recording or nonrecording precipitation gage.
- *Snowfall* The accumulation of new snowfall or other forms of frozen precipitation that have fallen and accumulated in the past day or other specified time period. (Glaze from rain that freezes on contact is not included in this category.) The measurement of snowfall has traditionally been done manually by trained observers using a simple measurement stick and their own good judgment.
- *Snow Depth* This is simply the total depth of snow and ice, including both freshly fallen and older layers. For shallow snows, a ruler or longer measurement stick is all the equipment needed. For deeper snows, fixed snow stakes or special calibrated probing bars are used. Electronic methods for measuring snow depth have been developed in recent years and are used at some sites.
- *Snow Water Equivalent* This term, commonly abbreviated SWE, is the total water content expressed as an equivalent depth of the existing snowpack at the date and time of observation. Since snow does not accumulate or melt uniformly, the SWE is often the average of a set of representative measurements taken in the vicinity of the point of interest. The measurement of SWE is often taken by weighing a full core sample (snow surface down to ground surface) or averaging the weights from several samples. Other methods will also be described.
- *Density* The mass per unit volume is an important variable for describing the nature and potential impacts associated with snow (Judson and Doesken,

2000). A common but often incorrect assumption used in the United States is that 10 in. of new snow has a liquid water content of 1 in., which is a density of 0.10. This could also be expressed as a percentage—10%. Under ideal conditions, the density can be obtained by dividing the measured precipitation amount (gage catch) for the time period of interest, by the measured depth of new snowfall for that same period. However, a direct measurement of water content per carefully determined volume is often more accurate since gage measurements of snowfall often do not catch all the precipitation that actually falls, especially under windy conditions.

- *Precipitation Type and Intensity* For purposes of weather forecasting and verification as well as airport operations and other aspects of transportation, continuous monitoring of the type of precipitation (rain, freezing rain, ice pellets, snow pellets, snow, hail, etc.), its intensity (light, moderate, or heavy), rate of accumulation, and how much the horizontal visibility is restricted are very important. For many years, a simple definition of snowfall intensity has been used in the United States at airport weather stations based on the degree to which the snowfall reduces visibility (U.S. Dept. of Commerce, 1996). For example, unless the horizontal visibility was restricted to less than $\frac{3}{4}$ mile, the snowfall intensity could only be reported as "light." Precipitation type, intensity, and visibilities were all determined manually until the mid-1990s. Electronic sensors have been introduced at many airport weather stations in recent years.
- *Snowcover Extent* This is an assessment of how much of a specified land area is covered by sufficient snow to whiten the surface at any specified time. Until the use of aircraft and satellites devoted to observing snow cover, this was accomplished simply by mapping individual weather station snow depth observations and approximating the location of the edge and area of snow-covered regions.

Other types of snow measurements are taken for basic research, for special applications such as water quality assessments and avalanche prediction, and for military applications in cold climates.

- Albedo
- Crystal types and evolution
- Insulation
- Acidity (snow and the first flushes of snowmelt have been found to be among the most acidic forms of precipitation in some areas)
- Electrical conductivity
- Trafficablity/compactability
- Layer structure and stability
- Forest canopy snow accumulation and sublimation

By no means is this an exhaustive set of measurements. An excellent source of additional information about snow properties and measurements is the *Handbook of Snow* (Gray and Male, 1981).

3 PROPERTIES OF SNOW THAT MAKE BASIC MEASUREMENTS DIFFICULT

All environmental measurements have difficulties and limitations. For snow, its dynamic changing features provide challenges for observations. Snow melts, sublimates, settles, and drifts. Its crystal structure changes from storm to storm and from time to time within a storm. Once on the ground, the crystals change again in the presence of surrounding crystals, temperature gradients, and vapor density gradients. Snow is not deposited uniformly on the ground, and it melts even more unevenly depending on factors such as shading, slope, aspect, wind exposure, vegetation height, color, and amount. Traditional precipitation gages that work fine for measuring rain are often grossly inadequate for capturing and measuring the water content of snow since the feather-light crystals are easily deflected around the precipitation collector even by light to moderate winds so some of the snow does not fall into the gage. Furthermore, snow may cling to the side of collectors, effectively changing the collection diameter of the gage. Additionally, the compressibility of snow makes it difficult to gather the appropriate core samples. When all is said and done, measuring snow is easy. Measuring it accurately and consistently is the problem.

The following example taken from *Weatherwise* magazine (Doesken, 2000) demonstrates the challenge of measuring snow.

> Snowyville, USA
>
> For an example of some of the difficulties that snow observers face, let's imagine we are in Snowyville, USA, on a cold February morning. At 7:00 a.m., it begins to snow. For 12 hours it snows hard and steadily, until it ends abruptly at 7:00 p.m
>
> Five volunteer weather observers all live in lovely Snowyville and all receive the same amount of snow. Each observer has an excellent location for measuring snow in wind-protected locations, and all take measurements using identical rulers and snowboards properly placed on the surface of the snow. [Details on how and where to set up an accurate snow measurement station will be discussed later in the chapter.]
>
> The first observer is a retired engineer and is home all day watching and enjoying the snow. He diligently goes outside every hour on the hour throughout the storm, measures the depth of the new snow on his snowboard, and then clears it in preparation for his next measurement. Each hour there is exactly one inch of new snow on the board. When the storm ends, the observer adds up the snowfall amounts for each hour and comes up with a total of 12 inches of new snow. He writes it down and calls the NWS with his report.
>
> The second observer is also very diligent, but had to go to work in an office downtown. She knows it's snowing hard, so she takes a few minutes after lunch and comes home at 1:00 p.m. to take a measurement. There are five inches of fresh snow on the snowboard when the observer takes her measurement. She clears the surface of the snowboard and places it back on the top of the new snow. She then goes back to her office. When the snow ends at 7:00 p.m., she goes back outside and measures the new snow on the snowboard. There is another five inches. She clears her snowboard and goes inside. She adds up the two measurements, and calls the NWS with her report of ten inches of new snow.

The third observer is a school teacher. She keeps an eye on the snow all day and knows it's snowing hard. She hopes that school will be canceled, but in little Snowyville, people are used to heavy snow. Not even the after-school events are canceled. Between teaching and after-school parent-teacher conferences, she doesn't get home until 7:00 p.m., just as the snow ends. The observer goes straight to her snowboard and measures 8.4 inches. She promptly calls the NWS with her daily report.

The fourth observer normally works in a distant town but instead decides to take the day off due to the heavy snow. He looks out his window nearly constantly, thrilled by the millions of snowflakes. Every hour or so, he journeys out to his snowboard and measures the depth. He just lets it accumulate, though, and does not brush it away. The snow is one-inch deep by 8:00 a.m., five inches deep by 1:00 p.m., and reaches a maximum depth of 9.3 inches around sunset. When the snow ends at 7:00 p.m., it has settled back to 8.4 inches. The observer gets on the phone and calls in his daily snowfall total of 9.3 inches to the NWS.

Finally, the official cooperative observer for the town leaves that morning for a meeting in the nearest town. When he returns that evening, he goes straight to bed. The next morning, precisely at his scheduled reporting time of 7:00 a.m., he goes out to his snowboard, inserts the ruler, and measures 7.2 inches of snow. He clears the board and goes inside. Not realizing how much snow had settled overnight, he records this on his observation form and calls in a 7.2-inch snowfall total to the NWS.

So how much snow actually fell in Snowyville? 12 inches? 7.2 inches? Or something in between? Each observer received the same amount and measured very carefully. Yet their reports were quite different. This is due to the fact that snow melts, settles, and compacts after it falls to the ground. Climatologists and many data users would say that the 9.3-inch report was the correct value. Weather forecasters might argue that the 12-inch report was accurate, but if measuring every hour is good, who is to say that measuring every 30 minutes or even every 15 minutes isn't even better. Then, the sum might have been 14 or 15 inches. Meanwhile, engineers and water resource officials may not care at all about these debates, as long as the observers each accurately reported the same water content of the snow.

This one example shows just how complicated measuring snow really is. And although it does not include the problems of drifting and melting snow, it does demonstrate how important it is to follow similar procedures if measurements are meant to be compared from one location to another.

4 PROCEDURES FOR MEASURING SNOWFALL, SNOW DEPTH, AND WATER CONTENT

As with all other measurements of our environment, the key to success is finding and preserving a consistent and representative location for measurement and maintaining strict standards for instrumentation and observing practices. For comparing data from many locations, consistent procedures and representative measurement locations are essential (Doesken and Judson, 1997).

Precipitation

The measurement of precipitation amount is arguably the most basic and most useful. The most common device for measuring precipitation is a straight-sided cylinder of a sufficient diameter and depth to effectively catch rain, snow, and other forms of precipitation. The National Weather Service (NWS) standard precipitation gage has a diameter of 8 in. and is approximately 2 ft tall. In addition to the outer cylinder, it comes with a funnel, inner measuring tube, and calibrated measuring stick as shown in Figure 1. The area of the opening to the funnel and the outer cylinder (overflow can) is precisely 10 times greater than the area of the top of the inner measuring tube. This allows greater precision in the measurement of precipitation. In the United States, observers measure precipitation to the nearest 0.01 in. while much of the rest of the world measures to the nearest millimeter. For the measurement of rain, the inner cylinder and funnel are installed in and on top of

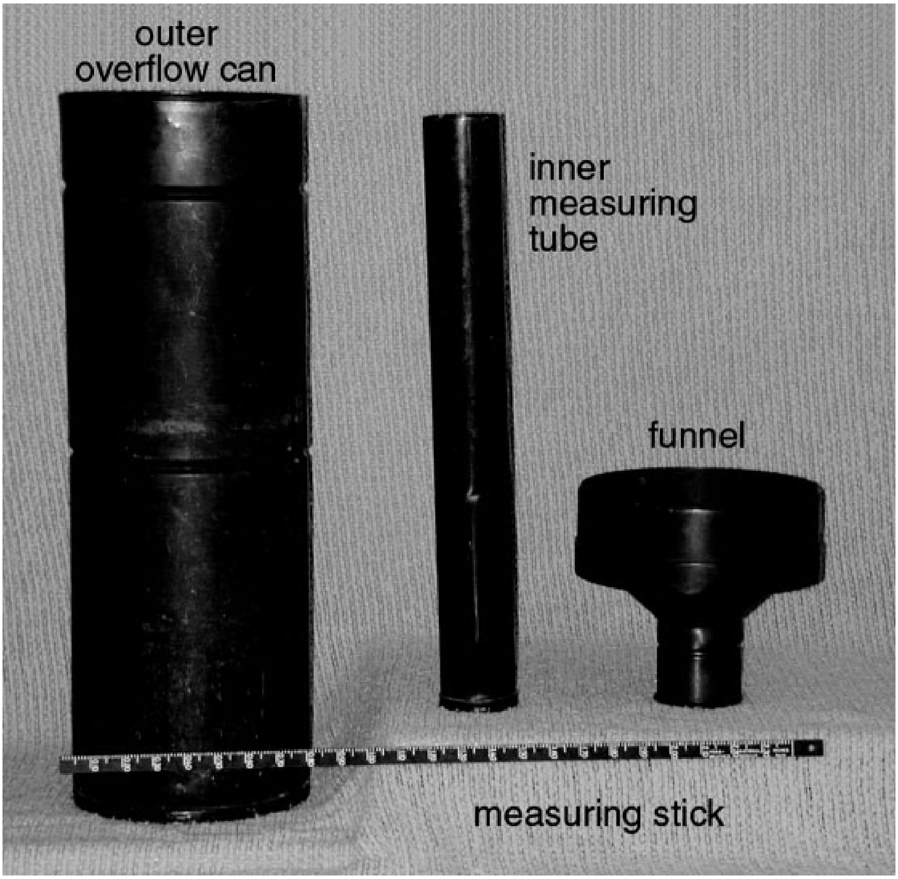


Figure 1 Standard precipitation gage consisting of funnel, inner measuring tube, outer overflow can, and calibrated measuring stick (photo by Clara Chaffin).

the large "overflow can," respectively. For capturing snow, the funnel and inner tube are removed.

Most manual weather stations read and record precipitation daily at a specified time of observation. A few manual stations may measure more frequently. When measuring liquid precipitation, the observer simply inserts the calibrated measurement stick straight down into the inner measurement tube and reads the depth of accumulated water. The observer then records that amount, empties the gage, and is ready for the next observation. Manual rain gages are very accurate for the measurement of liquid precipitation under most conditions. However, when winds are very strong during a rain event, the gage will not catch all of what falls from the clouds since wind movement over the top of the gage will deflect a portion of the raindrops.

For measuring the water content of snowfall only the large outer cylinder (overflow can) is left outdoors. Following a snow event, the standard observing procedure is to bring the gage inside at the scheduled time of observation. The snow sample must first be melted. This is accomplished either by setting the gage in a container of warm water until the snow and ice in the gage are melted or by adding a known amount of warm water directly to the contents of the gage to hasten its melt. Observers then pour the contents of the gage through the funnel into the inner cylinder for measurement, being careful to subtract any volume of water that was added to hasten the melt. In very snowy locations, some observers may be equipped with special calibrated scales so that the gage and its contents can be weighed to determine precipitation. This simplifies the observation considerably, especially for locations where warm water is not readily available.

A variety of other precipitation gages are also used. Weighing-type recording rain gages have been used for many years by the NWS for documenting the timing of precipitation (see, e.g., Fig. 2). For winter operation, an antifreeze solution is required. An oil film on the surface of the fluid reservoir is also recommended to suppress evaporation losses. The use of oil and antifreeze could be environmental hazards so care must be taken in the selection and use of these materials.

Storage precipitation gages, large gages with the capacity to hold several feet of snow water content, have been used for measuring total accumulated precipitation at remote locations. These gages also require oil and antifreeze. The volume of additives must be accurately measured since their density differs from water. Tipping bucket precipitation gages, a favorite gage because of its low cost, relative simplicity, and ease of use for automated applications, are not very effective for measuring the precipitation from snow (McKee et al., 1994). Heat must be applied to the surface of the funnel of these gages in order to melt the snow. Since most snow falls at rates of only a few hundredths of an inch per hour (1 or 2 mm per hour or less), even small amounts of added heat can lead to the sublimation or evaporation of much of the moisture before it reaches the tipping buckets. Furthermore, the addition of heat can create small convective updrafts above the surface of the gage, further reducing gage catch.

The National Weather Service continues to search for a satisfactory and affordable all-weather precipitation gage. As simple as it may seem, the measurement of precipitation amounts has yet to be perfected.

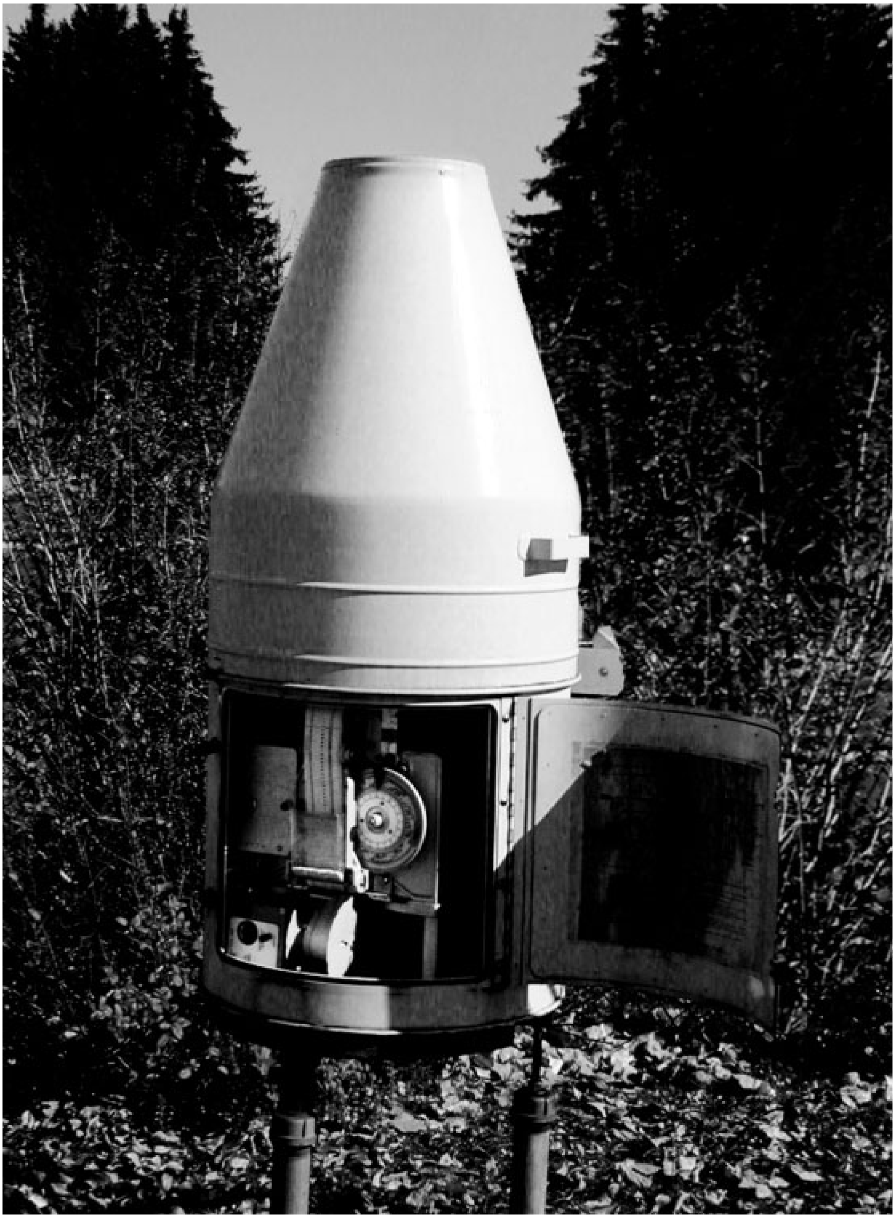


Figure 2 National Weather Service Fischer–Porter recording rain gage. Close to 3000 of these gages are in use in the United States for recording hourly and 15-min precipitation totals in increments of 0.10 in. (photo by Clara Chaffin).

Even if a perfect gage were available, there is still a major problem limiting the accuracy and introducing uncertainty into gage measurements of the water content of snow. *The problem is wind.* Gages protruding into the air present an effective obstacle to the wind resulting in a deflection. Lightweight snow crystals are easily deflected. The result is gage undercatch of precipitation. The degree of undercatch is a complex function of wind speed, snow crystal type, gage shape, and exposure (see Fig. 3). In the measurement of rain, gage undercatch is not significant until winds are very strong. However for snow, even a 5 mph breeze can have a very large effect on gage undercatch.

One approach to improving gage catch efficiency is the installation of a wind shield surrounding the gage to reduce the effects of wind-caused undercatches (see Fig. 4). The most common shield in use in the United States is called an Alter shield, named in honor of its designer. While the Alter shield clearly improves gage catch efficiency, it by no means solves the problem. The Nipher shield is a favorite in Canada, and the results above show why. Unfortunately, this shield does not perform well in heavy, wet snows common in many regions of the world and does not adapt to other types and sizes of precipitation gages, thus making it impractical for use with most precipitation gages in use in the United States. Currently, only a fraction of the U.S. precipitation gages are equipped with wind shields.

The World Meteorological Organization (WMO) has been diligently investigating the challenge of measuring solid precipitation. An extensive international study was completed during the 1990s thoroughly investigated the performance characteristics of a variety of gages and wind shields used in snowy regions of the world (Goodison and Metcalf, 1992). Many consider the Double Fence Intercomparison Reference (DFIR) to produce the most representative gage catch under a full range and snow-

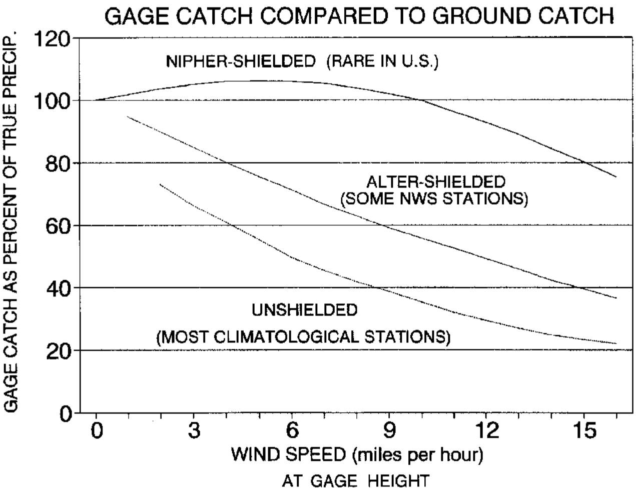


Figure 3 Gage catch efficiency in percent as a function of wind speed at the top of the gage for unshielded, Alter-shielded and Nipher-shielded gages (after Goodison, 1978).

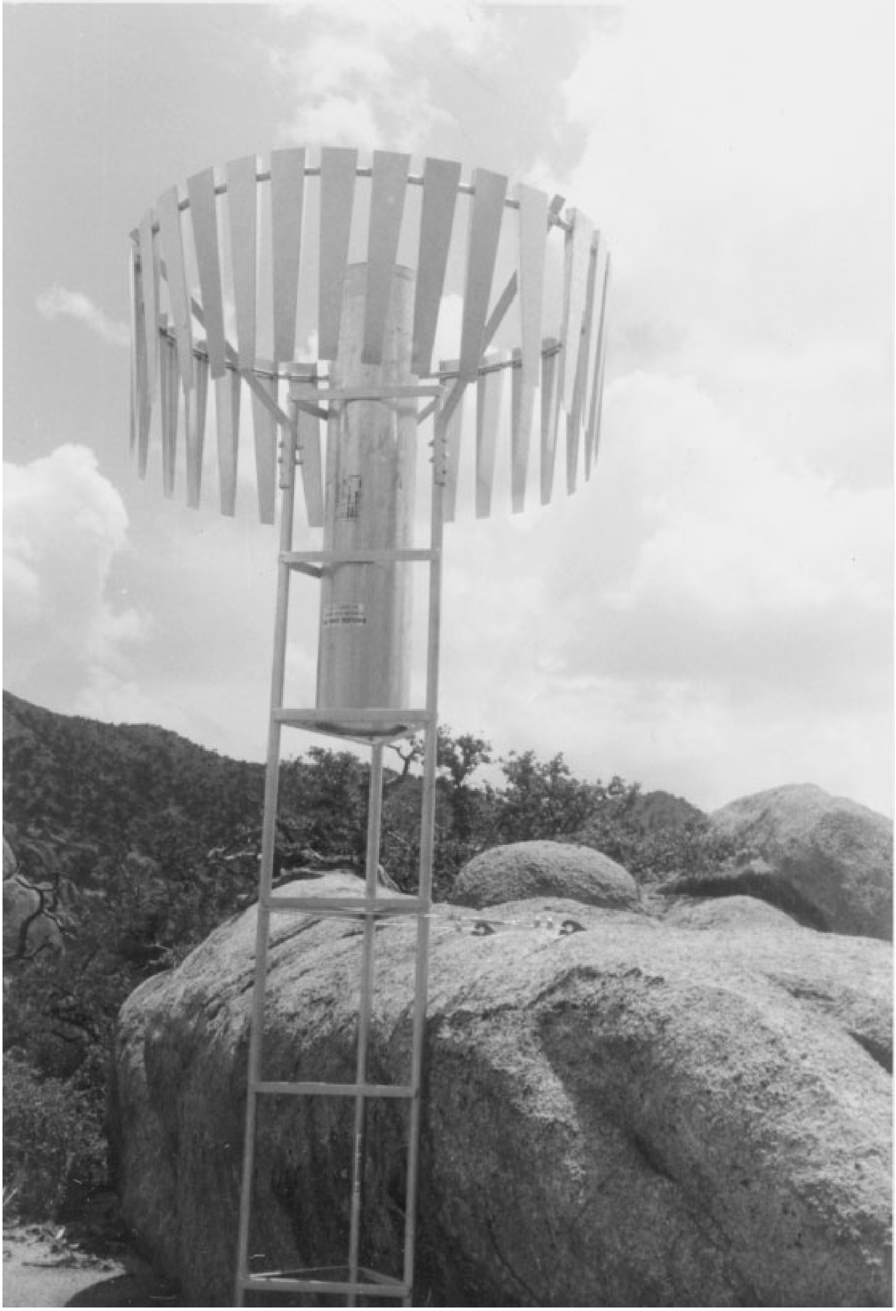


Figure 4 Alter wind shield on a tower-mounted precipitation gage (from Doesken and Judson, 1997).

fall and wind conditions (Yang et al., 1993). Unfortunately, the size (12-m diameter) alone makes this wind shield too bulky and expensive for most common weather stations. No simple solution exists, and few countries show a willingness to change their long-standing practices. But there is no excuse to ignore the problem of gage undercatch since it is now well documented.

A practical approach to the gage catch problem is to be as wise as possible when first deploying weather stations in order to achieve the best exposure for gages. The ideal exposure for a precipitation gage is a delicate compromise between an open and unobstructed location and a protected site where the winds in the vicinity of the gage are as low as possible during precipitation events. The center of a small clearing in a forest or an open backyard in a suburban neighborhood are examples of good sites. The closer to the ground the gage is, the lower the winds will be—a plus for improving gage catch. But the gage must also be high enough to always be above the deepest snow. Rooftop exposures are not recommended because of the enhanced wind problems and the potential for building-induced updrafts that will further reduce gage catch.

When gage undercatch is apparent, observers are encouraged to take a secondary observation by finding a location where the accumulation of fresh snow approximates the area average. A core sample is then taken, and this core is melted or weighed in order to determine its water content. Unless an unrepresentative sample is taken or melting has reduced the water in the remaining layer of fresh snow, the water content of a core often exceeds that of the gage. In practice taking supplemental core samples is seldom done to check for gage undercatch, but it could improve data quality considerably.

Snowfall

The traditional measurement of snowfall requires only a measurement stick (ruler) and a bit of experience. While this may be the simplest meteorological measurement, in practice, it may also be the most inconsistent. The public interest in snowfall is always great, but the measurements are often more qualitative than people realize. The inconsistency is a direct result of the dynamic nature of fresh snow. It lands unevenly, melts unevenly, moves with the wind, and settles with time. The location where measurements are taken, the time of day, the time interval between measurements, the length of time since the snowfall ended, the temperature, the cloudiness, and even the humidity affect snow measurements. The imaginary but realistic example from Snowyville dramatized these points.

For climatological and business applications, for spatial mapping, and for station-to-station comparisons, the best functional definition of snowfall is “the greatest observed accumulation of fresh snow since the previous day prior to melting, settling, sublimation, or redistribution.” The typical observer may only go out to measure snowfall once per day at a designated time. Depending on weather conditions and timing, that may or may not coincide with the time of greatest accumulation. If there had clearly been 4 in. of fresh snow on the ground early in the day but only 1 in. still remains at the scheduled observation, did it snow 4 in. or 1 in.? Four inches is obviously the better answer, but if the observer were not there to see it, he or she wouldn't know for sure. Ideally, the observer is available to continuously watch the snow accumulate, note the greatest accumulation, and then note the settling and melting that occurs later. But, in reality, this may not be the case, since much of the historic snowfall data in the United States has come from the National Weather Service Cooperative Observer Program, many of whom are volunteers who can take only one observation each day (National Research Council, 1998).

Because of the nature of fresh snow, perfectly consistent measurements may not be possible. However, there are several simple steps that lead to measurements that achieve a high degree of consistency.

- The location for taking measurements is critical. An unobstructed yet relatively protected location (such as a forest clearing or open backyard away from buildings, trees, and fences) is best since the goal is to measure where snow accumulation is as uniform and undrifted as possible and represents the average snow accumulation of the area.
- The use of a snowboard (square or rectangular flat, white surface positioned on the ground and repositioned daily on the top of the existing snow surface) for measuring the accumulation of new snowfall greatly helps to achieve consistency by providing a smooth, solid surface on which to measure and from which core samples can be taken (Doesken and Judson, 1997).
- When snow is falling, observers should periodically check the accumulation of new snow on the snowboard and note the greatest amount before settling or melting begin to reduce the depth of fresh snow. The greatest accumulation will typically occur just before the snowfall diminishes or changes to rain. Observers should only clear the new snow at the scheduled observation time and then reposition the snowboard on the top of the new snow surface.
- To account for blowing and drifting snow and the resulting uneven accumulation patterns, observers should assess the representativeness of the snowboard measurement by taking an average of several measurements in the surrounding environment, making sure to include only that snow that has fallen since the previous observation. If the snowboard measurement is found to not be representative, either because it has partially or totally blown clear, because drifts have formed in its immediate vicinity, or because snow melted on the snowboard but not on most ground surfaces, the reported snowfall should be the average of as many readings as are needed to obtain an appropriate average over an area including both moderate drifts and moderate clearings.

With the creation and expansion of airport weather stations from the 1930s through the 1950s, came a new emphasis on weather forecasting for air transportation and safety. Hourly weather observations were initiated that included manual measurements of precipitation type and intensity, visibility, and many other weather elements. These became the foundation of the surface airways weather observations (reference) and provided more details about snowfall characteristics than had been previously available. Every hour a complete report of weather conditions was used in a consistent manner from a large number of stations. Conditions were also monitored between hourly reports, and any significant changes were reported in the form of "special" observations. During snowfalls, depths were measured every hour, and special remarks were appended to observations whenever snow depth increased by an inch or more. These "SNOINCR" remarks always caught the attention of meteorologists since they signaled a significant storm in progress. But this also introduced a new complexity into the observation of snow. Instead of observing snowfall once daily, some weather stations reported more frequently. The instructions to airways observers stated that snowfall was to be measured and reported every 6 h. The daily snowfall was then the sum of four 6-h totals. Some weather stations then used the seemingly appropriate procedure to measure and clear their snowboards every hour and add these hourly increments into a 6-h and daily totals.

For some applications, short-interval measurements are extremely useful. However, for climatological applications, snowfall totals derived from short intervals are not the same as measurements taken once daily. For rainfall, a daily total can be obtained by summing short-interval measurements. However, for snowfall, the sum of accumulations for short increments often exceeds the observable accumulation for that period. To demonstrate this, volunteer snow observers were recruited from several parts of the country and measured snowfall for several winters at several locations in the United States. Each observer deployed several snowboards and measured snow accumulations on each board. One board was cleared each hour, one every 3 h, one every 6 h, and every 12 h, and finally one was only measured and cleared every 24 h (see Fig. 5). While results vary from storm to storm, it was very clear that snowfall totals are consistently and significantly higher based on short-interval measurements (Doesken and McKee, 2000). Based on 28 events where measurements taken every 6 h throughout the storms were summed and compared to measurements taken once every 24 h, the 6-h readings summed to 164.4 in., 19% greater than the 138.4-in. total from the once-daily observations. When hourly readings were summed and compared to once-daily measurements, the total was 30% greater. What this means to the user of snowfall data is that two stations, side by side, measuring the same snow amounts at different time intervals may report greatly different values.

In an effort to standardize procedures among different types of weather stations, the National Weather Service issued revised snow measurement guidelines in 1996 (U.S. Dept. of Commerce, 1996a). These guidelines stated that observers should measure snowfall at least once per day but could measure and clear their snowboards as often as but no more frequently than once every 6 h (consistent with long-standing airways instructions). These guidelines were promptly put to the test when an extreme lake-effect snowstorm in January, 1997, produced a reported 77 in. of snowfall in 24 h. Subsequent investigations by the U.S. Climate Extremes Committee (U.S. Dept. of Commerce, 1997) found that this total was the sum of six observations, some of which were for intervals less than 6 h. While the individual measurements were taken carefully, the summation did not conform to the national

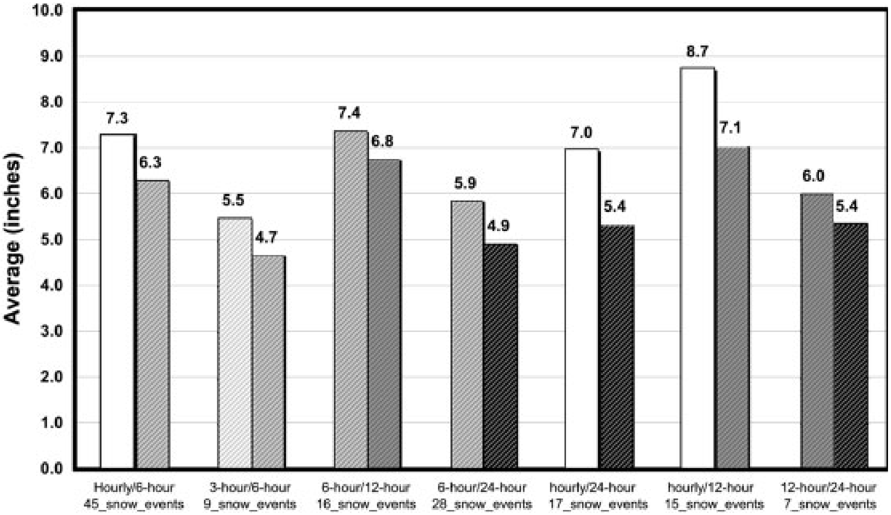


Figure 5 Snowfall comparisons for different measurement intervals. Values were normalized by dividing the total accumulated snowfall in each category by the number of events sampled (Doesken and McKee, 2000).

guidelines and hence could not be recognized as a new record 24-h snowfall for the United States.

The main conclusion here is that where station-to-station comparability and long-term data continuity are the goals, stations must measure in a consistent manner. There may be justification for different observation times and increments, but data from incompatible observation methods should not be interchanged or compared. Many of the data sets in common use today are not fully consistent, so end users may have the responsibility to determine the compatibility and comparability of various station data.

Snow Depth

The measurement of snow depth is not subject to the qualitative and definitional issues that plague the measurement of snowfall. The equipment is again very basic. For areas with shallow or intermittent snowpacks, a sturdy measurement stick is normally carried by an observer to the point(s) of observation and inserted vertically through the entire layer of new and old snow down to ground level. In areas of deep and more continuous snow cover, a fixed snow stake, round or square, is often used—clearly marked in whole inches or in centimeters and permanently installed in a representative location and read “remotely” by an observer standing at a convenient vantage point so that the snow near the snow stake is not disturbed by foot traffic.

The key to useful comparable measurements of snow depth is identifying and maintaining representative locations for taking measurements. Blowing and drifting are inevitable challenges. Uneven melting adds further complications. For uneven snow accumulation, measurements should be taken from several locations proportionally representing both the deeper and shallower areas. Since most traditional weather stations are long distances apart, it is imperative that each measurement represents the predominant conditions in the vicinity of each station.

Snow Water Equivalent (SWE)

This is an extremely important measurement for hydrologic applications. Both flood and overall water supply forecasting rely on accurate measurements of SWE from as many locations as possible to represent the spatial patterns of snow water content that will contribute to subsequent runoff.

The measurement of SWE is taken by only a fraction of National Weather Service stations and is usually accomplished by taking full core samples of total snow on the ground using the 8-in. diameter precipitation gage overflow can. Under deep snow conditions, the melting of snow cores requires considerable amounts of warm water and is tedious and time consuming for observers. Some stations are equipped with special scales that make it relatively easy to take a core sample and immediately estimate the SWE from the weight of the sample. When snow depths exceed 2 ft, the NWS overflow can is inadequate for effectively coring the snowpack.

The Natural Resources Conservation Service and other water resources organizations have a long history of measuring SWE in high snow accumulation areas. Records date back to the 1930s throughout the western mountains with even longer records from a few sites. A wealth of literature, much of it informal and non-peer-reviewed, exists detailing the results of years of experimentation and field testing of devices and techniques to measure snow water in the deep snow regions of North America. Proceedings of the Western Snow Conference and Eastern Snow Conference are great sources for information on the evolution of snow measurements and related research.

Over time, two devices have emerged as the standards for measuring snow water equivalent in deep snow accumulation regions. The Federal Snow Sampler, a portable set of tubes, handle, and a cutter to cleanly penetrate deep snow and ice layers, has been in use for many years (see Fig. 6). Core samples of the snowpack are taken with this instrument, and the sample is then weighed in situ with specially calibrated scales to determine the snow water equivalent of the core. To account for the nonuniform accumulation of snow, several cores are taken at each site. A measurement site is called a "snow course." The final SWE value for a snow course is the average of a set of measurements across the snow course. Each time a snow course is read, core measurements are taken at the same approximate set of individual points. Snow courses have been traditionally read once or twice a month beginning in midwinter and continuing into the spring until all annual snow has melted.

The second instrument, in wide use since the late 1970s, is called a snow pillow (see Fig. 7). This is essentially a scale built at ground level to measure the weight of

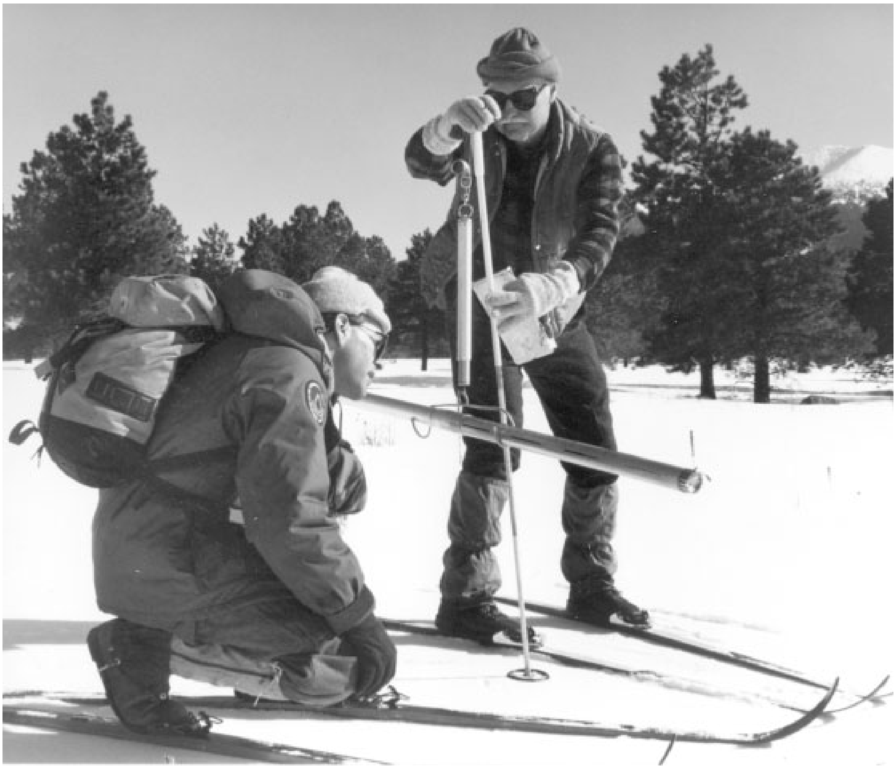


Figure 6 Federal snow sampler consists of a cutter, tubes, and scales for measuring snow water equivalent in moderate and deep snowpacks (from Doesken and Judson, 1997).

the snowpack as it accumulates and melts. Snow pillows were developed to provide remote measurements of SWE without requiring the huge expense of time and effort involved in sending teams of scientists and hydrologic technicians into the back country every month.

As with snowfall and snow depth, the utility and comparability of the observations are only as good as the representativeness of the measurement location and the averaging process. Snow pillow measurements have their own set of challenges. For example, "bridging" and other nonuniformities in load-bearing characteristics within the snowpack can compromise the accuracy of snow pillow measurements.

5 CONTRIBUTION OF TECHNOLOGY TO SNOW MEASUREMENTS

Many aspects of the measurement of snow continue to use only the simplest of instrumentation in the hands of trained and experienced observers. Increasingly,

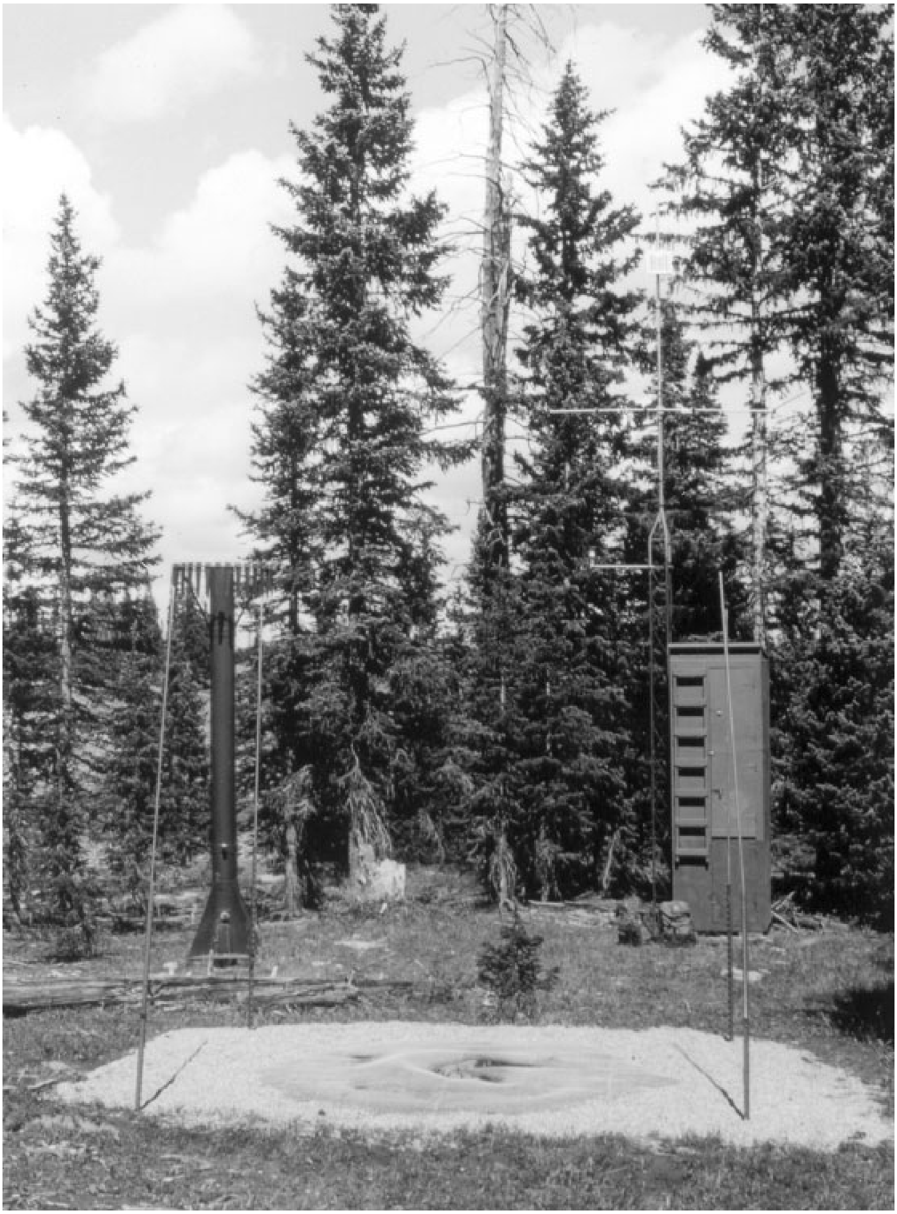


Figure 7 U.S. Department of Agriculture Natural Resources Conservation Service SNOTEL (Snow Telemetry) site in the western United States including snow pillow (in foreground), a shielded standpipe precipitation storage gage (left background), and radio telemetry equipment (from Doesken and Judson, 1997).

scientists and practitioners turn to new technologies and new measurement techniques to acquire the data needed to better understand and apply our knowledge of snow to improve forecasts. Remote sensing is providing data sets showing greater areal coverage and finer resolution information on the spatial distribution of snow. This leads to improved models of snowmelt processes and water supplies. At the same time, improved physical models have pointed out the deficiencies in surface data motivating greater efforts to employ technology to gather more and better data about snow.

In these few pages we cannot do justice to all the technological innovations that are helping improve the measurements of snow and its properties. Rather, we will touch on a few technologies and instruments currently in use. Some of these ideas have been around for many years, but the actual implemention of operational measurement systems is made possible by the remarkable expansion in low-cost computing power in recent years.

Visual Satellite Imagery

From the time the first satellites orbited Earth, it was apparent that snow cover was easily detectable from space. Depth and water content are not easily estimated from reflected visible light, but the extent of snow cover can be readily evaluated from space. The primary limitation is cloud cover, which makes it impossible to see the snow cover below using only visible wavelengths.

Meteorological Radar

Radar refers to the remote-sensing technique of transmitting microwaves of a specified wavelength and receiving, processing, and displaying that portion of the transmitted energy reflected back to the transceiver. Rain and mixed-phase precipitation reflect microwave energy relatively effectively. Snow crystals can also be detected but, depending on crystal structure and temperature, are not detected as well as "wetter" forms of precipitation. The U.S. National Weather Service routinely uses radar to monitor the development, movement, and intensity of snowfall. Improvements in radar realized by the NWS WSR-88D allow estimates of snowfall intensity (Holyrod, 1999) and potential accumulation rates. However, ground truth data remain essential for radar calibration. Detection efficiency varies greatly with distance from the radar, cloud height, and other factors. Still, this technology affords wonderful opportunities for studying snowfall processes in action.

Passive Microwave Remote Sensing

Energy in the form of microwaves is continuously emitted from Earth's surface. Snow crystals within the snowpack scatter and attenuate these microwaves. Aircraft or satellites overhead can detect these emissions and through a series of approximations and algorithms can estimate regional snow cover, water content, and approximate depth. There are various limitations of this technology. For example, shallow or wet snows are more difficult to detect than deeper snows containing little or no liquid water. However, a clear advantage of this technology is that it is not greatly affected by cloud cover, so measurements can be taken day and night during both clear and cloudy conditions.

Gamma Radiation Remote Sensing

The soil near the surface of Earth constantly emits radiation to space in the form of gamma waves. Snow cover attenuates this radiation in proportion to the water content of the snow on the ground. This form of radiation is best detected over relatively narrow bands by receivers mounted on aircraft. Levels of background gamma emissions must be measured in the fall prior to snow accumulation and then along the identical flight path at different times throughout the winter. This method of mapping snow water equivalent is used operationally in several parts of the United States where large river flooding from snowmelt is a common problem.

Acoustic Snow Depth Sensing

Point measurements of snow depth can be taken continuously and remotely. Sound waves from an above-ground transmitter reflect off the snow surface. By measuring the time for the reflected wave to reach the receiver, the distance can be measured that corresponds to a snow depth. The depth of older snow is easiest to measure since the surface tends to become smoother and harder with time. But recent improvements in signal processing have led to accurate measurements of the depth of fresh snow as well. While heavy snow is falling, measurements may be compromised as a portion of the sound wave is reflected by ice crystals in the air.

Satellite Snow Depth Measurements

High-resolution mapping of the elevation of Earth's surface is leading to the opportunity to do similar mapping of snow depth by computing the difference between the elevation of a current observation with the known elevation of the ground from previous satellite measurements. The accuracy of this method requires extremely high-resolution background data and nearly perfect navigation of the data.

Portable Depth/Water Content Sensors

New sensors are being developed for use by ski areas and others concerned about detailed spatial patterns of snow depth and water content. These devices can be sled mounted and pulled by a skier. Using geographical positioning systems to automatically map the sensors location, detailed maps of snow depth/water content can be made.

6 SUMMARY OF SNOW DATA CONTINUITY

More than 100 years of measurements of snowfall, snow depth, and water content are available at hundreds of locations in the United States providing a remarkable resource for meteorological, hydrological, environmental, engineering, and societal applications. The data, however, are far from perfect. All the challenges to accurate observations that are described here have been handled from the very beginning of quantitative observations with varying degrees of success. Observational consistency has been hard to maintain due, in part, to the fact the much of the data have been gathered by volunteers who have received only modest training and who often can take only one observation per day. Even at the nation's primary weather stations, many changes have occurred over time affecting observational consistency. The large natural variability in snowfall sometimes hides the impact of observing changes. Yet, in historical perspective, seemingly small observational changes such as station exposure, time and frequency of observation, and observing procedures do have profound impacts on historical time series. The example in Figure 8 for Sault Ste. Marie, Michigan, is fairly typical of snowfall time series found in the United States. It can be very difficult to distinguish between actual climate variations and true data inhomogeneities.

Perhaps the greatest change in surface observations was the deployment of the Automated Surface Observing System (ASOS) by the National Weather Service during the 1990s as a part of a very extensive national modernization effort. A change in precipitation gages and the automation of the measurement of visibility

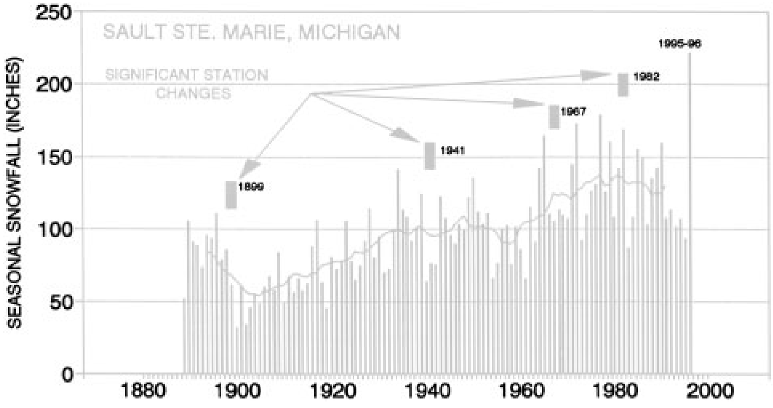


Figure 8 Historic seasonal snowfall totals at Sault Ste. Marie, Michigan. A combination of station moves, changes in exposure, and changes in observational procedures (some of which have not been documented are superimposed on what appears to be a very significant long-term increase in snowfall. Long-term snowfall time series for other parts of the country are plagued by similar problems. Improving the consistency in observations would make it easier to confidently identify important climate variations and trends (from Doesken and Judson, 1997).

and precipitation type have resulted in discontinuous records at several hundred stations. Furthermore, the measurement of snowfall was discontinued completely at many stations since it was not a requirement of the Federal Aviation Administration at that time and did not lend itself to automation.

7 SOURCES OF SPECIALIZED SNOW DATA

Incredible resources about snow, historic data, new measurement technologies, and applications are available, many via the World Wide Web. Some recommended sources include:

National Weather Service, National Operational Hydrologic Remote Sensing Center: *http://www.nohrsc.nws.gov*

National Snow and Ice Data Center (University of Colorado in collaboration with the National Oceanic and Atmospheric Administration: *http://nsidc.org*

National Oceanic and Atmospheric Administration, National Climatic Data Center: *http://lwf.ncdc.noaa.gov*

U.S. Department of Agriculture, Natural Resources Conservation Service, National Water and Climate Center: *http://www.wcc.nrcs.usda.gov*

U.S. Army Corps of Engineers Cold Regions Research and Engineering Laboratory: *http://www.crrel.usace.army.mil*

National Aeronautics and Space Administration: *http://www.nasa.gov*

Website addresses are as of autumn 2001.

REFERENCES

Bamzai, A. S., and J. Shukla (1999). Relation between Eurasian snow cover, snow depth, and the Indian summer monsoon: An observational study, *J. Clim.* **12**, 3117–3132.

Bergeron, T. (1935). On the Physics of Cloud and Precipitation. Proces Verbaux des Seances de l'Association de Metiorolgie, UGGI, 5th assemblee generale, Lisbon, Sept. 1933.

Doesken, N. J., and A. Judson (1997). *The Snow Booklet: A Guide to the Science, Climatology and Measurement of Snow in the United States*, Atmospheric Science Department, Colorado State University, Fort Collins.

Doesken, N. J., and R. J. Leffler (2000). Snow foolin', *Weatherwise* **53**(1), 30–37.

Doesken, N. J., and T. B. McKee (2000). Life after ASOS (Automated Surface Observing System)—Progress in National Weather Service Snow Measurement. Proceedings, 68th Annual Meeting, Western Snow Conference, April 18–20, Port Angeles, WA, pp. 69–75.

Goodison, B. E. (1978). Accuracy of Canadian snow gauge measurements, *J. Appl. Meteorol.* **27**, 1542–1548.

Goodison, B. E., and J. R. Metcalf (1992). The WMO Solid Precipitation Intercomparison: Canadian Assessment, WMO Technical Conference on Instruments and Method of Observation, WMO/TD No. 462, WMO, Geneva, pp. 221–225.

Gray, D. M., and D. H. Male (Eds.) (1981). *The Handbook of Snow: Principles, Processes, Management and Use*, Toronto, Pergamon.

Hobbs, P. V. (1974). *Ice Physics*, Bristol, Oxford University Press.

Holyrod, E. W. (1999). Snow Accumulation Algorithm for the WSR-88D Radar: Supplemental Report, Report #R-99-11, USDI, Bureau of Reclamation, Technical Service Center, Denver.

Judson, A., and N. J. Doesken (2000). Density of freshly fallen snow in the Central Rocky Mountains, *Bull. Am. Meteor. Soc.* **81**(7), 1577–1587.

Mason, B. J. (1971). *The Physics of Clouds*, 2nd ed., Oxford, Clarendon.

McKee, T. B., N. J. Doesken, and J. Kleist (1994). Climate Data Continuity with ASOS—1993 Annual Report for the Period September 1992–August 1993, Climatology Report 94-1, Atmospheric Science Department, Colorado State University, Fort Collins.

Minsk, L. D. (1998). *Snow and Ice Control Manual for Transportation Facilities*, New York, McGraw-Hill.

Nakaya, U. (1954). *Snow Crystals, Natural and Artificial*, Cambridge, MA, Harvard University Press.

National Research Council (1998). *Future of the National Weather Service Cooperative Observer Network*, Washington, DC, National Academy Press.

National Weather Service (2001). Snow Measurement Video, prepared by Department of Atmospheric Science, Colorado State University for National Weather Service, Silver Spring, MD, 30 min.

Takahashi, T., T. Endoh, G. Wakahama, and N. Fukuta (1991). Vapor diffusional growth of free-falling snow crystals, *J. Meteorol. Soc. Jpn.* **69**, 15.

U.S. Dept of Commerce (1996a). Snow Measurement Guidelines (rev. 10/28/96), NOAA, Silver Spring, MD, National Weather Service.

U.S. Dept. of Commerce (1996b). *Surface Weather Observations and Reports*, National Weather Service Handbook No. 7, NOAA, Silver Spring, MD, National Weather Service, Office of Systems Operations.

U.S. Dept. of Commerce (1997). Evaluation of the Reported January 11–12, 1997 Montague, New York, 77-inch 24-hour Lake-Effect Snowfall, NOAA, Silver Spring, MD, National Weather Service, Office of Meteorology.

Walsh, J. E., D. R. Tucek, and M. R. Peterson (1982). Seasonal snow cover and short-term climatic fluctuations of the United States, *Mon. Wea. Rev.* **110**, 1474–1485.

Yang, D., J. R. Metcalfe, B. E. Goodison, and E. Mekis (1993). An Evaluation of the Double Fence Intercomparison Reference Gauge, Proceedings, Eastern Snow Conference, 50th Meeting, Quebec City, Canada, pp. 105–111.

INDEX

Handbook of Weather, Climate, and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21490-6 © 2003 John Wiley & Sons, Inc.