Upper Atmosphere and Magnetosphere (1977)

Pages
180

Size
8.5 x 10

ISBN
0309026334

Geophysics Study Committee; Geophysics Research Board; Assembly of Mathematical and Physical Sciences; National Research Council

STUDIES IN GEOPHYSICS

The Upper Atmosphere and Magnetosphere

Geophysics Study Committee
Geophysics Research Board
Assembly of Mathematical and Physical Sciences
National Research Council

NATIONAL ACADEMY OF SCIENCES
Washington, D.C. 1977

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the Councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the Committee responsible for this report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The Geophysics Study Committee is pleased to acknowledge the support of the U.S. Geological Survey, the National Science Foundation, the U.S. Energy Research and Development Administration, the Defense Advanced Research Projects Agency, the National Oceanic and Atmospheric Administration, and the National Aeronautics and Space Administration for the conduct of this study.

International Standard Book Number 0-309-02633-4

Library of Congress Catalog Card Number 77-82812

Available from:

Printing and Publishing Office, National Academy of Sciences
2101 Constitution Avenue, Washington, D.C. 20418

Printed in the United States of America

Geophysics Research Board

Geophysics Study Committee

PHILIP H. ABELSON, Carnegie Institution of Washington, *Cochairman*
THOMAS F. MALONE, Holcomb Research Institute, *Cochairman*
LOUIS J. BATTAN, University of Arizona
CHARLES L. DRAKE, Dartmouth College
RICHARD M. GOODY, Harvard University
FRANCIS S. JOHNSON, University of Texas at Dallas
WALTER B. LANGBEIN, U.S. Geological Survey, retired
HUGH ODISHAW, University of Arizona

NRC Staff

PEMBROKE J. HART
DONALD C. SHAPERO

Liaison Representatives

JAMES R. BALSLEY, U.S. Geological Survey
EUGENE W. BIERLY, National Science Foundation
GEORGE A. KOLSTAD, U.S. Energy Research and Development Administration
CARL F. ROMNEY, Defense Advanced Research Projects Agency
WALTER TELESETSKY, National Oceanic and Atmospheric Administration
FRANCIS L. WILLIAMS, National Aeronautics and Space Administration

Upper Atmosphere Panel

FRANCIS S. JOHNSON, University of Texas at Dallas, *Chairman*
JAMES L. BURCH, National Aeronautics and Space Administration
HERBERT C. CARLSON, University of Texas at Dallas
EDWIN F. DANIELSEN, National Center for Atmospheric Research
THOMAS M. DONAHUE, University of Michigan
WILLIAM B. HANSON, University of Texas at Dallas
THOMAS W. HILL, National Oceanic and Atmospheric Administration
WILLIAM H. HOOKE, National Oceanic and Atmospheric Administration
DONALD M. HUNTEN, Kitt Peak National Observatory
JEAN-FRANCOIS LOUIS, National Oceanic and Atmospheric Administration
RAYMOND G. RÓBLE, National Center for Atmospheric Research
ROBERT G. ROPER, Georgia Institute of Technology
CHALMERS F. SECHRIST, JR., University of Illinois
RICHARD A. WOLF, Rice University

Preface

Early in 1974, the Geophysics Research Board completed a plan, subsequently approved by the Committee on Science and Public Policy of the National Academy of Sciences, for a series of studies to be carried out on various subjects related to geophysics. The Geophysics Study Committee was established to provide guidance in the conduct of the studies.

One purpose of the studies is to provide assessments from the scientific community to aid policymakers in decisions on societal problems that involve geophysics. An important part of such an assessment is an evaluation of the adequacy of present geophysical knowledge and the appropriateness of present research programs to provide information required for those decisions. Some of the studies place more emphasis on assessing the present status of a field of geophysics and identifying the most promising directions for future research. Topics of studies for which reports are currently in preparation include geophysical predictions and energy and climate. Recently completed studies have treated the geophysics of estuaries and water and climate.

Each study is developed through meetings of the panel of authors and presentation of papers at a suitable public forum that provides an opportunity for discussion. In completing final drafts of their papers, the authors have the benefit of this discussion as well as the comments of selected scientific referees. Responsibility for the individual essays rests with the corresponding authors.

The essays in this volume were presented in a symposium at an American Geophysical Union meeting that took place in San Francisco in December 1975. This report, which contains those essays, identifies aspects of the physics of the upper atmosphere and magnetosphere where progress can be expected, underscores the direct importance of improved understanding of this subject to mankind, and recommends means to obtain that understanding.

The introductory chapter provides an overview of the study summarizing the highlights of the essays and formulating conclusions and recommendations. In preparing it, the Chairman of the Panel had the benefit of meetings and discussions that took place at the symposium and the comments of the panel of authors and selected referees. Responsibility for its content rests with the Geophysics Study Committee and the Chairman of the Panel.

Overview and Recommendations

INTRODUCTION

The upper atmosphere and magnetosphere form a boundary region between the earth and space. In this region, a large and complicated system filling a volume of space about a million times greater than the volume of the solid earth, the earth's magnetic field dominates the motion of ionized gases. The variable components of the sun's radiation and its outer atmosphere interact with this system in often spectacular ways.

HISTORICAL DEVELOPMENT

Mankind has long had an interest in the lower atmosphere because of weather and climate. It is perhaps surprising to note how long some aspects of the upper atmosphere have also been of interest. More than 700 years ago an unknown Norwegian described auroras eloquently and speculated on their possible cause. However, serious scientific interest in the upper atmosphere did not arise until much later, near the beginning of the twentieth century, when improved geophysical measurements became available. At that time, small rapid variations in the earth's magnetic field were attributed to electrical currents in the upper atmosphere, and the existence of the distinct region of the atmosphere

that we now call the stratosphere was discovered in the course of measurements made by balloonborne instruments.

After this, advances in understanding emerged with ever-increasing speed. Early in the twentieth century, evidence for a conducting layer in the upper atmosphere (the ionosphere) was found when radio signals were propagated from Europe to America. The role of energetic charged particles in producing aurora and the importance of the geomagnetic field in localizing the effects were recognized. Direct evidence of the ionosphere was obtained from pulse sounding, a technique that was the forerunner of radar. Radio communications by means of ionospheric reflection came into widespread use.

Progress in understanding increased greatly when rockets and space vehicles made possible direct, *in situ* measurements, starting at the end of World War II. The Van Allen radiation belts were discovered during the International Geophysical Year, and the importance of the geomagnetic field in space became evident. It became clear that this huge volume where the motion of charged particles is dominated by the earth's field should be regarded as earth-space, not interplanetary space. Space vehicles leaving the earth and its magnetic field entirely found the motion of charged particles in interplanetary space to be dominated by the solar wind. This discovery was a great step forward in understanding one of the ways in which events on the sun influence the earth.

Since there now have been nearly three decades of probing with rockets, and two with satellites, one might wonder if most or all of the interesting and important problems have been solved. Indeed, the level of understanding of what is going on in the upper atmosphere and in space near the earth is impressive, as the following chapters will illustrate. However, *there are still many significant problems that remain unsolved* and many important areas in which our knowledge is incomplete or inexact. The rate of discovery of new and important physical processes remains high. Many aspects of this subject are related to important elements of man's environment, while others illustrate physical processes whose nature needs to be explored to increase our understanding of the universe. In this sense, the problems of the upper atmosphere and magnetosphere may be described as of high intrinsic interest because they are important problems in their own right and of high extrinsic interest because they are related to other areas of human activity.

The emphasis in the first decade or two of the space age has been on elucidating local processes such as photochemical reactions, diffusion, energy exchange, and excitation of atoms and molecules. The available measurement techniques have contributed to this preoccupation with microphysics. The satellite effort known as the Atmosphere Explorer Program represents the culmination of this phase of research into thermospheric processes. We now understand local phenomena in the thermosphere well enough to see that further progress will require an understanding of global processes, including circulation, waves on many scales, electric fields, and currents. Increasing interest can be expected in the macrophysics of the thermosphere. The Electrodynamic Explorer Program represents a first step in this direction, but *in situ* spacecraft measurements are not ideal for gathering synoptic information on conditions in the upper atmosphere. Simultaneous measurements are needed, and to get these a global network of ground-based upper atmospheric observatories and remote-sensing satellites will be needed. Appropriate remote-sensing techniques now exist; examples are airglow spatial scanning photometers, ionosondes, incoherent-scatter radar, lidar, and airglow Doppler measurements. There is, therefore, a need to develop and deploy, on the ground and

in space, intercalibrated, reliable, uniform, upper atmospheric observatories that would produce compatible data that could be rapidly processed and centrally archived. The worldwide standard seismograph network, which made notable contributions to solid-earth geophysics, provides an example of what is needed.

PRACTICAL APPLICATIONS

The problem of stratospheric pollution and its effects on the ozone layer has made clear how detailed our understanding must be to predict accurately such effects that may be of immense importance to man's future. The problem first attracted widespread interest in connection with the possible effects of supersonic transports, initially in terms of water vapor from the exhaust and later in terms of the effect of nitrogen oxides produced in the engines. Next, the effects of halocarbons became a matter of concern. Still another potentially serious problem arises from the release of nitrous oxide from chemical fertilizer. Although considerable research is now being done on all aspects of the ozone question, support for research was rather sparse prior to recognition of this problem. It was, and is, an area in which progress can be made, and such progress is largely dependent on the acquisition of the appropriate data relating to atmospheric behavior and chemical reactions.

One might assume that the importance of radio communications by ionospheric reflections has largely vanished with the advent of satellite communications. This is so in some fields, but much reliance is still placed on communication involving ionospheric reflection. It is important in some less developed parts of the world, but it is also important for some segments of advanced societies; for example, some parts of the military are unwilling to become wholly dependent on satellite communication because of their inherent vulnerability. The influence of the ionosphere on radio communications therefore remains important. It is even of importance to satellite communications, where small-scale ionospheric irregularities sometimes make communications unreliable even at the high frequencies that are used in such systems.

The upper atmosphere provides a natural plasma laboratory, one with parameters that are often difficult to reproduce in the laboratory. A great variety of plasma instabilities and physical processes have been discovered or demonstrated in the upper atmosphere, thus providing an interchange between the fields of geophysics (including the upper atmosphere, ionosphere, and magnetosphere) and plasma physics. Many of the plasma-physics effects in the upper atmosphere are important to the overall behavior of the ionosphere and magnetosphere. An understanding of these effects is essential in the prediction of many ionospheric effects, including some that are of importance in the field of radio communications.

The possible influence of solar activity and magnetic variations in space on weather and climate remains a controversial subject, although the preponderance of evidence indicates that relationships do exist. Physically satisfactory mechanisms have yet to be proposed. If a relationship is established and mechanisms are identified, it is almost certain that the mechanisms will involve the upper atmosphere and magnetosphere. The societal implications of weather and climate changes make it important that physical processes in the upper atmosphere and magnetosphere be studied with a view to finding the coupling mechanisms between solar activity, interplanetary magnetic variations, and weather and climate.

NEED FOR COORDINATION

The rest of this chapter highlights many of the important unanswered questions about the physical processes occurring in the upper atmosphere and magnetosphere, questions on which progress can be expected with improved observational data and additional theoretical effort. Advances in instrumentation techniques are just now providing the opportunity for the first time to make measurements that are vital to the accurate description of the magnetosphere. Perhaps the best example of this is plasma-drift instrumentation, which now makes measurements of electric fields feasible and trustworthy. Because electric fields are closely related to the fields of plasma motion, a means is now at hand for accurately observing motions in the magnetosphere; these observations are essential to the developing understanding of magnetospheric behavior.

Events taking place in one part of the huge volume of the magnetosphere are often closely related to other events occurring in distant parts of the same volume, posing a gigantic observational problem in developing a coherent picture of how the whole system operates. Because of the high degree of interrelation among various phenomena and because of the huge volume over which such phenomena occur, research efforts need to be carefully coordinated; the opportunity to make substantial progress in this field with isolated observational programs has largely passed. The relationships extending through this huge system are complicated and often unexpected. It is almost impossible to study any one part of the system without taking into consideration its interactions with at least some of the other parts. This also poses a problem in presenting the subject: how to organize the material and where to start.

ORDER OF PRESENTATION

The approach used in this presentation is to look at the outermost portion of the system first, where the solar wind (the extended and outward moving portion of the solar atmosphere) interacts with the magnetosphere and then to look at progressively deeper levels of the magnetosphere–atmosphere system. The rationale for this approach is that it follows the direction of the predominant energy flow, and it seems natural to describe first those regions that strongly influence the behavior of the other regions. However, the interactions are complex, and within the magnetosphere–atmosphere system the directions of energy flow sometimes reverse, so the prescription is imperfect. There is simply no natural order that is clearly best from all viewpoints.

In the approach adopted here, we neglect the largest flow of energy from the sun to the earth—the solar electromagnetic radiation that for the most part penetrates the atmosphere and heats the earth's surface. However, this energy flow is constant or at least so nearly so that claimed variations are controversial or debatable. It is a most important energy flow because it supports life on earth and produces weather. In this presentation, however, we focus on the variable influences of the sun on the earth: variations in the strength of the solar wind, ejection of energetic particles by the sun (low-energy cosmic radiation), and variations in the ultraviolet and x-ray emissions from the sun. These variable factors are the ones of most interest in the study of solar–terrestrial physics.

Although the upper atmosphere, magnetosphere, and solar–terrestrial effects are discussed in a fairly organized way, the treatment is far from complete. There are many important areas that have been left out entirely. Even the areas that have been covered have been treated very briefly. The intent has been to be illustrative rather than comprehensive and to convey a good perspective of the

state of knowledge and future opportunities in the field without presenting all the supporting details. The use of references is also limited, in keeping with the brevity of the presentations and with a preference for review articles that should be consulted for more complete treatments of the subject matter and for further references.

The various areas selected for separate descriptions in the following chapters are set forth briefly in the following sections. Important questions that need to be answered in each area are posed. Some of the international cooperative efforts that are important to this field are discussed in the last section. Conclusions and recommendations are set forth in the following section.

CONCLUSIONS AND RECOMMENDATIONS

The upper atmosphere and magnetosphere and their interaction with the solar wind constitute an incredibly complicated system, so it is not surprising that many aspects of its behavior are still not understood. What is surprising is the simple nature of some of the primary problems whose behavior still eludes understanding. The interaction of a moving (supersonic) ionized plasma with a fixed magnetic dipole is conceptually a very simple problem. Still, in the solar-wind/earth context, the basic physics of this problem is not understood; how does solar-wind plasma penetrate the magnetic field of the earth, and how is energy transferred from the solar wind so as to drive a global pattern of magnetospheric convection? These questions are about as basic as any that can be asked. *Answers to such questions about the physics of the upper atmosphere and magnetosphere will be forthcoming soon if appropriate efforts are expended.*

Because of the immense size of the magnetosphere–atmosphere system (relative to the size of the earth) and the extent to which processes in one part of the system influence distant parts of the system (mainly by electrical transfer of energy), *coordinated simultaneous measurements are of great importance.* Early in the space age, isolated measurements led to many important discoveries (including the radiation belts), but this opportunity has passed. *Coordinated planning of observations by spacecraft and other means is essential.* The International Magnetospheric Study (IMS) provides a good example of such coordination, making maximum effective use of available resources.

Continued investigations with spacecraft will be very productive even after the completion of the IMS. It is not possible to foresee the exact nature of investigations that will be needed in the decade of the 1980's, but it is a reasonable expectation that multiple spacecraft missions to the boundaries of the magnetosphere will be required, with no less attention to coordination than during the IMS. The AMPS Space Shuttle payload will provide an important base of operations in the lower magnetosphere, particularly valuable for active experiments and for coordinated measurements with satellites far out in the magnetosphere.

The overall structure of the upper atmosphere is reasonably well known, but many important phenomena within the atmosphere are not well understood. Progress in this area can be expected more or less in proportion to the efforts expended, provided they are well planned. Concern about the fragility of the ozone layer provides an excellent example; while much progress has been made since the ozone problem arose in connection with supersonic transports, a more confident understanding is essential and can be expected with continued re-

search.* Farther up, the interactions between the magnetosphere and that part of the atmosphere known as the thermosphere, and the overall behavior of the thermosphere as it is pushed one way by solar ultraviolet heating and another by energy input through the magnetosphere (both variable), are subjects on which important progress is to be expected. It has become increasingly clear that weather responds to solar events through physical mechanisms that have not yet been visualized; they probably involve the magnetosphere and the thermosphere. *The expectation of progress in the area of sun–weather relationships is speculative because physical links have not been identified, but it is an important area to pursue.*

The appropriate means of attack on problems of the thermosphere is with satellites of the Atmosphere Explorer type, two of which are in operation at this time. *The Committee concludes in agreement with the Space Science Board in its report,* Opportunities and Choices in Space Science, 1974, *that additional satellites of the Atmosphere Explorer type will be needed in the early 1980's, after the end of the IMS.* At that time, such satellites surely would be candidates for launch by the Space Shuttle, which may permit the use of a greater number of satellites built to less expensive standards than the current Atmosphere Explorers.

Balloons provide an important means of making in situ *observations in the stratosphere, and their continued use is important.*† However, it will not be feasible to get the needed geographically widespread coverage using balloons, and development of remote-sensing devices for worldwide monitoring of the stratosphere from satellites is essential, even though in some cases the remote-sensing observations will be less satisfactory than *in situ* measurements.

The ionosphere has proved to be much more variable than one would have ever expected on the basis of a highly stratified atmosphere irradiated with solar ultraviolet ionizing radiation. Part of its complicated structure results from motions of the thermosphere, but plasma instabilities and drifts are other sources of variations, especially small scale. Both types of variation produce important effects on some communication systems, and the plasma instabilities are often illuminating in the field of plasma physics. *Satellites such as those required for thermospheric investigations and the AMPS Shuttle payload are appropriate for attack on F-region problems, and incoherent-scatter radar observations provide a powerful technique on the ground.*‡ *The Committee recommends that steps be taken to reassert U.S. leadership in the latter area.* The very fact that other countries have moved vigorously into incoherent-scatter radar observations for atmospheric investigations is evidence of the expectations for continued productivity from such observatories.

Investigations of the lower ionosphere involve complicated chemistry; since satellites cannot fly at these altitudes, rockets or remote sensing must be used. The chemistry is an extension of that involved in the ozone layer, and the vulnerability of this atmospheric region to change resulting from pollution

*See the following reports published by the National Academy of Sciences, Washington, D.C.: Climatic Impact Committee, *Environmental Impact of Stratospheric Flight,* 1975; Panel on Atmospheric Chemistry, *Halocarbons: Effects on Stratospheric Ozone,* 1976; Committee on Impacts of Stratospheric Change, *Halocarbons: Environmental Effects of Chlorofluoromethane Release,* 1976.

†See Balloon Study Committee, Geophysics Research Board, *The Use of Balloons for Physics and Astronomy,* National Academy of Sciences, Washington, D.C., 1976.

‡See Ad Hoc Panel on a New Upper Atmosphere Observatory, Geophysics Research Board, *Upper Atmosphere Observatory Criteria and Capabilities,* National Academy of Sciences, Washington, D.C., 1971.

argues that *rocket programs must be preserved and remote-sensing techniques developed for use in satellites.*

Investigation of the earth's upper atmosphere and magnetosphere is an important element of planetary investigations. Of all the planets, the earth is, of course, by far the easiest to investigate. There is an amazing diversity in both atmospheres and magnetospheres among the planets. Confidence that we understand the behavior of upper atmospheres and magnetospheres of the other planets must rest in part in our understanding of the earth system. As long as we are in doubt about some of the most basic aspects of magnetospheric behavior on earth, such doubts must extend to our understanding of other planetary magnetospheres. As long as we do not understand the process of magnetic annihilation in the earth's magnetosphere, application of this concept in other astrophysical situations must be doubtful. In this sense, *investigation of the earth's upper atmosphere and magnetosphere should be considered a part of the foundation upon which planetary and astrophysical investigations are based.*

SOLAR-WIND INTERACTIONS

The solar wind is a plasma (ionized gas) that blows away from the sun. In passing the earth, it interacts with the earth's magnetic field and severely deforms it, producing the magnetospheric cavity in the solar wind, a vast region surrounding the earth from which the solar wind is largely excluded, as illustrated in Figure 1. A collisionless shock wave is produced in the solar wind upstream of the magnetosphere, usually referred to as the bow shock. Away from the sun, the magnetosphere stretches out into a long tail—the magnetotail—reaching far beyond the moon's orbit, and perhaps much farther than depicted in Figure 1.

The collisionless shock wave is a phenomenon that has wide applications in plasma physics but is not well understood theoretically. Detailed observations of the structure of the earth's bow shock during the next few years, particularly from multiple-satellite studies, will provide invaluable clues that are not avail-

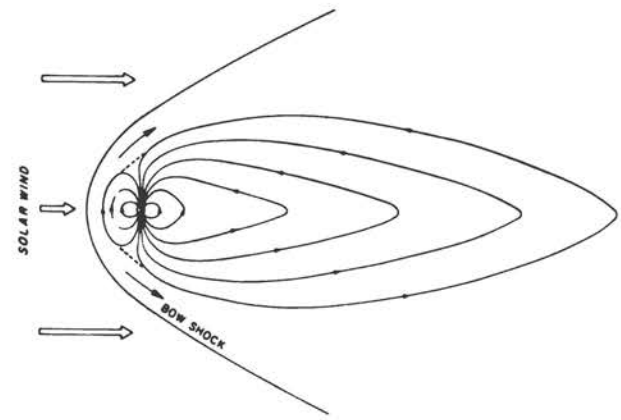


FIGURE 1 The solar wind blows away from the sun through interplanetary space, but it is diverted by the earth's magnetic field, producing the magnetosphere, a large volume of space around the earth from which the solar wind is largely excluded. A shock front, known as the bow shock, is formed in the solar wind upstream from the magnetosphere.

able from laboratory studies. Such shocks undoubtably play important roles in cosmic physics.

An important question about the interaction of the magnetosphere and the solar wind is the degree to which some magnetic-field lines from the earth join with some of the magnetic-field lines in the solar wind, something that would be brought about by a process known as magnetic merging. Although solar-wind magnetic-field lines are not shown in Figure 1, where all geomagnetic-field lines are shown as closing within the magnetosphere, the opportunity for connection would be best if the magnetic-field lines in the solar wind had a southward-directed component, for then they could be connected to geomagnetic-field lines near the equatorial plane by magnetic merging. (If the interplanetary field were directed northward, the two fields would be parallel where they come into juxtaposition near the equatorial plane, and connections could not be established.) In fact, geomagnetic activity is enhanced when the interplanetary field is directed southward, and this constitutes a strong argument that magnetic merging does occur. Another feature of this picture including magnetic merging is that all the geomagnetic-field lines from within a region around the magnetic pole, known as the polar cap, connect into the solar wind, and energy flow from the solar wind to the atmosphere occurs as a result of electrical currents flowing along magnetic-field lines between the solar wind and the ionosphere.

Magnetic merging is an essential process for interconnection between the geomagnetic field and the interplanetary magnetic field, but it is still a controversial and poorly understood process. It is also a process whereby charged particles in space plasmas are energized, the process sometimes being described as conversion of magnetic energy into charged-particle energy. Observations made during the next decade are expected to provide direct experimental evidence regarding the degree of interconnection of magnetospheric and interplanetary magnetic fields. This evidence is essential in establishing the importance of magnetic merging in the solar-wind–magnetosphere interaction, and hence its plausibility as an energy-conversion mechanism of general applicability in collisionless plasmas throughout the universe.

Although the picture involving interconnection between the interplanetary and geomagnetic fields is the generally accepted one (in spite of its difficulties), there is an alternative view, namely, that energy and momentum transfer between the solar wind and the magnetosphere occurs by a viscouslike mechanical interaction rather than by current flow between the solar wind and the polar cap ionosphere. In this case, the current flow would take place between the viscous layer near the surface of the magnetosphere and the polar-cap ionosphere.

A surprising feature of the magnetosphere is the relatively large degree to which the solar-wind plasma enters the magnetosphere, contrary to the oversimplified picture in Figure 1, which shows the magnetosphere diverting and excluding all the solar wind. It is further surprising that some of the plasma is energized or heated to some ten million degrees inside the magnetosphere—many times hotter than the solar-wind plasma itself. Considerable work will be needed, both in theoretical modeling and in obtaining more complete observational knowledge, in order to understand these recently discovered processes. Again, the results can be expected to be of cosmic significance.

A particularly intriguing problem is the implausible, but statistically significant, evidence indicating that the solar wind exerts a measurable influence on weather patterns, in spite of its negligibly small energy input to the lower atmosphere. Mechanisms that can explain this relationship must be investigated.

THE MAGNETOSPHERE

The magnetosphere is the region surrounding the earth where the movement of ionized gases (plasmas) is dominated by the geomagnetic field. Its lower boundary is determined by the altitude where neutral atmospheric gases become high enough in concentration (with decreasing altitude) so that collisions between ions and neutral particles dominate the motions of the ions. The ionized gases in the magnetosphere take part in a gigantic pattern of magnetospheric circulation or "convection" in which the plasma over the polar regions is swept away from the sun while the plasma outside the auroral zone (i.e., at lower latitudes) is swept toward the sun. (Alternative views of mechanisms for driving this convection were mentioned in the previous section.) Continuity requires that there be a return flow somewhere, and it is observed to be mainly just outside the auroral zone, thus causing a sharp reversal of flow near the auroral zone, especially at local magnetic times near 0600 and 1800 hours. Away from the auroral zone, the flow toward the sun is much weaker. At still lower latitude (below about 60° geomagnetic), the plasma does not have this convection pattern but, instead, corotates with the earth.

At times, the geomagnetic field is disturbed by the occurrence of magnetic storms and substorms. The substorm has been identified with large-scale changes in the magnetotail in which the total flux of magnetic field in the magnetotail is suddenly reduced, thus producing large induced electric fields. The magnetic storm, which involves a large number of substorms, causes a small reduction in the magnetic field over most of the earth's surface, something indicative of an increase in the total energy of ionized particles trapped in the geomagnetic field.

Winds in the neutral atmosphere tend to drag ions along with them, but this effect diminishes above the altitude at which the frequency of collisions of ions with neutral particles becomes small compared with the ion's gyration frequency in the magnetic field. The movement of the conducting plasma in the magnetic field gives rise to voltage differences that are felt through large volumes of the magnetosphere because of the high electrical conductivity along the magnetic-field lines. Thus the induced electric fields perpendicular to magnetic-field lines influence the motion of other charged particles in distant parts of the magnetosphere. (Corotation of the plasma with the earth may also be explained in this way.) Magnetospheric convection, driven by an interaction with the solar wind, involves a large-scale motion of magnetospheric plasma that, by collisions with neutral particles, may set the neutral atmosphere into motion. The question is: to what extent do the circulation patterns of the magnetospheric and ionospheric plasma couple into those of the neutral atmosphere?

Auroras constitute one of the most long-standing scientific problems confronting man. It seems clear that the auroral problem is not going to be solved by a flash of inspiration; instead it is yielding grudgingly as facts are accumulated. Auroras occur mainly in an oval pattern around the magnetic pole, and this auroral oval encompasses the polar cap. Magnetic-field lines extending outward from the polar cap are generally considered not to be closed in the vicinity of the earth but rather to extend into the solar wind. Thus solar-wind particles have some possibility of moving along magnetic-field lines into the polar-cap ionosphere, possibly entering into the auroral process. Other plasma populations exist in the magnetosphere, and one of these, the plasma sheet, connects to the ionosphere near the auroral zone. However, particles in the magnetosphere and in the solar wind have not generally been observed to have distributions of energies that agree with those of auroral particles. Thus, the basic question in

connection with auroras remains: by what processes are the electrons and ions that excite auroras accelerated to energies of several thousand electron volts? Evidence is now strong that there are limited regions in which a strong electric field exists along the direction of the magnetic field, something that contradicts the idea that electrical conductivity is high along the magnetic-field lines.

The outermost portion of the ionosphere extends well into the magnetosphere, thus providing magnetospheric plasma at least in a region close to the earth, including a region known as the plasmasphere that extends to about 4 earth radii in the equatorial plane. Some of the ionospheric plasma may be carried into distant parts of the magnetosphere by magnetospheric convection. In the polar regions, light ions are ejected upward from the ionosphere, giving rise to a phenomenon known as the polar wind. Although these particles might escape from the earth, it is most likely that they do not because the convection pattern probably reverses their direction of drift before they can escape; hence these particles may remain in the magnetosphere and contribute to the magnetospheric plasma. The degree to which the earth's ionosphere acts as a source of magnetospheric plasma poses an interesting question.

In many cases, energetic ions from the magnetosphere strike the atmosphere and produce important effects. In order for the effect to continue, some of the plasma particles that would otherwise have remained trapped in the magnetic field must have their trajectories disturbed so that they can follow the magnetic field down into the atmosphere without reflection. Sometimes disturbances occur that clearly change the trajectories of trapped particles and scatter them, allowing some of them to enter the atmosphere. Therefore, the question arises: what plasma processes are responsible for scattering otherwise trapped magnetospheric electrons and ions, allowing some of them to strike the atmosphere?

Electric currents between magnetosphere and ionosphere provide an important energy-transfer mechanism. This is especially true during the occurrence of substorms, when strong magnetic-field-aligned currents flow into the auroral zone. Any disruptive mechanism associated with magnetospheric convection will require the flow of current into and out of the auroral zone (or the edge of the polar cap). However, it is not clear what plasma populations carry the field-aligned currents that transfer electrical energy from the magnetosphere to the ionosphere.

Conditions favorable to the development of substorms include a southward-directed interplanetary magnetic field, which in turn leads to enhanced magnetic flux in the magnetospheric tail. Then, an instability suddenly occurs in which the current separating the two halves of the tail field is interrupted, possibly because of a plasma instability in the plasma through which the current flows. Auroras become active, extensive, and bright. The magnetic flux in the magnetotail is rapidly reduced, and the decrease in magnetic energy is the apparent source of the substorm energy. Large induced electric fields must occur. The key question is: what large-scale plasma instabilities are responsible for magnetospheric substorms? Understanding this phenomenon may contribute to the understanding of other explosive events in astrophysics, such as solar flares.

THE THERMOSPHERE

The thermosphere is that part of the upper atmosphere above about 80 km; it is characterized by rising temperatures up to about 300 km and nearly constant temperature at higher altitudes, as indicated in Figure 2. Except for its lower

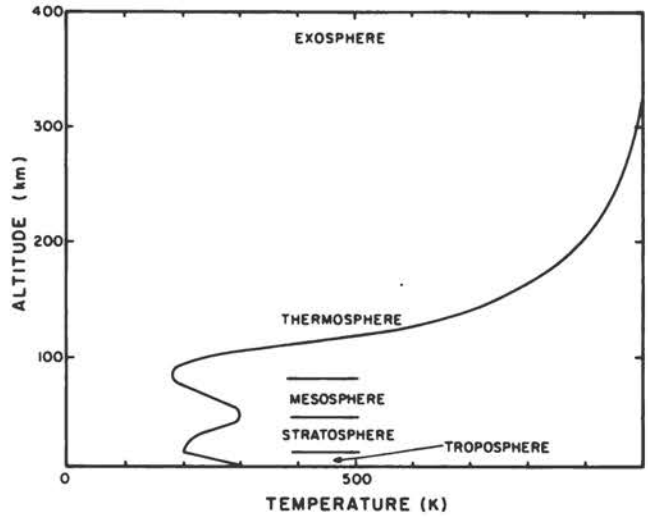


FIGURE 2 The temperature distribution through the atmosphere provides a basis for dividing the atmosphere into layers or spheres. The troposphere generates weather phenomena. The stratosphere is the location of the ozone layer. The exosphere is the outermost portion of the atmosphere where the gas is so rarified that collisions between gas particles occur only rarely, causing the temperature distribution there to be isothermal with height.

portion, it is dominated by atomic rather than molecular species, and each constituent is distributed largely as if the others were not present, a condition generally referred to as diffusive equilibrium. Because of limited information, most existing descriptions of the thermosphere tend to be oversimplified. In actuality, a great diversity of physical and chemical processes affects the behavior of the region.

The main characteristic of the thermosphere is its dynamically active state. The thermosphere is influenced by the two forms of solar energy that interact with the earth's outer environment: solar electromagnetic energy deposited mainly at low latitudes and in the summer hemisphere and the energy of the solar wind deposited within the thermosphere at high latitudes, mainly through intermittent auroral processes. As a result, the circulation, temperature, and compositional structure of the thermosphere exhibit large variations about a global mean state. The circulation is driven by absorbed solar electromagnetic energy during quiet geomagnetic times, whereas during geomagnetic storms a strong high-to-low-latitude circulation component is superimposed by auroral processes. The thermosphere thus has a highly variable motion structure that depends mainly on the level of geomagnetic activity.

The main source of thermospheric heating, and the one that largely explains the characteristic temperature distribution, is absorption of solar ultraviolet radiation. However, energy inputs in the auroral zone are also important, sometimes exceeding the average solar heat input, at least over a limited area of the globe. The distribution of heat input with altitude is important in determining atmospheric behavior, and this is reasonably well known in the case of solar ultraviolet radiation. The magnitude, distribution with altitude, and frequency of occurrence (or time history) of auroral heating all need to be known much better than they are now if the overall behavior of the thermosphere is to be well understood.

The fact that internal gravity waves (excited in the lower atmosphere, especially in mountainous regions) and tides propagate upward into the thermosphere is well known, but the importance of this energy dissipation to the overall heat budget of the upper atmosphere remains rather uncertain. Until a satisfactory evaluation of this energy input can be made, knowledge of the heat budget of the upper atmosphere will remain speculative. Gravity waves may also play a role in carrying energy away from the auroral regions to lower latitudes.

It is clear that large-scale circulations play an important role in the thermosphere. Diurnal (day–night) winds greatly reduce the magnitudes of the diurnal changes that would otherwise occur. On an annual cycle, enhanced helium and atomic oxygen concentrations are produced over the winter polar region by a fractionating effect due to an average wind from the summer to the winter hemisphere. Still another anomaly has been observed in the atomic oxygen concentration: lower values prevail over the equatorial region than in either the summer or the winter hemisphere. The cause of this variation has not been identified, but it is an example of the sort of major large-scale effect that needs explanation, and large-scale circulation may ultimately prove to be a factor. In any case, the general overall features of the circulation of the upper atmosphere need to be understood.

HYDROGEN

The exosphere of the earth's atmosphere is the region of the atmosphere where interatomic collisions are so improbable that the atoms for all practical purposes follow long segments of Keplerian orbits. The exosphere begins at a surface called the exobase or critical level about 500 km above the earth's surface. From this region, atoms traveling upward with sufficient velocity can even escape from the earth on hyperbolic orbits. The rest follow elliptical orbits with perigees below the exobase and hence re-enter the atmosphere. Still other atoms may be in orbits with perigees above the exobase. Atomic hydrogen is the most important gas in this region. In fact, a significant fraction of the hydrogen atoms in the high energy tail of the Boltzmann distribution at the exobase has more than enough energy to escape from the earth, a process known as Jeans escape.

Recent satellite studies have shown that the solar Lyman-alpha radiation pressure significantly affects that part of the exospheric hydrogen distribution in orbits entirely above the exobase. It causes rotation of the lines of apsides of a large class of very eccentric orbits so that their apogees lie away from the sun. Orbits that would have perigees far away from the earth toward the sun are depleted in population because of photoionization.

A diurnal variation in concentration of hydrogen at the exobase is known to exist. The maximum concentration occurs at night, and the minimum on the dayside of the earth. The amplitude of variation is a factor between 1.5 and 3, and the times of maximum and minimum are about 0400 and 1600 hours local time. The reason for this behavior, along with a decrease in the exobase concentration observed to accompany increasing solar activity and exospheric temperature, is to be found in an effect of the escape flux on the concentration in the upper thermosphere. To supply a significant upward flow of hydrogen, the concentration distribution must depart from diffusive equilibrium in the lower thermosphere. It turns out that the concentration must decrease with altitude very rapidly for several tens of kilometers above the turbopause, where eddy and molecular diffusion are equal, in order to produce flows of the order of about 5×10^7 cm^{-2} sec^{-1}. Such thermal escape fluxes must occur when the exospheric temperature is near 1000 K.

Only recently have the factors controlling the supply of escaping hydrogen atoms from the earth been reasonably well understood. The source has long been known to be the production of atoms from photolysis of water vapor below 100 km. However, it has generally been assumed that the source might vary because of changes in factors such as the flux of solar radiation and the transport properties of the lower thermosphere. Until recently, the relationship between

the distribution of hydrogenous species in the stratosphere and mesosphere and the escape flux has not been treated with appropriate attention to the boundary conditions. We now know that the escape flux of hydrogen depends almost entirely on the total mixing ratio of hydrogen atoms in the stratosphere, and it is quite insensitive to all other factors operating above the tropopause as long as the exospheric temperature is greater than about 800 K. Under these conditions any amount of flux supplied to the upper thermosphere from below can escape from the exosphere. The limiting flux is determined by the transport conditions near 100 km. The amount of hydrogen that can pass through this region (much of it in the form of H_2 that is later dissociated near 140 km by energetic oxygen) is set by the amount available in the stratosphere. Only if the exospheric temperature becomes too low (<800 K) does the exobase become the bottleneck and the flow become throttled there.

The amount of escape flux expected from the known total mixing ratio of hydrogen in the stratosphere (14.8 ppm) was 26.8×10^7 cm^{-2} sec^{-1}. However, recent measurements indicate that at 1000 K the Jeans flux is only 5×10^7 cm^{-2} sec^{-1}. This fact has led to the recognition that charge exchange between exospheric hydrogen atoms and energetic hydrogen ions in the exospheric plasma, producing fast neutral atoms, contributes about four times as much escape flux as does thermal escape at 1000 K, while there may also be some additional escape in the polar wind.

One of the most interesting implications of this increase in the estimated hydrogen escape flux by almost an order of magnitude is that it allows for the evolution of ten times as much oxygen in the earth's atmosphere during the time when the production of oxygen was mainly the residue of photodissociation of water vapor followed by hydrogen escape.

The major problems posed by these recent developments are to observe the fast hydrogen atoms produced by charge exchange and to determine their temporal and spatial distribution as well as to develop an adequate theory for the charge-exchange contribution to the escape flux. Observations of the distribution of hydrogenous compounds in the lower thermosphere, the mesosphere, and the stratosphere are also needed. Especially interesting would be an observation of molecular hydrogen near 100 km.

THE IONOSPHERE

The ionosphere can be roughly identified as that part of the atmosphere that is sufficiently ionized to affect the propagation of radio waves in some frequency region. The lowest portion, the D region, will be discussed separately in the next section. Most of the ionosphere is contained within the magnetosphere; only in the lowest portion are collisions with neutrals so frequent as to dominate the motion of the ionized plasma. The ionosphere is important to the field of radio communications, sometimes helping as in the case of propagation by means of ionospheric reflections and sometimes hindering by absorbing or bending the radio rays. The ionosphere is important as a plasma laboratory and also for its role in the flow of matter, energy, and momentum into and out of the neutral atmosphere and into and out of the magnetosphere.

A great many complicated phenomena occur in the ionosphere. Although chemistry is less complicated above the D region, other complications arise relating to the combined effects of magnetic and electric fields and a host of instabilities. Many of the large-scale effects in the ionosphere relate closely to physical phenomena taking place in the outer portions of the magnetosphere.

Many of the small-scale effects are apparently plasma instabilities and relate more to local plasma physics than to magnetospheric physics.

There are many important gaps in our understanding of the ionosphere that can very likely be filled only by resolution of major questions relating to the upper atmosphere and magnetosphere, to which the ionosphere is closely coupled in many important ways. By like token, proposed solutions for problems in these other regions must also pass the test of consistency with (and may at times be suggested by) the ionospheric behavior to which they would lead. A few examples are seasonal patterns of upper atmospheric circulation as they affect composition and concentration; the relative roles of solar photon, energetic particle, and Joule heating in controlling the temporal and global patterns of upper atmospheric heating and circulation; storm-time or other "weather" variations superposed on such longer time-scale patterns; and direct ionization transport effects due to atmospheric drag and magnetospheric electric fields.

There are many other questions that are more purely ionospheric—intrinsic to the understanding of either a magnetized plasma or a partially ionized planetary atmosphere. There is a broad range of both periodic and random irregularities found to exist naturally in the ionospheric plasma concentration, yet we do not know what mechanisms drive them. We know that man-made power levels of radio-wave energy directed into the ionosphere can, under some conditions, drive the plasma unstable; still, we do not have a useful basic knowledge of the processes by which the plasma then converts this radio-wave energy into the forms of charged particle and optical energy. Major questions remain as to the thermal budget of the ionospheric plasma; there is not yet an agreed upon set of heat sources and sinks that can lead to electron temperatures as great as those observed to exist in much of the ionosphere. As yet undetermined chemistry and transport processes appear to be called for to explain quantitatively the gross boundaries of the high-latitude ionospheric concentrations. The significance of excited-state constituents to the chemistry of a partially ionized planetary atmosphere has yet to be determined, although we now know they cannot be neglected.

IONOSPHERIC D REGION

The ionospheric D region is the lowest portion of the ionosphere, lying roughly between 60 and 90 km. It is responsible for the daytime absorption of radio waves in the normal broadcast band that would otherwise propagate to great distances by ionospheric reflection at the E or F region. It is characterized by an exceedingly complicated chemistry involving both positive and negative ions. Galactic cosmic radiation causes most of the ionization in the lower portion of the normal or quiet D region, and ionization of nitric oxide by solar ultraviolet radiation (in particular, the Lyman-alpha radiation from hydrogen) is most important in the upper portion. Below about 80 km, the positive ions are converted to water cluster ions $H^+ (H_2O)_n$ by a complicated series of reactions. The electrons form negative ions, although photodetachment limits the negative ion concentrations during the daytime. The D region is disturbed by a number of phenomena: hard x rays from solar flares, energetic particle radiation from the sun (solar cosmic radiation) at high latitudes, and energetic particles precipitated from the magnetosphere at times of magnetic disturbances. The chemistry associated with the ions formed during these disturbances is different from that for the quiet times; the primary ions that are formed during disturbances are mainly O_2^+ and N_2^+ rather than the NO^+ that characterizes quiet times. Strange-

ly, the sequence of chemical reactions is better understood for disturbed rather than quiet conditions because of uncertainties in the chemistry when the initial ions are NO^+.

The nitric oxide concentration in the D region has not been accurately determined, either by theoretical or observational techniques. Although the chemistry of odd nitrogen (a category that includes NO and some related constituents such as NO_2 and HNO_3) in the E region, where NO is produced, is becoming clearer, major uncertainties exist in the altitude variation of the eddy-diffusion coefficient and in the role of horizontal transport. Measurement techniques, in addition to the rocketborne NO gamma-band resonance fluorescence method that has been used, need to be developed.

The origin, formation, and distribution of metallic ion and neutral species in the D region are not well understood. Neither the production rates nor loss mechanisms of the metal ions are known. An interesting possibility is the hydration of metallic ions, for example $Na^+(H_2O)_n$, which has been observed by *in situ* experiments. The atomic sodium layer in the upper D region offers a wide variety of challenging problems, including the identification of sources and sinks and the role of transport processes.

The outstanding positive-ion problem in the D region is the identification of the chemical scheme for rapid conversion of NO^+ ions to the $H^+(H_2O)_n$ series of ions. This problem arose when it was realized that O_2^+ production from $O_2(^1\Delta)$ photoionization could not compete with NO^+ as a precursor for water cluster ion formation. However, this matter needs to be re-examined because there is now evidence that the NO^+ production rates may have been overestimated by a factor of approximately 10.

There is no paucity of negative-ion problems. Conflicting results have been obtained for the negative-ion composition, and this problem must be resolved. Mass-spectrometer sampling methods must be developed and improved to avoid heavy-ion fragmentation (this applies to both positive and negative ions). The observations of negative ions in the upper daytime D region require an explanation; there are indications that the negative-ion-to-electron density ratio (γ) may be approximately equal to 10 between 70 and 80 km in the daytime, something that is hard to understand in the presence of rapid photodetachment.

The role of particulates in the D region is attracting more attention because of the possibility that they may participate in both the positive- and the negative-ion chemistry. Perhaps they are responsible for the inexplicably fast electron loss rates deduced from eclipse measurements. Furthermore, the subject of ion-induced nucleation, involving the possible growth of $H^+(H_2O)_n$ ions to very large dimensions, is a fruitful area for research.

The interpretation of the winter variabilities of the upper and lower D regions in midlatitudes is a challenging topic that offers a broad range of problems in upper atmospheric dynamics and the chemistry of neutral and ionized species. These variabilities have been attributed to a redistribution of minor neutral constituents, which strongly influence the height distribution of the ionized species. Thus, winter variability is probably a manifestation of dynamical processes peculiar to the winter mesosphere in midlatitudes.

THERMOSPHERIC TURBULENCE

The lower thermosphere has generally been considered to be turbulent at least up to an altitude near 105 km; any turbulence above that level, if it really exists, must be obscured by rapid molecular diffusion, which plays a more important

role than turbulence in the spreading of tracers such as sodium vapor released from rockets. The boundary that indicates the upper level of turbulence, or the level above which molecular diffusion (due to molecular motions) is more rapid than eddy diffusion (due to turbulent eddies of scale sizes less than about a kilometer), has been designated the turbopause.

Recent studies of the lower atmosphere have shown the importance of the interplay between the troposphere and the stratosphere; our limited knowledge of the mesosphere (the region between the stratosphere and the thermosphere) and lower thermosphere indicates even more a complicated interaction there, with the global turbopause playing a vital, if not well understood, role. The interpretation of the dispersion of tracers, both natural and artificial, in the atmosphere in terms of three-dimensional diffusivity and atmospheric heating by turbulent dissipation is still controversial. A detailed knowledge of the global vertical temperature profile between 80 and 120 km is essential to the complete understanding of the generation and maintenance of the turbopause. Instruments capable of temperature measurement with the required resolution are not yet available, but recent advances in rocketborne pitot tubes, incoherent-scatter radar, and satellite radiometry look promising.

Most determinations of turbopause altitude and intensity have been made at middle latitudes. International programs, such as the Middle Atmosphere Program of the Special Committee on Solar-Terrestrial Physics, are emphasizing the necessity for coordinated observations that are truly global in nature, including the tropics and polar regions.

ATMOSPHERIC WAVES

Waves of many sizes are commonplace in the atmosphere, although it is not easy to see or even sense them. The most important waves from the standpoint of atmospheric behavior are influenced by gravity; this statement rules out sound waves, which are pure pressure waves in which gravity is not a factor in the restoring force for the wave. However, as the period of sound waves lengthens, gravity increasingly becomes a factor; for periods in excess of about 5 min, the effect of gravity on the mass distribution provides the restoring force that results in wave propagation. The decrease in atmospheric density with altitude requires that upward propagating waves increase their relative amplitudes in order to conserve energy. Thus waves that are generated with small amplitudes in the lower atmosphere, for example by winds blowing over mountains or by storms, attain large amplitudes in the upper atmosphere. Tides produced by the daily heating of the ozone layer also attain high amplitudes by the time they reach the thermosphere and the ionosphere. Sometimes the waves perturb clouds at very high levels—the noctilucent clouds that occur near 80 km at high latitudes in summer—and produce effects that can be seen visually from the ground, as shown in Figure 3.

Although the concept of internal gravity waves carrying energy upward in the atmosphere is well established, some important questions remain. On the theoretical side, better solutions to the wave-propagation problem are needed for realistic temperature and wind distributions in the atmosphere. On the observational side, more extensive observations are needed to establish the wave amplitudes in the upper atmosphere and the rates and mechanisms of attenuation that result in energy deposition in the upper atmosphere. More knowledge of wave motion, ranging from short-period traveling ionospheric disturbances to tides, is essential to the future development of our understanding of the upper atmosphere.

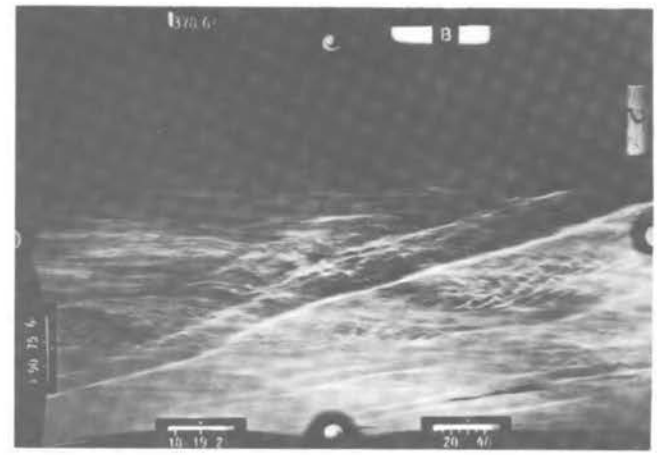


FIGURE 3 Noctilucent clouds are occasionally seen after sunset at high latitudes in summer; they occur at an altitude of about 80 km. This photograph, provided by Georg Witt, shows the complicated wave patterns that exist in the atmosphere, which are made visible in this case by the clouds.

TRANSPORT IN THE STRATOSPHERE

Stratospheric transport is complicated by a broad spectrum of atmospheric waves whose periods range from minutes to days and whose horizontal wavelengths extend from meters to the circumference of the earth. Only the larger of these waves (horizontal wavelengths >1000 km) can be resolved from conventional radiosonde data, and only the horizontal components of their velocities can be measured. The associated vertical velocities, although too small to be measured, cannot be ignored in stratospheric transport. Therefore, any reasonably accurate diagnosis or prediction of stratospheric transport must be based on three-dimensional displacements and three-dimensional trajectories, i.e., they must include derived values of the small but important vertical velocities. At present, these vertical velocities can be derived by a variety of methods, but each involves the solution of one or more partial differential equations and requires rather complicated numerical computation programs. However, with the aid of numerical models and the worldwide network of radiosonde stations (which provide synoptic data every 12 hours), it is now possible to diagnose or to predict stratospheric transport with considerable accuracy and detail.

Problems involving stratospheric photochemistry, particularly the net concentrations of stratospheric ozone, are of concern to the general public, to environmentalists, and to business communities as well as to scientists. The photochemical models used to solve these problems must incorporate transport processes, but there are simply not enough observations of trace gases and transient species to justify using a detailed, three-dimensional transport model.

To arrive at a more reasonable balance between the photochemical and transport modeling, it is convenient to reduce the number of independent variables by averaging the dependent variables over all longitudes and over a month or a season. Then one can use a two-dimensional transport and photochemical model to predict the zonal–seasonal mean concentrations as a function of altitude and latitude. A further reduction to a one-dimensional model, although convenient for photochemical computations, seriously overconstrains the transport by reducing it to just vertical diffusion. Stratospheric transport is not so constrained, therefore errors in the predicted concentrations due to errors in the transport must be expected.

For many problems, the two-dimensional transport model represents a rational compromise. It includes convective transport and diffusive transport caused by correlations between the deviations from the means of the velocity components and the concentrations. To avoid having to predict the deviations

themselves, their effects on the transport are expressed in terms of diffusion coefficients and the gradients (latitudinal and vertical components) of the mean quantity being transported. However, the diffusion coefficients must be components of a tensor because the diffusion is anisotropic, and they must vary with latitude and altitude because the diffusion is inhomogeneous. For example, the diffusive flux in the south-to-north ($+y$) direction depends on the vertical (z) gradient as well as the horizontal (y) gradient of the mean because the vertical and north–south motions are correlated. In fact, the stratospheric flux is often called countergradient because the net horizontal flux toward the north is positive even when concentrations increase toward the north, the diffusion term relating horizontal flux to the vertical concentration gradient dominating the horizontal flux.

In the two-dimensional model, diffusive transport usually opposes convective transport, and the k values (diffusion coefficients) are not independent of the mean velocities used in the model. It is necessary, therefore, to tune the model by analyzing the net transport of a quasi-conservative tracer and adjusting the grid-point values of the mean velocities and of the k's.

At present, the mean velocities are not known with sufficient accuracy to provide an objectively determined set of mean velocities and diffusion coefficients. The fact that the density of radiosonde observations decreases rapidly with height above 25–30 km also increases the uncertainty and the subjectivity in the specified grid-point values. Nevertheless, the two-dimensional models are sufficiently accurate to be useful and, with the current focus on the need for improvement, one can anticipate a trend toward more objective determinations. There is a definite need for synoptic observations of the winds in the upper stratosphere beween 25 and 55 km altitude.

OZONE

Ozone (O_3) is a poisonous, reactive gas, explosive in high concentrations, that is produced from ordinary atmospheric oxygen whenever energy is supplied in sufficiently concentrated form, as by lightning or ultraviolet light. As a direct consequence of its instability it is also an important absorber of solar radiation in the ultraviolet and an important, perhaps essential, shield for many biological systems, as indicated in Figure 4. Thus, the "ozone layer" of maximum concentration in the lower stratosphere is believed to be important to human health and economic welfare. Because of its reactivity, ozone is readily destroyed by a variety of natural and artificial agents.

Concern about damaging the earth's ozone layer arose first in connection with the supersonic transport. Although the initial concern was that water vapor from the exhaust might produce detrimental effects, the more important role of nitrogen oxides was soon recognized [*Statement of the* NRC *Ad Hoc Panel on (NO_x) and the Ozone Layer*, 1971]. After about two years of intensive investigation, the matter was described in considerable detail in 1975 in the report of the NRC's Climatic Impact Committee, *Environmental Impact of Stratospheric Flight*. The important role that might be played by chlorine was also recognized, and this is described in two further reports published by the National Academy of Sciences, *Halocarbons: Effects on Stratospheric Ozone* and *Halocarbons: Environmental Effects of Chlorofluorocarbon Release*, both issued in 1976.

Ozone is also important to the thermal balance of the stratosphere, and the possibility exists that changes of ozone could affect climate at the ground.

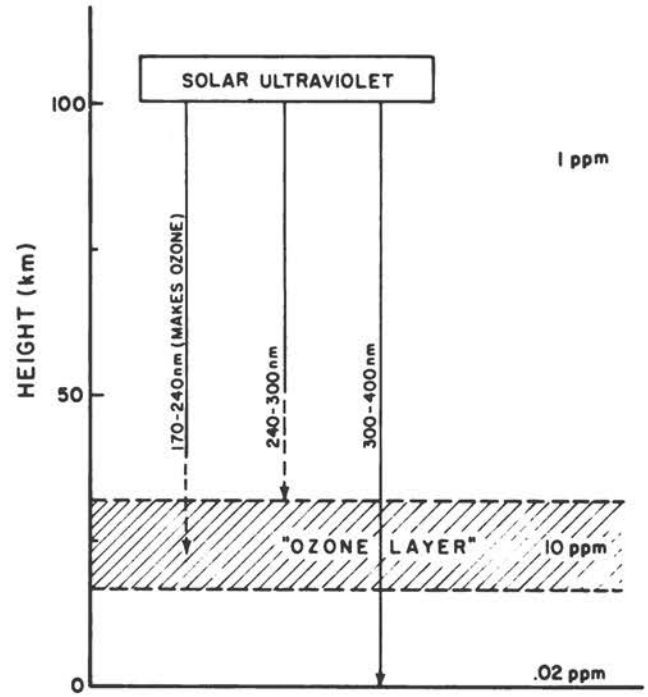


FIGURE 4 Sketch of the ozone distribution in the atmosphere up to 100 km, shown in terms of the volume mixing ratio in parts per million (ppm). Also shown are typical depths of penetration of various parts of the solar ultraviolet spectrum.

The small ozone concentration shown in Figure 4 for surface regions is greatly augmented in some kinds of urban pollution. Here ozone's chemical reactivity causes deterioration of rubber objects, the leaves of plants, and lungs.

In contrast to these concerns about human welfare, the study of ozone above 50 km is still a matter of scientific curiosity, part of the broad picture of the upper atmosphere. The same was true of the ozone layer itself for most of the century that it has been studied for its intrinsic interest. Our basic knowledge is still dangerously inadequate for today's concerns; however, past studies have provided a basis for judging the possible severity of the problem of possible ozone depletion, and without these it would not have been possible to suggest preventive measures.

Although the behavior of ozone in the stratosphere has been studied for many years, it is complicated by factors that are only beginning to be understood. Both meteorology and chemistry are involved. Because of meteorological phenomena, there is more ozone at high latitudes than at low, even though most of it is produced at the lower latitudes where there is more sunlight. Destruction of ozone is believed to be due primarily to a by-product of anaerobic decay in swamps and soil (nitrous oxide, which eventually finds its way to the stratosphere, where it is converted to a catalyst that destroys ozone). Mathematical models of these processes give an excellent account of the distribution and quantity of ozone. This good fit increases our confidence in the validity of our knowledge, but predictions of the effect of an environmental change are still risky. This problem is particularly severe in the realm of biological systems.

The biggest natural destruction mechanism for ozone is catalytic destruction by nitric oxide and nitrogen dioxide, known collectively as NO_x. Soil bacteria produce N_2O, which is converted in the stratosphere to NO_x. The natural N_2O budget is not well understood. Is there a major tropospheric or oceanic sink? What would the response be to a major increase of soil nitrogen or a change in

acidity? Could the lower-atmosphere/ocean system buffer such a change? Can increased use of chemical fertilizers have a serious detrimental effect on the ozone layer?

The effect of ozone changes on crops and ecosystems through the associated change in ultraviolet radiation is potentially more serious than the skin-cancer effect. We need to know much more about both damage and repair processes in biological systems.

We seem to have a good description of ozone chemistry in the stratosphere, although major questions about the troposphere remain unanswered. That description needs to be refined and tested by measurement of the variation with height, time of day, season, and latitude of key chemical species in the stratosphere. Such tests, if successful, will increase our confidence and point up needed improvements if discrepancies are found.

Meteorology, or transport, of ozone and of pollutants needs a much better description than that available at present. For some problems the available methods and data give usable operational predictions, although they are intellectually unsatisfying. Such predictions are often mere global averages; where geographical resolution is needed, the predictions are inadequate and their extension in this direction is on shaky ground. Problems of aircraft exhaust just above the tropopause are particularly intractable and will not be solved until we know much more about the behavior of the atmosphere.

INTERNATIONAL COOPERATION

In many geophysical problems, there is a greater need for international cooperation than in most other fields of science because coordination is required in the making of observations in different parts of the geophysical system that are interrelated by physical processes. Programs are usually arranged through the relevant International Scientific Unions, their parent body, the International Council of Scientific Unions (ICSU), or special committees of ICSU.*

THE INTERNATIONAL GEOPHYSICAL YEAR (IGY)

The IGY (1957–1958) in a sense launched the Space Age; it provided the incentive or the opportunity for the first launching of scientific satellites. From the point of view of upper atmospheric physics (magnetospheric physics was virtually nonexistent at that time), the IGY was more the beginning of a new era than a period of intensive observation; this was because of the tremendous increase in observing capability that resulted from the availability of satellite-based observations—a capability that did not substantially materialize during the IGY. The first dramatic discovery of the satellite era was the Van Allen radiation belts, and this is properly associated with the IGY. Even though developments in the post-IGY years overshadowed those in the IGY period

*The International Union of Geodesy and Geophysics (IUGG) is the Union most concerned with the upper atmosphere and magnetosphere, in particular two of its eight Associations, the International Association of Geomagnetism and Aeronomy (IAGA) and the International Association of Meteorology and Atmospheric Physics (IAMAP). Other Unions of interest in this connection include the International Union of Radio Science (URSI), and the International Union of Pure and Applied Physics (IUPAP). Two interunion committees organized under ICSU are of special interest—the Committee on Space Research (COSPAR) and the Special Committee on Solar-Terrestrial Physics (SCOSTEP).

itself as far as upper atmospheric physics is concerned, the IGY must be regarded as an important milepost in international cooperation. The IGY was timed to occur during a maximum in the solar cycle, a maximum that turned out to be the largest on record.

THE INTERNATIONAL YEARS OF THE QUIET SUN (IQSY)

The IQSY followed the IGY. It was a program designed to make use during a period of low solar activity of the observing capability developed during the IGY, especially satellite-based observations with their unprecedented capability for gathering data on a worldwide basis with a single set of instruments. The overall patterns of density and temperature in the upper atmosphere were determined during the IQSY, the global description of the ionosphere was improved, and the concept of magnetosphere became clearly established.

In both the IGY and the IQSY, there was relatively little coordination of separate programs compared with what is needed and can be accomplished today; the coordination was mainly through description of key days on which observational attention was concentrated.

THE INTERNATIONAL MAGNETOSPHERIC STUDY (IMS)

The IMS (1976–1979) is an international, interdisciplinary, cooperative enterprise whose chief objective is a comprehensive, quantitative understanding of the structure of the plasma-and-field environment of the earth and the dynamical processes operating in it, including its interaction with its external environment (the interplanetary medium, in particular the solar wind) and its internal environment (the atmosphere). The operational basis for this enterprise is a plan for the coordination of observations made from appropriate spacecraft with each other and also with observations made from the earth's surface or near it. The basic plan for coordination was developed in 1970–1972.

There is a dual aspect to the scientific goals of the IMS. One aspect is a quantitative understanding of the earth's magnetosphere and its dynamics just because the magnetosphere—like the atmosphere, the oceans, or the earth's interior—is an important feature of man's physical environment; the other aspect is the study of basic plasma physics, in which the magnetosphere plays the part of a huge natural plasma laboratory with densities, temperatures, field strengths, and especially dimensions not realizable in terrestrial laboratories. The scientific program for the IMS reflects this duality by first listing about a dozen specific sample questions about the structure and dynamics of the magnetosphere and then pointing out that the answer to these specific questions lies in an understanding of a considerable number of plasma-physical processes. These processes are grouped under several headings, which are intended to be illustrations rather than an exhaustive list: the maintenance of regions of abrupt reversals of the magnetic field; the maintenance of plasma convection; wave–particle interactions; propagation of waves in an inhomogeneous anisotropic medium; macroscopic plasma instabilities; instabilities driven by electric currents; and stochastic transport processes.

The following human and technical resources are involved in the IMS: thousands of scientists and technicians from more than 40 countries; more than

40 spacecraft of about half a dozen countries and the European Space Agency (ESA) with a wide variety of orbits [high-altitude, low-inclination circular; geostationary; low-altitude, high-inclination (polar) circular; high-altitude eccentric; interplanetary]; some 1000 individual or institutional programs, including theoretical and analytical studies and also including some hundred or more rocket and balloon campaigns; and some 700 ground-based stations from the extreme Arctic to the South Pole. About 200 kinds of instrumental techniques are used to measure various aspects of the wide variety of electromagnetic and particle phenomena involved.

Most of these resources were already in existence or planned independently of the IMS. Some countries and ESA, however, have added some programs especially for the IMS, to the benefit of scientists everywhere. For example, some 8 to 10 spacecraft dedicated to the IMS will have been launched during the IMS interval, and nets of appropriately instrumented geophysical observation stations have been or will be enlarged and modernized. In other words, the original and still basic rationale of the IMS is to maximize the scientific return of existing resources by coordinating the taking of data simultaneously at specific times and places, determined at least in part by the location of suitably instrumented spacecraft with respect to each other or to observational facilities on or near the surface, instead of leaving such coincidences to chance; but the IMS Steering Committee has also encouraged the development of facilities especially for the IMS where this possibility existed.

The international coordination of data-taking schedules to achieve the goals of the IMS far exceeds in complexity anything that has been tried before. The coordinating mechanisms include the following: (1) The Satellite Situation Center (SSC, located at the NASA Goddard Space Flight Center), which publishes orbital predictions in simplified graphical form for the principal IMS-related spacecraft, provides other predictions on request, and scans the predictions for scientifically interesting spacecraft configurations. (2) The IMS Central Information Exchange (IMSCIE) Office at Boulder with regional or national branches at the Meudon Observatory (for Western Europe) and in Japan, the Soviet Union, and other locations, which channel up-to-date information or plans, campaign schedules, recently completed observations, and other data to the IMSCIE Office. The IMSCIE Office publishes a monthly newsletter containing this information and also a calendar of selected special observational intervals (based on predicted spacecraft configurations or positions, major campaigns, and other data) and operates a referral service for special requests. (3) The computer-based IMS file (at the SSC), a directory of individual and institutional programs, the persons involved, their interests and objectives, measurements they are making, locations, timetables, and other data. (4) The existing IUWDS telex net for alerts and other quick transmissions. The objective of these mechanisms is to make available as much information as possible to the worldwide IMS community on a timely basis, so they can make their own decisions about detailed coordination with others. (5) National Committees and Coordinators for the IMS.

Arrangements for putting many classes of data at the disposal of scientific users everywhere either already exist (the World Data Center system) or have been made especially for the IMS (for example, spacecraft data for selected intervals, and continuous data samples from several ESA and NASA IMS-dedicated spacecraft, to be made available on data-pool tapes or in graphical displays). The IMS also takes advantage of the system for the real-time collection and distribution of data from several spacecraft and ground-station networks operated by NOAA's Space Environment Laboratory.

THE MIDDLE ATMOSPHERE PROGRAM (MAP)

The purpose of the MAP is to develop an adequate description and physical understanding of the atmosphere in the altitude range corresponding to the stratosphere and mesosphere (the "middle atmosphere," or from roughly 15 to 100 km), particularly with regard to the global fields of (1) density, pressure, and temperature; (2) compositions (including trace elements and aerosols); (3) radiation; (4) motions (on all scales); and (5) the interaction among these fields.

The MAP will require observations that are both intensive and extensive: intensive, so that interactions can be studied and defined; extensive, so that the global picture will be complete. The plan envisages a coordinated attack by all available techniques: surface stations, aircraft, balloons, rockets, and spacecraft. Theoretical models that adequately describe both the dynamical and aeronomical aspects of the middle atmosphere will be thoroughly explored. Observations and theory must also take into account interactions across the interface of the middle atmosphere with the troposphere below and the thermosphere above. All relevant physical and chemical processes will be considered in the light of significant space and time scales, although time scales appropriate to climatic change will almost certainly require longer-term observations than are planned for the MAP.

The operational stage of the MAP is expected to begin in about 1979 and continue through about 1985 or possibly longer, to take advantage of the opportunities offered by the NASA Space Shuttle and other future spacecraft.

The global distribution of temperatures, pressure, and density as a function of height and season and the influence of radiation and motion on their distribution should be determined in enough detail to account for such regional aspects as standing features of the pressure distribution. Atmospheric composition (including minor and trace species, gas and aerosols, excited species, ions, and electrons), the related photochemistry in the presence of radiation, and the effects of transport inside the middle atmosphere and across its boundaries need determination. Radiation of all kinds relevant to the heating and cooling of the middle atmosphere (solar photon fluxes, their variation, absorption, and scattering, and photochemical effects; the emission and absorption of terrestrial photon fluxes; particle fluxes of astronomical, solar, and magnetospheric origin and their chemical, dynamical, and heating effects) and its variation with latitude and time, and its relation to the establishment of or departures from local thermodynamic equilibrium need a comparison of observation and theory. A knowledge of motions and their implications for momentum and energy balance and for vertical and horizontal transport is extremely important. The space scales range from global circulation and its seasonal variation through large-scale planetary and gravity waves, tidal motions, semiannual to quasi-biennial oscillations, to features more localized in space or time, like orographic or tropospheric effects and sudden warmings, and to turbulent and eddy diffusion and finally molecular diffusion at the small end of the scale. A number of special topics (effects of natural and man-made pollutants and of changes in solar radiation, the characteristics of the middle atmosphere with implications for aviation and communications) are also included.

The MAP Planning Group draws its members from among experts affiliated with the international organizations interested in the middle atmosphere. These organizations have already approved the basic idea and have expressed an interest in participating. In addition to consulting with their colleagues, the MAP

Planning Group held a workshop in June 1976 attended by some 80 experts from all over the world. The resulting planning document describes the MAP's scientific goals and problems, including summaries of existing knowledge and techniques and types of instruments suitable for a coordinated attack on those problems. Like the IMS, the MAP will depend heavily on satellites for many types of measurements that cannot be done in any other way, e.g., remote sensing of some of the important variables on a global basis, repeated at frequent intervals, and monitoring of solar and other radiation at wavelengths inaccessible from low altitude.

Solar-Wind Interactions

1

THOMAS W. HILL
Environmental Research Laboratories, National Oceanic and Atmospheric Administration

RICHARD A. WOLF
Rice University

1.1 PROLOGUE

The outer limit of man's immediate environment in space is the magnetosphere, the sphere of influence of the earth's magnetic field. The geomagnetic field near the earth is basically a dipole field (like that of a short bar magnet), generated by some giant electromagnetic dynamo in the earth's liquid-metal core. The magnetic-field lines emanating from this subterranean electromagnet extend through the earth's surface and atmosphere and out into space to a distance of about 10 earth radii toward the sun, much farther away from it. The magnetosphere is not a geometrical sphere but is shaped more like a comet with the earth at its nucleus. At great distances, the earth's magnetic field is terminated by its interaction with the solar wind, which is the outward expanding corona (outer atmosphere) of the sun. The solar wind extends, although invisibly, past the earth's orbit and beyond, to the outermost reaches of the solar system.

The magnetosphere is, among other things, a buffer zone that shields the earth against potentially harmful charged-particle radiation from the sun, just as the atmosphere shields us against the harmful portions of the sun's electromagnetic radiation. The solar-wind–magnetosphere interaction produces a wide variety of geophysical phenomena that are observable at the earth's surface. The only visible manifestation of this interaction is the aurora, which frequently lights up the night sky at high latitudes with magnificent displays of color in motion. With suitable instruments, however, one can observe a number of other surface phenomena associated with the solar-wind–magnetosphere interaction, some of which have subtle but important effects on man's environment and technology.

It was known in the nineteenth century that magnetic variations on the earth's surface were associated with sunspots—visible disturbances on the sun's surface. In the first half of the twentieth century, these correlations attracted the attention of several scientists, who postulated the existence of a solar wind, in one form or another, many years before it was observed by early scientific

space probes. We will not attempt to summarize the fascinating history of early geomagnetic and solar-wind research—the interested reader is referred to the book by Chapman and Bartels (1940) and the review by Dessler (1967).

Only through spacecraft observations did it become possible to sort out the maze of untested and often conflicting ideas and to build a coherent, if still incomplete, theoretical picture of the solar-wind–magnetosphere interaction. In this chapter we attempt to provide a general overview of what has been learned from the first 15 years of combined observational and theoretical study and to focus on critical unresolved questions that one can reasonably hope to answer in the next decade.

1.2 SOLAR-WIND FLOW AROUND THE MAGNETOSPHERE

THE SOLAR WIND

The continual explosion or expansion of the solar corona generates a supersonically flowing plasma (ionized gas), which is continuously emitted into the solar system as a solar wind (see review by Dessler, 1967). This wind is observed to be quite variable, but its average properties, as measured near the earth's orbit, can be characterized as follows: composition—fully ionized gases, principally hydrogen, some helium, traces of heavier elements; number density—about 5 ions cm^{-3} (about 10^{-19} times the surface number density of the earth's atmosphere, or 10^{-5} times the lowest density practically obtainable in laboratory vacuum chambers); temperature—about 10^5 K (corresponding to an average thermal energy of about 10 eV); flow velocity—about 400 km/sec (corresponding to an ordered energy per proton of about 800 eV); magnetic field—about $5\gamma = 5 \times 10^{-5}$ G (10^{-4} times the earth's surface magnetic field). The flow speed is about eight times the local sound speed, i.e., a Mach-8 supersonic flow. The flow direction is usually observed to be radially outward from the sun (within a few degrees). The interplanetary magnetic-field lines in the ecliptic plane, on the average, follow a spiral pattern controlled by the sun's rotation; locally the magnetic field in the solar wind is highly variable and may be observed briefly in any direction (Brandt, 1970).

MAGNETOPAUSE

The energy density of the magnetic field at the earth's surface, or its ability to withstand pressure, is about a million times the energy density of the solar wind. The energy density in the earth's dipole field varies inversely as the sixth power of distance from the earth's center. Thus the magnetosphere—the region of space above the earth's atmosphere where the energy density is dominated by the earth's magnetic field—must extend to a distance of the order of $(1{,}000{,}000)^{1/6}$ earth radii (R_e), or about 10 R_e. Consequently, the magnetopause—the boundary between the solar wind and the magnetosphere—occurs sunward of the earth at an average distance of approximately 10 R_e.

To give the reader a rough idea of the physical nature of the magnetopause, we show in Figure 1.1 a simple model of a time-independent plane boundary between an unmagnetized solar wind and a vacuum geomagnetic field as proposed by S. Chapman and V. C. A. Ferraro more than 20 years before the space age. Since a fully ionized plasma is generally an excellent conductor, they assumed that the solar wind had no large-scale electric field. In their model, electrons or protons coming from the sun on straight-line trajectories are turned in opposite directions by the boundary-region magnetic field and are shot back into the solar wind. (As discussed in Appendix 1.A, a magnetic field generally deflects particles perpendicularly to their direction of motion.) No electrical current exists in the solar-wind region of Figure 1.1, because protons and electrons move together there. In the boundary layer, however, the magnetic field deflects the positive ions and the negative electrons in opposite directions, causing an electrical current in the boundary. This current, which is leftward in the figure, produces a magnetic field in the solar-wind region that is equal and opposite to the earth's dipole field at that location. Thus the two magnetic fields cancel each other outside the magnetopause.

When the solar wind was finally observed by spacecraft 30 years after this original theoretical work, it was found to contain a significant magnetic field; therefore, the Chapman-Ferraro model is not strictly applicable. However, its essential conclusion remains valid: the solar wind cannot easily penetrate deep into the magnetosphere, providing that the solar wind and outer magnetosphere can still be considered perfect conductors. The

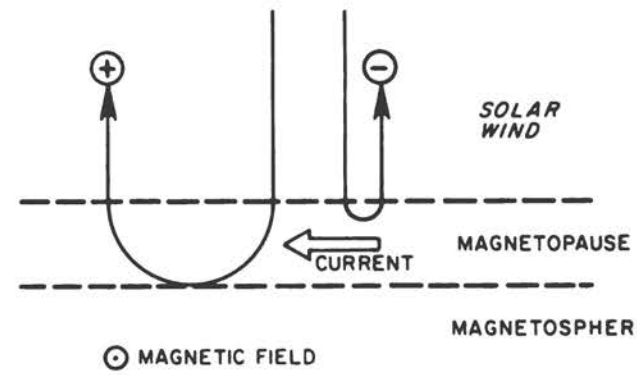


FIGURE 1.1 A Chapman-Ferraro-type picture of a small section of the magnetopause, the boundary of the magnetosphere. Electron and positive-ion trajectories are shown for the simplest case of perpendicular incidence on the plane boundary layer. The solar wind is assumed to have no electric or magnetic fields. The magnetic field in the magnetosphere is directed out of the page; in the magnetopause boundary layer, its magnitude drops smoothly from its magnetospheric value inside to zero outside.

"frozen-in flux theorem" (see Appendix 1.A) states that, in a perfectly conducting plasma, two plasma elements that initially share the same magnetic-field line forever lie on the same field line. This theorem implies that the magnetopause is impenetrable to the flow of the solar wind, assuming perfect conductivity. Referring to Figure 1.2, which represents a portion of the magnetosphere and the surrounding solar wind, the interplanetary field line through point A extends out to great distances in the solar system. It is inconceivable that *all* of its plasma particles could flow onto a closed geomagnetic field line (like the one through point B); therefore, none of its particles could do so if the conductivity were perfect.

In the infinite-conductivity limit, we are thus led to the magnetospheric picture shown in Figure 1.2—a closed magnetosphere that presents to the solar wind an impenetrable obstacle. The ram pressure of the solar wind compresses the geomagnetic field on the dayside and the partial vacuum in the solar-wind wake region allows the geomagnetic field to expand on the nightside.

The turning radius or the thickness of the magnetopause boundary in this model is typically about 100 km (one ion "gyroradius"—see Appendix 1.A). It is difficult for a single satellite to measure the thickness of a boundary like the magnetopause, because the boundary itself moves with an unknown velocity that may well exceed the velocity of the satellite in its orbit. However, most observed crossings of the dayside magnetopause indicate a sharp magnetic boundary and a nearly coincident particle boundary, in agreement with the impenetrable-magnetopause model. Although more detailed investigation indicates that the magnetopause actually leaks slightly (with important consequences), the impenetrable-magnetopause model is a useful approximation.

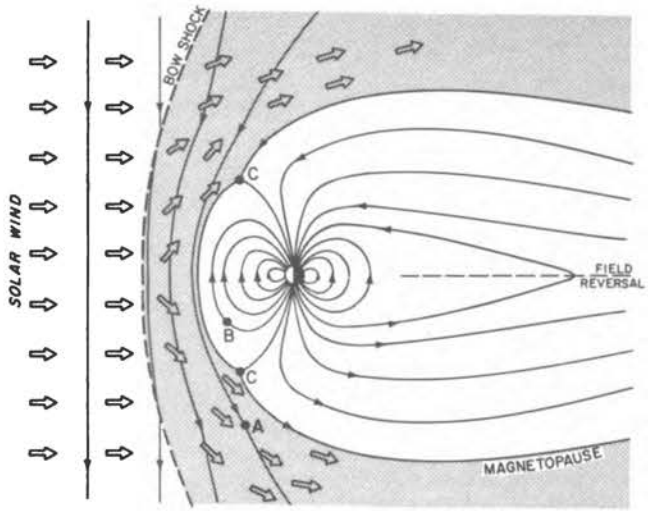


FIGURE 1.2 Picture of a closed, impenetrable magnetosphere as viewed in the noon–midnight meridian plane. North is up and the sun is to the left. Magnetic-field lines are depicted by solid lines with small arrowheads indicating field direction. Broad open arrows indicate plasma flow directions. The solar-wind region is shown dotted, as is the magnetosheath, which consists of solar-wind plasma that has been compressed, heated, and slowed down by the bow shock. The unshaded region is the magnetosphere. Points C are "neutral points," where the magnetic field goes to zero.

AERODYNAMIC ANALOGIES

The observation of a highly supersonic solar wind and of a well-defined magnetopause prompted theoretical suggestions that there might be a shock wave standing in front of the magnetosphere, since, in ordinary gases, a supersonic flow past an impenetrable obstacle usually results in such a shock, called a bow shock. Whether or not one should expect such a shock wave was unclear theoretically because of the difficulty involved in applying the aerodynamic analogy to the flow of the solar wind. The thickness of a gasdynamic shock is usually determined by the distance that particles can travel between collisions. The solar wind is so tenuous that the mean distance between particle–particle collisions is of the order of 10^4 R_e, far too large to be physically significant for interaction between the solar wind and the magnetosphere. Thus one might expect a huge compression region in front of the earth, rather than a well-defined shock wave. A counterargument was that the interplanetary magnetic field, "frozen in" the plasma, limits free particle motion perpendicular to the field (see Appendix 1.A) and makes the collisionless plasma act like a fluid, at least with regard to motion perpendicular to the field. Observations resolved this controversy: a thin, well-defined bow shock is usually observed upstream of the magnetopause, although, as might be expected, the shock is less well defined when the flow is parallel to the magnetic field than when it is perpendicular (see Section 1.3).

Between the bow shock and the magnetopause lies a region of compressed and heated solar-wind plasma called the magnetosheath. Although the magnetosheath is more turbulent than the solar wind, its measured flow velocities, temperatures, densities, magnetic fields, and other parameters are explained remarkably well by straightforward aerodynamic-flow calculations. These computations, which neglect magnetic forces in the solar wind and magnetosheath, are similar to those used in studies of the aerodynamics of vehicles re-entering the earth's atmosphere (Spreiter *et al.*, 1968). In these calculations, the shape of the magnetopause is determined by requiring that the magnetic pressure just inside the magnetopause, as determined from a magnetic-field model, balance the gas pressure outside; magnetopause shapes so calculated are in good agreement with observations made near the earth's equatorial plane. When the ram pressure of the solar wind increases, the magnetopause moves closer to the earth, where the geomagnetic field is sufficiently intense to withstand the increased force of the wind; a decrease in the solar-wind ram pressure likewise leads to expansion of the magnetosphere. The

aerodynamic calculations agree satisfactorily with observations in this regard, as they do also in regard to the shape and location of the bow shock.

The model of frictionless aerodynamic flow around an impenetrable obstacle fails, however, in dealing with the cometlike "tail" of the magnetosphere. Such a model predicts a relatively short, teardrop-shaped magnetosphere. Observations now indicate that the tail extends far beyond the orbit of the moon (at about 60 R_e distance). Taillike plasma and magnetic-field configurations have been seen occasionally more than 1000 R_e behind the earth, but spacecraft coverage of the region beyond 60 R_e is extremely sparse. It is clear that the earth's magnetic tail is long, but the question of how long it is and the physics that determines its length remain obscure.

OPEN AND CLOSED MODELS

The existence of the long magnetospheric tail is evidence of a transfer of momentum across the magnetopause to balance the tension in the stretched-out "slingshot" magnetic-field configuration. This and other evidence to be presented in Chapter 2 and also later in this chapter indicate that something like 0.1 percent of the mass and energy incident on the front of the magnetosphere, and of the order of 10 percent of the incident solar-wind electric field, leak through the magnetopause and cause a variety of dynamical phenomena in the magnetosphere. There are two basic models that have been used to describe this leakage through the magnetopause: the open and closed models. For more than a decade, the controversy between the open and closed pictures has dominated discussions of the interaction of the solar wind with the magnetosphere. The key question is this: does most of the transfer of particles, momentum, and energy result from interconnection of interplanetary and geomagnetic-field lines? In the open model, the answer is yes; in the closed model, interconnection either does not occur or is quantitatively unimportant.

Figure 1.2 shows a closed-model magnetosphere; no magnetic-field line that goes through the earth extends out into the solar wind. Figure 1.3 shows a portion of a corresponding open magnetosphere, with a few field lines from the earth's northern and southern polar caps threading out into the nearby magnetosheath.

It might seem strange that it is so important to know the exact shape of these abstract entities called "magnetic-field lines," but they are physically crucial because of the control that the magnetic field exercises over particle motions. As discussed more fully in Appendix 1.A, charged particles can move quite freely and independently *along* magnetic-field lines but have great difficulty moving independently *across* field lines; they are essentially "frozen onto" the field lines.

The controversy between the open and closed models is full of subtle questions, partly because both theory and observation are confusingly incomplete. At this time, the bulk of the circumstantial evidence favors the open model, and we shall discuss some of this evidence. But no one has yet found a "smoking gun," a conclusive piece of evidence to show that most polar-cap field lines connect to the interplanetary medium within a distance of a few hundred R_e of the earth or that substantial amounts of mass, momentum, and energy are transmitted from the solar wind to the magnetosphere along interconnected field lines. Such conclusive evidence may not be found until active experiments become possible involving large-scale release and observation of trace ions using, for example, the AMPS (Atmosphere, Magnetosphere and Plasmas in Space) laboratory aboard the Space Shuttle.

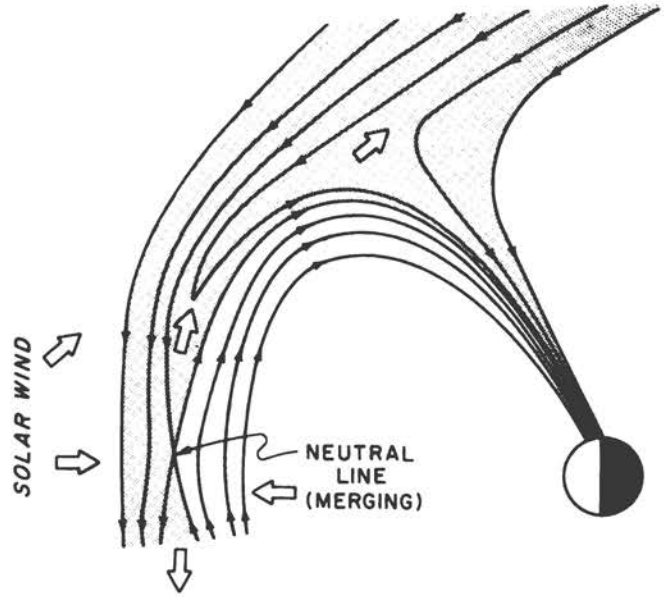


FIGURE 1.3 Magnetic-field topology in the interconnection region of an open magnetosphere. Open arrows indicate plasma flow directions. The view is in the noon–midnight meridian plane, with the sun to the left.

THEORETICAL DIFFICULTIES

The solar wind and the magnetosphere are examples of magnetized plasmas, i.e., ionized gases in the presence of magnetic fields. The dynamics of magnetized plasmas necessarily involves a combination of gasdynamics and electrodynamics, and the combination is often complicated. For one thing, charged particles can move much more freely along the magnetic field than perpendicular to it, and consequently plasma properties such as electrical conductivity are highly anisotropic, as discussed in Chapter 2. Secondly, a magnetized plasma is capable of supporting an astonishing variety of waves that can interact with one another and with the particles of the plasma. A third difficulty that is crucial in the solar-wind–magnetosphere interaction is that of self-consistency be-

tween the electromagnetic field and its sources. The electric and magnetic fields in a plasma are determined by local electrical charges and currents in the plasma as well as by sources outside the plasma. These charges and currents are produced by plasma motions that are, in turn, determined at least in part by the forces associated with the electric and magnetic fields in the plasma. Thus it is not always possible to assume either the fields or their sources as "given" and to derive one from the other; one must derive the fields and their sources simultaneously and self-consistently. Sometimes this self-consistency is simple to visualize; for example, in the solar wind where plasma energy far exceeds electromagnetic energy, one can ignore the electromagnetic field as a first approximation; or in the inner radiation belt (see Chapter 2), where the density of particle energy is negligible, one can ignore currents in the plasma as a first approximation. In the solar-wind–magnetosphere interaction region, however, the energy densities of the plasma and of the field are necessarily of comparable magnitude, and the self-consistency of fields and sources must be explicitly enforced as a part of any theory.

Considering these and other difficulties, it is not surprising that the study of solar-wind–magnetosphere interactions has produced diverse and often conflicting theoretical ideas. The number of theoretical possibilities has been greatly reduced by spacecraft observations, but there remain a number of fundamental questions of collisionless plasma physics to be answered by further observational and theoretical investigation of the solar-wind–magnetosphere interaction.

1.3 THE COLLISIONLESS BOW SHOCK

The earth's bow shock is a natural specialized laboratory for the observational study of collisionless shock waves, a topic of considerable interest in contemporary plasma physics. The details of the bow shock's complex structure may not have substantial effects on the earth's magnetosphere (and thus on the ionosphere and upper atmosphere). They may, however, provide important clues to the understanding of collisionless shock waves in general, clues unavailable in terrestrial laboratories because of the different range of parameters involved.

Theorists expect that any sharp, well-defined shock wave in a magnetized plasma should obey certain "jump conditions" based only on the conservation of mass, momentum, energy, and magnetic flux, along with the assumption of perfect electrical conductivity outside the shock. To within observational uncertainties, the earth's bow shock seems to obey these conditions, which do not depend on the details of the energy-transfer mechanisms operating within the shock.

There are a number of theories for the detailed structure of the shock, involving heat transfer by large-amplitude plasma waves of various types. None of these theories seems consistent with the observations. It is particularly difficult to explain theoretically the great amount of ion heating that occurs (see review by Greenstadt and Fredericks, 1974).

One impressive simplicity that emerges from the data is that when the direction of the solar-wind magnetic field is within about 40° of the shock's surface, the shock is thin and well defined; but when the solar-wind magnetic field is more nearly perpendicular to the shock wave, the shock structure is diffuse, turbulent, and ill-defined. It is then called a "pulsation shock." This simple observational fact remains unexplained by any quantitative theory, and there is, in fact, no such theory for a pulsation shock.

Various kinds of plasma waves have been observed in the bow shock, most of them predicted by one theoretical model or another. However, the waves are not always present, and there is no coherent observational picture of bow-shock structure. Part of the difficulty is caused by the erratic movements of the shock back and forth past an observing satellite, creating great difficulty in separating spatial and temporal variations. One never knows exactly where the satellite is relative to the moving shock structure. This basic problem, which has always plagued satellite observations of boundary structures, may be largely eliminated during the International Magnetospheric Study (1976–1979) because of the launch of satellites ISEE-A, -B, and -C; two of these satellites, called "Mother" and "Daughter," will fly in close formation through the bow shock and other boundaries, recording simultaneous data from two points and thus it is hoped allowing separation of spatial and temporal variations; the third satellite will remain upstream in the undisturbed solar wind, constantly monitoring conditions there—the "input conditions" whose fluctuations affect the bow-shock structure as well as influence the dynamics of the earth's magnetosphere.

1.4 MOMENTUM AND ENERGY TRANSFER

The central goal of the study of the solar-wind-magnetosphere interaction is to identify the mechanisms by which particles, momentum, and energy are transferred across the magnetopause from the solar wind into the magnetosphere. It is by considering these three closely related problems that one can best appreciate the important distinction between the closed and open models of the magnetosphere.

GENERAL REQUIREMENTS

The problem of particle transfer is more easily related to *in situ* observations, and it will be discussed separately below. The associated problems of momentum and energy transfer are more elusive in terms of observational constraints, but one can still make some useful statements about them.

For example, it is known that the solar wind causes magnetospheric convection, a persistent circulation of plasma within the magnetosphere and ionosphere. The solar wind must somehow inject enough energy into the magnetosphere to sustain this convection system against dissipative losses in the ionosphere, as discussed in Chapter 2. The power required to drive the steady-state magnetospheric convection circuit has been estimated to be a few times 10^{10} W. (The dissipation rate may briefly increase to 10^{12} W during substorms, as discussed in Section 1.7 and in Chapter 2.) If the energy is transported into the magnetosphere in the form of solar-wind protons with velocities of about 400 km/sec, the required rate of entry is roughly 10^{26} protons/sec. This required injection rate is close to other independent estimates that are based strictly on particle observations (see Section 1.6), and this "coincidence" suggests that perhaps the particle injection process, whatever it is, might be the dominant energy-transfer process as well.

The solar wind must also transfer enough momentum into the magnetosphere to maintain the long tail. The momentum-transfer rate required to accomplish this is probably not enormous, but it is hard to compute reliably because it depends on details of the field configuration that are not well established observationally. A more definitive requirement is that the momentum-transfer process must ultimately produce the observed magnitude and pattern of magnetospheric electric fields associated with convection, as discussed below.

CLOSED-MODEL TRANSFER PROCESSES

For a completely closed magnetosphere like the one illustrated in Figure 1.2, it is easy to make precise definitions of the terms "magnetopause" and "magnetosphere." The region occupied by field lines that go through the earth is the magnetosphere, and its boundary is the magnetopause. Magnetic-field lines do not cross the closed-model magnetopause.

It was suggested in the early 1960's that long-wavelength plasma waves ("hydromagnetic waves") propagating from the solar wind across the magnetopause could be the major mechanism for transferring energy and momentum into the magnetosphere. Wave propagation provided an attractive transfer mechanism within the closed model because it required no leakage of plasma particles across the magnetopause, i. e., no violation of the idea of frozen-in flux. However, satellite observations have shown that such long-wavelength waves are usually reflected at the magnetopause rather than transmitted; their amplitudes are generally observed to be small in the magnetosphere, although possibly quite large in the magnetosheath. Thus hydromagnetic waves now seem unlikely to be the dominant transfer mechanism, although the situation has never been fully resolved.

Thus we are led back to the conclusion that the energy- and momentum-transfer processes probably involve particle transfer as well. In the closed model there are two proposed mechanisms for injecting solar-wind particles into the magnetosphere: (1) particle drifts caused by gradients or curvature in the magnetic field (see Appendix 1.A) and (2) diffusion caused by wave–particle interactions. As discussed in Section 1.6 below, either of these mechanisms can probably produce the required particle injection rate.

However, one would also like to be able to explain the observed electric voltage difference (potential drop) across the polar cap. The magnitude of this potential drop is related, in the closed model, to the thickness of the boundary layer through which solar-wind particles penetrate the magnetopause.

This relationship is illustrated in Figure 1.4, which shows contours of plasma flow in the equatorial magnetosphere. A magnetized, perfectly conducting plasma moves perpendicular to the magnetic field by $\mathbf{E} \times \mathbf{B}$ drift, the flow velocity being perpendicular to both the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$ (see Appendix 1.A). Consequently, plasma flow lines are also contours of constant electrostatic potential (i.e., voltage). A specific feature of the closed model is that the magnetopause is an

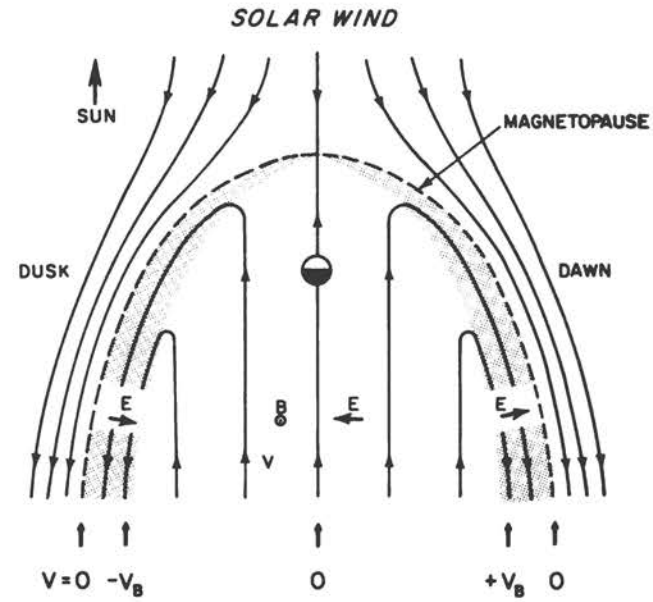


FIGURE 1.4 Sketch of the equatorial-plane $\mathbf{E} \times \mathbf{B}$ flow stream lines in the closed magnetosphere model. The stream lines are also contours of constant electric potential; the electric field $\mathbf{E}$ is perpendicular to these contours, the magnetic field $\mathbf{B}$ is directed out of the page, and the plasma flow $\mathbf{v}$ is in the direction indicated by arrowheads on the flow contours. Since the magnetopause is an equipotential surface in this model, the potential drop across the inner magnetosphere must be equal to twice the potential drop V_B across each boundary layer (shaded). For simplicity, we have not shown the electric field due to the earth's rotation (described in Chapter 2), which dominates the flow near the earth. The actual flow lines, including the effect of corotation, do not go through the earth.

electric equipotential surface and that, correspondingly, the $\mathbf{E} \times \mathbf{B}$ flow lines do not cross the magnetopause. There is a boundary-flow layer just inside the magnetopause wherein plasma flows away from the sun as the result of momentum transfer from the adjacent magnetosheath flow, this momentum transfer being accomplished probably by particle injection across the constant-potential contours by a process that deviates from the $\mathbf{E} \times \mathbf{B}$ drift approximation and the frozen-field theorem.

The average velocity in this boundary-flow layer is estimated from observation to be about 200 km/sec, i.e., slightly less than in the adjacent magnetosheath. The magnetic field strength in the boundary layer is also fixed at about 25γ. Consequently, the electric field is determined to be about 5 V/km. [The relationship between electric field and flow velocity is given in Eq. (1.A.2) of Appendix 1.A.] By reference to Figure 1.4, then, the potential drop across the boundary layer and therefore also the potential drop across the polar cap are determined by the thickness of the boundary layer.

The boundary-layer thickness itself can be estimated crudely for the two proposed mechanisms of particle injection. For the gradient-curvature drift mechanism, the boundary-layer thickness is just the drift-speed component perpendicular to the magnetopause multiplied by the time required for a particle to flow downstream at the boundary-layer flow speed to the point in question. These numbers are well known or easily estimated; the resulting average layer thickness is of the order of 100 km, resulting in a potential drop of the order of 500 V across the boundary layers, or a total of 1000 V = 1 kV across the magnetosphere, since the cross-magnetosphere potential drop must be equal to the sum of the potential drops across the dawn and dusk boundary layers (Figure 1.4).

The diffusion rate is much more difficult to estimate than the gradient and curvature drift speeds, because the process that drives the diffusion is not well understood. However, it is difficult to imagine a diffusion process that could cause particles to random walk (i. e., diffuse) across field lines much more than one gyroradius every gyroperiod. (The "gyroradius" and "gyroperiod" are the radius and period of the circular orbits in which charged particles move under the influence of a uniform magnetic field, as described in Appendix 1.A.) Adopting this large diffusion rate, and assuming that the average boundary-layer flow speed is half that of the solar wind, one can calculate a boundary-layer thickness of about 1000 km, corresponding to a total potential of about 10 kV across the magnetosphere. The potential drop associated with convection is, as we shall see, one of the stumbling blocks of the closed-magnetosphere model.

OPEN-MODEL TRANSFER PROCESS

In the open model of the magnetosphere, the magnetic-field lines from the earth's polar caps connect through the magnetosheath to interplanetary field lines, as shown in Figure 1.3. In the open model, it is difficult to define a sharp boundary between the magnetosphere and magnetosheath, because there are field lines that connect both to the earth and to the solar wind. For these open-field lines, we may say that the region where the magnetic energy density far exceeds the particle energy density is the magnetosphere and that those regions where the particle energy density dominates are the magnetosheath or solar wind. The region along these open-field lines where the two energy densities are comparable may be called the magnetopause boundary layer.

The crucial physical process occurring at the open-model magnetopause is "magnetic merging," which is diagrammed in Figure 1.5. Two regions of differently directed magnetic fields are separated by a slablike field-reversal region of relatively weak magnetic field, shaded in the figure. (For convenience, the fields on the two sides are pictured as antiparallel.) Plasma flows toward the field-reversal region from left and right; and the flux of electromagnetic energy is also directed into the field-reversal region, where electromagnetic energy is converted into particle kinetic energy. This can be visualized in terms of magnetic tension acting along the slingshot-

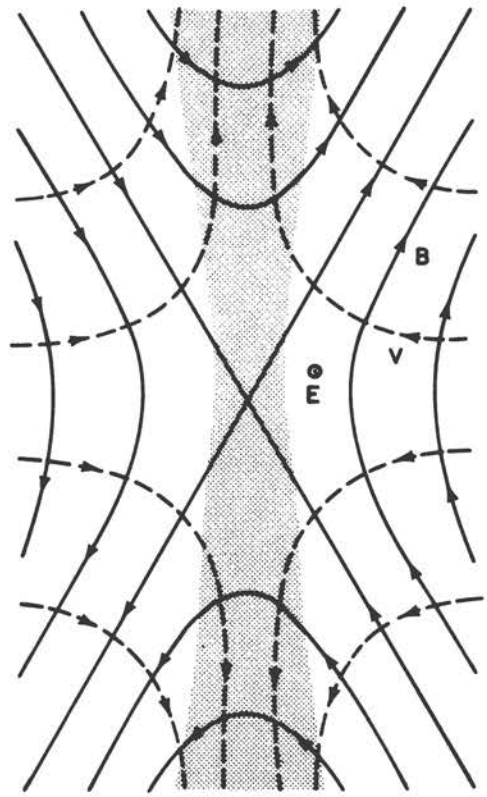


FIGURE 1.5 Diagram of the magnetic-merging process. Solid lines with arrows on them are magnetic-field lines. Arrowed dashed lines are stream lines describing particle flow. The electric field is directed out of the page. The shaded region is the field-reversal region.

shaped field lines to propel plasma away from the center of the field-reversal region. In the theory, there is always a small region in the center of the diagram where the magnetic field is negligibly small but the electric field is not; this implies that the plasma cannot be pictured as a perfect conductor in the central region. The theory of this central region requires invoking sophisticated and controversial plasma physics. However, extensive theoretical work over the last decade has yielded several pertinent results (reviewed by Vasyliunas, 1975); (i) The maximum possible merging speed, i.e., the maximum speed at which plasma can flow toward the field-reversal region, is of the order of the Alfvén speed as measured outside the field-reversal region. The Alfvén speed is the speed of propagation of a particular type of plasma wave along the magnetic field because of the tension in the field, in a way analogous to the vibrations of a stretched string (see, for example, Alfvén and Fälthammar, 1963). The Alfvén speed is of the order of 100 km/sec in the dayside magnetosheath. (ii) The actual merging speed may depend on conditions far away from the merging region and can be much smaller than the Alfvén speed. (iii) The magnitude of the magnetic-field component normal (perpendicular) to the field reversal increases as the merging rate increases; merging speeds much less than the Alfvén speed correspond to normal components that are very small compared to the field magnitude outside the merging region. As discussed in Section 1.5, this relationship may provide an important observational test of the merging theory in the near future. (iv) The merging speed is greatest if the magnetic fields on the two sides are antiparallel to each other but is nonzero for a wide range of angles between the two magnetic-field directions.

The quantitative application of the results to the dayside of the magnetosphere is still an active field of research. Part of the theoretical difficulty lies in the fact that the critical parameter involved in the maximum merging rate, namely, the Alfvén speed just outside the merging region, is a complicated function of the overall flow parameters. However, for purposes of making a rough estimate, we can overlook this subtlety and let the maximum merging speed be given by a typical value for the Alfvén speed in the dayside magnetosheath of about 100 km/sec. Flow speeds in the dayside magnetosheath are typically of the same order of magnitude. Thus the upper-limit merging rate, for the case where the interplanetary magnetic field is southward, opposite to the earth's, may correspond to such rapid flow into the merging region that the magnetosphere constitutes a very soft obstacle. If the magnetosphere is taken to be infinitely soft over a merging region whose width (perpendicular to the page in Figures 1.3 or 1.5) is about 15 R_e, the solar wind's motional electric field $-\mathbf{v} \times \mathbf{B}$ could penetrate directly into the magnetosphere, causing a potential drop of the order of $V_B \times 15\ R_e$—about 200 kV. This value constitutes an approximate theoretical upper limit for the magnetospheric potential drop caused by merging, for average conditions.

It should be mentioned that the merging theory quoted here contains a number of simplifying assumptions, the most important being that the plasma is assumed to behave like a classical isotropic fluid. Recall that this assumption led to the successful prediction of the existence of the earth's bow shock, but also that it cannot describe the detailed structure of that shock. Considerable theoretical work has also gone into attempts to describe magnetic merging from a purely collisionless point of view, the latter approach being more rigorous physically but less sophisticated mathematically. From this work it appears that the fluid-theory results are qualitatively reasonable, although quantitative corrections are to be expected as the collisionless theory progresses.

A field geometry similar to Figure 1.5 but rotated by 90° also applies to the field reversal in the magnetospheric tail; and self-consistency requires that, on the average, the potential drop across the tail must equal the potential drop across the dayside magnetosphere, or, equivalently, the average "reconnection" rate of field lines in the tail must equal the average "connection" rate or merging rate at the dayside magnetopause. This appears to be the case in a long-term average sense. Short-term imbalances do occur, however; and these imbalances are widely considered to be the cause of magnetospheric substorms, as discussed below.

1.5 EXPERIMENTAL EVIDENCE—OPEN VERSUS CLOSED

CRITICAL OBSERVATIONS

The most direct way to distinguish observationally between the open and closed models would be to measure the magnetic-field component normal to the boundary of the dayside magnetosphere. This approach has been pursued recently but so far has resulted in no conclusive evidence for or against magnetic merging. The expected magnitude of the normal component is only about 10 percent or less of the magnetosheath magnetic field, and although this is within the sensitivity range of existing satellite magnetometers, the identification of the normal component is difficult with a single spacecraft because the magnetopause twitches constantly and erratically, and there is no way to monitor the orientation and motion of the boundary.

It is hoped that the next decade will provide direct evidence concerning merging at the dayside magnetopause. During the International Magnetospheric Study, the Mother–Daughter pair of spacecraft will repeatedly fly in close formation through the magnetopause, providing simultaneous two-spacecraft measurements and a means of monitoring the twitching motion.

We now discuss presently available evidence that bears importantly, if indirectly, on the open-versus-closed magnetosphere controversy—evidence involving measured electric fields and particle distributions inside the

magnetosphere. These data bear on the mechanisms of transfer through the magnetopause, because magnetospheric electric fields and particles are believed to result from transfer processes at or through the magnetopause.

ENTRY OF SOLAR-FLARE PARTICLES

In solar-flare events, the sun suddenly emits streams of energetic protons and electrons into interplanetary space, particles with energies of 10^5–10^7 eV. These particles, 10^2 to 10^4 times more energetic than ordinary solar-wind protons, move rapidly and freely along the interplanetary magnetic field, but their motion perpendicular to the field is still given approximately by the $\mathbf{E} \times \mathbf{B}$ drift speed, which is orders of magnitude slower than their field-aligned motion. These energetic particles thus make good field-line tracers: they very quickly fill an entire magnetic flux tube, and they are readily distinguishable from normal magnetospheric and interplanetary particle populations. (The term "flux tube" refers to a magnetic-field-aligned tube containing a group of particles that move together as an entity in accordance with the frozen-in-flux theorem.) Thus it was suggested a decade ago that one could use flare particles as a test of magnetospheric models by comparing the time when the first interplanetary flux tubes filled with energetic flare particles swept past the earth with the times when the particles appeared at various places inside the magnetosphere.

Observations of solar-flare particles in the earth's ionosphere indicate that flare electrons quickly fill the entire polar cap (latitudes greater than about 75°), and thus most of the magnetospheric tail, as soon as their flux tubes reach the earth; this behavior is suggestive of an open magnetosphere. The flare protons, moving slower but with larger gyroradii, quickly fill a band around 75° latitude, which corresponds approximately to the region that is connected by magnetic-field lines to the field-reversal region of the geomagnetic tail and to either the neutral point (C) in Figure 1.2 or to the dayside merging region in Figure 1.3. Proton fluxes eventually fill the entire polar cap but with a typical time delay of a few hours, which, multiplied by the speed of the solar wind, corresponds to a solar-wind flux tube moving downstream a few hundred R_e. A tail length of this order of magnitude is clearly suggested in the open model.

Data on solar-flare-particle entry are generally easier to interpret using the open model than the closed model. Yet the open-model interpretation is not effortless enough for the solar-particle data to constitute a convincing proof of the correctness of the model. There are many subtleties that remain poorly understood, as described in a review by Paulikas (1974).

CONVECTION ELECTRIC FIELDS

Electric-field measurements in the outer magnetosphere are technically very difficult because of the high temperature and low density of the plasma there. Electric fields inferred from plasma-flow measurements [by Eq. (1.A.2) of Appendix 1.A] appear to be generally consistent with the convection theory and with ionospheric electric fields (see Chapter 2) mapped out to the outer magnetosphere by assuming that magnetic-field lines are equipotentials. The geometry of the mapping is illustrated in Figure 1.6. Large field-aligned potential drops along auroral field lines (see Chapter 2), associated with deviations from perfect conductivity, could make the mapping procedure inaccurate. Thus, confirming measurements near the equatorial plane are badly needed, and they may soon be available from several satellites to be launched during the International Magnetospheric Study.

The ionospheric electric fields predicted for open and closed magnetospheres are quite similar, and neither model is worked out in quantitative detail. However, we can mention several aspects of the polar-cap electric-field distribution that seem to bear on the open-versus-closed magnetosphere question. The magnitude of the observed potential drop across the polar cap averages about 50 kV and has been observed in one exceptional case to be as large as 225 kV. The observed ionospheric potential drops are thus comfortably less than that corresponding to the upper-limit merging rate of the open-model theory described above; apparently the merging speed is significantly less than the average Alfvén velocity in the dayside magnetosheath. On the other hand, we know of no closed-model mechanisms that can easily produce an average potential drop of 50 kV across the polar cap. The fast diffusion of particles discussed above (random walk of one gyroradius every gyroperiod) is inadequate by a

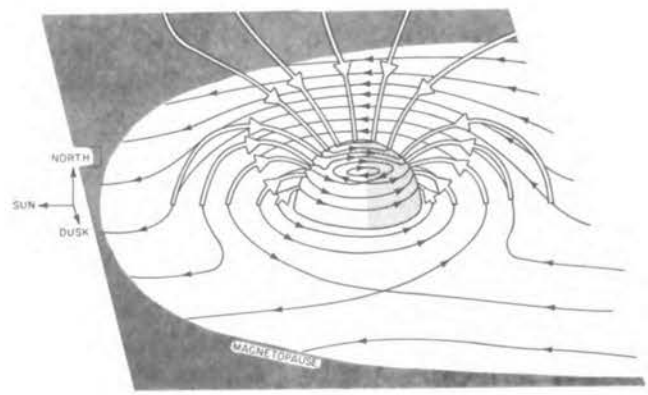


FIGURE 1.6 Typical convective flow pattern in the earth's magnetosphere and topside northern ionosphere, as viewed from a few R_e above the equatorial plane on the dusk side. The sun is to the left. The solid lines with solid arrows are stream lines on two selected surfaces representing flow lines for low-temperature plasma; one surface is the equatorial plane, and the other is a sphere concentric with the earth, in the topside ionosphere. The open lines with open arrows are sample magnetic-field lines in the noon–midnight meridian plane. Velocities are shown relative to a reference frame that does not rotate with the earth, so the earth's rotation dominates the flow near the earth in the equatorial plane and in the low-latitude ionosphere. Convective flow in the outer equatorial plane is sunward. On polar-cap field lines it is antisunward.

factor of the order of 5. Production of a 50-kV potential drop by closed-model particle transport would require that solar-wind protons drift systematically earthward across boundary-layer field lines at a speed of the order of their thermal speed. Thus the observed polar-cap potential drop is a limited victory for the open model, limited because the dayside merging rate has not been rigorously calculated and all possible closed-model transport mechanisms have not been carefully explored.

The average observed convection potential is about 10 percent of the average solar-wind electric field ($-\mathbf{v} \times \mathbf{B}$) multiplied by the total diameter of the magnetosphere ($\approx 40\, R_e$). This 10 percent electric-field penetration required to explain the observed magnetospheric convection should be compared with the 0.1 percent penetration of solar-wind particles required to explain the particle populations in the outer magnetosphere and the energy transfer from the solar wind to the magnetosphere. (See Section 1.6.) Consequently, explanation of the observed convection potential requires a powerful and direct-transfer mechanism like magnetic merging, while the weaker closed-model processes could transfer sufficient particles and energy.

Moreover, the convection pattern is observed to be strongly correlated with the direction of the interplanetary magnetic field. The electric-field magnitude is not uniform across the polar cap but tends to be larger near the edges of the cap than near the center. This in itself is consistent with either the closed or the open model. However, there is also a dawn–dusk asymmetry in the flow pattern (see Figure 2.6 of Chapter 2) that is correlated with the east–west component of the interplanetary magnetic field, i.e., the component in the equatorial plane, perpendicular to the earth–sun line. (The antisolar flow tends to be stronger on the dawn side of the polar cap than on the dusk side in the northern hemisphere, and vice-versa in the southern hemisphere, when the interplanetary magnetic field has a dawn-to-dusk component; the situation is reversed when the interplanetary field has a dusk-to-dawn component.) This correlation can be explained qualitatively in the open model in terms of the asymmetric tension exerted by the interplanetary magnetic field on geomagnetic-field lines to which it is connected. There is no obvious explanation for this correlation in the closed model, although it is not implausible that the efficiency of a diffusion process could somehow be affected by the direction of the interplanetary field.

Another significant correlation is the following: the northern and southern polar caps both tend to be larger (i.e., extend to lower latitudes) when the interplanetary magnetic field has a southward component than when it has a northward component (Burch, 1974). This is qualitatively what is expected in the open model. The predicted rate of merging at the dayside magnetopause, i.e., the rate of production of open field lines, is greatest when the interplanetary field is southward (opposite to the earth's field). Thus a southward interplanetary field would tend to increase the number of open field lines. In the open model, the polar cap is the region of the ionosphere that lies on open field lines, so that an increase in the number of open field lines corresponds to an increase in polar-cap area.

In general, the observations of solar-flare particle entry, convection electric fields, and polar-cap size tend to favor the open model of the magnetosphere over the closed model. We should, however, emphasize that neither model has been worked out in sufficient detail to allow a quantitative comparison with the observations.

1.6 SOLAR-WIND PLASMA IN THE MAGNETOSPHERE

OBSERVATIONS AND THEIR IMPLICATIONS

Figure 1.7 illustrates the locations of various plasma populations that are thought to be important in the solar-wind–magnetosphere interaction. It is widely believed that the solar wind is the principal source of these outermost magnetospheric plasma regions, but the mechanisms of solar-wind injection into the magnetosphere have not yet been established.

The plasma sheet was the first of these regions to be discovered, and it remains perhaps the least understood. The other labeled regions (cusp, mantle, boundary layer) are thought to be intermediate regions through which solar-wind particles travel on their way from the magnetosheath into the plasma sheet. The distinctions among these four regions derive in part from different times and methods of discovery and do not necessarily imply the existence of sharp boundaries. Plasma flows antisunward in both the mantle and the boundary layer. At lunar orbit (about $60\, R_e$ behind the earth, beyond the scope of Figure 1.7), these two regions expand to include a large fraction of the tail, including much of the region just above and below the plasma sheet. The ring current discussed in Chapter 2 is in part the earthward extension of the plasma sheet.

Recent satellite observations of heavy ions (notably helium and oxygen) in the ring current and plasma sheet indicate that the ionosphere may provide a larger fraction of the total magnetospheric plasma than was once supposed. Ionospheric particles can populate the outer magnetosphere by the solar-wind process described in Chapter 2. Deciding the relative importance of ionospheric and solar-wind plasma sources is presently an active problem in magnetospheric research. Here we describe the proposed mechanisms of solar-wind-plasma injection to populate the outer magnetospheric plasma regions of Figure 1.7.

The existence of the plasma sheet is intimately connected with the existence of the magnetospheric tail. The two bundles of oppositely directed magnetic flux in the tail must be separated by a sheet of electric current flowing at the field reversal, and this current is carried by

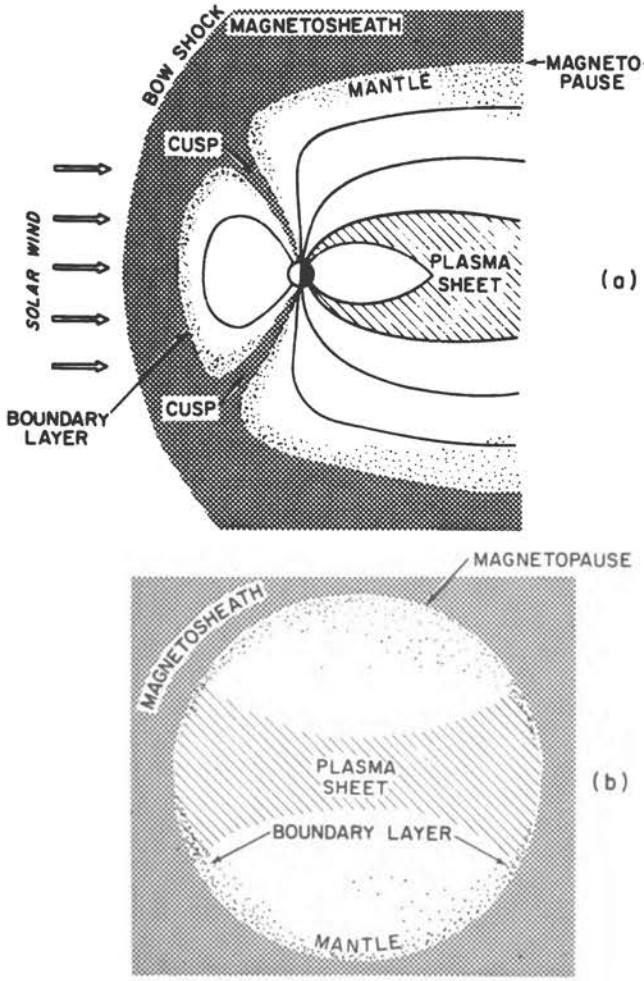


FIGURE 1.7 Illustration of the magnetosphere showing locations of various solar-wind plasma regions, in (a) the noon–midnight meridian cross section and (b) the tail cross section, at about 30 R_e behind the earth, as seen looking away from the sun.

the particles of the plasma sheet. These particles are typically observed with number densities of 0.1 to 1 particle cm^{-3} and temperatures of 5×10^7 K for protons and 5×10^6 K for electrons. Thus the plasma-sheet particles have an average energy ten times that of their counterparts in the magnetosheath and are less dense by a factor of 10–100.

These observations imply that the injection process from the solar wind must be diffusive, selective, or both; a simple adiabatic flow, for example, would change the density and temperature in the same sense, not in opposite senses as observed.

The plasma properties in the cusp, mantle, and boundary layer (in that order) are suggestive of a one-way transition from magnetosheath characteristics to plasma-sheet characteristics, but the situation is complicated. For example, at lunar orbit, observations often indicate a sharp boundary between the hot plasma sheet and the relatively cooler mantle plasma flowing above and below the plasma sheet.

The mantle and boundary layer are discoveries that were made in the last few years. Determination of their spatial relationships and other basic properties are exciting ongoing research problems now. To sort out their behavior and interrelationships will require further analysis of existing data and probably acquisition of new data as well.

Estimates of the various loss processes operating in the plasma sheet, as described in Chapter 2, indicate that a source injection rate of the order of 10^{26}/sec is required to maintain the plasma sheet. This estimate is uncertain by perhaps an order of magnitude, but it indicates that only a small fraction (about 10^{-3}) of the solar-wind flux incident on the magnetospheric cross section needs to be captured in order to supply the various plasma regions observed in the outer magnetosphere, with the remainder of the solar-wind particles flowing around the magnetosphere as in the simple closed model described above. Although the magnetosphere extracts only a small fraction of the available solar-wind particle flux (and the associated energy), this seemingly insignificant departure from the impenetrable-obstacle picture has profound influences on the dynamics of the magnetosphere and its interaction with the upper atmosphere, as discussed further in Chapter 2.

PLASMA INJECTION

There are three basically different mechanisms that have been proposed for the injection of solar-wind particles from the magnetosheath into the magnetosphere. These are (1) magnetic merging at the magnetopause, (2) diffusion of particles across the magnetopause, and (3) magnetic-field drift of particles across the magnetopause. The first of these applies explicitly to the open-magnetosphere model; the other two are basically closed-model concepts, although they are readily adapted to an open-magnetosphere model as well.

The merging model is particularly simple, in principle. When a field line from the dayside magnetosphere opens up and connects to an interplanetary field line, solar-wind particles are free to move along this field line from the magnetosheath into the magnetosphere. Such particles will move in the antisolar direction across the polar cap as the solar-wind particles on the same field line flow downstream. When the two halves of the geomagnetic-field line reconnect on the nightside of the earth, any solar-wind particle that has lingered too long on the earthward side of the reconnection point will become trapped on the newly closed field line and subsequently convect back toward the earth to become part of the plasma sheet. The frozen-in-flux condition is violated only near the points of field-line connection and reconnection; the flow over the polar cap is describable in terms of frozen-in flux. The polar-cap field lines must stretch to great distances in the tail during this convection

(several hundred R_e), since the plasma near the earth is traveling much slower that the solar-wind plasma on the same field line.

The cusp, mantle, and boundary layer can presumably be thought of in this model as intermediate steps in the convection across (or around) the polar caps, although no comprehensive theoretical model of this process is available. Similarly, no quantitative estimates are available of the injection rate and particle energy to be expected from such a model. The reason for the lack of a quantitative model is primarily the fact that the merging process itself is not well understood.

We should remark that, although about 10 percent of the interplanetary magnetic-flux tubes incident on the front of the magnetosphere must merge with geomagnetic-field lines if the open model is to explain magnetospheric convection, only about 10^{-3} of the incident solar-wind particles are captured by the magnetosphere. Thus most of the solar-wind particles on the merged field lines must escape from the magnetosphere before reconnection of field lines in the tail or must subsequently leak out the sides of the tail.

In the diffusion model, it is assumed that magnetosheath particles are kicked across the magnetopause onto closed-field lines by a random (diffusive) process occurring in the cusp or along the front and sides of the magnetopause or both. A total flux of 10^{26}/sec can be provided by a diffusion rate considerably smaller than the fast diffusion rate discussed earlier (random walk of one gyroradius per gyroperiod). Although there is no adequate theory for diffusion across the magnetopause, diffusion might well be capable of transmitting enough solar-wind particles to populate the plasma sheet.

The magnetic-drift model is similar to the diffusive one except that the particle transfer is adiabatic instead of stochastic, i.e., no random electromagnetic fluctuations are required. Charged particles in a nonuniform magnetic field experience a slow but systematic drift motion perpendicular to the field direction, as described in Appendix 1.A. The field inhomogeneity may be a gradient in field strength or a curvature of the field lines or, in general, a combination of the two. The gradient and curvature drift speeds are proportional to the particle energy per unit charge, and therefore they violate the frozen-in-flux condition since particles of different energies or charges drift at different rates and therefore end up on different field lines. These drift trajectories may well intersect the magnetopause, especially near the field-reversal region in the tail. It has been proposed that such drift motion accounts for injection of magnetosheath particles into the plasma sheet. In the closed model, gradient and curvature drifts are also especially large near the dayside neutral points marked C in Figure 1.2 (e.g., Willis, 1971).

As with the diffusion model, the required injection rate can be easily accounted for by the gradient/curvature drift model, but the implied distribution of magnetospheric particles has not been worked out.

PARTICLE ACCELERATION

The above three injection mechanisms have one thing in common: they all require an acceleration process after the particles gain access to the magnetosphere in order to account for the high temperatures observed in the plasma sheet. It is likely that this acceleration is provided by magnetic-field annihilation in the field-reversal region of the tail. We have already seen that the plasma sheet carries a dawn-to-dusk electric current across the center of the tail, and that, in addition, a dawn-to-dusk electric field exists there in association with the earthward convection. As a result, magnetic-field energy is constantly being converted into particle energy in a "field annihilation" process that is somewhat analogous to the electrical heating of a current-carrying resistor.

This acceleration has been analyzed extensively from a variety of plasma theoretical viewpoints. The most general result is that particles are energized by an amount equal to the magnetic energy per unit electrical charge in the plasma. This corresponds to an energy of the order of 5 keV (i.e., a temperature of 5×10^7 K) in the plasma sheet, which is consistent with the observed particle energies. Thus, while the detailed mechanisms of energy conversion are still a subject of considerable debate, there seems to be little doubt that the particle energies can be accounted for in terms of conversion of magnetic energy into particle energy.

While the acceleration of plasma-sheet particles can be qualitatively understood in terms of magnetic merging, the problem of their injection into the magnetosphere is still largely unsolved. We have mentioned three possible mechanisms above, but none of these mechanisms is consistent with all available observations (see review by Hill, 1974). It is likely that all three mechanisms play a role, but a considerable amount of both theoretical and observational work is apparently required to determine what those roles are.

1.7 TIME-DEPENDENT INTERACTIONS

At the beginning of this chapter we noted that time variations in the geomagnetic field provided the first clues to the existence of the solar-wind–magnetosphere interaction. Thus far we have described only the steady-state interaction without time variations. Although the magnetosphere rarely, if ever, achieves such a steady state, it is nevertheless necessary to attempt to understand the steady-state interaction before we can hope to understand the time variations that perturb that steady state.

STORMS AND SUBSTORMS

The solar wind is almost always changing in time, because of variations in the solar source regions. These variations in the solar wind produce time variations in the

solar-wind–magnetosphere interaction as the magnetosphere tries to adjust to a new steady state appropriate to the new solar-wind conditions. The large-scale time variations in the magnetosphere can usually be categorized either as "geomagnetic storms" or as "magnetospheric substorms." Geomagnetic storms are described in detail by Chapman and Bartels (1940); substorms are described in detail by Akasofu (1968) and in the more recent review by Russell and McPherron (1973).

Pressure variations in the solar wind often occur in the form of sharp discontinuities or shock waves. When those pressure discontinuities convect past the earth, the magnetosphere is suddenly compressed (or decompressed, depending on the sign of the pressure change). These sudden impulses, as they are called, are observed on the earth as sudden increases or decreases in the worldwide surface magnetic-field strength. We can understand the propagation of these pressure impulses by recalling that the earth's magnetic field exerts a pressure against the solar wind, the field pressure being proportional to the square of the field strength. The sudden impulse travels from the magnetopause to the earth's surface in the form of a propagating pulse of enhanced field strength, i.e., a pressure wave.

These sudden impulses often trigger large-scale magnetospherewide disturbances known traditionally as geomagnetic storms. The main phase of the geomagnetic storm is a period of several hours to several days during which the worldwide surface magnetic field is disturbed and generally depressed in magnitude because of impulsive injection of fresh plasma into the magnetospheric ring current (see Chapter 2). This plasma is thought to be injected from the tail plasma sheet in association with magnetospheric substorms.

The geomagnetic-storm phenomenon had been studied in great detail before the advent of the space age, using surface observations. The most complicated part of the geomagnetic storm is the magnetospheric substorm, and scientific interest in the substorm process has increased dramatically as satellite observations have increasingly demonstrated that the substorm involves a large-scale disturbance of the entire magnetosphere.

It is convenient to divide substorms into three phases: the growth phase, the expansion phase, and the recovery phase. The expansion and recovery phases will be described in Chapter 2; they are, apparently, internal magnetospheric processes that can, however, be triggered by time variations in the solar-wind–magnetosphere interaction. The growth phase, on the other hand, is essentially a solar-wind–magnetosphere interaction, which involves a large-scale change in the configuration of the outer magnetosphere.

It should be pointed out that observations of growth-phase phenomena have been reported without any ensuing substorm expansion activity, and conversely, substorm occurrences have been reported without any preceding growth-phase phenomena. Thus it may be more appropriate to consider the growth phase as a characteristic solar-wind–magnetosphere interaction that may lead to substorms but that is not in itself an essential part of the substorm process. To avoid confusion, we shall speak of the "tail-growth" phenomenon rather than the "growth-phase" phenomenon, emphasizing that it is essentially a time-dependent solar-wind magnetosphere interaction and leaving for future research the question of what role it plays in causing substorms.

GROWTH AND INSTABILITY OF THE TAIL

During the tail growth, which lasts a few minutes to an hour, the magnetic-field strength in the tail increases and the tail field geometry presumably becomes more stretched out in the antisolar direction. Meanwhile, the plasma sheet becomes gradually thinner in its north–south extent, shrinking toward the equatorial plane, although it rarely if ever disappears entirely. There is a decrease in the total magnetic-flux content of the dayside magnetosphere, corresponding to an increase of total flux in the distant tail. During this time, the earthward convection of plasma-sheet particles may be enhanced—but not dramatically so.

The growth of the tail is typically reversed in an explosive event that is commonly associated with a substorm expansion phase, lasting several minutes or tens of minutes, when the stretched-out tail field appears to "collapse" back toward a more nearly dipolar configuration, and plasma-sheet particles are impulsively injected into the inner magnetosphere and begin to form a new or enhanced ring current. The partial collapse of the tail magnetic field is associated with a partial disruption of the electric current across the tail; this current appears to be rerouted along the magnetic-field lines into the auroral-zone ionosphere, across the auroral zone in a dawn–dusk direction, and back out along the field lines into the tail, as illustrated in Figure 2.10 of Chapter 2. The section of this current loop that crosses the auroral-zone ionosphere (the auroral electrojet) causes the ground magnetic variations known as the polar magnetic substorm, and it is associated with the brightening and rapid motions of visible auroral arcs known as the auroral substorm. Meanwhile, the plasma from the plasma sheet that has been injected into the ring current causes the worldwide decrease of the surface field that characterizes the main phase of a geomagnetic storm. Shortly after the beginning of the expansion phase, the plasma sheet suddenly recovers to its presubstorm thickness and appears, if anything, hotter than before the thinning that occurred during tail growth.

The most important solar-wind parameter involved in the tail growth phenomenon seems to be the north–south component of the interplanetary magnetic field (although several other solar-wind parameters have been found to be correlated with substorm occurrence, and the picture is far more complicated than presented here). Specifically, tail growth is most likely, and most intense, when the solar wind has a large magnetic-field component

pointing southward, i.e., opposite the direction of the magnetospheric magnetic field at the dayside magnetopause.

This observed correlation is easily explained in terms of the open model. A southward interplanetary magnetic field causes enhanced merging of geomagnetic and interplanetary field lines at the dayside magnetopause, setting up a temporary net transfer of magnetic flux from the dayside magnetosphere into the tail. The sudden collapse of the tail is viewed as the process by which the reconnection rate in the tail adjusts itself to match the increased connection rate at the dayside magnetopause, so that the tail flux does not increase indefinitely. The period of tail growth is interpreted as the period of time after the dayside connection rate has increased but before the nightside reconnection rate has responded with a similar increase; hence the temporary increase in the tail magnetic flux.

It is not known why the plasma sheet thins during the tail growth; similarly it is not understood why the plasma sheet suddenly re-expands shortly after the collapse of the tail field. The variations in plasma-sheet thickness and temperature are often interpreted in a qualitative way as evidence of the rapid motions of the tail reconnection region (neutral line) toward and away from the earth, but there is no generally accepted theoretical model that explains why the reconnection rate should vary impulsively or what controls the location and motion of the reconnection region.

In any event, the impulsive collapse of the tail field is widely interpreted as a sudden increase in the reconnection rate in the tail, which seems to be initiated relatively near the earth (within 10–30 R_e). The consequent sudden injection of plasma into the inner magnetosphere then results in a myriad of substorm processes that are no longer directly related to the solar-wind triggering mechanisms.

As the reader may have guessed, there is no available theory of the tail growth and instability that can encompass all the available observations, even though the observations are far from complete. It is fair to say that the geomagnetic storm is reasonably well understood, given the essential element of magnetospheric substorms, but our empirical picture of what happens during a substorm is still being developed, and our theoretical understanding of the solar-wind triggering of substorms, and of the magnetosphere's response, is still in a primitive stage.

1.8 OUTSTANDING PROBLEMS

The main features of the outer magnetosphere and magnetosheath are now well established, including the existence and shape of the magnetopause and the bow shock, as well as the basic solar-wind flow pattern around the magnetosphere. There are, however, major holes in our knowledge and understanding. Since most of the large-scale aspects of the situation are well defined, our ignorance with regard to these key elements of the physics of the system becomes very glaring. Discussed below are some of these glaring problems on which significant progress may be expected in the next decade.

Magnetic merging is a process of general astrophysical significance because it converts electromagnetic energy into particle energy. Above the earth's atmosphere we directly observe relativistic particles accelerated in solar flares, as well as galactic-cosmic-ray particles of enormous energy. Furthermore, most of the radio and x-ray sources in the universe seem to result from fluxes of energetic particles far above thermal equilibrium. Determining the mechanisms that defy thermal equilibrium and energize these particles is one of the central problems of astrophysics, and magnetic merging is one of the most promising and most discussed acceleration mechanisms. This is a case for which the magnetosphere should be particularly useful as a plasma-physics laboratory. Measurements made in the last decade have placed important constraints on our picture of magnetic merging at the magnetopause, particularly with respect to limits on total merging rates and their dependence on the interplanetary magnetic field. We anticipate that observations in the next decade will provide direct experimental evidence regarding the degree of interconnection of magnetospheric and interplanetary field lines and more detailed data on the structure of the magnetopause boundary. We also lack direct observational proof that magnetic merging is the mechanism responsible for accelerating the solar-wind plasma up to the several kilovolt energies characteristic of the plasma sheet.

Particle diffusion across a magnetic boundary is another process of general astrophysical importance. More detailed observations of particle diffusion across the magnetopause (for example, by the Mother–Daughter satellite duo) should provide important information on the collisionless diffusion processes involved.

The collisionless shock wave is another general plasma phenomenon that is poorly understood theoretically. Measurements by the Mother–Daughter satellites, for example, will greatly increase our observational knowledge of the earth's bow shock and thus it is hoped provide theorists with the clues they need to solve the collisionless shock problem.

Magnetospheric substorms are in critical need of understanding for two reasons, aside from the intrinsically interesting nature of the subject: there is a practical need to predict their occurrence times and the resultant disruptions of radio communications and power transmissions, and their contributions to the energy budget of the neutral upper atmosphere (see Chapter 3) are important. If we understood their explosive onsets, we might then have physical models for solar flares and many other explosive astrophysical events. Research of the last decade has enormously increased our observational knowledge, particularly with regard to conditions in the solar wind that cause substorms. The internal dynamics of a

substorm remains a mystery, but at least the mystery has been reduced now to a rather well-defined set of specific questions that need to be attacked in future research.

1.9 SIGNIFICANCE OF SOLAR-WIND INTERACTIONS

Research into the nature of the outer magnetosphere and its coupling to the solar wind seems to us particularly interesting and important for three basic reasons.

The magnetosphere is an intrinsically interesting place. It is a large but conveniently located stage where nature frequently puts on dazzling shows, as when a quiet magnetosphere dramatically erupts in a large substorm. The magnetosphere is suddenly ablaze with plasma waves of many kinds and superfast particles of various types and energies, some of them lighting up the high-latitude atmosphere with the sort of auroral display that has awed men for centuries.

The magnetosphere is a part of man's environment. Spacecraft on various practical missions operate in the magnetosphere and ionosphere; it is helpful to understand and be able to predict various aspects of the spacecraft environments—radiation levels and satellite drag, for example. The neutral upper atmosphere, the ionosphere, the inner and outer magnetosphere, and the solar wind are so strongly coupled that it is difficult to understand one region without having a reasonably good understanding of the others. As discussed further in Chapter 6, changes in solar-wind conditions cause enhanced ionization in the high-latitude ionosphere and thereby disrupt radio communications between points on the earth's surface. There is even an increasing body of data indicating that solar-wind changes cause small but significant effects on the lower atmosphere's weather.

The magnetosphere is a laboratory for observation of large-scale plasmas. Much of the universe is filled with low-density plasma, and many observed astrophysical phenomena clearly involve subtle plasma-physics interactions. However, astrophysical data are very limited, and theories cannot be checked in detail with observations. The earth's outer magnetosphere provides a unique collisionless plasma-physics laboratory: it is a place where researchers studying large-scale, low-density plasmas can gain experience and have their theories tested with observations in a large system whose plasmas and magnetic fields are much closer to those of typical astrophysical media than to those attainable in the laboratory. Furthermore, magnetospheres appear to be quite common astrophysical phenomena, to which parts of our knowledge of the earth's magnetosphere are immediately applicable. Of the four other planets observed by spacecraft so far, Jupiter and Mercury clearly have magnetospheres, and Mars may well have one. Jupiter's magnetosphere, if it were visible, would be the largest object in the sky (appearing larger than the sun even though it is five times more distant); this giant Jovian magnetosphere apparently competes with the sun as a source of energetic-particle radiation in the solar system. At the other extreme, Mercury's magnetosphere is barely large enough to enclose the planet, but the recent discovery of even such a small magnetosphere at Mercury was a complete surprise that is profoundly reshaping theories of the interiors of planets.

On a more astronomical scale, pulsars are generally thought to involve neutron stars with rapidly rotating magnetospheres, analogous in some ways to that of Jupiter. Certain radio galaxies (ones that move supersonically through the intergalactic medium in clusters of galaxies) are observed to have long, ordered magnetic tails that emit at radio frequencies; these huge structures strongly resemble the earth's magnetospheric tail, despite being 10^{13} times larger in linear dimension. Evidently our magnetosphere provides a unique observational tool for the understanding of a wide variety of cosmic plasma phenomena.

APPENDIX 1.A: MOTION OF CHARGED PARTICLES IN ELECTRIC AND MAGNETIC FIELDS

Our discussion of the solar-wind–magnetosphere interaction relies on the following simple properties of charged-particle motion under the influence of large-scale electric and magnetic fields. We consider a highly rarified plasma like the solar wind, in which particle–particle collisions are so infrequent that their effect on particle motion is negligible. A more complete description is given, for example, by Alfvén and Fälthammar (1963).

An electric field **E** accelerates a positive charge parallel to itself, and a negative charge antiparallel. The magnetic force on a moving charge is perpendicular both to the magnetic field **B** and to the particle's velocity, the direction of the force being opposite for positive and negative particles. Written algebraically, the force on a particle of charge q and velocity **v** (called the Lorentz force) is given by

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \qquad (1.A.1)$$

Consider for a moment the simple case for which there is no electric field and the magnetic field is spatially uniform. Motion perpendicular to the magnetic field results in a magnetic force that is perpendicular to the particle's velocity; the motion therefore is circular, the radius of the circle (called the "gyroradius") being determined by the requirement that the magnetic force balance the centrifugal force. The period of the circular motion is called the "gyroperiod." Because the magnetic field deflects positive and negative particles in opposite directions, positive ions and negative electrons rotate about the magnetic field in opposite senses, as shown in Figure 1.8, Case a. Thus, given a magnetic field and no electric field,

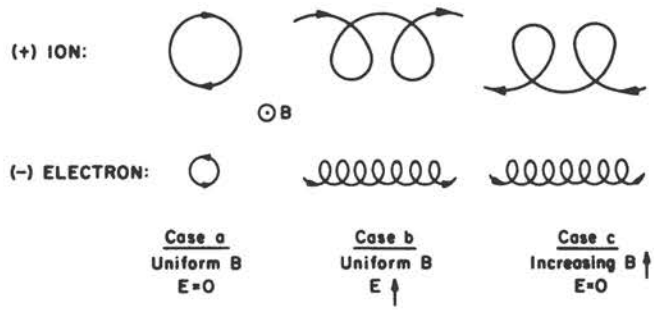


FIGURE 1.8 Motions of charged particles in the plane perpendicular to the magnetic field. Case a illustrates the circular motion that occurs when the magnetic field (directed out of the paper) is uniform and there is no electric field. Case b shows the effect of an upward electric field, namely, a rightward drift that is the same for both positive and negative particles. Case c shows the effect of having the magnetic-field magnitude increase upward on the page—positive and negative particles "gradient-drift" in opposite directions.

charged particles can move across magnetic-field lines only to the extent of moving in circles, generally quite small circles compared with most space-physics dimensions. However, they can move quite easily along a magnetic-field line, since there is no magnetic force in that direction. A steady electric field parallel to **B** steadily accelerates ions and electrons in opposite directions, building up larger and larger electrical current until the originally imposed electric field is neutralized, unless some more subtle process comes into play to restrict particle motion parallel to **B**.

Figure 1.8, Case b, shows the effect of a steady, uniform electric field perpendicular to **B**. On the side of the orbit where the particles have higher velocities due to acceleration by the electric field, they cannot turn so sharply as on the side of the orbit where they move more slowly. Consequently, the orbits do not form closed circles; the particles drift in a direction perpendicular to both **E** and **B**, at a speed E/B (in SI units). This drift, which is independent of the charge of the particle, is called $\mathbf{E} \times \mathbf{B}$ drift and can be expressed mathematically as follows:

$$\mathbf{V}_{E\times B\ \text{drift}} = \mathbf{E} \times \mathbf{B}/B^2. \qquad (1.A.2)$$

Considering the high mobility of charged particles parallel to magnetic-field lines, it is often assumed that the component of **E** parallel to **B** is zero. This condition, combined mathematically with Eq. (1.A.2), implies that, if the particles in a plasma all move at the $\mathbf{E} \times \mathbf{B}$ drift velocity, then the electric field is zero in the plasma's frame, the particles simply gyrate, as shown in Figure 1.8, perfect conductor, excluding an electric field from its interior, as seen in the plasma rest frame. In that rest frame, the particles simply gyrate, as shown in Figure 1.8, Case a. The assumption of perfect conductivity can be shown to imply that magnetic-field lines are "frozen into" the plasma; i.e., if two $\mathbf{E} \times \mathbf{B}$ drifting particles started out gyrating about the same magnetic-field line, they will continue forever sharing the same field line, no matter what drifts they experience.

Figure 1.8, Case c, illustrates that a gradient in the magnitude of the magnetic field causes charged particles to gradient-drift, i.e., to move systematically in a direction that is perpendicular to both the magnetic field and its gradient. The drift results from the fact that the orbit turns more sharply on the side with stronger magnetic field than on the side with weaker field. Gradient drift is usually slow in the solar wind and the outer magnetosphere, but it represents a slight thawing in the generally frozen relationship between the charged particles and the magnetic field.

Curvature of the magnetic-field lines causes a very similar kind of drift motion, called "curvature drift." Gradient and curvature drifts, unlike $\mathbf{E} \times \mathbf{B}$ drift, move positive and negative particles in opposite directions and thus produce an electrical current. It is, in part, these electrical currents carried by plasma particles that cause the self-consistency problem mentioned in Section 1.2 and result in the small but significant violations of the perfect-conductivity assumption that are so important in the solar-wind–magnetosphere interaction.

ACKNOWLEDGMENTS

We are grateful to F. S. Johnson, J. L. Burch, M. Harel, and P. H. Reiff for helpful comments on earlier drafts of this chapter. The work at Rice University was supported in part by the Atmospheric Sciences Section, National Science Foundation, Grant DES74-21185 and by NASA Grant NGL44-006-012. The work at NOAA was supported by a Resident Research Associateship from the National Research Council of the National Academy of Sciences.

REFERENCES

Akasofu, S. I. (1968). *Polar and Magnetospheric Substorms*, D. Reidel, Dordrecht, Holland.

Alfvén, H., and C.-G. Fälthammar (1963). *Cosmical Electrodynamics: Fundamental Principles* (2nd ed.), Oxford U. Press, London.

Brandt, J. C. (1970). *Introduction to The Solar Wind*, W. H. Freeman, San Francisco.

Burch, J. L. (1974). Observations of interactions between interplanetary and geomagnetic field, *Rev. Geophys. Space Phys. 12*, 363.

Chapman, S., and S. Bartels (1940). *Geomagnetism*, Oxford U. Press, London.

Dessler, A. J. (1967). Solar wind and interplanetary magnetic field, *Rev. Geophys. 5*, 1.

Greenstadt, E. W., and R. W. Fredericks (1974). Plasma instability modes related to the earth's bow shock, in *Magnetospheric Physics*, B. M. McCormac, ed., D. Reidel, Dordrecht, Holland, p. 281.

Hill, T. W. (1974). Origin of the plasma sheet, *Rev. Geophys. Space Phys. 12*, 379.

Paulikas, G. A. (1974). Tracing of high-latitude magnetic field lines by solar particles, *Rev. Geophys. Space Phys. 12*, 117.

Russell, C. T., and R. L. McPherron (1973). The magnetotail and substorms, *Space Sci. Rev. 15*, 205.

Spreiter, J. R., A. Y. Alksne, and A. L. Summers (1968). External aerodynamics of the magnetosphere, in *Physics of the Magnetosphere*, R. L. Carovillano, J. F. McClay, and H. R. Radoski, eds., D. Reidel, Dordrecht, Holland, p. 336.

Vasyliunas, V. M. (1975). Theoretical models of magnetic field-line merging, 1, *Rev. Geophys. Space Phys. 13*, 303.

Willis, D. M. (1971). Structure of the magnetopause, *Rev. Geophys. Space Phys. 9*, 953.

2 The Magnetosphere

JAMES L. BURCH
Marshall Space Flight Center, National Aeronautics and Space Administration

2.1 PROLOGUE

At altitudes near and above 100 km, the thinning air of the upper atmosphere begins to allow electrons and positive ions to travel significant distances between collisions with neutral atoms and molecules. As the close coupling between the ionosphere and the neutral atmosphere is relaxed, the earth's magnetic field is able to exert a strong influence on the motion of the ambient ions and electrons. This transition from neutral-gas control to magnetic-field control, which is complete for electrons at altitudes near 90 km and for ions at altitudes near 120 km, defines the lower boundary of the earth's magnetosphere. Extending upward some 60,000 km on the dayside and of the order of a million kilometers on the nightside, the magnetosphere terminates finally in a region of complex interaction with the expanding solar corona, or solar wind; this is discussed in Chapter 1. In this chapter, we are concerned with the vast region of magnetized plasma, or ionized gas, that is the earth's magnetospheric environment.

So rarefied as to be virtually collisionless, the magnetospheric plasma exists in a tremendous variety of density, temperature, and flow regimes. At distinct boundaries separating these various plasma regimes, and as some of the plasma populations flow through others, strong interactions often occur that result in the generation of intense plasma waves, the acceleration of electrons and ions to high energies, and the ultimate deposition of energy in the earth's atmosphere amounting to approximately 1 percent of the total solar-wind energy incident on the entire magnetosphere. During a typical magnetospheric plasma disturbance (or substorm) lasting about 1 hour, a power of between 10^{11} and 10^{12} W, or roughly that of a very strong earthquake, is continuously dissipated in the atmosphere. Since substorm disturbances occur on the average about every 4 hours, their frequency increasing markedly when sunspot activity is high, the magnetosphere is identified as a significant source of energy for the atmosphere above about 90 km. Statistical evidence is accumulating that this magnetospheric energy input may at times couple efficiently downward into the

troposphere, acting there as a trigger or modulator for some large-scale weather patterns. This evidence has been summarized by King (1975).

Magnetospheric disturbances are known to be directly responsible for the anomalous behavior of a number of man-made systems. For example, spacecraft in operating at stationary orbit, some 30,000 km above the earth, often become charged to voltages as high as 10 kV as they are enveloped by clouds of energetic electrons during magnetospheric substorms. The electrical discharges that may result from this charging are thought to be responsible for some of the malfunctions and failures that have occurred within such spacecraft. Closer to the earth, the increased ionization that is produced over the polar caps during magnetospheric disturbances interferes regularly with communication systems and over-the-horizon radars by altering the effective path length of radio waves reflected off the ionization layer. Intense ionospheric currents induce image currents in the ground that have seriously disrupted electric power and telephone-communication systems (Anderson *et al.*, 1974). In these areas it is clear, then, that better understanding of magnetospheric processes and the ultimate ability to forecast and compensate for their effects on man-made systems have become practical matters.

Following decades of dedicated ground-based experimentation and theoretical investigation into the nature of geomagnetic phenomena and the aurora by pioneering scientists such as Kristian Birkeland (1867–1917), Carl Störmer (1874–1957), Sydney Chapman (1888–1970), and Hannes Alfvén, the advent of rocket and satellite measurements in the late 1950's and early 1960's triggered a rapid growth in knowledge of the magnetosphere. This rapid growth continues today. Phenomenological knowledge of the magnetosphere and the geophysical phenomena occurring within it and because of it is truly vast. On the other hand, understanding of the underlying physical processes, while developing rapidly, is still meager. In many cases, new understanding is dependent on the making of specific sets of definitive measurements in the magnetosphere. To some degree these needed measurements either will be performed or are planned for suggested experimental programs in the next five to ten years. Many of these experiments are part of a coordinated international experimental program called the International Magnetospheric Study, which began in January 1976.

The purpose of this chapter is to present current knowledge of important phenomena that occur in the magnetospheric environment and to discuss some of the outstanding questions, answerable in the foreseeable future, upon which understanding of these phenomena hinges. As will become clear, these questions concern the most basic of magnetospheric processes. Generally stated, an understanding of how the energy carried by solar plasmas and magnetic fields is transmitted through the magnetosphere and into the ionosphere and upper atmosphere is needed. This understanding requires a determination of: how the solar wind and the earth's atmosphere act to populate the magnetosphere with plasma; how this plasma is energized and transported within the magnetosphere; and how energy and momentum are finally transferred to the upper atmosphere. Those aspects of the problem relating directly to the interaction of the solar wind with the magnetosphere were treated in Chapter 1 and hence are not emphasized here.

The discussion begins at the lower boundary of the magnetosphere, where it overlaps the upper atmosphere and ionosphere, and continues with a general description of the magnetic and electric fields, plasma populations, and current systems that make up the magnetosphere. Considerable attention is then given to attempts to identify the physical processes responsible for the acceleration of charged particles and their bombardment of the atmosphere to produce the aurora and other effects. This discussion is followed by a brief summary of magnetospheric wave phenomena and an outline of the present status of research on magnetospheric substorms. A final section treats the ionosphere as an important source of magnetospheric plasma.

2.2 THE INTERFACE BETWEEN THE NEUTRAL ATMOSPHERE AND THE MAGNETOSPHERE

Throughout the magnetosphere the motions of ions and electrons are determined by their initial velocities, the direction and strength of the magnetic field in their near vicinity, and any electric fields they encounter. Ions and electrons can move freely along the magnetic-field direction. However, when they move perpendicularly to the magnetic field they experience a magnetic force that causes them to move in a circle. The radius of this circular path (the gyroradius) increases in proportion to the particle's mass and velocity and in inverse proportion to the strength of the magnetic field. Generally, then, ions move in much larger circles than the much lighter electrons so that, for a given density of background neutral gas, the electrons may be able to circulate many times in the earth's magnetic field while the ions complete less than one full circle before colliding with a neutral atom or molecule. This is the reason for the difference in altitudes for full magnetic control of electrons and ions noted above. Moreover, as elucidated below, this difference is one of the reasons why large amounts of magnetospheric energy are dissipated in the region between 90 and 120 km, where at geomagnetic latitudes above about 55° there generally exist horizontal voltage drops or electric fields of some tens of volts per kilometer. The primary source of these electric fields is the interaction between the solar wind and the magnetosphere, discussed in the previous chapter.

Above about 120 km, both electrons and ions complete many circular orbits in the magnetic field between colli-

sions with neutrals, so that they are alternately accelerated and decelerated by any electric fields transverse to the magnetic field in their near vicinity. This results in a continuous oscillation of the radius of their circular paths, as shown in Figure 1.8, Case b, of Chapter 1. The net effect is a drift that is perpendicular to both the local magnetic field and the electric field and that on the average is the same for ions and electrons. Therefore, since all the positively and negatively charged particles drift together, no currents are produced by electric fields that exist at these higher altitudes. Such bulk motion or flow of plasma in the magnetosphere is referred to as magnetospheric convection. Although the simple picture described above has to be modified in the actual nonuniform magnetic field of the magnetosphere, the concept of plasma convection in the ionosphere and magnetosphere is one of the most valuable tools for understanding phenomena that occur in the earth's plasma environment.

Below altitudes of about 120 km, the motion of the ions changes to a pattern of short segments between collisions, in which they move nearly parallel to the electric field. The current that results, known as the Pedersen current, flows through the ionosphere from the positive terminal to the negative terminal of the magnetospheric electric power source (that is, the solar-wind–magnetosphere interaction), drawing energy from it and depositing heat in the ionosphere through ion-neutral collisions. The electrons simply drift or convect as they do at higher altitudes, thereby carrying a Hall current in the $-\mathbf{E} \times \mathbf{B}$ direction, (i.e., perpendicular to both the electric and magnetic fields), which does not dissipate energy from the magnetospheric electric power source. Although Hall currents can circulate as simple eddy currents in the ionosphere, both the Hall and Pedersen currents are in fact fed by currents that flow vertically between the ionosphere and the outer magnetosphere along magnetic-field lines, as discussed in the following section.

In the 90–120 km region, the neutral wind dominates the ion flow but not the electron flow, resulting in a dynamo effect that can modify significantly the overall magnetospheric electric-field distribution.

Above 120 km, the upper atmospheric neutral wind begins to change from a speed and direction determined by meteorological effects, such as solar heat input and the rotation of the earth, toward a speed and direction determined to a greater degree by magnetospheric electric fields. The extent to which the neutral wind is affected by magnetospheric electric fields depends on a process known as ion drag, in which ion–neutral collisions tend to set the neutral gas in motion along with the convecting ion gas.

Just how the earth's high-altitude neutral and plasma winds interact and combine to produce a global circulation pattern is one of the important outstanding questions of magnetospheric physics. Simultaneous measurements of these winds and of the associated electric fields at satellite altitudes (down to about 150 km), along with remote optical determinations of the neutral winds down to near 60 km, are now being planned for the proposed Electrodynamics Explorer Project. Successful completion of these measurements, along with several ongoing ground-based and rocketborne programs involving ionospheric backscatter radars and the tracking of artificial plasma clouds, should give us a better understanding of these important phenomena and their potential for affecting lower levels of the atmosphere. These topics are discussed in greater detail in Chapter 3.

2.3 MAGNETOSPHERIC PLASMA CONVECTION AND THE AURORA

The upper atmosphere in the 100- to 200-km-altitude range is excited by the bombardment of magnetospheric electrons and ions to produce the optical emissions known as aurora. The aurora is the only visible manifestation of magnetospheric plasma processes, and as such it was the motivation for man's first studies of the magnetosphere. However, several important questions remain unanswered, including how auroral particles are accelerated in the magnetosphere to energies of several thousand electron volts. Figure 2.1 is an auroral photograph generated by a scanning photometer on an Air Force satellite as it crossed over the northern polar cap in wintertime at an altitude of 800 km. In this particular picture the solar direction is toward the top of the page, the dusk and dawn meridians toward the left and right, respectively, and the magnetic north pole is in the upper center of the picture, encircled at magnetic latitudes of 60° to 70° by a very active auroral display. For comparison, a similar photograph taken during relatively quiet conditions is shown in Figure 2.2. The region of auroral emissions encircling the geomagnetic pole, known as the auroral oval, responds to increased magnetospheric activity by expanding to lower latitudes while also spreading out in latitude. As is apparent in Figures 2.1 and 2.2, the earth's atmosphere acts much like a large television screen, which images the low-altitude footprint of various magnetospheric plasma regions. If we knew precisely how to map magnetic-field lines from the global auroral displays out into the distant magnetosphere, and understood how this mapping changes during magnetospheric disturbances or substorms, we could identify the source regions for all the auroral particles. Some successful initial attempts have been made to render high-latitude magnetic-field lines visible by firing along them jets of barium ions, which emit visible spectral lines when illuminated by sunlight. Much work remains to be done in this area, particularly in developing techniques for increasing the velocity of the barium ions to minimize the effects of their drift across field lines due to electric fields.

We strongly suspect that the dayside auroras and perhaps the highest-latitude nightside auroras occupy field lines that are connected directly to the magnetic field of the solar wind, which would therefore be the

FIGURE 2.1 Air Force satellite image of auroral displays over western Europe during a magnetically disturbed period (18:25 UT on January 5, 1973). The magnetic pole is in the upper center of the picture. The noon–midnight meridian extends from the top to the bottom of the picture with dawn and dusk to the right and left sides, respectively.

source for these auroral particles. The lower-latitude diffuse nightside auroras, which actually extend at times around to the dayside (although being very weak and not easily detected there), are connected magnetically to a vast reservoir of electrons and ions known as the plasma sheet. Figure 2.3 is a sketch of the magnetosphere illustrating the approximate configuration of a few magnetic-field lines, the plasma sheet, the polar cusp (or cleft) through which solar wind plasma flows directly into the atmosphere, and the plasmasphere, which is populated by very-low-energy (or cold) plasma, which is the upward extension of the ionosphere. The radiation belts are contained within and just beyond the plasmasphere.

The configuration depicted in Figure 2.3 is typical of conditions that exist during moderate magnetospheric activity. During quiet periods the neutral line—the line between the tail lobes that marks the last or highest-latitude closed-field lines on the nightside—is located much farther down the magnetospheric tail; the plasma sheet is greatly expanded to a thickness of some 5 to 10 R_e in the north–south direction; and the equatorial earthward edge of the plasma sheet is at a higher altitude, generally well beyond the plasmapause (the boundary of the plasmasphere). The polar caps, which apparently are magnetically connected to the interplanetary medium, become larger during geomagnetic activity. This expansion results in a buildup of magnetic flux in the magnetospheric tail, and it also involves strong plasma convection that is driven by the interaction of the solar wind with the magnetosphere, the convection penetrating to lower latitudes within the magnetosphere.

A few contours illustrating the flow or convection of

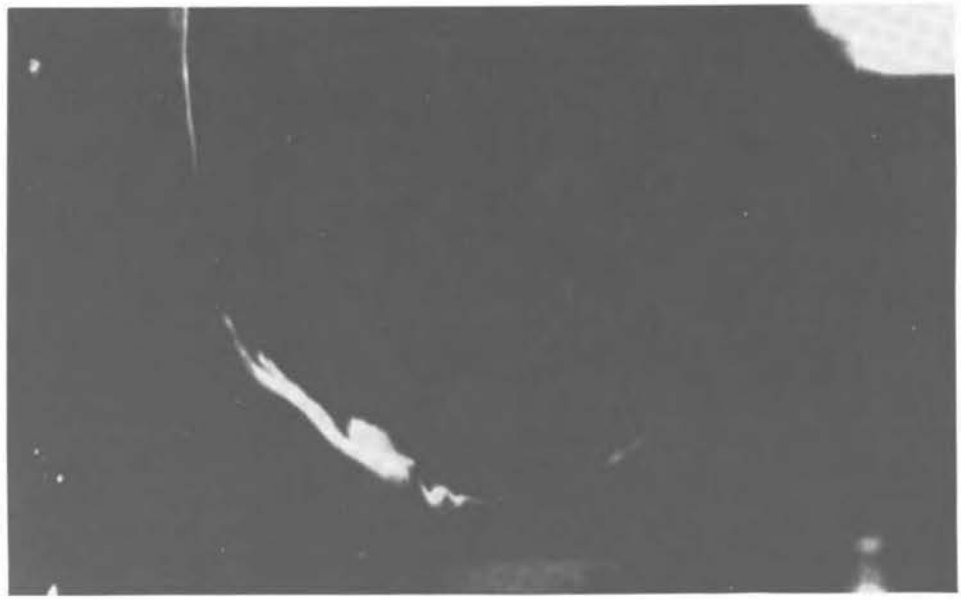


FIGURE 2.2 Same as Figure 2.1 but for a period of relative magnetic quiescence (17:57 UT on December 31, 1972).

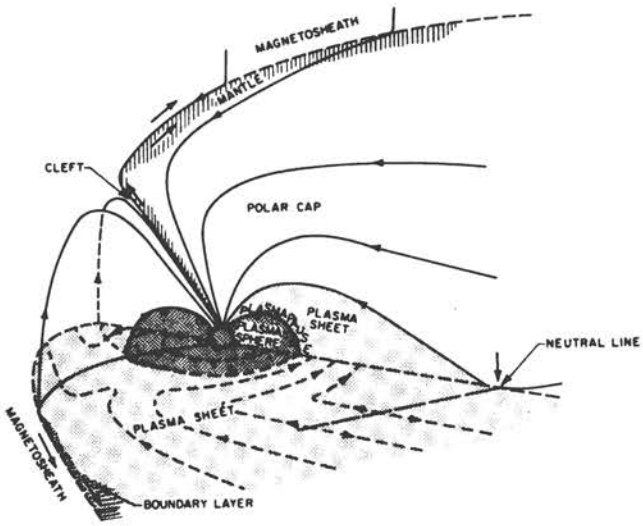


FIGURE 2.3 Sketch of magnetic-field lines and various plasma regions of the magnetosphere. Only the northern dusk quadrant is shown for simplicity. Also drawn are contours denoting the flow or convection of plasma in the equatorial plane. All magnetic-field lines shown in this figure except the two near the dusk meridian lie completely in the noon–midnight meridian plane. The Van Allen radiation belts lie within and just beyond the plasmasphere.

plasma are drawn in the equatorial plane in Figures 2.3 and 2.4. The flow on polar-cap magnetic-field lines and those through the polar cusp and the highest-latitude auroras is generally antisunward, or in the same direction as the flow of the solar wind. In the photographs of Figures 2.1 and 2.2 this antisunward flow exists in the dark region within the auroral oval, but it is concentrated toward the edges of the dark polar cap near the regions of the highest-latitude auroral emissions.

The convection of plasma on the closed-field lines of the outer magnetosphere, including the plasma sheet, is directed generally toward the sun, probably resulting in the eventual escape of plasma from the front of the magnetosphere into the solar wind. Nearer the earth, the plasma flow is dominated by the rotation of the earth. That is, frictional coupling causes the lower ionosphere to rotate along with the earth and its neutral atmosphere. This rotation involves the motion of the cold ionospheric plasma across magnetic-field lines, which can happen only if an electric field is present. The required electric field, which is produced by a polarization or rearrangement of electrons and ions in the ionosphere, extends out into the magnetosphere, there competing for control of plasma convection with the solar-wind-generated electric field. The region near the earth where plasma rotates with the earth and hence flows in roughly circular or trapped orbits defines the plasmasphere. As noted above, the plasmasphere is populated by ionospheric plasma extending upward along magnetic-field lines. During magnetospheric substorm disturbances, when the solar-wind-driven convection is stronger, the plasmasphere shrinks. This shrinkage is caused by a loss of cold plasma from its outer reaches and results in a new demand for refilling from the ionosphere. The readjustment of convection boundaries that occurs during substorms also results in the intermingling of the hot plasma-sheet and cold plasmasphere plasmas, producing important effects that are discussed in the next section.

The electric fields associated with convection are equivalent to the magnetospheric flow patterns; they generally map along magnetic-field lines into the high-latitude ionosphere, as shown in Figure 2.5, where the flow lines can be regarded as electric equipotentials. The inference of such a polar-cap flow pattern from the motion of auroras and from ground-based magnetometer measurements was the basis for our earliest concepts of magnetospheric convection (Axford and Hines, 1961). The flow patterns in Figures 2.4 and 2.5 are approximations to typical patterns, based on satellite electric-field measurements limited mostly to a single axis; satelliteborne three-axis ion-drift measurements in the ionosphere are only now becoming available for the first time on the Atmosphere Explorer spacecraft. Also, as discussed in Chapter 1, asymmetries in the overall polar-cap convection pattern, such as the concentration of flow toward the dusk hemisphere as shown in Figure 2.5, depend sensitively on the dawn–dusk component of the solar-wind magnetic field. Specifically, the pattern of Figure 2.5 is appropriate to the northern polar cap for a dusk-to-dawn component of interplanetary magnetic field, while a dawn-to-dusk component would tend to intensify the flow in the dawn region. This relationship is reversed for the southern polar cap. Because of the general "garden-hose" configuration of interplanetary magnetic fields, a dusk-to-dawn component can be interpreted as a magnetic field directed toward the sun.

The poleward boundary of the sunward flow region

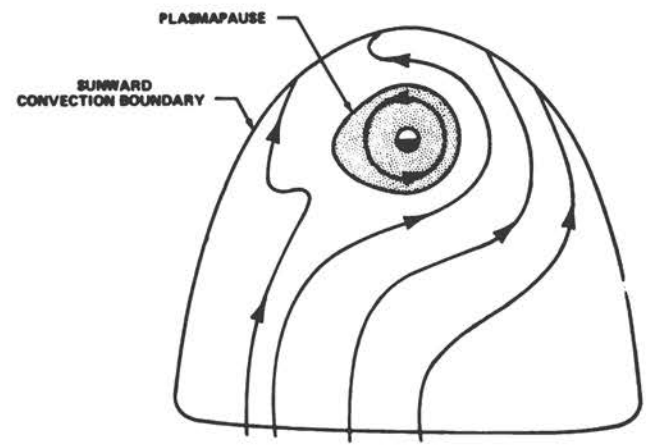


FIGURE 2.4 Idealized plasma convection pattern in the equatorial plane of the magnetosphere.

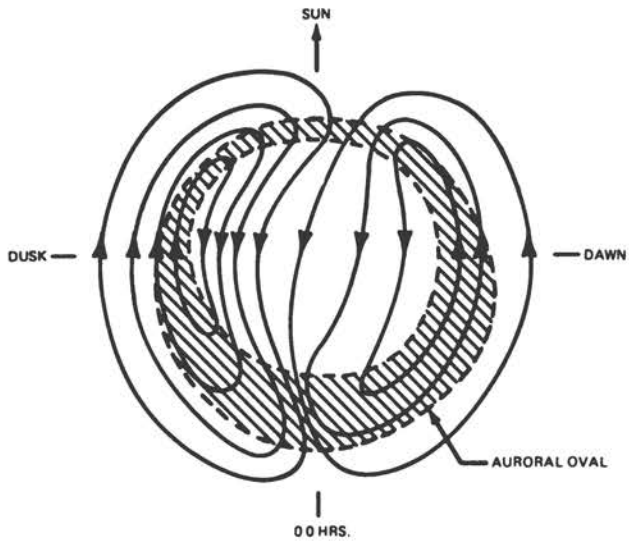


FIGURE 2.5 Idealized plasma convection pattern in the polar-cap ionosphere. The shaded region represents the average location of the auroral oval.

shown in Figure 2.5 (and the outer boundary of the sunward flow region in Figure 2.4) is an important feature of the magnetosphere about which little is known except in the ionosphere, where it maps roughly along the auroral oval. As suggested by the flow patterns sketched in Figures 2.4 and 2.5, parts of this boundary, which separates the regions of open- and closed-field lines, are penetrated by the plasma flow as it reverses from an antisunward to a sunward direction, or vice versa. These parts of the boundary are located generally near noon and in the nightside magnetospheric tail, and the term "neutral line" is sometimes used to describe them. The significance of these regions, where plasma flows between the closed- and open-field-line regimes, is discussed in detail in Chapter 1 and in Section 2.7 below.

The other parts of the antisunward–sunward convection boundary, extending around the sides of the magnetosphere, apparently exist as field-aligned walls separating regions of opposed plasma flow with at most only weak components of flow through them. As shown in Figure 2.6, this boundary between antisunward and sunward convection maps along magnetic-field lines into the ionosphere, where the convection electric fields drive currents (known as Pedersen currents) that must be fed by field-aligned currents flowing downward on the dawn side and upward on the dusk side of the polar cap. From satellite magnetic-field measurements, we know that a general field-aligned and ionospheric current system of this type exists at all times. However, we have not determined which magnetospheric plasma populations actually carry the current, except that downward-flowing auroral electrons with energies of a few thousand electron volts are the likely carriers of at least part of the upward-flowing current.

Also sketched in Figure 2.6 are oppositely directed field-aligned currents flowing along closed-field lines at lower latitudes and closing in some way across magnetic-field lines in the equatorial region of the magnetosphere. These lower-latitude currents, which also have not been identified as to the plasma that carries them, are thought to result from plasma stresses that occur in magnetospheric convection. That is, as plasma-sheet electrons and ions are convected toward the earth, they move into regions of rapidly increasing magnetic-field strength. Each individual particle then is subjected to a steadily intensifying magnetic field, which accelerates it just as in a betatron laboratory particle accelerator. But this increased velocity causes an increase in other drift velocities caused by the changing magnetic-field strength encountered by the particles in their circular paths and by the curvature of the magnetic-field lines they move along (see Chapter 1). Unlike the electric-field drift, these magnetic drifts are different for ions and electrons. The combination of electric and magnetic drifts determines the minimum geocentric distances to which electrons and ions of given energies can convect before they are diverted around the earth. An example of the asymmetries that result is the known tendency for ions to penetrate somewhat closer to the earth in the dusk hemisphere than do the electrons. The charge imbalance that results may cause the downward flow of field-aligned current in the dusk hemisphere, as shown in Figure 2.6.

Generally the high- and low-latitude field-aligned currents flow near the poleward and equatorward boundaries of the sunward convection region. The flow of the ionospheric parts of these field-aligned current systems through regions of varying conductivity determines the detailed distribution of electric fields over the polar cap and the auroral oval. Some evidence exists that the electric field is often reduced significantly inside bright auroral forms, where the conductivity is high, resulting in a pattern of plasma convection in which such regions are avoided, that is, with only weak flow within them and stronger flow at their edges. Evidence also exists for intensifications and reversals of the convective flow inside auroral forms, so the role of ionospheric conductivity

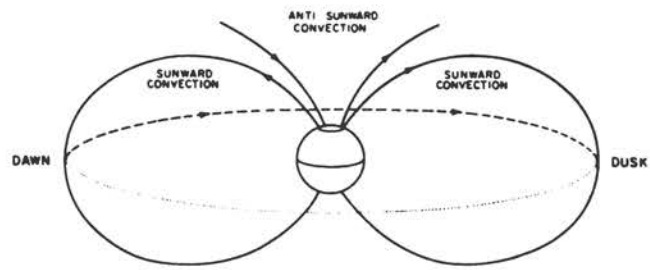


FIGURE 2.6 View from the sun of the approximate configuration of field-aligned currents in the magnetosphere. The dashed curves indicate the possible closure of the currents in the nightside equatorial plane.

in ionospheric and magnetospheric convection is by no means clearly understood.

As pointed out in the preceding section, the actual pattern of plasma convection in the magnetosphere depends to a large extent on the details of the coupling among the magnetosphere, the ionosphere, and the neutral atmosphere. Much theoretical and experimental effort is now being directed toward a better understanding of this coupling. The needed experimental techniques are now available, and the next few years should see much progress in this area.

2.4 WAVE–PARTICLE INTERACTIONS AS LOSS PROCESSES FOR ENERGETIC MAGNETOSPHERIC PLASMA

As noted in the preceding section, the atmosphere is continually bombarded by energetic electrons and ions from the magnetospheric plasma. This loss or precipitation of particles from the plasmas that exist in the magnetosphere is particularly efficient in the polar cusps and in the plasma sheet where it approaches the ionosphere. The processes responsible for this precipitation are not understood, although the growth of instabilities involving interactions between charged particles and plasma waves are thought to be of central importance. A brief digression into the concept of an atmospheric loss cone will help to elucidate these phenomena.

The motion of a charged particle as it moves along and spirals about magnetic-field lines can be described at any point by two velocities, $v_\parallel$ and $v_\perp$, or its components of velocity along and perpendicular to the magnetic-field direction. The pitch angle α of a particle is the angle between its total velocity vector and the magnetic-field direction, so that $\tan \alpha = v_\perp/v_\parallel$. As a particle travels down a magnetic-field line toward the atmosphere, it experiences a steadily increasing magnetic-field strength. Just as in the case of a particle being convected earthward in the equatorial plane, this increasing magnetic-field intensity is associated with an increase in the particle's perpendicular energy, that is, an increase in $v_\perp$ so as to keep constant its magnetic moment ($mv_\perp^2/2B$, where m is the particle mass and B the magnetic field strength). However, in this case there is no electric field or other mechanism forcing the particle to lower altitudes, so the only energy available is the particle's own parallel energy. Therefore, $v_\perp$ increases at the expense of $v_\parallel$, the ratio of the perpendicular energy to the magnetic-field strength remaining constant until $v_\parallel$ is decreased to zero. At this point, the particle mirrors, or reverses its direction, and moves back up the field line decreasing its perpendicular energy as it goes. Those particles whose mirror points are at altitudes of less than about 100 km are said to populate the loss cone, since they are effectively absorbed by the atmosphere after a single transit down the field line. While, by definition, the loss cone at 100 km includes all downward-moving particles, it shrinks rapidly toward higher altitudes where the magnetic field is much weaker. For example, near the equator at 6 R_e geocentric distance, the half-angle of the loss cone is about 3°. Farther out, in the distant plasma sheet, only those particles moving within 1° of the local magnetic-field direction will strike the atmosphere.

Low-altitude observations in the auroral oval show that electrons and ions in the few hundred eV to several keV energy range normally exhibit nearly isotropic angular distributions, that is, comparable fluxes at all downward-directed pitch angles. In addition, the localized occurrence of field-aligned angular distributions of electrons and ions suggests that a moderately low-latitude acceleration process exists that increases $v_\parallel$. Possible acceleration processes will be discussed in detail in the next section. In this section, we turn our attention to the possible mechanisms responsible for the nearly isotropic angular distributions.

The maintenance of a full (or isotropic) loss cone requires some strong process of pitch-angle scattering to exist somewhere along the auroral–oval field lines. The intense electromagnetic- and electrostatic-wave noise that is commonly observed along auroral–oval field lines suggests that pitch-angle scattering by waves may well be responsible. The theoretical and experimental work that needs to be done to confirm this is quite imposing. Sensitive electric- and magnetic-wave detectors and particle detectors capable of scanning through the small loss cones that exist at very high altitudes must obtain simultaneous measurements throughout the auroral–oval regions and along the entire length of auroral-field lines. Some programs now under study may satisfy these requirements; measurements to date have obtained fragmented, yet significant, information.

One situation that is favorable for exchange of energy between waves and charged particles exists when the particles experience oscillating electric fields near the frequency at which they spiral in the magnetic field (i.e., at their cyclotron frequency or gyrofrequency). A highly successful first-order theory of this type of interaction was developed by C. F. Kennel and H. E. Petschek. Their work is reviewed by Kennel (1969) and by Fredricks (1975). The theory involves the interactions of electrons and positive ions with electromagnetic whistler-mode (i.e., electron–cyclotron) and ion–cyclotron waves, respectively. These waves are circularly polarized with electric-field vectors that rotate in the plane perpendicular to the magnetic-field direction. The interaction, which is strongest near the equator, involves electrons or ions whose parallel velocities ($v_\parallel$) Doppler shift the frequencies of existing waves to frequencies near their own gyrofrequencies. The first-order calculations showed that resonant interactions of electrons with whistler-mode waves and of ions with cyclotron waves can occur for parallel energies above the value given approximately by the energy density of the magnetic field divided by the total plasma density. For sufficiently anisotropic angular distributions that increase toward larger pitch angles,

unstable wave growth occurs through the conversion of particle energy to wave energy. Approximations to the second-order theory showed that the waves in turn act to diffuse the particles in pitch angle such that the parallel energy is increased at the expense of perpendicular energy. A reduction in total particle energy also occurs but is important only at the very lowest energies.

The Kennel and Petschek theory was apparently successful in predicting that electromagnetic cyclotron interactions would act to limit the flux of trapped particles by precipitating resonant ions and electrons. Also, because of the dependence of the minimum resonant parallel energy on the plasma density, the region just inside the plasmapause was indicated as a likely place for this interaction to be of importance. J. M. Cornwall, F. V. Coroniti, and R. M. Thorne suggested that such interactions occurring in the region of overlap of the plasmasphere and the plasma sheet could result in damping of the waves by plasmaspheric electrons and the subsequent transfer of energy downward to the ionosphere. The resultant heating was postulated to be the cause of the stable auroral red (SAR) arcs that extend over wide swaths of longitude at subauroral latitudes during the recovery phase of many magnetic storms. Some experimental evidence for the positive-ion precipitation that should result from cyclotron interactions near the plasmapause has been found from low-altitude polar-orbiting and high-altitude equatorial-orbiting spacecraft. However, while the theory assumed that the electromagnetic ion-cyclotron interactions would result in full loss cones (i.e., strong pitch-angle diffusion), the measured ion fluxes fill only the edges of the loss cone, indicating the existence of only weak to moderate pitch-angle diffusion (see review by Williams, 1975).

Waves produced by cyclotron interactions near the plasmapause can also have important consequences in the dynamics of other particle populations in other parts of the magnetosphere. Recently L. R. Lyons, R. M. Thorne, and C. F. Kennel have suggested that whistler-mode waves in the few-hundred-hertz frequency range that are generated by energetic particles near the plasmapause may propagate across magnetic-field lines, filling the plasmasphere with plasmaspheric hiss, which has been observed. Subsequent interactions between the plasmaspheric hiss waves and trapped high-energy electrons well within the plasmasphere should result in an efficient precipitation loss of electrons that can explain the slot or gap that exists between the inner and outer electron radiation belts. Figure 2.7 illustrates the Lyons *et al.* mechanism schematically. In the left-hand part of Figure 2.7 are shown examples of 500-Hz whistler-mode ray paths as they fill the plasmasphere after emanating from the region of overlap between the outer radiation belt and the plasmasphere. Three typical electron trajectories in the slot region are shown in the right-hand portion of Figure 2.7 along with notations of where the various pitch-angle scattering processes (Landau damping and first- and higher-order cyclotron interactions) should predominate. The fluctuating electric and magnetic fields and the stronger convection that occur during magnetospheric disturbances (or substorms) act to repopulate the slot region with energetic electrons, this repopulation being followed in several days by the reappearance of the equilibrium (or slot) configuration. Detailed comparison of this theory with measured electron fluxes and pitch-angle distributions and plasmaspheric hiss intensities indicates that it can in fact explain quantitatively the separation of trapped electrons into inner and outer belts, which was one of the first magnetospheric observations of the satellite era.

Because of the low plasma density beyond the plasmapause, conditions are not expected to be favorable for the occurrence of the electromagnetic cyclotron instability in the body of the plasma sheet, where strong pitch-angle diffusion is known to occur. Attention has therefore been focused on electrostatic instabilities, which have

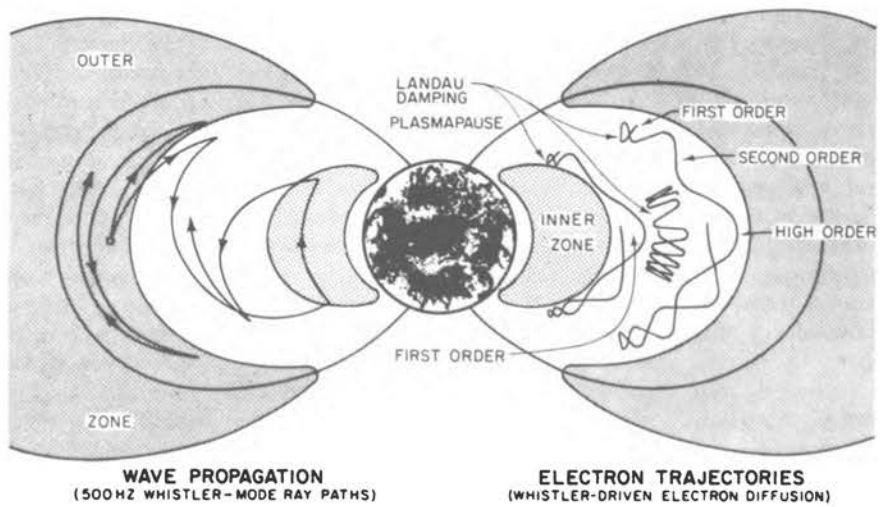


FIGURE 2.7 Schematic representation illustrating the theory of Lyons *et al.* (1972) developed to explain the existence of the slot between the inner and outer zones of trapped electrons. The left-hand side of the figure depicts the propagation of waves generated in the region of overlap between the outer zone and the plasmasphere so as to fill the plasmasphere with whistler-mode noise (plasmaspheric hiss). The right-hand side of the figure illustrates some typical trajectories of trapped electrons in the slot. The various modes of interaction between the waves and the electrons (Landau damping and first- and higher-order cyclotron interactions) that result in pitch-angle diffusion and loss of electrons through precipitation, or bombardment of the atmosphere, are noted alongside the electron trajectories.

been related to measurements of precipitating electrons in the keV energy range (see review by Fredricks, 1975).

The ability of the magnetosphere to amplify waves efficiently through wave–particle interactions has recently been strikingly demonstrated in the experiments conducted by R. A. Helliwell and co-workers at Siple Station, Antarctica. Located near the foot of a magnetic-field line that crosses the equatorial plane at about 4 R_e geocentric distance, Siple Station is well suited for the study of phenomena associated with the plasmapause. The injection from Siple of man-made whistler-mode waves of frequency 300 Hz to 30 kHz has resulted in their amplification by as much as 30 dB and in the triggering of waves at other frequencies. The electron-cyclotron interactions responsible for the amplification also result in the precipitation of energetic electrons, which have been detected indirectly by the x rays they produce on striking the atmosphere. Similar wave amplification and triggered emissions have been observed to result from waves inadvertently launched into the magnetosphere at high harmonics of the 60-Hz electric power that is generated at Canadian power stations near the other foot of the Siple field line at Roberval, Canada. While not so awe-inspiring as the high-latitude auroral displays, the effects of energetic particle precipitation from the radiation belts that may result from wave–particle interactions along field lines near and below the latitude of Siple Station are central to our understanding of upper-atmospheric and ionospheric phenomena at low and middle latitudes. A comprehensive review of the subject has been written by Paulikas (1975).

2.5 ACCELERATION PROCESSES FOR AURORAL PARTICLES

In addition to scattering by waves, which primarily changes a particle's pitch angle with little change in its energy, electrons and ions can also be precipitated by acceleration processes that act preferentially to increase $v_\parallel$ and not $v_\perp$. There is little doubt that such processes exist, although identifying and understanding them will require considerably more theoretical and experimental effort. In the various plasma regimes of the magnetosphere, the plasma is often describable by electron and ion temperatures, and these generally are somewhat different. Many measurements in the outer magnetosphere do exhibit a thermal, or Maxwellian, type of energy distribution, upon which convective flow velocities are often superimposed. However, as described above, these measurements generally are not of the loss-cone particles that impact the atmosphere. Low-altitude measurements above the aurora, on the other hand, reveal the common occurrence of electron distributions describable as thermal populations that have fallen through electric potential drops of several kilovolts parallel to the magnetic-field lines. Angular distributions showing strong peaks along the field line (reviewed by Arnoldy, 1974), and the apparent trapping of atmospheric secondary and backscattered primary electrons between the atmosphere and the assumed potential-drop or parallel-electric-field region, have provided mounting evidence for the existence of such acceleration/precipitation processes (see review by Evans, 1975).

Since the effects of these electron acceleration processes are observed to exist in the same general regions as the field-aligned currents above the auroral oval, attempts to explain the parallel electric fields have recently centered on processes resulting from the flow of currents through the ionospheric and magnetospheric plasmas. As noted above, electrons and ions can move freely along magnetic-field lines. Since collisions become unimportant above about 120-km altitude, the magnetospheric magnetic-field lines are generally assumed to possess zero electrical resistance, so that any potential drops along them are quickly shorted out and therefore are not maintainable. However, limits are expected on the amount of current flow that the ionospheric and magnetospheric plasmas can support. The current-driven instabilities that can result from such current overloading of a plasma have been suggested as possible mechanisms for producing field-aligned voltage drops, or electric fields, in limited regions above the auroral zones.

One such mechanism, discussed in detail by Block (1975), has been observed to develop in collision-dominated laboratory plasmas (actually gaseous discharges) when the discharge current increases to the point where the electron drift speed equals the thermal speed of the electrons. When this threshold is exceeded, rather thin layers appear in the plasma across which significant voltage drops occur. These layers are called double layers because they consist essentially of two parallel layers charged to opposite polarities, much like a parallel-plate capacitor. In the magnetosphere, conditions for the existence of double layers are most favorable at altitudes of a few thousand kilometers, although they might occur over a wide range of altitudes. An electron or ion moving down the field line can be accelerated, decelerated, or reflected by such a double layer, depending on the direction of the parallel electric field and its magnitude.

A major competing theory also involves a current-driven instability. The theory involves a process known as anomalous resistivity, which is thought to arise through interactions between a current-carrying plasma and various plasma waves, the electrostatic ion-cyclotron mode being one possibility. The threshold current required for anomalous resistivity is of the same order as that required for double-layer formation, both being approximately 10^{-5} A/m^2. Anomalous resistivity, predicted to occur at altitudes near 1000 km, acts to dissipate electrical energy, thereby heating the ambient plasma through which the field-aligned current flows. Just as in the flow of current through an ordinary resistor, the flow of field-aligned current through the anomalous resistance results in a field-aligned voltage drop. Particles moving through the

anomalous-resistivity region, which is predicted to extend over much greater distances along the field lines than an individual double layer, would experience an increase or decrease in parallel energy equal to the voltage drop traversed.

Crucial to our understanding of the potentially important role of such low-altitude auroral particle acceleration mechanisms are (1) the measurement of parallel electric fields; (2) the determination of their location and the vertical extent of the region over which they act; (3) the determination of which plasma populations carry the field-aligned currents that set up the parallel electric fields; and (4) the measurement of the wave modes involved in the current-driven instabilities and any plasma heating that results. Efforts are under way to make measurements in all of these areas, including an experiment designed to provide answers to (1) and (2) above by firing electron beams up field lines where they may be reflected by existing parallel electric fields.

The recent discovery that the earth is a strong emitter of kilometer-long (about 10^5 Hz) radio waves during magnetospheric disturbances may have important implications in the study of field-aligned electric fields. Using the direction-finding capabilities of the satelliteborne antennas that detected the kilometric radiation, D. A. Gurnett and co-workers have placed its source at an altitude of about 6000 km above the eveningside auroral oval. This result suggests a possible connection with the wave turbulence involved in the instabilities that may act to set up parallel electric fields.

2.6 LOW-FREQUENCY HYDROMAGNETIC WAVES

The discussion of the two preceding sections has centered on the interactions between magnetospheric plasma populations and waves with frequencies near the electron or ion gyrofrequencies. In addition to these gyroresonant, or cyclotron-resonant, phenomena, the magnetosphere apparently also exhibits a number of other important resonances at lower frequencies. Such resonances are thought to be associated with periodic phenomena such as transverse (or guitar-string-type) oscillations of individual field lines loaded with plasma and compressional oscillations of the entire magnetospheric cavity. Hydromagnetic waves of this type, known as geomagnetic micropulsations, have been studied with ground-based magnetometers for many years (Jacobs, 1970). Within the past few years, significant new information has been obtained with latitudinal chains of magnetometer stations (Lanzerotti and Fukunishi, 1974). Observations of similar waves have been made by satelliteborne magnetometers (McPherron *et al.*, 1972). The relationship between the waves observed in the magnetosphere and those observed on the ground is often difficult to determine, since the waves are modified significantly in their propagation through the ionosphere.

Recent results from the few existing magnetometer chains have supported the idea that surface waves generated at the magnetopause by solar-wind pressure variations or by the Kelvin-Holmholtz (like wind-over-water) instability can propagate deep within the magnetosphere. Near the plasmapause, where the plasma density increases rapidly, these waves may strike a local field-line resonance. Such localized field-line resonances have in several cases been indicated by ground-based magnetometer chains as reversals in the sense of polarization of the magnetic-field oscillations across the latitude of the resonant field line.

Resonant oscillations should also result from waves generated by wave–particle interactions, such as those discussed in the previous two sections, and in other plasma instabilities occurring in the magnetosphere. Comprehensive observations involving an expanded magnetometer network (chain) and ground–satellite correlative measurements are required to confirm our ideas about these important resonance phenomena in the magnetosphere. It is hoped that significant progress will be made in this area through the cooperative experiments that are planned for the International Magnetospheric Study, now under way.

2.7 MAGNETOSPHERIC SUBSTORMS

Several times a day, on the average, the nightside of the magnetosphere experiences a large-scale disturbance that we call the magnetospheric substorm. The ground signatures of substorms are dramatic at high latitudes. Previously narrow and stable auroral arcs, such as those shown in Figure 2.2, become disrupted in a region near midnight, brighten rapidly, and surge poleward, often filling the whole sky within 10 minutes or so. The global display of auroras during a strong substorm is extensive and highly structured, as in Figure 2.1. As a substorm develops, the ionospheric currents that flow in the auroral oval intensify rapidly, mainly because of the increased conductivity caused by the auroral particle bombardment. The power dissipations due to heating by ionospheric currents and by particle bombardment are of similar magnitude, each being between 10^{11} and 10^{12} W. The initial 10- to 20-min intensification (or expansion phase) of substorms is followed by a recovery phase lasting about 1 hour, during which quiet conditions are again approached. Often the recovery phase of one substorm is interrupted by the expansion phase of a new one, with a rapid sequence of substorms resulting in a magnetic storm, as described below.

The concept of a global or magnetospheric-wide substorm outlined above is a valuable one for the study of processes responsible for the dissipation of large amounts of magnetospheric energy. The actual physical phenomena involved may, however, be operative at virtually all times, although often weak and highly localized. These phenomena include the processes of auroral parti-

cle precipitation, field-aligned currents, and plasma convection discussed above. Also, the aurora, which rarely or never disappears altogether, commonly undergoes substorm-type disruptions over small regions, as evident in Figure 2.1. The essential differences between these small disturbances and the magnetospheric-wide substorms, if any, are the subject of much current debate.

The phenomenon thought by many researchers to be responsible for the substorm is a large-scale instability of the neutral sheet in the magnetospheric tail. The extended tail of the magnetosphere is maintained in part by a current that flows from dawn toward dusk in the neutral sheet—the region near the center of the plasma sheet, which separates domains where the magnetic field is directed toward and away from the earth. Field lines cross the neutral sheet at all distances earthward of the neutral line, which is the equatorial boundary between open and closed field lines (see Figure 2.3). The neutral sheet earthward of the neutral line is considered to be unstable to perturbations such as localized current flows, which tend to form new neutral lines. The formation of a new neutral line within the neutral sheet results in strong plasma flow into it from above and below the neutral sheet and in the creation of a new equatorial boundary

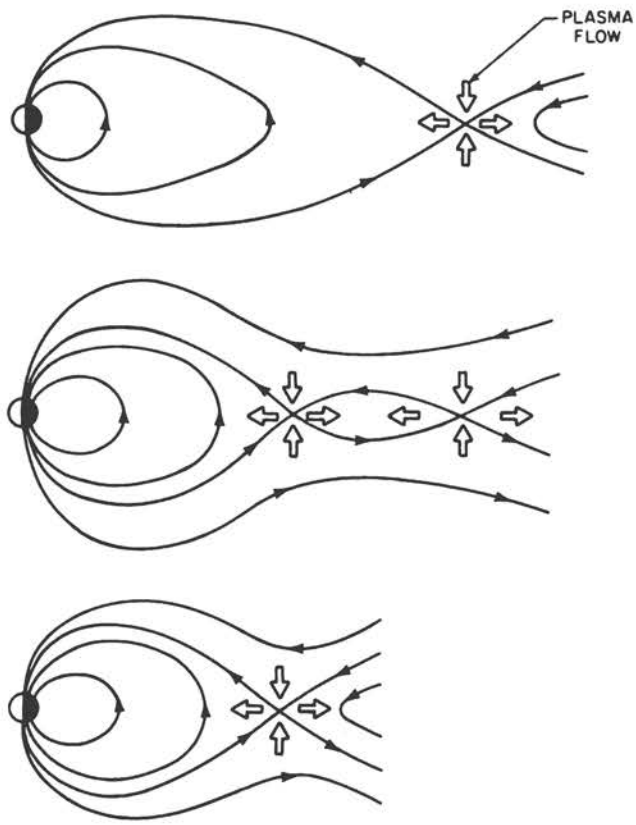


FIGURE 2.8 Possible sequence of events in the magnetospheric tail during substorms, resulting in the formation of a neutral line closer to the earth than during quiet time.

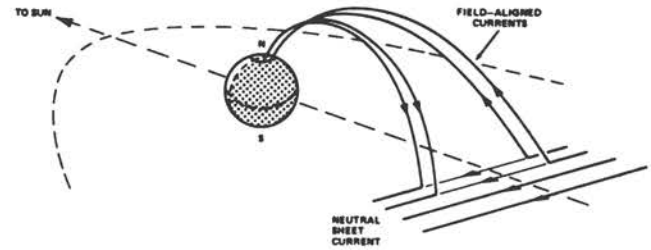


FIGURE 2.9 Suggested system of field-aligned currents associated with magnetospheric substorms, perhaps associated with the formation of a new neutral line, as in Figure 2.8.

between open- and closed-field lines closer to the earth, while remaining on the same field line (see Figure 2.8). One result of the formation of a new neutral line closer to the earth would be a relaxation of the stretched-out magnetospheric tail. This requires a weakening or diversion of the neutral-sheet current. It has been suggested that the neutral-sheet current may be short-circuited through the auroral–oval ionosphere by field-aligned currents, as sketched in Figure 2.9. Magnetic effects consistent with such field-aligned currents have been observed during substorms.

Magnetic perturbations and plasma flows consistent with the formation of neutral lines at geocentric distances of about 10–30 R_e in the tail have been observed near the beginning of the expansion phase of substorms. As the substorm recovers, this neutral line seems to move quickly back down the tail, although confirmation of this must await dual-spacecraft observations as planned for the International Sun–Earth Explorer Project mentioned above.

A sketch of the plasma flow in the vicinity of a neutral line is shown in Figure 2.8. The plasma flows in directions consistent with an electric field directed generally dawn to dusk. Very near the neutral line, some process must act to convert magnetic energy to plasma kinetic energy. This overall process, which is referred to as magnetic merging or reconnection, was discussed in more detail in Chapter 1. As pointed out in the review by Vasyliunas (1975), an understanding of the reconnection process is one of the greatest challenges facing researchers in magnetospheric physics and in solar physics as well, since it may also be responsible for energy conversion in solar flares (Akasofu and Chapman, 1972).

A possible consequence of the intense earthward plasma flow associated with the formation of neutral lines is the strong injection of plasma across the nightside sector of geostationary orbits (near 6.6 R_e geocentric distance) that accompanies the onset of the substorm expansion phase. These injections are observed to occur over a limited longitudinal sector near midnight during small disturbances but to spread to beyond the dusk meridian in strong substorms. Subsequently, a part of the substorm-injected plasma begins to form a ring current extending around the dusk hemisphere of the earth. A

rapid succession of substorms provides a ring-current source mechanism that exceeds the loss mechanisms (precipitation and charge-exchange interactions with the neutral hydrogen populating the same region). During the buildup of the ring current, the 10- to 50-keV positive ions that are its main contributors are thought to drift out through the dayside magnetopause since the convection electric fields are generally more intense during this time, which is referred to as the main phase of a magnetic storm. As the substorm activity recovers and the convection electric field relaxes, the ions carrying this partial ring current find themselves on trapped orbits that encircle the earth to form the symmetric ring current. This current decays over a period of a few days, which is called the magnetic-storm recovery phase. Since the magnetic-storm ring current exists at geocentric distances of 3 to 6 R_e, its effects are felt worldwide.

Although our understanding of the substorm as the source mechanism for magnetic storms is on fairly firm ground, the details of how the solar wind interacts with the magnetosphere to produce the substorm itself remains a mystery. The candidate physical processes were discussed in Chapter 1. As noted there, the existence of a southward component (that is, antiparallel to the low-latitude geomagnetic field) in the solar wind results in an increased energy input to the magnetosphere and in the occurrence of stronger and more extensive substorms. In other words, the size of a substorm depends on the overall magnetospheric conditions that exist at the time the neutral-sheet instability is initiated. These conditions depend in turn on what the conditions in the solar wind (particularly the north–south component of its magnetic field) have been for the previous half hour or so. Magnetospheric conditions favorable for large substorms include an expanded polar cap, a reduced total magnetic flux in the dayside of the magnetosphere, a reduced distance to the dayside magnetospheric boundary, a more extended magnetospheric tail, an increased magnetic energy density in the tail, and a thinner plasma sheet. The high probability for the occurrence of strong substorms that exists under such conditions has led to the identification of the large-scale magnetospheric reconfiguration that begins to evolve when southward components appear in the solar-wind magnetic field as the growth phase of substorms (referred to as tail growth in Chapter 2). Whether these effects are really a necessary element of substorms has been and is the subject of much controversy. Comprehensive reviews of substorm phenomena have been published by Russell and McPherron (1973), Rostoker (1972), and Akasofu and Chapman (1972).

2.8 TRANSPORT OF ENERGETIC PARTICLES INTO THE RADIATION BELTS

The plasma in the outer magnetosphere is generally thought to be the ultimate source of the energetic particles that populate the earth's radiation belts. The injection is sporadic, resulting apparently from the acceleration and earthward flow of particles and the turbulence that is associated with substorms and magnetic storms. A number of models have been constructed that can explain how fluctuating convection electric fields can produce a random walk of particles toward the earth known as inward radial diffusion. However, no satisfactory quantitative theory of diffusion has been proposed. A number of the proposed mechanisms involve a mass-dependent diffusion rate, with the heavier ions such as helium and oxygen diffusing somewhat more slowly than protons. For this reason, the heavier particles are viewed as valuable tracers for the investigation of radial diffusion processes.

In the case of the outer-belt electrons, magnetic storms produce a rapid increase of several orders of magnitude in the fluxes at the lower energies (sub-MeV), along with a loss at the higher energies. No satisfactory explanation of this apparent injection of low-energy electrons has been proposed.

Significant progress has been made in understanding the source mechanism for the energetic protons that populate the inner radiation belt. A mechanism proposed long ago, which subsequently fell into disfavor, is now recognized as able to explain the source of the protons with energies above about 30 MeV in the inner belt. This mechanism is known as cosmic-ray albedo neutron decay. It results from the interaction of cosmic radiation with the upper atmosphere, which produces secondary neutrons that can travel back toward the magnetosphere. The finite lifetime of each of these albedo neutrons for decay into a proton, an electron, and a neutrino then provides a sufficient source of energetic protons in the proper location to populate the inner belt as observed.

2.9 THE IONOSPHERE AS A SOURCE OF MAGNETOSPHERIC PLASMA

It has generally been thought that the more energetic magnetospheric plasma originates in the solar wind, while the cooler plasma with energies of a few electron volts to tens of electron volts has escaped from the ionosphere along magnetic-field lines. Recent measurements by R. G. Johnson, R. D. Sharp, and E. G. Shelley of the precipitation of energetic O^+ and He^+ ions along mid-latitude and high-latitude field lines during magnetic storms have demonstrated the need for a reassessment of these ideas. In Chapter 1, evidence for the solar-wind origin of the plasma in the polar cusp and plasma sheet was reviewed. Here we will concentrate on various mechanisms through which the ionosphere is known or thought to be a source of plasma for the magnetosphere.

Figure 2.10 shows schematically a number of flow regimes of ionospheric plasma. Over the poles, the low plasma pressure that exists on open-field lines results in an upward flow of H^+ and He^+ out of the ionosphere. This polar wind, which has been observed at altitudes of a few

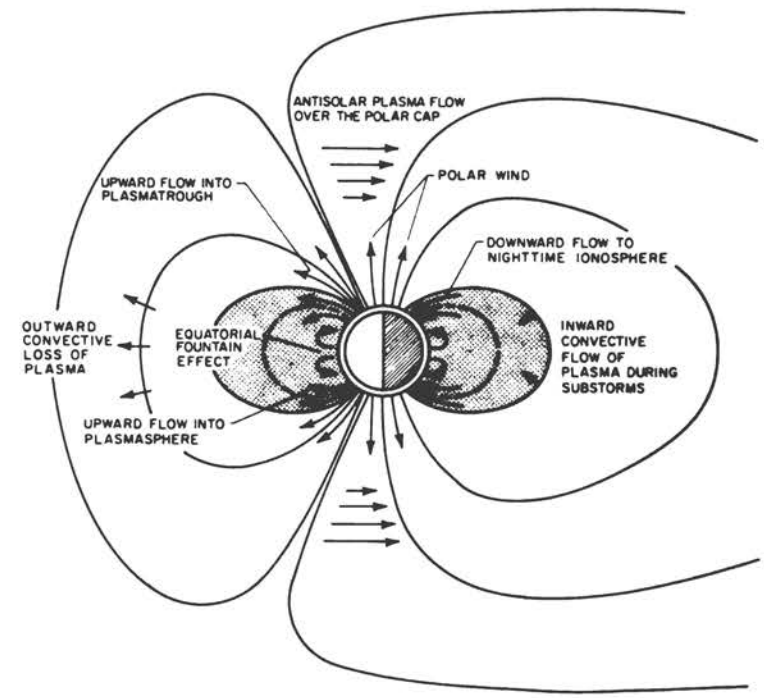


FIGURE 2.10 Flow regimes of cold plasma between the ionosphere and the magnetosphere. (Adapted from a figure by C. R. Chappell, NASA Marshall Space Flight Center, private communication.)

thousand kilometers to flow upward at velocities of several kilometers per second, is a potentially important source of plasma for the plasma sheet. Its subsequent convection into the plasma sheet and its expected energization as it flows back toward the earth on closed-field lines may be an explanation for the precipitation of energetic He^+ ions. Confirmation of this idea could come soon with the measurement of energetic plasma-sheet ions that is planned for the International Sun–Earth Explorer mission. Another possible origin of energetic He^+ and O^+ in the plasma sheet is the existence of electrostatic acceleration mechanisms discussed in the previous section. That is, a low-altitude parallel electric field that accelerates electrons downward into the ionosphere may also accelerate positive ions upward into the magnetosphere.

At lower latitudes, upward flow of plasma also results from solar heating of the ionosphere and from the low plasma pressure that exists on field lines threading the plasma sheet and the outer part of the plasmasphere. This upward flow acts continually to replenish the plasmasphere as it recovers from the sequence of events that apparently occur during substorms. As suggested by Figure 2.10, the downward flow of plasma from the nightside plasmasphere into the dark ionosphere may explain the maintenance of ionization in the nightside F-region.

Identification of the physical processes involved in the flow of plasma between the ionosphere and the plasmasphere requires that flow, temperature, and density measurements be made along the magnetic-field lines of the plasmasphere. Such measurements are planned for the Electrodynamics Explorer mission now under study.

2.10 SUMMARY AND CONCLUSIONS

The magnetosphere is now recognized as a key element of man's near-earth environment that determines how the energy carried by solar plasmas is deposited in the upper atmosphere. As discussed above, the complex interactions that occur among the neutral atmosphere, the ionosphere, and the magnetosphere must be understood before we can gain a complete understanding of the individual behavior of any one of these three elements of our environment.

In the limited space available here, it was not possible to discuss in detail a complete list of the important outstanding problems of the earth's magnetosphere. Instead, the discussion has focused on a limited number of the outstanding problems in magnetospheric research on which significant progress is anticipated in the foreseeable future. The list of important unanswered questions includes the following:

1. To what extent do the circulation patterns of the magnetospheric and ionospheric plasmas couple into those of the neutral atmosphere?
2. What plasma populations carry the field-aligned currents that transfer electrical energy from the magnetosphere to the ionosphere?
3. By what processes are the electrons and ions that excite the aurora accelerated to energies of several thousand electron volts?
4. What plasma processes are responsible for scattering magnetospheric electrons and ions into the atmosphere?

5. What processes act to populate the radiation belts with energetic particles?
6. By what means does the magnetosphere act as a source, amplifier, and resonant cavity for plasma waves?
7. What large-scale plasma instabilities are responsible for magnetospheric substorms?
8. To what extent and through what processes does the earth's ionosphere act as a source of magnetospheric plasma?

Other important questions relating more directly to the interaction of the solar wind with the magnetosphere have been treated in Chapter 1. These include questions concerning the entry of solar-wind plasma into the magnetosphere and the process that drives the large-scale circulation or convection of plasma in the magnetosphere and ionosphere.

The difficulties involved in efforts to answer these questions are due in large part to the tremendous range of scale sizes involved in magnetospheric phenomena and the problems encountered in relating measurements made at a few points to plasma phenomena that at times involve whole regions of the magnetosphere. To overcome these difficulties we must make coordinated simultaneous measurements of the appropriate plasma and electric- and magnetic-field parameters at a number of carefully chosen locations in the magnetosphere, using ground-based and balloonborne observatories and sounding rocket and satellite payloads. Such investigations, which are now being planned for the International Magnetospheric Study (1976–1979), the International Sun–Earth Explorer Project (1977–1979), and the Electrodynamics Explorer Project (about 1979–1981), promise to increase significantly our knowledge and understanding of magnetospheric processes.

Ultimately, theories should be tested wherever possible by introducing known inputs to the magnetosphere and observing the effects produced. This approach is necessary since many magnetospheric plasma phenomena tend to occur together, often preventing separation of cause from effect. Very specific plans for such input–output experiments using particle and plasma accelerators and the release of plasma clouds to probe the magnetosphere and stimulate some of its important processes are now being developed in the AMPS (Atmosphere, Magnetosphere and Plasmas in Space) Space Shuttle program. It is now possible to perform definitive experiments. A much better understanding of our near-earth environment may be expected in the foreseeable future. At the same time, initial exploration of the planets and in development of more sophisticated remote-sensing techniques for studying the sun and other astrophysical objects have begun. Many of these investigations also involve the study of large-scale magnetized plasmas, although the densities, temperatures, and magnetic-field strengths involved are often very different from those found in the magnetosphere. Nevertheless, knowledge of magnetospheric plasma processes has already proved to be of immense value in analyzing the initial measurements made in the magnetospheres of Jupiter (Formisano, 1975) and Mercury (Ness *et al.*, 1975).

Similarly, in solar physics research, a major outstanding question concerns the processes responsible for solar flares, in which large amounts of magnetic energy are somehow rapidly and efficiently converted to plasma kinetic energy. Although the densities, temperatures, and magnetic-field strengths are much lower, we have seen how the outer magnetosphere on the nightside of the earth exhibits this same capability for energy conversion. Moreover, it has been noted (Akasofu and Chapman, 1972; Akasofu and Lanzerotti, 1975) that the configuration of the magnetic fields in the two regions are similar, giving some hope that once we understand through direct observation how this important process works in the magnetosphere we can more effectively attack the difficult problem of understanding the processes that occur in the sun.

Finally, an ambitious program of laboratory research into the behavior of magnetized plasmas is under way. Among the phenomena under study are the occurrence of instabilities that result in the growth of plasma waves, the acceleration of particles to extremely high energies, and the rapid motion of plasma across magnetic-field lines. As discussed above, the same types of instabilities occur on a larger spatial scale in the low-density plasma of the magnetosphere. Therefore, while specific experimental results obtained in the magnetosphere cannot in themselves provide answers to the problems of laboratory plasma research, any advances in theoretical magnetospheric plasma physics that they make possible will be of value to both fields. It is important, then, that a strong theoretical program be maintained in which the unique experimental data available in the magnetosphere are effectively translated into models that are more directly applicable to laboratory plasma physics, as well as to planetary physics, solar physics, and astrophysics.

Initial exploratory programs have produced exciting discoveries and have provided a broad base of knowledge of the various magnetospheric plasma populations and how they interact in the magnetic and electric fields of the magnetosphere. Our understanding of these phenomena is in an embryonic but rapidly developing state. We know what questions to ask and what measurements must be made to answer them. The technology required is largely developed, and several very specific experimental programs are now under way or under study. The strong international cooperation that has evolved in these programs and in magnetospheric studies in general gives us great expectations that the next 10 years will see man gaining a confident understanding of the physics that controls his near-earth environment and making significant progress in the basic physics of collisionless plasmas.

ACKNOWLEDGMENTS

I am grateful to my colleagues of the Magnetospheric and Plasma Physics Branch, NASA Marshall Space Flight Center, C. R. Chappell, D. L. Reasoner, S. E. DeForest, P. H. Reiff, and W. Lennartsson, for their helpful comments. This chapter also benefited from discussions with F. S. Johnson, A. J. Dessler, T. W. Hill, and R. A. Wolf. Particular thanks go to C. R. Russell and D. J. Williams, who critically reviewed an earlier version of the manuscript.

REFERENCES

Akasofu, S.-I., and S. Chapman (1972). *Solar-Terrestrial Physics*, Oxford at the Clarendon Press, Oxford, England.

Akasofu, S.-I., and L. J. Lanzerotti (1975). The earth's magnetosphere, *Phys. Today 28*, 28.

Anderson, C. W., III, L. J. Lanzerotti, and C. G. MacLennan (1974). Outage of the L4 system and the geomagnetic disturbances of 4 August 1972, *Bell Syst. Tech. J. 53*, 1817.

Arnoldy, R. L. (1974). Auroral particle precipitation and Birkeland currents, *Rev. Geophys. Space Phys. 12*, 217.

Axford, W. I., and C. O. Hines (1961). A unifying theory of high-altitude geophysical phenomena and geomagnetic storms, *Can. J. Phys. 39*, 1433.

Block, L. P. (1975). Double layers, in *Physics of the Hot Plasma in the Magnetosphere*, B. Hultqvist and L. Stenflo, eds., Plenum, New York, p. 229.

Evans, D. S. (1975). Evidence for the low altitude acceleration of auroral particles, in *Physics of the Hot Plasma in the Magnetosphere*, B. Hultqvist and L. Stenflo, eds., Plenum, New York, p. 159.

Formisano, V., ed. (1975). *The Magnetospheres of the Earth and Jupiter*, D. Reidel, Dordrecht, Holland.

Fredricks, R. W. (1975). Wave-particle interactions and their relevance to substorms, *Space Sci. Rev. 17*, 449.

Jacobs, J. A. (1970). *Geomagnetic Micropulsations*, Springer Verlag, Berlin.

Kennel, C. F. (1969). Consequences of a magnetospheric plasma, *Rev. Geophys. 7*, 379.

Kennel, C. F., and H. E. Petschek (1966). Limit on stably trapped particle fluxes, *J. Geophys. Res. 71*, 1.

King, J. W. (1975). Sun–weather relationships, *Astronaut. Aeronaut. 13*, 10.

Lanzerotti, L. J., and H. Fukunishi (1974). Modes of magnetohydrodynamic waves in the magnetosphere, *Rev. Geophys. Space Phys. 12*, 724.

Lyons, L. R., R. M. Thorne, and C. F. Kennel (1972). Pitch-angle diffusion of radiation belt electrons within the plasmasphere, *J. Geophys. Res. 77*, 3455.

McPherron, R. L., C. T. Russell, and P. J. Coleman, Jr. (1972). Fluctuating magnetic fields in the magnetosphere, II. ULF waves, *Space Sci. Rev. 13*, 411.

Ness, N. F., K. W. Behannon, R. P. Lepping, and Y. C. Whang (1975). The magnetic field of Mercury, 1, *J. Geophys. Res. 80*, 2708.

Paulikas, G. A. (1975). Precipitation of particles at low and middle latitudes, *Rev. Geophys. Space Phys. 13*, 709.

Rostoker, G. (1972). Polar magnetic substorms, *Rev. Geophys. Space Phys. 10*, 157.

Russell, C. T., and R. L. McPherron (1973). The magnetotail and substorms, *Space Sci. Rev. 15*, 205.

Vasyliunas, V. M. (1975). Theoretical models of magnetic field-line merging, I, *Rev. Geophys. Space Phys. 13*, 303.

Williams, D. J. (1975). Hot plasma dynamics within geostationary altitudes, in *Physics of the Hot Plasma in the Magnetosphere*, B. Hultqvist and L. Stenflo, eds., Plenum, New York, p. 159.

3 The Thermosphere

RAYMOND G. ROBLE
National Center for Atmospheric Research

3.1 PROLOGUE

On a clear dark night in the polar regions of our planet, one may often see the awesome and dazzling display of the overhead aurora (see Figure 2.1 of Chapter 2). The intensity and extent of the optical emissions are a sign that enormous amounts of energy are being transferred into our atmosphere from the near-space environment. While the auroras appear to have no noticeable effect on conditions near the ground, statistics on the relationships between solar activity and terrestrial weather indicate that perhaps auroras may have some relationship to weather systems. Where does the auroral energy go, and how does it influence basic atmospheric structure? Nowhere in the atmosphere is the effect of auroral activity on the neutral gas structure more apparent than it is in the earth's thermosphere—the neutral atmosphere above 80 km.

Solar energy reaches the earth mainly in the form of electromagnetic radiation and in the form of solar wind, which is a charged-particle (plasma) flow from the sun. Both electromagnetic radiation and the energy from the solar wind interact with the earth's thermosphere in a number of ways. The main interactions are shown schematically in Figure 3.1. Most of the solar electromagnetic energy is in the visible region of the spectrum; it passes directly through the tenuous regions of the upper atmosphere and is either absorbed at the ground or reflected back to space. The solar electromagnetic energy at ultraviolet wavelengths shorter than 200 nm strongly interacts with the gases in the earth's thermosphere. The high-latitude thermosphere is also the region where important amounts of solar-wind energy are deposited through auroral processes. Within the high-latitude thermosphere, the deposited solar-wind energy can at times locally exceed the absorbed solar electromagnetic energy, thereby making the thermosphere a region that is competitively forced by the two forms of solar energy. The thermosphere is, therefore, a dynamically active region with large variations about its global mean state.

The atmosphere below 80 km is almost entirely in the molecular state and consists of approximately 78 percent molecular nitrogen, 21 percent molecular oxygen, and 1 percent other minor gas constituents such as ozone, car-

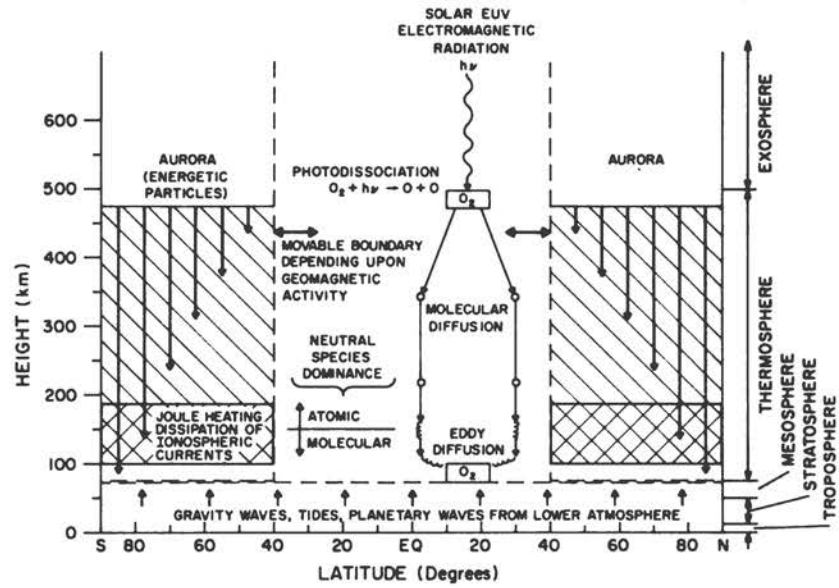


FIGURE 3.1 The main processes and energy sources that govern thermospheric structure and dynamics.

bon dioxide, and argon. Above 80 km, in the thermosphere, the molecular constituents are subjected to intense solar ultraviolet radiation, which can break the molecular bonds and create atomic species. The photodissociation of molecular oxygen into atomic oxygen is particularly effective in the thermosphere. The chemical reaction for recombination of atomic oxygen back into molecular oxygen is very slow above 90 km; therefore the compositional structure changes dramatically from a molecular mixture in the lower thermosphere to an atomic-oxygen-rich mixture in the upper thermosphere. One of the fundamental problems in thermospheric research is the study of this oxygen cycle—its formation, transport by molecular and eddy processes, transport by global circulation systems, and recombination back to molecular oxygen, for these processes largely determine the compositional and thermal structure of the atmosphere above 80 km. Within the thermosphere every absorbed solar photon eventually dissociates an oxygen molecule into two oxygen atoms, even though many complex chemical reactions may be involved in intermediate processes. Of the total absorbed solar energy, approximately 35 percent appears as local heating of the ambient neutral particles, 45 percent is radiated out of the thermosphere as ultraviolet airglow, and 20 percent is in the stored chemical energy of the oxygen atoms, which is regained on recombination to molecular oxygen.

Since there are no effective radiators to remove the thermal energy absorbed above 90 km, the global mean vertical temperature structure is determined by a balance of the local heating with downward molecular and eddy thermal conduction to the region of the mesopause—the temperature minimum near 80 km. In this region, the energy is radiated to space by optically active minor neutral constituents such as carbon dioxide and ozone. Thermal-energy conduction requires a positive temperature gradient, therefore the thermospheric temperature increases from a low value of about 200 K near 80 km to as high as 1500 K at 300 km, where molecular conduction is so fast that it prevents large vertical temperature gradients from developing.

The two main processes by which solar-wind energy is transferred into the atmosphere are auroral particle precipitation and the frictional dissipation of ionospheric current systems driven by the interaction of the solar wind with the earth's magnetic field. These processes are mainly confined to high latitudes, whereas the solar ultraviolet heating is greatest near the subsolar point. The relatively steady solar ultraviolet heating drives a thermospheric circulation that basically flows from the subsolar point to the antisolar point. In a latitudinal average, the flow is from the equator toward the poles at the equinoxes and from the summer to the winter hemisphere at the solstices. The high-latitude heat source drives a latitudinally averaged circulation from high to low latitudes. The strength of this circulation is dependent on the magnitude of the heat source, which in turn varies with geomagnetic activity.

During large storms, this energy source associated with geomagnetic activity may in the global mean be comparable to or even exceed the global mean heating due to solar ultraviolet radiation above 150 km. As a result, the thermosphere exhibits large variations in its circulation, being driven by heating due to the absorption of solar electromagnetic energy during quiet geomagnetic conditions and then driven by auroral processes during large geomagnetic storms, with some intermediate circulation at other levels of geomagnetic activity. As a result of these variations, the temperature and compositional structure are also in a dynamic state of adjustment with large seasonal, diurnal, and geomagnetic-storm time components.

The interaction of solar electromagnetic radiation with the earth's neutral thermosphere is reasonably well understood, and therefore the global distribution of such solar heating can be determined. On the other hand, the heat and momentum sources associated with auroral processes are not well understood, and the next major effort in thermospheric research will be to determine the complex magnetospheric–ionospheric–thermospheric coupling processes and their mutual interactions. Finally, the coupling between the thermosphere and lower atmospheric levels must be investigated, perhaps during the Space Shuttle era when remote probing instruments will be able to obtain data with full coverage of the mesosphere and lower thermosphere. It should then be possible to determine the ultimate fate of the auroral energy and to understand how auroral processes interact with the lowest regions of the atmosphere.

Recent reviews on this subject have been given by Carignan (1974), Dickinson (1975), and Evans (1975).

3.2 STRUCTURE OF THE THERMOSPHERE

The atmosphere above 80 km has been explored by rockets, satellites, and ground-based remote sensing (radar, radio, and optical) stations for over 25 years. It is now evident that dynamical phenomena on just about all scales are present in the neutral thermosphere. Since the region is subject to forcing from below 80 km in the form of upward-propagating planetary, tidal, and gravity waves and from above by magnetospheric–ionospheric interactions (in addition to strong solar forcing), there is indeed good reason for large variations about a global mean state in the thermal and compositional structure. The causes of these variations are mainly dynamical; chemical processes may modify the variations that are produced, but chemical processes do not appear to initiate the variations nor to dominate them.

One of the important problems, then, is to determine the global mean thermal and compositional vertical structures about which variations take place. In the global mean, the influence of motions on the thermal structure is averaged out above some altitude, and the thermal structure is determined by a balance between the known energy sources and sinks. It is possible to evaluate current knowledge of these sources and sinks by comparing calculations and observations.

The main force that changes the orbit of a satellite is the drag exerted by the neutral atmosphere. The drag is related to the atmospheric density. Thus, by observing the orbital changes of many satellites, Jacchia (1965) was able to determine a global distribution of neutral gas density within the thermosphere above 120 km. With further theoretical constraints and assuming constant boundary conditions at 120 km, he was then able to express the density measurements in terms of the exospheric temperature T_∞ that is associated with the vertical profile of temperature and is described later. Thus, a great quantity of satellite-derived density data could be easily summarized with an analytic mathematical representation of global distributions of T_∞. This empirical model of the global distribution of neutral temperature and neutral composition has proved to be a valuable resource for thermospheric research. It has worked well in problems of ionospheric physics, neutral gas dynamics, and wherever a neutral background gas must be specified to study some atomic, chemical, or collisional process.

The vertical profiles of the global mean temperature within the thermosphere, as determined by Jacchia, are shown in Figure 3.2. It is strongly controlled by the absorption of solar extreme ultraviolet (EUV) radiation (wavelengths less than 102.5 nm, the ionizing portion of the solar spectrum). This radiation cannot be measured at ground level because it is all absorbed in the upper atmosphere. However, it was shown about 10 years ago that solar EUV radiation is closely related to the solar decimeter radio emission at a wavelength of 10.7 cm, which can be measured from the ground with radio telescopes. Thus, the global mean neutral temperature of the thermosphere can be expressed as a function of the solar radio emission. The notation for this function is F10.7. During solar minimum the sun is relatively quiet, and solar F10.7 values are typically 80×10^{-22} W m^{-2} Hz^{-1}. Near solar maximum, values of 160×10^{-22} W m^{-2} Hz^{-1} or greater are observed. The magnitude of the solar EUV output roughly doubles between solar minimum and solar

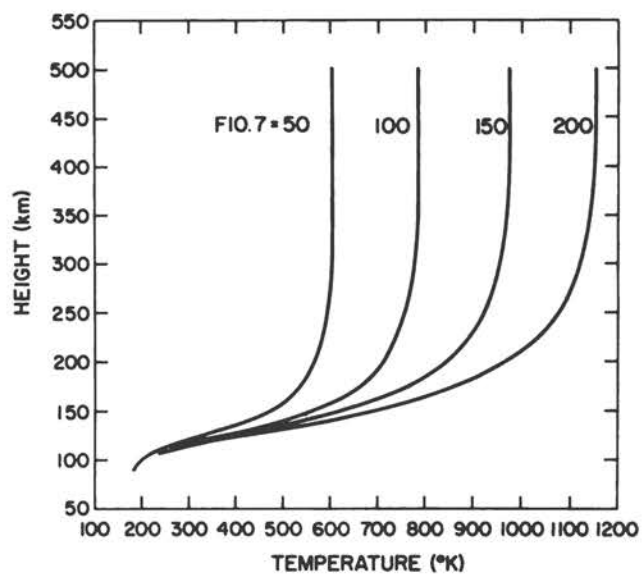


FIGURE 3.2 Global mean neutral gas temperature profiles determined from the Jacchia (1965) empirical model of neutral temperature and composition based on satellite drag measurements as a function of the solar decimeter radio emission at 10.7-cm wavelength (F10.7). The solar decimeter radio emission F10.7 ($\times 10^{-22}$ W m^{-2} Hz^{-1}) is an indicator of the solar EUV electromagnetic radiation output.

maximum, and there is also considerable day-to-day variation.

As shown in Figure 3.2, the neutral gas temperature has a profile that increases from a minimum of about 180 K near the mesopause at 80-km altitude to a constant value above 300 km that is called the exospheric temperature T_∞. The exospheric temperature occurs by definition at levels sufficiently high for molecular thermal conduction to be fast enough in the rarified atmosphere to eliminate any vertical temperature gradient, thus keeping the neutral gas temperature constant with altitude. The global mean neutral-gas exospheric temperature increases with the solar EUV output and exhibits a several-hundred-degree variation between solar minimum and solar maximum.

There are two main periods associated with variations of solar activity: one corresponds to the sun's rotation period of 27 days and the other to the 11-year sunspot cycle. The main centers of solar EUV activity are related to active regions associated with sunspots, plages, and coronal condensations that appear and disappear on the visible disk as the sun rotates. The solar EUV output is also related to the number of sunspots. There are also short-time fluctuations associated with flares and other solar noise, and these also induce a response in thermospheric temperature and density. Generally, as solar activity increases so does the neutral thermospheric temperature.

The thermosphere has a diurnal variation in temperature with an amplitude of roughly 20–30 percent. The exospheric temperature has a maximum value near the subsolar point and minimum value near the antisolar point, as shown in Figure 3.3 for northern hemisphere summer solstice conditions. As the large diurnal variation of temperature proves, the absorption of solar electromagnetic radiation is normally the dominant forcing mechanism within the thermosphere. The temperature maximum also migrates in latitude with the subsolar point during the course of a year, giving a seasonal variation within the thermosphere. The diurnal temperature variation depends on latitude and season in a manner that

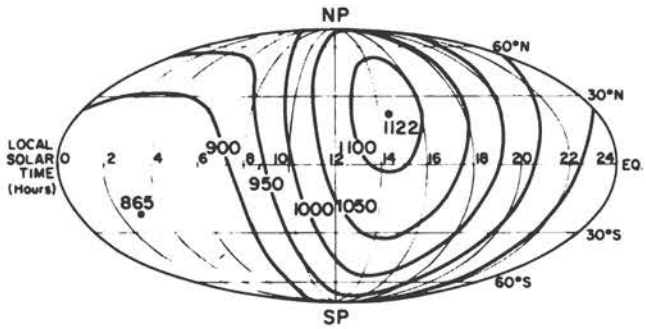


FIGURE 3.3 Global distribution of the neutral exospheric temperature (K), T_∞, determined from the Jacchia (1965) model for an F10.7 value of 150 at northern summer solstice. The diurnal distribution is a function of latitude and local time.

indicates that it is strongly dependent on the length of the day and hence on the total amount of absorbed solar EUV radiation.

The composition of the neutral gas constituents in the thermosphere also responds to the solar EUV radiation and ultraviolet radiation with wavelengths between 102.5 and 200 nm. Photodissociation of molecular oxygen by solar radiation at wavelengths below 200 nm occurs to an important degree above 90 km;

$$O_2 + h\nu \rightarrow O + O,$$

where $h\nu$ represents an ultraviolet photon. For oxygen atoms to recombine back to molecular oxygen in a chemical reaction, the two oxygen atoms must collide with a third body, M,

$$O + O + M \rightarrow O_2 + M,$$

and measured values of the reaction rates for such chemical reactions indicate that they can proceed only very slowly above 90 km. It is necessary that the atomic oxygen move downward into a denser region of the atmosphere below 90 km, where collisions occur rapidly enough to provide a recombination rate that matches the photodissociation rate. (The various processes in the photodissociation/recombination chain of atomic oxygen within the thermosphere are illustrated schematically in Figure 3.1.) At altitudes greater than 90 km, photodissociation of molecular oxygen acts to convert the molecular neutral atmosphere into an atomic neutral atmosphere; atomic oxygen is the major atmospheric constituent above about 150 km, as shown in Figure 3.4.

In the tenuous upper region of the thermosphere above 110 km, atmospheric turbulence dies away and molecular diffusion processes dominate. Molecular diffusion dominates the vertical transport of the atmospheric constituents at these levels and, to a first approximation, the atmospheric constituents adjust themselves in the earth's gravitational field according to their molecular weights. Heavier gases occupy the lower thermosphere, and the lighter gases extend to higher altitudes. Such a diffusive equilibrium distribution of atmospheric constituents is shown in Figure 3.4. The distribution of neutral constituents in diffusive equilibrium depends on the neutral gas temperature, which has a large diurnal variation; the neutral gases expand and contract with the temperature variations. The magnitude of a typical neutral composition diurnal variation and, for comparison, the diurnal variation of the electron density of the ionosphere, are shown in Figure 3.4. Particles in the lower thermosphere are mostly neutral; however, the neutral gas number density decreases more rapidly than the electron density with altitude, causing a gradual change to an essentially pure plasma atmosphere at very high altitudes within the magnetosphere.

Analysis of satellite drag data and the corresponding empirical thermospheric models reveals variations on

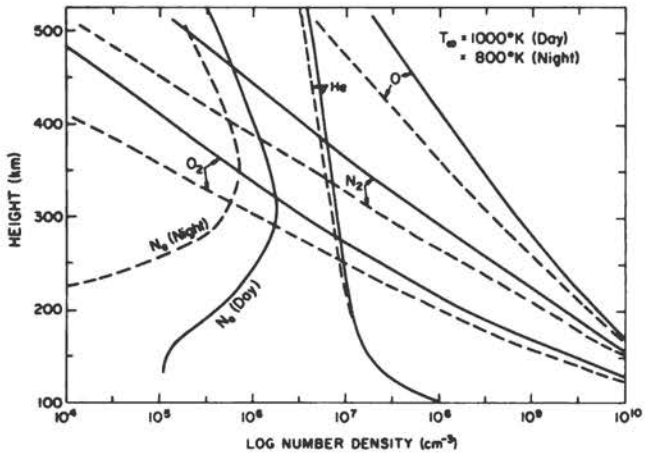


FIGURE 3.4 Typical diurnal variation of the major neutral gas constituents in the thermosphere: atomic oxygen, O; molecular nitrogen, N_2; molecular oxygen, O_2; helium, He; and electron density, N_e, in the ionosphere. Solid lines indicate daytime values, dashed lines nighttime values.

other temporal and spatial scales within the thermosphere. These include semiannual and semidiurnal variations as well as variations related to geomagnetic activity. Maximum values of temperature and density occur semiannually near the spring and autumn equinoxes, and minimum values occur near the summer and winter solstices. The amplitude of the semiannual variation is about 20 to 40 percent of the annual variation. The cause of the semiannual variation is not known, although some evidence indicates that it may be excited in the lower atmosphere and propagate into the thermosphere. Other causes have been suggested, including semiannual variations in Joule heating caused by the dissipation of ionospheric current systems and tidal- and gravity-wave energy transfer from the lower atmosphere. The actual mechanism is still uncertain, and semiannual variations in the thermosphere involving couplings between the upper and lower atmosphere will remain a lively research subject.

Semidiurnal variations occur with an amplitude of 10 to 30 percent of the diurnal variation near 300 km. The semidiurnal variation, however, is observed to dominate the diurnal variation below 200 km. Recent studies have shown that the semidiurnal variation is caused mainly by a tidal oscillation excited by the absorption of solar energy by ozone in the stratosphere that propagates up to thermospheric heights. There are all sorts of tidal components that are excited in the lower atmosphere. Not all of these reach upper thermospheric heights. Some are trapped within the stratosphere and mesosphere, and others are damped in the lower thermosphere by the increasing molecular viscosity and molecular thermal conductivity, transferring their energy and momentum to the region where wave absorption occurs. The semidiurnal and diurnal variations also interact with the ionospheric plasma near 100 km to produce a global-scale ionospheric current system that can be detected by ground-based magnetometers.

Rapid fluctuations of atmospheric density and temperature with a typical duration of one or two days are connected with geomagnetic disturbances. The amplitude of the observed effect is dependent on the size of the geomagnetic storm; it increases with altitude in the thermosphere, and the maximum effect in density lags the maximum geomagnetic disturbance by several hours. The delay increases with distance from the auroral zone. The main effect of geomagnetic activity is a global increase in the exospheric temperature that depends on the strength of the geomagnetic storm. This clearly implies that the large amount of energy transferred to the thermosphere through auroral processes at high latitudes is later distributed globally through dynamic processes.

Other empirical thermospheric models have been developed that are based on data obtained from incoherent-scatter radars and on satellite measurements of neutral composition. Each has revealed complex dynamic processes on a much smaller scale than those resolved by satellite drag. For example, the incoherent-scatter radar data showed that the neutral gas temperature peaked at a later time in the afternoon than the temperature in the model based on satellite-drag data. This suggested that the local time variations in temperature and density over a given geographic location were not in phase. Subsequently, it was shown that each constituent has its own time of maximum amplitude that also depends on altitude; at 300 km, molecular nitrogen peaks with the neutral gas temperature near 1600 local time (LT), atomic oxygen near 1000 LT, helium early in the morning near 0600 LT, and hydrogen near 0300 LT. These observations suggest that important variations occur in the lower thermosphere below 120 km that influence the temperature and composition at higher altitudes and that in some degree invalidate the temperatures determined from satellite drag (since they depend on models assuming constant boundary conditions at 120 km).

3.3 ENERGY AND MOMENTUM SOURCES THAT DRIVE THE THERMOSPHERIC CIRCULATION

Solar extreme ultraviolet and solar ultraviolet energy (with wavelengths below 200 nm) absorbed in the earth's thermosphere about 80 km because the absorption cross sections of the atoms and molecules are large at these wavelengths. Figure 3.5 schematically illustrates the basic vertical distribution of the absorption process in the atmosphere. The incident solar energy at the top of the atmosphere encounters an increasing concentration of absorbing particles as it penetrates to lower altitudes and so is absorbed at an increasing rate. A peak absorption rate is attained at unit optical depth; at lower altitudes, the rate declines as the solar EUV and ultraviolet flux become negligible. The height of the maximum energy

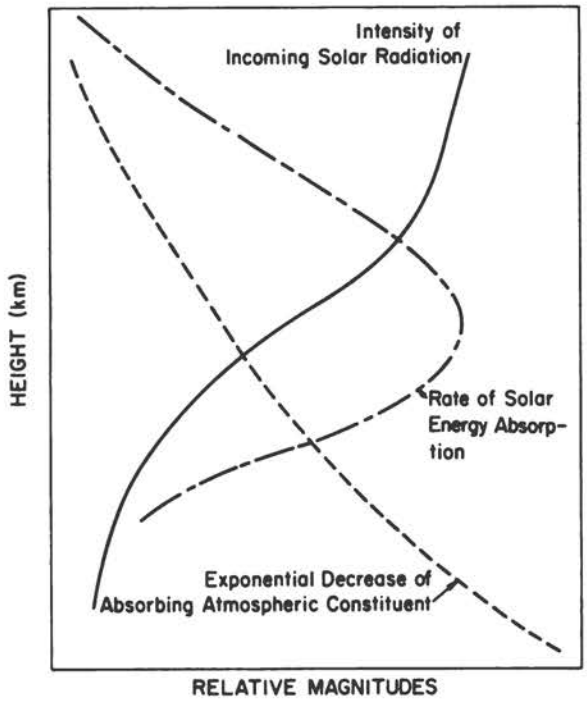


FIGURE 3.5 Schematic of the basic energy absorption process in the earth's atmosphere. The energy absorption has a peak altitude that depends on the cross section of the absorbing species.

absorption depends on the product of the absorption cross section and total column number density of the absorbing species along the path to the sun; if the cross section is large, solar energy is absorbed high in the atmosphere, and if it is small, the absorption takes place lower down. Thus, the absorption of a complete spectrum of solar radiation occurs over a broad region determined by the summation of all the individual layers established by different wavelengths.

Just how the absorbed solar energy is then channeled through the myriad of atomic and molecular chemical and physical processes that can occur has been a subject of considerable interest for a number of years. It is necessary to understand the processes by which absorbed solar energy is redistributed into radiated airglow, stored chemical energy, and direct heating of the neutral gas.

The partitioning is dependent on the atmospheric composition because different radiative, chemical, and atomic processes occur for each gas species and each gas mixture. Thus, each planet within the solar system will have a different thermospheric structure depending on how the solar energy is apportioned among airglow, chemical energy, and heating and also upon how the absorbed energy is lost to space by radiative processes.

For the earth, virtually all solar photons at wavelengths less than 102.5 nm are absorbed by the major neutral constituents of the thermosphere—N_2, O_2, and O—leading to ionization of the absorbing species. The photoionization event splits the energy of the absorbed photon approximately equally into two channels, shown schematically in Figure 3.6, the kinetic energy of the ejected fast photoelectron and the chemical energy of the ion production. The latter can be regained in recombination with an electron. The ejected fast photoelectron slows down by collisions with the ambient thermal electrons of the ionosphere and by collisions with neutral particles. Fast-photoelectron/thermal-electron collisions result in a local heating of the background ionospheric electron gas, which in turn is directly transferred to the neutral gas by collisions, thus providing thermal energy

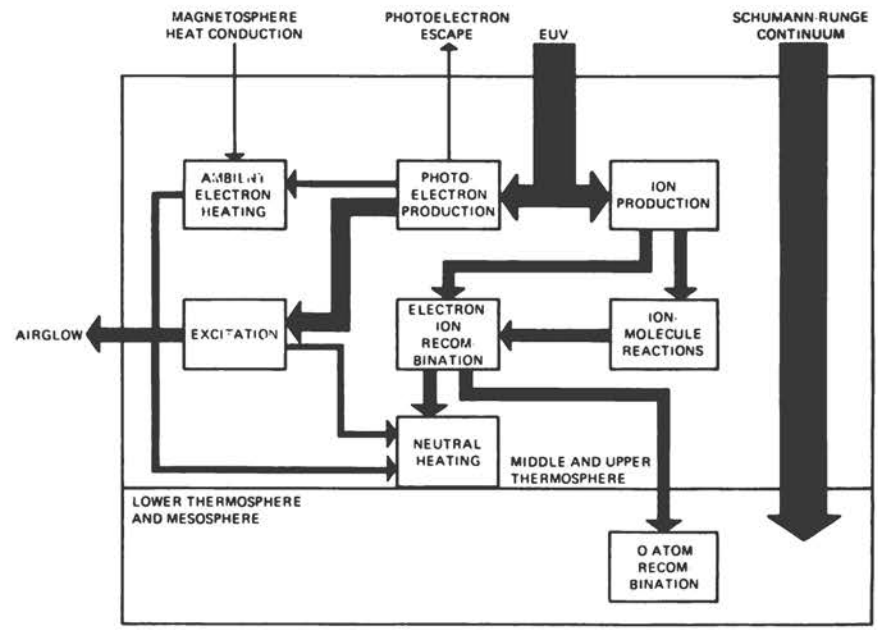


FIGURE 3.6 Energy flow diagram of the processes leading to the conversion of absorbed solar EUV electromagnetic radiation into local thermal energy of the neutral gas (courtesy of R. S. Stolarski, NASA). The width of the arrow illustrates the relative energy in each process.

to the neutral gas. Also, the fast photoelectrons may be sufficiently energetic that, on collision with neutrals, they can cause further ionization, creating more electron–ion pairs, or they may excite the internal atomic levels of an atom or molecule. These excited levels can then be either deactivated by collisions with neutral particles, the excess energy appearing as local thermal energy of the particles, or, if the optical transition probability is sufficiently high, radiated as airglow. The airglow can be either lost by radiation to space or reabsorbed—possibly causing dissociation or additional ionization, depending on the wavelength and the optical depth.

The chemical energy of the ion, on the other hand, is shuffled from one ion to another by a series of charge-transfer or ion–molecule interchange reactions, eventually ending in a dissociative recombination (molecular ion + electron → two atoms) yielding two atoms of oxygen or nitrogen. The N atoms participate in local neutral chemistry that leads to the production of O atoms by breaking the molecular oxygen bond, with the N atoms ending up as nitric oxide, NO, or molecular nitrogen, N_2. The O atoms do not recombine locally to O_2 because, as discussed previously, the reaction for atomic oxygen recombination is a three-body reaction that requires a high neutral particle density to proceed rapidly. Atomic oxygen is therefore transferred by molecular and eddy diffusion downward to about 90 km. As a result, the energy required to dissociate O_2 is lost from the region of EUV absorption, as it is transported to lower altitudes and released there through recombination, reappearing as thermal energy below 90 km.

Transport of the photoelectron or ion species can further complicate the heating process because the created products may move away from the region of solar energy absorption and participate in either a collisional process or a chemical reaction at some other altitude. Although the details of the neutral gas heating are complex, a rather simple argument can be made to show that the neutral gas heating efficiency in the earth's thermosphere is 30–40 percent of the absorbed solar radiation. The average solar EUV photon energy is approximately 30 eV. The initial photoionization event splits this energy into two channels—about 50 percent or 15 eV each, for the ion formation and the fast photoelectron. Of the 15-eV potential energy of recombination in the ion formation, 5 eV—the O_2 binding energy—is transported by atomic oxygen downward to the lower thermosphere, where recombination to O_2 occurs and the binding energy is released. The remaining 10 eV provides direct neutral heating near the altitude where the photon was absorbed. Of the 15 eV that goes into the photoelectron channel, only about 1 eV heats the ambient electron gas, and this eventually appears as neutral gas heating because of collisional processes. The rest of the energy goes into atomic and molecular excitation that radiates as airglow and, depending on the radiation optical depth, may be lost as far as local heating is concerned.

Combining the two channels, an overall local neutral gas heating efficiency (defined as energy appearing as local thermal energy of the neutral gas divided by the solar EUV radiation locally absorbed) of 36.6 percent is obtained. Allowing for uncertainties in the processes, a heating efficiency between 30 and 40 percent is expected. Approximately 20 percent of the absorbed solar energy is transported by atomic oxygen to the lower thermosphere below 90 km, and 40 to 50 percent appears as airglow radiation that is either lost to space or transferred out of the thermosphere and reabsorbed in the lower atmosphere. Much progress has been made in understanding the gross aspects of this energy-partitioning problem, yet much uncertainty exists about the details of the many atomic and radiational processes. Atmosphere Explorer satellite data are being used to examine the interaction of solar EUV radiation with the earth's atmosphere, and the data will help greatly in unraveling the details of these important physical and chemical processes.

Neutral gas heating by solar EUV as described above dominates in the 150–300 km altitude range. Above 300 km, the same processes occur, but the neutral gas heating there is also affected by collisional processes involving ions. The electrons and ions are heated by energy flowing by conduction in the plasma down along the earth's geomagnetic-field lines from the magnetosphere. The neutral gas heating at high altitudes is controlled by complex magnetospheric–ionospheric coupling interactions and is relatively unimportant in controlling the thermospheric dynamics, which is mainly driven by heating in the 150–300 km region.

Below 150 km, molecular oxygen absorption of solar ultraviolet radiation in the Schumann-Runge continuum region (the O_2 absorption continuum between 130 and 175 nm), is the dominant heat source, with a heating efficiency of about 33 percent. The absorbed energy breaks the O_2 bond, producing two atomic oxygen atoms that are transported to the lower thermosphere below 90 km, where they recombine. The average solar energy per photon in the continuum is about 7.5 eV (at 160 nm), and the excess of this energy over the O_2 dissociation energy (5 eV) results in neutral gas heating with an efficiency of 33 percent.

At still lower altitudes, below 100 km, neutral gas heating due to absorption of solar energy in the Schumann-Runge band system (175–200 nm), heating due to the solar energy absorption by ozone, and heating due to the recombination of all the atomic oxygen transported downward from the thermosphere above 100 km become important heat sources for determining the thermal and compositional structure of the lower thermosphere and mesosphere. The relative importance of these heat sources is shown in Figure 3.7. Thus, within the earth's thermosphere, the oxygen cycle—the breaking of the O_2 bond and the subsequent transport of atomic oxygen out of the region of formation—is an important process in determining the neutral gas heating efficiency and compositional structure.

Energetic charged-particle precipitation from the mag-

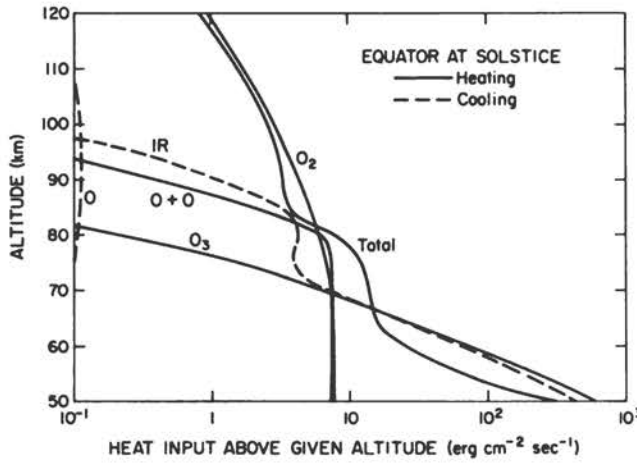


FIGURE 3.7 Heat input into the upper atmosphere above given altitudes evaluated over the equator at solstice (Johnson and Gottlieb, 1970). O_2 indicates local heat release due to absorption of solar radiation by molecular oxygen in the wavelength range below 187.5 nm; O_3, that due to absorption of solar radiation by ozone; O + O, the heat release due to recombination of atomic oxygen into molecular oxygen; O, the heat loss by 63-μm radiation from the ground level of atomic oxygen; IR, the infrared losses due to optically active minor constituents.

netosphere and Joule (resistance) heating due to electrical current systems within the ionosphere are the next most important heat sources that control the thermospheric structure and composition. These sources derive their energy from the solar-wind interaction with the earth's magnetic field. Most of this energy deposition occurs in the vicinity of the auroral oval at high latitudes (see Figure 2.5 of Chapter 2), and it appears to be about equally partitioned between particle precipitation and the Joule dissipation of ionospheric currents. The most intense heating occurs in the sunward portion of the auroral oval near the region of the polar cusp. Within this region, large electric fields perpendicular to the geomagnetic-field lines have been observed by satellites and, because of the high conductivity in the dayside ionosphere, large Joule heating rates occur. The major currents of the coupled magnetospheric–ionospheric regime are the equatorial current, the field-aligned currents, and the auroral electrojets (described in Chapter 2). These current systems, each carrying roughly a million amperes, are directly involved in energy transport from the solar wind to the neutral atmosphere. The auroral oval is also a region of intense particle precipitation, which appears as a bright optical emission in satellite photographs of the aurora (see Figures 2.1 and 2.2 of Chapter 2). Recent studies indicate that the horizontal distribution of neutral heating has roughly a horseshoe shape along the auroral oval with maximum heating in the dayside polar cusp region and minimum values near the midnight sector.

There are theoretical studies suggesting that the magnitude of the solar-wind energy entering the atmosphere (and thus being transferred to the neutral atmosphere) is proportional to the size of the auroral oval as it expands and contracts with geomagnetic activity. During geomagnetically quiet conditions, the global magnetospheric energy input to the thermosphere through auroral processes is of the order of 5×10^{17} to 10^{18} erg sec^{-1}, whereas during large geomagnetic storms the energy input increases by a factor up to perhaps 10 or even more. The variability of the magnitude and spatial extent of the high-latitude heat source can be appreciated by studying the morphology of the aurora. For comparison, the global mean solar heating due to the absorption of solar ultraviolet radiation is 4×10^{18} erg sec^{-1} above 120 km and about 10^{19} ergs sec^{-1} above 100 km. Thus, in contrast to the relatively steady heat source provided by solar EUV and ultraviolet radiation, the optical aurora exhibits large temporal and spatial variations that undoubtedly cause large variations in the neutral gas heating.

Optical auroras are produced mainly by energetic electron bombardment exciting the neutral particles in the thermosphere through collisional processes. The processes that absorb most of the energetic electron energy are ionization ($O + e \rightarrow O^+ + e$); dissociation ($O_2 + e \rightarrow O + O + e$); and dissociative ionization, in which the molecular bond of O_2 or N_2 is broken in the ionization process, yielding an atom and an atomic ion ($N_2 + e \rightarrow N^+ + e$). The primary bombarding electrons have energies of a few keV, so they are energetic enough to cause many ionization, dissociation, dissociative ionization, or optical excitation events as they slow down in their progress through the neutral gas of the thermosphere. The more energetic the particle, the deeper into the atmosphere it can penetrate before depositing the bulk of its energy. A 10-keV electron will cause most of its ionization and optical excitation near 105 km, whereas a 0.5-keV electron will have a broader energy deposition profile peaking near 250 km altitude. Once an energetic electron slows down to about 100 eV, processes occur that are similar to those described earlier (Figure 3.6) for the absorption of solar EUV energy and its partitioning among airglow, chemical energy, and kinetic energy. The wide range of energy spectra of energetic particles observed by satellites indicates that the energy deposition altitude profile must also be variable, adding considerable complexity to the question of understanding the earth's high-latitude heat source. How this complex heating structure couples dynamically with the earth's thermospheric circulation system will be a challenging problem for the future, not only for the earth but also for understanding the mechanisms by which solar-wind energy is coupled to the atmospheres of other planets.

The interaction of the solar wind with the earth's magnetic field produces a magnetospheric convection flow of the earth's plasma at high latitudes, as shown in Figure 2.3 of Chapter 2. The charged-particle flow has associated with it an electric field perpendicular to the geomagnetic-field lines. Because of the high conductivity

along magnetic-field lines, the magnetosphere-generated electric field can penetrate down to the ionosphere and provide energy and momentum to the ambient electrons and ions. Both the electrons and the ions above 120 km drift in a direction perpendicular to both the electric and magnetic field vectors **E** and **B**, where the magnitude of the drift is linearly related to the magnitude of the electric field. For example, an electric field of one millivolt per meter causes the charged particles above 120 km to drift at 20 m/sec. Typical values of electric fields associated with the high-latitude magnetospheric convection pattern can vary between 10 and 100 mV m^{-1}, implying drifts of 200 to 2000 m sec^{-1}. The strength of the field is variable and also depends on the magnitude of the convection process; it is greatest during geomagnetic substorms.

Ions drifting through the neutral gas experience a frictional drag force. Collisions between ion and neutral species limit the ion drift velocity and convert part of the kinetic energy of the macroscopic ordered motion into random thermal motion, resulting in frictional heating. The total heating of the ions and neutral gas by their relative drift in the presence of an electric field is proportional to the product of the ionospheric Pedersen conductivity and the square of the electric field in the frame of the neutral gas ($\sim\sigma_p E_\perp^2$). This is simply an ionospheric equivalent to Joule (resistance) heating of current flowing through a resistor.

The ions, through collisions with neutrals, also "drag" the neutral atmosphere in their direction of motion by giving up some of their momentum to the neutral gas in the collisional process. The "ion drag" is an important momentum source for the neutral gas in stirring the atmosphere in the direction of the convection pattern. However, the atmosphere will adjust hydrodynamically to this forcing in a complex fashion, either flowing with the drifting ions or building up a back pressure to resist the momentum forcing. Thus the electric fields are both a heat source for the neutral atmosphere, converting the ordered motion of the $\mathbf{E} \times \mathbf{B}$ charged-particle drift into increased random thermal motion through collisional processes, and a momentum source also, because of the transfer of charged-particle momentum to the neutral gas in the collisional processes.

There are other heat and momentum sources acting in the thermosphere. These include heat conduction in the plasma from the magnetosphere to the ionosphere and eventually to the neutral atmosphere at high latitudes and midlatitudes; dissipation by thermal conduction, molecular viscosity, and compositional damping of tidal and gravity waves that are excited in the lower atmosphere and propagate upward; Joule heating by tide-driven ionospheric current systems; and all sorts of plasma energy and momentum interactions with the neutral atmosphere. The global impact of all of these interactive processes and the coupling of the small-scale phenomena with global-scale processes is not completely known at this time. Many of these processes are discussed in other chapters of this volume. In this chapter, only the large-scale processes associated with solar electromagnetic energy absorption and the gross features of the high-latitude heat and momentum sources are considered.

3.4 GLOBAL MEAN THERMAL STRUCTURE

The global mean temperature distribution in the thermosphere above 120 km is approximately determined by a balance of the global mean radiative heat sources with vertical molecular conduction. The only significant thermal radiative loss process for the neutral gases is the 63-μm radiation from the ground level of the oxygen atom, but it is not very effective in radiating atmospheric energy to space, as shown in Figure 3.7. The only way to rid the thermosphere of the 30 to 40 percent absorbed solar electromagnetic radiation that appears as heat is by transport to the lower atmosphere by molecular thermal and eddy heat conduction. Then in the region near 90 km the thermospheric heat is radiated to space by the optically active minor neutral constituents, carbon dioxide and ozone.

Figure 3.7 indicates the various heating and cooling processes that are active in the lower thermosphere and mesosphere, and their relative magnitudes, calculated at the equator during solstice. If just the global mean temperature above 120 km is calculated, it can be determined by considering a balance between the global mean solar and high-latitude heating and the downward molecular thermal conduction. One calculation of the partitioning between the global mean solar energy and the high-latitude energy sources is shown in Figure 3.8, along with the calculated global mean temperature for these distributions. The dashed line is the global mean temperature derived from an empirical model of temperature and composition in the thermosphere based on satellite measurements, shown for comparison. The calculations apply to quiet geomagnetic conditions during solar maximum. The main point here is that the high-latitude heat source appears to be an important component of the global thermospheric energy budget, and it must be considered in bringing the calculated and observed global mean temperatures into agreement even during geomagnetic quiet conditions. During large geomagnetic storms, the high-latitude heat source becomes even more important in the global thermospheric energy budget.

3.5 DYNAMIC STRUCTURE OF THE THERMOSPHERE

Direct measurements of the neutral gas temperature and winds in the thermosphere are difficult to make. Nevertheless, it has been possible to derive information on the global dynamic structure of the thermosphere from ground-based optical measurements, incoherent-scatter radar data, and various types of satellite data. These data, in addition to providing specific information on a

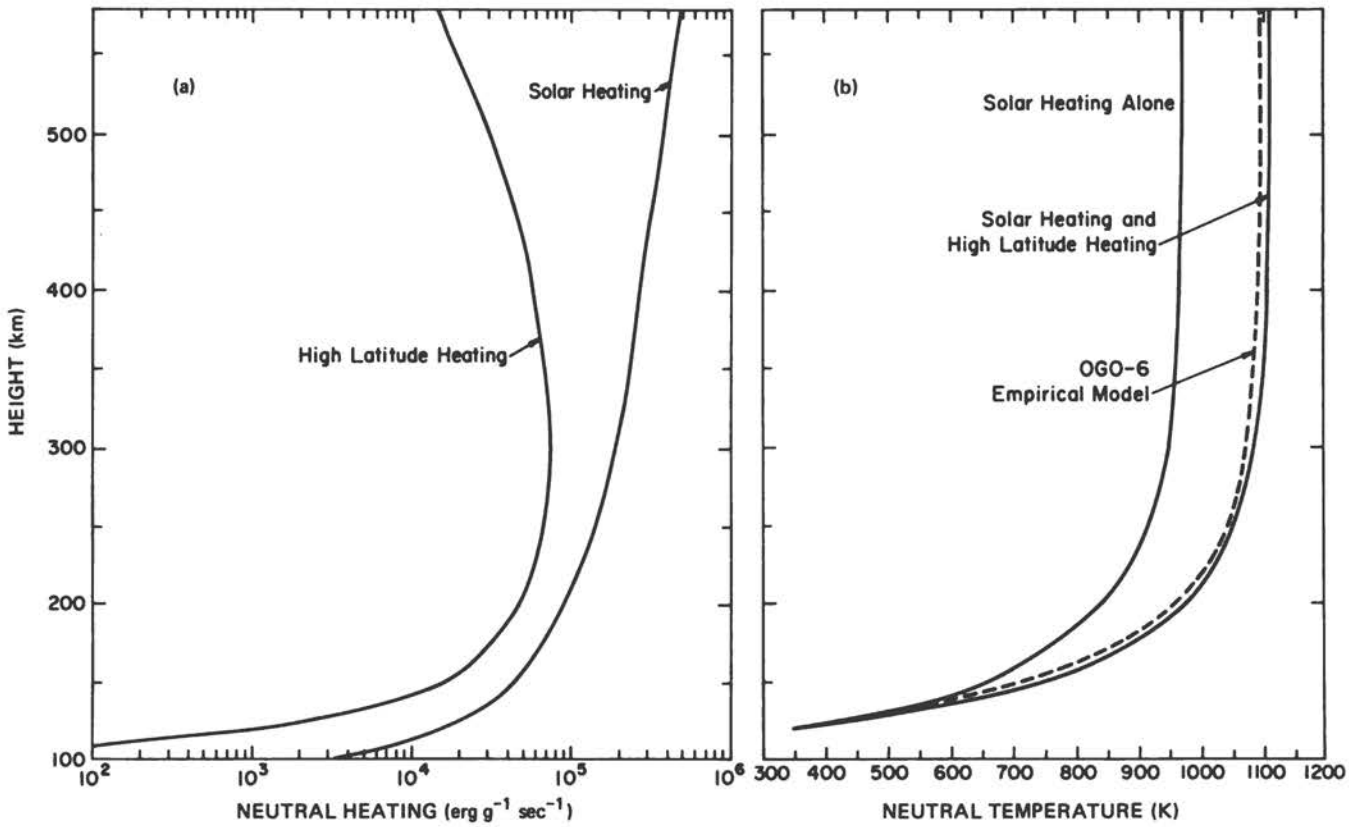


FIGURE 3.8 (a) A calculated global mean neutral gas heating rate (erg g^{-1} sec^{-1}) for solar ultraviolet heating (EUV + UV) and a global mean heating distribution for high-latitude auroral processes. (b) The calculated global mean temperature for the given global mean heating rates. The dashed curve is the global mean temperature determined from the OGO-6 empirical model for similar conditions.

geophysical situation, have also been used to construct global empirical models of thermospheric properties. About a decade ago, it was recognized that the empirical thermospheric models were a valuable resource for those interested in studying global thermospheric wind systems. The model distributions of temperature and composition equivalently specified the distribution of pressure that forces the global circulation. Therefore, it was possible to integrate the equations of fluid motion and determine the wind system that should result from the pressure distribution. These equations are the same as those used by meteorologists studying weather systems in the lower atmosphere; however, they were modified to include two additional forces that are important at thermospheric heights: the viscous force and the ion drag force.

Kinematic molecular viscosity increases exponentially with altitude by several orders of magnitude within the thermosphere. Its main effect is to transfer momentum between the various altitude regions and thus to smooth out vertical gradients in wind velocity. This "stickiness" increases with altitude to such an extent that the upper thermosphere has essentially a bulk flow or slab motion above 300 km. It is also large enough above 300 km to prevent the development of large horizontal shears.

The other force is the ion drag force, discussed previously, which is a collisional interaction between the charged particles and the neutral particles. At thermospheric heights above 120 km, the ion–neutral collision frequency is much smaller than the ion gyrofrequency (the frequency at which a charged particle spirals around a magnetic-field line), and therefore ions are locked to the geomagnetic-field lines and can only move across them when driven by an electric field. Outside of the auroral zone, electric fields are small and the ions can to a first approximation be considered to simply corotate with the earth. A neutral wind flowing through the ions experiences a collisional drag that becomes a maximum at the peak of the ionospheric layer and provides the resistance force that balances the driving pressure force. During the day, when the ionization and hence ion drag are large, the wind flows across constant-pressure surfaces from the subsolar high-pressure area to the antisolar low-pressure area shown in Figure 3.9, which depicts the calculated neutral winds at 300 km. At night, the ionospheric density decays and the thermospheric pressure forces drive much larger winds because of the reduced ion drag.

The thermospheric winds are three dimensional in space and also a function of time. However, to a good

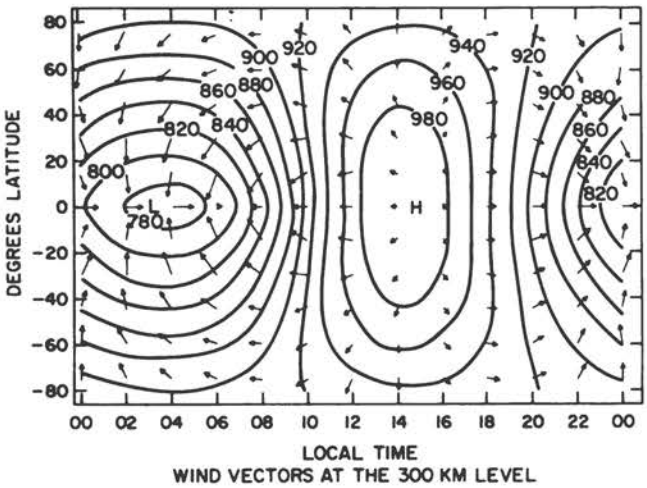


FIGURE 3.9 The global distribution of neutral gas temperature (K) at 300 km as determined from the Jacchia (1965) empirical model. The arrows are the wind pattern determined at 300 km. The longest arrow represents a wind speed of 225 m sec^{-1} (Dickinson and Geisler, 1968).

approximation, variations in longitude are equivalent to variations in local time. It is difficult to visualize the mean cellular motions (e.g., between the equator and poles within the thermosphere) when such a large diurnal variation is present. Therefore, just as the global mean temperature-profile calculation was instructive in evaluating current knowledge of the thermospheric heat sources and sinks, the next simplest model for examining the thermal and compositional structure is a two-dimensional zonal mean (longitudinally averaged) model. From studies with such a model, it is possible to examine the basic cellular structure operating within the thermosphere and to determine whether the currently understood distribution of thermospheric heat and momentum sources yields a calculated structure that agrees with the observations. Such a study allows the dynamicist to examine the thermospheric equivalent of the Hadley, Ferrel, and polar circulation cells of the tropospheric circulation, long studied by meteorologists.

Dickinson and Geisler (1968) examined the structure of temperature and circulation that was derived from the empirical model based on satellite drag measurements. They found upward motions on the dayside of the earth and downward motions at night, just as would be expected from a solar-driven circulation. However, the real surprise came when they examined the mean meridional circulation. At equinox, there were upward motions and high temperatures at high latitudes and downward motions and low temperatures at low latitudes. That is, the empirical models based on satellite drag measurements required a circulation during equinox that was opposite to what one would expect of a solar-radiation-driven circulation—upward motion over the equator and sinking motion over the pole.

It had been recognized for a long time that there probably is a strong heat source at high latitudes due to auroral processes, but the global area where the heating occurred—the auroral oval—seemed small compared with the rest of the globe that is under direct solar control. The theoretical question was: could the heating at high latitudes due to auroral processes be large enough, even during geomagnetically quiet conditions, to reverse the expected solar-driven circulation? Analysis of additional satellite and ground-based data confirmed that the reverse circulation existed in the zonal mean (i.e., zonally averaged, or mean meridional circulation) down to mid-latitudes during geomagnetically quiet times, and evidence also existed that indicated that it extended as far as the equator during geomagnetic storms.

This apparent discrepancy between theoretical expectations and observational indications has recently been analyzed with a two-dimensional, time-dependent numerical model of the thermosphere in studies done at NCAR by R. E. Dickinson, E. C. Ridley, and the author. When the temperature and circulation structure was calculated using only solar ultraviolet heating as the forcing function, it was found that indeed an equator-to-pole circulation and poleward temperature decrease was attained, as expected. But when a high-latitude heat source due to auroral processes was included in the calculation, it produced the circulation reversal. Its magnitude was found to be comparable to energy provided to the magnetosphere by the solar wind during geomagnetically quiet times, about 10^{18} erg sec^{-1}.

The high-latitude heat source thus proved to be the mechanism responsible for reversing the mean meridional circulation. Even though the high-latitude heat source could be small in the global mean compared with solar EUV heating, its confinement to high latitudes produces a departure from the global mean that is large compared with the solar heating deficit at high latitudes. Consequently, its inclusion in the dynamic model greatly modified the latitudinal temperature gradient and gave an equatorward mean meridional circulation in middle and high latitudes quantitatively in agreement with the observed structure.

The reverse circulation is almost linearly related to the magnitude of the high-latitude energy input, with the result that a wide range of circulation patterns can exist, depending on the magnitude of the geomagnetic activity. The circulation during equinox is shown in Figure 3.10, where the mass flow schematically illustrates the cellular motion in the thermosphere between 80 and 500 km. [The details of the wind and temperature latitudinal structure for this circulation are given by Dickinson *et al.*(1975).] Figure 3.10(a) shows the circulation during a very quiet geomagnetic period; upward motion occurs over the equatorial subsolar point and flows poleward to high latitudes, where sinking motion occurs. The small reverse circulation at high latitudes is due to a small high-latitude heat source. During average geomagnetic activity, the size of the equatorward circulation is larger, as shown in

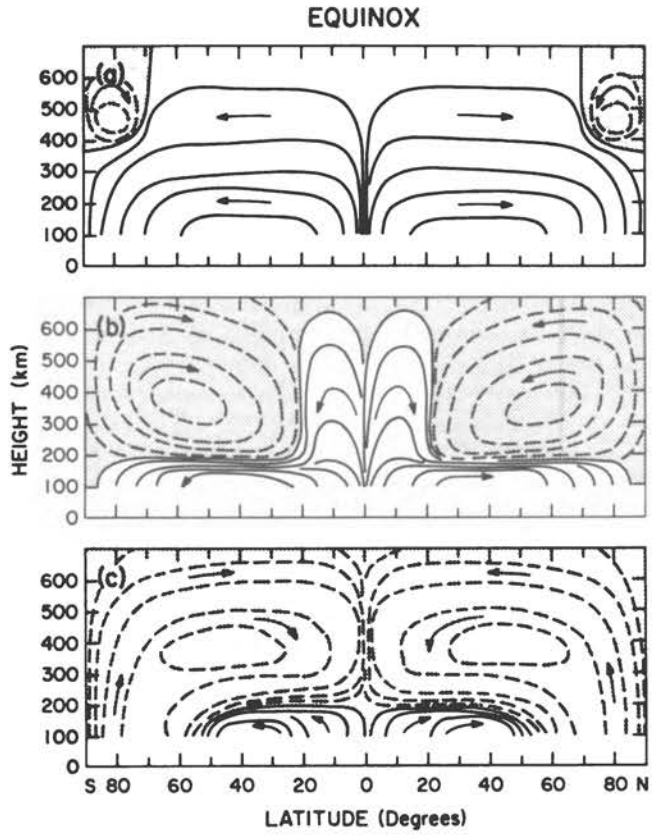


FIGURE 3.10 Schematic diagram of the zonal mean meridional circulation in the earth's thermosphere during equinox for various levels of auroral activity (a) extremely quiet geomagnetic activity, (b) average activity (10^{18} erg sec^{-1}), and (c) geomagnetic substorm (10^{19} erg sec^{-1}). The contours schematically illustrate the mass flow, and the arrows indicate the direction of the motion.

Figure 3.10(b). Below about 150 km, the strong solar ultraviolet heating still maintains a poleward flow; however, above that altitude the flow is equatorward to about 20° latitude. During sizable geomagnetic storms, the equatorward circulation increases greatly, extending right down to the equator, as shown in Figure 3.10(c). The circulation below 150 km is also affected then by the auroral heating.

Thus it is evident that the mean motions within the thermosphere are in a constant state of agitation, depending on the magnitude of the high-latitude heat source. Since this source is dependent on magnetospheric storms and substorms that are energized by solar-wind particles interacting with the magnetosphere, it is quite clear why the thermospheric circulation is forced by both solar electromagnetic radiation and the solar wind.

The situation is similar during solstice. There is a surplus of solar energy in the summer hemisphere and a deficit in the winter hemisphere, by comparison with global mean values. Under geomagnetically quiet conditions, the asymmetry in solar heating drives a summer-to-winter circulation, as shown in Figure 3.11(a). In Figure 3.11(b) the high-latitude heat source associated with average geomagnetic activity reinforces the solar-driven summer-to-winter circulation in the summer hemisphere but forces a reverse circulation in the high-latitude winter hemisphere; the two cells converge in the midlatitude winter hemisphere. Below about 150 km, however, the summer-to-winter circulation is maintained at all latitudes. As the magnitude of the high-latitude heat source increases during geomagnetic storms, the circulation is equatorward in both hemispheres above 300 km with an asymmetry at lower altitudes, as shown in Figure 3.11(c). Thus, even during solstice conditions, the high-latitude heat source plays a strong role in modifying the thermospheric circulation. The factors controlling the magnitudes of the winds associated with the indicated flows are complex. Near 300 km at 40° latitude, the meridional winds during average conditions are 50, 20, and 5 m sec^{-1} equatorward during summer solstice, equinox, and winter solstice, respectively.

The patterns shown here are the mean motions. Certainly during geomagnetic storms there is some division of the energy between the mean motions and waves that can be excited by auroral processes and propagate away from the region of excitation. The partitioning of large-

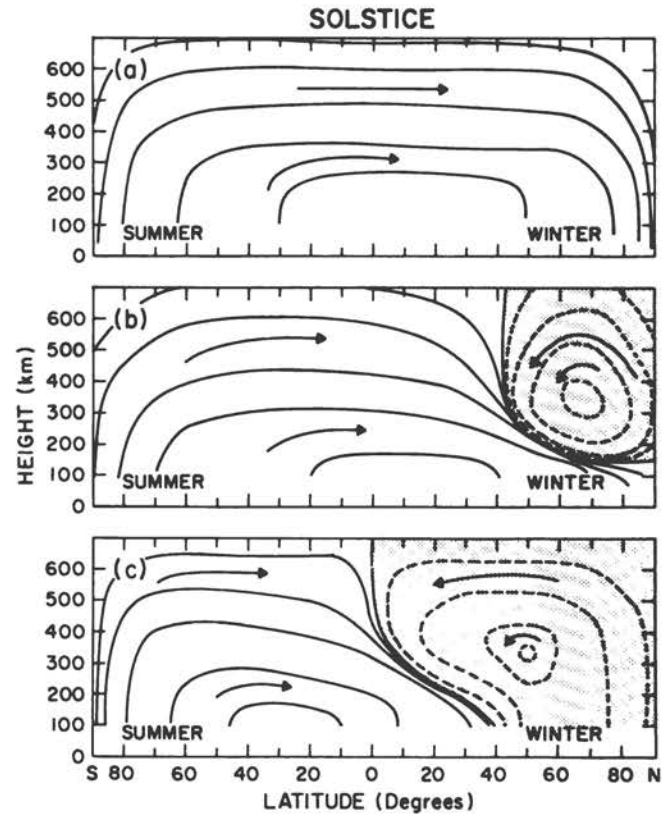


FIGURE 3.11 Same as Figure 3.10 but at solstice.

and small-scale variations and their interactions is a study that should occupy dynamicists for a long time. The important processes that control the thermospheric circulations are beginning to be understood. The study of the three-dimensional circulation within the earth's thermosphere is the next step. There are models that can be used to study the circulation, and more detailed ones will be developed within the next few years. Current knowledge of the solar heating processes seems adequate to prescribe accurately the global distribution of heating by the absorption of solar electromagnetic radiation. The main obstacle to a detailed three-dimensional study of the thermospheric circulation is the need for a three-dimensional specification of the high-latitude heat and momentum sources associated with auroral processes. Accurate descriptions of these processes will improve our understanding of the global circulation of the earth's thermosphere.

3.6 TEMPERATURE AND COMPOSITIONAL STRUCTURE

The temperature and composition latitudinal structures associated with the circulations just described are complex and variable. However, there is an average seasonal behavior that is clearly evident in the data. The zonally averaged neutral gas temperature, mean mass, and composition are shown in Figure 3.12; the values were obtained from the OGO-6 empirical model of neutral composition and temperature based on spherical harmonic analysis of satellite neutral mass-spectrometer measurements (Hedin *et al.*, 1974). During solstice both the temperature and mean molecular weight exhibit a large summer-to-winter hemispherical difference, whereas during equinox the latitudinal variation is small. During equinox both the temperature and the mean molecular weight at 300 km have maximum values in the polar regions and minimum values at midlatitudes. The maximum at high latitudes is apparently due to the high-latitude heat source. This effect is also present during solstice; however, the hemispherical difference due to solar ultraviolet heating dominates. The increase in mean molecular weight is caused by an increase in the relative ratio of the molecular species to the atomic species, whose seasonal distributions are also shown in Figures 3.12(b) and 3.12(c) at 120 and 300 km. The heavy molecular constituents are enhanced in the summer hemisphere, whereas the lighter atomic constituents are enhanced in the winter hemisphere. This type of variation cannot be explained by (molecular) diffusive equilibrium alone, in which the concentrations of all constituents, molecular and atomic, increase with the neutral temperature. According to diffusive equilibrium with fixed lower-boundary conditions near 120 km, all neutral constituents should be enhanced in the summer hemisphere and depleted in the winter hemisphere by comparison with the global mean values. Clearly, another mechanism is required to alter the diffusive equilibrium structure, and it must involve a change in composition at the base of the region of rapid molecular diffusion.

For over a decade, evidence has been gathered that

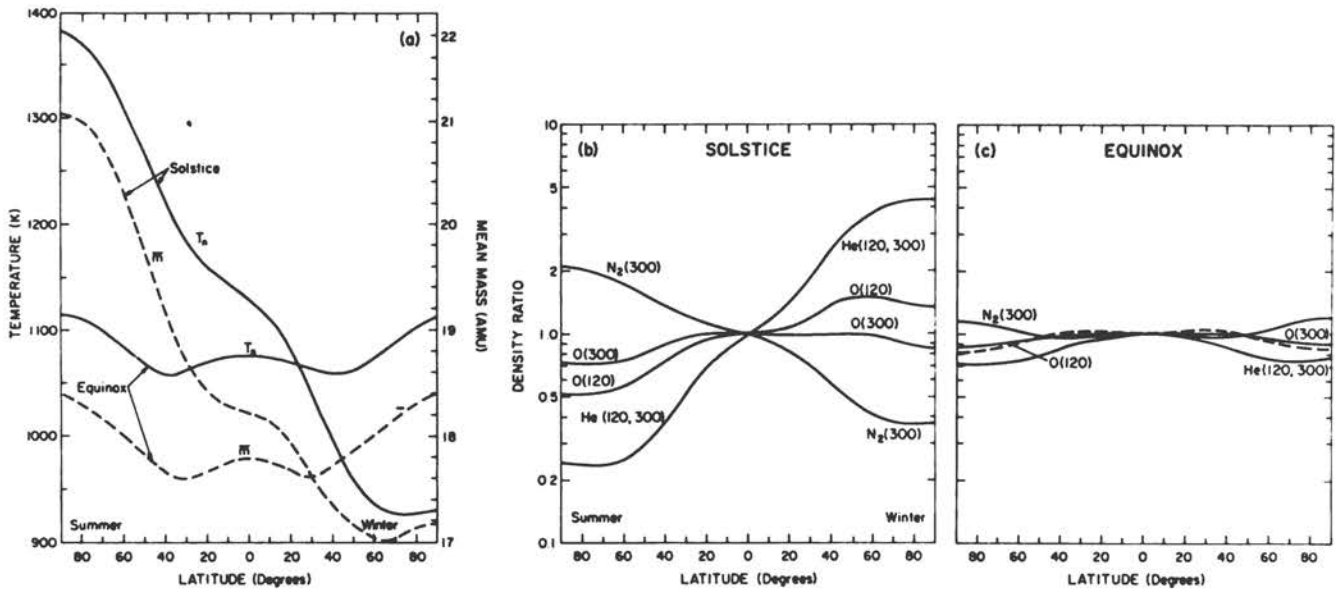


FIGURE 3.12 (a) The zonal mean latitudinal distribution of neutral temperature, T_n (K), and mean molecular weight, $\bar{m}$, at 300 km from the OGO-6 empirical model for solstice and equinox conditions. (b) The density ratios (relative to the equatorial values of the zonal mean distributions) of composition during solstice: molecular nitrogen, N_2, at 300 km; atomic oxygen, O, at 120 and 300 km; and helium, He, at 120 and 300 km. (c) The density ratio for equinox.

indicates that neutral thermospheric winds play an important role in determining the compositional distribution (e.g., Johnson, 1973). In the lower thermosphere, horizontal winds just above the region where diffusive equilibrium begins favor transport of lighter constituents of the atmosphere over the heavier. The lighter constituents essentially constitute thicker atmospheres than do the heavier constituents, and the transport is proportional to the scale height of the constituent being transported (the scale height of the constituent is a measure of the thickness or altitude difference over which its density decreases by the factor $1/e$). This effect is pronounced in the case of helium, whose concentration is almost entirely controlled by transport and not at all by chemistry. Thus, the summer-to-winter circulation causes the gases that are lighter than the mean mass (atomic hydrogen, helium, atomic oxygen) to be depleted over the summer pole with a corresponding buildup over the winter pole. The neutral gas constituents that are heavier than the mean mass, such as molecular nitrogen, molecular oxygen, and argon, all have summer maximum and winter minimum values at a given altitude. If this transport mechanism were not operating, there would be a pronounced summer-pole-to-winter-pole decrease in atomic oxygen because of the reduced oxygen photodissociation at solstice over the winter pole. Such a decrease has not been observed, and actually a winter maximum develops in the lower thermosphere that is associated with transport.

In the upper thermosphere, the changing temperature profile in the 120–200 km region with season has a compensating effect on the concentration changes established by circulation in the lower thermosphere. This is particularly true for atomic oxygen, whose scale height and departure from the mean mass are small enough that it is not so effectively circulation-controlled as helium. While the circulation tends to enhance the atomic oxygen concentration in the lower thermosphere in winter, the latitudinal temperature distribution—colder in the upper winter thermosphere—tends to reduce the concentration of atomic oxygen. Thus, the atomic oxygen seasonal variation in the high thermosphere gives a small summer-to-winter-pole decrease, as shown in Figure 3.12(b).

During geomagnetic storms, satellite observations show dramatic changes in neutral composition from geomagnetically quiet values. The molecular nitrogen concentration at 400 km has been observed to increase by an order of magnitude. The helium concentration has been observed to decrease dramatically. Strangely, the atomic oxygen concentration, however, remains nearly constant at satellite altitudes (near 400 km) during geomagnetic storms. Theoretical studies have shown that the causes of these composition changes are changes in atmospheric circulation and structure brought about by the atmospheric response to intense heating during auroral events. Near the center of the maximum heating, upward vertical motions develop with a resulting change in the thermal structure. Horizontal pressure forces are also generated that cause horizontal neutral winds to flow away from the region of intense heating to subauroral latitudes. The same circulation and thermal effects that explain the seasonal variations occur here but on a much smaller scale. The vertical and horizontal transport depletes constituents whose mass is less than the mean mass and increases the constituents whose mass is greater than the mean. Thus, hydrogen, helium, and atomic oxygen decrease, and molecular nitrogen, molecular oxygen, and argon increase. While atomic oxygen is decreased by circulation in the lower thermosphere, this is compensated for by thermal expansion at higher altitudes, with the net effect of no observed change at the satellite altitudes.

The observed thermospheric compositional response to a geomagnetic storm is illustrated schematically in Figure 3.13, in which the compositional displacements are seen to propagate out from the auroral zone. After the storm, the neutral atmosphere relaxes as the heating subsides. The displaced composition then generates an equivalent pressure force that causes a return circulation to take place until the prestorm composition distribution is reestablished.

The compositional pattern described above seems to develop whenever there is strong thermospheric heating. It is apparent in satellite data as a persistent feature in regions where heating is known to occur at all times, such as in the vicinity of the polar cusp where solar-wind particles have direct access to the earth's thermosphere and heating is always present. Another strong heating region is the auroral oval; the neutral composition there differs from that of midlatitudes, where the heating is absent.

Recently, compositional effects have been shown to be associated with small-scale atmospheric waves at thermospheric heights. It is known that waves are prevalent in the thermosphere, and observations have shown that the heavy atmospheric constituents are in phase with each other and out of phase with the constituents that are lighter than the mean mass.

3.7 OUTSTANDING PROBLEMS

Much progress has been made in understanding the gross features of the temperature, composition, and dynamic structure of the earth's thermosphere. Yet, as more data are analyzed, it is becoming evident that dynamic phenomena on just about all scale sizes are present within the thermosphere with large variability. The most difficult problem that will limit progress is in understanding the cumulative interactions of small- and large-scale dynamics that occur within the thermosphere, and with the magnetosphere at the upper boundary and with the stratosphere and mesosphere at the lower boundary.

The Atmosphere Explorer satellite program is making an excellent contribution to our understanding of the interaction of solar electromagnetic energy with the gases within the thermosphere. Many long-standing problems

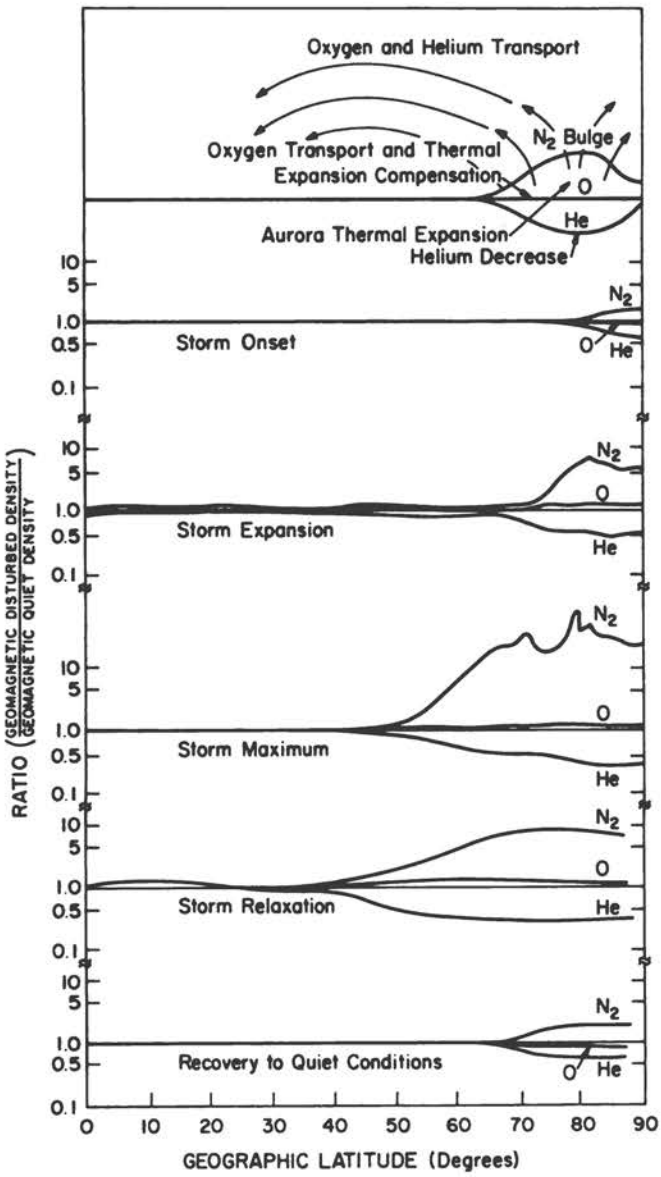


FIGURE 3.13 Schematic of the thermospheric compositional response to geomagnetic storms as a function of latitude at various times during the storm. The ratio is storm-time departure relative to geomagnetically quiet conditions.

are being solved by this effort, which is occurring during solar minimum, a time when the solar electromagnetic radiation effects have the maximum influence on the basic thermospheric structure.

The next most important problem for thermospheric dynamics is to investigate the energy, momentum, charge, and mass exchange problems that occur within the magnetospheric–ionospheric–thermospheric system in auroral processes. This problem will be attacked during the next few years by the International Magnetospheric Study (IMS) and the planned Electrodynamics Explorer satellite program. Both of these efforts will occur as solar maximum is approached, a time when the high-latitude energy and momentum sources have their greatest influence on the thermospheric structure.

Finally, the couplings between the thermosphere and lower atmosphere will emerge as a new scientific frontier of thermospheric research during the Space Shuttle era. The effects of upward propagating planetary, tidal, and gravity waves on the thermospheric structure present an important problem that is poorly understood at present, but this will be examined as the middle atmosphere is explored by the Space Shuttle remote-sensing instruments. Another important problem is to determine how the auroral energy that is deposited within the thermosphere is ultimately dissipated by radiative processes in the lower thermosphere and mesosphere. The chemistry and exchange transport of minor neutral constituents and various excited species between the thermosphere and the mesosphere will most likely emerge as a major scientific problem, much as did the role of minor constituent chemistry in the stratospheric ozone problem. These are necessary studies to determine whether the highly variable upper-atmospheric processes can influence processes within the lower atmosphere and thus affect our immediate environment.

There undoubtedly exists a rich variety of solar–planetary interactions within our own and other solar systems. By understanding the interaction of the solar-electromagnetic and solar-wind energy with our own planet we should gain the insight necessary to speculate more effectively on these interactions with other planetary bodies.

The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

Carignan, G. R. (1974). Thermospheric composition, *Rev. Geophys. Space Phys. 13*, 885.

Dickinson, R. E. (1975). Meteorology of the upper atmosphere, *Rev. Geophys. Space Phys. 13*, 771.

Dickinson, R. E., and J. E. Geisler (1968). Vertical motion field in the middle thermosphere from satellite drag densities, *Mon. Weather Rev. 96*, 606.

Dickinson, R. E., E. C. Ridley, and R. G. Roble (1975). Meridional circulation in the thermosphere. I. Equinox conditions, *J. Atmos. Sci. 32*, 1737.

Evans, J. V. (1975). A review of F region dynamics, *Rev. Geophys. Space Phys. 13*, 887.

Hedin, A. E., H. G. Mayr, C. A. Reber, N. W. Spencer, and G. R. Carignan (1974). Empirical model of global thermospheric temperature and composition based on data from the OGO-6 quadrupole mass spectrometer, *J. Geophys. Res. 79*, 215.

Jacchia, L. G. (1965). Static diffusion models of the upper atmosphere with empirical temperature profiles, *Smithsonian Contrib. Astrophys. 8*.

Johnson, F. S. (1973). Horizontal variations in thermospheric composition, *Rev. Geophys. Space Phys. 11*, 741.

Johnson, F. S., and B. Gottlieb (1970). Eddy mixing and circulation at ionospheric levels, *Planet. Space Sci. 18*, 1707.

4 Hydrogen

THOMAS M. DONAHUE
The University of Michigan

4.1 PROLOGUE

At one time, it was believed that the atmosphere of the earth at altitudes greater than 75 km consisted predominantly of molecular hydrogen gas, H_2. The reason was simple and as sound as some arguments produced today for many respectable atmospheric models. By 1910, it was known from measurements made on sounding balloons that above 10 km the atmosphere becomes calm, stratified, and isothermal at a temperature of 220 K. Hence it was argued by authorities as renowned as Jeans, Humphreys, and Wegener (see Tinsley, 1974; Chamberlain, 1973) that turbulent (eddy) mixing ceased and gravitational diffusive separation began at 10 km. Since the mixing ratio of H_2 in the troposphere was believed to be about 2×10^{-5}, it followed that H_2 should begin to be the dominant constituent of the atmosphere at about 75 km.

Not much higher than 100 km, collisions between H_2 and other gases would become so infrequent that the base of the exosphere or (as Wegener called it) the geocorona would be reached. This is the region in which molecules and atoms can be assumed to travel on Keplerian orbits. It was under the assumption that the exospheric temperature was only 220 K that Jeans developed his classical theory of thermal escape of gases and concluded that under those circumstances the escape rate of hydrogen from the earth would be insignificant.

In fact, Jeans, with whose name the mechanism of thermal escape of a gas from a planet is associated, was anticipated in the formulation of the concept and its publication as well by some 50 years by Waterston (Tinsley, 1974). Although Waterston knew nothing about the kinetic theory of gases and considered his atmospheric particles to be monoenergetic, G. J. Stoney (Tinsley, 1974) in three papers published during the nineteenth century did understand that a few particles in the high-velocity tail of the Boltzmann distribution would have energy sufficient to escape even if the average particle did not.

Thus the concepts of an extensive geocorona of light gases and the escape of energetic members of their population from the earth were developed long before the

rapid growth of knowledge that accompanied the use of sounding rockets to explore the upper atmosphere following the Second World War. In 1955, the Naval Research Laboratory (NRL) group working under Herbert Friedman detected hydrogen Lyman-α emission above 75 km in the night sky, and in 1957 they constructed a map, shown in Figure 4.1, of its intensity as measured from about 120 km when the sun was 44° below the horizon. The emission rate decreased from the horizon to a minimum located not far from the antisolar direction. In 1959, a high-resolution spectrum of the broad solar Lyman-α emission, obtained from a rocket near 134 km by Richard Tousey's group at NRL, showed a deep narrow absorption feature, shown in Figure 4.2, that had to be attributed to relatively cool hydrogen between the rocket and the sun.

By this time, information concerning atmospheric densities in the upper thermosphere obtained from the effect of atmospheric drag on satellites, and from other sources as well, had shown that the exospheric temperature was greater than 1000 K, far above the 220 K assumed by Jeans. At such temperatures, the escape flux of hydrogen from the earth would be very large compared with the values calculated by Jeans.

4.2 EXOSPHERIC DENSITY MODELS

Just after the Lyman-α nightglow (wavelength 121.6 nm) was discovered, a controversy developed concerning its source—whether it was produced by resonance scattering of solar photons by interplanetary or by geocoronal hydrogen. The geocoronal hypothesis was developed by Francis Johnson and his collaborators and led to the development of models for the distribution of hydrogen in the exosphere by Johnson and Fish and by Öpik and Singer in 1961. The theories that led to these models were based on the concept of the existence of a critical surface some 500 km above the earth's surface, defining the altitude at which collisions between hydrogen atoms and oxygen became so infrequent that the atoms could be considered to follow ballistic trajectories above that altitude. Those with greater than escape velocity constituted the part of the Boltzmann distribution that would provide the Jeans escape flux if they were traveling upward at the time of their last collision. Figure 4.3 shows the velocity distribution calculated at two altitudes. Portions of the full Maxwell-Boltzmann distribution corresponding to orbiting or re-entering particles on hyperbolic orbits are missing. From these distribution functions, a density profile could be calculated if the exospheric temperature, the density at the base of the exosphere, and the fraction of the population on elliptic trajectories whose perigees lay above the base of the exosphere, or critical level, should be known.

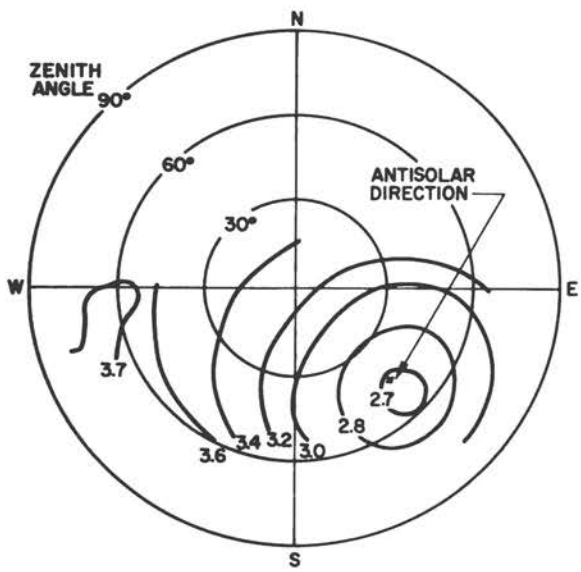


FIGURE 4.1 Directional intensity contours of Lyman-α radiation obtained above 120 km with the sun 44° below the horizon. The units are 10^{-10} W cm^{-2} sr^{-1}.

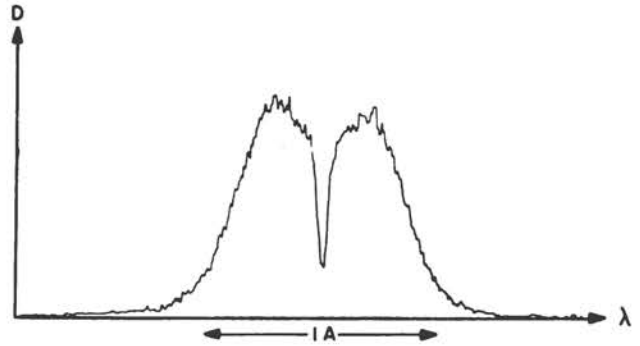


FIGURE 4.2 Microphotometer tracing of the solar Lyman-α profile obtained above 134 km, showing the narrow absorption core due to hydrogen between the spectrometer and the sun.

The theory of the exospheric distribution was developed later in an elegant form based on Liouville's theorem by Chamberlain (1973). His theory treats the orbiting particles by assuming that they are effectively at the same temperature as the other atoms but deals with the problem of accounting for their origin and presence by introducing the concept of a satellite critical level R_{sc}. All possible orbits allowed by the Boltzmann distribution with perigees that lie at or below a specified R_{sc} are assumed to be present in the exosphere. Physically they would find themselves on such trajectories as a result of collisions that they undergo after leaving the critical level on another orbit. Chamberlain's exospheric distributions for various values of temperature at the critical level are shown in Figure 4.4 for the case in which there are no satellite orbits (i.e., $R_{sc} = R_c$).

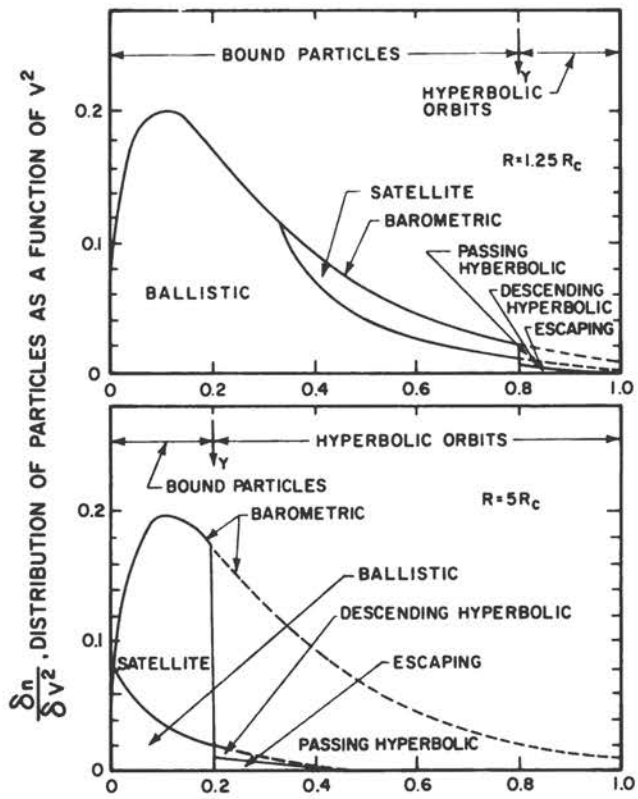


FIGURE 4.3 Distribution of exospheric particles at two altitudes as a function of V^2, the ratio of the square of the particle velocity to the square of the escape velocity at the critical level. $V^2 = y = R_c/R$ separates bound from free particles.

EARLY RADIATIVE TRANSPORT CALCULATIONS

To determine what must be the characteristics of the hydrogen geocorona that would permit it to produce the Lyman-α nightglow radiation as a result of resonance scattering of solar photons, it was necessary to develop a method of solving the radiative transport equation for resonantly scattered photons in a spherical geometry. This was done by Gary Thomas in his doctoral thesis at the University of Pittsburgh in 1961 with a calculation based on the transport theory of Holstein. Thomas and Donahue found the resonance scattering to be inadequate to account for the nightglow observed by the NRL group if the distribution of hydrogen was spherically symmetrical and contained only as much hydrogen (about 3×10^{12} atoms/cm²) as indicated by the absorption core in the solar Lyman-α line, and if the solar flux was that measured at the center of that line. They pointed out that the nightglow results in 1957 and the absorption spectrum in 1959 could be reconciled in a resonance scattering model if there were about 3 times as much hydrogen present in the atmosphere at night as during the day. Such a distribution would permit multiple scattering to transport an adequate number of Lyman-α photons around to the nightside of the earth.

THERMOSPHERIC HYDROGEN DISTRIBUTION WITH FLOW

That such a diurnal variation in hydrogen density might occur was suggested at about that time in the work of Mange, of Bates and Patterson, and of Kockarts and Nicolet (see Tinsley, 1974). These authors attacked the problem of determining the distribution of hydrogen in the thermosphere between a lower boundary at 100 km and the base of the exosphere in the presence of a given flow of hydrogen. This flux must be the same as the escape flux determined by the exospheric temperature and the hydrogen density at the critical level. The calculations assumed either a flux and a fixed density at the lower boundary (Kockarts and Nicolet) or a density and temperature at the critical level that was consistent with the variation in hydrogen abundance inferred from the absorption measurements between 134 and 163 km. These measurements implied that there were 5.7×10^{12} atoms cm^{-2} above 100 km and 1.8×12 atoms cm^{-2} above 200 km. The calculations were performed by solving the molecular diffusion equation for the vertical flux of a minor constituent such as hydrogen in a two-component atmosphere in a gravitational field. A diffusive flow of gas

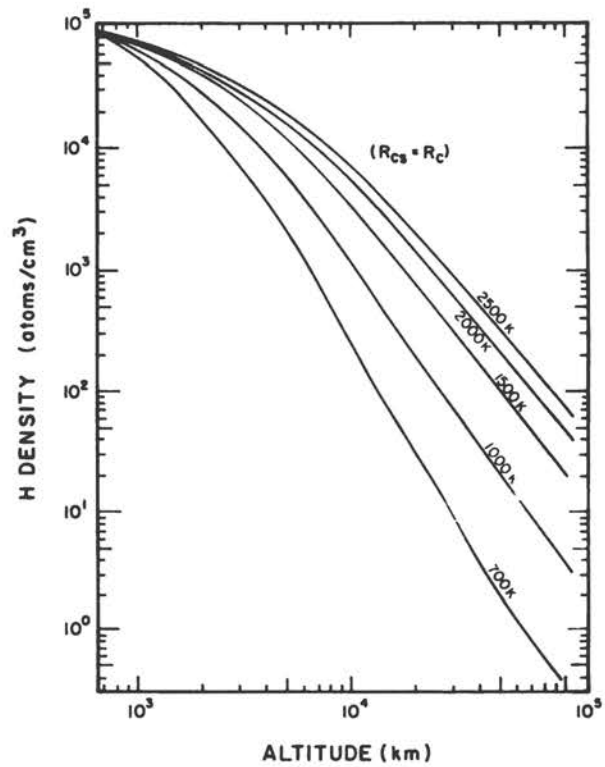


FIGURE 4.4 Exospheric density without orbiting particles for various values of the exospheric temperature.

species i through another in a field-free environment occurs when there is a density gradient of the ith species. But in a gravitational field the vertical distribution, even when the species i is not flowing, calls for a density gradient to counteract the force field. The gradient for zero flow is given by the familiar barometric equation

$$\frac{dn_i}{dz} = \frac{n_i}{H_i} + \frac{1 + \alpha_i}{T} \frac{dT}{dz}, \tag{4.1}$$

where $H_i = kT/m_i g$ is the scale height of species i and α_i is the thermal diffusion coefficient, n_i and m_i are the concentration and particle mass for the ith species, T is the temperature, g the acceleration of gravity, and z the altitude. A departure of the gradient from the value indicated in Eq. (4.1), will set up a flow given by the product of a diffusion coefficient D_i and the difference between the gradient for zero flow, as indicated by Eq. (4.1), and the actual density gradient. If the density decreases more rapidly than indicated by Eq. (4.1), the flow will be upward, and if less rapidly, it will be downward. The diffusion coefficient for hydrogen in the lower part of the thermosphere, which depends inversely on the density of the dominant gas, is small, and large fluxes can be supported only by large departures of the hydrogen density profile from diffusive equilibrium profile, i.e., by large gradients relative to the equilibrium gradient. On the other hand, at high altitude (250 km and above), D_i becomes very large and the same fluxes call for very little departure from the diffusive equilibrium gradient. In these early calculations, the upward flux through the thermosphere was supposed to match the Jeans escape flux at the base of the exosphere. This escape flux ϕ_J can be visualized as a product of the hydrogen density at the critical level and an effusion velocity w_c, which increases rapidly with the exospheric temperature T_c,

$$\phi_J = n_c w_c. \tag{4.2}$$

Thus, with n_c and T_c (or w_c) specified, the flux is fixed and the value of n_c determines just how the hydrogen density must vary through the thermosphere in order to support the flux ϕ_J. If ϕ_J is to be independent of T_c and determined by the source below, then n_c must decrease with increasing T_c—a situation that tends to require lower hydrogen densities near the subsolar point than near the antisolar point and an increase in average exospheric hydrogen densities as solar activity wanes and the upper atmosphere cools.

Kockarts and Nicolet assumed that the hydrogen density at 100 km was 10^7 atoms cm^{-3} and the flux 2.5×10^7 atoms cm^{-2} sec^{-1}. Radiative transport analysis of Lyman-α airglow data from a large number of rocket flights was summarized by Donahue (1966); he suggested that the results could be accounted for if the hydrogen distribution followed that of the Kockarts and Nicolet model normalize to 3×10^7 atoms cm^{-3} at 100 km and a standard exospheric model (e.g., Öpik and Singer) above the critical level, provided that there were a sizable diurnal variation at the exobase. Because of the nonsystematic way in which the early exploratory rocket data were gathered and questionable sensor calibrations, it was far from clear at that time how large a diurnal variation was required to account for the airglow.

4.3 SATELLITE OBSERVATIONS OF LYMAN-α AIRGLOW

The early observations of Lyman-α radiation from sounding rockets and the attempts to derive information concerning thermospheric and exospheric hydrogen densities from them have been supplemented in the years since then by a long series of detailed observations of the Lyman-α airglow from satellites (including the moon) and by other techniques. For the most part, the observations have elucidated the nature of the exospheric distribution rather than the thermospheric distribution. The lower thermosphere remains even today a sphere of ignorance as far as hydrogen is concerned, as it is for many other atmospheric properties.

The earliest satellite observations and their analyses roughly confirmed that the exospheric distributions were similar to those predicted by Chamberlain, although there was considerable uncertainty concerning the value to be assigned to the critical satellite radius R_{sc}. Values of R_{sc} in the neighborhood of 2.5 R_e (earth radii) were favored, however. Diurnal variations in density at the exobase by a factor of about 1.7 with the maximum at about 0500 hours and minimum about 1600 hours were indicated. Recent observations by Bertaux from OGO-5, by Vidal-Madjar and Blamont from OSO-5, by Thomas and Bolin from OGO-5, and by Carruthers and his colleagues (1976) from Apollo 16 have, however, yielded such a rich harvest of detail concerning the exosphere that it is sufficient for our purposes to discuss what they have found.

RECENT RESULTS: THE PERTURBATION OF ORBITING ATOMS

The Service d'Aeronomie experiment flown aboard OGO-5 by Bertaux and Blamont consisted of a Lyman-α photometer in front of which was an atomic hydrogen absorption cell. The optical depth of hydrogen in the cell could be varied. The variation in the transmitted airglow Lyman-α intensity with optical depth in the filter yielded information from which could be deduced the atmospheric line width and thus the temperature of the exospheric hydrogen. The orbit of OGO-5 was very eccentric. Detailed information concerning the variation of hydrogen with distance from the earth was obtained out to very large distances (16 R_e). The inferred density distribution was obtained by radiative transport analysis of the observed Lyman-α emission rates. Three times while OGO-5 was near apogee, the orientation of the spacecraft was varied so that much of the celestial sphere, including the

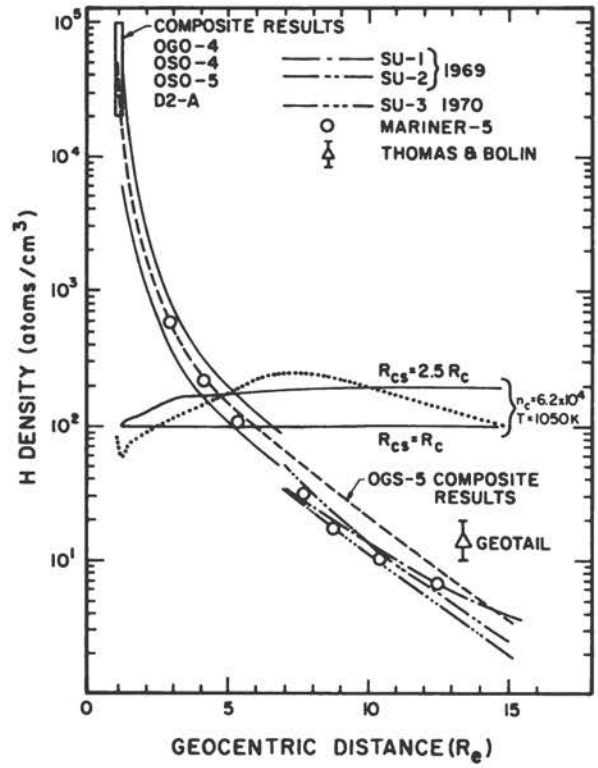


FIGURE 4.5 Hydrogen density in the geocorona according to the measurements of Bertaux and Blamont on OGO-5, indicated by dashed lines, the nearly solid lines indicating the variance. Measurements by others are also known. SU stands for "spinup" of OGO-5.

geocorona, was scanned. Results, along with those obtained by some other experimenters, are summarized in Figure 4.5. The composite OGO-5 results are shown by the dashed line. The two solid lines surrounding the dashed line out to 7 R_e indicate the variance in the results. Data obtained during the three "spinups" (SU-1, SU-2, and SU-3), together with those gathered during the Mariner 5 exit from earth and the Thomas and Bolin (OGO-5) observation near 14 R_e in the antisolar direction, are also plotted. The approximately horizontal line marked $R_{sc} = 2.5\,R_c$ illustrates how the density should have varied relative to the Chamberlain model density with no satellite particles ($R_{sc} = R_c$) if all allowed orbiting particles with perigees below 2.5 R_c were present, assuming an exospheric temperature of 1050 K and a column density $n_c = 6.2 \times 10^4$ atoms cm^{-2}. The observed average density relative to the Chamberlain model with no satellite-orbiting particles (represented by the dotted line) did not behave at all like any model with a fixed R_{sc} for satellite particles. There was a deficiency in the total distribution below 3 R_e, that is, there were fewer particles there than would have been expected even if there were no satellite atoms at all. At 7 R_e, the critical satellite radius needed to produce the observed density was higher than 2.5 R_c, and at 15 R_e again no satellite particles with perigees above the critical level were required. Furthermore, the observations showed a deficiency of orbiting atoms in the solar direction and an excess in the antisolar direction. Somewhat similar results indicating that the geocorona has a kind of tail, or excess of atoms in the antisolar direction, were obtained by Thomas and Bolin, who also observed Lyman-α radiation with a spectrometer on the same satellite.

Graphic and dramatic illustration of the same effects were produced by the Naval Research Laboratory far-ultraviolet camera/spectrograph operated on the lunar surface during the Apollo 16 mission on April 21–23, 1972 (Carruthers *et al.*, 1976). Figure 4.6 shows a composite image of the geocorona obtained by this camera in Lyman-α radiation. Analysis of the data indicates, as do the OGO-5 results, good agreement with the Chamberlain model with no satellite atoms near 3 R_e, with a deficiency of atoms in the far up-sun direction and a pronounced geotail in the antisolar direction. Figure 4.7 compares the emission rates as a function of distance radially toward the sun and away from the sun with those predicted by radially symmetric Chamberlain models for various

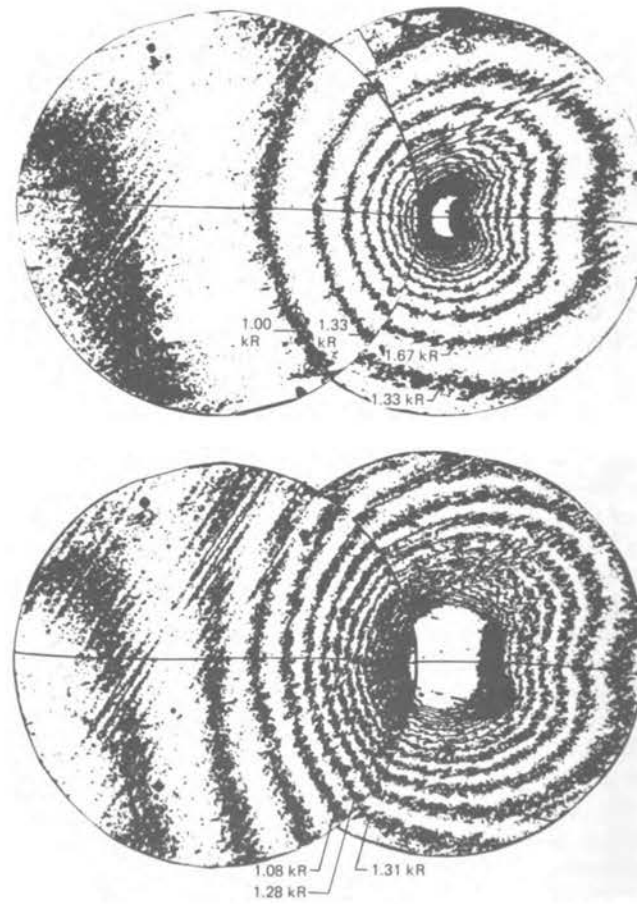


FIGURE 4.6 Mosaic of isointensity contours in the band 105–160 nm obtained by Carruthers *et al.* (1976) from the moon during the mission of Apollo 16. The sun is to the left.

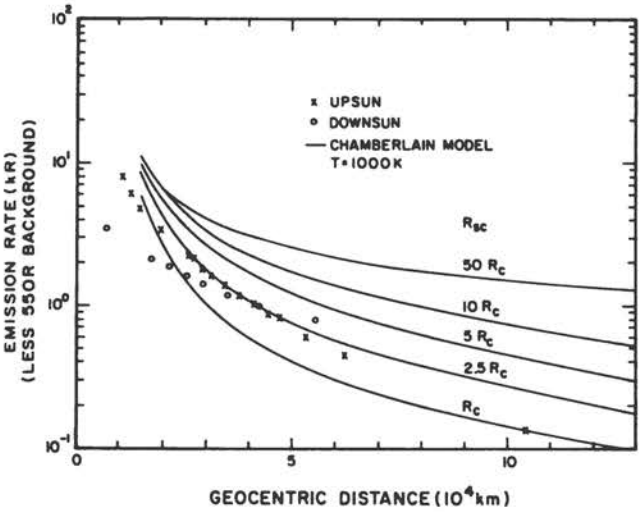


FIGURE 4.7 Observed Lyman-α emission rate in kiloraleighs from the moon during Apollo 16 compared with the expected values for radially symmetric Chamberlain models at a T_c of 1000 K and various values of R_{sc}.

values of R_{sc}; this clearly shows the deficiency in the direction of the sun and the excess in the antisolar direction. The asymmetry about the sun–earth line at distances far from the earth toward the sun is curious, and no explanation for it has been offered. There is a possibility that it is not a geocoronal effect but represents a variation in background interplanetary Lyman-α emission. The model densities that were used to compute Lyman-α emission rates for comparison with those observed are shown in Table 4.1, where T_c is taken to be 1000 K.

The explanation offered by Bertaux and by Thomas and Bolin for the departures of the observed geocoronal densities from those predicted by Chamberlain is based on the effect of solar Lyman-α radiation pressure on the hydrogen atoms executing satellite orbits. The most thorough analysis has been presented by Bertaux in his thesis,

TABLE 4.1 Model H Densities

Altitude (km)	H Density ($R_{sc} = R_c$)[a]	H Density ($R_{sc} = 2.5\,R_c$)[a]
100	3(7)	3(7)
200	1.6(5)	1.6(5)
500	8(4)	8(4)
1000	4.9(4)	5(4)
5,000	4.1(3)	4.8(3)
10,000	8.2(2)	1.2(3)
50,000	1.4(1)	2.7(1)
100,000	2.6	5.2

[a]The numbers in parentheses indicate, in powers of 10, what the preceding number should be multiplied by.

and it can be discussed most readily with the help of Figure 4.8. Radiation pressure has the effect of causing orbits such as that marked 1, with the line of apsides nearly normal to the sun–earth line to be slightly compressed with a lowering of both perigee and apogee, as indicated by the orbit marked 1′. Orbits such as that marked 2, with their lines of apsides initially near 60° from the solar direction, have their lines of apsides rotated by about 120° in the antisolar direction, and their apogees are increased as indicated by orbit 2′. Orbits with lines of apsides very close to the sun–earth line, like that marked 3, are severely deformed, becoming much less eccentric with their apogees pushed in toward the earth.

These radiation pressure effects predict qualitatively the newly observed features of the geocorona. The virtual absence of any orbiting particles below 2.5 R_e is accounted for because of the rapidity with which radiation pressure pushes the perigees of atoms in this region back into the thermosphere. The geotail, or the excess of particles beginning at 4 R_e and continuing far down sun from the earth, is explained by the tendency of the lines of apsides of orbits such as that marked 2 of Figure 4.8 to align themselves with the sun–earth line and for the apogees of such orbits to be raised. The compression of orbits like those marked 3 and 1 results in a relative increase in the atomic population about 3 R_e, particularly in the direction toward the sun, and a decrease in the population at greater distances near 9 R_e, an effect also noted by Wallace and his co-workers in connection with the Mariner 5 observations.

The volume of the region in space containing orbits that are pushed eventually toward the subsolar point (such as $3 \rightarrow 3'$ in Figure 4.8) is greater than that containing particles where orbits like that marked 1 are in a relatively stable configuration nearly perpendicular to the sun–earth line. Hence, there is a tendency for a bulge to develop in the solar direction near 3 R_e. Finally, there is a tendency for photoionization to reduce the population of

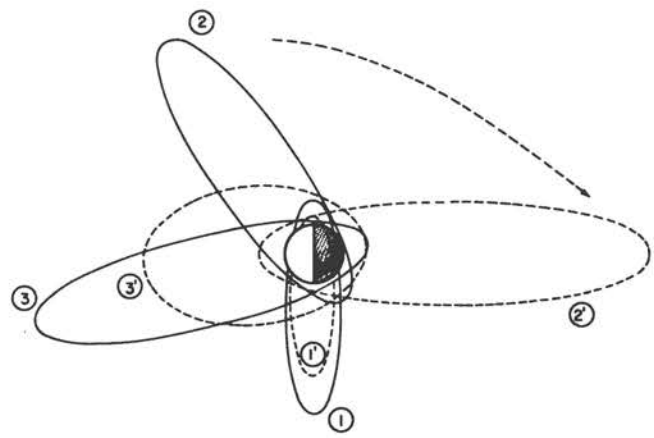


FIGURE 4.8 The three principal effects of radiation pressure on hydrogen satellite orbits in the exosphere, where the primed orbits show the effects of perturbation of the unprimed orbits.

atoms in very eccentric orbits whose perigees lie in the antisolar direction and that spend much of their lifetime in the sunlight while near apogee in the direction of the sun.

CHARGE EXCHANGE BETWEEN H AND O^+

In addition to the interpretation of Lyman-α scattering and absorption data, one other important source of information concerning the hydrogen distribution in the thermosphere has been exploited. This method makes use of satellite measurements of the H^+ density, the O^+ density, and either inferences of the O density from models or direct measurements of this quantity as well. The density of H may then be obtained in regions where the large cross section (3×10^{-15} cm^2) at thermal energies for the reaction

$$H + O^+ \leftrightarrows H^+ + O \qquad (4.3)$$

keeps the ions and neutrals involved in a charge-exchange equilibrium, as W. B. Hanson and his colleagues have explained (Tinsley, 1974). Then detailed balancing requires

$$\frac{n(H)}{n(O)} = K \frac{n(H^+)}{n(O^+)}, \qquad (4.4)$$

where K is 8/9 when the neutral and ion temperatures are equal and high enough.

4.4 DIURNAL VARIATIONS

MEASUREMENTS

Brinton and Mayr and also Breig *et al.* (1976) have used satellite data, and Ho and Moorcroft have used incoherent-scatter data, to infer the diurnal variation of hydrogen densities near the critical level from the relationship discussed in the previous section. Others, Meier and Mange as well as Vidal-Madjar and Bertaux, have also determined the diurnal variation using satellite Lyman-α observations. From our earlier discussions it will be recalled that diurnal concentration variations by a factor of 3 or larger appeared to be required to account for the early sounding-rocket airglow results. However, at least until recently, the observations just cited could be summarized as indicating that the ratio is 1.8 ± 0.3 for latitudes near 25° for a minimum diurnal T_c value of about 900 K (i.e., near solar maximum) and 2.0 ± 0.4 for a minimum diurnal temperature in the neighborhood of 650 K (i.e., near solar minimum). The maximum occurs at 4 ± 2 hours, and the minimum at 16 ± 1 hours local time. Except for the recent results of Breig *et al.*, this work is summarized by Tinsley (1974).

THEORY

The development of models designed to account for the diurnal variation in H concentration has a long history. It begins with calculations of the lateral flow in the exosphere of hydrogen driven by two conflicting tendencies. One is the temperature gradient that tends to send particles from the dayside of the earth toward the nightside because of their higher velocities, and the other is the density gradient that sends more atoms in the opposite direction. The first ambitious calculation of this effect using spherical geometry was carried out by McAfee. McAfee, in his thesis, following a suggestion by Donahue, introduced the concept that the steady-state distribution at the exobase might be one in which the density adjusted itself so that the net ballistic flow would vanish everywhere. These early efforts were followed by time-dependent studies that have grown more and more comprehensive. The latest is one due to Tinsley *et al.* (1975). Their treatment equates the net flux of atoms into a column whose base is in the thermosphere somewhere below the critical level to the horizontal divergence of the hydrogen contained in the column as it is transported laterally by thermospheric winds, rotation of the thermosphere, and exospheric trajectories. The flux into the column is the diffusive flux arising in the stratosphere, and the efflux consists of lateral flow and thermal escape plus contributions from charge exchange of H with H^+ and O^+, which will be discussed in some detail later in this chapter. The results are shown in Figures 4.9 and 4.10.

The models that fit the data best seem to be those that call for a charge-exchange flux that varies diurnally from 0 to 2×10^8 cm^{-2} sec^{-1} and a thermospheric wind velocity

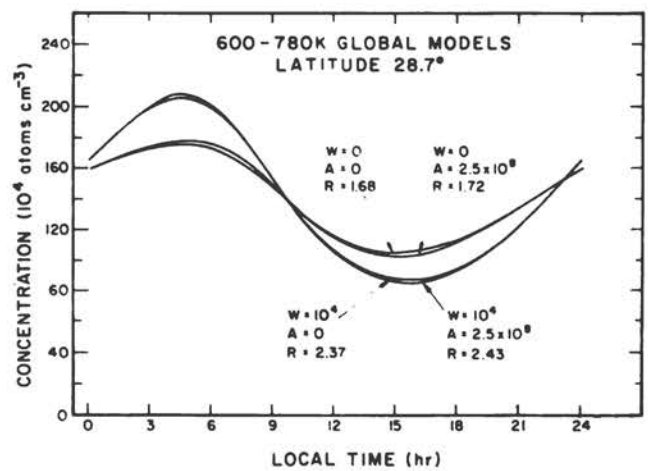


FIGURE 4.9 Diurnal variation of exobase H concentration according to Tinsley *et al.* (1975) at solar minimum. *A* is the maximum diurnal value of the charge exchange flux (cm^{-2} sec^{-1}), *w* is the tidal wind amplitude in cm sec^{-1}, and *R* is the ratio of maximum to minimum hydrogen densities.

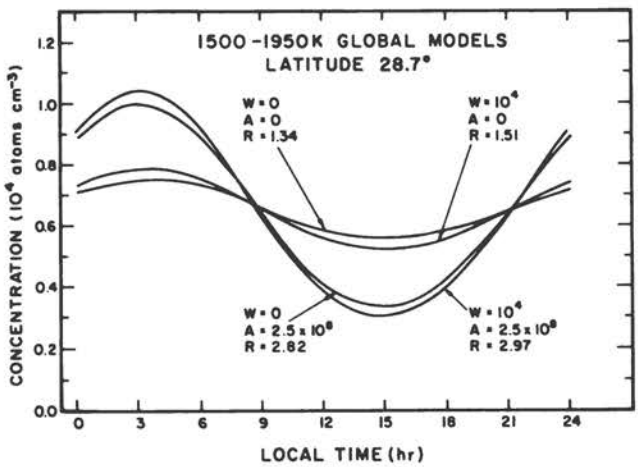


FIGURE 4.10 Same as Figure 4.1 but for solar maximum conditions.

of at most 50 m sec^{-1} at the critical level. This wind value is rather low compared with the calculated wind systems. In the following section we shall see that such large charge-exchange fluxes have been predicted theoretically for other reasons. Recently Brinton *et al.* (1975) have reported a new set of observations using data on simultaneously measured values of H^+, O^+, and O at 250 km obtained in the winter of 1974 by Atmosphere Explorer C. The amplitude of the equatorial variation for an average temperature of 800 K was 3.2. At 40° latitude, it was 2.6. There was a phase lag of 1 hour in the variation of H concentration compared to the temperature. These results, strikingly at variance with earlier measurements, appear to fit the theory of Tinsley *et al.* (1975) if the horizontal wind velocity is raised to 100 m sec^{-1}.

4.5 FACTORS THAT DETERMINE THE ESCAPE FLUX

INTRODUCTION

We return to the subject with which we began the discussion. The unique property of hydrogen in planetary atmospheres is that it can escape at a significantly high rate from the atmospheres of terrestrial planets and the satellites of the major planets. The rate of thermal escape from the base of the exosphere is given by the Jeans formula, Eq. (4.2), where w_c, as we pointed out, is determined by the relative number of atoms in the Boltzmann distribution that have energies large enough to escape the gravitational field. The critical level is usually assumed to be a well-defined surface. Corrections in this relationship for the effects caused by the assumption of the existence of a simple surface defining the critical level above which all atoms follow ballistic trajectories have been discussed by Brinkman and by Chamberlain and Smith. In reality, the population of escaping atoms is cooled near the escape level by the escape of the most energetic particles. This phenomenon (a sort of evaporative cooling) leads to a reduction in escape flux by a factor of 0.72 for earth.

A popular misconception concerning the escape flux is that the density of hydrogen at the critical level is determined by atmospheric structure, and the escape flux is thus determined by the product $n_c w_c$. Those who have treated the problem more carefully have assumed more or less explicitly that the escape rate was set by the upward flux of hydrogen emerging from the lower thermosphere and upper mesosphere following the photolysis of water vapor. It has been rather vaguely assumed that a certain net upward flux would be determined at the source and that the hydrogen distribution in the thermosphere and the density at the critical level would adjust themselves to support this flux, which would be carried away at the effusion velocity, w_c, set by the exospheric temperature. But those who have dealt with the problem of the distribution of minor constituents in the mesosphere have never imposed an upward flux boundary condition on their models. Thus the link between production of H from photolysis and the apportionment of the amount produced between upward flow and recombination was not established until recently.

LIMITING FLUX: HUNTEN'S PRINCIPLE

It is to D. M. Hunten that credit is due for the conceptual breakthrough that has changed our way of looking at steady-state escape. He did so by introducing the concept of limiting flux, emphasizing the control exercised by vertical transport near the homopause. He argues that, for a wide range of conditions, the upward flux toward and through the critical level is determined by the total mixing ratio of hydrogen atoms as they are assembled in various compounds in the stratosphere. According to Hunten, the flux should be relatively insensitive to details of photochemistry, the photolysis rate, the strength of mixing, and the value of the exospheric temperature for a wide range of conditions, some of which will become apparent in this section. Hunten's idea was verified and elucidated in detailed calculations performed in the case of the earth's atmosphere by Hunten and Strobel and by Liu and Donahue, and in the case of Mars and Venus by Liu and Donahue. A review of this work has recently been prepared by Hunten and Donahue (1976).

To understand what is involved, let us assume to begin with that hydrogen exists in one form only, atomic hydrogen, even in the stratosphere. In the lower atmosphere vertical flow of a minor constituent is controlled by eddy diffusion rather than by molecular diffusion, and the flux of a species according to the eddy-diffusion construct is determined by the gradient in the volume mixing ratio f_i of that species:

$$\phi_i = -Kn \frac{df_i}{dz}, \tag{4.5}$$

where K is the eddy-diffusion coefficient and n is the atmospheric density. In the stratosphere and mesosphere, Kn is so large that practical fluxes are maintained with only trifling departures in f_i from a thoroughly mixed condition. At higher altitudes near 100 km, where K and D_i (the molecular-diffusion coefficient for species i) become comparable, it might be thought that gradients in f_i would begin to occur, because it is the gradient in species concentration n_i, not f_i or n_i/n, that determines the molecular diffusive flow. However, D_i varies as $1/n$ so the ratio of n_i to n really is involved in the molecular diffusion contribution to the flow. Hence, even at the homopause, where K and D_i are equal, the flux can be determined essentially by

$$\phi_i = Kn_i/\mathrm{H} = D_i n_i/\mathrm{H}. \tag{4.6}$$

Hence n_i/H is approximately the hydrogen density gradient under circumstances where a significant departure of n_i from complete mixing cannot be tolerated, i.e., $dn_i/dz \simeq n_i/\mathrm{H}$. This flux [Eq. (4.6)] is called the limiting flux ϕ_i, and it can be written in the form

$$\phi_i = b_i f_i/\mathrm{G}. \tag{4.7}$$

where

$$D_i = b_i/n.$$

If we call n_i at this altitude n_{ih}, where h stands for homopause, and recall that limiting flux results when df_i/dz remains close to zero near the homopause, then a necessary and sufficient condition for the flux to be ϕ_i is that

$$n_{ih} >> n_{ic}. \tag{4.8}$$

Since continuity requires that

$$\phi_J = n_{ic} w_c \cong n_{ih} D_{ih}/\mathrm{H}, \tag{4.9}$$

this condition amounts to the requirement that

$$w_c >> w_{ih}, \tag{4.10}$$

where the diffusive flow velocity w_{ih} of constituent i at the homopause is given by

$$w_{ih} = D_{ih}/\mathrm{H}. \tag{4.11}$$

In other words, when escape at the critical level is easy compared to flow through the homopause, the flow is limited by a bottleneck at the homopause. This bottleneck limits the flux to ϕ_i, and $\alpha\phi_i$ depends on f_i at the homopause. However, f_i at the homopause will be about the same as f_i much lower in the atmosphere (above the next lower bottleneck if there is one), since, where eddy flow Kn (df_i/dz) dominates, Kn is so large that f_i remains nearly constant. For earth, w_{ih} is K_h/H. If K_h (the value of K at the homopause) is 5×10^5 cm^2 sec^{-1}, then w_{ih} is 1 cm sec^{-1}.

Now, this description of the flow has been written as though the minor constituent, say atomic hydrogen, maintains its identity uniquely throughout the region in which we are interested. In fact, on earth, much of the flowing H (in the stratosphere and mesosphere) is bound up in other forms: H_2, H_2O, H_2, OH, and HO_2. Thus in this region of the atmosphere there is a flow equation

$$\phi_j \cong -Kn \frac{df_j}{dz} \tag{4.12}$$

for each hydrogeneous constituent, and photochemistry keeps shuffling these constituents so that their relative populations vary. However, as R. Thomas has pointed out, there are no sources or sinks for total hydrogen, that is, for the sum of all hydrogeneous constituents weighted by the number of H atoms each contains. Therefore, the flux equation for H alone is equivalent to that for total hydrogen

$$\phi_t = \phi_t = -Kn \frac{df_t}{dz} \tag{4.13}$$

where f_t is the mixing ratio for total hydrogen. Difficulty with this concept will develop when there is an appreciable admixture of heavy hydrogeneous molecules near the homopause, since Eq. (4.13) assumes that the effect of D_i is unimportant. As long as H and H_2 are dominant near the homopause and the critical level, the arguments just presented are sound. Hence, the picture of hydrogen flow and escape that develops is the following. If the exospheric temperature is high enough so that $w_{ic} >> w_{ih}$, hydrogen will flow up from the earth's surface in forms such as CH_4, H_2O, and H_2, and into the stratosphere maintaining continuity, i.e., with ϕ_t constant. It is clear that there is a bottleneck for H_2O flow at the tropopause. Above that level in the stratosphere, the mixing ratio of total hydrogen is f_t; and this mixing ratio will also be equal to f_{ih}, so that the limiting flow through the homopause bottleneck is

$$\phi_t \cong \frac{b_i f_i}{\mathrm{H}}. \tag{4.14}$$

The hydrogen distribution in the thermosphere will be such as to maintain this flux, and the density at the critical level will adjust itself so that

$$\phi_t = \phi_i = n_c w_c. \tag{4.15}$$

However, the final part of this picture assumes that hydrogen escape is governed by the Jeans mechanism. We shall see that this in fact is not true. When the exospheric temperature is not sufficiently high and $w_c \leq w_h$, the bottleneck will shift to the critical level. Hydrogen will "pile up" in the thermosphere, and a sizable backflow

through the homopause region must develop. The atomic hydrogen will in part flow downward toward a sink in the mesosphere or stratosphere as well as upward toward the escape level.

LIMITING FLUX: MODEL CALCULATIONS

Quantitatively, these ideas have been tested by model calculations performed by Hunten and Strobel and by Liu and Donahue (Hunten and Donahue, 1976). In the stratosphere, hydrogen is mainly in the form of H_2O, CH_4, and H_2. The measured mixing ratios for these constituents are, respectively, 4 to 7.6 ppm, 0.25 ppm, and 0.5 ppm, causing f_t to lie somewhere between 9 and 17.2 ppm. The model calculations of Hunten and Strobel and those of Liu and Donahue agreed. The limiting flux turned out to be 1.8×10^7 cm^{-2} sec^{-1} for each 1 ppm of stratospheric "total hydrogen." Liu and Donahue varied many parameters in the model over wide ranges; T_c from 400 K to 1900 K; K_h from 0.5×10^6 to 3×10^6 cm^2 sec^{-1}; the solar flux by a factor of 6; and the rate constants for some important reactions by an order of magnitude. Several other characteristics of the model were also varied. All these changes caused the ratio of escape flux to f_t to change at most by 25 percent, from 1.7×10^7 to 2.2×10^7 cm^{-2} sec^{-1} for each 1 ppm of hydrogen as long as T_c was above 1000 K. The ratio of H_2 to H turned out to be large at 100 km, ranging from 0.5 to 2.5 depending on the chemistry involved. The H_2 is produced near the mesopause by a reaction between H and HO_2, and near 140 km it is converted efficiently to H by hot oxygen atoms in the tail of the Boltzmann distribution,

$$O + H_2 \rightarrow O + 2H. \qquad (4.16)$$

Below 1000 K, at a value of T_c depending on K, the escape flux dropped rapidly toward zero, as shown in Figure 4.11. Physically, what happened was that the relative flux of H_2 and H_2O through the homopause became more and more important as the escape of H was choked off. The latitude at which H_2O changed to H by photolysis and H_2 to H by reaction (4.16) increased, and so accordingly did the altitude at which the flux of H reversed. In the limiting case, the situation is similar to that of O_2 and O. The upward flux of H_2O and H_2 to the dissociation region is balanced by a downward flux of H to the recombination sink.

Diurnal variations in T_c have little effect in producing a deviation from the limiting flux as long as the maximum temperature is large enough. The reason for this fact is that the time constant to empty the reservoir between the diffusion and escape regions by escape at the higher temperature is very much smaller than the time constant for filling it as T drops below the critical value. The effective temperature for determining deviations from ϕ_l is close to the maximum diurnal value of T_c for times of the order of 100 days if T_c becomes large enough during the day.

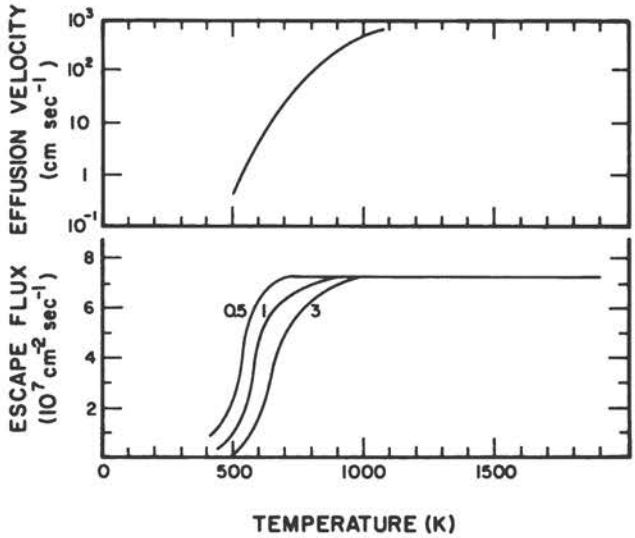


FIGURE 4.11 Escape flux of hydrogen and effusion velocity as a function of T_c for three values of the eddy-diffusion coefficient in units of 10^6 cm^2 sec^{-1}.

The chemistry and transport processes involved in the stratosphere and mesosphere turn out to be quite complex. The models suggest that the stratosphere is actually a source of water vapor produced by oxidation of methane and recombination of radicals such as OH and HO_2. Water vapor flows downward from the stratosphere to the troposphere. The water vapor that is eventually subject to photolysis near 100 km is not water vapor that came out of the troposphere at all. The small amount of water that passes the cold trap at the tropopause is turned into H and OH by metastable oxygen atoms in the stratosphere.

A problem immediately became apparent as a consequence of these calculations. If f_t lies between 9 and 17.2 ppm, the limiting flux should be between 16×10^7 and 31×10^7 cm^{-2} sec^{-1}, and this should be the escape flux for all values of T_c above 800 K. In fact, the consensus of the measurements of exospheric and thermospheric hydrogen discussed already is that when T_c is in the 1000–1200 K range, the calculated Jeans escape flux is in the neighborhood of 7×10^7 cm^{-2} sec^{-1}; Liu and Donahue concluded that 10×10^7 cm^{-2} sec^{-1} was an upper limit. They regarded the range of f_t quoted as adequately covering the range of uncertainty in the H_2O measurements, and thus they believed that the lowest credible theoretical value for the diffusive flow and hence the escape flux was 16×10^7 cm^{-2} sec^{-1}. Therefore they argued that there must be very important mechanisms for hydrogen escape other than thermal or Jeans escape.

ESCAPE MECHANISMS OTHER THAN JEANS ESCAPE

Liu and Donahue proceeded to evaluate other possible escape mechanisms. Two stood out. One was the polar

wind, discussed in Chapter 5. Here the protons that escape from the polar caps are supplied by thermospheric hydrogen that has undergone charge exchange with O^+. Banks and Holzer have calculated that the escape flux of protons over the polar caps varies from 5.6×10^8 to 5.8×10^8 to 3.8×10^8 cm^{-2} sec^{-1} as T_c varies from 750 to 1000 to 1500 K. However, they assumed that the neutral hydrogen density in the upper atmosphere was that appropriate to an escape flux $n_c w_c$ of 10^8 cm^{-2} sec^{-1}. Not all the hydrogen atoms that undergo charge exchange and escape in the polar wind are supplied by local vertical flow. Some travel into the polar cap on ballistic trajectories, but most arrive from a ring about 20° wide in latitude about the polar cap. So the entire planet certainly does not contribute equally to the polar wind flux. Nevertheless, the planetwide average flow to supply the outflow, assuming that the hydrogen distribution in the upper atmosphere at 1000 K is that appropriate to a Jeans flux of 5×10^7 cm^{-2} sec^{-1}, is 1.8×10^7 cm^{-2} sec^{-1}.

The other potentially important loss mechanism was suggested by Cole in 1966, and its magnitude was estimated crudely by Tinsley in 1973. This efflux is the result of charge exchange between the normal thermal atoms in the exospheric population and hot protons trapped in the plasmasphere. These protons, according to Serbu and Maier, have temperatures ranging from 5000 to 20,000 K and densities as high as 10^4 protons cm^{-3} at $1.5\,R_e$ on the dayside of the earth. The cross section for a charge-exchange collision, which leaves the neutral atom with energy large enough to escape, is 4×10^{-15} cm^2. Tinsley calculated the escape flux by first determining the rate at which fast atoms are produced, $knn^+(r)$, where k was computed by Dalgarno introducing a geometrical factor to accommodate only outward traveling atoms and n and n^+ are the concentrations of H and H^+. He then reduced the flux to the critical level using a factor $(r/r_c)^2$ and integrated from $1.5\,R_e$ to ∞ to obtain the escape flux for the mechanism,

$$\phi_E = \int_{1.5\,R_e}^{\infty} g\,(r/r_c)^2 knn^+ dr. \qquad (4.17)$$

Tinsley used a value for n_c of 3×10^4 atoms cm^{-3} and for T_c of 1250 K in a model in which the thermal flux was 10^8 cm^{-2} sec^{-1}. He found ϕ_E to be 4×10^8 cm^{-2} sec^{-1} in this case, a rate that clearly indicated that charge exchange between H and H^+ was potentially the most important of all mechanisms for escape of hydrogen from the earth.

Liu and Donahue proposed that the difference between the Jeans flux of about 5×10^7 cm^{-2} sec^{-1} at 1000 K and the limiting flux, which had to lie somewhere between 16×10^7 and 31×10^7 cm^{-2} sec^{-1}, be made up of contributions from the polar wind, ϕ_p, and charge exchange, ϕ_E. They pointed out that assuming the Jeans flux, ϕ_J, to be constant (independent of temperature) would force ϕ_E and ϕ_p to vary, since they scale with n_c. Thus, to keep the total flux constant, they concluded that ϕ_J would have to increase with increasing temperature. They computed the contributions of ϕ_J, ϕ_p, and ϕ_E as functions of temperature. The conclusion was that ϕ_J, ϕ_E, and ϕ_p all would vary with T_c. At 900 K in units of 10^7 cm^{-2} sec^{-1}, they would be 2.6, 22.2, and 1.4, respectively; at 1100 K, 7.1, 16.5, and 2.9; and at 1500 K, 13.5, 8.9, and 4.4. These fluxes do not disagree with those required by current models for the diurnal variation discussed in the previous section.

OBSERVED VARIATION IN ϕ_J

The same conclusions were reached independently by Bertaux as a result of his observations of hydrogen densities and temperatures obtained from the Lyman-α photometer/absorption-cell apparatus that he and Blamont placed aboard OGO-5. Results obtained with this apparatus have been described in connection with the exospheric densities. Because he observed the densities and temperatures of thermal atoms only, Bertaux was able to calculate the Jeans escape flux. He found that ϕ_J increased with T_c, as shown in Figure 4.12, where it is compared with the predicted values of Liu and Donahue. Realizing that this variation conflicted with the Hunten principle of limiting flux if ϕ_J were the sole escape term, Bertaux also proposed that ϕ_p and ϕ_E must make important contributions to the escape flux.

Previously Vidal-Madjar and his co-workers had noted a variability of ϕ_J during the lifetime of their Lyman-α experiment on OSO-5. This they attributed to a variation with solar activity in the density of hydrogen in the source region near 100 km, an explanation no longer viable. Their results are also plotted in Figure 4.12. Agreement between the model predictions and the two sets of observations is fairly good.

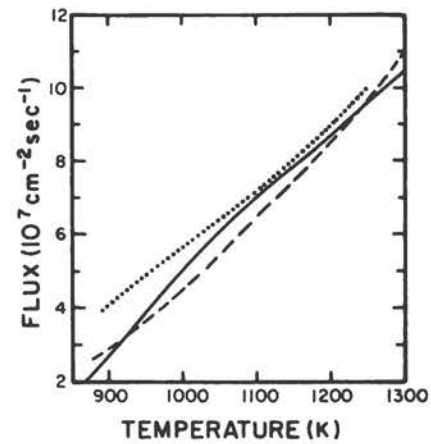


FIGURE 4.12 Jeans escape fluxes plotted as functions of exospheric temperature. The dotted line is evaluated from experiments conducted on OGO-5. The dashed line is deduced from measurements made on OGO-5. The solid line represents the calculated values.

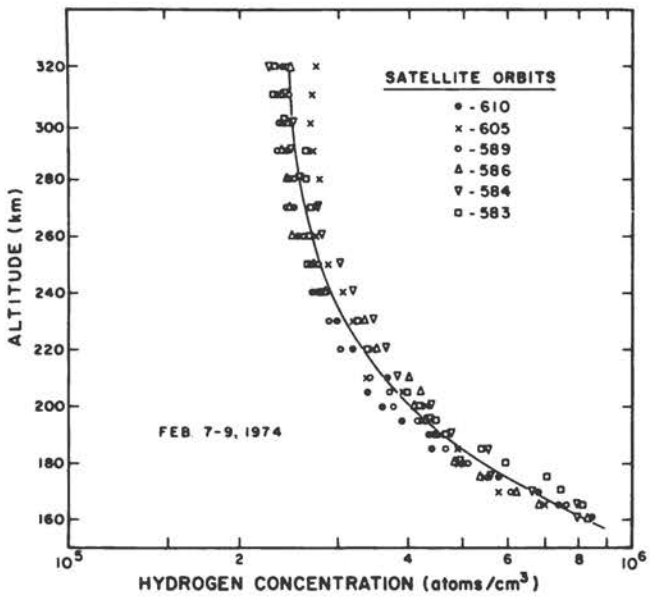


FIGURE 4.13 Hydrogen concentrations determined from Atmosphere Explorer C data for ascending portions of six orbits between low and middle winter latitudes in the early afternoon. The solid line has been selected as representative of the data.

It should be noted that in the real world the relative contribution of ϕ_J, ϕ_E, and ϕ_p should vary locally. In particular, ϕ_p will be large compared to ϕ_E on and near the polar cap but virtually vanish in the region of closed-field lines. As we know, however, the H^+ density in the plasmasphere increases steeply near the boundary between these two regions, and there ϕ_E will increase correspondingly.

Very recently, Breig *et al.* (1976) using data for H^+, O^+, and O densities obtained by the satellite Atmospheric Explorer C have been able to compute the hydrogen density between 160 and 300 km on the dayside of the earth in early February 1974. They used the appropriately corrected version of Eq. (4.4) in their calculations. The results, shown in Figure 4.13, allowed them to calculate the hydrogen flux from the diffusion equation. The value they obtained was $(32 \pm 10) \times 10^7$ cm^{-2} sec^{-1}. This is gratifying confirmation of the total flux $(16–31) \times 10^7$ cm^{-2} sec^{-1} predicted according to the limiting flux principle. It is far higher than the Jeans flux. An extrapolation of the data in Figure 4.13 to 500 km would suggest that n_c is in the neighborhood of 2×10^5 atoms cm^{-3} when T_c is 900 K. This would predict a Jeans flux of about 4×10^7 cm^{-2} sec^{-1}, in good agreement with previous observations and the Liu and Donahue prediction.

4.6 FUTURE WORK

The recent developments discussed here that indicate charge exchange of H and H^+ in the plasmasphere are a dominant source of hydrogen escape demand careful study. This study should include both observational and theoretical work. A good measurement of the proton density and energy distribution in space is required. A complete theory of the charge-exchange mechanism involving H^+ and O^+ is needed. It should be possible actually to observe the energetic atoms produced by charge exchange. Lower in the atmosphere, in the region between 140 km and the stratosphere, the observation of the distribution of H, H_2, OH, H_2O, CH_4, H_2, and if possible HO_2, is highly desirable. Of these it will be most interesting to obtain the distribution of H_2 between 90 and 140 km.

In this review, several topics have been deliberately slighted. Most of these relate to the Balmer-α radiation produced as a result of transport of Lyman-β radiation to the nightside of the earth (see Tinsley, 1974). There has been a persistent problem in understanding the absolute values of the emission rates observed in terms consistent with all the other information that we have concerning the geocorona. It would appear that the flux in the line center of the solar Lyman-β line, as measured, is too low by a factor of about 5 to account for the Balmer-α emission rates that are observed (Tinsley, 1974).

REFERENCES

Breig, E. L., W. B. Hanson, J. H. Hoffman, and D. C. Kayser (1976). *In situ* measurements of hydrogen concentration and flux between 160 and 300 km in the thermosphere, *J. Geophys. Res. 81*, 2677.

Brinton, H. C., H. G. Mayr, and W. E. Potter (1975). Winter bulge and diurnal variations in hydrogen inferred from AE-C composition measurements, *Geophys. Res. Lett. 2*, 389.

Carruthers, G. R., T. Page, and R. R. Meier (1976). Apollo 16 Lyman-alpha imagery of the hydrogen corona, *J. Geophys. Res. 81*, 1664.

Chamberlain, J. W. (1973). Planetary coronae and atmospheric evaporation, *Planet. Space Sci. 11*, 901.

Donahue, T. M. (1966). The problem of atomic hydrogen. *Ann. Geophys. 22*, 175.

Hunten, D. M., and T. M. Donahue (1976). Hydrogen lows from the terrestrial planets, *Annu. Rev. Earth Planet. Sci. 4*.

Tinsley, B. A. (1974). Hydrogen in the upper atmosphere, *Fund. Cosmic Phys. 1*, 201.

Tinsley, B. A., R. R. Hodges, Jr., and D. F. Strobel (1975). Diurnal variations of atomic hydrogen: observations and calculations, *J. Geophys. Res. 80*, 626.

5 The Ionosphere

WILLIAM B. HANSON and HERBERT C. CARLSON
Center for Space Sciences, The University of Texas at Dallas

5.1 PROLOGUE

The fact that earth is enveloped by a plasma of ionized gas was unknown at the turn of the century. Today, as a consequence of many years of remote probing by optical and radio devices, and more recently by *in situ* measurements utilizing rockets and satellites, we know a great deal about the morphology of this plasma. It is formed in the sunlit hemisphere by ionization of the earth's tenuous upper atmosphere by solar radiation and at high latitudes by energetic particles streaming in directly from the sun or accelerated in the magnetosphere.

Only in the last several years has the dynamic nature of the near-earth plasma, termed the ionosphere, begun to be appreciated. Both large and small-scale motions take place, but we are only beginning to accumulate detailed measurements of the plasma velocity field. We do know that very large plasma velocities (hypersonic—up to a few kilometers/second) can occur both perpendicular and parallel to the magnetic field and that these motions give rise sometimes to collisional heating and sometimes to expansion cooling of the plasma and that they also appreciably affect the plasma density distribution. Ions with long lifetimes (up to days) serve as useful tracers of some of these motions, while shorter-lived species provide clues to the complicated chemistry of the region.

The ionosphere consists of a series of layers or regions. The lowest portion, the D region, extends from about 60 to 90 km, and it is separately described in Chapter 6. The E region extends from about 90 to 140 km and is caused mainly by soft solar x rays. The F region lies above the E region, and during the day it is divided into two subregions, F1 (140–200 km) and F2 (above 200 km, with peak ionization near 300 km); during the nighttime the F1 region disappears. The F region is caused by extreme ultraviolet radiation from the sun (wavelengths from 20 to 90 nm).

Irregularities in the ion concentrations occur at all latitudes—nearly always at high latitudes, often at night at low latitudes, and rarely at midlatitudes. Scale sizes from hundreds of kilometers to tens of centimeters can be present, with amplitudes usually increasing with scale

size. Variations in concentration as large as a factor of 10^2 or 10^3 in a few kilometers are not uncommon near the equator, and they give rise to fading of communication-satellite signals even in the gigahertz frequency range. The ionosphere offers a unique workshop for the study of a multitude of natural and manmade interactions of a partially ionized medium with a broad spectrum of radiation and particles, and as a by-product of all these studies we are learning a great deal that will assist in investigations of laboratory plasmas, other planetary environments, and celestial bodies.

This chapter discusses briefly the historical development of knowledge of the ionosphere, together with a rudimentary explanation of the role of various parameters that affect its behavior. Somewhat more detailed descriptions of various observed ionospheric phenomena, and our present understanding of them, are then presented. The latter discussions are organized geographically (low, mid, and high latitudes), since, in general, different physical phenomena prevail in these different regions. The boundaries between the regions are somewhat indistinct but lie at approximately 20 and 60° magnetic latitude. A few references are noted at the end of the chapter relating to points of special interest in the development of the field or presenting current reviews on various aspects of the ionosphere.

5.2 DISCOVERY AND RECOGNITION

In 1901, shortly after the invention of radio, Marconi transmitted a coded signal from England to Newfoundland. The success of this experiment, the propagation of a wireless radio signal roughly 3000 km around the curved earth, could not be explained at that time. However, it triggered a revolution in long-distance communications and attracted much scientific attention Initial efforts to explain the effect in terms of atmospheric diffraction were shown by Lord Rayleigh to be invalid. In 1902, Kennelly and Heaviside independently hypothesized the correct explanation of a conducting layer in the upper atmosphere that reflected or refracted the radio waves back to earth.

Interestingly, 20 years earlier Balfour Stewart had suggested that an electrically conducting upper atmosphere, driven by diurnally varying solar heating to oscillate across the earth's magnetic-field lines, would have electrical currents induced in it, like the rotating armature of a huge dynamo, leading to the weak diurnal magnetic-field modulations that had been observed at the ground since the eighteenth century. By the end of the nineteenth century, it emerged that to be a conductor a gas needed to be ionized, not merely highly rarefied as Stewart first thought on the basis of early laboratory experiments. It then seemed likely that the sun would produce the needed ionization by means of x rays, ultraviolet radiation, or streams of particles.

5.3 MEASUREMENT TECHNIQUES

PREROCKET PROBING OF THE IONOSPHERE

In 1925–1926, reflection of vertically incident radio waves was demonstrated from the nighttime ionosphere. These critical experiments opened the way to the use of radio waves as a tool to study the nature and morphology of upper atmospheric ionization and stirred early thoughts about which atmospheric constituents were ionized and what the (presumably solar) radiation responsible for the ionization might be.

The radio device that emerged for probing the ionosphere is the ionosonde; it transmits a wave pulse and translates the delay time of the vertically incident reflected echo into the apparent height of the reflecting layer. Development of a "magneto-ionic theory" to explain radio-propagation effects, and measurements of the polarization of the reflected echoes, showed that free electrons rather than positive ions (or negative ions formed by electron attachment) led to the observed echoes. When the range of sounding frequencies was extended, qualitatively new layers of ionization were found. Detailed analysis of refraction effects led to quantitative estimates of electron concentration profiles that increase from approximately 10^{10} to 10^{12} electrons m^{-3} in the altitude range from 100 to 300 km. Signals that penetrate the layer of maximum concentration are not reflected at higher levels, so the ionosonde provides no information on the topside of the ionosphere.

By 1931, Sidney Chapman developed a chemical equilibrium theory whereby solar radiation ionized atmospheric gases that were exponentially distributed in altitude, and the resulting electrons were lost by recombination with positive ions. Ionosonde data showed that this theory explained the basic behavior of two ionospheric layers, the E layer with its peak in electron concentration near 120 km and the F1 layer with its peak near 160 km, but not the highest layer with the largest electron concentration, the F2 layer. It was a quarter century later before it was generally recognized that plasma diffusion could account for the altitude of the F2 layer, even though its importance was early suggested by Hulburt (1928).

When a long wire is attached to an audio amplifier, a class of electromagnetic radiations known as "whistlers" can be heard, characterized by a pitch descending through the audible range over a time scale of a few seconds. Although these signals were known in the late nineteenth century, it was 1953 before Owen Storey showed that they were caused by electrical discharges in lightning strokes from which the radiated electromagnetic energy propagated along a line of the earth's magnetic field to the opposite hemisphere. The time dependence of the pitch can be related to the electron concentration along the magnetic-field line, particularly when the signal persists for more than one "hop" between hemispheres. Surprisingly, large electron concentrations ($\sim 10^9$

electrons m^{-3}) near the earth's magnetic equatorial plane were measured in this way out to several earth radii, and the existence of a strikingly sharp drop in electron concentrations near a distance of 4 R_e (the plasmapause) was later discovered by this technique.

Radio star scintillations seen on two separate radio telescopes were found to correlate well for a sufficiently closely spaced pair. By about 1950, such observations by receivers at different spacings, interpreted using diffracting screen theory, had shown that electron concentration irregularities of a few kilometers scale size and several percent amplitude were a common feature of the ionospheric F region.

The early history of ionospheric observation has been reviewed by Waynick (1975).

ROCKET AND SATELLITE MEASUREMENTS

The availability of rockets after 1946 led to major advances in understanding ionospheric photochemistry and thermal balance. Ion mass spectrometric data showed molecular ions to dominate the E and F1 regions, giving way to atomic oxygen ions in the F2 layer, as explained by Havens *et al.* (1955). Such measurements also showed metallic ions (of meteoric origin) to be important ionospheric constituents, particularly of the nighttime E region. The role of light ions in the topside ionosphere was recognized. It was also shown that electron, ion, and neutral particle temperatures can be significantly different from one another, with the flow of energy predominantly from photoelectrons to thermal electrons and then to ions and neutral particles (Hanson and Johnson, 1961). Measurement of the intensity of ionizing radiation versus altitude confirmed the theory that the F2-layer peak height was determined by a balance between downward diffusion and electron loss rather than by where the electron production rate peaked.

Satelliteborne instrumentation made possible a level of geophysical understanding that could only follow from a large data base that contained information from previously inaccessible latitude and longitude sectors. The division between a high-latitude "polar ionosphere," dominated by energetic particle and magnetospheric processes, and a lower-latitude ionosphere, dominated by solar radiation, also emerged. Seasonal effects were more clearly evident, and a general magnetic control of ionospheric behavior, even at low latitudes, was firmly established. Longitudinal anomalies, many still unexplained, have also been observed.

These satellite efforts have culminated recently in an effective complement of sensors that can measure solar radiation and neutral and charged-particle temperatures, concentrations, compositions, and drifts; these have been mounted on the uniquely versatile Atmosphere Explorer (AE) series of satellites. These spacecraft have an onboard propulsion system that provides unprecedented low-altitude coverage in an elliptic-orbit phase, as well as a later circular-orbit phase; during both phases the satellites can be made to spin or be three-axis stabilized. Data from the first of these satellites (AE-C) has shown the importance of various excited neutral species to the ionospheric photochemistry, and it has illuminated the importance of global transport and dynamics of both neutral and ionized species to ionospheric behavior. Many qualitative advances in high-latitude ionospheric physics have been made.

INCOHERENT-SCATTER RADAR MEASUREMENTS

Radio waves at frequencies much greater than those that are used by ionosondes (whose transmitted radio waves are reflected by the ionosphere) still lead to very weak reradiation by free electrons in the ionosphere. With the development after World War II of high-power radars, it became technically feasible to construct instruments capable of measuring the very weak backscattered echoes from ionospheric electrons (Evans, 1975). The signal strength, spectral shape, and Doppler shifts of such ionospheric echoes have permitted measurement of plasma bulk velocity and electron and ion temperatures and concentrations in the altitude range from 100 km to greater than 1 R_e. The technique also provides neutral-particle concentrations and photoelectron flux information over useful altitude regions. Particularly exciting have been recent plasma velocity measurements that allow the collection of continuous information on electric-field vectors and meridional neutral winds. The combination of these time-continuous profiles with satellite (particularly the AE series) global coverage is now beginning to be exploited.

5.4 IONOSPHERIC THEORY

FORMATION OF THE IONOSPHERE

Equatorward of the auroral zones, the only important source of energy available to ionize the atmospheric gases is solar ultraviolet and x radiation. The primary ions resulting from this absorption are O^+, N^+, He^+, N_2^+, and O_2^+. The primary molecular ions are relatively short-lived because they can recombine rapidly with electrons, forming two neutral atoms, e.g.,

$$O_2^+ + e \rightarrow O^* + O^* + \Delta E, \qquad (5.1)$$

where the asterisks indicate that the atoms may be internally excited, and the ΔE indicates that the atoms may be released with translational energy. Recombination of atomic ions with electrons is a much slower process (by a factor of the order of 10^4), and the likely fate of these ions is that they will react chemically with a neutral molecule to form a molecular ion, e.g.,

$$O^+ + N_2 \rightarrow NO^+ + N. \qquad (5.2)$$

The fact that NO^+, which has the lowest ionization potential of all atmospheric ions, is the dominant daytime molecular ion was unanticipated before the first rocket mass spectrometer measurements. Since the concentration of molecules available for the above reactions diminishes rapidly with increasing altitude, the lifetime of the atomic ions becomes longer at greater heights. This causes an increase in the ratio of atomic to molecular ions with altitude, such that the region near and below approximately 180 km (the F1 region) is dominated by molecular ions, whereas the overlying F2 region consists mostly of atomic ions, principally O^+.

The main source of O^+ (photoionization of atomic oxygen) decreases less rapidly with increasing altitude than does its loss rate with molecules, because of the smaller mass and consequent greater scale height of atomic oxygen compared with the molecules. Thus the absolute concentration of O^+ increases with altitude. This increase continues until the tenuous nature of the atmosphere allows rapid vertical diffusion to take place, so that ions formed at greater heights are pulled down by gravity to where they can charge exchange with the molecules. This downward diffusive transport of ions stops the increase of ion concentration with altitude and causes a maximum to form at approximately 300 km. Not far above this peak, the ion concentration decreases with altitude z approximately as $\exp -[\overline{m}_i g z/k(T_e + T_i)] = \exp -(z/H_i)$, where T_e and T_i are the electron and ion gas temperatures, $\overline{m}_i$ is the mean ion mass, k is the Boltzmann constant, and g is the local value of the gravitational acceleration. The denominator in the second form of the exponential expression is H_i, the ion scale height. It differs from the scale height as normally defined by the inclusion of T_e in addition to T_i. Charge-separation electric fields force very near equality in the ion and electron concentrations; and since the gravitational forces on the electrons are negligible, the gradient in the electron pressure (kT_e) is effectively added to that of the ion pressure in balancing the gravitational force on the ions. Actually, because of a slight separation between the ion and electron gases, a vertical electrical field $E = \overline{m}_i g T_e/(T_e + T_i)\,\epsilon$, exists, where ϵ is the electron charge. Thus the net vertical force $F = \epsilon E - m_i g$ on an ion depends on its mass; and for ions whose mass is less than $\overline{m}_i/2$ this net force is upwards! Light ions float to the top of the ionosphere like cream on milk!

A representative midlatitude ion distribution with altitude is shown in Figure 5.1 for daytime. There is a general decrease of average ion mass with increasing height, and the scale height of the H^+ ions in the upper portion of the F region is so large that H^+ ions extend far into the magnetosphere, as first indicated by the early whistler measurements.

There are often metallic ion layers (not shown in Figure 5.1) whose concentrations and altitude distributions are greatly influenced by upper-atmospheric motions. Because of the low ionization potentials of metals, their ions cannot react by simple charge-exchange processes with the atmospheric gases. Since they also recombine very slowly with electrons, they have very long lifetimes (several days) above 100 km, and they can be transported great distances. Their concentrations are relatively small in the daytime, except for occasional intense "sporadic-E" layers just above 100 km, but they may play important roles at night after partial recombination of the atmospheric ions has taken place. It is presumed that meteors are the source of the metallic ions, which can be formed upon impact with the atmosphere, or by charge exchange of neutral metallic atoms with ambient ions or by photoionization.

Ionospheric composition and reactions have been discussed by Donahue (1968).

IONIZATION TRANSPORT

The presence of the earth's magnetic field greatly influences the motions of plasma in the ionosphere, so that it is perhaps not too surprising that many ionospheric features are better described in magnetic rather than geographic coordinates. The ions spiral around the magnetic field at a characteristic frequency, ω_i (the angular gyrofrequency), such that

$$\omega_i = \epsilon B/m_i \quad \text{(rad/sec)}, \tag{5.3}$$

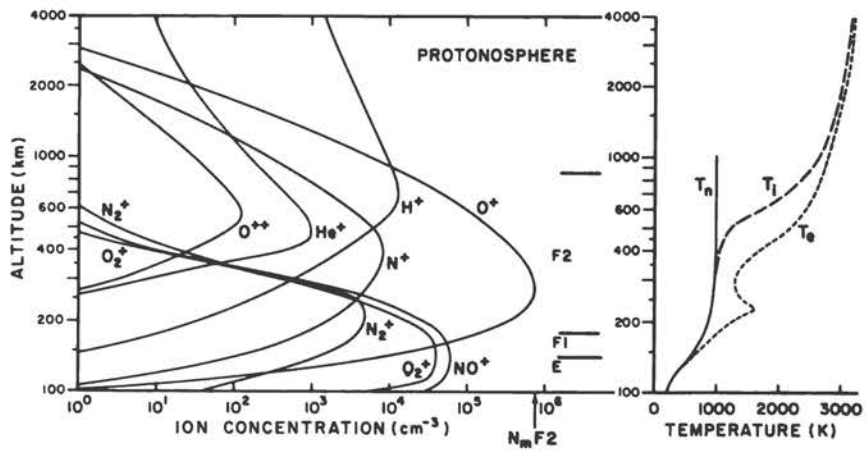


FIGURE 5.1 Profiles of gas temperatures and ion concentrations representative of ionospheric daytime conditions. The dip near 300 km in electron temperature T_e toward the ion and neutral temperatures T_i and T_n is less pronounced or absent for lower values of maximum total ion concentration at the F2 peak. The distributions and concentrations of metallic ions (not shown) are quite variable.

where B is the magnetic-field strength. Near 120 km, the collision frequency of ions with neutral particles, ν_{in}, is equal to the angular gyrofrequency; hence below this altitude the ions cannot move in uninterrupted spirals about magnetic-field lines but instead move relatively freely across the magnetic field, and, in the absence of electric fields, they tend to follow the neutral gas motions. Much above 120 km, the ions are constrained to move along magnetic-field lines unless an electrical field **E** is present, in which case the ions take up a velocity component perpendicular to both the electric and the magnetic fields, according to the relation

$$\mathbf{v}_\perp = \frac{\mathbf{E} \times \mathbf{B}}{B^2}. \tag{5.4}$$

For electrons, the electron collision frequency with neutrals equals the gyrofrequency near 90 km, so that neutral winds have little direct influence on the electron motions above 90 km, the motions being governed almost entirely by the electric and magnetic fields according to the relationship defined above for the ions.

Neutral winds blowing perpendicular to the magnetic field (e.g., zonal winds) are quite ineffective in moving ions directly at high altitudes, where ν_{in}/ω_i is small. Collisions between neutral particles and the plasma do tend to polarize the plasma, leading to a polarization electric field

$$\mathbf{E}_p = -\mathbf{W} \times \mathbf{B}, \tag{5.5}$$

where **W** is the neutral wind speed. If no currents flow to short out this polarization electric field, it will attain the value necessary to make the plasma drift at a velocity equal to the component of the neutral wind velocity perpendicular to **B**. The wind component along magnetic-field lines tends to move the plasma in the direction of the wind, and because of the slope of magnetic-field lines this interaction moves ions upward if the wind is equatorward and downward if the wind is poleward, even though the wind itself is horizontal.

THERMAL STRUCTURE

Photoionization of the upper atmosphere by solar ultraviolet absorption gives rise to photoelectrons that have an average energy on release of approximately 18 eV in the F region. This photoelectron energy is dissipated by interactions with the neutral, ion, and electron gases. Inelastic collisions with neutral particles are efficient in removing energy from the more energetic photoelectrons, and some additional ionization is produced in this manner, although most of the energy absorbed in such collisions is radiated away or utilized in subsequent chemical reactions.

Elastic collisions of photoelectrons with ambient electrons and ions heat these gases above the neutral gas temperature. Because the momentum transfer from photoelectrons is faster to electrons than to ions, and because momentum transfer is faster to neutral particles from ions than from electrons, the three gas temperatures usually satisfy the relationships $T_e > T_i > T_n$. Below 120 km the collision frequencies are so large that the three gas temperatures are nearly equal, but above this altitude in sunlight T_e increases rapidly up to approximately 200 km, where it is usually about twice the neutral temperature. The detailed behavior of T_e above 200 km depends strongly on the electron concentration height profile, and in general T_e varies inversely with the electron concentration. The ion temperature stays close to the neutral gas temperature up to approximately 350 km, but above this altitude T_i begins to increase until ultimately $T_i \approx T_e$; above 1000 km both can be several thousand degrees Kelvin above the neutral gas temperature.

At night, photoionization ceases and the three gas temperatures all tend toward the same value. This equilibration proceeds rapidly at low altitudes and latitudes, but considerable heat can be stored in the plasma contained on the high-latitude magnetic-field lines that extend far out into space (several earth radii). It takes some time to conduct all this heat away, and lack of thermal equilibrium is observed at high altitudes and latitudes long after sunset. In addition, even though the atmosphere may be locally in darkness, the atmosphere at the opposite, or conjugate, end of the local magnetic-field lines may not be. When this is the case, direct impingement of photoelectrons released in the opposite hemisphere can heat the local plasma; this direct heating is augmented by thermal conduction along the magnetic tube from the hot sunlit conjugate ionosphere.

The above simple picture is drastically modified at high latitudes, where large energy inputs are made by energetic particles and electric fields even when the atmosphere is in darkness at both ends of the field lines.

AIRGLOW

As a result of the absorption of solar radiation in the upper atmosphere there is considerable internal excitation of the ambient gases. This internal energy can be quenched by collisions or it can be radiated. When the latter happens, the resulting radiation is called airglow, and the emissions are characteristic of the gas composition.

Some of the absorbed energy is stored chemically with rather long time constants, e.g., the ionization energy of atomic ions at high altitudes and the dissociation energy of molecules of oxygen and nitrogen. Chemical recombination of these fragments occurs throughout the night, and this may also lead to airglow that persists through the night, although it is usually much less intense than the daytime airglow. Figur 5.2, an ultraviolet picture of the earth taken from the moon, illustrates some important types of airglow.

FIGURE 5.2 The earth as seen from the moon in far-ultraviolet (125–160 nm) light. The light includes radiation from atomic oxygen and molecular nitrogen. One sees bright solar EUV-excited dayglow to the left, particle-excited polar cap emission to the lower right, and bands extending into the nighttime hemisphere of airglow excited by $O^+ + e$ recombination. The recombination airglow maps out the very high electron concentrations in the Appleton anomaly regions bracketing the magnetic equator. Background stars are also seen. (Naval Research Laboratory photograph.)

5.5 LOW-LATITUDE PHENOMENA

IONOSPHERIC CURRENT SYSTEM AND ELECTROJET

The magnetic-field components at the earth's surface have been known for over 250 years to undergo small (~0.1 percent) diurnal variations about their mean values. The diurnal variation of solar heating drives the conducting upper atmosphere across the earth's magnetic-field lines, and these motions induce currents that generate the solar daily magnetic variations, called Sq variations for magnetically quiet days. Under these conditions, charged-particle mobilities perpendicular to the magnetic field maximize when the collision and angular gyrofrequencies are equal, and this condition is satisfied near 90 km for electrons and some 30 km higher for ions. Since conductivity is proportional to both mobility and charged-particle concentration, the main conducting layer is near 120 km, where the ion and electron concentrations are much larger than at 90 km, and ions are the principal current carriers.

At the equator, the magnetic field is horizontal, and this special geometry leads in the E region to a very high east–west "Cowling" conductivity (perpendicular to the magnetic field) that is comparable with the conductivity along the field. As would then be expected, a correspondingly intense current sheet, known as the equatorial electrojet, is found to flow along the magnetic equator near 100 km altitude in a strip only a few degrees wide in latitude. The current flows toward the east by day and the west by night, but the westward currents are nearly undetectable in magnetometer measurements because of the small electron concentrations at night. The large particle velocities in the electrojet drive plasma instabilities that are of special interest in their own right.

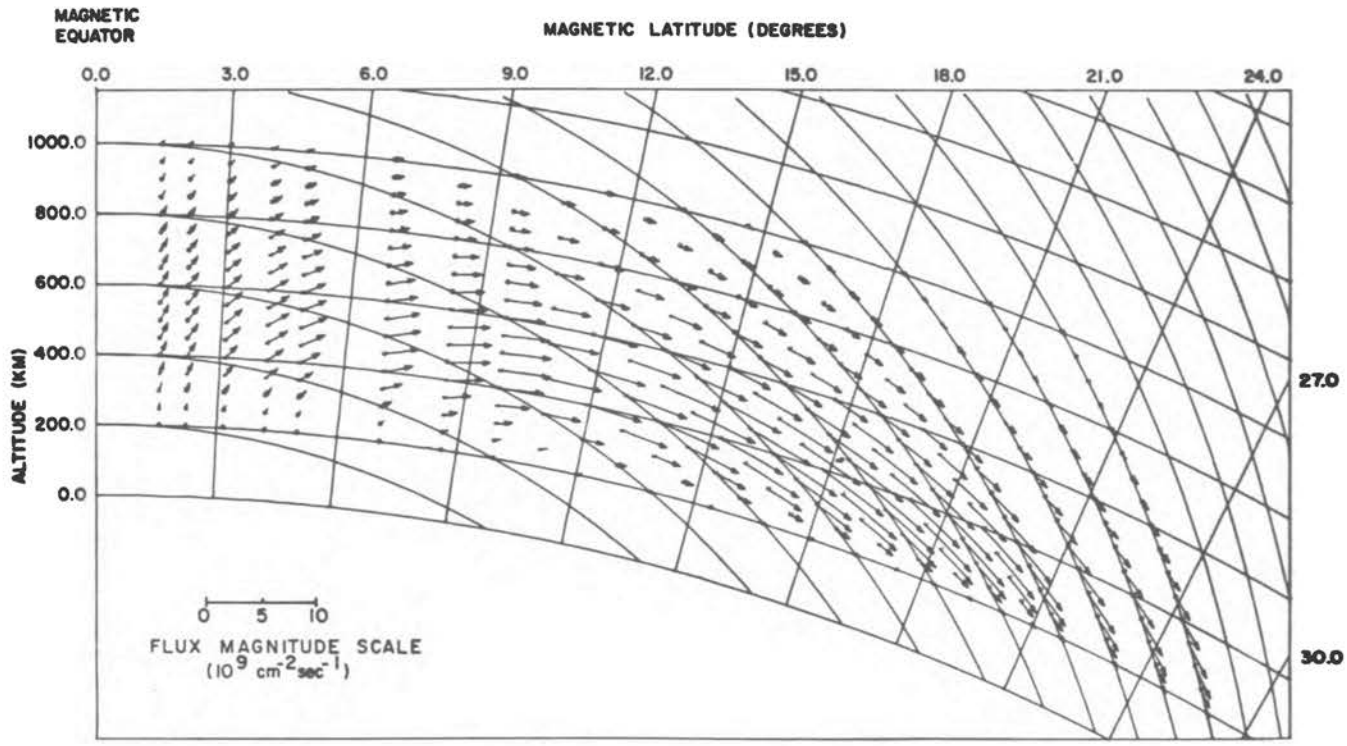
EQUATORIAL ELECTROJET PLASMA INSTABILITIES

The equatorial electrojet current drives a plasma turbulence that has been studied with the complementary techniques of very-high-frequency backscatter radars and rocketborne probes. The rockets travel much faster than the plasma-wave phase velocities and provide spatial spectral information. Very-high-frequency radars are sensitive only to plasma waves with a propagation vector component along the line of sight and of a scale size matched to the probing wavelength; they see two characteristically different type echoes. Type 1 echoes, with a relatively constant Doppler shift corresponding to the acoustic velocity of the medium, are believed to arise from plasma waves that are directly excited at the observed wavelength by a modified two-stream instability. Type 2 echoes, generally weaker and with smaller and quite variable Doppler shifts, are returns from plasma irregularities that result from a gradient-drift plasma instability.

While the linear plasma theory of electrojet instabilities is well developed, major progress in the nonlinear theory is required before the high potential of this diagnostic tool for plasma turbulence and instabilities in the ionospheric "laboratory plasma" can be realized. Extension of these results to auroral conditions, where different altitudes are strongly coupled because of the high conductivity along the magnetic-field direction, poses further challenges.

APPLETON ANOMALY

The electric fields associated with ionospheric currents generally drive a plasma convection in the F region at low magnetic latitudes that is upward and westward in the daytime and downward and eastward at night. The upward motion in the daytime raises freshly ionized plasma near the equator to great heights, where recombination is slow. Subsequent diffusion or flow down the magnetic-field lines under action of gravity adds this extra plasma to that produced locally at higher latitudes. The result of this plasma transport, illustrated in Figure 5.3, is that ioniza-



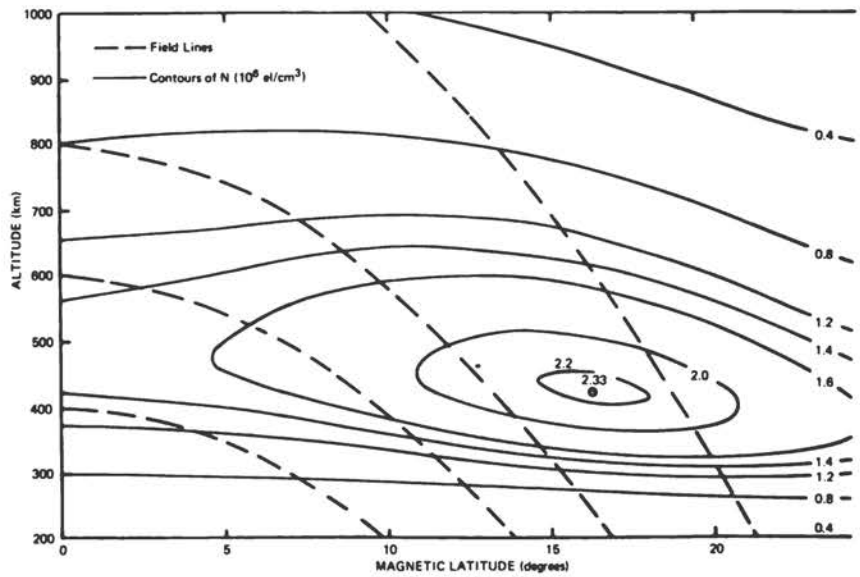


FIGURE 5.3 Upper portion: the pattern of plasma drift due to electric fields and gravitationally induced drift along magnetic-field lines, leading to the production of the Appleton anomaly. Lower portion: contours of electron concentration that result from the combined drifts due to electric field and gravity.

tion peaks are formed in the subtropics on each side of the magnetic equator. These regions of anomalously high electron concentration (often called Appleton's peaks) were originally attributed to the above described "fountain effect" by D. F. Martyn.

The latitude of the peak formation and the ratio of the peak value of electron concentration to the maximum electron concentration N_mF2 over the equator both increase as the vertical plasma drift velocity increases. The two peaks are often not symmetrical, and this fact is due principally to plasma transport along the magnetic-field lines induced by a neutral wind component along the magnetic meridian. The neutral winds usually cause plasma to be pushed from summer to winter near midday, so that the winter peaks are larger. There are longitudinal anomalies in this seasonal behavior related to large dif-

ferences in magnetic declination that affect the magnitude and even the sign of the wind component along the field lines.

METEORIC IONS AT LOW LATITUDES

Meteorites provide a continuous supply of metallic and semimetallic atoms to the upper atmosphere near and below 100 km. The low ionization potential of these atoms leads to their ionization either on impact, by photoionization, or by charge exchange with atmospheric ions. It had been demonstrated with rockets that such ions (Mg^+, Fe^+, Si^+, Na^+, Al^+) are rather permanent residents of the altitude region near 100 km. Recent satellite data have shown them to be commonly present also at much higher altitude, up to F2 peak heights, particularly at night. These ions, which are essentially chemically inert above 100 km, partake in the above-described drift motions and serve as excellent tracers of these movements. Near the magnetic equator, where vertical diffusion is suppressed by the horizontal magnetic field, they penetrate to altitudes above 1000 km. In general, they appear sporadically in time and place, for reasons that we do not yet fully appreciate.

At moderate to high latitudes solar heating tends to drive poleward winds in the daytime, and these winds tend to push ions to lower altitudes. At night, this tendency is reversed by generally equatorward winds, and meteoric ions are pushed up into the F region.

EQUATORIAL SPREAD F

Pulsed-radar probing of the ionosphere often produces multiple, or spread, return echoes. The resulting ionosonde records are distinctly different from normal conditions, and they have been used historically to identify and characterize the F-region phenomenon called "spread F." Equatorial spread F is almost exclusively a nighttime phenomenon that is more prevalent near the equinoxes. Communication satellites have revealed that equatorial spread F can seriously distort radio waves even in the gigahertz frequency range. *In situ* satellite measurements of the equatorial ion concentration have now been carried out with rather good time resolution, and the severity of the radio-wave transmission difficulties is no longer surprising (Hanson, 1975). The total electron concentration, and hence the index of refraction, is very highly irregular with changes in concentration as large as a factor of 10^2 or 10^3 in only a few kilometers, as can be seen in Figure 5.4. Scale sizes as small as 60 m have been observed from Atmosphere Explorer satellites, but the radar at Jicamarca, Peru, extends this limit down to 3 m. Not only are the radar echoes at 3 m enhanced by up to 60 or 70 dB during spread F, but these abnormal signals can appear in a time of less than 8 msec over regions tens of kilometers across. Sometimes the abnormal echoes persist for hours, and during this time the altitude of a particular irregularity patch will often increase by hundreds of kilometers, as shown in Figure 5.5. Detailed examination of the echoes also reveals a frequency spreading that corresponds to a turbulent velocity of the order of several hundred meters per second. A similar behavior is shown by ion concentration and mean ion velocity data recorded by the Atmosphere Explorer near the equator at night. It is found that the depleted ion regions tend to move upward, as bubbles rise in a denser medium, and indeed it has been suggested that spread-F structure is caused by such buoyancy forces.

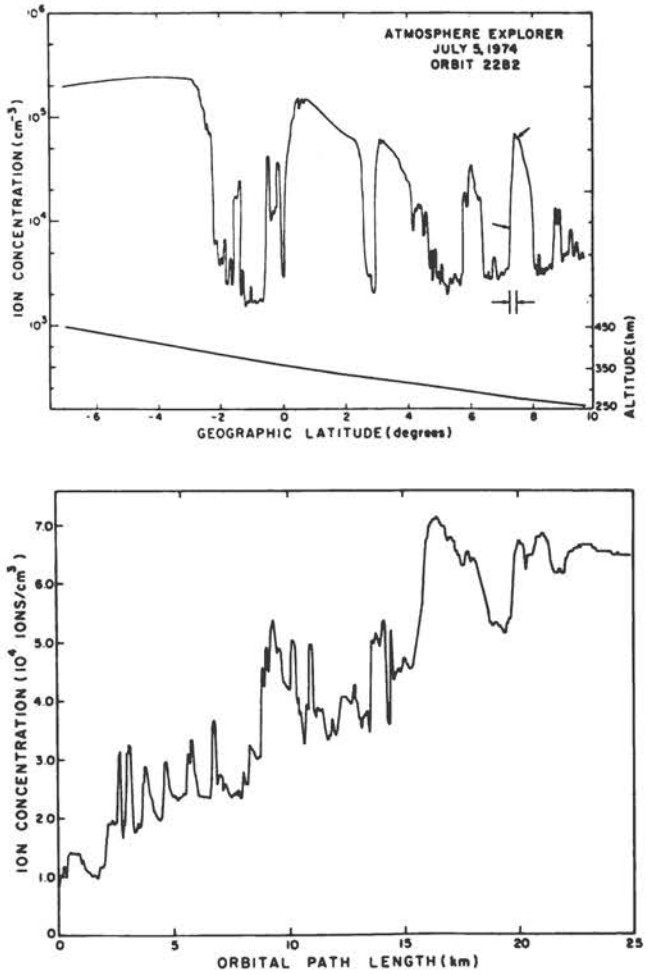


FIGURE 5.4 The equatorial ionosphere during spread-F conditions as seen by Atmosphere Explorer. Large-scale depressions in ion concentration by factors up to 100 are illustrated in the upper figure. The lower trace shows an expanded view of the small segment of the upper trace between the two arrows. Large fractional changes in ion concentration are seen to occur down to scale sizes of hundreds of meters.

Before *in situ* velocity measurements were available, it was noted that a strong correlation existed between irregularities in ion concentration and the presence of metallic ions. It was suggested that there might be a causal relationship between the two, possibly associated

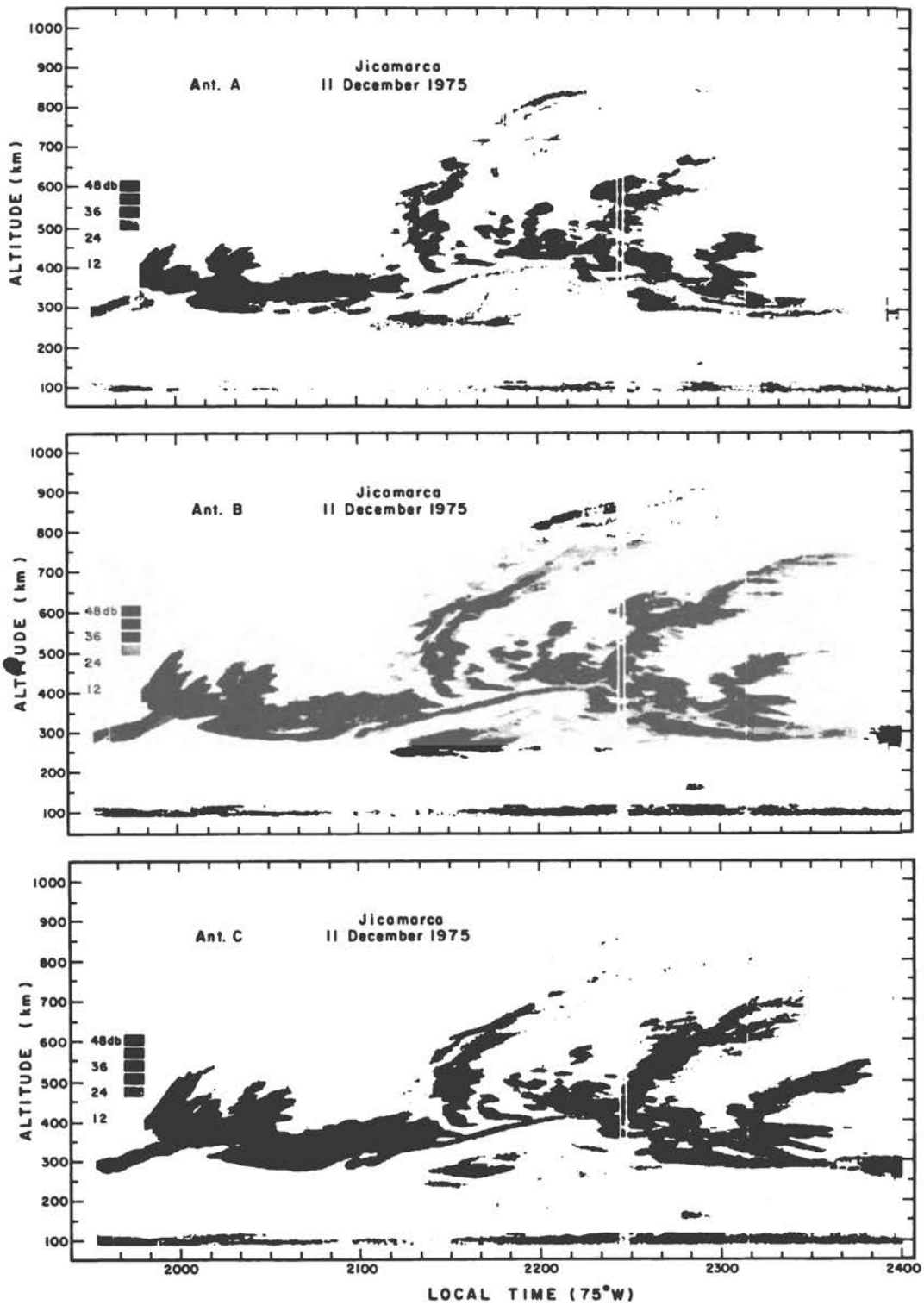


FIGURE 5.5 Equatorial F-region radar echoes seen at Jicamarca from 3-m scale-size irregularities generated by plasma instabilities. These echoes can appear in a time of less than 8 msec over regions of tens of kilometers in scale size. Each time–intensity–altitude plot shows the persistence of disturbed regions and traces out their vertical motions in time. Adjacent plots for slightly different lines of sight by the radar provide information on horizontal structure.

with conductivity irregularities induced by the metallic ions in the F1 region. It now appears that an alternative explanation for the correlation may be that the metallic ions are transported to great heights by the large buoyancy-induced vertical drift, and their presence is merely an indicator of such activity rather than the cause. From time to time there have been claims of observations of large fluxes of energetic particles at the times and places where equatorial spread F occurs, but there are serious questions about the validity of these claims and whether such particles could be related to spread F.

ION SUPERCOOLING

It was observed on OGO-6 that the ion temperature at night near the magnetic equator often shows drastic changes with latitude in the altitude range above 600 km, as illustrated in Figure 5.6. A minimum in ion temperature usually appears on the summer side of the magnetic equator, and a maximum appears on the winter side. There are wide variations in the magnitudes of these maxima and minima, and the minimum value of the ion temperature is often less than half of the temperature of the neutral gas within which the ions are embedded. The ratio of T_i(max) to T_i(min) can exceed a factor of 5. This unexpected behavior can be explained in terms of a large interhemisphere flow of plasma induced by summer-to-winter neutral winds in the F region. In the summer hemisphere these winds lift the plasma along the magnetic-field lines to altitudes where O^+ ions are destroyed much less rapidly by charge exchange with neutral molecules. Conversely, in the winter hemisphere these same winds move the F-region ionization to lower altitudes and increase its loss rate. The resultant imbalance in the plasma pressure at the ends of the field lines causes plasma to flow from the summer toward the winter, and in doing so it must rise against gravity to go over the apex of the magnetic-field line at the magnetic equator. As it rises in the gravitational field, the ion concentration decreases, and this expansion of the gas causes it to cool. Since the plasma flux along the field tubes is conserved, its flow velocity must vary inversely as the ion concentration. The expansion cooling is not entirely adiabatic because of the thermal coupling between the ions and the neutral gas and the thermal conductivity of the electrons, which are in good thermal contact with the atmosphere near and below the F2 peak; the latter effect is especially important at high altitudes. Thus the plasma gains energy from its surroundings when it is cooled below the neutral gas temperature, and when it descends in the winter hemisphere it is superheated and overshoots the neutral gas temperature.

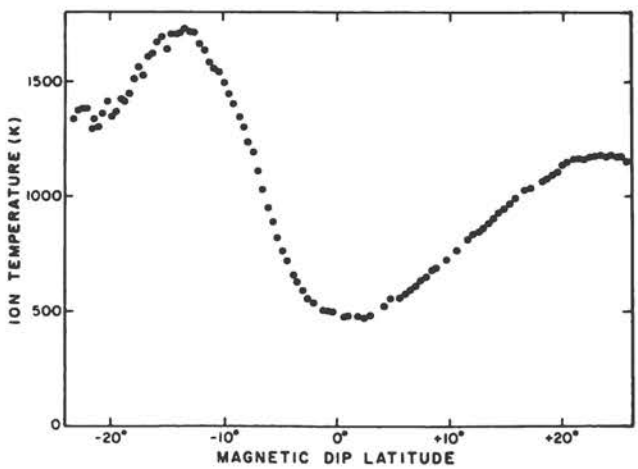


FIGURE 5.6 The ion temperature at a nearly constant 1000-km altitude, seen here to vary across the magnetic equator by over a factor of 3 despite the background neutral gas temperature remaining at about 1200 K. It is hypothesized that this is due to interhemispheric plasma transport upward along magnetic-field lines in the north, involving nonadiabatic expansion cooling, conductive heating while at high altitudes, and compressional heating while descending along the magnetic-field lines in the south. Interhemispheric gradients in plasma concentrations, six times higher to the north than the south for this example, are presumed to drive the flow.

Quantitative explanations of the observed ion temperature behavior require flow velocities of the order of the mean ion thermal speed, and so far there is only crude evidence to support these large velocities. The high-latitude boundary for these interhemisphere flows is not well established, but one might be expected related to the apex height of the field lines at the equator. When this height becomes too great, the atomic oxygen ion concentration falls to where it is too small to support the required flux at reasonable velocities (i.e., not supersonic). To some extent this difficulty can be overcome by charge exchanging O^+ to H^+ and allowing H^+ to carry the flow at high altitudes, but there are complications with this possibility also. New measurements from the Atmosphere Explorer should provide adequate detail in all the pertinent parameters to assure a quantitative understanding of the phenomenon.

EQUATORIAL He^+ BITEOUT

The concentration of He^+ ions has been observed to be quite variable near the magnetic equator, and it often shows very large relative depressions at low latitudes. This behavior is probably related to the activity of the ion transport mechanisms discussed above, both the interhemisphere transport and the drift motions due to electric fields. The chemical time constant for the establishment of an equilibrium He^+ distribution is quite long at high altitudes, where diffusive equilibrium must be established in magnetic tubes in which the photoionization source is small. The actual distribution of He^+ depends on the outcome of the competition between transport and photochemistry. At high altitudes where their lifetimes are long, He^+ ions are better tracers of motions than even the metallic ions because their source is quan-

titatively known, but to date little theoretical effort has been expended to capitalize on this fact.

5.6 MIDLATITUDE PHENOMENA

The midlatitude region as defined here includes that part of the plasmasphere beyond the tropical region just discussed. The plasmasphere terminates at the plasmapause, which is often near 60° magnetic latitude. It is free of the direct influence of phenomena associated with the horizontal magnetic-field geometry peculiar to the equatorial region. It is generally also free of the direct influence of energetic particle precipitation and large electric fields associated with solar wind–magnetosphere interactions. The energy deposited into the neutral atmosphere at high latitudes, however, can profoundly influence the midlatitude ionosphere by changing the atmospheric circulation patterns that determine the neutral gas composition, as discussed in Chapter 3.

The midlatitude F region is generally thought to be the best understood region of the ionosphere. The maximum electron concentration N_mF2 occurs at the level where downward diffusion and electron loss by recombination are of comparable importance. At that altitude, the electron concentration is nearly in photochemical equilibrium, with a balance between the photoionization of atomic oxygen and electron recombination with molecular ions (formed via ion–neutral charge exchange of O^+ and N^+ with molecular gases). Variations in the ratio $O/(N_2 + O_2)$ thus can lead to important variations in N_mF2. Vertical ion drift (due to neutral winds or electric fields) can shift the altitude of the F2 peak and also the value of N_mF2. Exchange of ionization along the magnetic-field lines between the ionosphere and the protonosphere (the high-altitude region where H^+ ions predominate) is of importance to the maintenance of the nighttime ionosphere. Appropriate combinations of these various factors have been applied to explain a variety of phenomena initially viewed as "anomalous." Yet there are other gross features of the global morphology of the midlatitude ionosphere that still await basic understanding.

NIGHTTIME TRANSPORT EVENTS

The upper portion of Figure 5.7 illustrates a night for which the qualitative evolution of the ionosphere was quite simple. Through sunset and following through the

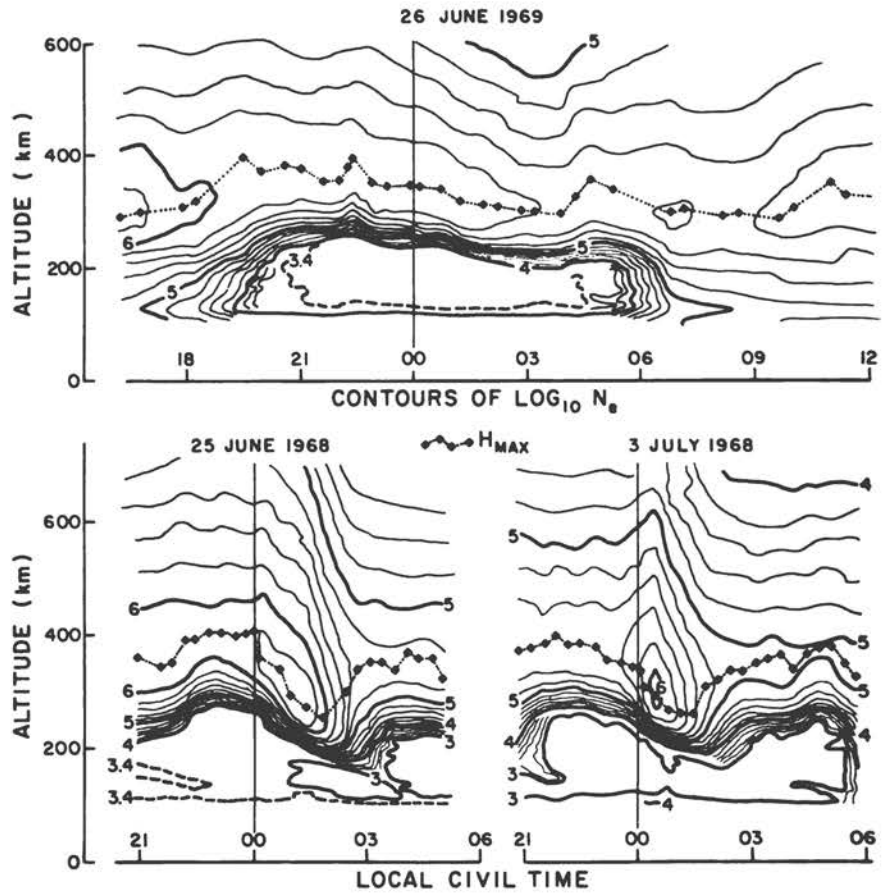


FIGURE 5.7 Contours of $\log_{10}$ of electron concentration (electrons cm^{-3}), on an altitude versus local time grid, showing patterns of behavior in the nighttime midlatitude ionosphere over Arecibo at 18.3° N geographic (29° N geomagnetic) latitude. Equatorward neutral winds are important in maintaining the nighttime ionosphere, as in the upper half of the figure. Transient poleward wind components of unknown origin generally sweep across the entire midlatitude sector of the globe, manifested here by rapid order-of-magnitude drops in topside plasma concentrations and transient bottomside increases of two to three orders of magnitude.

night, N_mF2 falls monotonically until its sunrise increase. Quantitatively, however, the electron concentration maintains nearly its daytime value throughout the night. Even for nights when N_mF2 falls much more dramatically, a long-standing dilemma has been how the midlatitude N_mF2, in the absence of sunlight for roughly 10 hours, maintains its high observed values. In the case seen here, an equatorward wind, which has an upward component along the magnetic field, raised H_{max} (the altitude at which N_mF2 occurs) at sunset to a level where the molecular concentration and resultant recombination rate were an order of magnitude lower than during the daytime. In fact, the general F-region atmospheric wind pattern tends to blow from the hot subsolar to the cold antisolar point, as well as from the summer to the winter hemisphere. This effect, combined with a downward flow of plasma from the plasmaspheric reservoir, has been found adequate to maintain the nighttime F region when tested against the very limited body of data for which the necessary measurements were available.

However, a much more common nighttime evolution includes an abatement or reversal of the equatorward wind in the middle of the night, allowing the plasma to slide down the magnetic-field lines to regions of a greater recombination rate. As illustrated in the lower portion of Figure 5.7, this "midnight collapse" can lower N_mF2 by an order of magnitude over a time period of approximately an hour and enhance electron concentrations (and conductivities) below 200 km by 2 to 3 orders of magnitude. This event spans the entire midlatitude sector, although less pronounced and occurring later in time with increasing latitude. This ionospheric effect was discovered over two decades ago by workers in high-frequency communication prediction services. Only now, however, are the measurements becoming available that may identify the driving mechanism for this wind. Rishbeth (1975) and King and Kohl (1965) have discussed upper atmospheric circulation and its ionospheric effects.

DAYTIME ANOMALIES

A variety of daytime midlatitude N_mF2 anomalies have been known for the past 10 to 40 years; while many have been explained, others remain puzzles. Although the solar ionizing flux is 6 percent greater in January than July because of the annual variation in the earth–sun distance, the global average value of N_mF2 exhibits about triple this annual variation (greater in December than June). A further semiannual variation yields equinoctial maxima in daytime N_mF2 comparable in amplitude with the annual variation, throughout the solar cycle. This semiannual variation is presumably associated with the semiannual upper atmospheric temperature variation now known to exist.

There is also a seasonal or winter anomaly in daytime N_mF2; although the solar radiation is more nearly overhead in local summer, N_mF2 achieves significantly greater daytime values in local winter! This anomaly is absent at sunspot minimum but becomes increasingly apparent with increasing solar activity. The effect appears to correlate best with geomagnetic latitude in the northern hemisphere but with geographic latitude in the southern hemisphere. N_mF2 is observed to increase sooner and more rapidly at sunrise in winter than in summer, while sunset rates of decrease are comparable for the two solstices.

NIGHTTIME SUB-F-REGION IONIZATION

Development of highly sensitive incoherent-scatter measurement techniques has recently permitted continuous measurement (at 30° N geomagnetic latitude) of a persistent nighttime layer of ionization, shown in Figure 5.8, in the 10^9–10^{10} electrons m^{-3} concentration range, lying between the bottomside of the F region and the nighttime E layer near 100 km. Detailed analysis of the time evolution of the layer has shown electron production rates within the layer to be much higher than would follow from scattered Lyman-α and Lyman-β radiation, which produce the normal nighttime nonmetallic E-region ionization. Thus this newly found weakly ionized layer takes on special significance, as it requires a qualitatively new and as yet unknown ionization source. Energetic particle fluxes have been suggested, but there are no clues as to what their source might be.

TRAVELING IONOSPHERIC DISTURBANCES

For many years, ground observers have seen ionospheric disturbances travel from high latitudes toward the equator. We now know that these traveling ionospheric disturbances (TID's) are produced by auroral heating events that send out high-altitude atmospheric waves. Near the directly heated regions, large upwellings of gas from the lower thermosphere drastically alter both the neutral and ion compositions at high altitudes. The disturbances propagate as waves in the neutral atmosphere that can be detected through their effects on the ions.

A class of small-scale TID's with periods of approximately 15 min has also undergone intensive study, and within the past decade these disturbances have been shown to be a manifestation of gravity waves propagating in the neutral atmosphere. Strong viscous dissipation of these waves much above 200 km constitutes an appreciable energy source that can modify temperature profiles, and thus the neutral and ion compositions, over appreciable areas of the globe. The sources of such TID's have variously been suggested to be particle precipitation and electric currents in the auroral zone, earthquakes, atmospheric tides, winds, and large-scale weather perturbations. A classic example of TID's equatorward of the midday auroral zone as seen by an Atmosphere Explorer satellite is shown in Figure 5.9, where periodic variations in ion concentration correlate closely with those in the vertical ion velocity, with a spatial scale size of about 150 km. The auroral-zone energetic-particle and Joule heat-

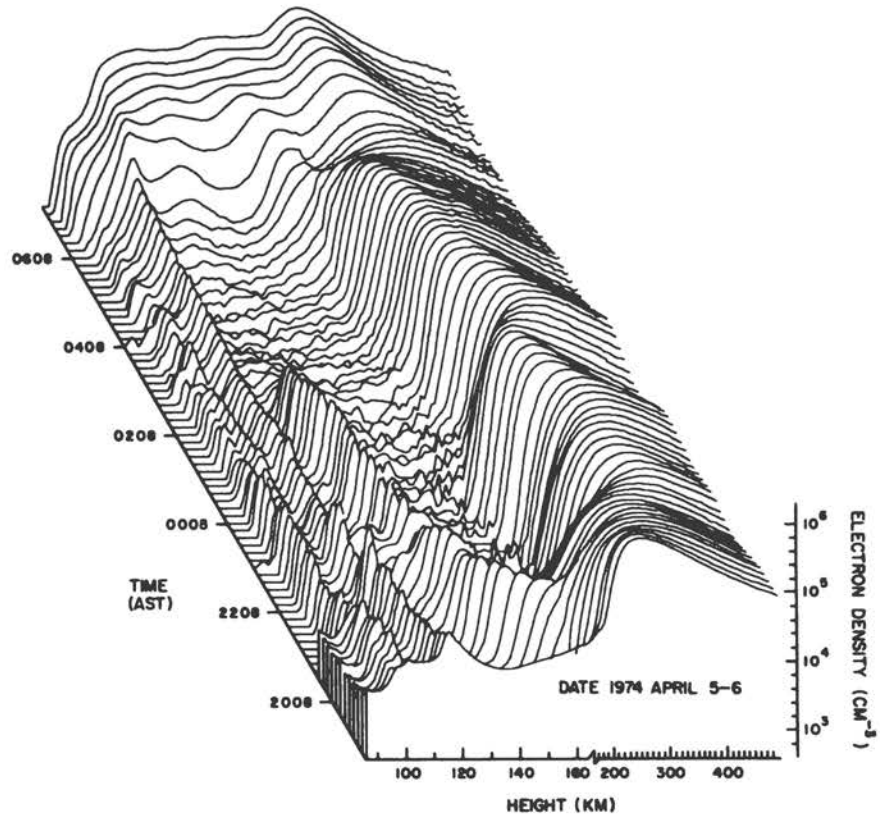


FIGURE 5.8 New high-sensitivity Arocibo incoherent-scatter radar data. These have unveiled a weak nighttime sub-F-region layer of ionization, found commonly from sunset till midnight. Present analysis indicates that its presence requires a previously unsuspected nighttime ionizing source. The characteristic downward motion exhibited here is attributed to "trapping" in the mode of an atmospheric tidal oscillation. The layer is bracketed above by the long-lived F layer tracing large-scale atmospheric oscillations and below by the relatively stable E region predominantly due to deeply penetrating ionizing radiation.

ing shown in Figure 5.10, as deduced over a several-hour period from ground-based incoherent scatter and photometric data at Chatanika, Alaska, exemplifies the magnitude of these energy inputs, peaking nearly an order of magnitude greater than the solar energy input above 95 km.

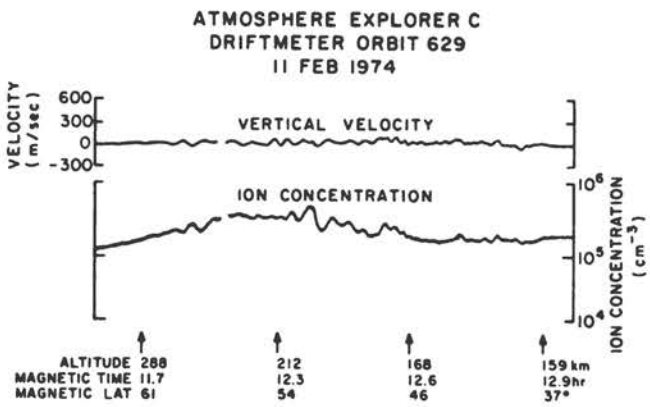


FIGURE 5.9 Satellite measurements showing ion concentrations periodically covarying with vertical plasma velocities, indicating the presence of a 150-km-wavelength gravity wave propagating equatorward from about a 60° magnetic latitude.

STABLE AURORAL RED ARCS

At times of geomagnetic disturbance, strongly enhanced, although subvisual, upper atmospheric airglow emissions are often observed at midlatitudes over a wide longitude sector. These 630-nm emissions, called stable auroral red arcs or SAR arcs, are now known to follow from impact excitation, from the ground state to the (O^1D) first excited state of atomic oxygen, by electrons in the high energy tail of the velocity distribution of a hot thermal plasma. The ambient ionospheric electrons exciting the arc are heated by energy conducted down magnetic-field lines from the magnetosphere. The source of this energy is thought to be energetic ring-current protons, although present research still seeks to distinguish the role of various possible modes of wave–particle interactions involved in transferring energy from the several keV protons to the ambient thermal electrons for conduction downward.

There is, in fact, a persistent increase in the topside F-region plasma temperature at night near the plasmapause, suggestive of a continual interaction with trapped particles in the magnetosphere near this boundary. Somehow, it seems, this temperature anomaly becomes large enough to excite the red emissions, but just how this comes about we do not know.

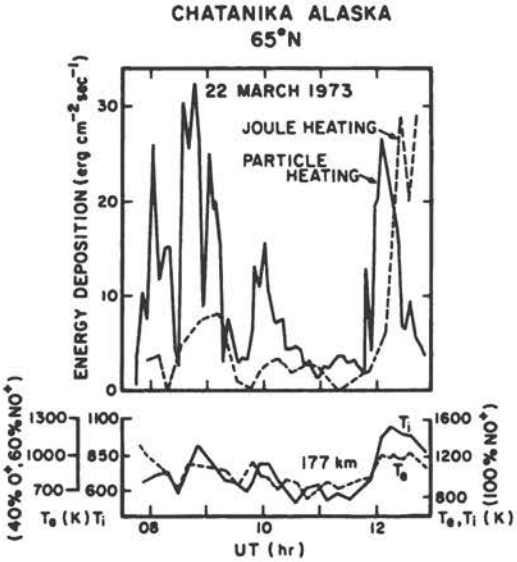


FIGURE 5.10 Variations in heating rates and gas temperatures as measured by the backscatter radar at Chatanika, Alaska, at 65° N latitude. At high latitudes, Joule and particle heating can greatly exceed the solar ultraviolet heating rate (roughly 0.5×10^{-3} and 3×10^{-3} J m^{-2} sec^{-1} above 120 and 95 km, respectively) and can become important to even the global-scale thermospheric heat budget. Part of this energy drives gravity waves such as those noted in Figure 5.9.

CONTROLLED IONOSPHERIC MODIFICATION

A new frontier in ionospheric physics has recently been opened by experiments that modify the ionosphere using intense high-frequency radio waves (Perkins *et al.*, 1974). Rather than simply observing the ionospheric conditions offered by nature, the opportunity is presented to perform actual ionospheric experiments by examining the ionospheric response to controlled changes over a range of conditions.

The oscillating electric field of a radio wave propagating through a plasma accelerates the electrons. Those electrons suffering collisions before reradiating the stored rf energy will thereby convert their ordered component of motion into a random component, or heat. When the electron collision time approaches the time for the wave to propagate through the medium, a substantial fraction of the rf energy can go into electron heating. Even though powerful facilities were constructed to examine this effect, very modest ionospheric distortions were anticipated in the initial experiments. However, the initial experiments produced striking and unexpected modifications far richer in information than anticipated. Some of these are indicated in Figure 5.11. The high-power densities excited plasma instabilities, enhancing plasma wave amplitudes over 10^5 times their normal thermal level, thus opening a new phase of plasma-physics studies in the ionosphere. The ionosphere, as a low-pressure, low-magnetic-field-strength "laboratory plasma" without walls, has many experimental advantages over conventional laboratory plasmas.

The plasma instability principally heats the tail of the electron velocity distribution at a narrowly defined ionospheric altitude through a process whose details are under current theoretical investigation. The several eV electrons thus generated allow studies of their transport in the atmosphere; these suprathermal electrons also excite neutral species. Controlled electron heating, with or without the instability and energetic electron heating component, allows direct measurement of electron cooling rates by various processes. This should soon lead to significant advances in the important problem area of the thermospheric and ionospheric heat balance. It also will allow useful work in the atmospheric (rather than labo-

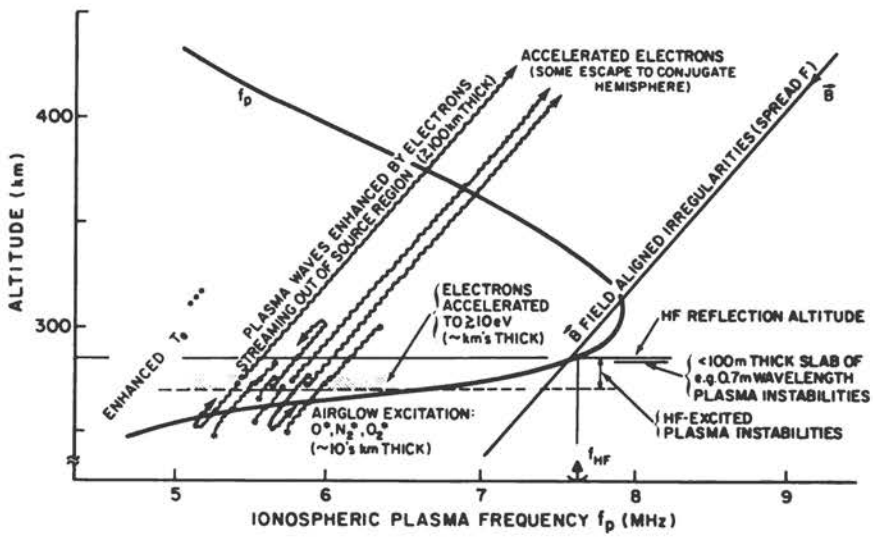


FIGURE 5.11 Effects produced by a ground-gassed transmitter of power aperture of the order of 10^4 MW m^2 in the 4–12 MHz frequency range. Energy deposited in the ionospheric plasma alters both the thermal and nonthermal properties of its charged-particle population. Controlled experiments have applications to aeronomy, chemical rates, atomic cross sections, communications, and a number of areas of plasma physics.

ratory) environment on selected temperature-dependent reaction rates. The intense high-frequency radio waves also produce spread F over a wide range of scale sizes and with measurable evolution times. Study of controlled experiments of this nature could contribute to a better understanding of naturally occurring ionospheric spread F.

5.7 HIGH-LATITUDE PHENOMENA

The high-latitude ionospheric region is defined to be that region where interaction with the solar wind, through magnetospheric coupling, has a strong influence on the behavior of the ionosphere. This influence is exerted through the impingement of energetic particles, by the imposition of electric fields, and by providing a sink for ambient ionospheric plasma. It is essentially the region poleward of the plasmapause.

ION-CONCENTRATION PERTURBATIONS

Ionization of the upper atmosphere by energetic particles can appreciably affect the ion concentration both at night and during the daytime. The apparently direct entry of plasma-sheath particles into the dayside cusp region causes increases in the ambient ion concentrations above the levels established by solar photoionization that can approach an order of magnitude, as indicated in the lower portion of Figure 5.12. Since the particle precipitation is not uniform, neither is the resultant ion concentration distribution. The high-latitude region is the only place on earth that has large ion-concentration irregularities (spread F) on a regular basis while sunlit.

In the dark nightside auroral zone, the ion concentration is even more variable, particularly at lower altitudes where recombination is faster. In the F2 region, the mean-square deviation of ion concentration from its mean value, for scale sizes less than 30 km, is nearly always greater than 10 percent in the auroral oval and over the polar cap, being larger over the winter than the summer pole.

Very-small-scale size (tens to hundreds of meters) irregularities may be observed in the ion currents swept up by a satelliteborne Faraday cup at high latitudes. Since these currents are proportional to $N_i V_r$, where N_i is the ion concentration and V_r is the relative velocity between the satellite and the ions, it is not possible from these measurements alone to determine whether the observed structure is in N_i, V_r, or both parameters. On some occasions it has been established that for scale sizes of the order of 1 km there exists a dispersion in V_r of a few

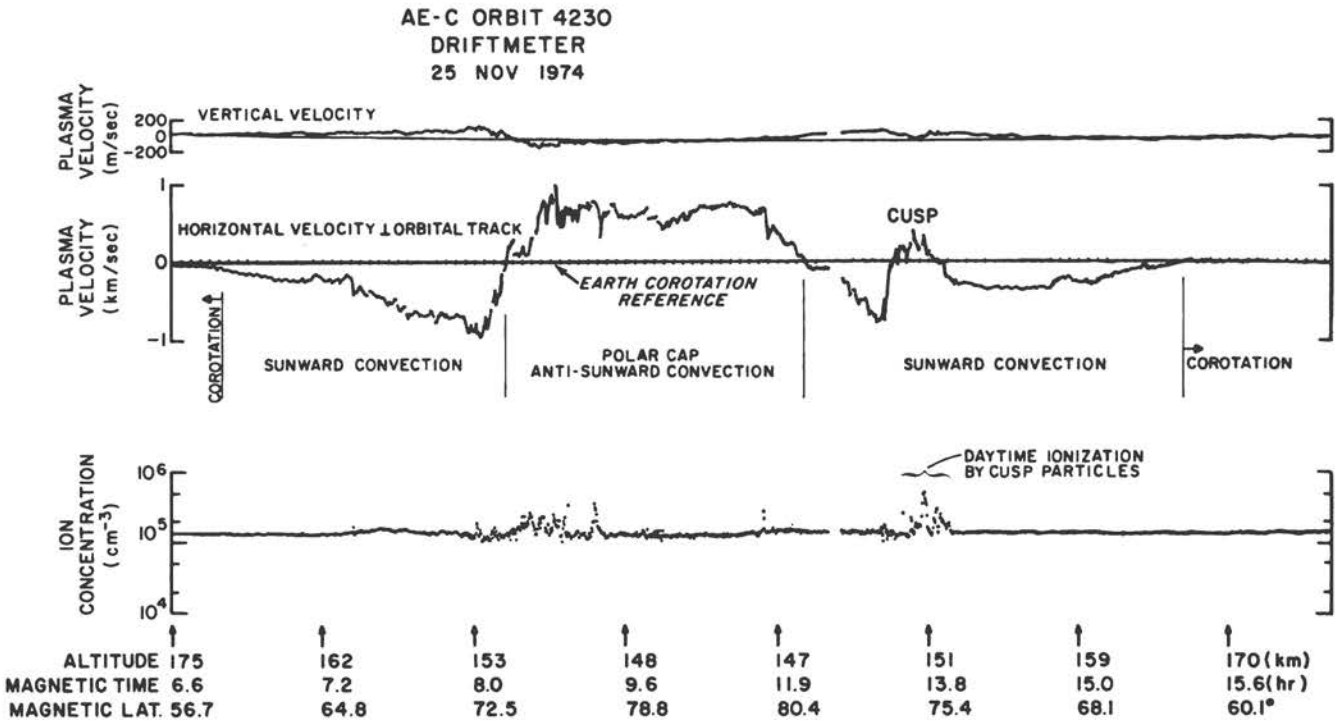


FIGURE 5.12 Upper portion: the vertical velocity component of ionospheric plasma and the horizontal velocity component perpendicular to the satellite velocity vector as deduced from an ion drift meter on Atmosphere Explorer-C in a pass across the auroral zone (altitude, magnetic local time, and magnetic invariant latitude are indicated). Lower portion: simultaneous measurements of ion concentration. The ionization rate is proportional to the square of the ion concentration. Enhancements in ion concentration can be seen for which the particle ionization rates exceed that of solar EUV by at least an order of magnitude.

percent, approximately the same magnitude as the simultaneously observed irregularities in the Faraday ion current. There is no explanation for these small-scale irregularities in N_i in V_D or in both, but they may well result from a plasma instability, perhaps associated with magnetic-field-aligned currents.

POLAR CONVECTION

It has been well established that there exists a large-scale high-latitude electric field that convects plasma away from the sun over the polar caps. This flow diverges near the auroral oval in the midnight sector and returns sunward at lower latitudes just outside the auroral oval through both the morning and evening sectors, forming a double-cell convection pattern. Typical flow velocities are of order 0.5 km sec^{-1}, but values as great as 3 km sec^{-1} have been observed. The upper portion of Figure 5.12 shows an example with convection velocities as high as 1 km sec^{-1}. Figure 5.13 shows convection patterns over both polar regions. Details of the overall convection patterns are quite variable; the relative symmetry between the morning and evening cells, and probably between the north and south polar regions, depends on solar-wind parameters, particularly the direction of the interplanetary magnetic field. Usually superimposed on this general convection pattern are small-scale random flow components with amplitudes of approximately the same size as the main convection velocity.

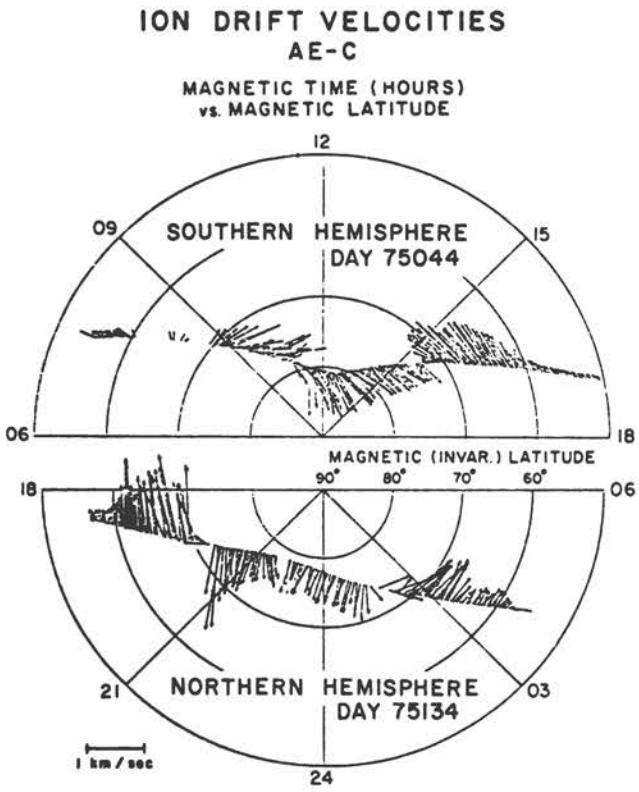


FIGURE 5.13 Plasma-drift patterns observed from satellites. Midlatitude ionospheric plasma approximately corotates with the earth. Poleward of roughly 60° geomagnetic latitude, electric fields due to solar wind–magnetospheric interaction drive the ionospheric plasma in a "convection pattern" of antisunward flow over the pole, with a return lower latitude flow toward the sun through both the morning and evening sectors. This flow can now be mapped, as illustrated here, in the form of the horizontal component of the plasma velocities perpendicular to the magnetic field along satellite tracks.

These large electric fields and drift velocities have important consequences for the ambient plasma. Collisions between the ions and the neutral gas tend to raise the ion temperature to the value

$$T_i \approx T_n \left[1 + \frac{2}{3} \left(\frac{V_D}{v_t} \right)^2 \right] , \qquad (5.6)$$

where V_D is the relative velocity between the plasma and neutral gas and v_t is the mean neutral-particle thermal velocity. This relationship is established approximately in an ion–neutral collision period τ_{iN}. In a neutral–ion collision period τ_{Ni}, the neutral particles are also affected; they not only become heated, but they also tend to acquire the ion drift velocity. It should be noted that $\tau_{iN} = (N_i/N_n)\tau_{Nb}$ and that while τ_{iN} is approximately 1 sec at 250 km, τ_{Ni} is at least 10^3 times greater. When $V_D \geq v_t$, which is often the case, the ion velocity distribution departs appreciably from Maxwellian and the concept of ion temperature can only be defined in terms of the mean ion energy. In this sense, F-region ion temperatures greater than 4000 K have been observed where $T_n \approx 1000$ K. Near 120 km, τ_{iN} is approximately equal to the ion angular cyclotron period, and the energy dissipation per ion is a maximum there for a given electric field. It is this altitude range that is subject to large Joule-dissipation heat inputs in the daytime—and even at night in the auroral regions where new ionization is supplied by energetic particles. This Joule heating provides an energy source over the polar regions that is comparable to the entire global absorption of solar ultraviolet energy above 110 km. Figure 5.10 shows a case of strong Joule and particle heating.

The rate-limiting process in the recombination of O^+ ions is their reaction with neutral molecules. These reactions are quite energy-dependent, and their rates increase with temperature above approximately 1000 K. The large convection velocities increase these charge-exchange reaction rates not only because of the increased ion temperature due to collisional heating but also because the energy of each collision is increased by the relative drift motion itself. Large increases in the (NO^+/O^+) ratio are in fact observed to correlate with large ion drift velocities.

It is apparent that horizontal ionization transport across the terminator at these huge speeds must be reckoned with in the plasma continuity equation. Perhaps even more important, however, is the vertical component of the electrically induced convective motion in the $\mathbf{E} \times \mathbf{B}$ di-

rection. Vertical displacements of the ions can drastically increase or decrease their recombination time and their mean collision period with neutral particles; the latter parameter controls the amount of Joule heating that will occur. It has recently been discovered that metallic ions are relatively abundant below 200 km at high latitudes. While their concentrations do not exceed 10^{10} ions m^{-3} above 135 km, these ions can have a great influence on the quantity of Joule heating that occurs because of the high ion–neutral collision frequency in their height range.

At great heights, the magnetosphere is essentially devoid of ions ($<10^7$ ions m^{-3}) at magnetic latitudes higher than the plasmapause boundary. Since there is a continuous source of H^+ ions in the upper F region at all latitudes by charge exchange of O^+ with H, some loss mechanism is required to depopulate this region of space. One mechanism that appears adequate to this task is the F-region ion convection pattern described above, provided that these ion motions follow magnetic-field lines that become open over the polar cap (i.e., that they become connected to the interplanetary magnetic field). Thus, during the time that the tubes of magnetic force are open, the low-energy plasma can expand into the earth's magnetotail beyond the reconnection distance and become lost. If this flushing occurs frequently enough, the field lines will remain essentially empty. Since the H^+ source in the ionosphere is of the order of 10^{12} ions m^{-2} sec^{-1} and the volume of the field tubes per m^2 is approximately 10^{17} m^3 at 60° magnetic latitude (just beyond the normal plasmapause boundary), the exposure would have to take place every 10^4 sec in order to keep $N_i < 10^7$ ions m^{-3}. This escape of H^+ ions at high latitudes has been named the polar wind by Axford (although the principal source of the escaping ions is not at polar latitudes, and there is even some evidence that the escape mechanism fails at times over the winter pole when there is insufficient energy available to evaporate the hydrogen ions).

There must also be times and places where large fluxes of O^+ ions evaporate from the ionosphere, because they are seen at times to precipitate into the ionosphere with energies of several kilovolts. These energetic O^+ fluxes (10^{11} to 10^{12} ions m^{-2} sec^{-1}) were totally unexpected, and as yet no obvious source mechanism for the O^+ ions has emerged. It has even been suggested that they come from the solar wind, but this seems unlikely. Most probably they are extracted from the ionosphere in some manner, and then energized by a magnetospheric acceleration process before they are returned to the ionosphere; they will be the subject of intense investigation in future magnetospheric programs.

HIGH-LATITUDE BOUNDARIES

We have used the plasmapause to mark the boundary between the midlatitude and high-latitude regions. This boundary is not distinct at F-region heights nor even at 1000-km altitude. At altitudes greater than 2000 km, the transition is often quite sharp, with the H^+ concentration decreasing by one or two orders of magnitude in a few degrees of magnetic latitude.

Poleward of the plasmapause, there is often another distinct boundary, usually near 60° geomagnetic latitude, that has been termed the "midlatitude" trough, although here we consider it a high-latitude feature. This trough in ion concentration is mainly a nighttime phenomenon that is seen all the way down to the F2 peak. The F2 peak ionization decreases gradually beyond the plasmapause to values near 10^{10} electrons m^{-3} and then often abruptly increases by up to an order of magnitude in approximately 1° of latitude. This abrupt increase coincides with the sharp equatorward edge of the soft auroral zone, whose energetic particles are identified with the plasma sheet in the magnetospheric tail. The cause of the "midlatitude" trough is not known, but it has been observed on many occasions that a very large (>1 km/sec) sunward plasma drift coincides with this region, as shown in Figure 5.14, and that the concentration ratio (NO^+/O^+) often maximizes there.

Usually the plasma drift in the soft auroral zone is sunward for several degrees poleward of its equatorward boundary. Poleward of this region in the morning and evening sectors, a distinct flow reversal usually occurs, accompanied by intense fluxes of auroral particles. This boundary is associated by many researchers with the transition from closed magnetic-field lines to the open geomagnetic tail, higher latitudes being identified with the polar cap. The boundaries of the polar cap are less distinct in the noon and midnight sectors, where plasma is convected into and out of this region. The influx appears to take place over a rather narrow throat on the dayside, and the flow then expands rapidly so that the largest flow velocities occur near the reversal boundaries, as shown in Figure 2.6 of Chapter 2. There is much confusion in the flow patterns near midnight.

5.8 PRESENT AND FUTURE PROGRAMS

Aside from limited rocket program, the principal ionospheric measurements being performed are from the Atmosphere Explorer and ISIS satellites, and some satellites in the COSMOS and INTERCOSMOS series. ISIS II makes *in situ* measurements only near 1400 km, but the topside sounder and auroral optical scanners provide useful data down to and below the F2 peak. Atmosphere Explorers provide *in situ* data from 4000 km down to 130 km from a sophisticated set of sensors, and they also perform some optical probing and solar monitoring. These satellite data are supplemented in many ways with measurements from ground-based backscatter radars and optical sensors. It seems likely that considerable progress will be made in a number of the problem areas outlined in this chapter as the data now being accumulated are scrutinized and digested.

The principal new satellite program that will shed

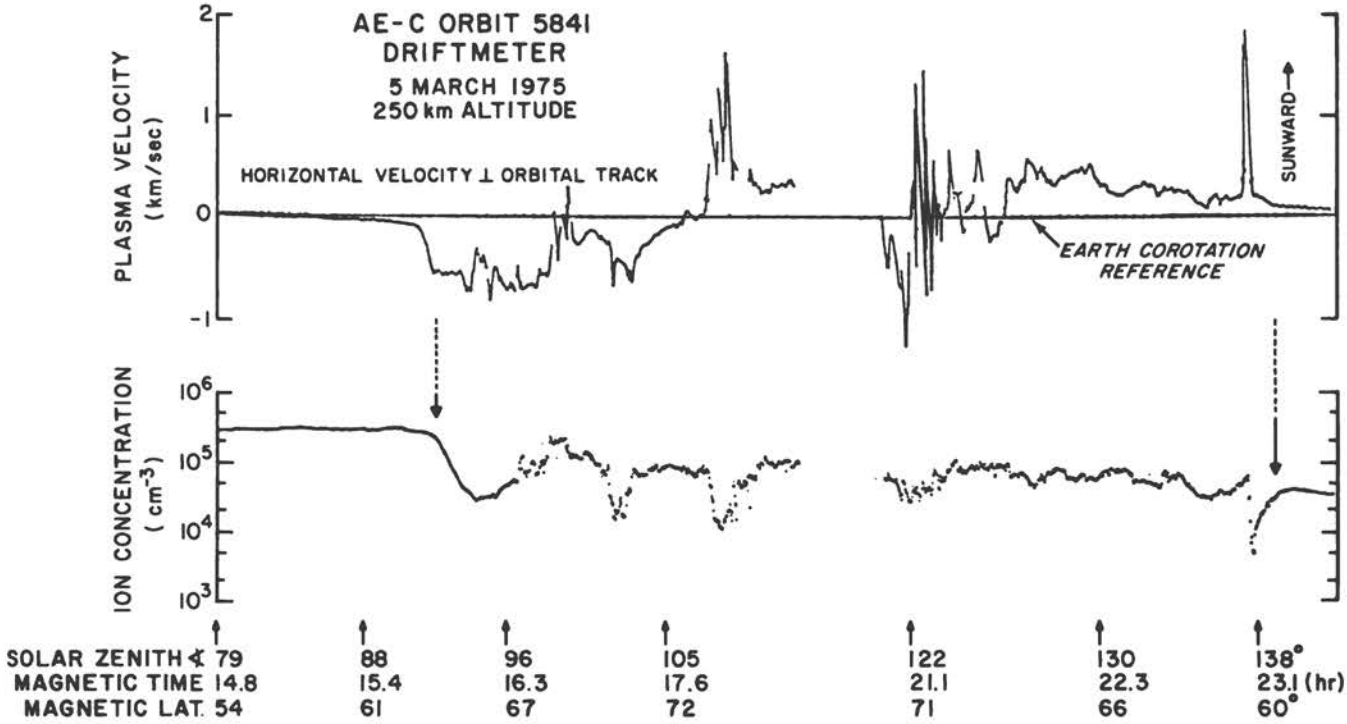


FIGURE 5.14 Lower portion: the "midlatitude trough," in which the F-region maximum ion concentration gradually decreases poleward of the plasmapause latitude to values near 10^{10} ions m^{-3} and then abruptly increases by up to an order of magnitude in approximately 1° of latitude. Upper portion: the horizontal plasma drift. While the cause of this (predominantly nighttime) trough phenomenon is not yet known, it is often noted to coincide with very large (> 1 km/sec) sunward plasma drift and peak (NO^+/O^+) concentration ratios. A clear nighttime trough is seen here near the right-hand arrow; a trough phenomena is also manifested where the ion concentration rolls off poleward of the left-hand arrow.

appreciable light on ionospheric behavior is the Electrodynamic Explorer (EE) mission now under active consideration. This multiple satellite effort is designed specifically to examine the coupling between the solar wind and the ionosphere (atmosphere) via the intervening magnetosphere. Many of the questions set forth here will not be answered in the near future if this mission is not undertaken.

The construction of a moderately high-latitude sophisticated backscatter station (called the Upper Atmosphere Observatory, or UAO) is also under consideration, and it, together with other high-latitude ground-based instrumentation being installed for the International Magnetospheric Studies (IMS), could greatly enhance the scientific benefits from the EE satellites.

In the Space Shuttle era, there are no firm plans for investigations of the ambient ionosphere. The principal planning effort is being devoted to the AMPS payloads, which are designed mainly to carry out active experiments in the ionosphere–magnetosphere laboratory. It is, of course, possible that smaller "free flyers" of the Atmosphere Explorer category could be launched from the Space Shuttle, provided that (as seems likely) sufficiently challenging problems still exist then to warrant such an effort.

REFERENCES

Donahue, T. M. (1968). Ionospheric composition and reactions, *Science 159*, 489.

Evans, J. V. (1975). High-power radar studies of the ionosphere, *Proc. IEEE 63*, 1636.

Hanson, W. B. (1975). Earth's dynamic thermosphere, *Astronaut. Aeronaut. 13*, 16.

Hanson, W. B., and F. S. Johnson (1961). Electron temperatures in the ionosphere, in *Les Spectres des Astres dans l'Ultraviolet Lointain, 20*, Mem. Soc. R. Sci. Liège.

Havens, R. J., H. Friedman, and E. O. Hulburt (1955). The ionospheric F2 region, in *The Physics of the Ionosphere*, The Physical Society, London, pp. 237–244.

Hulburt, E. O. (1928). Ionization in the upper atmosphere of the earth, *Phys. Rev. 31*, 1018.

King, J. W., and H. Kohl (1965). Upper atmospheric winds and ionospheric drifts caused by neutral air pressure gradients, *Nature 206*, 699.

Perkins, F. W., C. Oberman, and E. J. Valeo (1974). Parametric instabilities and ionospheric modification, *J. Geophys. Res. 79*, 1478.

Rishbeth, H. (1975). F-region storms and thermospheric circulation, *J. Atmos. Terrest. Phys. 37*, 1055.

Waynick, A. H. (1975). The early history of ionospheric investigations in the United States, *Phil. Trans. R. Soc. Lond., Ser. A 280*, 11.

The Ionospheric D Region

6

CHALMERS F. SECHRIST, JR.
Aeronomy Laboratory
University of Illinois at Urbana–Champaign

6.1 PROLOGUE

This chapter describes the D region of the earth's ionosphere, arbitrarily defined to be the lower ionospheric region lying approximately between 60 and 90 km. This chapter emphasizes some interesting facets of knowledge about the D region and several problems presently of high interest to scientists studying the D region. No attempt is made to review all aspects of D-region research or to discuss the details of all outstanding problems. Rather, the purpose of this chapter is to introduce to the general scientific community the rich variety and complexity of processes and problems that are currently under study in this unique region of the upper atmosphere. Without doubt, the D region is the least understood portion of the ionosphere, and it offers the scientific community a stimulating and challenging assortment of strikingly interdisciplinary unsolved problems. To understand it completely will probably require the efforts of meteorologists, radio scientists, atmospheric electricity workers, laboratory physicists and chemists, aeronomers, and magnetospheric physicists, among others. In a sense, one could think of the D region as a microcosm of the middle atmosphere (15–100 km), because the problems there are somewhat representative of those encountered in lower altitudes.

Most early knowledge of the D region came from ground-based experiments involving radio-wave propagation. The field intensity and phase path of low-frequency (lf) and very-low-frequency (vlf) radio waves were observed to vary in accordance with a weakly ionized layer below the ionospheric E region, the D region. The absorption of high-frequency (hf) radio waves was attributed to the daytime D-region ionization, which exhibits regular diurnal and seasonal variations at low and middle latitudes. In the 1950's, the wave interaction (cross-modulation) and partial-reflection experiments were introduced to study D-region electron density profiles; direct rocket measurements appeared in the late 1950's and early 1960's. It is generally agreed that *in situ* rocket measurements of electron density offer the best height resolution and greatest absolute accuracy. In recent years,

the incoherent-scatter radar method has been successfully applied to the measurement of electron densities in the upper D region above about 80 km.

The undisturbed daytime upper D region is produced mainly by the photoionization of nitric oxide and metastable molecular oxygen, $O_2(^1\Delta)$, by solar ultraviolet radiation. During periods of high solar activity, hard x rays are an important source of ionization. Also, energetic electrons precipitating from the magnetosphere with energies greater than 40 keV may serve as a significant source of ionization in middle latitudes during and following certain geomagnetic storms. Galactic cosmic rays (GCR) provide the main source of ionization in the lower D region.

The relatively high neutral atmospheric pressure ($\sim 10^{-1}$ to 10^{-3} Torr) in the D region is responsible for the complexity of the ion chemistry there compared with the ion chemistry at higher levels. Three-body reactions occur efficiently. This is a region of transition between massive water cluster ions and simple molecular ions. Also, negative ions are present, mainly in the lower D region, because the relatively high neutral gas density allows rapid three-body attachment of electrons to molecular oxygen. The resulting O_2^- ions react rapidly with various minor neutral gases to form other negative ions.

The winter D region in middle latitudes exhibits a high degree of variability in electron density as determined by numerous rocket- and ground-based measurements. There is presently wide interest in the subject of the "winter anomaly," which refers to the abnormally high values of medium frequency (mf) and hf radio-wave absorption measured on certain days and groups of days in winter. It is now generally accepted that the anomalously high electron densities are caused by the redistribution of one or more key minor neutral gases by means of vertical or horizontal transport processes or both peculiar to the winter mesosphere in middle latitudes. Thus, the winter anomaly is a manifestation of dynamical processes in the neutral atmosphere, and very striking associations have been detected between various measured stratospheric meteorological parameters and D-region observations such as radio-wave absorption, electron density, sodium airglow emissions, winds, and temperature. These associations provide evidence for stratosphere–mesosphere–ionosphere coupling.

The winter anomaly has also been ascribed in part to electron-density enhancements induced by the effects of precipitating energetic electrons from the radiation belts of the magnetosphere. Following certain geomagnetic storms, there is precipitation of energetic electrons into the middle-latitude D region, and this magnetic-storm aftereffect may persist for several days or weeks. This aftereffect has been observed on numerous occasions in winter, probably because the undisturbed or normal D-region ionization production rates are relatively low then because of the high solar zenith angle. Thus, ion production by particles then is competitive with the ion production rates due to solar radiation. In summer, the midday zenith angles are relatively small at middle latitudes and the solar EUV (extreme ultraviolet) and x radiations dominate the energetic electrons as sources of ionization. Thus, it is likely that in winter at middle latitudes the D-region ionization variability may be attributed to a combination of dynamical processes in the stratosphere and mesosphere and precipitation of energetic electrons from the magnetosphere following geomagnetic disturbances. Therefore, the D region must be viewed as a portion of a very complex atmospheric system that is hydrodynamically coupled to the stratosphere and mesophere and electrodynamically coupled to the magnetosphere.

Dynamical or meteorological effects in the D region are significant for minor neutral species possessing long chemical lifetimes. Nitric oxide is an excellent example, because model calculations show that the altitude distribution of the NO concentration is relatively sensitive to the magnitude of vertical eddy transport as parameterized by an effective vertical eddy-diffusion coefficient (see Chapter 7 for a description of vertical eddy transport). The altitude distribution of long-lived minor neutral species is a factor in the determination of the height profile of the ionized species. Thus, transport processes in the neutral mesosphere influence the ionization profiles through the redistribution of the minor neutral gases.

Presented in the following sections of this chapter are some excerpts of knowledge of D-region electron densities, minor neutral constituents, ionization sources, positive ions, negative ions, and particulates. Additionally, the important subjects of the winter and disturbed D regions are briefly discussed. It will become evident that the D region is indeed a fruitful area for future upper atmospheric research and that the chemical and dynamical processes specific to this region are currently far from being clear. This chapter concludes with a summary of several examples of outstanding problems whose solutions must be actively pursued in the future if progress is to continue on this challenging region of the upper atmosphere.

6.2 ELECTRON DENSITIES

D-region electron density profiles are of great interest to aeronomers and radio scientists. Electron densities influence the propagation of elf (extremely low-frequency) to hf (high-frequency) radio waves through their effect on reflection height, absorption, and wave polarization. At night, the D-region electrons below ~90 km practically disappear because of attachment to molecules. Techniques for measurement of electron concentrations have been discussed and the results intercompared by Sechrist (1974). Electron density profiles have been measured by ground-based radio techniques such as partial reflection (differential absorption) and wave interaction (cross modulation); in recent years the incoherent-scatter radar method has been used to measure electron density pro-

files above about 82 km. Rocket methods such as differential absorption and/or Faraday rotation together with dc probes have been used extensively by several experimental groups with the result that high resolution and accurate electron density profiles are available for a wide variety of solar and geophysical conditions. In comparison with other D-region neutral and ionized parameters, it should be noted that measurements of electron density are relatively simple and inexpensive. Furthermore, measured electron densities of high absolute accuracy serve as useful constraints in theoretical model studies of concentrations of neutral and ionized species.

The temporal and geographical variations of D-region electron concentrations have been described in detail by Thomas (1974a), who discussed the diurnal, seasonal, latitudinal, and solar-cycle variations; the anomalous behavior in middle latitudes during winter; and the effects associated with geomagnetic storms. Figure 6.1 shows median electron concentration profiles for active- and quiet-sun conditions, derived from rocket measurements of electron current, differential radio-wave absorption, and differential phase or Faraday rotation.

The sunrise variation of electron concentration profiles is illustrated in Figure 6.2, which depicts the formation of the ionospheric C layer in the lower D region. It is currently believed that the C layer forms relatively rapidly a little before ground sunrise and that solar radiation is responsible for the release of electrons from negative ions by an unknown mechanism. Recently, it has been suggested that photodissociation of complex negative ions, followed by photodetachment of electrons from simpler negative ions, may be an important presunrise mechanism in the C-layer formation. Model studies of the sunrise D region have been met with limited success.

The solar zenith-angle variation of daytime electron concentration has been measured by ground-based and rocket experiments. Generally, the data for the summer and equinox months show a clear solar control and a definite asymmetry about noon, with electron concentrations in the afternoon larger than in the morning. The solar zenith angle variation of D-region ionized species has been modeled with some success.

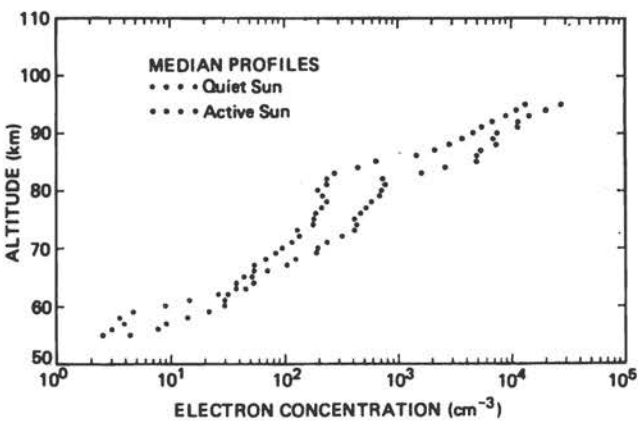


FIGURE 6.1 Median profiles of electron concentration representing five quiet-sun profiles and five active-sun profiles (Mechtly *et al.*, 1972a).

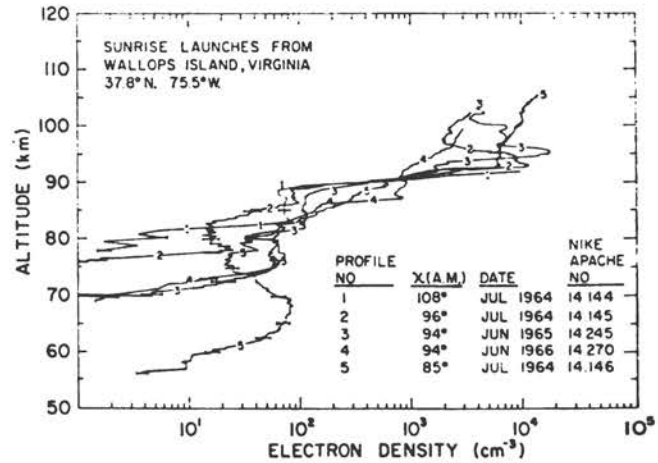


FIGURE 6.2 Electron density profiles measured by rockets during the sunrise period (Mechtly and Smith, 1968).

Results of extensive partial-reflection measurements in middle latitudes indicate very little day-to-day variation in daytime electron concentrations during the summer months. However, there is ample evidence from ground-based and rocket measurements that winter electron densities in middle latitudes exhibit considerable variability from day to day. Enhancements of the electron concentration above 80 km cause corresponding abnormally large values of hf and mf (medium-frequency) radio-wave absorption in winter, and this phenomenon is the well-known winter anomaly in the D region. The day-to-day changes in winter electron densities observed below 80 km have been labeled winter variability, and there is evidence that it is not associated with the variability of electron concentration above 80 km. Evidently, both types of winter variability or anomaly are caused by the redistribution of one or more minor neutral gases by transport process. For example, nitric oxide, atomic oxygen, ozone, metastable molecular oxygen $O_2(^1\Delta)$, and water vapor all play important roles in the positive- and negative-ion chemistry of the D region; and the altitude distributions of species having relatively long photochemical lifetimes are determined by the vertical or horizontal transport processes or both present in the region.

The limited number of electron-density measurements has not provided a completely satisfactory description of geographical variations. In particular, the latitudinal variation of electron densities is not clear. Rocket measurements between 13° S and 58° S along the west coast of South America during March and April 1965 revealed the same general features noted in other electron-concentration profiles. These are a rapid increase with height in the lower D region, a slower increase at intermediate heights, and a rapid increase in the upper D region. Although the magnetic dip angle at the four dif-

ferent latitudes where the measurements were made varied from 0° to 58°, no evidence of a systematic latitudinal variation in electron concentration was found. There was an anomalous latitudinal variation for electron concentrations in the 70–90 km height range; that is, the densities observed at middle latitudes and near the magnetic equator were about one half of the values at intermediate latitudes. It is possible that this behavior is related to the anomalous variation of mf radio-wave absorption, which exhibits maxima near ±20° magnetic dip angles and a minimum near the magnetic equator.

Perhaps the most significant feature of the daytime D-region electron-concentration profile is the steep gradient or "ledge" that occurs at an altitude between 80 and 90 km. The height of this ledge varies with solar zenith angle, season, solar activity, and magnetic conditions. The exact cause of this ledge is unknown, although it appears to occur near the height at which atomic oxygen and atomic sodium concentrations increase abruptly and where the positive cluster-ion and water-vapor concentrations decrease abruptly. These points will be mentioned in more detail in Section 6.5.

6.3 MINOR NEUTRAL GASES

Minor neutral gases are those having volume mixing ratios less than $\sim 10^{-3}$ relative to the ambient atmosphere. The minor neutral gases are definitely not minor in terms of their chemical reactivity, and they play important roles in the neutral and ion chemistry of the D region. Anderson and Donahue (1975) have reviewed the neutral composition of the mesosphere. Emphasis in this chapter will be placed on nitric oxide, atomic oxygen, metastable molecular oxygen $O_2(^1\Delta)$, ozone, oxides of hydrogen, and sodium.

NITRIC OXIDE

This gas plays several important roles in the neutral and ion chemistry of the D region. It is photoionized by solar Lyman-α radiation (121.6 nm) and serves as a dominant source of NO^+ ions and free electrons in the daytime middle and upper D region, that is,

$$NO + h\nu \rightarrow NO^+ + e.$$

Nitric oxide participates in numerous ion–molecule reactions, including those involving constituents such as O_2^+, O_3^-, and O_4^-. Much of our knowledge about the NO concentration stems from rocket measurements of the intensities of certain gamma bands produced by resonance fluorescence in sunlight. Essentially, rocketborne scanning-type spectrophotometers have been used to measure these intensities as a function of altitude. The intensity versus altitude profile leads to a determination of the NO column density. However, this method possesses limited accuracy in the D region because the bulk

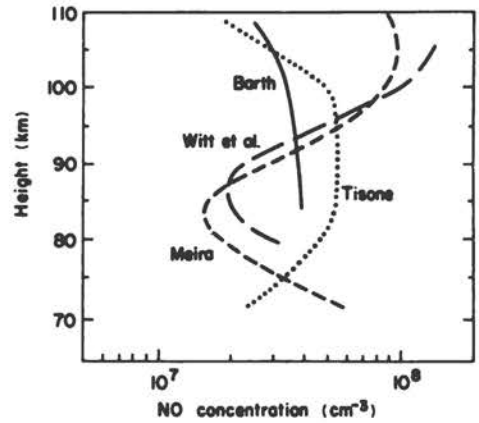


FIGURE 6.3 Altitude distributions of nitric oxide concentration deduced from rocket observations (Thomas, 1974b).

of the NO lies above the mesosphere. The NO concentration profiles derived from several rocket experiments are shown in Figure 6.3. It should be emphasized that Meira's measurement was conducted on January 31, 1969, at Wallops Island, Virginia, which was a winter day of anomalously high radio-wave absorption in the D region. Thus, it is possible that Meira's NO concentration profile is not typical for all seasons. In fact, in Section 6.7 some evidence is presented indicating that summer NO concentrations in the region may be an order of magnitude less than Meira's values. Furthermore, model computations of the NO number density profile suggest that enhanced vertical eddy transport could explain Meira's results. Thus the winter anomaly in the upper D region may be caused in part by an NO enhancement. This notion is plausible because the relatively long chemical lifetime of NO allows enhanced winter dynamical processes to transport more NO downward from the E region, where NO is produced, into the D region. Of course, horizontal transport of NO must be considered also. Nitric oxide concentrations at high latitudes are highly variable in space and time; the average values are 3 to 4 times greater at high latitudes than at midlatitudes, suggesting that energetic charged particles play an important role in its formation.

ATOMIC OXYGEN

This gas is another important constituent that plays vital roles in the neutral, positive-ion, and negative-ion chemistry. Atomic oxygen participates in several ion–molecule reactions, including those involving heavy cluster ions. The altitude distribution of atomic oxygen in the upper D region has been computed in several one-dimensional studies. Measurements of the atomic oxygen profile have been carried out by the silver-film sensor, neutral mass spectrometer, nitric oxide release, resonant scattering, absorption, and airglow emission techniques. Figure 6.4 illustrates height distributions of O (^3P), i.e.,

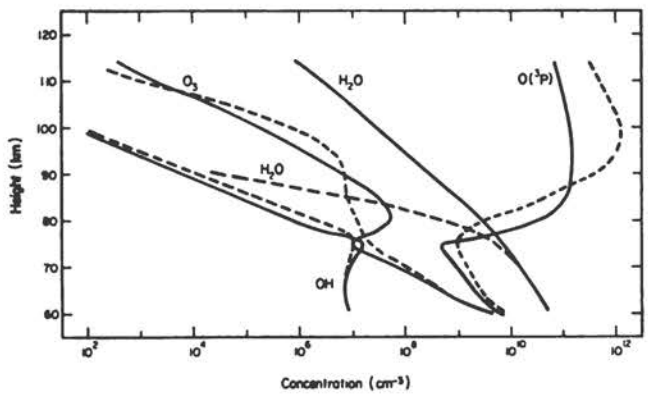


FIGURE 6.4 Height distributions of O, O_3, OH, and H_2O computed with eddy and molecular diffusion (solid curves) and without diffusion (dashed curves) (Bowman *et al.*, 1970).

ground-state atomic oxygen, computed in the presence of eddy and molecular diffusion (solid curves) and in the absence of diffusion (dashed curves). Atomic oxygen concentrations calculated from rocket measurements of absorption and resonance fluorescence during nighttime revealed a peak at 95 km of 5×10^{11} atoms cm^{-3} and an abrupt decrease to 10^9 atoms cm^{-3} at 80 km. Global maps of the atomic oxygen density near 97 km, deduced from the atomic oxygen green line (557.7 nm) nightglow observations carried out on OGO-6, show strong variations with latitude, longitude, universal time, and time of year but not much diurnal variation. The latitudinal distribution implies important meridional flows. From the observed horizontal scale of variations and their duration, a transport coefficient associated with horizontal motions was estimated to be 10^{11} cm sec^{-1}. Corresponding transport velocities of the order of 10 m sec^{-1} are implied, and this is the approximate velocity required for the meridional flows that balance atomic-oxygen production and loss.

OZONE AND METASTABLE MOLECULAR OXYGEN

Mesospheric ozone is important because ozone photolysis yields $O_2(^1\Delta)$, or metastable molecular oxygen, which can then be photoionized by 102.7- to 111.8-nm radiation to form O_2^+ ions and electrons in the daytime D region. This is a major source of O_2^+ ions in the middle and upper D region. $O_2(^1\Delta)$ is formed by photodissociation of ozone (200 to 300 nm); it loses its energy either by radiation at 1.27 μm or by deactivation collisions with ground-state O_2. Measurements of the 1.27-μm emission intensity have been used to derive the altitude profiles of $O_2(^1\Delta)$ concentration. O_3 also reacts with numerous neutral and ionized species in the mesosphere.

WATER VAPOR

This gas is an important minor neutral constituent in the D region for several reasons. Hydrogen-containing species in the mesosphere serve as indicators of the behavior of constituents such as CH_4 and H_2O from ground-level sources. It is now clear that H_2O transported vertically upward from the lower atmosphere by eddy diffusion serves as a source of exospheric atomic hydrogen, H, which is produced in the lower thermosphere through the photodissociation of H_2O. The chemistry and transport of H-containing species have been studied intensively in recent years. As the result of these studies, there is a better understanding of the role of vertical eddy transport and the relation between stratospheric odd hydrogen (mainly H and OH) and the exospheric escape flux. Figure 6.4 includes a calculated H_2O concentration profile. The H_2 number densities in the stratosphere and lower mesosphere have been measured. *In situ* measurements of the atomic hydrogen and hydroxyl radical concentrations have been carried out. Water vapor in the mesosphere is a subject under active investigation for several reasons: it is probably involved in the formation of noctilucent clouds at high latitudes in summer, it plays vital roles in the positive and negative cluster-ion chemistry, and it may be related to the light-scattering particulate layer in the mesosphere.

CARBON DIOXIDE

It should be mentioned that CO_2 probably participates in the cluster-ion chemical reactions of the D region. Neutral mass-spectrometric measurements of the CO_2 concentration indicate that it is probably well mixed up to the turbopause (near 105 km). Photodissociation is the main CO_2 loss mechanism above that altitude.

METALLIC SPECIES

The subject of neutral metallic species in the D region is a fruitful area for future research (Brown, 1973). It continues to provide challenging problems for theoretical and experimental aeronomers. Evidently, the neutral metal atoms from which metallic ions arise are deposited in the upper mesosphere and lower thermosphere by the ablation of meteorites entering the earth's atmosphere. However, there is limited evidence from recent lidar (laser radar) observations of sodium and potassium abundances that seawater is a possible source of some of these constituents. The seasonal variation of sodium concentration profiles has been obtained by use of lidar. The winter enhancement of sodium abundance is striking and possibly related to transport effects that are believed to be stronger and more variable than during summer conditions. Interesting associations have been discovered between sodium airglow emission intensities and meteorological parameters in the stratosphere. These relations are especially striking during stratospheric warmings. Ionized sodium atoms are probably involved in the positive cluster-ion chemistry of the D region. $Na^+(H_2O)$, $Na^+(H_2O)_2$, NaO^+, and $NaOH^+$ ions have been tentatively identified based on *in situ* measurements above Sardinia in December 1971. An interesting feature of the

atomic sodium layer is the small-scale height (~3 km) on the bottomside, and it should be noted that the steep gradients in sodium, electron, atomic oxygen, and water-cluster ion densities, all observed in the 80–90 km range, might possibly be related. It is certainly clear that much more attention must be focused on studies of the sources, sinks, chemistry, and transport of metallic neutrals in the future.

6.4 IONIZATION SOURCES

The photoionization of NO by solar Lyman-α radiation (121.6 nm) is a major source of NO^+ and free electrons in the daytime D region. The calculation of this ion-pair production rate requires information on the NO concentration, the ionization cross section, and the photon flux of solar Lyman-α radiation throughout the D region. As mentioned earlier, there is some evidence suggesting that the NO density profile deduced by Meira is not typical for the normal, undisturbed midlatitude D region. Thus, it is possible that NO^+ production rates computed using Meira's NO profile are unrealistically large and apply only to the anomalous winter situation. Figure 6.5 illustrates the relative roles of solar Lyman-α, solar uv, hard x rays (0.2–0.8 nm), and galactic cosmic rays (GCR) as sources of ionization in the daytime D region below ~90 km. Above 90 km, it is apparent that solar Lyman-β, solar EUV (extreme ultraviolet) and soft x rays become major sources of positive ions such as N_2^+ and O_2^+. Generally, the x rays and GCR produce N_2^+ and O_2^+ ions. The N_2^+ ions are quickly converted to O_2^+ ions, which in turn may react with NO, electrons, or O_2, depending on the relative concentrations and reaction rate constants, which vary with altitude.

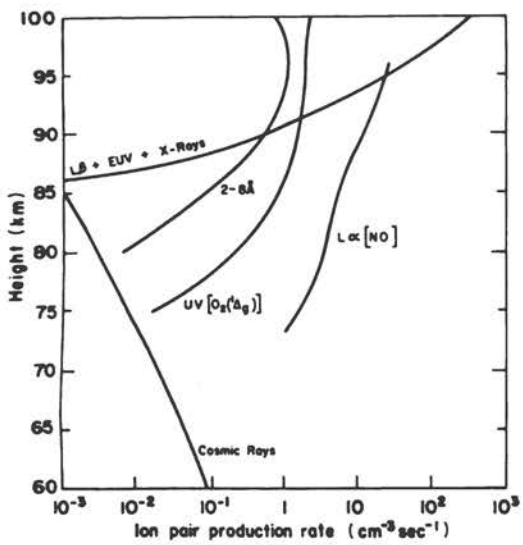


FIGURE 6.5 Ionization rates in the quiet daytime D region for solar minimum conditions and solar zenith angles near 60° (Thomas, 1974b). Lα [NO] represents the effect of Lyman-α radiation on nitric oxide, UV [$O_2(^1\Delta g)$] the effect of 102.7–111.8 nm radiation on metastable oxygen molecules, Lβ + EUV + x-rays the combined effect of Lyman-β, EUV, and soft x-ray radiation, and 2–8 Å (0.2–0.8 nm) the effect of hard x rays.

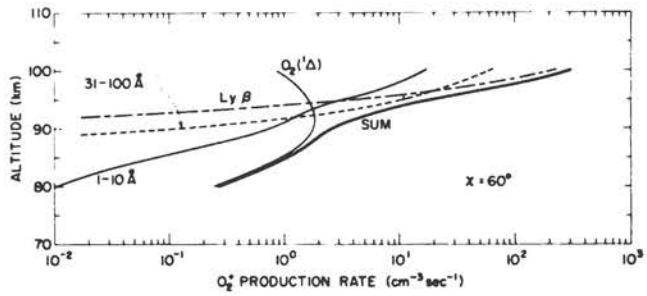


FIGURE 6.6 O_2^+ production rates for quiet-sun conditions and a solar zenith angle of 60°.

As mentioned in Section 6.3, $O_2(^1\Delta)$ is a major daytime source of O_2^+ ions below 90 km through the photoionization process

$$O_2(^1\Delta) + h\nu\,(102.7\text{–}111.8\text{ nm}) \rightarrow O_2^+ + e.$$

The initial estimates of the O_2^+ production rate by this process were revised downward when it was recognized that 102.7–111.8 nm solar radiation is absorbed by CO_2 in the upper mesosphere and lower thermosphere. Because of this development, it was thought that O_2^+ production from $O_2(^1\Delta)$ could not possibly compete with NO^+ production from NO. However, since it is now realized that the NO^+ production from Meira's NO densities may be 10 times larger than the normal NO^+ production rates, photoionization production rates of NO^+ and O_2^+ ions may be comparable in the normal, undisturbed, daytime D region. Figure 6.6 illustrates the relative importance of Lyman-β, hard and soft x rays, and 102.7–111.8 nm solar radiations in the photoionization of O_2 and $O_2(^1\Delta)$. These O_2^+ production-rate profiles are valid for a solar zenith angle of 60° and quiet sun conditions.

Rates of ion production from solar EUV and x rays for the ionization of N_2 and O_2 by galactic cosmic rays have been computed. It should be noted that solar x-ray fluxes may vary by several orders of magnitude from quiet to active sun conditions. Furthermore, these fluxes are highly variable during active sun years and may increase by several orders of magnitude during solar flares. An interesting point is the variation of ionization by GCR in the lower D region. The ion production rate by GCR is proportional to the neutral atmospheric density in the mesosphere and increases approximately exponentially with decrease in altitude. Clearly, GCR are a major source of N_2^+ and O_2^+ ions at the base of the D region. The ion-production rates vary with geomagnetic latitude such that polar rates are about 5 times greater than equatorial rates. Also, another feature is the solar-cycle variation of ion production from GCR. During years of high solar activity, the GCR flux in

the earth's atmosphere is smaller than it is during quiet sun years, with the result that ion-pair production in the lowest D region varies inversely with solar activity.

Nighttime sources of ionization include photoionization of NO by Lyman-α radiation arising from geocoronal scattering and precipitating energetic electrons. Although the precipitating electron fluxes are variable, it is quite possible that even the lowest fluxes represent an important nighttime ionization source in the undisturbed midlatitude D region.

Photoionization sources may become relatively unimportant at high latitudes where the average solar zenith angle is large. During the polar night, it is likely that ionization by energetic particles dominates all other possible sources. In the auroral zone, precipitating energetic particle fluxes are intense and quite variable, with the result that ionization production and densities are sporadic and enhanced above midlatitude levels. The resultant high levels of ionization are responsible for hf radio-wave absorption events sometimes called auroral zone absorption (AZA) events. During years of high solar activity, occasional solar proton events (SPE) cause polar-cap absorption (PCA) of radio waves, and intense fluxes of energetic protons (and other heavy particles) lead to extremely large ion-production rates in the polar D and sub-D regions. These particle events are associated with certain types of major solar flares and generally affect the D region for several days or weeks following flare onset. Peak ion-production rates during intense SPE's may be several orders of magnitude larger than normal daytime rates in the undisturbed D region at midlatitudes.

Another source of D-region ionization at midlatitudes arises from precipitating energetic electrons during and following certain geomagnetic storms. This phenomenon has been labeled the "geomagnetic storm aftereffect." It has been observed mainly during the winter months, probably because of the larger average value of the solar zenith angle during this season and the resulting low rates of photoionization.

6.5 POSITIVE IONS

Progress in understanding the positive-ion composition of the D region has been made through the combined results of *in situ* measurements, laboratory studies of ion–molecule reactions and rate constants, and numerical model studies. Positive-ion composition measurements using rocketborne ion mass spectrometers have been made at equatorial, middle, and high latitudes under a variety of geophysical conditions by Narcisi *et al.* (1971). Figure 6.7 shows the results obtained in Brazil. Note the water-cluster ions $H^+(H_2O)_n$ in the lower D region and the abrupt transition to NO^+ and O_2^+ above about 80 km. There is a continuing uncertainty about the ion mass-spectrometer sampling problem; it is not clear whether $H^+(H_2O)$ and $H^+(H_2O)_2$ ions are ambient D-region ions or the result of fragmentation of more massive hydrated

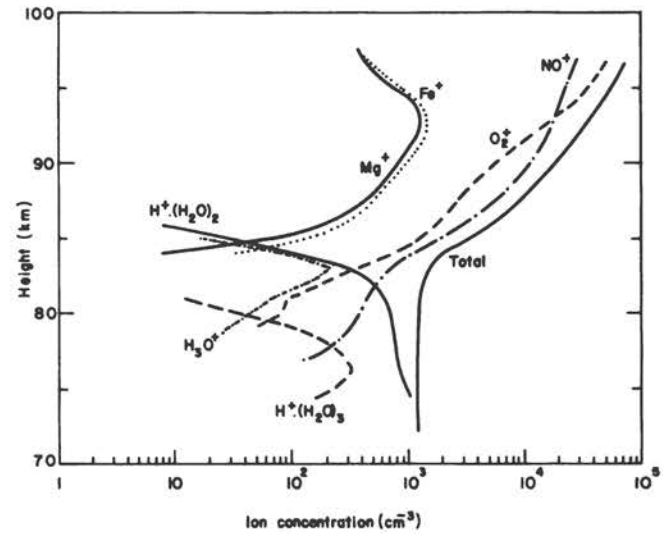


FIGURE 6.7 Rocketborne mass-spectrometer observations of the positive-ion composition for a solar zenith angle of 20° at Cassino, Brazil.

clusters such as $H^+(H_2O)_3$, $H^+(H_2O)_4$, and higher-order cluster ions. Rocket measurements using subsonic ion mass spectrometers and downleg wake sampling have revealed the presence of heavier cluster ions, thus lending support to the notion that $H^+(H_2O)$ and $H^+(H_2O)_2$ may be produced by fragmentation. Also, experiments in which the ion draw-in electric field is varied have demonstrated the importance of this parameter in producing cluster ion fragments. *In situ* measurements at high latitudes in summer have revealed hydrated protons $H^+(H_2O)_n$ up to order of hydration $n = 8$. It must be emphasized, however, that the degree of hydration is temperature-dependent, and *in situ* measurements only establish the lower limits of hydration. The higher-order hydrated and cluster ions have been found only in the cold mesopause region at high latitudes in summer, using improved sampling techniques. Also, between 85 and 100 km, cluster ions of the type $A^+ \cdot N_2$, $A^+ \cdot O_2$, $A^+ \cdot CO_2$, and $A^+ \cdot H_2O$ have been detected, where A^+ denotes an atomic, molecular, or cluster ion.

Metallic ions are observed in the upper D region. At times, layers of metallic ions such as Fe^+ and Mg^+ may reach peak concentrations comparable with the electron concentration around 95 km. Other observed metallic ions include Na^+, Al^+, K^+, Ca^+, Ni^+, Cr^+, Mn^+, and Si^+. $Na^+(H_2O)$, $Na^+(H_2O)_2$, NaO^+, and $NaOH^+$ have been tentatively identified; these ions may serve as sinks for Na^+ ions, whose fate is not understood.

Laboratory measurements have provided a wealth of information on positive and negative ion–molecule reactions relevant to D-region chemistry. Based on composition measurements and laboratory-determined reaction rate constants, it is possible to construct positive-ion chemical schemes leading from precursor ions to cluster

ions. The two possible precursor ions are O_2^+ and NO^+, which are produced in the D region by the photoionization of $O_2^+(1\Delta)$ and NO, respectively. The $O_2^+ \rightarrow H^+(H_2O)_n$ reaction chain begins with reactions

$$\begin{aligned} O_2^+ + 2O_2 &\rightarrow O_4^+ + O_2, \\ O_4^+ + H_2O &\rightarrow O_2^+(H_2O) + O_2, \\ O_2^+(H_2O) + H_2O &\rightarrow H_3O^+(OH) + O_2 \\ &\rightarrow H^+(H_2O) + OH + O_2, \\ H_3O^+(OH) + H_2O &\rightarrow H^+(H_2O)_2 + OH, \\ H^+(H_2O)_2 + H_2O + N_2 &\rightarrow H^+(H_2O)_3 + N_2. \end{aligned}$$

The above reaction sequence has been studied in the laboratory, and rate constants for the reactions have been measured. Another important reaction in the upper D region is

$$O_4^+ + O \rightarrow O_2^+ + O_3.$$

This reaction short circuits the water cluster–ion reaction chain at altitudes where the atomic oxygen concentration exceeds that of water vapor. This transition is expected to occur between 80 and 90 km, the altitude range in which water cluster–ion densities decrease abruptly and NO^+ and O_2^+ become dominant.

Unfortunately, the above reaction chain may not be applicable to the quiet, midlatitude D region where the major source of O_2^+ ions is the photoionization of $O_2(^1\Delta)$; this source is apparently inadequate to produce the concentrations of O_2^+ needed to explain the observed concentrations of water-cluster ions. Also, if Meira's NO profile is assumed, the NO^+ production rate is roughly 10 times larger than the O_2^+ rate. The above state of affairs led aeronomers to propose that NO^+ must be the precursor ion for the hydrated cluster ion reaction chain. Unfortunately, there are several problems with this chain. A fast path is needed between NO^+ and the hydrated protons, and a possibility is

$$\begin{aligned} NO^+ + N_2 + M &\rightarrow NO^+(N_2) + M, \\ NO^+(N_2) + CO_2 &\rightarrow NO^+(CO_2) + N_2, \\ NO^+(CO_2) + H_2O &\rightarrow NO^+(H_2O) + CO_2, \\ NO^+(H_2O) + N_2 + M &\rightarrow NO^+(H_2O)(N_2) + M, \\ NO^+(H_2O)(N_2) + CO_2 &\rightarrow NO^+(H_2O)(CO_2) + N_2, \\ NO^+(H_2O)(CO_2) + H_2O &\rightarrow NO^+(H_2O)_2 + CO_2. \end{aligned}$$

The above sequence continues until $NO^+(H_2O)_3$ is formed. This is followed by

$$NO^+(H_2O)_3 + H_2O \rightarrow H^+(H_2O)_3 + HNO_2.$$

The lowest-order hydrate produced by this chain is $H^+(H_2O)_3$. However, *in situ* observations indicate that $H^+(H_2O)$ and $H^+(H_2O)_2$ are present. It is not clear how these lower-order hydrated protons can be produced by the above scheme, enhancing the probability that these are fragments of the higher-order water-cluster ions. The $NO^+ + N_2 + M$ reaction is very difficult to measure in the laboratory because of the fast breakup of $NO^+(N_2)$ and its fast loss by switching with impurities such as NO, CO_2, or H_2O. During disturbed conditions at high latitudes, the O_2^+ production rate exceeds the NO^+ rate and the $O_2^+ \rightarrow H^+(H_2O)_n$ scheme is applicable.

Thus, two very basic questions in D-region positive-ion chemistry concern the NO^+ ion production rate and the conversion of NO^+ ions into the hydrated $H^+(H_2O)_n$ cluster ions. As mentioned earlier and discussed more fully in Section 6.7, NO concentrations are variable and probably are normally significantly less than those reported by Meira. Model studies have demonstrated that D-region NO densities are sensitive to the mesospheric eddy-diffusion rate, the NO predissociation rate, and the NO concentration in the E region, where NO is produced. Thus, variations of D-region NO are expected to occur with season, latitude, solar and geomagnetic activity, and meteorological activity. As a consequence, the positive-ion and electron abundances exhibit associated variabilities.

Because O_2^+ ions are produced mainly by photoionization of $O_2(^1\Delta)$, it is essential to have much more information on the $O_2(^1\Delta)$ density profile and its temporal and spatial variations. Hence, it seems apparent that O_3, the major source of $O_2(^1\Delta)$, has the potential to influence markedly the O_2^+ production rates and thus the O_2^+ and cluster-ion populations. Because CO_2 is responsible for most of the absorption of the solar uv radiation that can photoionize $O_2(^1\Delta)$ to produce O_2^+ ions, it is necessary to know its altitude distribution in and above the D region.

6.6 NEGATIVE IONS

Most of our knowledge about negative ions has evolved through a relatively small number of *in situ* measurements and prodigious laboratory studies of negative-ion/molecule reactions and measurements of rate constants. Negative-ion reaction schemes have been proposed on the basis of the laboratory measurements, but there has been no substantial test of the schemes. Several negative-ion model studies based on mass-spectrometric measurements have been carried out, but these are preliminary, and they produce somewhat conflicting or inconclusive results. Figure 6.8 depicts the negative-ion results from *in situ* measurements made during a daytime AZA (auroral zone absorption) event at Fort Churchill, Canada. Heavy negative ions were found in largest concentrations below 92 km and consisted of species tentatively identified as $NO_3^-(H_2O)_n$ with $n = 0$–8, with some possible admixture of $CO_3^-(H_2O)_n$ and CO_4^-. Ions with mass numbers of 56 ± 1 and 121 ± 2 AMU were also found, but these have not been identified. Species as heavy as 203 ± 4 AMU and even more massive ions were indicated, suggesting that conglomerates may play an important role in the negative-ion chemistry. It should be noted that the ions O_2^-, CO_3^-, NO_3^-, and CO_4^- were expected on the basis of current negative-ion reaction schemes.

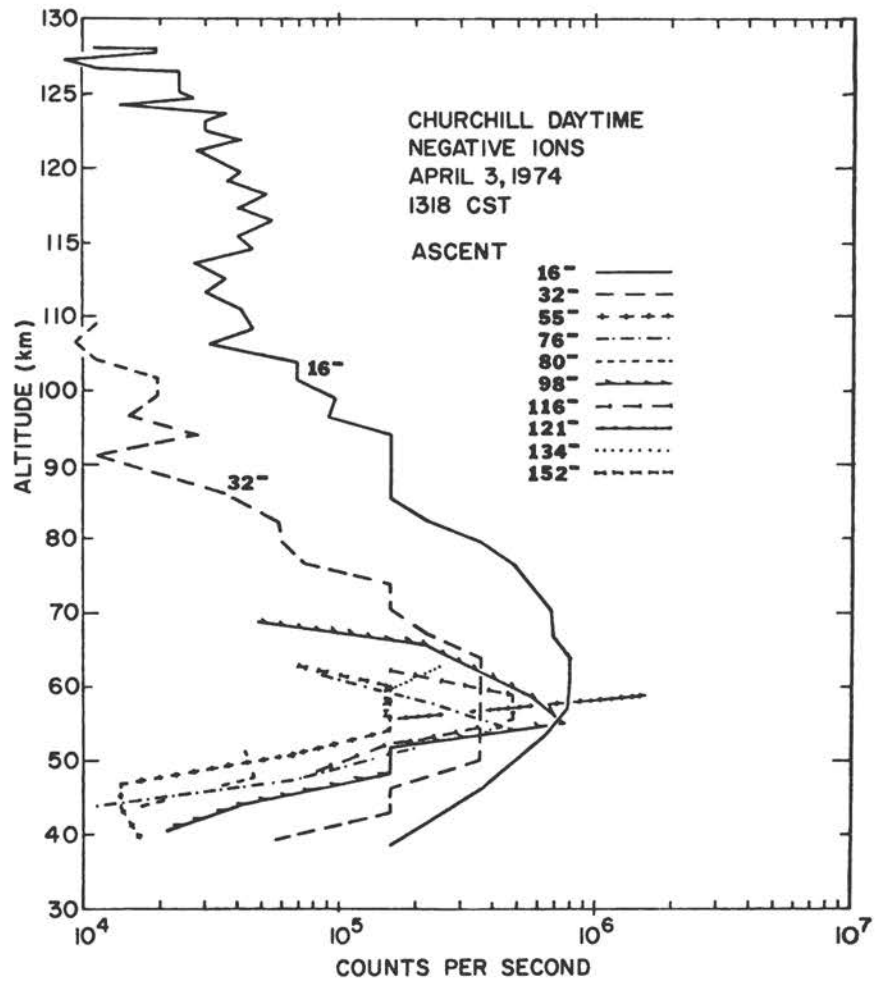


FIGURE 6.8 Results of D- and E-region negative-ion composition measurements obtained during a daytime auroral zone absorption event (Narcisi, 1975).

A relatively large number of negative-ion reactions relevant to the D region have been studied in the laboratory. According to these reactions, O_2^- and O^- are the precursors or primary ions. Their formation is followed by a sequence of ion–molecule reactions leading to ions such as CO_4^-, CO_3^-, NO_2^-, and the terminal or stable ion NO_3^-. Model computations of negative-ion concentration profiles depend markedly on accurate determinations of minor neutral constituent concentrations.

The major primary negative ion O_2^- is formed in a three-body attachment process:

$$O_2 + e + M \rightarrow O_2^- + M.$$

The O_2^- ions in the lower D region, where the O/O_3 concentration ratio is small, can either charge transfer with O_3 to form O_3^- or associate with O_2 to form O_4^-. Both these ions react rapidly with CO_2, forming CO_3^- and CO_4^- ions. Nitric oxide then converts the CO_3^- and CO_4^- ions to NO_2^- and $(NO_3^-)^*$, where $(NO_3^-)^*$ is a less stable form of NO_3^-. Eventually the stable, terminal ion NO_3^- is produced, and this probably hydrates to yield $NO_3^-(H_2O)_n$ ions, which have been observed by Narcisi *et al.* (1971). The effects of hydration on the negative-ion reaction scheme have been considered, and laboratory measurements suggest that hydration probably does not significantly affect the reaction scheme.

The O_2^- ions observed above 80 km are inconsistent with current negative-ion reaction schemes. These ions were not expected in the upper D region, where the O/O_3 concentration ratio exceeds unity, because of the fast associative-detachment reaction

$$O_2^- + O \rightarrow O_3 + e.$$

Narcisi has proposed that the rate constant for this reaction must be lowered by two orders of magnitude in order to match the computed and measured concentrations of O_2^- ions in the upper D region.

Generally, the *in situ* measurements reveal massive negatively charged ions in the upper D region where these were not expected. Thus, it is necessary to reexamine current notions about the importance of negative ions in the daytime D region. Originally, it was accepted

that the negative ion-to-electron concentration ratio was less than unity above ~70 km in the daytime, and the main electron loss process above that altitude was probably electron–ion recombination. Now, it appears that serious thought must be given to the possibility that attachment to heavy particles (molecular conglomerates or neutral clusters) is a possible electron-loss process in the middle D region. Above the altitude of the electron density ledge (~82 km), it is clear that dissociative recombination through NO^+ and O_2^+ is the dominant electron-loss process.

6.7 WINTER D REGION

During the winter months in middle latitudes, there are marked variations in D-region radio-wave absorption and electron densities. The well-known D-region winter anomaly refers to anomalously high values of mf or hf radio-wave absorption observed on certain days or groups of days in winter. It is now generally agreed that the winter variability may be ascribed to at least two quite different physical mechanisms. One possibility is D-region meteorology or dynamical processes in the neutral mesosphere and lower thermosphere. A second possibility is precipitation of energetic particles from the magnetosphere during and following certain geomagnetic storms. In this section, emphasis is placed on meteorological disturbances in the midlatitude D region. This topic is sometimes labeled forcing from below, stratosphere–mesosphere–ionosphere coupling, or interaction between the neutral and ionized atmospheres. Essentially, it is now believed that atmospheric waves may propagate from the lower atmosphere into the thermosphere when the wind and temperature fields of the winter stratosphere and mesophere possess the proper magnitudes, directions, and spatial variations.

Figure 6.9 illustrates the winter variability of electron densities. There is some evidence that two different types of variability or anomaly may occur in the winter D-region electron-density profile. One anomaly occurs above the altitude of the electron density ledge near 82 km, the other is below it. The current notions are that enhanced electron densities above the ledge produce anomalously high mf and hf radio-wave absorption and that electron density variabilities below the ledge (observed mainly by the partial-reflection radio method) are not closely associated with those above.

Several dynamical mechanisms have been suggested as causes for the anomalous winter behavior. Both gravity and planetary waves propagating upward from the lower atmosphere may influence the transport of minor neutral species and thus alter their vertical and horizontal distributions. Turbulence-induced transport is a possibility, and the vertical eddy-diffusion coefficient profile is probably more variable in winter than in summer. Vertical and horizontal motions associated with planetary-scale disturbances in the neutral mesosphere and lower thermosphere may substantially alter the vertical and horizontal distributions of long-lived minor neutral species such as nitric oxide. Redistribution of these species would then alter the spatial distribution of D-region ions and electrons.

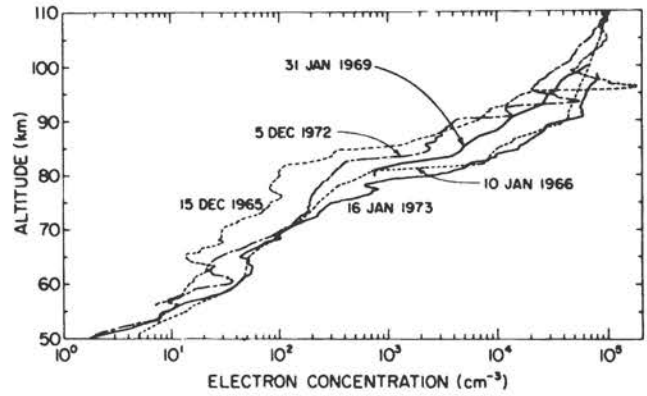


FIGURE 6.9 Electron-density profiles measured by rockets, illustrating winter variabilities in upper and lower D regions.

In situ measurements of the positive-ion composition were conducted on an anomalous winter day over Sardinia in December 1971 and over Wallops Island in January 1969. The results from both experiments are presented in Figure 6.10, which shows altitude profiles of electron density and the NO^+/O_2^+ concentration ratios. Also shown in Figure 6.10 are profiles derived from summer measurements in Spain. The larger concentration ratios of NO^+/O_2^+ obtained from the winter measurements are striking and suggest that NO may be enhanced on anomalous winter days in the D region. In order to esti-

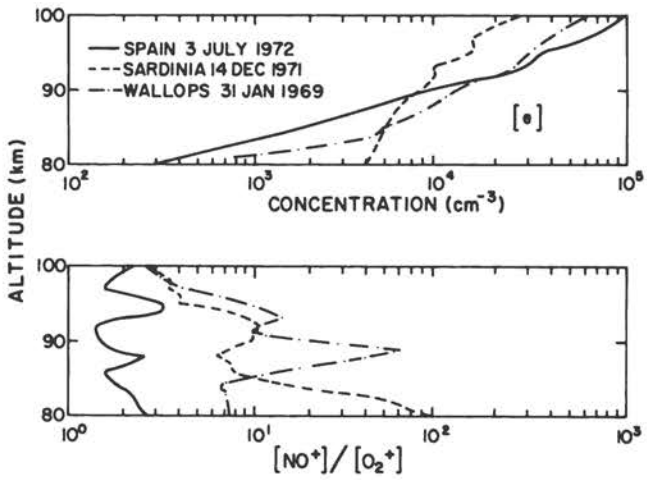


FIGURE 6.10 Electron density and NO^+/O_2^+ concentration profiles for a summer day and two winter days at middle latitudes.

mate NO densities present in the upper D region, a simple three-reaction ion-chemical scheme may be used:

$$O_2^+ + e \rightarrow O + O,$$
$$NO^+ + e \rightarrow N + O,$$
$$O_2^+ + NO \rightarrow NO^+ + O_2.$$

A quadratic expression for the NO concentration may be derived from the NO^+ and O_2^+ continuity equations. Computed values of the NO concentration are shown in Figures 6.11 and 6.12, corresponding to summer and winter conditions in Spain and at Wallops Island, respectively. Shown for comparison is Meira's NO profile for January 31, 1969, at Wallops Island. The computed summer NO densities are substantially less than Meira's at altitudes below 90 km, and this lends support to the notion that Meira's NO profile is not a typical one for the quiet, undisturbed region. The computed winter NO densities are comparable with Meira's values above 90 km. Below this height, the computed NO profile tends to follow the lower limits of Meira's error bars. Thus, there is some evidence suggesting that NO concentrations may be enhanced on anomalous winter days.

There are several other possible explanations of the ionization enhancements on certain winter days. Ozone may be enhanced at certain times. This would imply that $O_2(^1\Delta)$ densities would be greater, which in turn would lead to larger O_2^+ and electron production rates. Several ion–molecule reaction-rate constants are temperature-sensitive, and thus mesospheric temperature fluctuations induced by dynamical processes may alter the relative ion composition and concentration. There is limited evidence from ground-based meteor-radar and partial-reflection

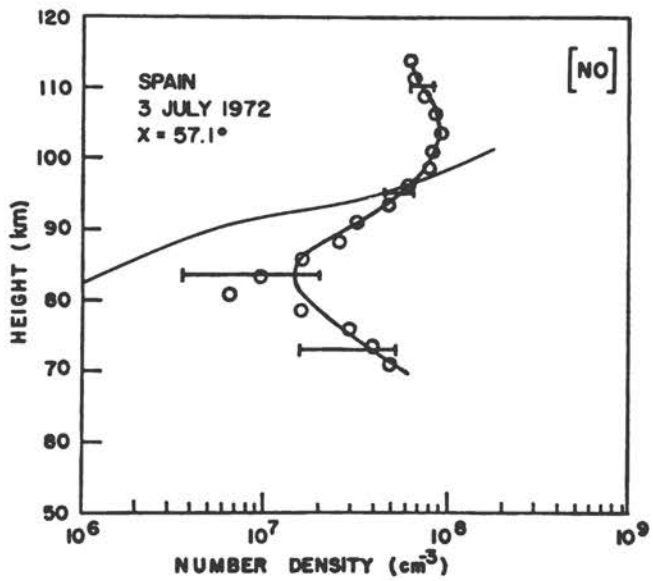


FIGURE 6.11 Nitric oxide concentration profile deduced from summer measurements of NO^+ and O_2^+, compared with Meira's profile, for a solar zenith angle of 57.1°.

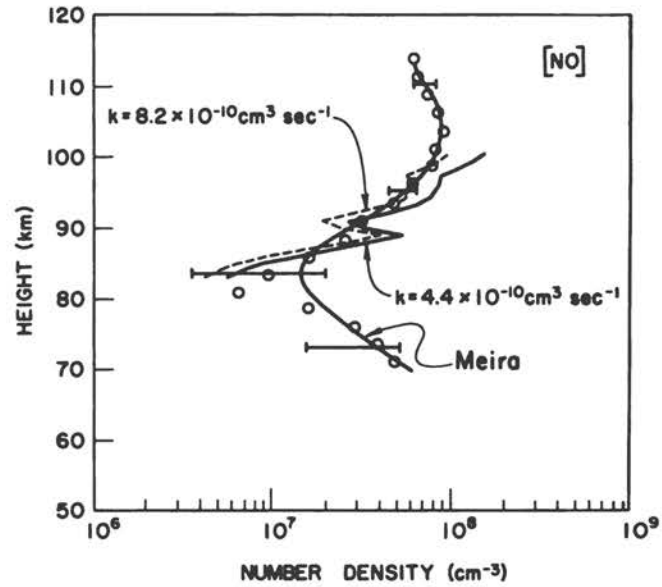


FIGURE 6.12 Nitric oxide concentration profiles deduced from winter measurements of NO^+ and O_2^+ over Wallops Island, compared with Meira's profile. k is the rate constant for the charge-transfer reaction $O_2^+ + NO \rightarrow NO^+ + O_2$.

observations that enhanced electron densities in the D region below 80 km tend to be associated with equatorward motion and vice versa. This result is consistent with the notion that planetary-scale disturbances may transport long-lived species such as NO from the auroral zone to middle latitudes. It should be noted that NO is produced in the E region and is mixed downward into the D region by means of eddy diffusion. Therefore, a combination of vertical and horizontal transport of NO might be expected at certain times in winter.

6.8 DISTURBED D REGION

Section 6.7 dealt with the disturbed winter D region in middle latitudes. In this section, other disturbances are considered; these include disturbances induced by solar flares, solar eclipses, solar proton events, auroral zone absorption events, and geomagnetic-storm aftereffects. Generally, this topic may be labeled extraterrestrial influences on the D region or forcing from above. Solar flares are accompanied by enhancements of soft and hard x-ray fluxes. These enhancements begin quite abruptly and then decay with various time constants depending on the class of solar flare. X-ray flares may have durations of several minutes to tens of minutes. Exceptionally large flares may be accompanied by x-ray fluxes that are several orders of magnitude greater than normal, undisturbed levels. Obviously, flares increase the ion production rates in the D region; in particular, the N_2^+ and O_2^+ rates are enhanced with the N_2^+ rapidly converted to O_2^+, and the electron densities are elevated throughout the D

region. During a flare, the O_2^+ production rate may greatly exceed the NO^+ production rate, with the result that positive hydrated cluster ions are probably produced through the $O_2^+ \rightarrow O_4^+ \rightarrow$ hydrates reaction sequence described in Section 6.5. Because of the fast reaction

$$O_4^+ + O \rightarrow O_2^+ + O_3,$$

there is a short-circuiting of the $O_2^+ \rightarrow$ hydrates scheme at altitudes where the O concentration exceeds that of H_2O. Thus, water-cluster ions $H^+(H_2O)_n$ are not expected to occur above the steep bottomside of the atomic oxygen distribution in the upper D region. Because the $O_2^+ \rightarrow$ hydrates sequence is understood better than the $NO^+ \rightarrow$ hydrates sequence, it is ironic indeed that the disturbed D-region positive-ion composition is explicable whereas that of the quiet D region is not. During disturbed conditions, hydrated protons $H^+(H_2O)_n$ exhibit a cutoff (or steep gradient) near the altitude of the bottomside of the atomic oxygen distribution. The altitude of this cutoff during disturbed conditions occurs at an altitude several kilometers lower than that for the undisturbed situation. This causes the electron density ledge to move downward in altitude because the effective electron–ion recombination coefficient for $H^+(H_2O)_n$ ions is at least an order of magnitude greater than the dissociative recombination coefficients for simple molecular ions such as NO^+ and O_2^+.

The other sources of D-region disturbances (solar proton events, auroral-zone particle events, and geomagnetic-storm aftereffects) tend to cause D-region modifications similar to those produced by solar-flare x-ray events. That is, the water-cluster-ion ledge is depressed in altitude, and the positive-ion composition and concentrations are explicable in terms of the well-understood $O_2^+ \rightarrow O_4^+ \rightarrow$ hydrates reaction scheme. A major difference between solar proton effects and flare effects is that during proton events large ionization densities can also exist at night. Also, during proton events it appears that the normally predominant heavy negative cluster ions are replaced by simpler ions such as O_2^- ions during the daytime and O^- ions during the night at the lower altitudes. Thus, negative ions, like the positive ions, tend to be simpler nonclustered forms during flare and energetic particle events. Energetic electron precipitation other than in auroras may influence the D region, the magnetic-storm aftereffect being a good example. Weak electron precipitation events are most significant at night because solar ionizing radiations are absent except for the Lyman-α radiation scattered from the geocorona. It has been reported that precipitating electrons at South Uist (Hebrides) are the dominant nighttime D-region ionization source 15 ± 11 percent of the time around solar minimum and about 35 ± 20 percent of the time during solar maximum.

The cause of magnetic-storm aftereffects in the middle-latitude D region is related to the enhancement of fluxes of outer-zone radiation-belt electrons during the main phase of geomagnetic storms. During the storm recovery phase the electrons diffuse radially toward the earth, and precipitation loss is a major source of D-region ionization. This precipitation is latitude-dependent and is probably unimportant below 45° magnetic latitude. Also, ion production rates usually maximize a few days after the storm main phase, and enhanced ionization persists for approximately a week following the storm, consistent with electron precipitation lifetimes.

Another important D-region disturbance is caused by a total eclipse of the sun. This has a strong impact on electron densities below about 85 km, as shown in Figure 6.13. Mechtly *et al.* (1972b) presented electron density profiles measured during eclipses of March 1970 and November 1966. During totality, the electron densities decayed markedly below the ledge. Mechtly *et al.* characterized the loss of electrons during totality by a recombinationlike loss coefficient, A, and by an attachmentlike loss coefficient, B. The calculated value of A is roughly 1×10^4 cm^3 sec^{-1}, and B is practically constant at 1×10^{-2} sec^{-1} below 87 km. Because the recombination coefficients of electrons with $H^+(H_2O)_n$ ions probably do not exceed 1×10^{-5} cm^3 sec^{-1}, it is likely that negative ions will have to be considered in the effective recombination coefficient ψ, given by

$$\psi = (1 + \lambda)(\alpha_d + \lambda\alpha_i),$$

where λ is the negative ion-to-electron density ratio and α_d and α_i are the ion–electron and ion–ion recombination coefficients, respectively. It has been suggested that positive-ion and electron densities are not equal in the normal daytime D region between 70 and 80 km and that λ is almost 10 in this altitude interval.

Positive-ion composition was measured by Narcisi *et al.* during the November 1966 solar eclipse in Brazil. The D-region results yielded evidence for the existence of a fast process for the conversion of NO^+ ions to water-cluster ions. Also, the decrease in water-cluster ion con-

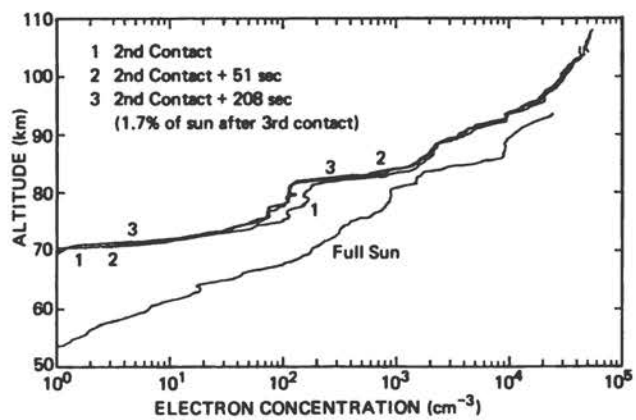


FIGURE 6.13 Rocket profiles of electron concentration from the solar eclipse of March 7, 1970, at Wallops Island (Mechtly *et al.*, 1972b).

centration at totality was less than a factor of 4 near 80 km. Negative-ion composition measurements were obtained over Wallops Island between 70 and 111 km near totality during the March 1970 eclipse. Heavy negative cluster ions were predominant below 92 km. Between 90 and 98 km, there were relatively large concentrations of ions such as O^-, O_2^-, NO_2^-, and NO_3^-.

It is becoming clear that, on the basis of the results described above, it is necessary to re-examine the widely held view that negative ions are not important in the daytime D region above ~70 km. Indeed, there is now evidence that negative ions may play an important role in the electron loss process below the ledge in the daytime electron-density profile.

6.9 PARTICULATES

Noctilucent clouds have been observed for many decades at high latitudes in summer. It is generally agreed that these are composed of particulates or aerosol particles in the cold mesopause region. An aerosol may be considered to be a grouping of a few molecules, a small solid particle, or a liquid droplet suspended in the atmosphere. The formation and physical properties of noctilucent cloud particles remain an unsolved problem in mesospheric research. There has been much speculation about the phenomena associated with their formation, and the mechanisms are still uncertain (Castleman, 1974). Hypotheses on formation vary from those that suggest extraterrestrial dust sources to mechanisms involving nucleation about ions or molecules. The nucleation mechanism, with a water-vapor supersaturation condition at mesopause heights, is perhaps more likely; and Castleman believes the most likely mechanisms are either nucleation on pre-existing particles or heteromolecular nucleation about ions. Apparently noctilucent cloud particles evaporate rapidly at lower altitudes, where temperatures are higher, suggesting that they consist of water or ice. Nucleation may arise about ions, and ion hydration provides a mechanism for the initial step in the formation of visible particle (aerosol) layers. At low mesopause temperatures (~160 K) and an H_2O concentration of about 10^9 cm^{-3}, sufficient supersaturation exists for this nucleation process to be operative. Because $H^+(H_2O)_n$ ions have been detected consistently at mesopause heights, nucleation about these must be considered as a possible mechanism for noctilucent cloud formation. Another possibility is that negative ions result in nucleation.

Donahue *et al.* (1972) have described observations of circumpolar particulate layers near the summer mesopause. Observations with a horizon-scanning airglow photometer on OGO-6 revealed the presence of a dense scattering layer near the mesopause over the geographic pole in summer. The temporal and latitudinal variations in the radiance and altitude of the scattering layer were determined. The average altitude was 84.3 km, with higher values on the nightside than on the dayside of the polar cap. It was surmised that this particulate layer is probably a poleward extension of noctilucent clouds, i.e., noctilucent clouds are weak sporadic manifestations of this persistent polar layer near its low-latitude edge. Figure 6.14 shows the variation of slant emission rate as a function of the photometer line-of-sight altitude at closest approach. The scattering layer effect near 80 km is clearly discernible. At present, the relationship between the light-scattering layer and the circulation of the summer polar atmosphere is unknown. Perhaps there is a relationship between the disappearance of the particulate layer and the dynamics of the neutral stratosphere and mesosphere, particularly during the reversal of the polar vortex between summer and winter.

Reid considered the physical properties that mesospheric ice clouds ought to possess and developed a simple model of an ice cloud at the mesopause in which eddy diffusion was assumed to transport H_2O vapor vertically upward. He also assumed that the ice-cloud particles grow from initial sizes that are comparable with ionic dimensions, i.e., that positive water cluster ions $H^+(H_2O)_n$ might serve as nuclei for particle growth. He found that needle-shaped and disk-shaped particles can grow readily to linear dimensions of the order of 100 nm. Clearly, there is a need for further work on the nucleation process

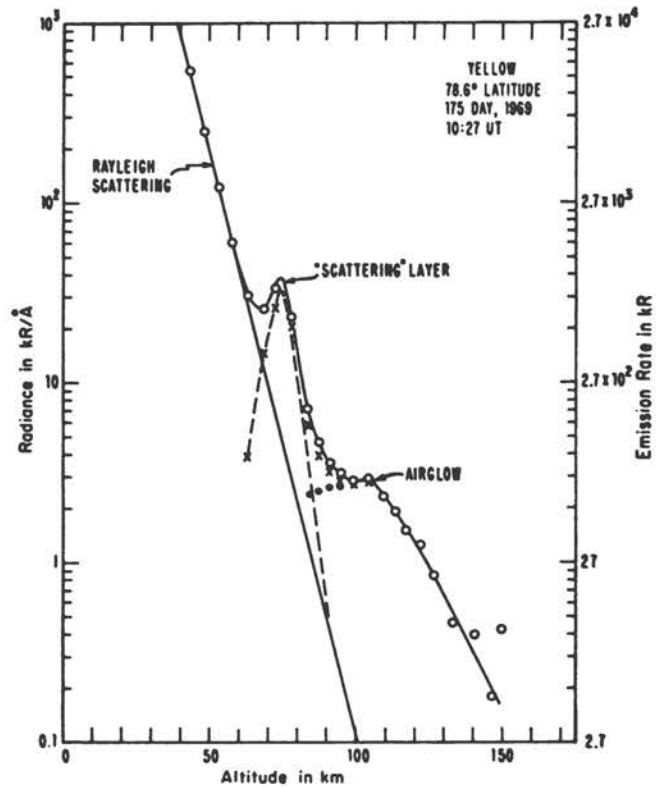


FIGURE 6.14 Slant emission rates observed by OGO-6 airglow photometer at 589 nm above the horizon as a function of the altitude of closest approach of the line of sight for 78.6° N latitude (Donahue *et al.*, 1972).

in the mesopause region and its relation to water-cluster ions. It is essential to clarify the proposed transition between $H^+(H_2O)_n$ ion clusters and the heavy particles or conglomerates containing thousands of molecules.

There is evidence for the existence of heavy ions in the lower D region. Two groups of positive ions have been found in the 58- to 72-km range using rocketborne instrumentation. The heavy group was about 4 times less mobile than the extrapolated mobility of light ions at ground level, and the light group was about 8 times more mobile. The mobility of the negative ions was about equal to the mobility of the light group of positive ions. These results implied that the negative ions and light positive ions detected between 58 and 72 km were probably molecular ions. Rose and Widdel proposed that the heavy, positively charged group were charged aerosol particles. Thus, if the light ion mass was about 30 AMU, it would suggest a heavy positive ion mass of the order of 1000 AMU.

There is additional evidence for submicrometer particulates in the mesosphere. It has been proposed that ice crystals may occur throughout the mesosphere over a wide range of latitudes during all seasons. These authors argued that 10^3 to 10^4 ice crystals per cm^3 of order 10 nm in diameter can dominate the ionization loss processes in the mesosphere. If so, then this would be an alternative explanation for the large values (1×10^{-4} cm^3 sec^{-1}) of effective recombination coefficient below the electron-density ledge, deduced from eclipse measurements. Their measurements also indicate that positive ions above 50 km occur in two relatively distinct mobility groups, separated by an order of magnitude in mobility, and with a preponderance of the ions in the low-mobility group. In accord with the results of Rose and Widdel, it is believed the heavier ions have a mass of several thousand AMU, assuming the light or high mobility ions to be about 30 AMU. These heavier ions may be formed by the attachment of light ions to particulates having masses of the order of 1000 AMU.

Heavy negative-ion clusters have been observed in the upper D region by Narcisi *et al.* (1971), and this is further evidence that negatively charged particulates must be seriously considered in future studies. Also, Narcisi argued that significant negative-ion concentrations must be considered above 80 km. This is additional evidence that λ, the negative ion-to-electron density ratio, is probably greater than unity in the daytime D region above ~70 km. Furthermore, negative-ion composition and electron-density measurements conducted during solar eclipses by Narcisi *et al.* have indicated that electron attachment processes considerably faster than

$$O_2 + O_2 + e \rightarrow O_2^- + O_2$$

are necessary to explain the extremely large electron loss rate observed during totality in the 82–87 km region. Perhaps the heavy negative-ion clusters observed in this altitude region are responsible for the large rate of electron loss. It is possible that there are fast two-body electron attachment reactions with neutral clusters (molecular conglomerates) that may be present in the cold mesopause region. Thus, it is becoming evident that particulates may play significant roles in the D-region positive- and negative-ion chemistry; they may be responsible for the large electron loss rates that seem to be present in the daytime D region and the heavy positive and negative ions observed there.

6.10 SOME OUTSTANDING PROBLEMS

In this concluding section, several problems of the D region are mentioned briefly. The solution of one or more of these interrelated problems would have a significant impact on the status of D-region research.

Nitric oxide concentration in the D region has not been accurately determined, either by theoretical or observational techniques. Although the chemistry of odd nitrogen [NO, $N(^2D)$, $N(^4S)$, etc.] in the E region where NO is produced is becoming clearer than before, there remain major uncertainties such as the altitude variation of the eddy-diffusion coefficient and the role of horizontal transport. Measurement techniques, in addition to the rocketborne NO gamma-band resonance fluorescence method, should be developed.

The origin, formation, and distribution of metallic ion and neutral species in the D region are not well understood. Neither the production rates nor loss rates of the metal ions are known, and no detailed quantitative analyses of the metal ion chemistry have been carried out. In particular, the loss mechanisms are not well known. An interesting possibility is the hydration of metallic ions, such as $Na^+(H_2O)_n$. The atomic sodium layer in the upper D region offers a wide variety of challenging problems, including the identification of sources and sinks and the role of transport processes.

The outstanding positive-ion problem is the identification of the ion-chemical scheme for rapid conversion of NO^+ ions to the $H^+(H_2O)_n$ series of hydrated ions. This problem arose when it was concluded that O_2^+ production from $O_2(^1\Delta)$ photoionization was relatively slow and that, therefore, O_2^+ could not compete with NO^+ as the precursor for water-cluster ion formation. However, this matter should be re-examined because there is now evidence that the NO^+ production rates were overestimated.

There is no paucity of negative-ion problems. The conflicting results obtained for the negative-ion composition must be resolved. Mass-spectrometer sampling methods must be developed and improved to avoid heavy-ion fragmentation; this applies to both positive and negative ions. The observations of negative ions in the upper daytime D region require an explanation, as well as do the indications that the negative ion-to-electron density ratio between 70 and 80 km in the daytime may be equal to about 10.

The role of particulates in the D region is attracting

more attention because of the possibility that they participate in both positive- and negative-ion chemistry. Perhaps they are responsible for the inexplicably fast electron loss rates deduced from eclipse measurements. Furthermore, the subject of ion-induced nucleation, involving the possible growth of $H^+(H_2O)_n$ ions to very large dimensions, is a potentially fruitful area for future research.

The interpretation of the winter variabilities of the upper and lower D regions is a challenging topic that offers a broad range of problems in upper atmospheric dynamics and the chemistry of neutral and ionized species.

It is appropriate to conclude this chapter with the reminder that the high-latitude disturbed D region is well understood in comparison with the quiet D region in middle latitudes. This ironic situation exists because the O_2^+ production rate is substantially enhanced above the NO^+ rate during energetic particle precipitation events, and thus the well-known $O_2^+ \rightarrow O_4^+ \rightarrow H^+(H_2O)_n$ ion reaction chain is operative. As a result, there is a remarkable agreement between model calculations and measurements of the ionized species. Therefore, it is reasonable to suggest that future research on the D region should concentrate on the problems of its quiet, mid-latitude part.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by NASA and NSF under Grants 14-005-181 and DES 73-06485-A02, respectively.

REFERENCES

Anderson, J. G., and T. M. Donahue (1975). The neutral composition of the stratosphere and mesophere, *J. Atmos. Terr. Phys. 37*, 865.

Bowman, M. R., L. Thomas, and J. E. Geisler (1970). The effect of diffusion processes on the hydrogen and oxygen constituents in the mesosphere and lower thermosphere, *J. Atmos. Terr. Phys. 32*, 1661.

Brown, T. L. (1973). The chemistry of metallic elements in the ionosphere and mesosphere, *Chem. Rev. 73*, 645.

Castleman, A. W., Jr. (1974). Nucleation processes and aerosol chemistry, *Space Sci. Rev. 15*, 547.

Donahue, T. M., B. Guenther, and J. E. Blamont (1972). Noctilucent clouds in daytime: Circumpolar particulate layers near the summer mesopause, *J. Atmos. Sci 29*, 1205.

Mechtly, E. A., and L. G. Smith (1968). Growth of the D region at sunrise, *J. Atmos. Terr. Phys. 30*, 363.

Mechtly, E. A., S. A. Bowhill, and L. G. Smith (1972a). Changes of lower ionosphere electron concentrations with solar activity, *J. Atmos. Terr. Phys. 34*, 1899.

Mechtly, E. A., C. F. Sechrist, Jr., and L. G. Smith (1972b). Electron loss coefficients for the D region of the ionosphere from rocket measurements during the eclipses of March 1970 and November 1966, *J. Atmos. Terr. Phys. 34*, 641.

Narcisi, R. S. (1975). Negative-ion composition of the D and E regions, presented at IAGA Symposium on Solar Fluxes and Photochemistry, Grenoble, France.

Narcisi, R. S., A. D. Bailey, L. Della Lucca, C. Sherman, and D. M. Thomas (1971). Mass spectrometric measurements of negative ions in the D and lower E regions, *J. Atmos. Terr. Phys. 33*, 1147.

Sechrist, C. F., Jr. (1974). Comparisons of techniques for measurement of D-region electron densities, *Radio Sci. 9*, 137.

Thomas, L. (1974a). The temporal and geographical variations of D region electron concentrations, in *Methods of Measurements and Results of Lower Ionosphere Structure*, K. Rawer, ed., Akademie-Verlag, Berlin, pp. 153–167.

Thomas, L. (1974b). Recent developments and outstanding problems in the theory of the D region, *Radio Sci. 9*, 121.

Turbulence in the Lower Thermosphere

7

ROBERT G. ROPER
Georgia Institute of Technology

7.1 PROLOGUE

A static, or motionless, atmosphere can be adequately described on a macroscopic scale by three parameters: pressure, temperature, and composition. If the atmosphere were macroscopically motionless for a sufficiently long time (its molecules in purely random motion), one would find a decrease in mean molecular weight with altitude, the lighter species having separated from the heavier species by the process of molecular diffusion. Such a process is not observed in the earth's atmosphere until one reaches an altitude of from 100 to 110 km; up to this altitude the mean molecular weight remains approximately constant. This comes about because the earth's atmosphere is not motionless on a macroscopic scale but is continually mixed by atmospheric turbulence and large-scale circulation. The beginning of diffusive separation at 100–110 km does not mean that there are no winds to cause turbulence and mixing above these altitudes but rather that at such low densities molecular diffusion rates are much greater than the wind-induced mixing rates. This chapter concentrates on the region known as the lower thermosphere, between the mesopause at approximately 80 km, where the atmosphere is always mixed, and the altitude of 130 km, where diffusive separation is always observed.

Our knowledge of mixing processes in the upper atmosphere has been accumulated mostly in the years since World War II. While the lower thermosphere has been probed for decades using radio techniques, detailed knowledge of its structure has only come with the use of rocket soundings.

Before considering the techniques of measurement of mixing processes in the upper atmosphere and the interpretation of those measurements, a definition of "mixing process" is appropriate. On a global scale, the atmosphere can be said to be mixed by the transport of constituents from one location to another by large-scale wind systems. This mixing is important. Here, however, we will concern ourselves with more localized mixing, usually brought about in the free atmosphere by shears in the wind system and called atmospheric turbulence. A turbulent atmo-

sphere is mixed at rates often hundreds or more times faster than its molecules can diffuse by means of their thermal motions, and thus turbulence can be very effective in maintaining homogeneity in an atmosphere composed of several molecular species; hence the term "homosphere" is often applied to describe the homogeneous atmosphere, with the heterosphere above. The level at which turbulent mixing ceases to be effective is sometimes referred to as the homopause, but here we will use the term turbopause.

In addition to mixing the atmosphere, locally intensive turbulence also causes local heating. Turbulence is dissipative, extracting energy from the total flow and transferring it by a cascading process to scales so small that the random motion of the molecules, which determines the temperature, is increased. Thus turbulence in the lower thermosphere is important because its intensity affects both the relative concentrations of constituents of the thermosphere and the heat budget in the lower thermosphere. The source of the turbulent energy resides ultimately in the lower atmosphere, and this energy is transferred to the thermosphere by the upward propagation of internal atmospheric gravity waves and tidal winds (see Chapters 8 and 9).

7.2 INTRODUCTION

In the parlance of ionospheric physicists, the lower thermosphere is known as the E region. The E region has been probed from the ground almost since the inception of radio. In particular, irregularities in the structure of the ionization that produces reflections of radio waves at this altitude have been observed for decades. Because these irregularities could be observed only through reflection from the continually changing ambient ionization imbedded in the neutral atmosphere, their interpretation in terms of the turbulent structure of the neutral atmosphere was highly speculative. However, in the early 1950's, a technique was devised that enabled the characteristics of the neutral atmosphere to be determined by direct observation of relatively well understood radio reflectors—the ionized trails formed by meteorites entering the earth's atmosphere.

Measurements of the radio-frequency Doppler shifts produced as meteor trails were blown along by the neutral winds between 80 and 100 km (the altitude range over which most meteorites burn up in the earth's atmosphere) were, and still are, used to provide knowledge of the neutral wind and its variation in both height and time. At the same time, considerable progress was being made by scientists working in the analysis and interpretation of atmospheric turbulence in the troposphere. The interpretation was assisted in no small measure by the progress made in turbulence theory, in particular by Batchelor's (1953) *Theory of Homogeneous Turbulence*. After this theory was applied successfully to the explanation of the tropospheric scattering of very-high-frequency radio waves, Booker and Cohen (1956) attempted to explain the fading observed on long-duration meteor echoes in terms of turbulence in the neutral atmosphere at E-region altitudes. From their data, they deduced that energy was extracted from the large-scale wind motions at meteor altitudes and dissipated at a rate $\epsilon = 25$ W/kg. While the underlying theory was sound, their paper was attacked on the basis of their interpretation of the echo-fading process.

In the late 1950's, the chemical-release rocket technique was perfected and used to introduce a visible tracer, initially sodium, into the atmosphere over the altitude range 80 to 200 km. Such a release, made at twilight while the trail was illuminated by sunlight and the ground was in darkness, could be photographed from several camera sites on the ground, and a time series of exposures recorded simultaneously could be used for triangulation of the release, thus determining its motion with time. Winds could be determined for as long as the trail remained visible, sometimes for as long as 15 min. Since the early 1960's, trimethyl aluminum (TMA), which reacts with atomic oxygen in the ambient atmosphere to produce products that not only scatter sunlight but also glow in the dark (making nighttime as well as twilight measurements possible), has been used in addition to sodium. For daylight releases in particular, lithium has been used.

One outstanding feature of these trails is the fact that below an altitude of some 105 km the release is "obviously turbulent," whereas above that altitude the trail expands smoothly, as it should under the action of molecular diffusion alone. Considerable controversy still exists over the interpretation of data from rocket-released vapor trails as evidence for turbulence of the ambient atmosphere. Some interpretations of sodium vapor trails, for example, have led to anomalous results that are thought to be related to the energetics of the thermite burn used to produce the sodium vapor. However, interpretation of the breakup of vapor trails below what has become known as the turbopause is not the only evidence for the existence of turbulence at these altitudes. Theoretical studies of the atmospheric heat budget in the high atmosphere by Johnson and Gottlieb (1970), for example, require vertical diffusivities below the turbopause that are much larger than molecular. Calculations of the diffusivities responsible for the measured constituents above the turbopause also require similar values of diffusivity below the turbopause; composition measurements made using rocketborne mass spectrometers show the level of diffusive separation to be considerably higher than would be the case in a nonturbulent atmosphere. Further, a completely different class of measurements, based on the shearing of radio meteor trails by the winds in the 80- to 100-km region rather than the fading of individual meteor echoes, yields values consistent with those deduced from vapor-trail observations.

7.3 THE INTERPRETATION OF MEASUREMENTS

Is it possible for rocket-wake effects and chemical energy released by contaminants (from the thermite canister) to influence the subsequent dispersion? Certainly. Such effects have been documented. For energetic releases such as rocket exhausts or large quantities of explosive, the "release phase" at thermospheric altitudes lasts only for some 10 sec or less. Figures 7.1 and 7.2 show an excellent example of a release phase anomaly in a TMA trail, as photographed by the Smithsonian Institution's Baker-Nunn camera at Woomera, Australia (31° S). Figure 7.1 shows a portion of the trail 2 sec after release in the 100- to 110-km height range; vortex shedding on the right-hand side of the trail is clearly visible. In Figure 7.2, a frame taken 4 sec later, the effects of this motion have disappeared. The subsequent breakup of the trail, with the production of the characteristic "obviously turbulent" appearance, did not occur until some 30 sec later, as illustrated in the isodensitrace montages of Figure 7.3. These unique Baker-Nunn photographs are an example of the rewards of international cooperation. The Smithsonian Institution for many years operated a worldwide network of Baker-Nunn cameras for the photographic determination of satellite positions. Relationships between the satellite station staff and the Australian and British rocket experimenters at Woomera were such that the Baker-Nunn camera would be used to photograph the rocket releases when such use did not interfere with the primary mission of the observatory.

FIGURE 7.1 Portion of a trimethyl aluminum trail showing a release phase instability with vortex shedding on the right-hand edge 2 sec after release at 103 km.

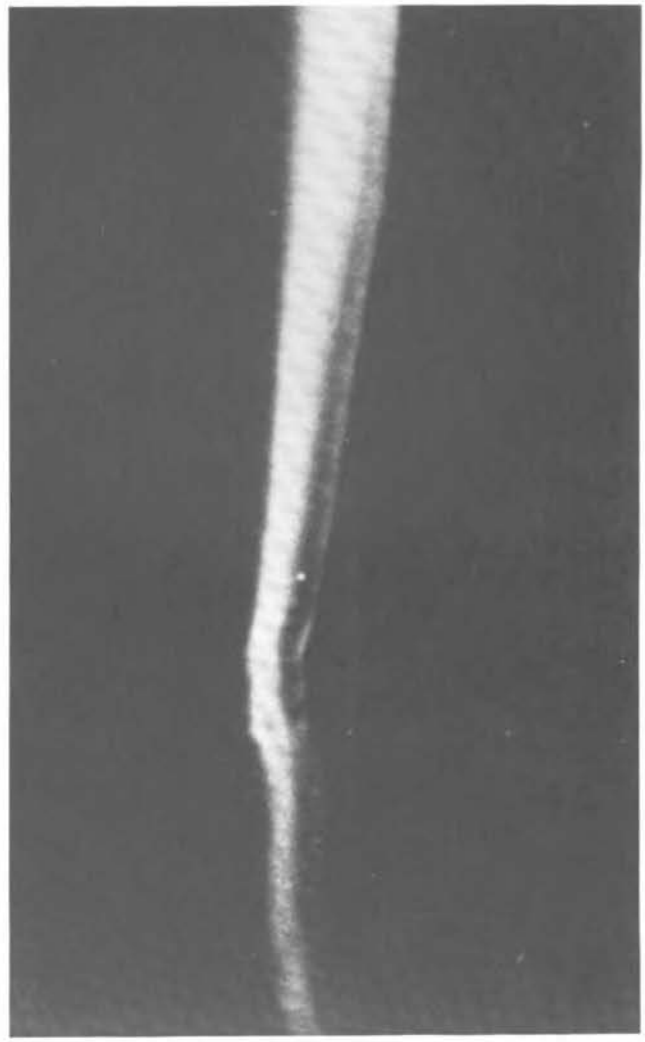


FIGURE 7.2 Same as for Figure 7.1, but 6 sec after release. Note that the release phase instability has been damped out.

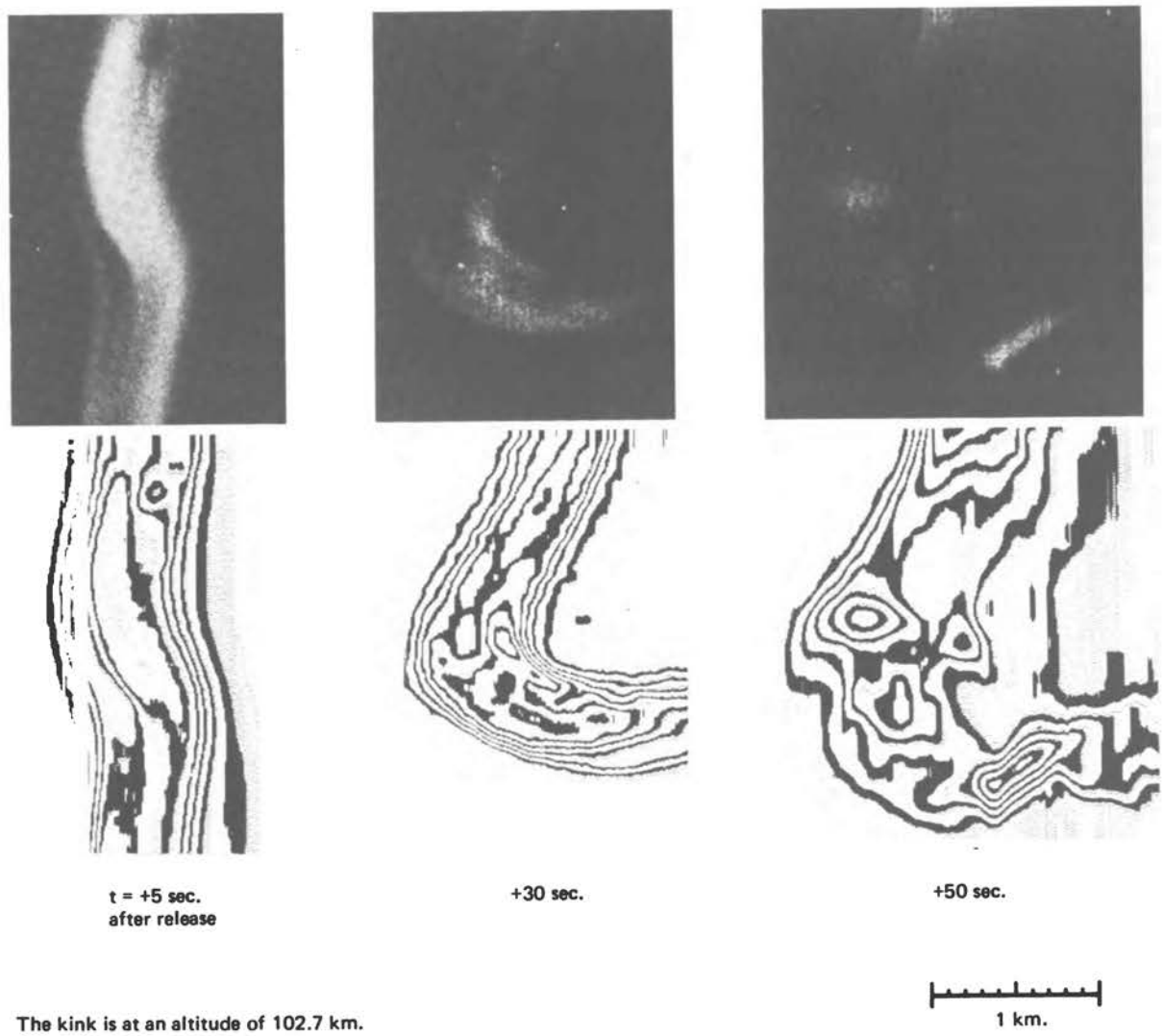


FIGURE 7.3 Time history of a portion of the trail of Figure 7.2 showing the initial laminar behavior and subsequent breakup induced by ambient atmospheric turbulence. Quantitative measurements of dispersion can be made from the isodensitrace contours. The kink is at an altitude of 102.7 km. Note that the trail is laminar at +30 sec and that the trail diameter at this time is greater than the scale of the eddies at +50 sec.

The abrupt cutoff in ambient turbulence (the turbopause), which almost always underlies a region of high wind shear, is illustrated in Figure 7.4. That the breakdown from "obviously laminar" to "obviously turbulent" represents a dramatic change, easily recognized by visual inspection of the film, is shown in Figure 7.5, in which the square of the effective radius of the trail is plotted against time after release at 105 km. The growth in the first few seconds after release is an order of magnitude faster than the subsequent growth up to the time of trail breakup, and this is regarded as representing the release phase, in which the energetics are definitely nonambient. The growth between 8 and 32 sec after release could be molecular diffusion of a cloud with an initial radius of 130 m, i.e., the release produced a cloud with, effectively, this radius at zero time. However, one can only say that the growth during this phase could be molecular—the 15 percent error in the determination of the effective radius from each film frame, and the fact that the trail cross section is only approximately Gaussian, precludes a measurement of the diffusion coefficient to better than a factor of 3. The spreading due to turbulence commences 33 sec after release and then proceeds to follow the dispersion relation predicted by the theory of homogeneous turbulence:

$$r_e^2 \propto t^3,$$

until, at 54 sec, the trail becomes too irregular for an estimate of radius based on a Gaussian distribution to be

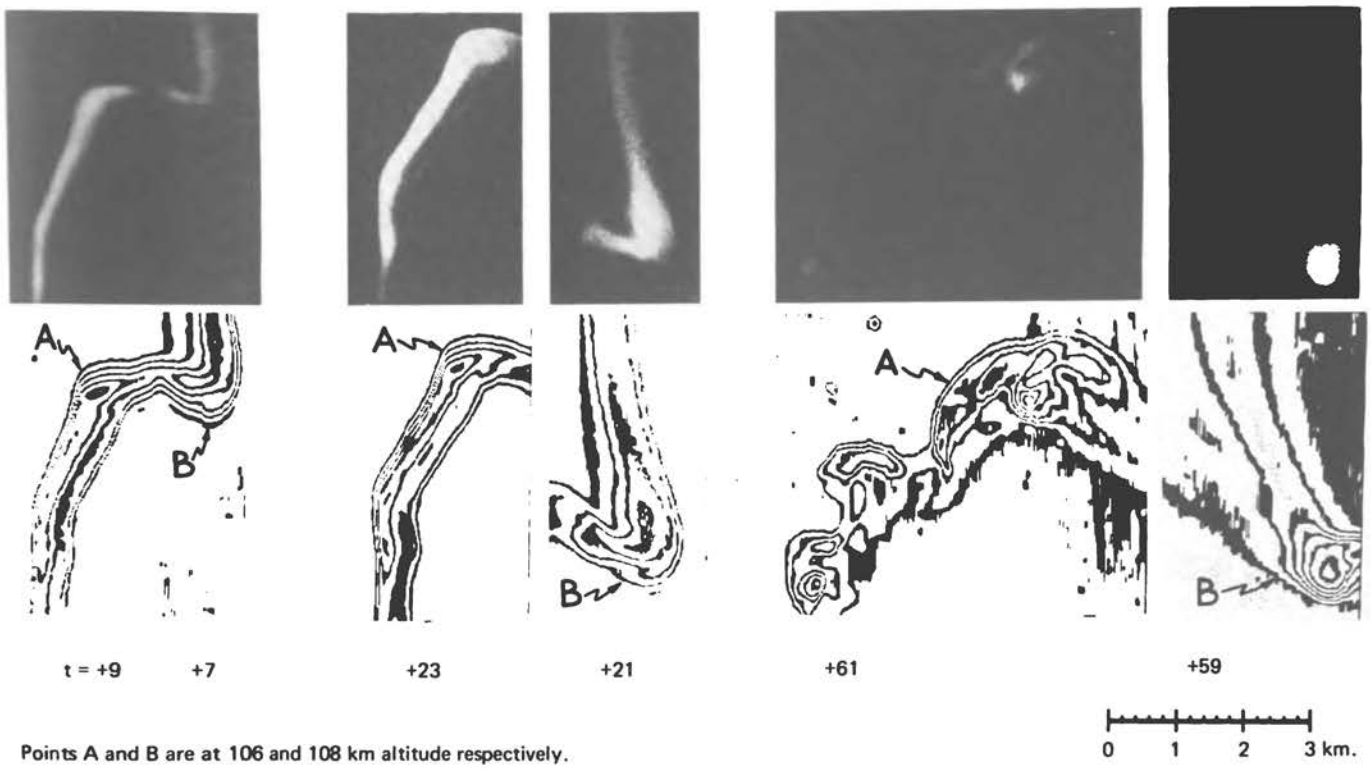


FIGURE 7.4 Montage of the behavior of the trail above and below the turbopause. Points A and B are at 106 and 108 km, respectively. Note the transition from laminar to turbulent at A between +23 and +61 sec. This has not occurred at B.

meaningful. However, Figure 7.5 does provide two parameters: the effective radius at transition and its time of occurrence after release, which should characterize the small-scale end of the turbulence spectrum. The concept is that until the trail spreads to a size equal to that of the smallest scale eddies, it is distorted but not spread by the eddies. Thus the scale size at transition is interpreted as indicating the size of the smallest eddies.

The most important parameter of any turbulence spectrum is ϵ, the rate of dissipation of turbulent energy. In order to calculate ϵ from any set of space–time correlations, it is necessary to use a model. The simplest is that of A. Kolmogoroff, as elaborated on by Batchelor. Kolmogoroff put forward the hypothesis that, in any turbulent flow field in which energy is extracted from the large-scale or mean motion and cascaded to smaller scales before eventual dissipation at scales where molecular viscosity becomes important, there exists a range of scales sufficiently removed from the large-scale anisotropic eddies, and yet not appreciably damped by molecular viscosity, that is both homogeneous and isotropic. The assumptions of homogeneity and isotropy inherent in this model are open to question in the case of lower thermospheric turbulence, but a considerable amount of work by several experimenters has shown these assumptions to be reasonable, at least for length scales less than 1 km.

Batchelor defines the basic length parameter of the viscous region (that is, the size of the small-scale eddies that are responsible for the conversion of the eddy kinetic energy to heat) as

$$\eta = (\nu^3/\epsilon)^{1/4},$$

where ν is the kinematic viscosity (a measure of molecular diffusivity), and η is the scale size in meters/radian (thus the characteristic wavelength of these turbulent eddies is

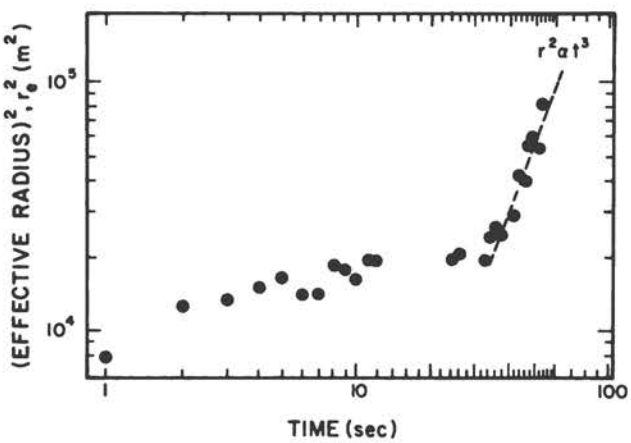


FIGURE 7.5 Variation of the square of the radius of the trail of Figure 7.1 with time after release at 105 km.

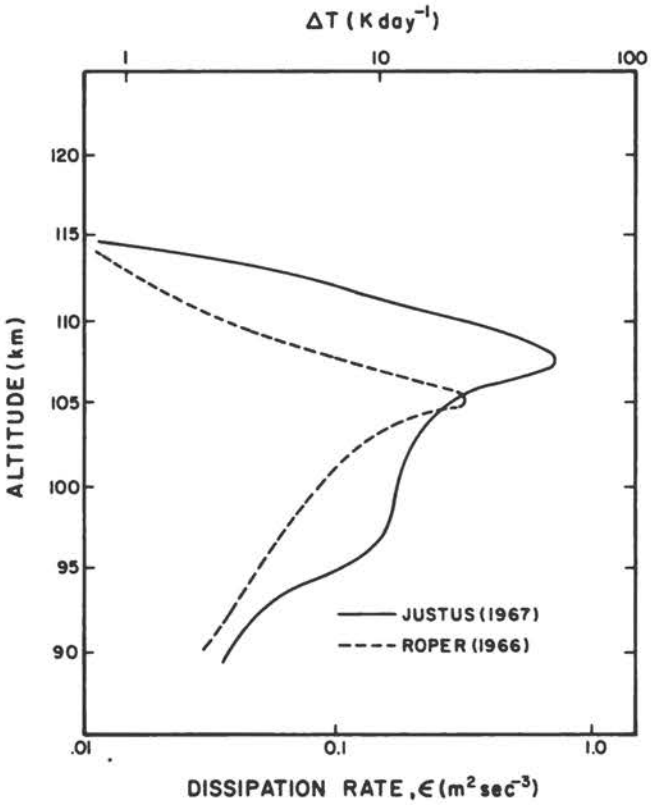


FIGURE 7.6 Profiles of the rate of dissipation of turbulent energy deduced from wind shears measured for 25 trails over Wallops Island (38° N) by Roper (1966b) and from the dispersion of a similar number of trails over Eglin Air Force Base (29° N) by Justus (1967). The upper scale gives a measure of the atmospheric heating resulting from turbulent dissipation.

$2\pi\eta$). Batchelor also defines a characteristic time constant t^* corresponding to this unit length scale such that

$$\eta = (\nu t^*)^{1/2}.$$

t^* is the lifetime of the smallest eddies, essentially the time required for their dissipation by molecular viscosity. Combining these two equations to eliminate η yields

$$t^* = (\nu/\epsilon)^{1/2}.$$

In terms of the length scale η,

$$\epsilon \propto \eta^{-4}.$$

In terms of the time scale t^*,

$$\epsilon \propto t^{*-2}.$$

Wind shears have been used to calculate the variation with height of the rate of dissipation of turbulent energy (Roper, 1966b). The kinematic viscosity ν may be determined from the viscosity and density published in the U.S. Standard Atmosphere Supplements for 1966. The function

$$\nu\ (\mathrm{m^2\ sec^{-1}}) = \exp\,[0.17\ (z - 80.0)],$$

with z in kilometers, fits the data at the tabulated altitudes. The ϵ profile determined in this way is shown in Figure 7.6, together with the profile inferred from measurements made by Justus (1967), calculated from the velocity fluctuations observed on 18 TMA trails. These values of ϵ and ν have been coupled to produce what can be regarded as average height-dependent characteristic length scales η_R (= $2\pi\eta$, in order to express the length scale in the more usual wavelength notation, meters/cycle). η_R and t^* are plotted, together with values calculated from Justus's ϵ profile, in Figures 7.7 and 7.8, respectively. Also plotted are the values determined from two Skylark-released TMA trails photographed by the Baker-Nunn camera at Woomera at dawn and dusk on May 31, 1968.

The similarity in form, and the order of agreement, between the predicted "average" trail time constants t_R^* and t_J^* (Roper's and Justus's values, respectively) and the measured values for each of the two releases is surprisingly good when one considers the approximations involved in (a) the model atmosphere viscosity, which is based on an average atmospheric model; (b) the variation, both diurnal and seasonal, in the turbulent dissipation rate, which has been averaged out in the construction of the t_R^* and t_J^* profiles; and (c) the fact that, while the

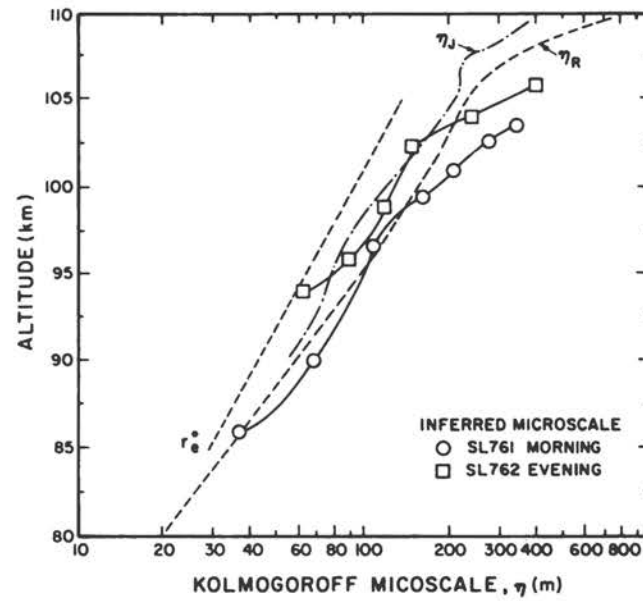


FIGURE 7.7 The variation with altitude of the Kolmogoroff microscales η_J and η_R inferred from the turbulent dissipation profiles of Figure 7.6. Also shown are the microscales inferred from the May 1968 trails. r_e^* is the Gaussian radius of the evening trail at the time of onset of trail breakup, t^*.

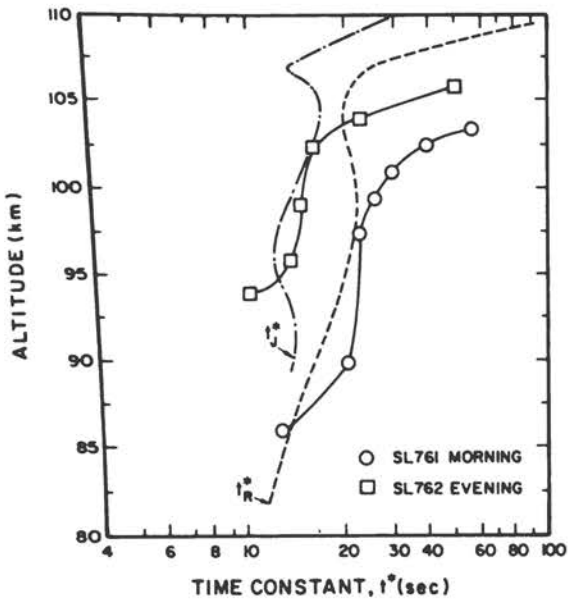


FIGURE 7.8 The variation with altitude of the time constant of the Kolmogoroff microscale for the ϵ profiles of Figure 7.6 and the May 1968 trails.

parameter t^* appears to be a measure of the time taken for trail breakup to occur, it is not clear why it should be.

The definition of η is a mathematical expedient characterizing the scale size for turbulence at which viscous dissipation becomes important. There is no obvious reason why either η or t^* should be physically measurable features of the motion. Nevertheless, the close similarity in the shapes of the t^* curves strongly indicates that the time delay in the onset of turbulence should be related to the time constant of the Kolmogoroff microscale. Furthermore, these observations suggest a reason why the turbopause, defined as the boundary between the regions that break up and those that remain laminar, should manifest itself so abruptly. Above 105 km, the time constant t^* for the onset of turbulence increases so rapidly with altitude that the trail is not, in general, observed for a sufficient length of time for visible breakup to occur.

Attempts have been made to explain the existence of the turbopause in terms of a critical value of some parameter, generally the Reynolds number R_e or the Richardson number R_i. These attempts have met with marginal success at best, partly because of the difficulty in defining the characteristic lengths that occur in these parameters and partly because it is not evident *a priori* what critical value the parameter should have at the turbopause. Without a detailed knowledge of the temperature gradient at the scale characteristic of the vertical mixing process (a few hundred meters or less), the Richardson number just cannot be specified. Johnson (1975) has considered the relative importance of buoyancy and dissipation in some detail, referring particularly to the work of J. D. Woods, who determined that there was hysteresis in the criterion—laminar flows become turbulent when $R_i \leq 0.25$, while turbulent flows become laminar when $R_i \geq 1$. While the mean (undisturbed) atmospheric temperature profile is pertinent to the breakdown from laminar to turbulent flow, once turbulence is established its cessation will depend on the temperature gradient as modified by the turbulence. As yet, there is no technique available for measuring such temperature gradients in the lower thermosphere.

Another error that has often been made in attempting to explain turbulence in the lower thermosphere is the assumption that the turbopause corresponds to an altitude at which turbulence ceases abruptly. The results presented here, on the contrary, show that the turbopause is the altitude at which the time constant of the Kolmogoroff microscale of the turbulence increases very rapidly with altitude. This viewpoint resolves the paradox that regions above the turbopause, which were thought of as nonturbulent, have diffusion coefficients based on the measured laminar trail growth that are greater than molecular. We now see that turbulence does exist above the turbopause but that its efficacy in transport relative to molecular diffusion decreases with altitude. At an altitude of 130 km, the contribution of turbulence to diffusivity is insignificant, even though its absolute value may be as large as it is at the turbopause.

The rate of dissipation of turbulent energy may also be calculated from η, the length scale at which eddy diffusion becomes effective. However, since this length depends critically on the shape of the cloud (the assumed Gaussian variation across the cloud is rarely realized in practice below the turbopause) and because $\epsilon \propto \eta^{-4}$, Rees *et al.* (1972) chose to use the relatively precisely determined t^* values to calculate the variation of ϵ with height. This is shown for a pair of dawn and dusk releases above Woomera (30° S) on May 31, 1968, in Figure 7.9. Note that

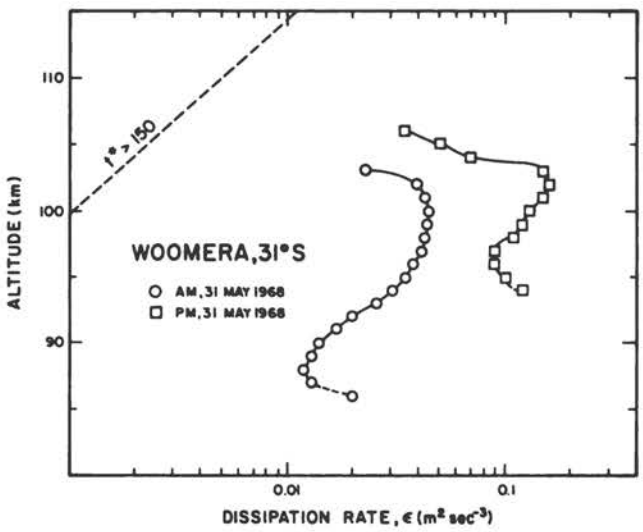


FIGURE 7.9 Estimates of the rate of dissipation of turbulent energy ϵ from the time of onset of turbulence for the release of May 1968.

the turbopause is higher in the evening than in the morning and that the higher turbopause is associated with a greater overall turbulent intensity. This substantiates the suggestion of a variation in midlatitude turbopause height made by Elford and Roper (1967), which was based on seasonal variations in turbulent intensity at 93 km as determined from the wind shear measured simultaneously on individual radio meteor trails.

The results from two further TMA releases made above Woomera at dawn on October 16 and dusk on October 17, 1969, are presented in Figure 7.10. For this pair of trails, the turbulent intensity is higher in the morning than in the evening—opposite to the May 1968 releases. This diurnal variation with season is the same as that measured for the large-scale turbulent velocity component from radio meteor winds at Adelaide (35° S). These October releases are of particular interest in that they show alternating laminar and turbulent regions similar to those previously reported by Blamont and Barat (1967). The various layers observed in these releases do not seem to be quite so simply related to the wind profile as those of Blamont and Barat. However, the regions of prolonged laminar behavior all seem to be located at altitudes where the wind shear is high.

In an attempt to explain why turbulence in the lower thermosphere should be stratified at times, and why in fact an ostensibly highly stable region of the earth's atmosphere should be turbulent at all, Lloyd *et al.* (1972) proposed a model in which random internal gravity waves produce turbulence accompanied by a considerable modification of the temperature profile. The creation of temperature inversions by turbulence is commonplace in the troposphere. It is proposed that a similar effect occurs in the more stable lower thermosphere, with part of the vertical component of the gravity-wave velocity as the source of the turbulent energy. If random, internal gravity waves propagate at a significant angle to the vertical, then their vertical velocity component will be able to contribute to the destabilization of the stably stratified lower thermosphere (Hodges, 1967). Similar destabilization mechanisms have been discussed for the oceans by Phillips (1971) and for the lower stratosphere by Roach (1970). Radio meteor studies have already established that gravity waves are the source of the turbulent energy in the lower thermosphere. This has been confirmed quite independently in radio-meteor studies by Spizzichino (1972) and by Blamont and Barat (1967) through observations of chemical releases.

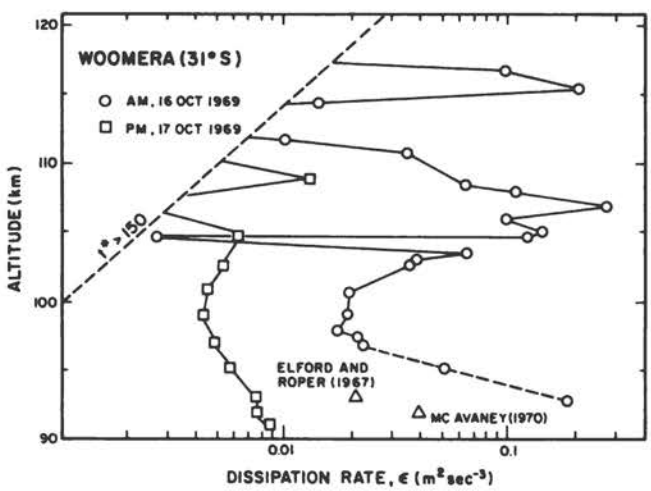


FIGURE 7.10 Same as for Figure 7.9 but for the releases of October 1969. Note the presence of turbulent "sheets" alternating with laminar layers. Diurnally averaged dissipation rates for the month of October 1961 (Elford and Roper, 1967) and October 1969 (McAvaney, 1970) as measured by the radio-meteor technique at Adelaide (35° S) are shown for comparison.

The lower thermosphere is stabilized against vertical motion by the mean temperature gradient; above the mesopause near 80 km the mean temperature increases with height. By equating the work done in the vertical displacement of a parcel of gas to the fraction of the vertical component of the gravity-wave spectrum that is responsible for the measured turbulence spectrum, Lloyd *et al.* deduced the modification of the mean temperature profile that would be produced by the turbulence measured on the October 16 trail in the height range 104 to 110 km. An isodensitrace montage of this portion of the trail is shown in Figure 7.11. The choice of the turbulent layers within which the analysis can be applied is somewhat subjective, being based on regions where growth is "obviously" different from that above and below.

The solid lines of Figure 7.12 show the results of the application of the Lloyd *et al.* model to each of the layers delineated in Figure 7.11. The dashed lines are necessary for profile continuity and must represent laminar sheets. Because of the discontinuous nature of the determination of the modified profile and the forced fitting of the midpoint temperatures, the magnitudes of the positive and negative gradients are open to question. However, it is interesting to note that the existence of similar gradients in the lower stratosphere, a region of similar mean-temperature gradient, is well documented from aircraft observations, as can be seen from the project HICAT determination shown in Figure 7.13 from Mitchell and Prophet (1969). Unfortunately, the flight of an aircraft at constant altitude cannot reveal a height profile of turbulence, but at least in the encounter with CAT (clear air turbulence) at 21 km, the temperature profile has been modified in a manner consistent with the present model.

Several deficiencies exist in the model, since the finite time constants of the processes involved (the period of the destabilizing gravity wave, for example) have not been considered. The basic energy-budget equation can be made more general by the inclusion of terms describing energy sinks (e.g., heat conduction). One promising model being developed uses the reversible heating associated with propagating gravity waves (Hines, 1965) as the initial destabilizing energy. Even with this criticism, the above semiempirical approach allows deduction of

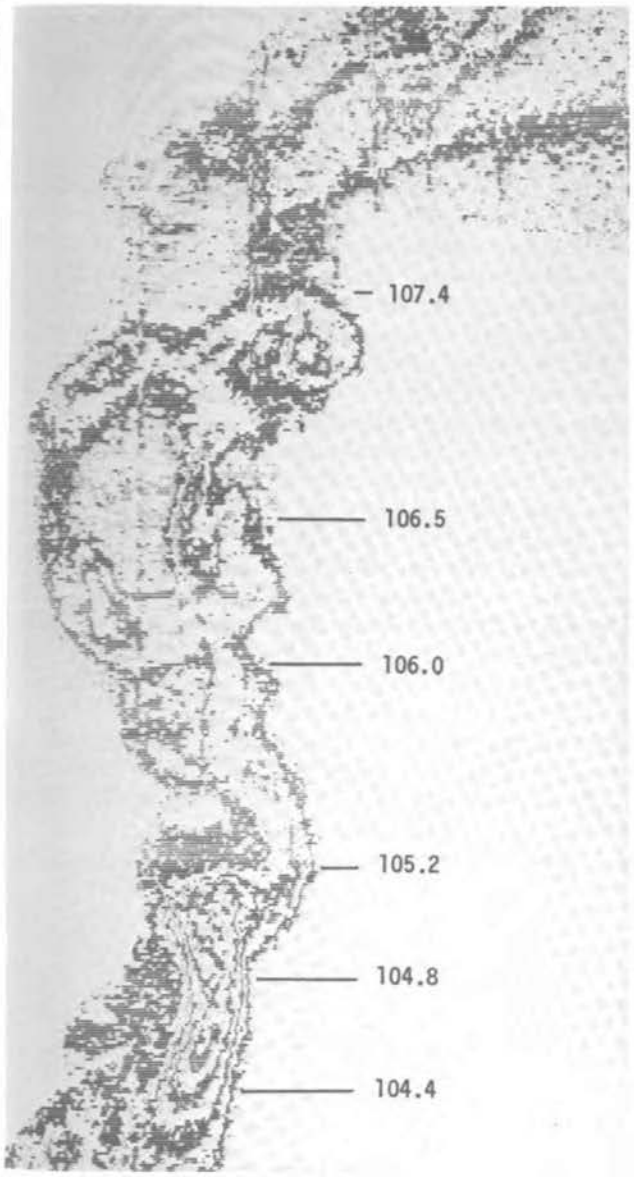


FIGURE 7.11 Isodensitrace of portion of the morning trail of October 16, 1969, 50 sec after release, showing the division into laminar and turbulent layers. Altitudes are shown in kilometers.

many reasonable properties of the atmosphere. In particular, the model counters the objections raised to the existence of turbulence in what ostensibly is a highly stable region; the presence of turbulence itself tends to destabilization by modification of the temperature profile.

Up to this point, major emphasis has been placed on the fundamental parameter ϵ, the rate of dissipation of turbulent energy. There is an equally important although not so easily defined parameter, K_z, the coefficient of turbulent eddy diffusion in the vertical, which is the transport parameter incorporated in all one-dimensional models of the chemistry and constituents of the lower thermosphere. For some time, there was considerable discrepancy between the diffusion coefficients calculated from the growth of rocket-released tracers and those inferred from measurements of diffusive separation and atmospheric heat-budget calculations. With the discovery that the turbulence responsible for the enhanced diffusivity of chemical releases in the lower thermosphere was highly

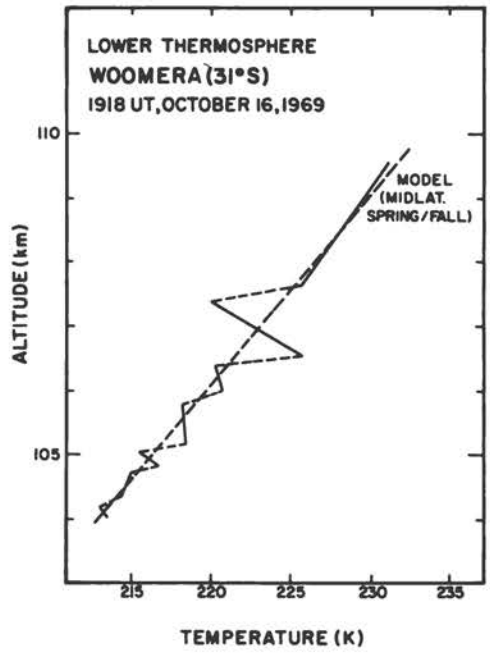


FIGURE 7.12 Theoretical morning temperature structure deduced from the profile of Figure 7.11.

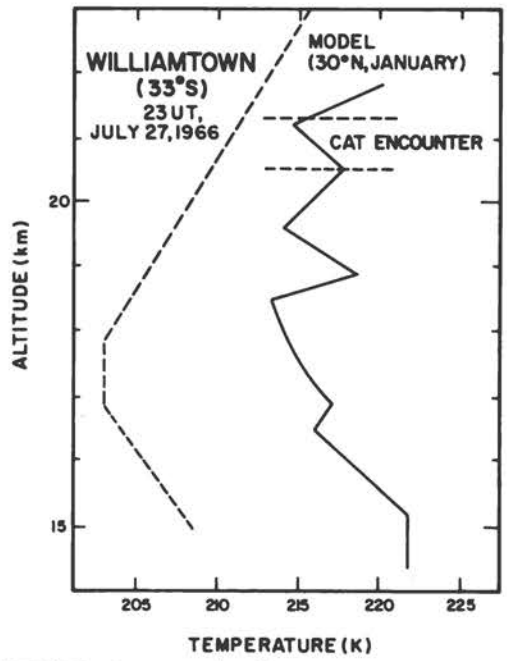


FIGURE 7.13 An example of temperature structure in the lower stratosphere (Mitchell and Prophet, 1969).

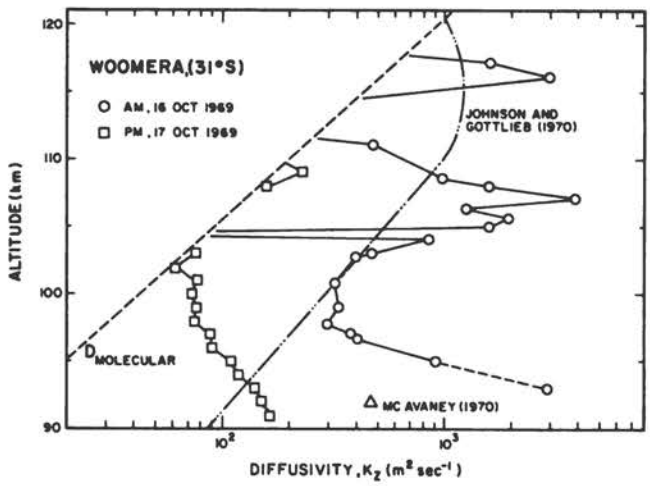


FIGURE 7.14 The turbulent vertical eddy-diffusion coefficient profiles deduced from the structure on the October 1969 releases. Δ indicates the coefficient deduced from simultaneous radio-meteor observations at Adelaide (35° S), approximately 450 km southeast of Woomera.

anisotropic with horizontal scales ten times those of the vertical, this discrepancy was readily explained. Since the time constant t^* for the observed onset of turbulence as used by Rees *et al.* (1972) is characteristic of the small-scale, isotropic turbulence spectrum, Lloyd *et al.* (1972) used the ϵ values thus determined to estimate the vertical diffusion coefficient. By an extension of their temperature profile modification model, they equated the vertical destabilizing influence of the turbulence to the stabilizing influence of the mean-temperature profile to obtain a vertical diffusion coefficient

$$K_z = \frac{\beta\epsilon}{\frac{g}{T_0}\left(\frac{dT_0}{dz} + \Gamma\right)},$$

where g is the acceleration due to gravity, T_0 is the temperature at altitude z, dT_0/dz is the undisturbed mean temperature gradient, and Γ is the adiabatic temperature lapse rate, 9 K/km. The above relationship is based on the fact that as long as the vertical temperature gradient remains greater than the adiabatic, there is an inhibition of vertical transport by vertical motion. It is interesting to compare this relationship with that determined independently by Lilly *et al.* (1973) for the lower stratosphere. The constant β above was determined from turbulence theory to be 10. Lilly arrived at the value of 1/3! This discrepancy is not so serious as it appears at first sight, since Lilly's temperatures and temperature gradients were the measured values, leading to a considerably smaller denominator than that produced by the use of the undisturbed mean temperature and gradient values by Lloyd *et al.* The variation of K_z with height for the October 1969 releases is shown in Figure 7.14. The theoretically deduced upper limit for eddy diffusion, based on the heat flux model of Johnson and Gottlieb (1970), is shown for comparison. Note that the maximum value of eddy diffusivity estimated by Johnson and Gottlieb is the integrated global maximum averaged over all seasons and can be exceeded by a particular measurement if there is any temporal, latitudinal, diurnal, or seasonal variation. Note also that all the turbulent intensity and diffusivity profiles presented here reflect the apparent sharp cutoff in turbulence as observed on chemical trails. With the time constant in onset of turbulence increasing so rapidly at and above the turbopause, most trails are not observable long enough for breakup to occur. However, observed diffusivities between the turbopause and approximately 130 km are usually greater than molecular, in agreement with the Johnson and Gottlieb calculations.

The only two sets of data so far produced that are amenable to analysis in terms of month-by-month variation of ϵ or K_z were determined from meteor trail shear measurements made at Adelaide (35° S) by Roper (1966a) and McAvaney (1970). The variation of the monthly mean K_z at 93 km is shown in Figure 7.15. Also plotted in this figure are the results deduced from a far more rigorous treatment of the 1961 ϵ data by Zimmerman; some modifications of absolute values is evident, but the overall variation, with equinoctial maxima, remains. There is a real difference between the 1961 and the 1969 data that is

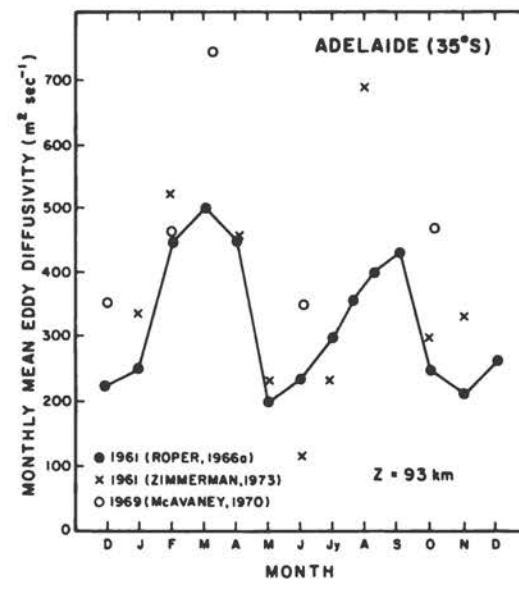


FIGURE 7.15 The southern hemisphere midlatitude variation of vertical eddy diffusivity deduced from two years of radio-meteor wind-shear observations. Zimmerman's values result from a more rigorous analysis of the same 1961 data.

not one of location, interpretation, or technique. The higher values measured in 1969 may be a consequence of the higher solar activity that year.

7.4 SOURCES OF TURBULENCE ENERGY

The fact that random internal gravity waves are the source of the turbulence energy in the lower thermosphere has already been mentioned. While this subject is covered in more detail in Chapter 8, one further correlation is pertinent to this discussion. In looking for possible sources of turbulence in the meteor region, comparisons were made with the magnitudes of the prevailing and tidal winds and shears as measured simultaneously by the radio-meteor method. While no apparent relationship existed between the turbulence parameters and the prevailing and semidiurnal winds, the seasonal variation of the amplitude of the diurnal oscillation was highly correlated with the turbulent intensity. The monthly means of the amplitude of the diurnal tide for several years of observation are shown in Figure 7.16. Note that equinoctial maxima in the diurnal tidal wind amplitude are not observed in all years and, therefore, that the measured variation in turbulent intensity may not be characteristic of all years.

The global-scale diurnal tidal wind does not produce turbulence directly but by a cascade process in which the tidal wind becomes unstable and generates *in situ* a spectrum of random internal gravity waves. This instability in the diurnal tidal wind may come about either because its amplitude becomes large or because of nonlinear interactions with gravity waves generated below and propagating upward through the lower thermosphere. Such gravity waves propagating from below can themselves be a direct source of turbulent energy. Thus the dominant feature of turbulence in the lower thermosphere, even if it is present at all times, will be the intermittency of its intensity. Little is known of the role played by large-scale motions such as planetary waves in the stability of this region of the atmosphere. In fact, since practically all measurements of turbulent intensity have been made at middle latitudes, the properties of the global turbopause and the influence of the turbopause on, for example, the hemispherical asymmetries in minor constituents measured globally at satellite altitudes can hardly be estimated at this juncture. Our knowledge of the temperature structure of the turbopause region at scales less than 1 km, which is crucial to the understanding of turbulence and diffusivity, is woefully inadequate. Present measurement techniques are quite incapable of producing such detail at these altitudes.

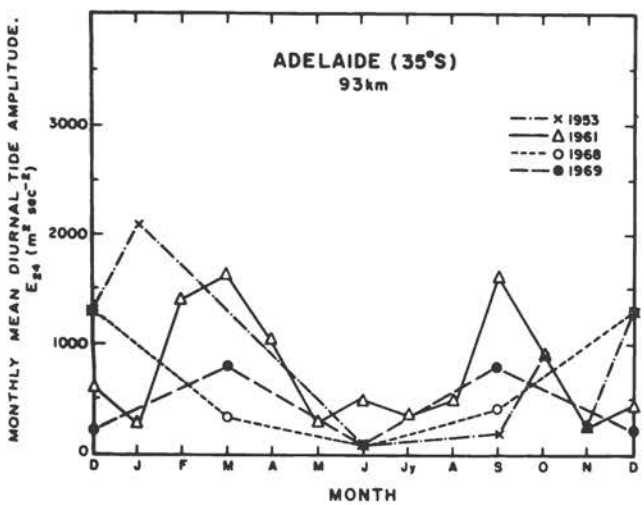


FIGURE 7.16 The seasonal variation of the southern hemisphere midlatitude diurnal tidal wind energy per unit mass determined from radio-meteor data. Equinoctial maxima are not a feature of all years.

In addition to the modifications to the temperature profile already discussed, the dissipation of turbulent energy produces heating of the ambient atmosphere. For the mixed atmosphere of mean molecular weight 29, a rate of turbulent dissipation ϵ of 1 W/kg produces heating at a rate of 85 K per day. The top scale of Figure 7.6 gives the heating rates appropriate to the inferred turbulent dissipation rates. In the light of the considerable variability indicated by the measurements presented here, even more emphasis must be given to the rate of dynamical heating of the lower thermosphere, as proposed by Hines (1965). The dissipation of wind energy at these altitudes may, at times, give rise to local heating rates in excess of those due to the solar input, which has usually been considered to be the major source of heating in this region. The relative importance of turbulence in mixing and dynamical heating has been summarized by Johnson (1975).

7.5 CONCLUSIONS

While the turbulence in the lower thermosphere is isotropic to scales of a few hundred meters, the transition from turbulence to the nondissipative scales of the gravity-wave spectrum is gradual, and therefore horizontal diffusivity is greater than vertical diffusivity. Based on diffusivity criteria alone, one would say that the turbulence is anisotropic, with a vertical scale of the order of 1 km and a horizontal scale of a few kilometers. The turbulent intensity is intermittent in both space and time, with large diurnal and significant seasonal and possible solar-cycle variations. For this reason alone, it is essential that simultaneously measured *in situ* values of atmospheric parameters be used in any attempt at meaningful comparisons. Since practically all measurements of turbopause altitude and turbulent intensity have been made at middle latitudes, and those low and high latitude measurements that have been made have not been coor-

dinated with simultaneous observations elsewhere, variations with latitude are largely unknown. However, it has been established by rocket-grenade and falling-sphere measurements that the winter polar mesopause is warmer than that at the summer pole and that in the absence of solar input a significant atmospheric dynamical heating source is responsible. As has been shown, the turbulent dissipation of wind energy in the 90- to 125-km height range is significant and should therefore be included in any model of the thermospheric heat budget.

While knowledge of the vertical diffusivity in the lower thermosphere is vital to the understanding and modeling of thermospheric constituents, the role played by the global turbopause in the dynamics of the thermosphere has yet to be determined. High-resolution photography, with a frame rate of at least one every 2 sec, is able to produce data on turbulent intensities and vertical diffusivities from rocket vapor trails, but such measurements are highly localized in time and space. A recent development, the use of a high-flying aircraft as a camera platform, overcomes two ground-based camera problems: atmospheric haze and clouds are avoided, and photographs can be taken in the daytime. Daytime releases of lithium can also be observed from the ground using narrow band filters and electronic scanning systems (recording on video tape), but interpretation of the dispersion of the highly energetic release in terms of turbulence parameters is difficult. Vertical diffusivities and turbopause altitudes can be inferred from rocketborne mass-spectrometer measurements, regardless of hour of day or cloud cover but with location limitations similar to the vapor-trail method (both require reasonable range facilities). Multistation meteor wind radars can provide a continuous measurement of the turbulent intensity just below the turbopause, but only a few stations, all in midlatitudes, are currently capable of this type of operation, and none operates continuously. Incoherent-scatter radars are able to measure the temperature structure in the neighborhood of the turbopause with about 2-km height resolution, but these measurements are subject to constraints similar to, but even more severe than, those of the radio-meteor technique. A suggestion by Bencze (1970) that an ionosonde can be used to measure turbopause altitude warrants further investigation, since a global network of ionosonde stations is already in existence.

A proper understanding of the nature of the turbulence in the lower thermosphere requires a knowledge of the temperature profile with better than 1-km (preferably 100-m) resolution—a resolution that is not technically feasible at this time. Nevertheless, global variations in turbopause altitude and intensity can be determined using well-established techniques but only with international cooperation. Simultaneous global observations could be coordinated through the Middle Atmosphere Program (MAP) of the Special Committee on Solar Terrestrial Physics (SCOSTEP) and the International Mesospheric and Ionospheric Structure Parameter Interaction program (MISPI) proposed by the Soviet Union. Particular emphasis is being placed by these programs on the encouragement of the Arctic, equatorial, southern hemisphere, and Antarctic observations so badly needed for global coverage.

ACKNOWLEDGMENT

The writing of this paper was supported in part by the National Aeronautics and Space Administration under Grant NGL 11-002-004.

REFERENCES

Batchelor, G. K. (1953). *The Theory of Homogeneous Turbulence*, Cambridge U. Press, New York.

Bencze, P. (1970). An analysis of the virtual height of ionospheric sporadic $E(h'E_s)$, *Acta Geodaet. Geophys. Montanist. Acad. Sci. Hung. 5*, 223.

Blamont, J. E., and J. Barat (1967). Dynamical structure of the atmosphere between 80 and 120 km, in *Aurora and Airglow*, B. McCormac, ed., Reinhold, New York, p. 159.

Booker, H. G., and R. Cohen (1956). A theory of long duration meteor echoes based on atmospheric turbulence with experimental confirmation, *J. Geophys. Res. 61*, 707.

Elford, W. G., and R. G. Roper (1967). Turbulence in the lower thermosphere, in *Space Res. VII*, North-Holland, Amsterdam, p. 42.

Hines, C. O. (1965). Dynamical heating of the upper atmosphere, *J. Geophys. Res. 70*, 177.

Hodges, R. R., Jr. (1967). Generation of turbulence in the upper atmosphere by internal gravity waves, *J. Geophys. Res. 72*, 3455.

Johnson, F. S. (1975). Transport processes in the upper atmosphere, *J. Atmos. Sci. 32*, 1658.

Johnson, F. S., and B. Gottlieb (1970). Eddy mixing and circulation at ionospheric levels, *Planet. Space Sci. 18*, 1707.

Justus, C. G. (1967). The eddy diffusivities, energy balance parameters, and heating rate of upper atmospheric turbulence, *J. Geophys. Res. 72*, 1035.

Lilly, D. K., D. E. Waco, and S. I. Adelfang (1973). Stratospheric mixing estimated from high altitude turbulence measurements, Paper 73-497, AIAA/AMS Intern. Conf. Environmental Impact of Aerospace Operations in the High Atmosphere, Denver, June 11–13, 1973.

Lloyd, K. H., C. H. Low, B. J. McAvaney, D. Rees, and R. G. Roper (1972). Thermospheric observations combining chemical seeding and ground based techniques. I. Winds, turbulence and the parameters of the neutral atmosphere, *Planet. Space Sci. 20*, 761.

McAvaney, B. J. (1970). Small scale wind structure in the upper atmosphere, Ph.D. Thesis, U. of Adelaide, Australia.

Mitchell, F. A., and D. T. Prophet (1969). Meteorological analysis of clear air turbulence in the stratosphere, in *Clear Air Turbulence and Its Detection*, Y.-H. Pao and A. Goldberg, eds., Plenum, New York, p. 144.

Phillips, O. M. (1971). On spectra measured in an undulating layered medium, *J. Phys. Oceanog. 1*, 1.

Rees, D., R. G. Roper, K. H. Lloyd, and C. H. Low (1972). Determination of the structure of the atmosphere between 90 and 250 km by means of contaminant releases at Woomera, May 1968, *Phil. Trans. R. Soc. London, Ser. A271*, 631.

Roach, W. J. (1970). On the influence of synoptic development on the production of high level turbulence, *Q. J. R. Meteorol. Soc. 96*, 413.

Roper, R. G. (1966a). Atmospheric turbulence in the meteor region, *J. Geophys. Res. 71*, 5785.

Roper, R. G. (1966b). Dissipation of wind energy in the height range 80 to 140 kilometers, *J. Geophys. Res. 71*, 4427.

Spizzichino, A. (1972). Wind profiles in the upper atmosphere deduced from meteor observation, in *Thermospheric Circulation*, W. L. Webb, ed., MIT Press, Cambridge, Mass.

Rossby-Planetary Waves, Tides, and Gravity Waves in the Upper Atmosphere

8

WILLIAM H. HOOKE
Wave Propagation Laboratory
National Oceanic and Atmospheric Administration

8.1 PROLOGUE

The outstanding structural feature of all planetary atmospheres is their effectively exponential density decrease with height under the action of gravity. In the earth's atmosphere, the density decreases by 12 orders of magnitude between the surface and the base of the exosphere. As we have seen in previous chapters, this enormous variation of density with height and the changes in composition associated with it combine to produce an atmospheric photochemistry correspondingly rich in variety. The qualitative nature of the photochemistry presents a different aspect with every change in altitude of a scale height or so, and the chemistry as a whole changes from an essentially neutral regime in the lower atmosphere to a plasma regime in the magnetosphere. Thus the upper atmosphere presents us with a sophisticated photochemistry and plasma-physics laboratory, available to us to the extent that we are able to probe and monitor it using *in situ* sensors on rocket or satellite platforms or ground- and satellite-based remote-sensing devices such as ionosondes, radars, lidars, radiometers, and spectrometers. In recent years, our utilization of this laboratory has become increasingly active. We are now able to modify a number of upper atmospheric processes—stimulating artificial airglow and aurora and heating the ionospheric plasma in controlled ways.

But the upper atmosphere is more than a laboratory. It serves as a shield, protecting us from the harsh particle and radiative environment of the interplanetary medium. In recent years, we have come to learn that this shield is possibly more fragile, more vulnerable to man's depredations than we had previously supposed. Because our understanding of the processes sustaining this shield and our impact upon it is so fragmentary, the photochemistry and plasma physics of the upper atmosphere are now urgent problems that we must face.

As the earlier chapters of this volume serve to indicate, the complexity of these problems would severely tax our modeling capabilities for some time to come even if the atmosphere were stationary. In fact, however, the air is in a pronounced state of motion, ranging in scale from circulations that are truly global to turbulent eddies that may be only centimeters in dimension. The result is that the

continuity equations describing the photochemistry and plasma physics of both major and minor atmospheric constituents must contain divergence or transport terms associated with these motions. For example, any model of the ozone balance of the upper atmosphere, to be complete, must incorporate or parameterize stratospheric dynamics, as indicated in Chapters 9 and 10. Similarly, if we are to understand the effects on high-frequency radiowave propagation of phenomena such as traveling ionospheric disturbances and ionospheric storms, we must include dynamical processes in the models.

Thus the subject of atmospheric dynamics forms a large part of this study, being considered explicitly or implicitly in every chapter. Other chapters treat the electrically neutral atmospheric dynamics in its largest and smallest aspects—from the global circulations engendered by nonuniform heating (Chapter 3) to the turbulent motions producing local transport and diffusion, as well as an energy cascade into molecular dissipation (Chapters 7 and 9)—and the plasma dynamics involved in the interaction between the charged particles of the upper atmosphere and the earth's magnetic field (Chapters 1, 2, and 5). But we observe that the large-scale circulations account for only part of the transport in the upper atmosphere. We find similarly that small-scale, turbulent motions do not account for the rest. The remainder—an important fraction—occurs as a result of neutral atmospheric motions on six orders of magnitude of intermediate scale—motions between a few meters and a few thousand kilometers in spatial dimension and having temporal scales between a few seconds and several days.

It happens that most of the atmospheric motions on these scales can be interpreted as wavelike in nature. The time series they display on various records are often nearly sinusoidal. They exhibit a high degree of spatial correlation. They propagate with well-defined phase and group speeds that appear to exhibit appropriate dispersion. The study of these wave motions is thus attractive from two points of view. On the one hand, the waves appear to be an important part of the dynamics of the upper atmosphere, contributing significantly to the transport processes that occur there. On the other, a wide variety of mathematical tools are available for analyzing atmospheric wave motions; the wave approach to the study of atmospheric dynamics has consistently proven to be a powerful one. In this chapter we present in broad outline a summary of the progress of this subject and its prospects for the future.

In general, wave motions involve an interchange of energy of various forms—kinetic, internal, electromagnetic, and gravitational potential, for example. Because the earth's upper atmosphere provides a wide variety of forms of energy storage, it sustains a rich spectrum of wave motions of different types, including at various ends of the spectrum hydromagnetic waves, inertial oscillations, acoustic waves, and the like. However, the observationally important wave motions of the electrically neutral upper atmosphere fall for the most part into one of the three dominant classes—the Rossby-planetary waves, the tides, and the gravity waves. In this chapter we shall describe briefly the physical characteristics basic to each of these three wave types, the various complicating factors affecting their propagation in the real atmosphere, and the processes governing their generation and dissipation. In addition to the wave kinematics, we consider wave dynamics, which plays a greater role in atmospheric physics than most people realize. One of the striking features of geophysical fluid dynamics is that wave motions are so often nearly monochromatic. All of us have often seen displayed, either on the ocean surface or in clouds, patterns revealing wavefront after wavefront, evenly spaced so that the wave nature of the phenomenon leaps out at us in a way that would not be nearly so apparent if the wave spectrum were fairly broad and all we saw was the chaotic superposition of many waves. An example is shown in Figure 8.1, which shows wave motions on two scales modulating noctilucent clouds at 85-km altitude—the height of the mesopause. With examples such as this before us, and with the mathematics of linear monochromatic plane waves being relatively simple and straightforward, it is no wonder that theorists have found the interpretation of such events so tempting. But it turns out that atmospheric wave motions are much more than a mere curiosity of the atmospheric motion field; because waves so effectively transport momentum and energy without requiring a concomitant mass transport, they are dynamically quite important, even when they are of relatively small amplitude. In the troposphere, planetary waves provide significant meridional transports of momentum and energy, while gravity waves produce corresponding vertical transports of these quantities. The effect of these transports is to produce a global climate that is much more temperate than it would otherwise be. In the upper atmosphere, analogous processes are at work but with greater intensity. Because the waves propagate

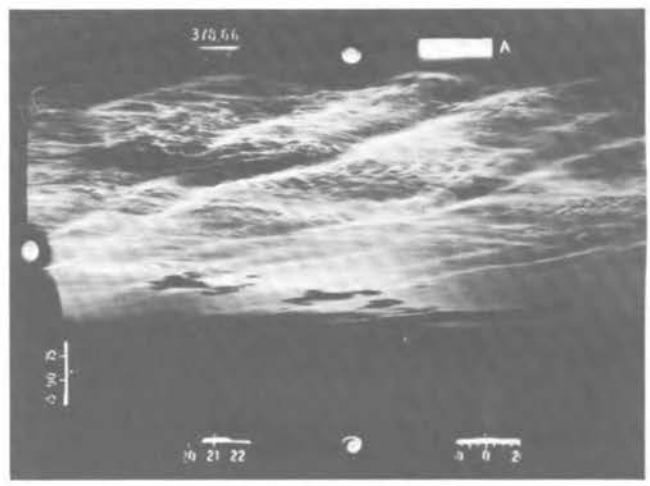


FIGURE 8.1 Noctilucent clouds at the mesopause revealing wave motions (after Witt, 1962).

their energy upward into regions of lesser density, conservation of energy requires that the wave amplitude increase to compensate. The effect is so extreme that while lower atmospheric wave energy fluxes are small compared with solar radiative energy fluxes, in the upper atmosphere the dynamical and radiative fluxes can be of comparable magnitude; more is said about this effect in Section 8.5.

The decrease in atmospheric density with height affects the wave dynamics, and indeed the fundamental structure of the upper atmosphere, in yet another way, through the associated increase in atmospheric kinematic viscosity (inversely proportional to the density) with height. Just above the tropopause this viscosity is small in the sense that the observed atmospheric motions can be considered inviscid to good approximation; not only are the motions inviscid, but they tend to engender turbulence and concomitant mixing. Turbulent, or eddy, viscosity is orders of magnitude larger than its molecular analog; the latter plays a negligible role in the observable dynamics. However, molecular kinematic viscosity increases steadily with height; at 100 km, it is some six orders of magnitude greater than its surface value. Not far above this altitude, turbulence can no longer be maintained; turbulent mixing ceases, and the different atmospheric constituents diffusively separate according to their molecular weights (Chapters 3 and 7). At the same time, the atmospheric motions become increasingly subject to viscous dissipation. For the same reason, the time constants for radiative transitions eventually become shorter than the collision times for the particles involved, so that the atmosphere can no longer be considered in local thermodynamic equilibrium. Similarly, the increasing relative plasma density introduces a new dimension to the fluid motion. Thus the same density decrease that causes the upper atmosphere to present us with a richly varied photochemistry and plasma-physics laboratory provides a diversity of fluid-dynamics regimes, ranging from inviscid and turbulent to laminar and from electrically neutral to plasma, for our scrutiny.

8.2 ROSSBY-PLANETARY WAVES, TIDES, AND GRAVITY WAVES

The upper atmosphere supports a wide variety of oscillations and wave modes, ranging on the large-scale end from the quasi-biennial oscillation of the tropical upper atmosphere and the atmospheric semiannual oscillation through Rossby-planetary waves and the tides to inertial oscillations, gravity waves, and acoustic waves with periods of a few seconds at the smallest scale. The most important of these from a fluid-dynamical standpoint are the Rossby-planetary waves, the tides, and the gravity waves. These motions are, of course, implicit in, and represent solutions to specialized forms of, the equations of motion for the upper atmospheric fluid. One form of these equations is given in Table 8.1 (for a complete and thorough development of the equations the interested reader is referred to Batchelor, 1967); they are simple in appearance but complex in implication. Three of these equations (one vector equation and two scalar equations) express conservation of momentum, energy, and matter, respectively; these are *prognostic*, in the sense that they

TABLE 8.1 Equations of Motion

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u}\cdot\nabla)\mathbf{u} + 2\rho\boldsymbol{\Omega}_E \times \mathbf{u} = -\nabla p + \rho\mathbf{g} + \mu\left[\nabla^2\mathbf{u} + \nabla(\nabla\cdot\mathbf{u})/3\right] + \rho\mathbf{F}$$

$$\frac{\partial e}{\partial t} + (\mathbf{u}\cdot\nabla)e = -p\frac{\partial \rho^{-1}}{\partial t} - p(\mathbf{u}\cdot\nabla)\rho^{-1} + \phi + \frac{1}{\rho}\nabla\cdot(K_T\nabla T) + J$$

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) = 0$$

$$p = \rho RT$$

ρ = density
$\mathbf{u}$ = velocity
$\boldsymbol{\Omega}_E$ = earth's angular rotation vector
p = pressure
$\mathbf{g}$ = acceleration of gravity
μ = coefficient of viscosity
$\mathbf{F}$ = body force/unit mass (e.g., Lorentz force, solar, lunar gravitational forces)
e = internal energy/unit mass
ϕ = rate of mechanical energy dissipation/unit mass due to viscosity
K_T = thermal conductivity
J = heat input/unit mass (e.g., radiation transfer, chemical reactions, water-phase transformations)
R = atmospheric specific gas constant
T = temperature

predict evolutions of the system resulting from its present state. The fourth—an equation of state—is *diagnostic*, in that it only states a necessary constraint relating instantaneous values of the system pressure, density, and temperature; for most purposes the ideal gas law suffices. Together with the appropriate boundary conditions, these equations are capable of describing the whole of atmospheric dynamics (exceptions arise in those cases for which the dynamics of each of the atmospheric constituents is substantially different, as is the case with ions and neutrals; then separate sets of equations must be written for each constituent). This comprehensiveness may be welcome in some respects, but it is most unwelcome in others, in that solution of the equations in their fullest generality requires more initial information than we are able to provide and would yield more information than we could assimilate. Thus the science of atmospheric dynamics involves a search for simplifications and idealizations to these equations and their boundary and initial conditions that drastically reduce the amount of input required and output retained while reproducing the essential features of the particular phenomena of interest.

A major source of such simplification derives from the fact that the different terms in the equations depend in different ways on spatial and temporal derivatives of the pressure, density, temperature, and velocity fields. The result is that the different terms of the equation may or may not be important to a given phenomenon, depending on its spatial and temporal scales. It is for this reason that the Rossby-planetary waves, tides, and gravity waves each have such a distinct character, as indicated by Tables 8.2 and 8.3.

Rossby-planetary waves are oscillations of the atmosphere having wavelengths of the order of several days or more. These motions are global in scale and have periods exceeding the period of the earth's rotation; they arise as a result of the earth's sphericity and rotation, which combine to produce a Coriolis force that increases with latitude. The inertial term in the equation of motion is small compared with the Coriolis term; hence the wave-associated motions are the result of an approximate balance between Coriolis and pressure-gradient forces, i.e., they are nearly geostrophic. Similarly, vertical accelerations are small in the sense that there is a nearly complete balance between the vertical pressure gradient and the buoyancy term; i.e., the motions are hydrostatic. Their horizontal wave phase speeds are the order of 10 m sec^{-1}; these are comparable with the mean zonal wind speed at temperate latitudes, and the sensitivity of planetary waves to and their interaction with the zonal flow is strong. Inspection of the wave dispersion equation in Table 8.3 indicates this; the wave-phase speed is always smaller than the mean zonal flow u_0 and decreases with increasing wavelength, vanishing when $k^2 = 2\Omega_e \sin \theta / u_0 R_0$; for $\theta = 45°$ and $u_0 = 10$ m sec^{-1}; this occurs for $\lambda = 2\pi/k \approx 5000$ km. Because Rossby-planetary waves produce at alternate phases of their cycles, the familiar "highs" and "lows" of weather maps, their discovery and the understanding of their motion constituted a major advance in our understanding of the weather. In the troposphere, Rossby-planetary waves contain orders of magnitude more energy than their tidal or gravity-wave counterparts; however, their theoretical study is much more recent, dating from pioneering work by Rossby in the late 1930's.

The tides, although producing only weak oscillations in the atmospheric surface pressure (fluctuations ~ 1 mbar as opposed to changes ~30–100 mbar that are associated with the Rossby-planetary waves), are strongly evident in the oceans. As a result, they attracted early theoretical interest; late in the seventeenth century Newton explained the so-called "equilibrium" tide that would develop on a nonrotating earth under the gravitational influ-

TABLE 8.2 Summary of Wave Properties

Wave	Horizontal Wavelength, λ	Period, τ	Horizontal Wave Phase Speed, λ/τ	Motion	Source Mechanisms
Rossby-planetary waves	$\sim R_E$	$\sim 5\tau_E$	$\ll R_E/\tau_E$	Geostrophic Hydrostatic	Instability Topographical forcing Differential heating
Tides	$\sim \frac{2\pi R_E}{n}, n = 1,2,\ldots$	$\sim \frac{\tau_E}{n}, n = 1,2,\ldots$	$\sim 2\pi R_E/\tau_E$	Nongeostrophic Hydrostatic	Solar, lunar gravitational fields Solar heating
Gravity waves	$\ll R_E$	$\ll \tau_E$	$< C_0$	Nongeostrophic Nonhydrostatic (except for largest τ)	Instability Topographical forcing Differential heating Wave–wave interaction

R_E = radius of the earth
τ_E = solar day
$2\pi R_E/\tau_E$ = 460 m sec^{-1}
C_0 = speed of sound (~300 m sec^{-1} at altitudes <100 km, increasing to 10^3 m sec^{-1} above)

TABLE 8.3 Wave Dispersion Equations

Rossby-planetary waves:

$$\psi = \psi_0 \exp i\,(\omega t - kx)$$

$$k^2 = \frac{2\Omega_E \sin\theta/R_E}{u_0 - \omega/k}$$

Tides:

$$\psi^{m,s} = \sum_n \Theta_n^{m,s}(\theta) \exp i\,(\omega t - s\phi - nz) \exp(z/2H)$$

$$\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\frac{\sin\theta}{\omega^2/4\Omega_E^2 - \cos^2\theta}\frac{\partial}{\partial\theta}\right) - \frac{1}{\omega^2/4\Omega_E^2 - \cos^2\theta}\left(\frac{s}{\omega/2\Omega_E}\cdot\frac{\omega^2/4\Omega_E^2 + \cos^2\theta}{\omega^2/4\Omega_E^2 - \cos^2\theta} + \frac{s^2}{\sin^2\theta}\right)\right]\Theta_n^{m,s} = -\frac{4R_E^2\Omega_E^2}{gh_n^{m,s}}\Theta_n^{m,s}$$

$$n = \left(\frac{\gamma - 1}{\gamma H h_n^{m,s}} - \frac{1}{4H^2}\right)^{1/2}$$

Gravity waves:

$$\psi = \psi_0 \exp(z/2H) \exp i\,(\omega t - ky - nz)$$

$$\omega^4 - \omega^2 C_0^2(k^2 + n^2) + (\gamma - 1)g^2k^2 - \gamma^2 g^2\omega^2/4C_0^2 = 0$$

x = zonal (east–west) coordinate
y = arbitrary horizontal coordinate
z = vertical coordinate
t = time
Ω_E = earth's angular rotation velocity
R_E = earth's radius
θ = colatitude
u_0 = mean zonal wind
ϕ = longitude
g = acceleration of gravity
$h_n^{m,s}$ = "equivalent depth"
γ = ratio of atmospheric specific heats
H = atmospheric scale height
C_0 = speed of sound

ence of the sun and the moon; his work was extended in the 1800's by Laplace, who treated the dynamic case of tides on a rotating planet. Except in the tropics, the weak tidal pressure fluctuations are masked by planetary-wave fluctuations in the daily records; thus the determination of their amplitudes has required extensive, careful statistical analysis. As this was done, it was discovered that the dominant tidal component in the surface observations was the solar semidiurnal mode. This was quite a puzzle; if the tidal forcing were predominantly gravitational, then it would appear that the lunar semidiurnal tide should be dominant, while if the forcing were primarily thermal, then it would appear that the solar diurnal mode should be larger in amplitude. In the 1880's, Thomson (later to become Lord Kelvin) postulated an explanation in terms of a resonant response of the atmosphere to the solar gravitational forcing. This theory was vulnerable to the criticism that the resonance had to be quite sharp to enhance the solar semidiurnal tide relative to the corresponding lunar mode, which should be stronger by a factor of 2 in the absence of resonance. The result was that the popularity of the theory waxed and waned for nearly 80 years, its fortunes changing with each new discovery concerning new details of the temperature structure of the upper atmosphere, to which the theory was quite sensitive; during this period some of the best minds in atmospheric science applied themselves to this problem.

As the resonance theory was finally seen to be untenable, the problem was approached from the opposite point of view, namely, that the predominant forcing was thermal and that the amplitude of the solar diurnal mode at the surface was somehow suppressed. Upper atmospheric research was an important factor in resolving the problem. In the 1950's and 1960's, researchers showed that stratospheric ozone absorption of solar radiation could account for most of the observed surface amplitude of the solar semidiurnal tide. High-altitude rocket data provided another clue when their analysis revealed that in the stratosphere and mesosphere the dominant tidal component was indeed a solar diurnal mode. The final resolution to the mystery of the tides was provided by Lindzen in the mid-1960's. Lindzen considered the set of eigen-

values of the tidal problem known as the "equivalent depths" (Table 8.3); he showed that this set was complete only if it included negative values for the square of the vertical wavenumber, corresponding to wave energy trapping in the vertical. He then showed that most of the solar energy input goes into such a vertically trapped mode. At one stroke, this explained both the dominance of the solar semidiurnal mode at the earth's surface and the dominance of the diurnal mode in the upper atmosphere (for a complete development of tidal theory as well as an interesting historical account, the reader is referred to the book by Lindzen and Chapman, 1970). The tides, like the planetary waves, are global in dimension; however, for tidal motions the inertial terms are comparable to the Coriolis terms and the motions are therefore nongeostrophic. The vertical accelerations remain small, so that the motions are hydrostatic. Wave-phase speeds are much larger than zonal wind speeds; hence the vertical propagation of the important tidal modes is only weakly affected by the mean zonal winds.

The gravity waves, the third major wave type, occur on spatial scales much smaller than the earth's radius and with periods much less than a day; as a result, the earth's sphericity and rotation play a negligible role in their description. These, together with their closely related oceanic counterparts, the wind waves, have been studied for more than a century. However, the interpretation of upper atmospheric motions in terms of gravity-wave theory effectively begins only with Hines (1960; Hines has recently published a collected works including his 1960 paper and subsequent research together with interesting annotation, 1974). Gravity waves exist because of a subtlety of the atmospheric density decrease with height; for the most part, that density decrease occurs in such a way that the atmospheric stratification is *statically stable*. Consider the motion of a parcel of air in such a region after it has been displaced upward from its equilibrium position. Finding itself denser than its surroundings, it decelerates, and sinks back to its equilibrium level, and overshoots, whereupon it finds itself warmer and lighter than its surroundings, so that it decelerates, and so on. In this way it finds itself in oscillation about its equilibrium position. It does this with a characteristic oscillation frequency N called the Brunt-Väisälä frequency after its discoverers. In the Boussinesq (incompressible) approximation, N is given in terms of the density stratification and the acceleration of gravity g by

$$N^2 = -\frac{g}{\rho}\frac{d\rho}{dz},$$

where z is the height; more properly, since the atmosphere is compressible, N^2 is expressible in terms of the potential temperature θ as

$$N^2 = -\frac{g}{\theta}\frac{d\theta}{dz}.$$

The Brunt period has a value of the order of 10 min in the atmosphere between the surface and 100 km; above that height it decreases to a value of 5 min.

The Brunt-Väisälä frequency has a critical importance for gravity waves. Examination of the wave dispersion equation of Table 8.3 shows that in an isothermal atmosphere, gravity waves with frequency $\omega < N$ are internal; the vertical wavenumber n is real, and there is vertical phase and energy propagation. Gravity waves having frequencies $\omega < N$ are evanescent; for these the vertical wavenumber n is imaginary and there is no phase or energy propagation in the vertical. (It should be noted that the dispersion equation given in Table 8.3 is a general dispersion relation encompassing both acoustic and gravity waves, as suggested by the fact that it is fourth order in ω; care must be taken to avoid confusion between the two sequences.) Gravity-wave motions are in no way geostrophic, since in their description the curvature and rotation of the earth are ignored. Waves having frequencies $\omega << N$ are nearly hydrostatic, but for those waves for which $\omega \to N$ the vertical accelerations are important enough to invalidate the hydrostatic assumption. For some gravity-wave components, the horizontal wave-phase speeds may approach the speed of sound; however, for the most part the wave-phase speeds are a few tens of meters per second or less. In an inviscid atmosphere, the wave-phase speeds, and indeed the wavelength, can in principle be arbitrarily small. It is appropriate to note at this point another property of gravity waves satisfying the Boussinesq approximation (phase speeds $V_{ph} << C_0$, vertical wavelengths $\lambda_z \ll H$, where H is the scale height); to good accuracy these obey the dispersion relation $n^2 \doteq k^2[(N/\omega)^2 - 1]$, where n is the vertical wavenumber and k the horizontal wavenumber. This is an approximate form of the general result in Table 8.3 and shows that in this limit the vertical tilt of the wave-phase fronts (related to the ratio n/k) is dependent only on ω/N. This result applies for the tides and the Rossby-planetary waves as well; it shows that for all waves having periods much greater than the Brunt period, the wavefronts are nearly horizontal. For a review of the theory of gravity waves in the atmosphere, the reader is referred to the book by Gossard and Hooke (1975).

8.3 FACTORS COMPLICATING WAVE PROPAGATION IN THE UPPER ATMOSPHERE

In the preceding section we reduced the physics of each of the three wave types to its barest essentials; however, to apply the wave theory successfully to interpretation of the upper atmospheric observations we must incorporate a number of complicating factors in our models. These complications are many and varied, but for the most part they may be thought of as producing (a) wave refraction, (b) wave coupling, or (c) wave dissipation. Since most of these complications have proved sufficiently severe to defy comprehensive analysis, and since

observational data user requirements are becoming increasingly demanding, investigations of the effect of these processes on wave propagation constitute an active area of current research.

Most familiar of the refractive influences is the temperature structure of the upper atmosphere, specifically that associated with the mean height structure: the stratospheric temperature increase, the mesospheric temperature decrease, and the thermospheric temperature increase with height. These produce both wave refraction and partial or total wave energy trapping. The equations governing the wave motion in an isothermal atmosphere have constant coefficients; introduction of the vertical atmospheric temperature structure introduces height-variable coefficients. The mathematical difficulties thus encountered are usually surmounted by (a) WKB or ray-tracing methods, which assume that the temperature varies slowly over a wavelength, so that the dispersion equations of Table 8.3 hold locally; (b) use of multilayer models in which the temperature is constant within each layer and matching of boundary-conditions determines the wave solutions at layer interfaces; (c) analytical solutions for special profiles; and (d) a variety of numerical techniques. All illustrate the wave refraction, trapping, and partial ducting; they serve to identify the trapped modes and in many cases lead to good comparison with observation.

Wind structure of the upper atmosphere introduces analogous complications, producing wave refraction and trapping, and its effects can be handled in exactly the same way as the effects of the temperature structure, with one important exception. That exception occurs in the immediate vicinity of so-called critical levels, which mark heights at which the component of the background wind speed in the direction of wave propagation matches the horizontal wave-phase speed. Dispersion equations incorporating the effects of atmospheric winds all reveal singularities at such heights—singularities associated with the fact that at such heights the wave frequency measured by an observer moving with the mean wind vanishes. Careful analyses show this to have a number of interesting implications. Inspection of the group velocity of wave packets in the WKB approximation reveals that the packets approach such critical levels asymptotically. Full-wave calculations, providing matching across the critical level, tend to confirm the WKB picture, showing that for slowly varying shear flows the wave energy penetration through such critical levels is very slight. Furthermore, the WKB approach indicates that in the vicinity of a critical level, the vertical wavelength becomes vanishingly small while amplitudes of wave-associated fluctuations in horizontal velocity and shear tend to infinity; this implies that the usual linearized, inviscid models of the wave motions must break down there. Critical-level encounters are consequently quite difficult to treat in the models. It is apparent from our earlier discussion of wave phase speeds that both Rossby-planetary waves and gravity waves, with their low phase speeds, are quite susceptible to critical-level encounters as they propagate through the upper atmosphere; by contrast most tidal components are relatively immune.

Two other refractive effects deserve passing mention. Above about 100 km, the atmosphere ceases to be well mixed; instead it approaches a state of diffusive separation in which lighter elements and molecules tend to overlie heavier ones; as a result the atmosphere scale height increases with height quite independently of the thermospheric temperature increase. (In addition to the refractive effects, diffusive separation produces dissipative effects, which are discussed separately below.) Similarly, the $1/r^2$ decrease of gravity with radial distance r from the center of the earth also influences the scale height and the Brunt-Väisälä frequency. The result is some contribution to wave refraction; however, this effect is slight compared with those listed above.

A second set of complications arises from the fact that the decrease in atmospheric density with height generally causes the various wave modes to grow in amplitude with height. For plane waves, constancy of vertical energy flux implies constancy of the quantity $\rho u'^2$, where ρ is the atmospheric density and u' is the wave-associated perturbation velocity. Since ρ varies as $\exp(-z/H)$, u' must vary as $\exp(z/2H)$. Thus, in the absence of wave reflection and energy dissipation, wave amplitude increases exponentially in the upper atmosphere, by orders of magnitude between the tropopause and the base of the thermosphere. Thus even the smallest-amplitude wave motions in the troposphere tend to violate the linearization assumption at some point in the upper atmosphere; at this point the effect of the nonlinearities must be taken into account explicitly. This is done only with difficulty.

The simplest result of the nonlinearity is a self-interaction of the wave; this turns out to be weak, because the wave motions considered here happen to be nearly incompressible, but it does result in steepening of the waves into shocks. In addition to the self-interaction produced by nonlinearity, there arises coupling between different wave modes. In its simplest form, this manifests itself because small-scale, short-period waves find themselves propagating in an atmosphere in which the dominant temperature and wind structure is not that associated with the mean but rather that associated with larger-scale, longer-period waves of large amplitude. In its more complex form, coupling arises between different wave modes that happen to form resonant triads, satisfying the conditions

$$\mathbf{k}_3 = \mathbf{k}_1 \pm \mathbf{k}_2$$

and

$$\omega_3 = \omega_1 \pm \omega_2,$$

where the $\mathbf{k}$'s represent the wave vectors of three waves, while the ω's represent the corresponding wave frequencies, and the ω's and $\mathbf{k}$'s individually satisfy the dispersion

equation. Higher-order resonances are also possible, as are forced couplings of wave energy into modes that do not satisfy the dispersion equation. There is a continuous interchange of wave energy among modes in this way. This energy exchange is more rapid, the larger the amplitudes of the wave involved. Considerable study has been devoted to the generation of gravity waves by the atmospheric tides through this means. At altitudes the order of 100 km, tidal-associated density perturbations become a significant fraction of the ambient, while tidal velocity fluctuations become a substantial fraction of the speed of sound. Above this height, tidal-wave amplitudes cease their growth, while gravity-wave amplitudes increase; there is considerable interest in the question of whether wave–wave interaction (as opposed, for example, to simple tidal dissipation by viscosity) is responsible for this.

Associated with the increase of wave amplitude with height is the increase potential for wave-induced instability. On the small scale, the relevant stability is that in the vertical; this is characterized by the so-called Richardson number R_i, defined as

$$R_i \equiv \frac{N^2}{(du/dz)^2};$$

here du/dz is the vertical shear of the horizontal wind u. Stability theory shows that shear flows are dynamically stable to infinitesimal perturbations as long as $R_i > \frac{1}{4}$. However, wave-associated velocity and temperature perturbations may become so large in the upper atmosphere as to cause R_i to fall below $\frac{1}{4}$ locally, inducing dynamic instability, or even cause R_i to become negative, inducing convective instability (in this extreme case the atmospheric density structure is not even statically stable). This results in the generation of small-scale motions at a rate much faster than that allowed by conventional wave–wave interaction. The result may be a cascade of wave energy into smaller and smaller scales and eventually into molecular dissipation; at the same time, the unstable modes produce enhanced diffusion and transport of energy in such a way as to limit the growth of the originally unstable mode. The problem has only been treated successfully numerically, and very little is currently known about wave dynamics once wave amplitudes reach unstable levels. It has been suggested, however, that nonlinear, unstable wave fields may account for much of atmospheric diffusion and transport.

In addition to nonlinearities, wave–wave coupling, and instability, which may be thought of in one sense as processes tending to limit wave amplitudes, there are a number of other dissipative processes operative in the upper atmosphere. Radiative damping is a major source of wave energy dissipation; this complication is sufficiently taxing that it is usually entered into the analysis not in a general way but through a parameterization. The most common approach is to add a term linear in the wave-associated temperature perturbation, and of the opposite sign; this is the so-called Newtonian cooling. Analyses of the effect of this cooling show that it is of major importance in damping Rossby-planetary waves in the stratosphere and mesosphere.

Molecular relaxation also produces wave energy damping. This process arises as a result of the failure of a molecular gas to equipartition energy among translational, vibrational, and rotational states of the molecules in a time short compared with a wave period. At the earth's surface, the relevant time constant is the order of 10^{-5} sec; but because the time constant is inversely proportional to the pressure, this process becomes relevant to gravity waves at altitudes the order of 100 km or so. Only preliminary analyses have been carried out on its precise effects, but it appears that molecular relaxation could be a major source of wave damping in the lower thermosphere (above that height the dominant constituents are atomic and the process no longer so important).

Atmospheric viscosity and thermal conductivity have a major influence on the upper atmospheric wave physics; their influence on wave propagation increases with height, and, when it becomes significant, it introduces radical changes in the wave behavior. This may be seen from the equations of motion of Table 8.1, which show that introduction of the viscous and thermal-conduction terms produces equations with nonconstant coefficients, in addition to raising the order of the differential equation. The result is that exact analytic solutions no longer obtain. Several methods exist for sidestepping this difficulty. The first involves parameterization of the viscous effects by the substitution of a term that is simply linearly proportional to the wave-associated velocity perturbations and oppositely directed. This is the so-called Rayleigh friction; it is useful in the sense that it removes a number of mathematical difficulties, but it is deficient in the respect that it suppresses the scale-dependence of the viscous effects, which are greatest for the smallest-scale waves. A second method of solution involves arbitrarily requiring that the viscosity μ and the thermal conductivity K_T vary with height in the way necessary to reduce the equations to equations with constant coefficients. This approach is most useful in analyzing the propagation of waves having vertical wavelengths small compared with a scale height, the scale of variation of the viscosity, and the thermal conductivity. Analysis of the wave propagation in this case shows that the lowest-order effect of viscosity and thermal conduction is wave damping, but that higher-order effects involve refraction as well. In addition, the effect of raising the order of the differential equations is to introduce new modes of oscillation: viscous (or shear) waves and thermal conduction waves. A third approach is numerical. Analyses of this type show that the atmosphere can be thought of as consisting of four regions as far as wave motions are concerned. The lowermost region is a boundary layer in contact with the earth's surface; viscous and thermal conduction modes coexist with the other wave modes here in order to satisfy boundary conditions at the surface. Above this relatively thin layer there exists a region in which viscosity and

thermal conductivity are relatively unimportant; in this region the wave motion is dominated by the Rossby-planetary waves, tides, and gravity waves. With increasing height, the viscous and thermal conduction terms increase in importance, until finally a height region is reached in which there is significant coupling between the nondissipative modes and the former; above this height region is a fourth, in which only the viscous and thermal conduction modes are important. This is a highly idealized picture, but it does appear to reflect processes at work in the thermosphere, where theoretical analyses indicate the mode conversion to be substantial. Such analyses permit an examination of the competing effects of dissipation and wave amplitude growth with height to determine whether the waves in question will ever become nonlinear before they are damped by viscosity.

From the tropopause to the mesopause, the molecular kinematic viscosity and thermal conductivity are relatively small; in particular their effects are small relative to the effects of small-scale waves and turbulence in dissipating energy of the larger-scale modes through either wave–wave interaction or instability. The effects of these interactions on the waves are parameterized by effective "eddy" viscosities and thermal conductivities that are orders of magnitude larger than the molecular values. It is these that produce the mixing and vertical transports in this region of the atmosphere; their distribution with height is uncertain, and indeed any attempt to make these distributions precise would necessarily be artificial, but they can be taken to be roughly constant with altitude. To the extent that the parameterization is valid, then, the approximation of constant coefficients should also be valid. However, at altitudes not far above 100 km, molecular viscosity and thermal conductivity have become so large that they suppress turbulence and the related mixing. In this region, the Reynolds number of the flow is small but sufficiently large for the development of turbulence; nevertheless, molecular viscosity and thermal conductivity hold full sway and grow exponentially with height.

The photochemistry of the upper atmosphere may play some role in wave dynamics. This occurs by wave generation through absorption of solar radiation and the heat release associated with it. Wave-associated density and temperature changes modify the photochemistry and produce certain phase-related heat release and absorption; if this heat release is of the correct phase, then instability results and there may be wave growth. Heat release of the opposite phase may work to suppress wave amplitudes. This is known to happen in the lower atmosphere as the result of water-vapor condensation; the effects of analogous photochemical reactions in the upper atmosphere have been only cursorily treated. Much more work needs to be done in this research area, which requires a synthesis of dynamics and photochemistry. Diffusive separation of molecular constituents above 100 km also produces wave energy loss, because the different constituents respond to the impressed wave motion in different ways. Thus the different species tend to move with different velocities, and wave energy loss results from collisions between molecules of the different species.

A final factor complicating the propagation of Rossby-planetary waves, tides, and gravity waves in the upper atmosphere is the increasing concentration of ionized particles relative to neutral particles with increasing height. The relevant quantity here is the neutral–ion collision frequency—the frequency with which neutrals experience collisions with ions. In the E region, this collision frequency is of the order of 1 day^{-1}, so that for motions of Rossby-planetary wave or tidal period, the effects of ion drag are comparable with inertial effects here. In the F region, near the ionization density maximum, the collision frequency rises to about one collision each 30 min. Thus here, even waves of gravity-wave period experience significant ion drag. In addition, the introduction of charged particles and the earth's magnetic field into the equations of motion introduces new modes of oscillation—mixed hydromagnetic modes that couple with the incident wave energy.

8.4 WAVE SOURCES

In the previous sections, we took for granted the existence of the waves of interest here and discussed the basic aspects of their physics and propagation and modification of this propagation by various complicating factors. However, for waves to exist, they must be generated, and in this section we treat the manner of their generation.

Rossby-planetary waves are primarily generated by three mechanisms: instability of the mean zonal flow, topographical forcing, and differential heating. The relevant instabilities are of two types: the barotropic instability, which is analogous to the shear-flow instability of a homogeneous fluid, and baroclinic instability, which results in a statically stably stratified shear flow in the presence of a north–south temperature gradient. For a further description of these instabilities, the reader is referred to the introductory text by Holton (1972). Planetary waves generated by such instabilities have relatively short wavelengths of several thousand kilometers. The topographical forcing results from the vertical velocity component imposed on the zonal airflow over the continents. Similarly, the differential heating arises from the difference in the heat input to the atmosphere encountered over the continents and over the oceans. Both the topographical forcing and the differential heating result in wavelengths of the order of half a global circumference at 45° latitude, i.e., about 2×10^4 km.

The tides, as their name implies, are generated by lunar and solar gravitational forcing and by solar heating. The dispersion equation shown in Table 8.3 represents free tidal oscillations; in analyses of the tidal forcing, inhomogeneous terms appear in the vertical structure equation, and the response of the atmosphere to these is

determined. As mentioned in the previous section, it is found that the predominant forcing is thermal; the major diurnal thermal mode is trapped, while that of the semidiurnal mode is propagating.

Sources of gravity-wave excitation are many and varied. Gravity waves, like the Rossby-planetary wave, may be generated by the shearing instability of larger-scale flows for which the Richardson number drops below the critical value of ¼. This is probably the single most important source mechanism for such wave motions, since wave generation by shear instability is the process effective in limiting the intensification of large-scale circulations. The waves may also be generated in similar ways by the airflow over irregular and unevenly heated topography. They may also be generated by photochemical destabilization. In addition, gravity waves are generated in the auroral zone by the heat and momentum input associated with variations in energetic particle precipitation. We mentioned in the previous section that many waves owe their existence to wave–wave interactions, involving either other gravity-wave modes or the interaction of tidal modes. The distinction between gravity waves and turbulence may be artificial (the latter representing a highly nonlinear, interactive, and broad spectrum of the former), but to the extent that the distinction has some value, gravity waves may also be considered to be generated by turbulent fields. The process is analogous to that of the generation of acoustic waves by turbulent fields. However, there is a considerable difference in the efficiency of the two processes. Calculations made of acoustic-wave generation by low-Mach-number turbulence that assume that the turbulence is unaffected by the wave generation confirm that the generation is relatively weak and inefficient, while similar calculations of gravity-wave generation by turbulent fields show that the power output per unit volume is infinite. This shows that the original assumption that the turbulence is unaffected by the wave generation is invalid but follows physically from the fact that acoustic waves, unlike the turbulent eddies, are longitudinal motions propagating with the speed of sound, while the gravity waves are qualitatively similar to the turbulence in that they are rotational and may have arbitrarily small phase speeds. Finally, gravity waves may be launched by penetrative convection of air parcels from convectively unstable regions of the atmosphere (those for which $N^2 < 0$) into stable regions.

It should be recognized in reviewing this catalog of sources that from the standpoint of those of us interested in the upper atmosphere, they can be cataloged in another way, i.e., as arising locally or in the lower atmosphere. Much has been made of this fact, but it cannot be stressed too often. All else being equal, atmospheric wave sources below the height of observation will tend to dominate because of the decrease of atmospheric density with height and the concomitant increase in wave amplitude. For the three wave types considered here, phase propagation and energy propagation have oppositely directed vertical components, so that waves that are observed to have a downward phase propagation are associated with upward energy propagation. Such waves do indeed dominate the observations of every level in the upper atmosphere.

8.5 WAVE DYNAMICS

The study of atmospheric wave physics has progressed substantially beyond the point of a mere analysis of wave propagation characteristics to the study of wave dynamics and its role in determining the atmospheric motion field (and indeed atmospheric structure). At issue is the handover of energy and momentum from the mean atmospheric flow to the waves during wave generation, the transport of this energy and momentum by the waves as they propagate, and the return of that energy and momentum to the mean flow accompanying the wave damping. Wave-dynamical processes are important even in the troposphere, but they assume additional importance in the upper atmosphere because of the decrease in atmospheric density with height and the concomitant increase in wave amplitudes required by energy conservation. If we consider the upward propagation of energy from a Rossby-planetary wave whose amplitude at the tropopause is ~10 m sec^{-1}, at the 100-km level the corresponding wave amplitude would be ~10^4 m sec^{-1} if the wave propagates its energy vertically. This is a factor of 10 greater than the molecular thermal velocity at these heights and would result in thermospheric temperature of the order of 10^5 K, some two orders of magnitude greater than that observed. The earth's gravitational field would be too weak to prevent a rapid boiloff of such a hot atmosphere. Explaining why this process does not occur is therefore a matter of considerable interest. Research on this subject reveals that Rossby-planetary wave energy is prevented from reaching the thermosphere either by the effects of atmospheric zonal winds, which, depending on the season (there is a seasonal wind reversal), trap the wave energy by causing the waves to encounter a critical level or levels below 50 km (Charney and Drazin, 1961) or by the effects of radiative damping (Dickinson, 1969). Even at these lower altitudes the effect of the waves on atmospheric temperatures and mean flows is substantial. Stratospheric warmings may have their origin in dynamic processes of this sort.

In contrast to the Rossby-planetary waves, the tides are only slightly affected by atmospheric temperature and wind structure and are thus efficient in propagating their energy vertically upward. These have amplitudes negligible compared with the planetary waves at the earth's surface, but at thermospheric heights the tidal energy flux is nevertheless substantial. Lindzen and Blake (1970) estimate that tidal heating of the upper atmosphere is sufficient by itself to maintain the temperature of the thermosphere at some 400 K; this is about one third of the observed value and suggests that this dynamical heat flux is the same order of magnitude as the solar EUV heat flux

responsible for the strong day–night thermospheric temperature variation.

Similarly, gravity waves provide substantial heat input into the upper atmosphere. Hines (the relevant papers are contained in his 1974 volume) estimated that gravity waves should provide heat inputs comparable to the tidal flux, but the exact contribution is difficult to assess because of uncertainties about the variability of this flux with latitude, season, time of day, and other parameters. Hines went so far as to suggest that gravity-wave heating might account for the observed warm temperature of the polar wintertime mesopause, an explanation that remains intriguing to this day.

Associated with the large wave-energy fluxes just described there are substantial momentum fluxes as well. Hines has estimated that the tidal fluxes to the upper atmosphere should be so large as to provide accelerations of the mean zonal wind of the order of 20 m sec^{-1} day^{-1}. The gravity waves, if all propagating in the azimuthal direction (an admittedly unlikely assumption), should lead to even larger local thermospheric accelerations of ~70 m sec^{-1} h^{-1}. Fluxes of this magnitude could account for the observed upper-thermosphere superrotation and certain other anomalies following magnetic substorms.

The quasi-biennial oscillation, a unique feature of the tropical stratosphere, is now believed to be the result of the interaction between gravity waves and the semiannual oscillation. The former encounter critical levels created by the latter and deposit their momentum and energy there, changing the wind profile in such a way as to lower the height of the critical level at subsequent times. It might be noted that within a few degrees of the equator, gravity waves may have periods of the order of days before Coriolis effects become important. Waves of such long period are internal within a narrow equatorial region; outside that latitude belt they are trapped (by the increased value of the Coriolis parameter). It is these long waves, termed Kelvin and Yanai waves depending on their symmetry properties, that are relevant to the quasi-biennial oscillation. Wallace (1973) reviews this subject and gives pertinent references.

The major wave types discussed here also produce important dynamical effects in the ionized upper atmosphere. These range from tidal driving of the so-called *Sq* current system producing magnetic fluctuations observed at ground level to gravity-wave production of traveling ionospheric disturbances and ionospheric storms. These topics are considered in other chapters.

8.6 THE UPPER ATMOSPHERE AS A FLUID-DYNAMICS LABORATORY

The upper atmosphere is an interesting astrogeophysical fluid-dynamics laboratory, for several reasons. First and foremost, it exhibits the full range of wave phenomena we wish to study for astrogeophysical application, including the Rossby-planetary waves, tides, and gravity waves, which we have considered in detail, as well as a number of wave types that for reasons of space we could not discuss—inertial oscillations and acoustic waves being among them. The earth's upper atmosphere mirrors all the observed tropospheric wave motions but usually with greater amplitudes so that they are easier to see. Wave–wave interactions and wave-associated instabilities mirror analogous processes in the oceans. The earth's thermosphere is truly a "geocorona" in the sense that the dynamical heating provided by the tides and gravity waves provides a close analog to the heating of the solar corona by wave processes. The density stratification of the upper atmosphere allows us to study these phenomena under a wide range of conditions from small-amplitude to highly nonlinear, from inviscid and turbulent to laminar, and from electrically neutral to ionized. And it permits us to study these phenomena on a geophysical scale free from the severe compromises in scaling imposed by laboratory models.

In addition, the upper atmosphere facilitates the study of fluid dynamics by providing tracers of the neutral dynamics. These range from constituents with unique optical properties such as ozone, to airflow chemistry for revealing gravity-wave motions, to ionization for revealing the motions of the neutral atmosphere on all scales.

REFERENCES

Batchelor, G. K. (1967). *An Introduction to Fluid Dynamics*, Cambridge U. Press, New York.

Charney, J. G., and P. G. Drazin (1961). Propagation of planetary-scale disturbances from the lower and the upper atmosphere, *J. Geophys. Res. 66*, 83.

Dickinson, R. E. (1969). Vertical propagation of planetary Rossby waves through an atmospheric with Newtonian cooling, *J. Geophys. Res. 74*, 929.

Gossard, E. E., and W. H. Hooke (1975). *Waves in the Atmosphere*, Elsevier, New York.

Hines, C. O. (1960). Internal atmospheric gravity waves at ionospheric heights, *Can. J. Phys. 38*, 1441.

Hines, C. O., *et al.* (1974). *The Upper Atmosphere in Motion*, American Geophysical Union, Washington, D.C.

Holton, J. R. (1972). *An Introduction to Dynamic Meteorology*, Academic Press, New York.

Lindzen, R. S., and D. Blake (1970). Mean heating of the thermosphere by tides, *J. Geophys. Res. 75*, 6868.

Lindzen, R. S., and S. Chapman (1970). *Atmospheric Tides*, D. Reidel, Dordrecht, Holland.

Wallace, J. M. (1973). General circulation of the tropical lower stratosphere, *Rev. Geophys. Space Phys. 11*, 191.

Witt, G. (1962). Height structure, and displacements of noctilucent clouds, *Tellus 14*, 1.

9 Transport in the Stratosphere

EDWIN F. DANIELSEN
National Center for Atmospheric Research

JEAN-FRANCOIS LOUIS
National Oceanic and Atmospheric Administration

9.1 INTRODUCTION

We present here a physical description of transport as it applies to the stratosphere, troposphere, mesosphere, or to any portion of the atmosphere. It involves the reversible processes of translation, rotation, and deformation and the irreversible process of molecular diffusion. All organized motions with length scales from global dimensions to submillimeter contribute to the reversible transport. Their relative importance to translation or to deformation depends on the ratio of the length scales that describe the velocities and the quantity being transported.

In principle, only the random motions of the molecules produce diffusion, but in practice only the larger scales of motion are known or predictable. Therefore, we must also consider and discuss a mathematical–statistical description of transport. The latter introduces a time–space averaging to define the mean (reversible) motions. All motions whose scales are smaller than those of the averaging volume remain interdeterminant. Their effects on the transport are determined from statistical correlations between deviations from the means.

In effect, the transport due to a range of scales extending from the molecular to that of the averaging volume are rendered irreversible because of our inability to resolve or predict these scales of motion. As will be shown, this range can vary from small-scale limits imposed by the observations to the large-scale limits of the spherical earth.

Located between the troposphere and mesosphere, the stratosphere is a thin, spherical layer whose horizontal dimensions are about 1000 times its vertical dimension. Given this strong asymmetry, it is reasonable to expect that the motions and the transport will be predominantly horizontal. In fact, the vertical velocities are small—too small to be observed or measured by conventional methods. Although the vertical motions are small, their effects on the vertical transport are not small and must be included. One reason for their importance is related simply to the similarity between length and speed scales. Horizontal-to-vertical lengths and horizontal-to-vertical speeds are both about 1000:1, so typical transit times are comparable when velocities of the same sign are sustained. Another reason is related to the intermittent but systematic mass exchange between the lower strato-

sphere and the troposphere. As air descends into the lower stratosphere, it is occasionally transported directly into the troposphere by a phenomenon known as tropopause folding.

Tropopause folding, i.e., a three-dimensional folding of the boundary between the troposphere and stratosphere, reminds us that the troposphere and stratosphere form a strongly coupled system. The coupling, particularly strong beneath the 23–25 km level, is caused by a broad spectrum of propagating and interacting waves, many of which amplify in the troposphere but dampen as they propagate into the stratosphere. These waves are gravity-modified shear, inertial, and Rossby waves, and therefore are mainly transverse waves. For example, as the waves propagate toward the east, the air parcels oscillate north–south on stream surfaces that are slightly inclined in the north–south direction. The inclinations are small, so the corresponding vertical velocities are small; and, as indicated above, these inclined stream surfaces and wavefronts are important to stratospheric transport.

The waves affect not only the velocities and the transport; they affect also the thermal structure and, therefore, the quantity normally used to distinguish the stratosphere from the troposphere. It is convenient to express this quantity in terms of a derived temperature θ, called potential temperature, rather than the actual temperature T.

$$\theta = T\left(\frac{1000}{p}\right)^{0.286}, \tag{9.1}$$

where the air pressure p is expressed in millibars. An advantage of θ is that it is independent of compression and hence is analogous to temperature in a liquid. Another advanatage of using θ rather than T to identify an air parcel is that θ is conserved in isentropic transport. To a reasonable first approximation, we can consider θ as a stationary coordinate to compute 12- to 24-h trajectories.

The quantity used to distinguish the stratosphere, troposphere, and mesosphere is based on the vertical derivative of θ or on an equivalent quantity σ, called the stability, which is defined by

$$\sigma = \frac{g}{\theta}\frac{\partial\theta}{\partial z} = N^2. \tag{9.2}$$

Qualitatively, the stability of the stratosphere is much larger than that of the troposphere and mesosphere. (The prefix "strato" implies a suppression of vertical motions and of vertical mixing.) The stability σ refers to a parcel of air displaced vertically and is also called the hydrostatic stability. If σ is positive a displaced air parcel will oscillate vertically with a frequency N, the so-called Brunt-Väisälä frequency. Frictional forces will dampen this oscillation, so the air parcel will tend to return to its undisturbed position. If σ is negative, as it sometimes is in the troposphere, the Brunt-Väisälä frequency will be imaginary, and the displacements will amplify and lead to active vertical mixing. (The prefix "tropo," which means turning, implies that vertical mixing is active in the troposphere.)

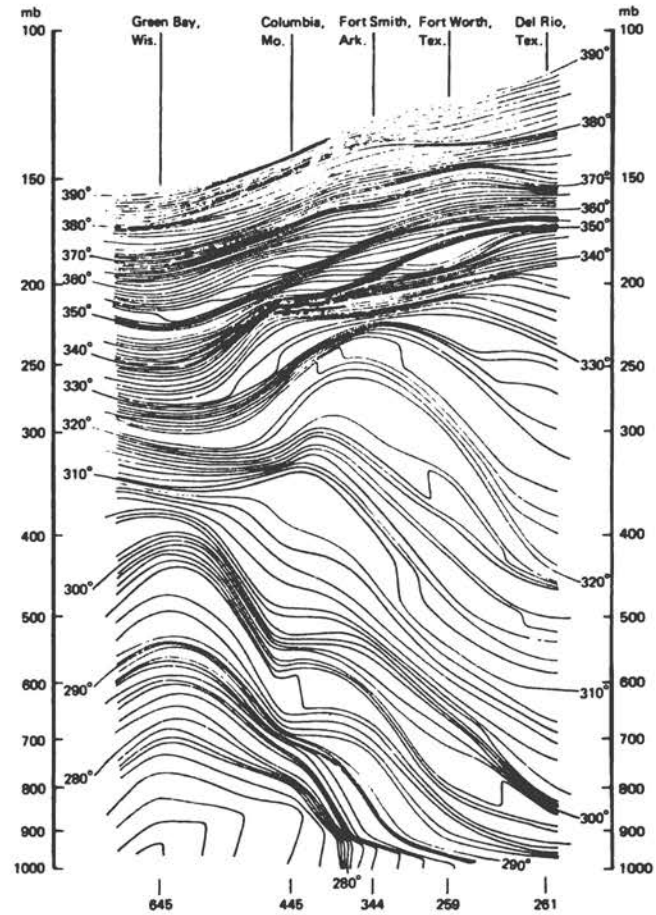


FIGURE 9.1 Detailed vertical cross section from Green Bay, Wisconsin, to Del Rio, Texas, 1500 GMT, March 29, 1956.

If these terms, troposphere and stratosphere, are to have any meaning, they must be identified with mean values of σ not with actual point values. The reason for this generalization returns us to the effects produced by the spectrum of atmospheric waves. The divergence associated with the waves increases, and the convergence decreases, the value of σ. It is, therefore, almost impossible to distinguish the stratosphere from the troposphere when the effects of all waves are included. Such an example is presented in Figure 9.1, based on a complete retrieval of all temperature data observed on the original radiosonde records.

The laminar structure of the actual atmosphere, due to the ever-present waves, complicates and tends to obscure the simpler, large-scale organization. Figure 9.2 shows the simplified tropospheric–stratospheric structure that emerges when the smaller-scale waves are removed by filtering the radiosonde observations (Danielsen, 1959).

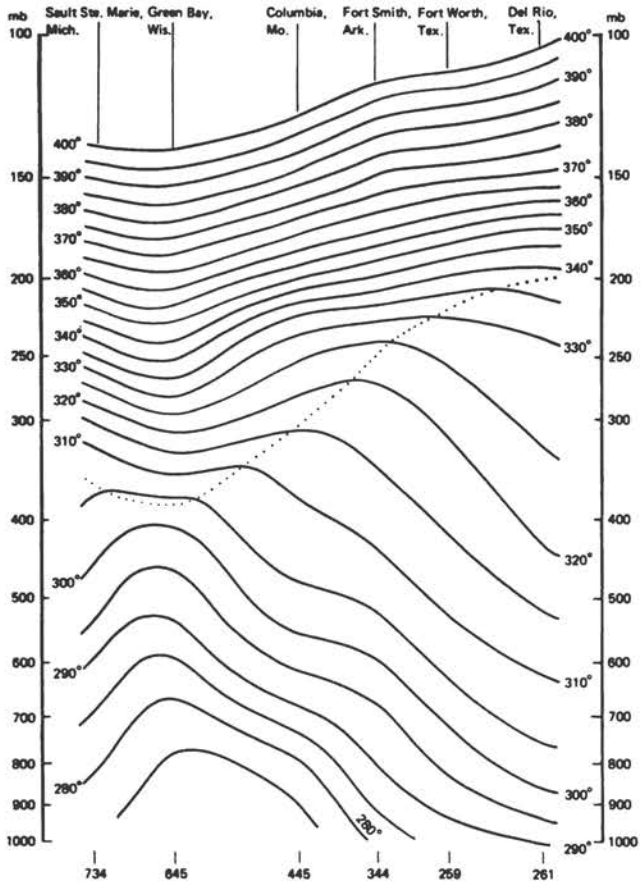


FIGURE 9.2 Macroscale thermal structure. Smoothed vertical cross section from Sault Ste. Marie, Michigan, to Del Rio, Texas, 1500 GMT, March 29, 1956.

Therefore, the concept of a troposphere and a stratosphere and of a boundary between them—the tropopause—depends on filtered data, that is, on large-scale mean quantities. In general, the mean stability $\overline{\sigma}$ is much larger in the stratosphere than in the troposphere. However, the existence of the smaller-scale waves implies that vertical mixing is not completely suppressed even in the stratosphere. As seen in Figure 9.1, regions where σ is small or slightly negative can be produced in the stratosphere by large-amplitude waves. Some of these waves are generated by mountains, others by imbalances in the momentum and pressure fields. Vigorous mixing can result where σ is negative. The phenomenon of clear air turbulence in the stratosphere attests to its effects on high-flying aircraft.

We must incorporate these vertical mixing events when we compute the long-range transport of a trace constituent. Each event is anisentropic, so if we compute isentropic trajectories we must, of necessity, underestimate the diffusion or the dispersive transport. Nevertheless, the isentropic trajectory method enables us to obtain a first approximation to the actual three-dimensional transport. The net 12-h vertical displacement

$$z(12) - z(0) = \int_0^{12\,\mathrm{h}} \left(\frac{\mathrm{d}z}{\mathrm{d}t}\bigg|_\theta + \frac{\partial}{\partial\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t} \right) \mathrm{d}t \qquad (9.3)$$

includes the diabatic heating rate or the increase in entropy due to mixing. These rates can be estimated where they are sufficiently systematic to be significant. To obtain a first approximation, we neglect $\mathrm{d}\theta/\mathrm{d}t$ and determine Δz from the corresponding heights of the θ surface at 0 and 12 h.

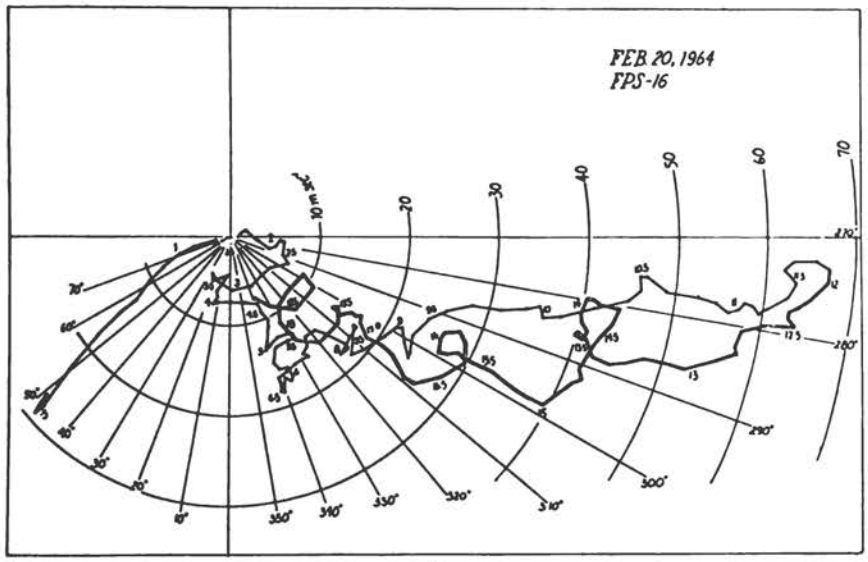
On the basis of many trajectories computed in the lower stratosphere, we know that the motions in the large-scale waves in the lower stratosphere slope downward toward the north in the northern hemisphere. This slope implies that kinetic energy is being converted to available potential energy. The latter is reduced by horizontal gradients of radiational cooling. It follows that the major mixing surfaces also slope downward toward the north in the lower stratosphere. This negative slope reverses to positive in the upper troposphere. That is, as air mixes northward in the lower stratosphere, it also mixes downward, toward the tropopause. Then, when it enters the troposphere, it tends to mix southward and downward, toward the earth's surface.

Additional evidence of the three-dimensional transport in the atmosphere can be obtained from the numerical simulations of general circulation models and from the observed dispersion of radioactive isotopes from nuclear explosions in the stratosphere. The former is limited by the representativeness of the numerical solutions, and the latter by the sparseness of aircraft or balloon measurements. Although the isentropic trajectory method, the numerical simulations, and the radioactive tracer measurements each have deficiencies, the combined evidence permits one to describe the transport phenomenon with reasonable accuracy from a physical and statistical viewpoint.

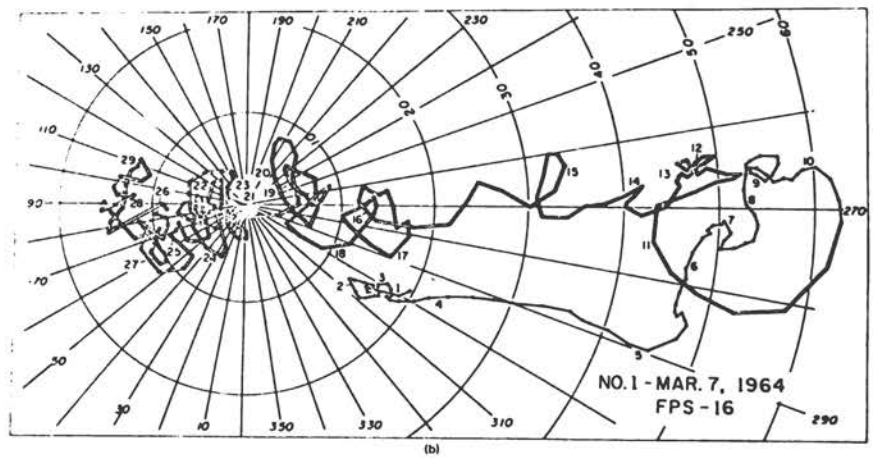
9.2 PHYSICAL DESCRIPTION OF TRANSPORT

The compressible atmosphere with its vertical and horizontal gradients of potential temperature or potential density surrounding a spherical, rotating earth is free to oscillate in a variety of limiting and mixed modes. The limiting modes include sound, internal gravity, inertial, and Rossby waves, whose periods increase from fractions of a second to several days. (See Chapter 8 for more detailed information on the wave modes.) The mixed modes, such as gravity-inertial and gravity-modified Rossby waves, are also prevalent, and each and all of these waves affects the transport of a tracer in the stratosphere.

With conventional data, such as those supplied twice



(a)



(b)

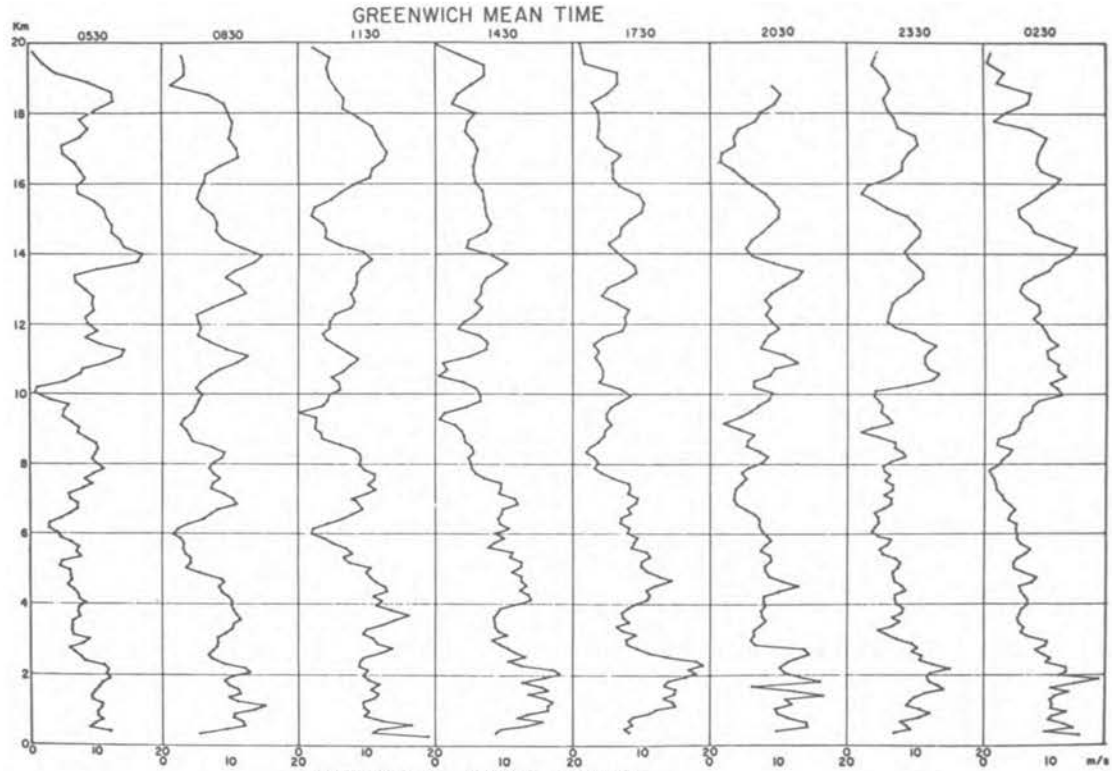
FIGURE 9.3 (a) and (b) Detailed hodographs derived from precision FPS-16 radar tracking of radiosonde balloons. Heights are plotted every 0.5 km, speeds are in meters per second.

daily by the radiosonde balloon ascents, we can only resolve the horizontal velocities associated with waves whose horizontal wavelengths are greater than about 1000 km. However, from aircraft equipped with inertial navigation systems and from radar tracking of radiosonde balloons we know that the smaller-scale modes are ever present.

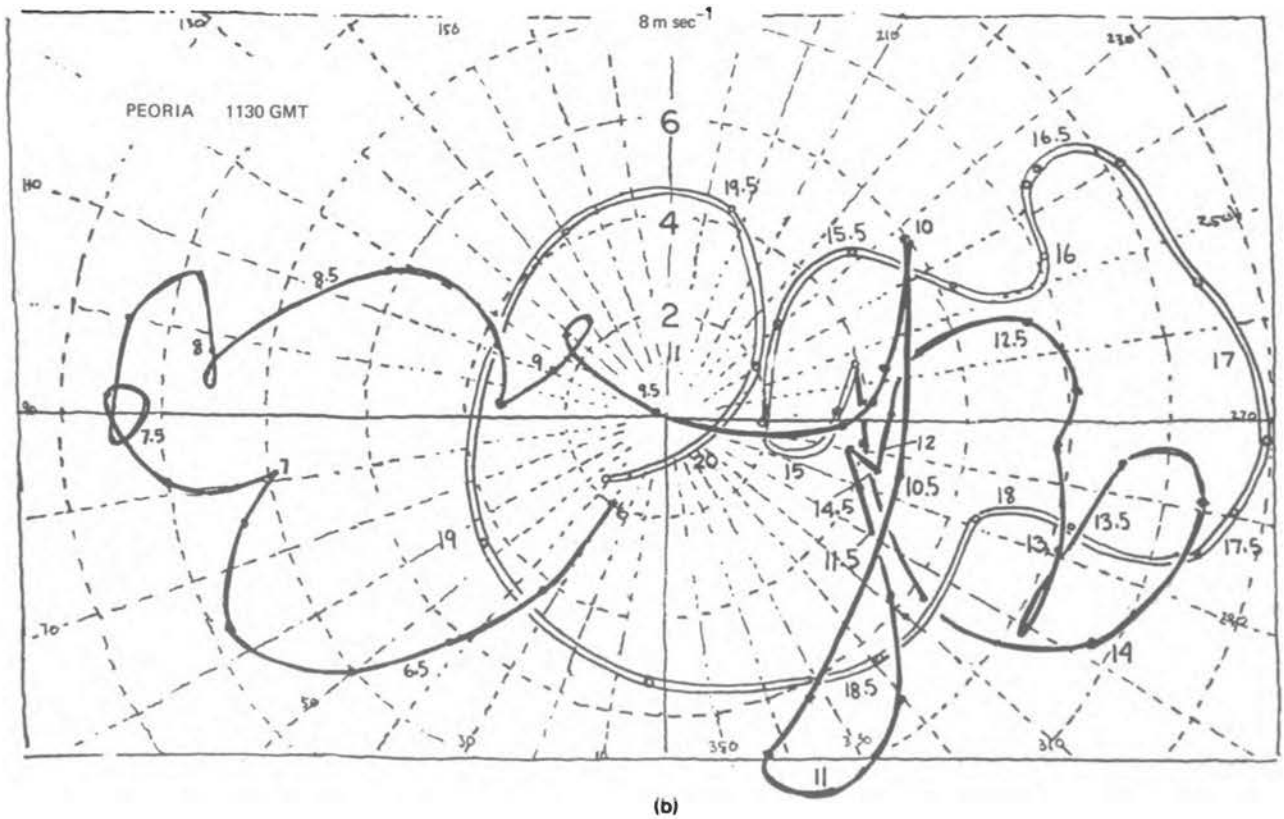
Figures 9.3(a) and 9.3(b) show how these smaller-scale waves affect the vertical profile of the horizontal wind vector by imposing undulations or loops on the hodograph. The total wind vector at a given height is the vector from the origin to the point on the line or hodograph. For example, the wind vector at 12 km in Figure 9.3(a) is from 274° and the speed is 66 m sec^{-1}. Note the predominance of anticyclonic loops in the stratosphere (above the maximum wind speed) in both figures, especially the large loop from 9.5 to 11.5 km in Figure 9.3(b). The amplitude of this loop is 10 m sec^{-1}. It is produced by adding an anticyclonically (clockwise) rotating vector to the mean wind. An internal gravity wave is linearly polarized and produces undulations not loops. The latter are produced by inertial waves, gravity-inertial waves, and some shear-gravity waves.

The effect of adding the small-scale vectors to the larger-scale winds is to change both the direction and speed of the total winds. When the large-scale winds are strong, the percentage variations are small; but when the large-scale winds are light, the hodograph and the wind-speed profile can be dominated by the small waves. An example of the 3-h changes in the wind speed profile caused by vertically propagating waves (1–4 km vertical wave lengths) is shown in Figure 9.4(a). Note the downward phase propagation, which implies an upward or

FIGURE 9.4 (a) Vertical profiles of wind speed at Peoria, Illinois, from 3-h serial ascents. Note apparent negative phase propagation. (b) Hodograph for Peoria corresponding to the speed profile at 1130 GMT. Note presence of both linearly and circularly-polarized velocity perturbations. Anticyclonic rotation is consistent with negative vertical phase propagation.


GREENWICH MEAN TIME
0530
0830
1130
1430
1730
2030
2330
0230
Km
m/s
WIND SPEED - PEORIA, ILLINOIS

(a)


8 m sec⁻¹
PEORIA 1130 GMT

(b)

positive vertical energy flux for gravity-inertial waves. Figure 9.4(b) indicates that these speed changes with height are due to predominantly anticyclonic loops in the hodograph.

In general, the regularly processed radiosonde winds do not resolve the velocity perturbations produced by waves with 1- to 3-km vertical wavelengths, and, even if they could, the approximately 300-km spacings between radiosondes would make it difficult or impossible to resolve their horizontal wavelengths. Despite these obvious limitations, we can assess the relative importance of all wavelengths to the three-dimensional transport by separating three effects:

1. When the scale of the velocity gradient is much larger than the scale of a tracer's distribution, the tracer moves as a unit with little or no distortion about its center of mass.
2. When the scales are comparable, the distribution is distorted or deformed relative to the center of mass.
3. When the velocity scales are much smaller than the tracer scales, the velocity fluctuations and random molecular motions dilute the concentrations of the tracer by irreversible mixing with the surrounding air.

The first effect is a translation of the center of mass with no significant change in area or volume occupied by the tracer. The second effect is an increase in its area with no significant change in its volume. The third effect is a systematic increase in its volume. The first two effects are, in principle, reversible. The last is definitely irreversible. Effect 2 contributes to effect 3 by systematically increasing the surface area, thereby exposing more and more of the tracer to the small-scale irreversible mixing. Using these three effects as abstractions we can apply them first to tracers like radioactive isotopes, which are produced or injected into a small volume by nuclear explosions, and then to tracers like ozone, which are photochemically produced in a large part of the stratosphere.

9.3 TRANSPORT OF RADIOACTIVE ISOTOPES FROM POINT INJECTIONS

The transport of tracers injected into the stratosphere depends on the altitude, latitude, longitude, and season of the injection. These differences are basically caused by direct or indirect thermal forcing in both the troposphere and the stratosphere, but we can characterize them by different zonal wavenumbers and different mass exchanges.

Below 21 km, there is considerable energy in zonal wavenumbers 5 to 20 and the mass exchange between the stratosphere and troposphere is predominantly intrahemispheric, although there is some transport across the equator. In response to strong forcing by and direct coupling with the troposphere, there is an intermittent but systematic flow from the troposphere into the stratosphere at low latitudes and from the stratosphere into the troposphere at middle and high latitudes.

The center of mass of a tracer injected into the lower stratosphere will describe a complicated, undulating, or looping path with occasional large meridional excursions. Its surface area will rapidly increase and small-scale mixing will dilute its concentrations. Also, when it reaches the lower stratosphere at high latitudes, it can be transported into the troposphere in the descending flow west of amplifying troughs or vortices (cyclonic storms). Its residence time in the stratosphere can vary from a few days to more than a year.

The intermittent transport from the lower stratosphere to the troposphere is associated with a folding of the tropopause. This phenomenon, first described by Reed (1955) and Reed and Danielsen (1959), has been well documented and verified by aircraft measurements of ozone and radioactivity (Danielsen 1964, 1968; Danielsen *et al.*, 1970.)

When viewed in vertical cross section, from the side, as in Figure 9.5, the folding produces an inclined layer of stratospheric air that can be identified by its rich ozone concentration, its radioactivity (natural or man-made), and its potential vorticity (a meteorological scalar whose large values are generated in the stratosphere). When viewed from above, the stratospheric air entering the troposphere fans out as it descends to the west and south of each major cyclone. The anticyclonic branch of the flow descends behind the main cold front, where it is mixed down to the surface in the turbulent boundary layer.

Above 35 km, the flow is dominated by zonal wavenumbers 0–4, and interhemispheric mass exchange is important. There is direct seasonal coupling between the northern and southern hemispheres. In response to the solar heating in the summer hemisphere and the radiational cooling in the winter hemisphere, a warm anticyclone dominates the summer hemisphere, and mass flows across the equator, spiraling into a cold cyclone in the winter hemisphere. The predominantly zonal flows are illustrated schematically in Figure 9.6. Anticyclonic summer circulations are quite stable and tend to be centered at the pole. The wintertime cyclonic circulations increase in intensity and become unstable, resulting in periodic breakdowns or transformations from predominantly zonal flow (wavenumber 0) to flow with important meridional components (wavenumbers 1, 2, etc.) with cyclonic flow usually over northern Europe and Asia and anticyclonic flow over Alaska. Examples of these two characteristic flow patterns are illustrated in Figures 9.7(a) and 9.7(b). These analyses, limited to the western half of the northern hemisphere, are based on a sparse network of rocketsonde data. They were prepared by the Upper Air Branch, National Meteorological Center (ESSA, 1969).

Tracers injected into the upper stratosphere will be transported into the winter hemisphere, where they will circulate rapidly in the strong, cyclonic flow. They will

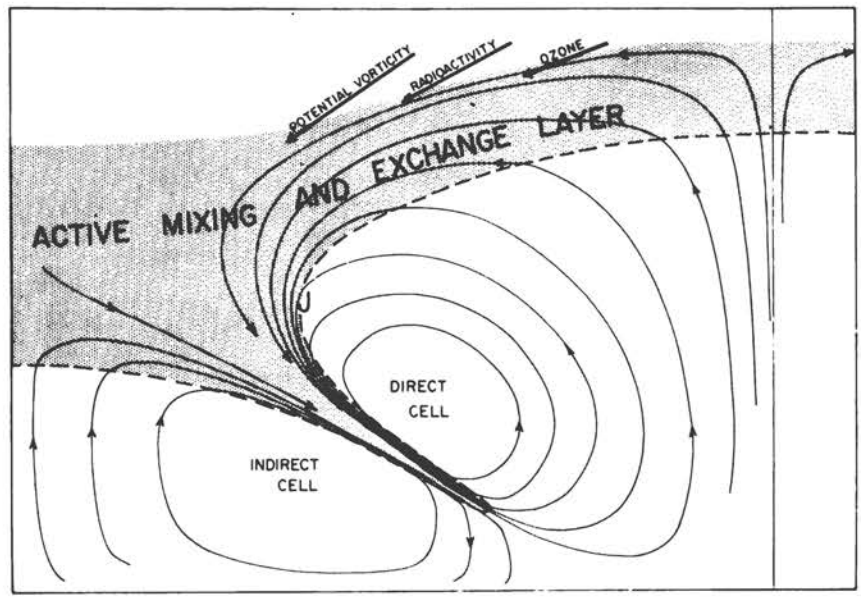


FIGURE 9.5 Vertical cross section of tropopause folding. The shaded zone corresponds to stratospheric air in the lower stratosphere due to the stratospheric extrusion.

also descend at high latitudes in the winter hemisphere. Because of the large speed gradients normal to the flow, the tracer's distribution will be stretched out along the direction of the flow, while smaller-scale diffusion spreads the tracer normal to the flow. If the tracer is injected into the cyclonic zonal flow, its center of gravity will make several circumpolar transits, while the tracer spreads into and fills the vortex. During the transition from wavenumber 0 to 1 or 2, portions of the tracer will be rapidly transported to lower latitudes and the center of gravity of the tracer will shift toward Europe. Diffusion into the anticyclonic circulation over Alaska is impeded, so large, longitudinal variations are produced. The residence time for these upper stratospheric injections is several years.

Direct thermal forcing in the middle stratosphere, 21–35 km, is a minimum, so one would expect little interhemispheric flow. What evidence there is from bomb experiments and constant-volume balloon trajectories tends to support this expectation. The equatorial flow is predominantly zonal, reversing from easterlies to westerlies with a quasi-biennial period. A tracer injected in the middle stratosphere near the equator will tend to make several transits around the globe with only small portions occasionally "peeling off" to higher latitudes. This behavior in the midstratosphere is distinctly different from the behavior of a similar injection into the upper or lower stratosphere, where the cross-equatorial flow during solstice is significantly larger. During the transition from westerlies to easterlies, or vice versa, a tracer injected over several kilometers of depth can also be rapidly extruded longitudinally by the large vertical shear of the zonal winds. This phenomenon occurred during the 1958 tungsten tracer experiments when the lower stratospheric debris moved from the west, while the middle stratospheric debris moved around the globe from the east. Also, the northward and southward excursions of tungsten-185 from the equatorial zone were much larger in the lower stratosphere than they were in the middle stratosphere, so the axes of maximum concentration sloped downward both to the north and to the south of the equator.

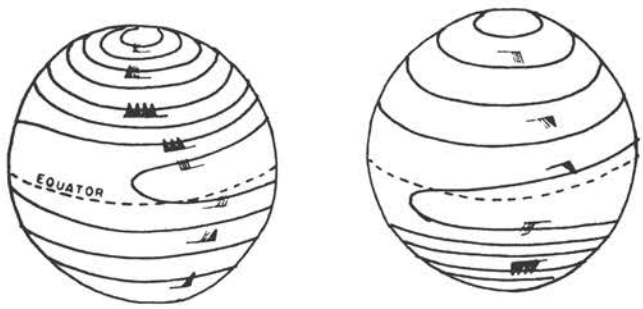
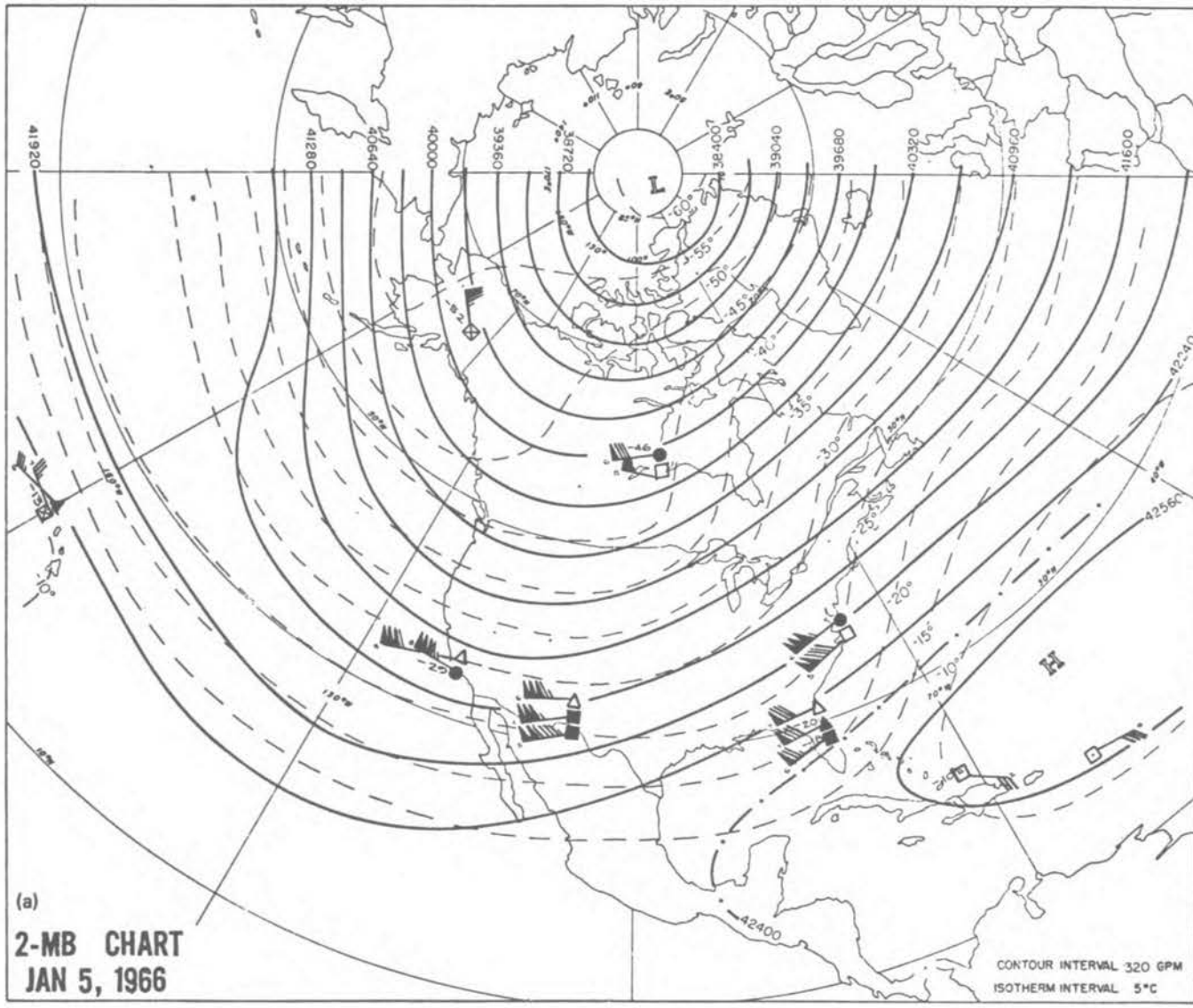


FIGURE 9.6 Schematic of upper-stratospheric flow showing winter cyclonic vortex, summer anticyclonic vortex, and cross-equatorial flow.

9.4 MATHEMATICAL–STATISTICAL CONCEPT OF STRATOSPHERIC TRANSPORT

Up to the present, measurements of trace constituents in the stratosphere have not been made routinely nor at uniformly distributed points. Probably the most extensive measurements have been made of ozone. There are not sufficient measurements of any trace constituent to de-



termine its four-dimensional (space–time) distribution in the stratosphere. However, on the basis of aircraft and balloon measurements of ozone (Danielsen *et al.*, 1970), it has been demonstrated that the ozone mixing ratio is positively correlated with the potential vorticity in the lower stratosphere. Because the potential vorticity can be directly computed from large-scale meteorological analyses, we can quite accurately estimate the ozone mixing ratio distribution from the potential vorticity plus a few direct measurements that establish the proportionality factor.

Unfortunately, as shown by Hering (1965), the positive correlation does not extend into the middle stratosphere. Above 21 km, the largest ozone mixing ratios shift to low latitudes and the largest values of potential vorticity shift toward high latitudes. These shifts are in response to their distinctly different latitudinal sources. The fact that the ozone mixing ratio and the potential vorticity are positively correlated in the lower stratosphere despite their different sources is direct evidence of the dominant importance of mixing processes and indirect evidence of tropospheric forcing in the lower stratosphere.

In general, the lack of sufficient observations makes it necessary to reduce the number of independent variables by space–time averaging. The reduction to only one spatial dimension, height z, by integrating over all latitudes and longitudes is a gross oversimplification, which is strictly applicable only when photochemical processes predominate and three-dimensional transport is negligible. The difficulty is that all the transport must be expressed in terms of a vertical diffusion coefficient and this overrestricts the atmospheric degrees of freedom.

The simplest physically reasonable reduction that can be applied to the lower, middle, and upper stratosphere is

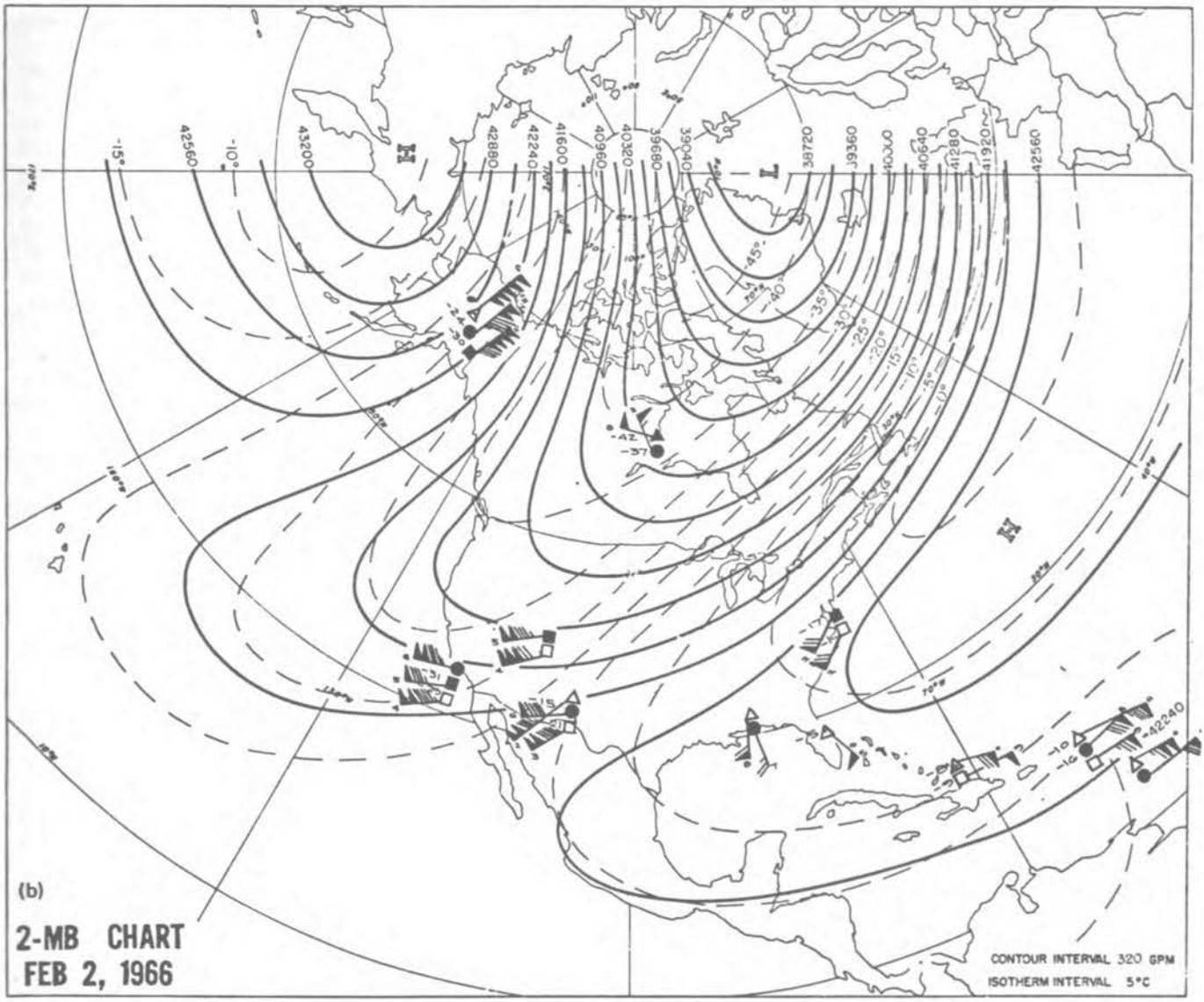


FIGURE 9.7 Examples of flow at 2-mbar level (~40 km height) determined from rocketsonde data; (a) predominantly zonal, wavenumber zero; (b) asymmetrical flow, wavenumber one. Each barb represents 10 knots, black triangle represents 50 knots.

obtained by averaging an observed property χ over all longitudes and time interval τ according to the following equation:

$$\overline{\chi}(\phi, z, \tau) = \int_0^{\tau}\int_0^{2\pi} \rho\chi(\phi, \lambda, z, t)\,d\lambda\,dt \left[\int_0^{\tau}\int_0^{2\pi} \rho(\phi, \lambda, z, t)\,d\lambda\,dt\right]^{-1}, \quad (9.4)$$

where ϕ is latitude, λ is longitude, ρ is atmospheric density, z is altitude, and t is time. In some cases, the air density weighting is not included, in which case the ρ in Eq. (9.4) is replaced by unity. It is also customary to set τ equal to three months and to consider four seasonal means.

The advantages of analyzing or predicting average values $\overline{\chi}$ rather than instantaneous values χ are that $\overline{\chi}$ requires fewer observations to be representative, is more slowly varying, and requires significantly fewer computations and computer time. The major disadvantage is that the deterministic, reversible transport produced by all the resolvable wave motions must be represented by statistical, irreversible transports of turbulent eddies. We stress this

distinction by referring to this type of imagined turbulence as "mathematical" turbulence rather than physical turbulence. The portion of the wave spectrum that is designated as turbulence depends directly on the limits of the integrals in Eq. (9.4). If, as indicated in Eq. (9.4.), the limits on longitude λ are 0 to 2π, all transport produced by zonal wavenumbers greater than zero must be considered as turbulent transports. The inner integral in Eq. (9.4) then corresponds to $\overline{\chi}(\phi,z,t)$ associated with zonal wavenumber zero, as determined from a double Fourier analysis.

An analysis of $\overline{\chi}(\phi,z,t)$ at constant time can be made from the Fourier component, zonal wavenumber zero, at several heights. This analysis, independent of λ, is a meridional-height cross section of $\overline{\chi}(\phi,z,t)$. If $\overline{\chi}$ varies with time, and generally it does, then when the time integration is completed, $\overline{\chi}(\phi,z,\tau)$ is stationary for a season and the fluctuations in $\overline{\chi}(\phi,z,t)$ must also be considered as turbulent fluctuations.

In three-dimensional numerical prediction models, a similar distinction between nonturbulent and turbulent motions must be made. The averaging integrals are then over space (x, y, z) and time, and the limits are determined by the grid spacings and time interval of integration. In general, all waves whose wavelengths are shorter than two grid spaces cannot be resolved, and their effects on $\overline{\chi}$ must be computed from statistical formulations that relate the effects of correlations between deviations from the mean to gradients or functions of the mean variables.

Because the grid intervals and, therefore, the limits of averaging, vary from one model to another, the distinction between nonturbulent, deterministic transport and turbulent transport will also vary. The method used to account for the effects of the "turbulent" portion of the wave spectrum also varies from modeler to modeler. To these variations imposed by the grid and subjectively chosen diffusion formulations must be added another variation imposed by spurious diffusion caused by the truncation errors in the finite difference method used by the modeler. If one considers all these variations, it is obvious that each model is unique in respect to its implicit and explicit "turbulence," and each must be tuned against observations. There is no set of diffusion coefficients that is universal in its applicability for three-dimensional modeling.

At present, if a three-dimensional, general circulation model (GCM) is used to predict the distribution of a trace constituent, Eq. (9.4) must be used to average the many predictions, because only $\overline{\chi}(\phi,z,\tau)$ data are available for verification. Reliable GCM models are rare and costly to run, so the reality of data limitations and operational costs forces us to focus on the two-dimensional (ϕ,z) model and mathematical turbulence.

Because the effects on the transport of $\overline{\chi}(\phi,z,\tau)$ by the complete spectrum of waves must be evaluated from the statistics of the correlations between deviations from the means, we shall now discuss the characteristics of the correlations associated with the large waves. These waves, with the largest deviations from the mean, dominate the "turbulent" transport in the two-dimensional model.

The large waves are modified gravity waves and, therefore, modified transverse waves. With short vertical-to-horizontal wavelength ratios, the wavefronts and the velocities are quasi-horizontal, but the correlations between the vertical and meridional velocity deviations from the mean $\overline{w'v'}$ are nonzero. v' and w' are the deviations from the mean wind component toward the north and in the vertical, respectively, so $\overline{w'v'}$ indicates the correlation between upward and northward motions. In the stratosphere, $\overline{w'v'}$ is predominantly negative. This negative correlation is well established from isentropic trajectories and from the observed transport of trace constituents. It indicates that as stratospheric air moves northward in the northern hemisphere, it also descends to lower elevations, and when it moves southward it ascends. The effect of these inclined stream surfaces on the transport of a trace constituent depends on the slope of the stream surface relative to the slope of the isolines and the gradient of the trace constituent. For a lucid explanation of the phenomenon, the reader is referred to the article by Reed and German (1965).

In the lower stratosphere, the negative slope of the stream surface often exceeds the negative slope of the isoline of the tracer's mixing ratio $\overline{\chi}$. Therefore, if $\partial\overline{\chi}/\partial z$ and $\partial\overline{\chi}/\partial y$ are positive, the horizontal flux $\overline{\rho_0}\,\overline{v'\chi'}$ is positive, i.e., it is up the gradient of the mean (z is the vertical coordinate and y the horizontal coordinate toward the north in the northern hemisphere). This phenomenon is usually called a countergradient transport. It refers only to the horizontal transport. In the ϕ,z plane the transport is actually downgradient. If, as discussed below in Eqs. (9.8) and (9.9), the flux vector

$$\overline{\rho}\,\overline{v'\chi'}\,\mathbf{j} + \overline{\rho}\,\overline{w'\chi'}\,\mathbf{k} = -\mathbf{K}\cdot\nabla\overline{\chi} \tag{9.5}$$

is assumed to be proportional to the negative of the gradient of $\overline{\chi}$, then there are three independent coefficients of the diffusion tensor $\mathbf{K}$: K_{yy}, $K_{yz}(=K_{zy})$, and K_{zz}. Representative magnitudes of these terms are illustrated in Figure 9.8 (Danielsen, 1974). The values of $\mathbf{K}$ increase with the time or space scales from the lower molecular limit to their global, wave-turbulent limits. The latter apply to the two-dimensional, meridional-height models. K_{yy} has the maximum, K_{zz} the minimum, and K_{yz} intermediate magnitudes. Also, K_{yz} can be positive or negative. In the northern hemisphere it is usually negative in the lower stratosphere, positive in the troposphere.

Appropriate $\mathbf{K}$ values for most two-dimensional models are within the ranges indicated by the shaded bands in Figure 9.8. Because the mean (i.e., associated with mean meridional wind) and turbulent transports usually oppose each other but sometimes reinforce each other, the actual values to be used cannot be selected independently of the mean circulations. Mahlman (1973) has used the predictions of a general circulation model to demonstrate that the net transport is critically determined by the residual

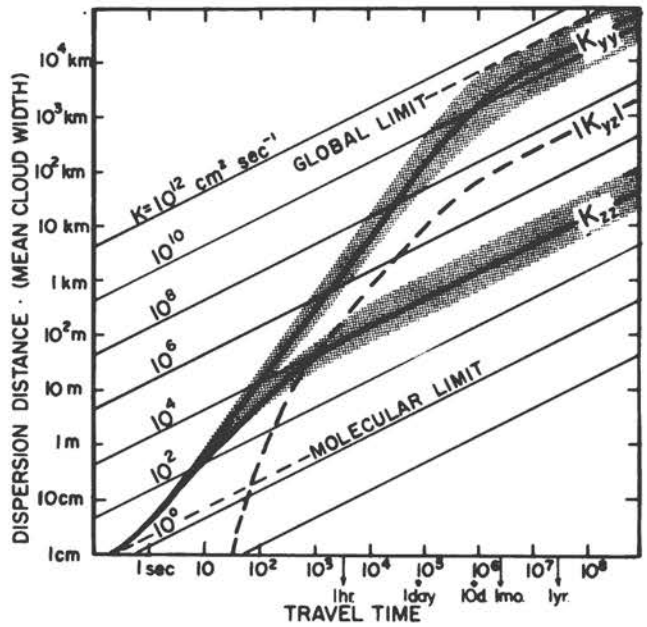


FIGURE 9.8 Coefficients of turbulent diffusion tensor as a function of dispersion distance and travel time. The range in values for a given travel time is given by the shaded area. The dashed line represents the upper bound for K_{yz}.

of the mean and turbulent transports. There is no alternative to testing or tuning the model by using either ozone or radioactive tracers from nuclear bomb tests.

9.5 MODELING THE STRATOSPHERIC TRANSPORT

The relative advantages and disadvantages of one-, two- and three-dimensional models, discussed in the previous section, are subject to each individual's subjective appraisal. Although the one-dimensional models have been used with considerable success in conjunction with photochemical models to compute the vertical distributions of various trace constituents in the upper stratosphere, the details of transport in the middle and lower stratosphere and from the stratosphere to the troposphere cannot be accurately described by a simple vertical diffusion coefficient. The obvious choice for generality and accuracy is a three-dimensional model, but the complexities and cost of introducing photochemical computations into a three-dimensional model are currently prohibitive. Therefore, we will limit our discussion in this section to a two-dimensional model, which we believe to be a logical compromise for combined photochemical and transport computations.

In two dimensions, the equation of continuity for a tracer is written

$$\frac{\partial}{\partial t}(\bar{\rho}\bar{\chi}) + \nabla \cdot \bar{\rho}\bar{\mathbf{V}}\bar{\chi} + \nabla \cdot \bar{\rho}\,\overline{\mathbf{V}'\chi'} = \bar{S}. \tag{9.6}$$

χ represents the tracer mixing ratio, $\mathbf{V}$ the velocity of the air, ρ the air density, and S the rate of change of the tracer concentration due to *in situ* sources or sinks. ∇ is the two-dimensional "del" operator. The over-bar means that the quality is averaged over longitude and time, as in Eq. (9.4), and the primed quantities are the deviations from this average.

The second term in Eq. (9.6) is the advection by the mean meridional circulation, and the third term represents the effect of waves of all scales. Both terms are important, and their effects often oppose each other, so the answer is often determined by their residual. As mentioned earlier, most of the energy is contained in waves or eddies of low wavenumber, hence the third term in Eq. (9.6) is usually called the large-scale eddy-diffusion term. This term is an unknown in Eq. (9.6), and a first-order closure is normally used, relating the eddy term to the gradient of the mean value of the tracer's mixing ratio through a diffusion tensor $\mathbf{K}$:

$$\overline{\mathbf{V}'\chi'} = -\mathbf{K} \cdot \nabla\bar{\chi}. \tag{9.7}$$

Since the large-scale eddies are not isotropic, $\mathbf{K}$ must be a second-rank tensor

$$\mathbf{K} = \begin{pmatrix} K_{yy} & K_{yz} \\ K_{zy} & K_{zz} \end{pmatrix}. \tag{9.8}$$

y is the horizontal coordinate, positive northward, and z the vertical coordinate, positive upward. $\mathbf{K}$ is symmetric with $K_{yz} = K_{zy}$. Equation (9.6) can then be written

$$\frac{\partial}{\partial t}(\bar{\rho}\bar{\chi}) = -\nabla \cdot \bar{\rho}\bar{\mathbf{V}}\bar{\chi} + \nabla \cdot (\bar{\rho}\mathbf{K} \cdot \nabla\bar{\chi}) + \bar{S}. \tag{9.9}$$

The advection and diffusion terms in Eq. (9.9) can be combined either as an "effective advection" term

$$\frac{\partial}{\partial t}(\bar{\rho}\bar{\chi}) = -\nabla \cdot \bar{\rho}\bar{\mathbf{V}}^{\circ}\bar{\chi} + \bar{S}, \tag{9.10}$$

where

$$\mathbf{V}^{\circ} = \bar{\mathbf{V}} - (\mathbf{K} \cdot \nabla\bar{\chi})\bar{\chi}^{-1}, \tag{9.10'}$$

or, using the fact that

$$\nabla \cdot \bar{\rho}\bar{\mathbf{V}} \simeq 0, \tag{9.11}$$

as an "effective diffusion" term

$$\frac{\partial}{\partial t}(\bar{\rho}\bar{\chi}) = \nabla \cdot (\bar{\rho}\mathbf{K}^{\circ} \cdot \nabla\bar{\chi}) + \bar{S}, \tag{9.12}$$

where

$$\mathbf{K} = \begin{pmatrix} K_{yy} & K_{yz} - \phi/\bar{\rho} \\ K_{zy} + \phi/\bar{\rho} & K_{zz} \end{pmatrix}, \tag{9.12'}$$

and ϕ is the stream function for $\bar{\rho}\bar{\mathbf{V}}$, defined as follows:

$$\rho\mathbf{V} = \nabla \times (\phi\mathbf{i}),$$

$\mathbf{i}$ being the unit vector perpendicular to the meridional plane. The two formulations (9.10) and (9.12) are equivalent to Eq. (9.9). Both formulations appear in the literature, but if Eq. (9.12) is used, $\mathbf{K}$ is no longer a symmetric tensor.

In order to solve Eq. (9.9) for $\bar{\chi}$, the coefficients of the diffusion tensor $\mathbf{K}$ must be determined. This is the main difficulty in two-dimensional modeling. It cannot be determined directly from observations because there is no passive tracer whose distribution is known with enough resolution to compute $\overline{\mathbf{V}'\chi'}$. Furthermore, the vertical component $\bar{w}$ of the wind velocity is not observed. However, using the observed distribution of ozone, which is the best-documented tracer, and some reasonable assumptions, it is possible to deduce a set of diffusion coefficients that simulate rather well the transport of a trace constituent in the atmosphere.

Let us consider the lower stratosphere, where ozone can be considered a passive tracer and where the effects of motions rather than photochemistry dominate. As mentioned earlier, the net horizontal flux of ozone in this region is upgradient, i.e., it is directed from the lower latitudes toward the polar regions, even though the maximum concentration of ozone is in the polar regions and the upper branch of the midlatitude Ferrel cell of the mean circulation tends to carry ozone toward the equator, as shown in Figure 9.9 for the December–February time period.

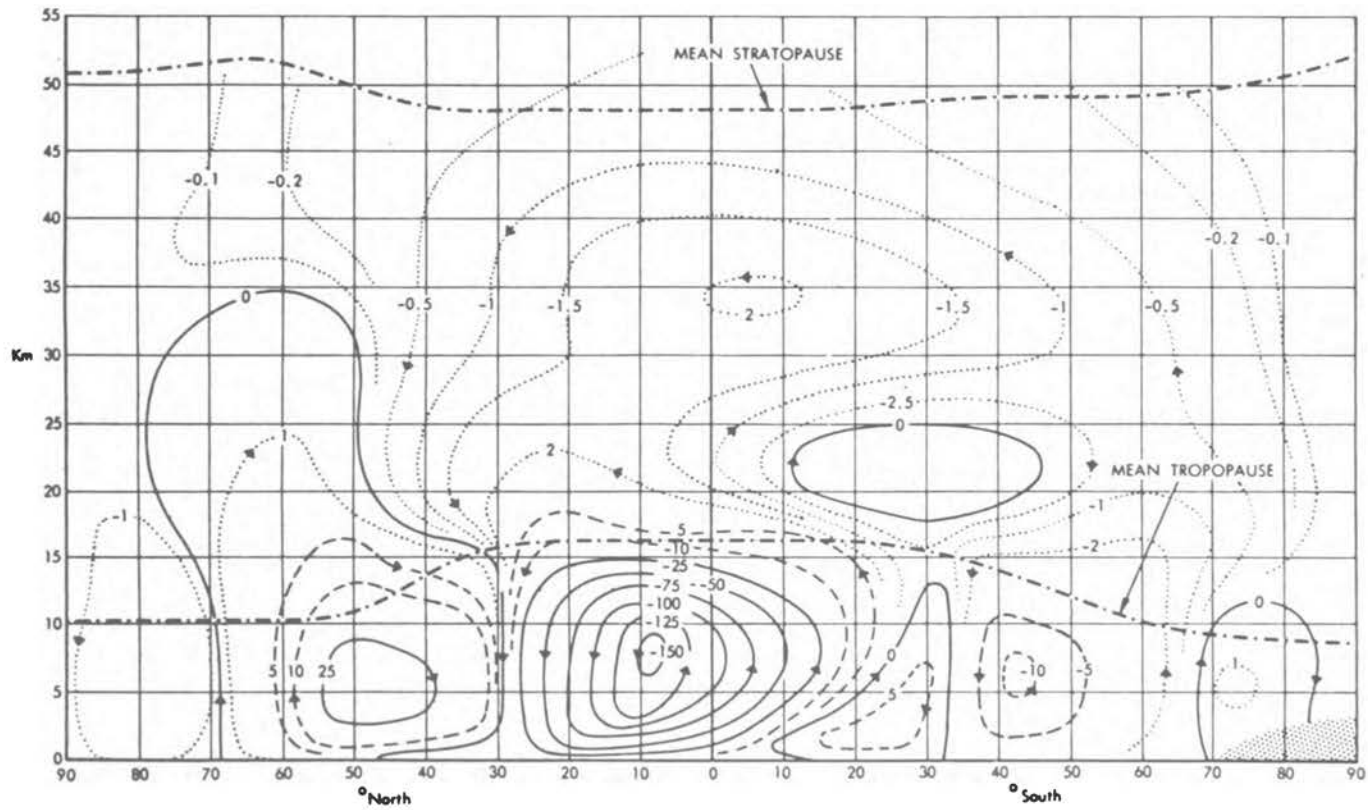


FIGURE 9.9 Mean meridional circulation, mass flow in units of 10^{12} g/sec, December–February.

The expression for the horizontal flux of ozone is

$$F_y = \bar{\rho}\left(\bar{v}\,\bar{\chi} - K_{yy}\frac{\partial\bar{\chi}}{\partial y} - K_{yz}\frac{\partial\bar{\chi}}{\partial z}\right). \qquad (9.13)$$

Assuming that, in the principal axes system, the diffusion is downgradient, then K_{yy} must be positive. In order for F_y to be positive (in the northern hemisphere), the third term on the right-hand side of Equation (9.13) must be the largest one of the three, with K_{yz} negative. This reflects the fact that the principal axis of diffusion in the lower stratosphere must be inclined at an angle $\alpha = K_{yz}/K_{yy}$, which is greater than the angle $\beta = (-\partial\bar{\chi}/\partial y)(\partial\bar{\chi}/\partial z)^{-1}$, the angle of inclination of the isopleths of the ozone mixing ratio. If we assume that the horizontal eddy flux of ozone is proportional to its horizontal mean flux, we get a first relation between K_{yy} and K_{yz}.

A second relation can be derived from the vertical flux. If we assume negligible sources or sinks for ozone below 21 km, the net globally integrated vertical flux of ozone must be independent of altitude and equal to the global rate of destruction of ozone at the ground. Because the principal axes of diffusion are tilted, one of the terms in the vertical flux is $-\alpha^2 K_{yy}\partial\bar{\chi}/\partial z$. We have just seen that α has to be greater than the angle of the ozone isopleths, but if we make α too large, then this term in the vertical flux will completely dominate the other terms and, by itself alone, will produce a global flux greater than the destruction at the ground. To avoid this problem we can assume that $\alpha^2 K_{yy}$ has to be minimum. This gives a second relation between K_{yy} and K_{yz}, and we can then compute these two coefficients. The third coefficient, K_{zz}, can then be determined by solving Eq. (9.9).

In the upper stratosphere, there is no passive tracer that can be used to derive the diffusion coefficients, not only because of the lack of observations but also because the effect of photochemical reactions tends to dominate the dynamical effects. Thus the diffusion coefficients have to be extrapolated into the upper stratosphere. For example, one can make K_{yy} proportional to the variance of the meridional wind. But it is very difficult to check the model at high altitudes.

The method of deriving the diffusion coefficients that we have suggested for the lower stratosphere (by no means the only possible method) has the advantage of making the coefficients consistent with the mean meridional circulation, which, incidentally, is known with only a low accuracy (within about a factor of 2). Even though this derivation has not made direct use of the dynamical properties of the stratosphere, the diffusion coefficients appear to agree well with the dynamics (see Table 9.1 for winter values of K_{yy}, K_{yz}, and K_{zz}). High values of K_{yy} at high latitudes correspond to the decaying baroclinic waves in the lower stratosphere and to the common, small wavenumber components of the polar-night jet circulation in the upper stratosphere. Low values of K_{yy} in the equatorial region and in the summer hemisphere agree with the predominantly zonal flow in these parts of the stratosphere. Small values of K_{zz} in the stratosphere (as compared with the troposphere) are consistent with the large static stability of the region. The sign of K_{yz}, which determines the sign of the diffusion angle α, was derived from the transport of ozone. It is also consistent with eddy transport of heat, which is known to be directed from the cold, equatorial lower stratosphere (about −80°C) to the warmer, polar regions.

Using the mean velocities shown in Figure 9.9 and the diffusion coefficients listed in Table 9.1 (plus corresponding values for other seasons, as given by Danielsen, 1974), a numerical model was developed to simulate the transport of radioactive debris (zirconium-95) from a series of Chinese nuclear bomb tests (1968–1972), a high-yield 1968 French test, and the transport of tungsten-185 from the U.S. Hardtack tests, 11° north of the equator, in 1958. With this model (Model I), both the meridional transport and the depletion rates of the stratospheric burden were too large. Significantly better agreement for the total stratospheric burden was obtained by dividing both the mean velocities and the diffusion coefficients by two (Model II), as illustrated in Figure 9.10 for zirconium-95. A comparison between the observed and computed distributions of zirconium-95 is shown in Figure 9.11. The maximum in the northern hemisphere is from the December 12, 1968, injection, and the maximum in the southern hemisphere is from the French test.

The differences between Models I and II for the tungsten-185 tracer injected at 11° N latitude are shown in Figures 9.12 and 9.13. Because discrete spot sampling of the tungsten-185 by B-57 aircraft or balloons could miss the actual maximum, the maximum concentration observed should be equal to or less than that predicted. Model II meets these conditions. Model I predicts values smaller than those observed; therefore, it fails to meet the required conditions. Also, since Model I overpredicts the dispersion and removal, the amount available for deposition during April and June 1959 (illustrated in Figure 9.13) is significantly lower than that observed. The deposition rates for Model II are closer to the observed but still deficient, especially north of 30° N latitude. The location of the observed maximum in the deposition rate between

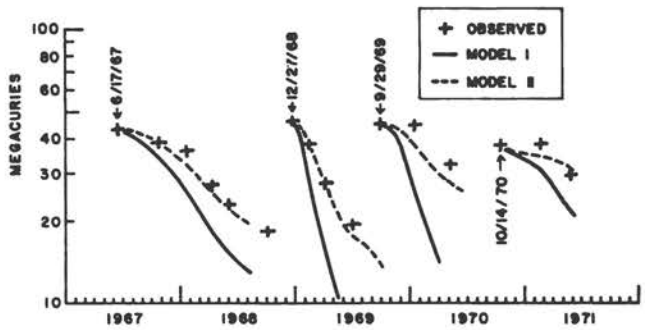


FIGURE 9.10 Global stratospheric burden of zirconium-95 following four Chinese nuclear tests (MCi, decay-corrected to the date of each injection).

TABLE 9.1 Mean Eddy-Diffusion Coefficients K_{yy}, K_{yz}, K_{zz} for Northern Winter Season, December–February

Alt (km)	Latitude 90	80	70	60	50	40	30	20	10	0	−10	−20	−30	−40	−50	−60	−70	−80	−90
K_{yy} (10^8 cm^2/sec)																			
50	17.88	14.10	11.00	8.00	4.99	3.97	3.34	2.92	2.87	3.03	2.96	2.54	1.98	1.40	0.91	0.44	0.21	0.04	0.03
45	15.61	13.54	10.53	7.41	4.73	3.75	2.83	2.40	2.25	2.33	2.42	2.20	1.85	1.34	0.91	0.41	0.20	0.05	0.03
40	12.83	11.96	9.76	6.39	4.41	3.20	2.27	1.73	1.50	1.56	1.76	1.77	1.54	1.31	0.88	0.40	0.24	0.18	0.15
35	9.75	9.82	8.17	5.20	3.90	2.70	1.71	1.46	1.19	1.20	1.36	1.51	1.46	1.19	0.85	0.48	0.38	0.30	0.26
30	5.61	6.32	5.76	4.43	3.41	2.23	1.56	1.06	0.97	0.97	0.98	0.99	1.01	1.00	0.98	0.80	0.75	0.63	0.56
25	3.66	4.01	5.07	4.36	3.38	2.04	1.51	1.45	1.25	1.48	1.46	1.17	0.97	0.95	1.14	1.72	1.78	1.33	1.09
20	2.59	3.32	5.12	5.40	3.93	2.37	1.57	1.51	1.86	2.51	2.24	1.69	1.05	1.07	2.10	4.27	3.53	1.87	1.29
15	2.42	2.36	3.92	3.98	4.84	3.26	1.78	2.10	4.47	6.10	5.14	2.82	1.31	1.13	2.07	2.51	2.16	1.57	1.41
10	2.98	3.24	3.93	4.90	5.02	3.35	1.80	2.33	6.23	10.39	8.55	3.78	1.25	1.31	1.64	2.11	2.14	1.97	1.90
5	2.55	3.91	4.50	5.27	4.68	2.85	1.16	1.48	3.64	4.80	4.12	1.89	0.87	1.42	2.35	3.55	3.59	2.69	2.29
0	2.88	3.34	4.18	4.29	3.55	2.02	0.85	0.76	1.48	1.78	1.64	0.88	0.58	1.69	2.09	2.35	2.37	2.41	2.05
K_{yz} (10^8 cm^2/sec)																			
50	0.00	−7.93	−8.50	−4.92	−2.58	−1.25	−0.10	0.35	0.29	−0.06	−0.22	−0.24	−0.06	0.25	0.58	0.93	1.07	0.84	0.00
45	0.00	−10.01	−10.30	−4.80	−2.36	−0.86	0.22	0.49	0.30	−0.07	−0.30	−0.37	−0.23	0.05	0.41	0.55	0.84	0.54	0.00
40	0.00	−8.74	−10.00	−4.19	−2.07	−0.29	0.41	0.66	0.28	−0.10	−0.46	−1.00	−0.36	−0.16	0.17	0.48	0.55	0.51	0.00
35	0.00	−5.45	−6.28	−1.16	−1.31	0.09	0.35	0.45	0.27	−0.09	−0.42	−0.97	−0.54	−0.29	−0.06	0.23	0.49	0.47	0.00
30	0.00	−1.86	−2.49	−1.33	−0.04	0.15	0.30	0.24	0.18	−0.05	−0.30	−0.34	−0.31	−0.21	−0.02	0.35	0.45	0.37	0.00
25	0.00	0.26	−0.18	−1.08	−0.92	−0.29	0.10	0.22	0.35	0.05	−0.25	−0.25	−0.18	0.27	0.92	1.07	0.99	0.67	0.00
20	0.00	−0.99	−3.65	−5.58	−6.27	−5.02	−3.01	−0.09	0.25	0.12	−0.27	−0.20	0.20	−1.34	2.28	3.07	1.84	1.00	0.00
15	0.00	−1.91	−3.91	−7.42	−12.44	−9.77	−4.92	−0.26	0.31	0.10	−1.50	−0.68	0.32	1.29	2.24	2.40	0.79	−0.06	0.00
10	0.00	0.14	−0.45	−2.23	−3.27	−2.11	0.08	0.97	3.33	−0.17	−4.52	−1.52	−0.36	0.14	0.62	0.52	−0.14	−0.25	0.00
5	0.00	0.58	1.05	1.50	2.76	1.79	1.04	2.27	3.74	−0.13	−2.65	−1.64	−0.68	−1.04	−1.39	−1.96	−1.60	−0.84	0.00
0	0.00	0.37	0.37	0.23	0.02	0.21	0.03	0.02	0.33	−0.01	−0.21	−0.01	−0.01	−0.04	−0.08	−0.35	−0.31	−0.03	0.00
K_{zz} (10^8 cm^2/sec)																			
50	8.17	12.63	14.73	11.19	9.50	8.56	8.17	8.21	8.20	8.17	8.18	8.19	8.17	8.21	8.52	10.12	13.75	24.56	8.17
45	4.95	12.35	15.03	8.06	6.13	5.15	4.97	5.05	4.99	4.96	4.99	5.02	4.98	4.95	5.13	5.68	8.43	10.68	4.95
40	3.00	9.40	13.25	5.75	3.97	3.03	3.08	3.26	3.06	3.01	3.12	3.57	3.09	3.02	3.04	3.59	4.75	4.39	3.00
35	1.82	4.84	6.65	3.74	2.26	1.83	1.90	1.96	1.88	1.83	1.95	2.45	2.02	1.89	1.83	1.93	2.45	2.56	1.82
30	1.11	1.65	2.18	1.50	1.11	1.11	1.16	1.16	1.14	1.11	1.20	1.22	1.20	1.15	1.11	1.26	1.38	1.32	1.11
25	2.12	2.17	2.19	1.98	1.52	0.97	0.83	1.05	1.56	2.60	1.60	1.16	0.87	0.91	1.87	2.10	1.88	1.29	0.95
20	3.53	3.89	6.25	8.35	11.62	11.49	6.38	1.04	2.07	4.61	2.29	1.25	0.67	2.31	3.77	4.19	2.69	1.44	0.88
15	4.93	6.32	9.03	12.70	33.92	30.05	13.94	1.08	2.64	6.61	3.39	1.52	0.51	1.90	3.87	4.83	2.43	0.86	0.82
10	6.14	8.67	10.27	11.40	7.99	3.94	1.77	3.48	16.93	26.87	17.53	5.47	2.20	1.94	3.30	3.45	3.14	2.27	2.11
5	8.33	13.52	26.14	35.06	40.87	33.48	16.70	19.82	35.86	91.84	43.57	21.60	12.80	15.94	20.89	14.56	8.97	5.24	4.36
0	4.48	6.71	13.10	13.90	15.51	14.29	11.50	9.59	12.31	13.70	13.37	8.72	9.61	9.02	9.63	5.85	4.92	4.28	4.20

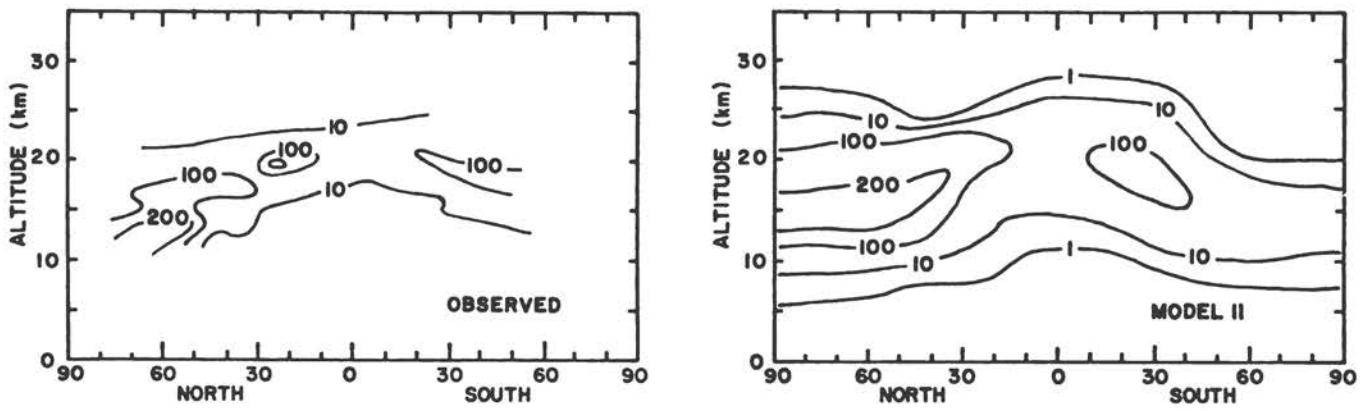


FIGURE 9.11 Second computation of the distribution of zirconium-95 (pCi/standard cubic meter) April 1969.

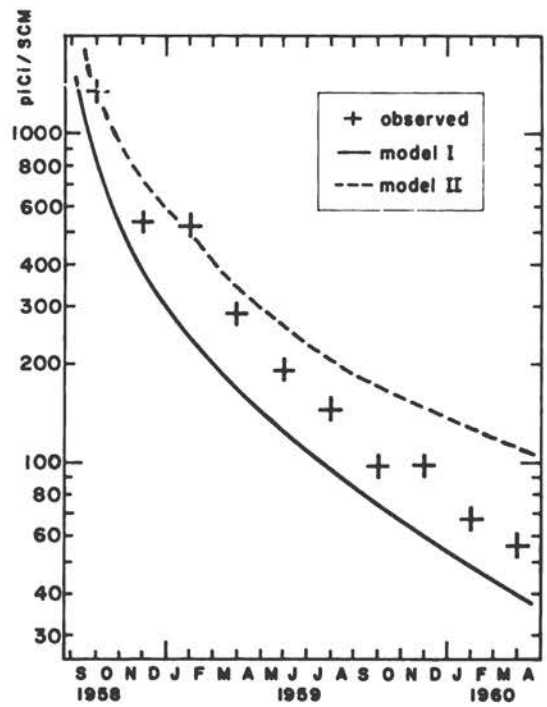


FIGURE 9.12 Maximum concentration of tungsten-185 in the stratosphere, following the 1958 U.S. nuclear tests (pCi/standard cubic meter, decay-corrected to August 15, 1958).

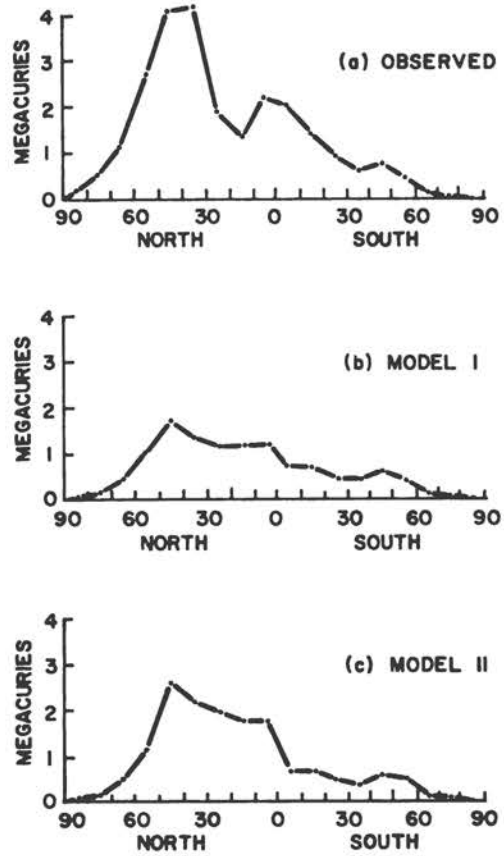


FIGURE 9.13 Latitudinal distribution of the deposition of tungsten-185 during April–June 1959 (MCi/10°-latitude band).

30° and 50° N latitude clearly indicates that the eddy diffusion must dominate the transport by the mean circulation. The tungsten-185 was injected into the Hadley circulation and would have been deposited by this circulation cell between 30° N and the equator if, in fact, the mean circulations dominated in the transport.

On the basis of the general improvement in predictions demonstrated by Model II for injections at both high and low latitudes, we suggest that the originally derived mean circulation and diffusion coefficients be decreased by a factor of 2. It is our opinion that the accuracy of the mean circulation is within this limit. We stress that if the mean circulations are decreased, the diffusion coefficients must also be decreased.

The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

Danielsen, E. F. (1959). The laminar structure of the atmosphere and its relation to the concept of the tropopause, *Arch. Meteorol. Geophys. Bioklimatol., Ser. A, B11*, 293.

Danielsen, E. F. (1964). *Report on Project Springfield*. Defense Atomic Support Agency Contract DA-49-XZ-079, DASA 1517, Washington, D.C.

Danielsen, E. F. (1968). Stratospheric-tropospheric exchange based on radioactivity, ozone and potential vorticity, *J. Atmos. Sci.* 25, 502.

Danielsen, E. F. (1974). The national stratosphere of 1974, in CIAP Monograph 1, Department of Transportation, Washington, D.C., pp. 6–9.

Danielsen, E. F., R. Bleck, J. P. Shedlovsky, A. Wartburg, P. Haagenson, and W. Pollock (1970). Observed distribution of radioactivity, ozone and potential vorticity associated with tropopause folding, *J. Geophys. Res.* 75, 2353.

ESSA (1969). Tech. Rep. WB 9, *Weekly Synoptic Analyses, 5-, 2-, and 0.4-Millibar Surfaces for 1966*, U. S. Department of Commerce, Environmental Science Services Administration, Contract P-32290 (G), Washington, D.C.

Hering, W. S. (1965). Ozone measurements for diagnostic studies of atmospheric circulation, AIAA Second Annual Meeting No. 65-462.

Mahlman, J. D. (1973). Preliminary results from a three-dimensional, general-circulation/tracer model, in *Proc. Second Conference on the Climatic Impact Assessment Program*, November 14-17, 1972, pp. 321–337.

Reed, R. J. (1955). A study of a characteristic type of upper-level frontogenesis, *J. Meteorol. 12*, 226.

Reed, R. J., and E. F. Danielsen (1959). Fronts in the vicinity of the tropopause, *Arch. Meteorol. Geophys. Bioklimatol., Ser. A, B11*, 1.

Reed, R. J., and K. E. German (1965). A contribution to the problem of stratospheric diffusion by large-scale mixing, *Mon. Weather Rev.* 93, 313.

10 Ozone

DONALD M. HUNTEN
Kitt Peak National Observatory

10.1 PROLOGUE

Ozone is a form of oxygen containing three atoms per molecule instead of the normal two. Stable if undisturbed, it is a violent explosive in concentrated form, releasing almost as much energy as the same mass of TNT when the atoms revert to their normal arrangement. It is a strong absorber of ultraviolet radiation; the amount present in the atmosphere cuts off all wavelengths below 300 nm. This absorption is annoying to astronomers in their professional capacity but beneficial to many organisms, including people, because solar radiation around 300 nm is dangerous to them. Ozone is not a good thing to breathe; fortunately most of it is located in the stratosphere at an altitude in the vicinity of 25 km, or 80,000 feet, where we can get the benefit of its absorption without having to breathe it.

Almost a century ago, Cornu found that the spectra of all astronomical sources, including the sun, show a common cutoff at 300 nm and inferred that some absorber in the atmosphere was responsible. Two years later, Hartley observed the absorption spectrum of ozone in the laboratory and suggested this gas as the absorber. For the last several decades, measurements of the strength of this absorption have been used to measure the ozone amount from the ground and from rockets, balloons, and satellites.

In the past few years, two things have begun to be clear: the ozone shield is important to us, and it can be damaged by more than one of man's activities. Only an energy input can make ozone; this energy can be released, and the ozone destroyed, by a variety of catalytic cycles involving common pollutants. One such cycle, based on oxides of nitrogen, operates in the natural atmosphere, where the source of the pollutant is N_2O from anaerobic decay at the earth's surface. The stratospheric ozone is in balance, or steady state, between its creation by sunlight and its destruction, chiefly by the nitrogen oxides. This destruction can be increased by the release of additional nitrogen oxides, for example by high-flying aircraft. Another potent catalyst is chlorine, which can be carried to the stratosphere in compounds such as methyl chloride, carbon tetrachloride, and chlorofluoromethanes such as Freon. Concern about such potential problems has led to a large acceleration of research on the chemis-

try and meteorology of ozone. The following sections summarize the knowledge so obtained.

10.2 OBSERVATIONAL METHODS

Early work on atmospheric ozone has been described by Mitra (1952), the source for the following summary. Cornu's 1878 discovery, and Hartley's identification of ozone, have been mentioned above. Shortly after, Huggins strengthened the identification on the basis of the diffuse bands at somewhat longer wavelengths. Attempts to defeat the absorption by observation from mountaintops were unsuccessful, and it became clear that the ozone must be primarily at higher altitudes. (Even today there is a tendency to suppose that citizens of a city like Denver, and skiers in the nearby mountains, are exposed to more solar ultraviolet than people at lower altitude. If this is so, it is because the air is clear and outdoor activity is popular, not from any lessening of the ozone shield.)

In Figure 10.1, the scale on the left shows the ozone absorption cross section from 220 to 310 nm. The right-hand scale shows a typical transmission at 30° zenith angle for a typical midlatitude abundance of ozone: 0.3 cm-atm or 8×10^{18} molecules cm^{-2} (the conversion factor is Loschmidt's number, 2.687×10^{19}). A measurement of ultraviolet transmission near 300 nm leads to a straightforward inference of the ozone abundance above the instrument and is routinely used for this purpose. At shorter wavelengths, the transmission rapidly becomes negligible. Variations of ozone abundance, or changes in the solar zenith angle, cause significant shifts of the cutoff wavelength. This wavelength is right in the middle of the biologically damaging band, and these shifts are therefore of great practical concern.

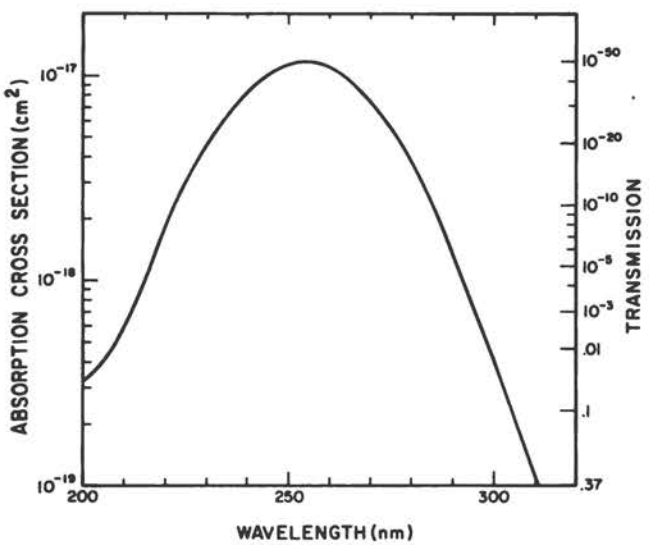


FIGURE 10.1 Ozone absorption cross section from 200 to 310 nm. At the right is shown the transmission for a typical midlatitude amount (0.3 cm-atm) at a zenith angle of 30°.

The first quantitative estimate of height was made in 1917 by Strutt (the fourth Lord Rayleigh). He measured the growth of ozone absorption toward the horizon and its departure from a secant law, obtaining 40–60 km. The true value, 25 km, was not obtained until 1929. Götz, Meetham, and Dobson used the *Umkehr* method, which takes advantage of light scattered from high in the atmosphere at low sun; and in 1934, E. and V. H. Regener sent a spectrograph aloft on a balloon. Both techniques are still in use.

Ozone abundances were measured from 1921 by photographic spectrophotometry, following Fabry and Buisson, and from 1931 by Dobson's photoelectric instrument. The basis is the use of pairs of wavelengths at which the ozone absorption differs considerably but atmospheric extinction does not. By 1929, the results of a latitude survey had been published by Dobson, and a year later Chapman's photochemical theory appeared.

Figure 10.2 shows one of the first, and still one of the best, high-altitude ozone profiles obtained in 1949 from a V-2 rocket by F. S. Johnson *et al.* H. Trinks has extended measurements to higher altitudes by mass spectroscopy, and W. F. J. Evans and E. J. Llewellyn by a day-airglow technique.

Photolysis of ozone by absorption of the Hartley continuum leaves the O_2 molecule in its $^1\Delta_g$ excited state, which radiates at 1.27 μm. Rocket observations of this band can therefore be converted to ozone abundances. Figure 10.3 shows mean profiles up to 30 km for various latitudes and two seasons, spring and fall (Dütsch, 1971). There is an obvious latitude and seasonal effect, with the most total ozone always at high latitude and the seasonal maximum in the spring. These variations occur below the main photochemical peak, located at 25 km, and are due

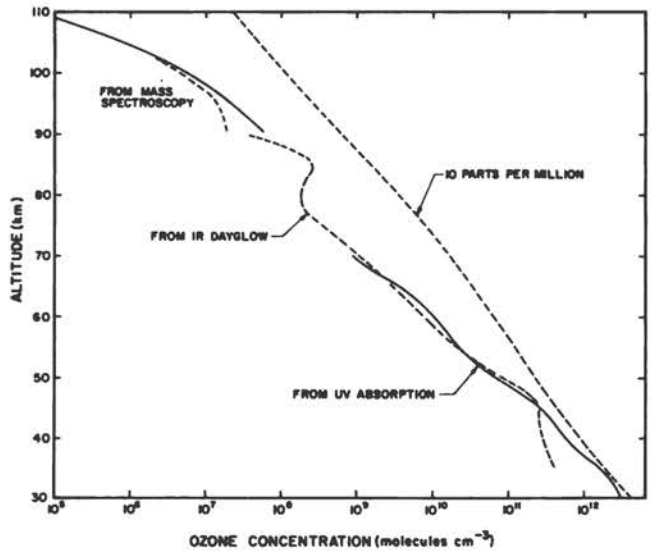


FIGURE 10.2 Daytime ozone profiles up to 110 km by three different rocketborne experiments. A molecular mixing ratio of 10 parts per million is shown for comparison.

almost entirely to horizontal transport of ozone, as discussed below in Sections 10.3 and 10.5.

Various instruments have been used aboard balloons and aircraft for ozone measurement. They include an updated ultraviolet method and two chemical methods operating on a stream of air pumped through the instrument. Detection can use an electrochemical technique with a potassium iodide solution or a chemiluminescent device. Satellite soundings use the "backscattered ultraviolet" (BUV) technique, which employs several wavelengths between 250 and 340 nm and takes advantage of the fact that each wavelength is scattered from a different height.

10.3 TYPICAL BEHAVIOR

Figure 10.3 shows that the latitude gradient is of opposite direction above and below about 25 km. The upper part is photochemically controlled, as discussed below, and shows the expected equatorial maximum. The lower part dominates the variations of total ozone and is therefore responsible for the considerable latitude and seasonal variations of this quantity. The lower region is strongly shielded from photochemical destruction and reflects quasi-horizontal mixing by atmospheric motions.

This example of stratospheric meteorology has been known since Dobson's early work, and much of the interest of meteorologists has been in the variations or departures from the mean. Aeronomers tend to emphasize the mean state and ignore the departures. Fortunately, we now have explanations of both aspects, although the remaining sections of this paper concentrate on the aeronomical side.

The term "quasi-horizontal" is used for the meridional motions because, at least in the lower stratosphere, they appear to run predominantly parallel to the tropopause. This boundary, visible in Figure 10.3 by breaks in the ozone profiles, slopes from 17 km at the equator to 9 km at the poles. The corresponding quasi-horizontal motion of radioactive tracers was measured in a large-scale program during the 1960's (Machta *et al.*, 1970). We can, therefore, expect that the redistribution of ozone will take place in a similar manner.

At approximately 30 km, the photochemical time constant of ozone is 15 days, and it rapidly increases at lower altitudes (McElroy *et al.*, 1974). The ozone mixing ratio along the quasi-horizontal surfaces should, therefore, be controlled primarily by the low-latitude region, where the time constant is shortest. This is the semiquantitative explanation of the low-altitude bumps at high latitudes seen in Figure 10.3. The same data (supplemented above

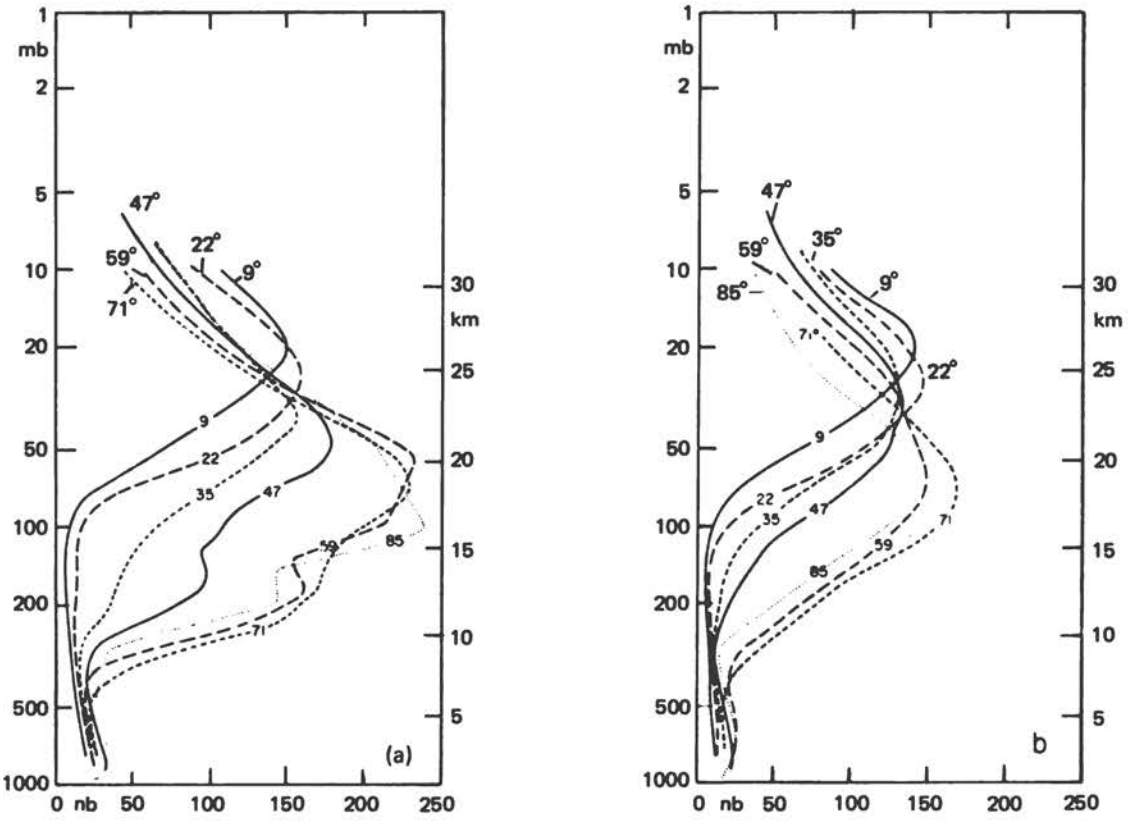


FIGURE 10.3 Mean vertical distributions of ozone to 30 km by balloon sounding. Each curve is tagged with the latitude. Panel (a) represents spring and (b) autumn (Dütsch, 1971).

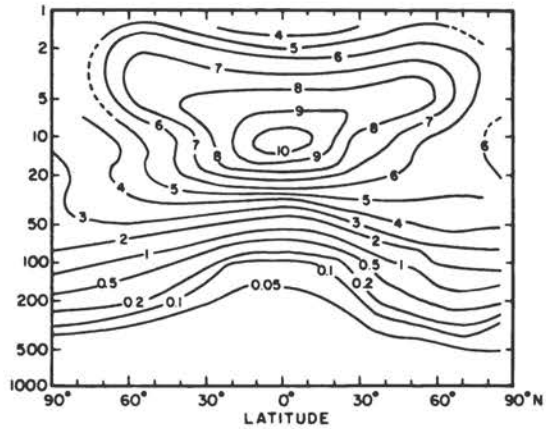


FIGURE 10.4 Pole-to-pole cross section of ozone/air mixing ratio for March–April (Dütsch, 1974).

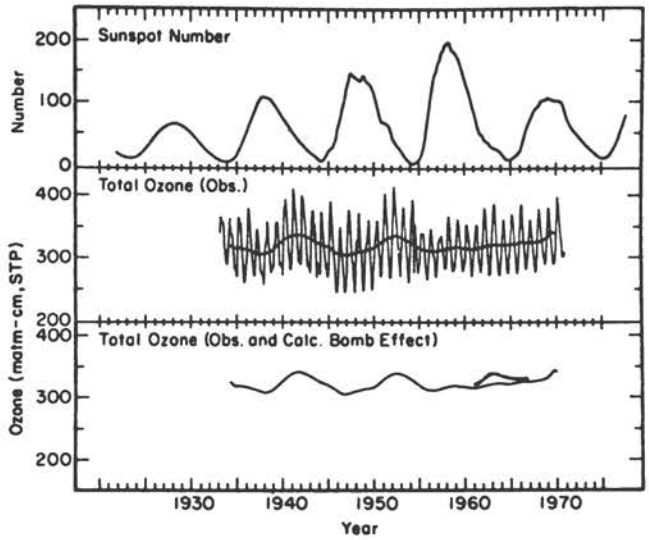


FIGURE 10.5 Average ozone and sunspot number. In the bottom panel, the smoothed ozone data are shown again, with a computed correction for nuclear testing. (Climatic Impact Committee, 1975).

30 km by *Umkehr* measurements) are shown in Figure 10.4 in the form of a meridional cross section of mixing ratio, which is the quantity that is conserved during transport. The sloping surfaces are clearly brought out in this kind of presentation. Apparently, the quasi-horizontal mixing is stronger, or the slope of the surfaces is somewhat greater, in the winter hemisphere. Thus, the mixing processes are entirely responsible for the behavior of total ozone at middle and high latitudes.

Many workers believe that the above description needs to be supplemented by large-scale (but low-velocity) mean motions. This issue is succinctly discussed by Dütsch (1974).

Vertical mixing is also important in determining the ozone below 20 or 25 km; this process is discussed in the aeronomical context in Section 10.5.

Total ozone, like any meteorological quantity, is subject to fluctuations that interfere with any attempt to recognize a trend. Christie (1973) has compiled the available data in the middle panel of Figure 10.5, which are heavily weighted toward the northern hemisphere. The annual oscillations are the seasonal effect; when they are filtered out, the relatively smooth curve is obtained. There is an obvious anticorrelation with solar activity, except that the 1963 peak is missing. However, it was just at this time that a significant loss of ozone due to nuclear bomb tests should have occurred. The bottom panel of Figure 10.5 shows an estimate of how the ozone would have behaved in the absence of the bomb effect; the 1963 peak is now plainly present. Both effects are linked to NO_x production and catalytic ozone destruction, as discussed below.

Although the features of Figure 10.5 can be explained in principle as just outlined, the explanations are not completely quantitative. If one wishes to establish the presence of a small trend due to the effect of a pollutant, the oscillations interfere seriously. This problem has been analyzed quantitatively by Pittock (1974). Even with an ideal global network, a trend of 2 percent per decade can be established to 95 percent confidence only after 10 years of observation. By the time an ozone reduction of this order can be confirmed, it may be much too late to reverse the actions that caused it. Decisions must therefore be based on predicted effects; unsatisfactory as this may be, there is no alternative.

10.4 PHOTOCHEMICAL PRODUCTION AND LOSS

Figure 10.6 illustrates the global ozone budget, following H. S. Johnston and G. Whitten. Production by solar ultraviolet radiation is 5×10^{31} molecules/sec, or 4.0 metric tons/sec. The absolute minimum energy needed to produce two ozone molecules in the stratosphere is 5.1 eV, the dissociation energy of O_2. Even without the additional

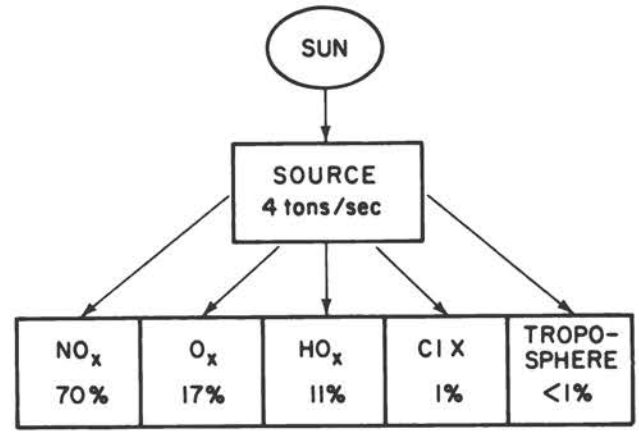


FIGURE 10.6 Ozone budget for the stratosphere.

inefficiency from use of higher-energy photons, the required power input is 2×10^{13} W. This exceeds man's total power consumption for 1970, including the heating of buildings, by a factor of 3. An even larger amount of energy is absorbed by the ozone itself, leading to a small amount of atomic oxygen in equilibrium with the ozone. We now discuss the individual terms of Figure 10.6 in more detail, focusing on the dominant processes only. The budgets of the pollutants are considered in the next section. A much more thorough treatment is given by Nicolet (1975).

THE STRATOSPHERE

Ozone Production

Oxygen atoms are formed by photodissociation of O_2 at wavelengths below 240 nm:

$$O_2 \xrightarrow{h\upsilon} O + O, \tag{10.1}$$

$$O + O_2 \underset{h\upsilon}{\overset{M}{\rightleftharpoons}} O_3. \tag{10.2}$$

Formation of ozone is mediated by a third body, *M*. The reverse of Reaction (10.2) is accomplished by absorption of sunlight in the Hartley continuum and also to a smaller extent by the Chappuis continuum in the yellow. Ozone and atomic oxygen are therefore interchangeable, and a convenient name for their sum is "odd oxygen" or O_x, where x is 1 or 3 only.

The same wavelengths that create odd oxygen in Reaction (10.1) are also absorbed by ozone in the reverse direction of Reaction (10.2). Because of this competition, there is an important "self-healing" process: any lessening of ozone amount channels more energy into Reaction (10.1) and increases the odd-oxygen production. This effect must be (and normally is) included in any estimate of the effect of an added pollutant that destroys ozone. A quantitative illustration is given by McElroy *et al.* (1974).

The O_x Sink

The principal sink within the odd-oxygen system is

$$O + O_3 \rightarrow O_2 + O_2. \tag{10.3}$$

There is also a three-body association of O atoms, but it is not important below 80 km.

The system discussed so far is usually called "Chapman chemistry"; it was formulated in 1930 and accepted as the principal mechanism until the late 1960's. Measured rate coefficients finally became accurate enough at this time to reveal that the Chapman system predicts about twice too much ozone. Since the loss rate by Reaction (10.3) varies as the square of the ozone density, its actual contribution is less than one quarter of the total.

The HO_x Sink

Odd hydrogen, or HO_x, consists of the radicals H, OH, and HO_2. They exist in a steady state with water vapor, just as does O_x with O_2. Their aeronomy was first discussed in 1950 by Bates and Nicolet, stimulated by Meinel's discovery of an intense infrared airglow due to OH. The importance of HO_x to ozone seemed slight until 1964, when Cadle and Hampson independently pointed out that a likely source was the reaction of H_2O with $O(^1D)$, a metastable excited atom produced in ozone photolysis. Before long, however, it became clear that the HO_x system can only be adequately treated with allowance for vertical transport of long-lived constituents such as CH_4 and H_2 and also the upward flux due to hydrogen escape. Treatments including all these factors were not published until 1973, by Hunten and Strobel and by Liu and Donahue, building on earlier studies of the Martian atmosphere. Fortunately it turned out that the escape term is not important to the problem of stratospheric ozone.

Once the odd-hydrogen amount has been ascertained, the odd-oxygen sink is

$$HO_2 + O \rightarrow OH + O_2, \tag{10.4}$$

$$OH + O \rightarrow H + O_2, \tag{10.5}$$

$$H + O_2 \rightarrow HO_2. \tag{10.6}$$

The first two reactions destroy the odd oxygen, and the third recycles the catalyst to its original form. This cycle is traversed in a very short time, typically a few seconds. It is thus highly appropriate to treat odd hydrogen as a single entity.

Detailed calculations (e.g., McElroy *et al.*, 1974) show that the HO_x sink dominates the ozone loss in the mesosphere (above 50 km) but is only a minor contributor in the stratosphere.

The NO_x Sink

The possible importance of odd nitrogen was suggested in 1970 by P. J. Crutzen. Wide interest was generated the following year when H. S. Johnston, as well as Crutzen, pointed out that NO_x from aircraft exhausts could possibly lead to large reductions of the ozone abundance. At about the same time, J. E. McDonald pointed out a probable connection between ozone reductions and increases of skin cancer. His basis was the observation that this disease is much more probable in the southern United States than the northern, by a factor currently estimated to be about 5 (Climatic Impact Committee, 1975). He pointed out that ultraviolet radiation near 300 nm behaves the same way because of the latitude gradient of ozone and the change of mean solar elevation. All these suggestions

were controversial when they were made, but detailed study has shown them to have been remarkably accurate. According to Figure 10.6, the natural NO_x is responsible for 70 percent of the ozone sink; the expected effect of a large fleet of supersonic transports is indeed considerable.

Odd nitrogen is principally composed of NO, NO_2, and HNO_3. Some writers use the symbol NO_x for the first two and NO_y for all three, but this distinction is ignored here. The cycle that destroys ozone is

$$NO + O_3 \rightarrow NO_2 + O_2, \tag{10.7}$$

$$NO_2 + O \rightarrow NO + O_2. \tag{10.8}$$

Again, we find rapidly interchangeable constituents, NO and NO_2. Nitric acid is formed and destroyed principally by

$$NO_2 + OH \underset{h\upsilon}{\overset{M}{\rightleftharpoons}} HNO_3. \tag{10.9}$$

The acid form of NO_x is catalytically inactive; the fraction tied up in this way depends on height but is around half in the most chemically active region near 30 km.

Natural NO_x is produced from N_2O by reaction with $O(^1D)$. It is not significantly destroyed within the stratosphere; it must be physically removed by atmospheric motions until, mainly as HNO_3, it dissolves in water drops and is rained out. These processes are the subject of the next section.

For assessment of likely effects of an artificial perturbation, it is desirable to have a formula relating the change of ozone, $\delta(O_3)$, to the change of NO_x mixing ratio $\Delta f(NO_x)$. The results of McElroy *et al.* (1974) can be fitted by

$$\delta\,(O_3) = (1.405\,\Delta f - 0.0105\,\Delta f^2)\ \text{percent}$$

for Δf in ppb and less than 24.

The ClX Sink

The aeronomy of chlorine, first studied for Venus in 1971 by Prinn, was transferred to the earth's stratosphere three years later. Potential sources are still being turned up, but the most likely ones are compounds that, like N_2O, are insoluble in water and can be carried to the stratosphere where they are photolyzed and release their chlorine. Examples are CH_3Cl, CCl_4, CF_2Cl_2, and $CFCl_3$. The first is probably of natural origin; the second may or may not be; and the last two are certainly artificial. Under such trade names as Freon, they are used as refrigerants and propellants in spray cans. Their atmospheric concentration is still far below the steady-state value to be expected on the basis of current production rates, as pointed out by Molina and Rowland (1974), and the present small ClX sink for ozone is likely to become much larger if the current rate of release is continued.

The sink operates by the following reactions:

$$Cl + O_3 \rightarrow ClO + O_2, \tag{10.10}$$

$$ClO + O \rightarrow Cl + O_2. \tag{10.11}$$

The similarity with the NO_x cycle, Reactions (10.7) and (10.8), is noteworthy. A further similarity arises in the formation of HCl as the major inactive species:

$$Cl \overset{H_2}{+ CH_4} \rightarrow HCl \overset{H}{+ CH_3}, \tag{10.12}$$

$$HCl + OH \rightarrow Cl + O_2. \tag{10.13}$$

An important difference, however, is that, in comparison with NO_x, a far smaller fraction of the ClX is found in the active forms, Cl and ClO. Computations by Crutzen (1974) suggest

$$\delta(O_3) = 1.6\,\Delta f\ \text{percent},$$

where Δf is now the total Cl mixing ratio in ppb.

Other Halogens

Fluorine is far less effective in destroying ozone than is chlorine, because the analog of Reaction (10.13) is endothermic. The fluorine atoms are all tied up in HF. Bromine, on the other hand, is more effective. Only its much smaller abundance (in the present atmosphere) keeps it from being a serious problem.

Downward Transport

The last box in Figure 10.6 represents loss of ozone to the troposphere. It can be estimated from the gradient of mean ozone mixing ratio just above the tropopause, as discussed in the next section. Although this flux may be significant to the ozone density of the upper troposphere, it is insignificant to the stratospheric budget.

Comments

This section has outlined only the major features of the ozone equilibrium for the stratosphere. Minor reactions have been omitted for the sake of clarity. Even in this simplified framework, there are obvious interactions among the various systems. The $O(^1D)$ that produces odd hydrogen and odd nitrogen is the result of ozone photolysis. OH, a prominent member of the odd-hydrogen family, produces nitric acid [Reaction (10.9)] and recycles hydrochloric acid [Reaction (10.13)]. Such interactions are normally included in computer models that may contain more than 50 temperature-dependent chemical reactions and 10–20 transport equations. Further complications include the "self-healing" effect

described above: any lessening of ozone amount channels more energy into Reaction (10.1) and increases the total production of ozone. Ozone is also intimately involved in the thermal balance of the stratosphere, and any change of its amount will affect the temperature.

THE MESOSPHERE AND THERMOSPHERE

In the mesosphere, the role of dominant ozone sink is taken over by the HO_x system. By coincidence, here too the ozone densities are about half of what they would be with only the O_x sink. The dominant form of odd oxygen is O, and the amounts of O and O_2 become comparable at the mesopause, at about 90 km. A large diurnal variation of O_3 is therefore present, with O converted to O_3 at night and the reverse in the daytime. The regions of the mesopause and lower thermosphere are dominated by transport, with O_2 being mixed upward and O downward. Ozone takes on a similar role to O in the stratosphere, a minor form in almost instantaneous equilibrium through the cycle described by Reaction (10.2).

THE TROPOSPHERE

Until recently it was assumed that ozone in the troposphere was entirely due to downward flow from the stratosphere, with destruction at the surface. An excellent study based on this viewpoint has been given by Junge (1962). In 1973, Crutzen pointed out that odd oxygen is probably produced during the oxidation of methane to carbon dioxide and water in the lower troposphere. A tentative model based on this idea was given by Chameides and Walker (1973). The basic source of odd oxygen is photolysis of NO_2:

$$NO_2 + h\nu \rightarrow NO + O. \qquad (10.14)$$

The role of the methane oxidation chain is to produce HO_2, which recycles the NO to NO_2:

$$HO_2 + NO \rightarrow OH + NO_2. \qquad (10.15)$$

There is a rather delicate balance between this source and a sink that also involves NO_x:

$$NO + O_3 \rightarrow NO_2 + O_2, \qquad (10.7)$$

$$NO_2 + O_3 \rightarrow NO_3 + O_2, \qquad (10.16)$$

$$NO_3 + h\nu \rightarrow NO + O_2. \qquad (10.17)$$

As usual, other reactions are present, but this set gives the flavor of the mechanism. The photochemical lifetimes estimated for ozone production and loss are a few tenths of a day at the ground, 1 day at 5 km, and 10 days at 10 km. These are shorter than the vertical mixing time of a month or so and suggest that local photochemical equilibrium should obtain except perhaps just below the tropopause.

Chameides and Walker showed that their model gives as good an account of certain observations as the traditional one. A maximum abundance of ozone in spring and summer at 30° N is explained by faster photolysis rates with more sunlight. Fabian (1974) points out, however, that data for other latitudes are more consistent with the older view that requires greater mixing rates out of the stratosphere at certain seasons and locations. And Crutzen does not agree with the short photochemical time constant suggested by Chameides and Walker. Their picture must therefore continue to be regarded as tentative and subject to refutation or confirmation by new evidence.

Urban photochemical pollution involves ozone, among many other objectionable compounds. The principal source of odd oxygen is again Reaction (10.14); the NO_2 is a primary exhaust product of automobile engines. There may be some recycling by Reaction (10.15).

10.5 ATMOSPHERIC TRANSPORT

As discussed in Section 10.3, major features of the latitude and seasonal variations of ozone are to be explained in terms of quasi-horizontal transport. In this section, the emphasis will be on vertical transport. Although the downward motion of ozone plays some part in determining its distribution below 25 km, the really important effects lie elsewhere: the minor constituents and pollutants that destroy ozone must be continually carried into and out of the stratosphere, and their abundances are often critically dependent on the speed of these transports.

In the last few years, it has become fashionable to classify discussions of atmospheric transport by the number of dimensions explicitly carried, with averages taken over the remaining dimensions. This classification is useful but should not be taken as reflecting a corresponding ranking of validity or applicability. It is possible to have too much information contained in the model as well as too little. Two- and three-dimensional models have useful insights to give us and are briefly discussed at the end of this section. Zero-dimensional models use the concept of a residence time and do not explicitly consider transport at all. They are most valuable as an aid to mental visualization and to help with choices of dominant processes. Most of the present discussion will be in the context of one-dimensional transport. More details can be found in recent articles by Hunten (1975b) and Wofsy (1976).

Atmospheric motions are observed and expected to be dominantly horizontal. It is, therefore, reasonable to assume that, to first order, inert minor constituents or "tracers" are stratified, with globally uniform distribution at any level. A one-dimensional model assumes that vertical transport of a globally averaged quantity takes place at a rate proportional to the gradient of its mixing ratio f_i. The flux is given by

$$\phi_t = -K\, n_a \frac{df_t}{dz}, \tag{10.18}$$

where n_a is the number-density of the background atmosphere and K is the eddy-diffusion coefficient. This name should not be misunderstood; the "eddies" that are responsible for most of the vertical mixing are of very large scale. They can be regarded in part as small vertical components of the strong horizontal motions. Small-scale eddies, such as those seen in plumes and trails, are very ineffective at large-scale mixing. Although there has been some success in attempts to calculate K from first principles (Mahlman, 1973), it must normally be obtained from empirical data on the fluxes and densities of a tracer. Historically, the most important tracer for the lower stratosphere has been radioactive debris, but recently it has become possible to use CH_4 and N_2O, which originate in the ground and are irreversibly destroyed in the stratosphere (Wofsy and McElroy, 1973; Hunten, 1975c). It seems likely that the earlier studies of radioactive debris were biased to short residence times and large K's by particle sedimentation (Hunten, 1975a). Recently, however, Johnston *et al.* (1976) have made use of additional data referring to ^{14}C, which resides principally as CO_2, and have confirmed the eddy coefficients derived from the chemical tracers. A typical set is shown in Figure 10.7, which includes data for heights up to 100 km. (The dashed region, 50–80 km, is a mere interpolation.)

It is possible to augment the strict one-dimensional treatment outlined above with the help of empirical data. It is observed that transport across the equator is strongly inhibited in both the troposphere and the lower stratosphere. Correspondingly, one may use hemispheric, instead of global, averages for constituents whose sources are biased to one hemisphere. Second, the quasi-horizontal surfaces discussed in Section 10.3 may be adopted instead of strictly horizontal ones. This is appropriate for the lower stratosphere, but we have little or no information for heights above 30 km, where we might as well revert to horizontal surfaces. Wofsy (1976) has shown that the "tilted one-dimensional" viewpoint gives a good account of such things as the latitude distributions of CH_4 and CO, and its application to ozone has already been mentioned. The height scale of Figure 10.7 refers to a latitude of 30°. For a source at some other latitude and between 10- to 30-km altitude, an adjustment should be made in accordance with the difference in tropopause heights.

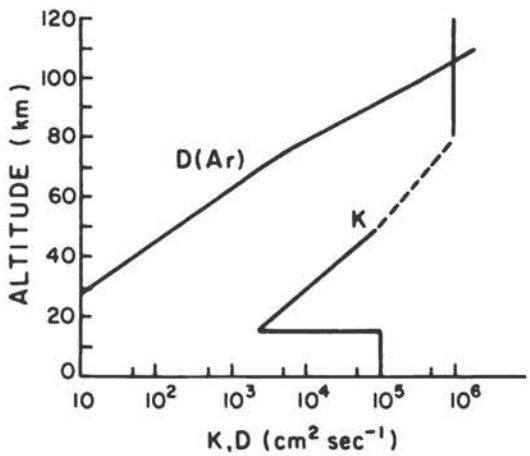


FIGURE 10.7 Eddy-diffusion coefficient K as a function of height and the molecular-diffusion coefficient for argon in air (Hunten, 1975b).

GROUPED CONSTITUENTS AND TOTAL MIXING RATIOS

In the previous section it was noted that many constituents naturally fall into groups (for example, NO_x). The individual compounds are rapidly interchangeable, but the group as a whole may have a very long lifetime. A flux equation such as Equation (10.18) may, therefore, be written and solved for the group, with a great saving of computing time. The distribution among the participating species can then be obtained separately. Thomas (1974) has shown that the "total mixing ratio" is a very useful concept. A good example is provided by the chlorine compounds. The mixing ratio of total chlorine, f_t, is obtained by adding all chlorine atoms, no matter what their chemical form. A flux equation is written for each species, multiplied by the number of chlorine atoms it contains. When all these equations are added, the result is

$$\phi_t = -K\, n_a \frac{df_t}{dz}, \tag{10.19}$$

since the total flux ϕ_t is the algebraic sum of the individual fluxes, and K and n_a are the same for all the species. In the steady state, ϕ_t and df_t/dz must be zero, and therefore f_t must be strictly independent of height. In general, ϕ_t must be found by means of the continuity equation, but this too is greatly simplified because there are no sources or sinks of total chlorine except at or near the ground. Applications to the hydrogen and nitrogen systems are discussed below.

THE ClX SYSTEM

As far as transport is concerned, the ClX system is the simplest of those discussed here and is therefore considered first. Figure 10.8 (Wofsy *et al.*, 1975) has been chosen from the various published computations, because it best illustrates the total-chlorine principle. The dashed line in Figure 10.8(b) represents a situation close to steady state for odd chlorine or ClX. (There is an additional component near the surface, representing chlorine of marine origin; being soluble in water, it is removed from the upper troposphere and makes a negligible contribution to the stratosphere.) The tropospheric mixing ratios of $CFCl_3$ and CF_2Cl_2 are 0.7 and 1.2 ppb, and f_t is therefore 4.5 ppb; the odd Cl in the stratosphere is 4.2 ppb, in

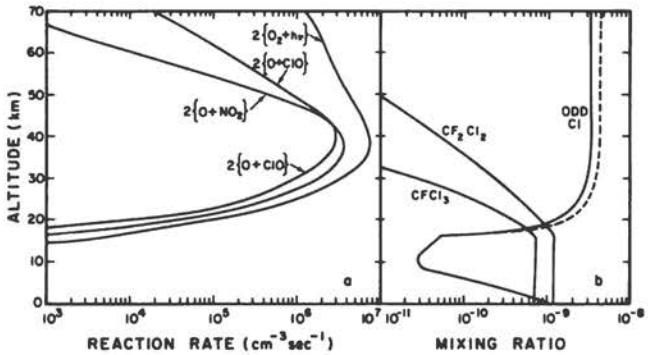


FIGURE 10.8 The right panel shows computed mixing ratios in 1995 for various chlorine compounds by Wofsy *et al.* (1975). On the left is the profile of ozone production, $2(O_2 + h\nu)$ and of destruction by NO_x and ClX.

agreement to the accuracy of scaling the figure. The decrease of fluorocarbon mixing ratio in the stratosphere is controlled by a competition between photolysis and upward mixing, and it is readily calculated. It has also been observed. The stratospheric profile of ClX can then be obtained, if desired, by simple subtraction. A result obtained by detailed computation, such as Figure 10.8(b), can be checked for consistency. What is more important is the virtual independence of the details of vertical transport. Faster transport, for example, will slightly raise the height of transition between tropospheric and stratospheric forms but will not appreciably change the amount of odd chlorine at 30 km, given a fixed mixing ratio of fluorocarbons.

Figure 10.8(a) shows the destruction rates of odd oxygen by the NO_x and ClX cycles and the creation rate by O_2 photolysis. The O_x and HO_x processes make up the difference. This particular model assumed that the rate of release of fluorocarbons increased at 10 percent per year until 1995, the year for which the profiles are shown, and then ceased abruptly. The dashed curve is for 2001, when peak ClX concentrations were found. The perturbation then dies away with a time constant ($1/e$) of 60 years. This is the time required for the tropospheric reservoir to be mixed up to the stratosphere, suffer photolysis, and move back as HCl to be rained out. This time *is* dependent on transport rates, as is the steady-state amount for a given release rate.

An assessment of the present situation has been given by Cicerone *et al.* (1975). They find a stratospheric ClX mixing ratio of 0.8 ppb; the principal sources are fluorocarbons (15 percent), CCl_4 (35 percent), and $CHCl_3$ (50 percent). The last may be of natural origin; carbon tetrachloride is probably artificial, although a natural origin has been argued. In any case, it is the fluorocarbons whose mixing ratio is known to be rapidly increasing.

Cicerone *et al.* compare their predicted HCl mixing ratios with observations by Lazrus and others (Figure 10.9). The agreement is excellent and strongly suggests that no major factors are missing from the models. Even without a model, the steep gradient of mixing ratio directly implies a downward flow of HCl and a source at or above 25 km. These observations, made by filter collection, have been confirmed by an entirely different method, infrared spectroscopy, by Farmer *et al.* (1976).

THE NO_x SYSTEM

The principal natural source of odd nitrogen is

$$N_2O + O(^1D) \rightarrow 2\ NO. \qquad (10.20)$$

However, the major sink of N_2O is photolysis,

$$N_2O + h\nu \rightarrow N_2 + O. \qquad (10.21)$$

The tropospheric mixing ratio of N_2O is about 260 ppb and the stratospheric NO_x of the order of 15 ppb. The efficiency (atoms in NO_x)/(atoms in N_2O) is therefore about 3 percent, and the concept of total mixing ratio is useful only if the huge background of N_2 is omitted. Since the rates of Reactions (10.20) and (10.21) vary differently with height, the nominal 3 percent yield depends somewhat on the eddy coefficient. Even so, there is a strong tendency for the stratospheric NO_x to be independent of K, as it would strictly be if the efficiency did not vary (cf. McConnell and McElroy, 1973).

Artificial NO_x has been studied intensively during the Climatic Impact Assessment Program. Its Report of Findings (Grobecker *et al.*, 1974) and the report of the Climatic Impact Committee (1975) contain many of the details. Jet engines oxidize a small fraction of the N_2 in the air that passes through their flame. Current wide-bodied aircraft produce about 15 g of NO_2 per kg of fuel burned; the *emission index* is said to be 15 g/kg (or 1.5 percent by mass). Current supersonic transports (SST's) produce 18 g/kg. For reasonable assumptions about hours of flight at cruising altitude per year, the mass of NO_2 produced per aircraft is

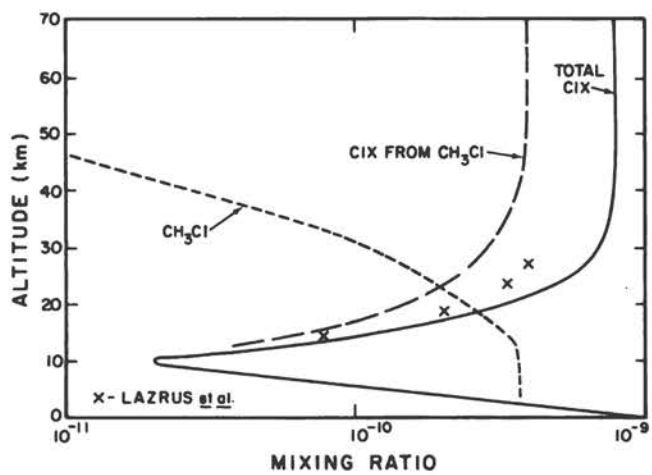


FIGURE 10.9 An estimate of total ClX for 1974 (Cicerone *et al.*, 1975).

390 and 630 metric tons/year, respectively. Spread over a hemisphere, the source strengths are, therefore, 6.3×10^4 and 1.0×10^5 molecules cm^{-2} sec^{-1}.

Given such a source Q at a specific height, the mixing ratio of NO_x at any greater height in steady state can be found by a simple procedure (Hunten, 1975c). We define an "injection coefficient" α, which gives this mixing ratio by

$$f(NO_x) = \alpha Q. \tag{10.22}$$

Figure 10.10 shows α as a function of source height for three profiles of the eddy coefficient. Curve 1 corresponds to Figure 10.7, and Curve 2 to a modification that has an extra layer of intermediate K (10^4 cm^2 sec^{-1}) from 10 to 14 km. The four points are by McElroy *et al.* (1974). The curves are obtained from an analytic solution of Eq. (10.19) when K and n_a vary exponentially with height.

A large fraction of long-distance air traffic flies across the North Atlantic, where the mean tropopause is about 2 km lower than it is at 30° latitude. The effective heights for subsonic and supersonic aircraft are, therefore, 14 and 18 km, and the injection coefficients 0.35×10^{-17} and 4.45×10^{-17} cm^2 sec for Curve 2 of Figure 10.10. (If the subsonic aircraft were to fly only 1 km higher, its effect would be nearly quadrupled.) Substitution in Eq. (10.22) gives NO_x mixing-ratio increases per aircraft in continuous service of 3.5×10^{-13} and 7.2×10^{-12}. A 1 percent ozone reduction requires a mixing-ratio increase of 3.6×10^{-10} according to Section 10.4, and would be produced by just over 1000 subsonic, or 50 supersonic, aircraft. Small increases of cruising height would rapidly increase the effect. In 1974, about 400 wide-bodied airliners were in use, but many of them presumably did not fly so high or so far north as assumed above. Nevertheless, it is clear that we are on the verge of a significant problem from both kinds of aircraft, especially if these estimates should turn out to be too low.

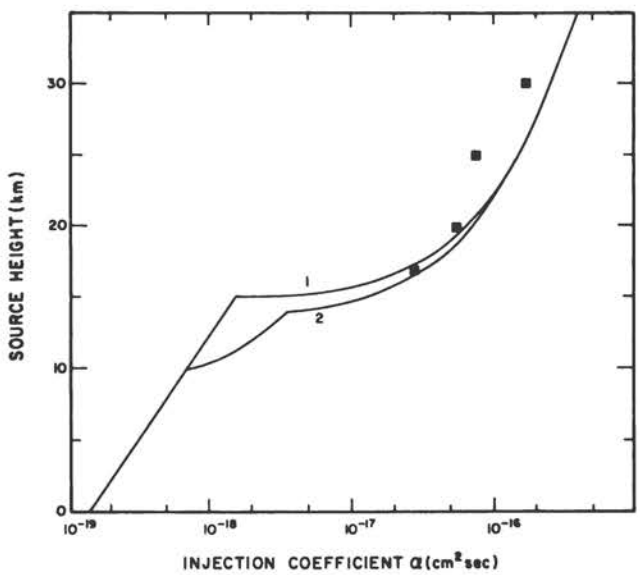


FIGURE 10.10 Injection coefficient for supersonic transports as a function of flight altitude (Hunten, 1975c).

Another artificial source of NO_x is the hot cloud from a high-yield nuclear explosion. As is indicated in Figure 10.5, a small effect of this sort may already have been detected, although the result is marginal. A side effect of a large nuclear war would be a much larger ozone reduction.

Since NO_x resides only in the stratosphere, recovery of the ozone after removal of the source takes place with a time constant of 3 to 5 years, the stratospheric residence time.

Ruderman and Chamberlain (1973) have proposed an explanation of the apparent inverse correlation (Figure 10.5) of ozone amount and solar activity. Low solar activity permits increased penetration of cosmic rays to the vicinity of the earth; these cosmic rays produce additional NO_x. A related correction of opposite sign is suggested by Crutzen *et al.* (1975): solar proton events, commonest during high solar activity, should produce short-lived NO_x enhancements.

The most recent concern is for the possibility of an augmentation of the natural NO_x source through the use of artificial fertilizer. As this is written, a lively debate is in progress (McElroy, 1975; Liu *et al.*, 1976; Crutzen, 1976), and the final conclusion cannot be predicted. The N_2O that goes into Reaction (10.20) is produced in the soil, and also in the oceans, by denitrification of fixed nitrogen. There is evidence for a major sink, which may or may not be in the oceans. Within a few decades, the artificial and natural sources of fixed nitrogen may become comparable. Much of the debate is about the magnitude of possible reservoirs and the rate of exchange with them. The reason for concern is clear: if such a source of pollution should become significant, it would be essentially impossible to turn off and even once turned off might require a long time to decay away.

THE HO_x SYSTEM

Water vapor and its derivatives form the most complicated and least understood of the systems discussed here. One of the major problems, to account for the extreme dryness of the stratosphere, can fortunately be bypassed because we have a good observational base. The other, involving the rate of conversion of odd hydrogen back to H_2O, represents one of the major uncertainties in stratospheric aeronomy. The direct effect on ozone is fairly minor (Figure 10.6), but OH participates in a major way in all systems. Fortunately, this problem too is beginning to have some observational input.

The mixing ratio of H_2O at 30 km is close to 4.0 ppm, and there are indications that it increases by a few ppm at higher altitudes. H_2 and CH_4 are each about 0.5 ppm and after oxidation would contribute an additional 1.5 ppm of H_2O, in reasonable agreement. Total H is, therefore, of order 11 ppm. The ratio of surface to stratospheric humidity is 10^4, and a powerful mechanism is clearly at work to

maintain such a steep gradient. The primary, and perhaps only, mechanism is condensation and fallout. Whether this mechanism is quantitatively adequate is somewhat doubtful. Ellsaesser (1974) has discussed the problem and tentatively concluded that one or more additional processes may be needed. Conventionally, all stratospheric air is assumed to have passed through the cold (−80°C) tropical tropopause and to have taken up that dew point. Correspondingly, the flow must be downward at all other latitudes. There is, however, some doubt that the cold trap is quite cold enough, and the dominance of such a coherent flow pattern could also be questioned. Ellsaesser's suggestion is that condensation and fallout occur within the stratosphere as well, during the winter over the Antarctic. He rejects the possibility that the stratospheric aerosol (concentrated sulfuric acid) plays a substantial role, but the writer believes that this rejection may be premature. The question is of practical as well as academic interest. In the conventional ("Brewer") picture, the stratospheric humidity could be readily increased by additional sources of H_2O. If active mechanisms operate within the stratosphere to keep it dry, its humidity is much harder to perturb.

The total hydrogen mixing ratio is dominated by H_2O up to 70–80 km (e.g., Hunten and Strobel, 1974). Odd hydrogen in this height range is in equilibrium, dominated by

$$H_2O + O(^1D) \xrightarrow{h\nu} OH + OH, \qquad (10.23)$$

$$HO_2 + OH \rightarrow H_2O + O_2. \qquad (10.24)$$

Reaction (10.24), the dominant sink of HO_x and therefore crucially important for all reaction systems, has been observed in the laboratory, but there are serious questions about the value of the rate coefficient. Once this is settled, and there are several observations of the actual OH concentration, the situation will be much clearer. Even now, the uncertainty in the rate of Reaction (10.24), about a factor of 10 total, reflects only a factor of 3 into the HO_x density, since Reaction (10.24) goes as its square.

OZONE

The downward flux of ozone through the tropopause can be estimated by application of Eq. (10.18). At 15 km, from the data of Figure 10.3, df/dz is +0.13 ppm/km, or 1.3×10^{-12} cm^{-1}; with $K = 2500$ cm^2 sec^{-1} and $n_a = 4.9 \times 10^{18}$ cm^{-3}, we obtain $\phi = -1.6 \times 10^{10}$ molecules cm^{-2} sec^{-1}. This is about 0.16 percent of the total production, as suggested in Figure 10.6.

MULTIDIMENSIONAL TRANSPORT

One-dimensional models, even with the modifications discussed above, are too limited for many purposes. There have been various attempts to remove the limitations and even to eliminate altogether the need for an arbitrary, empirical eddy-diffusion coefficient. Global-circulation models, or GCM's, exist, which try to solve in detail the dynamical and thermal equations for the atmosphere. Nevertheless, they still do not lack arbitrary parameters; eddy-diffusion coefficients (or viscosities) must be used to damp instabilities and to describe processes occurring at scales smaller than the computational grid. For some purposes, the coarseness of this grid can also be an important limitation. Mahlman (1973) has nevertheless made some experiments in which a passive tracer was injected into a model and its motion followed. His results encourage further use of this technique as models improve. Another approach, using spherical-harmonic expansions instead of grid points, has been discussed by Cunnold *et al.* (1975).

The chief difficulty with three-dimensional models is their sheer magnitude. The dynamical part of the problem taxes today's biggest computers; addition of a large number of reactive constituents has not even been attempted and may not be for many years. A more fruitful approach for the nearer future is to study the motion of passive tracers and form averages that can be used to derive eddy coefficients for one- and two-dimensional models.

A two-dimensional model results from taking of averages around latitude circles and describes vertical and meridional motions by a combination of eddy transport and mean motions. This approach, although conceptually attractive, suffers from the difficulty of obtaining the large number of transport parameters needed. Such problems are well described by Danielsen and Louis in Chapter 9 and by Hesstvedt (1974). It is probable that observations of the actual atmosphere will never suffice for a satisfactory determination. What seems far more promising, although not yet attained, is the use of a suitable three-dimensional model to calculate the required "observations."

10.6 CONCLUDING REMARKS

It is often asked, "Why does everything seem to destroy ozone in the stratosphere; aren't there some things that make it instead?" Concentrated ozone is an explosive almost as powerful as TNT; it is easily decomposed but can only be made with a large input of energy. More than a thousandth of the solar energy reaching the earth is already going into ozone production. To capture significantly more would be difficult indeed. Reaction (10.14), photolysis of NO_2, can be a significant source of odd oxygen in the troposphere because it absorbs in the blue and near ultraviolet. The same reaction in the stratosphere is not a net source, because formation of NO_2, according to Reaction (10.7), uses up an ozone molecule.

The other terrestrial planets offer striking examples of control of ozone by minor constituents. Both Mars and Venus have atmospheres dominated by CO_2, whose photolysis is a large source of odd oxygen. Yet the amounts of O_2 and O_3 are remarkably small. The Martian troposphere is

photochemically analogous to the terrestrial stratosphere, but it is not nearly so dry. Correspondingly, HO_x aeronomy dominates the destruction of both odd and even oxygen and the recycling of CO_2. In the polar regions, H_2O, and probably at least one component of HO_x, are frozen out, and large increases of the ozone abundance are observed (cf. Hunten, 1974).

On Venus, O_2 is still rarer, and ozone presumably also. Here a major influence appears to be HCl, which has been observed spectroscopically. Indeed, chlorine aeronomy was studied in some detail for Venus before the ideas were applied to earth. Details may be found in a recent article by Sze and McElroy (1975).

It is an overstatement to suggest that Mars and Venus are examples of what earth could become in an age of pollution. Nevertheless, they are real examples of major ozone reduction by trace constituents, and at the very least they help us to keep our intellectual tools sharp.

Kitt Peak National Observatory is operated by the Association of Universities for Research in Astronomy, Inc., under contract with the National Science Foundation.

REFERENCES

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

Christie, A. D. (1973). Secular or cyclic change in ozone, *Pure Appl. Geophys. 106–108*, 1000.

Cicerone, R. J., D. H. Stedman, and R. S. Stolarski (1975). Estimate of late 1974 stratospheric concentration of gaseous chlorine compounds (ClX), *Geophys. Res. Lett. 2*, 219.

Climatic Impact Committee (1975). *Environmental Impact of Stratospheric Flight*, National Academy of Sciences, Washington, D.C.

Crutzen, P. J. (1974). Estimates of possible future ozone reductions from continued use of fluoro-chloro-methanes (CF_2Cl_2, $CFCl_3$), *Geophys. Res. Lett. 1*, 205.

Crutzen, P. J. (1976). Upper limits on atmospheric ozone reductions following increased application of fixed nitrogen to the soil, *Geophys. Res. Lett. 3*, 169.

Crutzen, P. J., I. S. A. Isaksen, and G. C. Reed (1975). Solar proton events: stratospheric sources of nitric oxide, *Science 189*, 457.

Cunnold, D., F. Alyea, N. Phillips, and R. Prinn (1975). A three-dimensional dynamical-chemical model of atmospheric ozone, *J. Atmos. Sci. 32*, 170.

Dütsch, H. U. (1971). Photochemistry of atmospheric ozone, *Adv. Geophys. 15*, 219.

Dütsch, H. U. (1974). The ozone distribution in the atmosphere, *Can. J. Chem. 52*, 1491.

Ellsaesser, H. W. (1974). Water budget of the stratosphere, in *Proceedings of the Third Conference on the Climatic Impact Assessment Program*, A. J. Broderick and T. M. Hard, eds., U.S. Department of Transportation, DOT-TSC-OST-74-15, pp. 273–283.

Fabian, P. (1974). Comments on "A photochemical theory of tropospheric ozone" by W. Chameides and J. C. G. Walker, *J. Geophys. Res., 79*, 4124.

Farmer, C. B., O. F. Raper, and R. H. Norton (1976). Spectroscopic detection and vertical distribution of HCl in the troposphere and stratosphere, *Geophys. Res. Lett. 3*, 13.

Grobecker, A. J., S. C. Coroniti, and R. H. Cannon, Jr. (1974). *Report of Findings. The Effects of Stratospheric Pollution by Aircraft*, U.S. Department of Transportation, Report DOT-TST-75-50, December.

Hesstvedt, E. (1974). Reduction of stratospheric ozone from high-flying aircraft, studied in a two-dimensional photochemical model with transport, *Can. J. Chem. 52*, 1592.

Hunten, D. M. (1974). Aeronomy of the lower atmosphere of Mars, *Reviews of Geophys. Space Phys. 12*, 529.

Hunten, D. M. (1975a). Residence times of aerosols and gases in the stratosphere, *Geophys. Res. Lett. 2*, 26.

Hunten, D. M. (1975b). Vertical transport in atmospheres, in *Atmospheres of Earth and the Planets*, B. M. McCormac, ed., D. Reidel Publishing Co., Dordrecht, Holland, pp. 52–72.

Hunten, D. M. (1975c). Estimates of stratospheric pollution by an analytic model, *Proc. Nat. Acad. Sci. U.S.A. 72*, 4711.

Hunten, D. M., and D. F. Strobel (1974). Production and escape of terrestrial hydrogen, *J. Atmos. Sci. 31*, 305.

Johnston, H. S., D. Kattenhorn, and G. Whitten (1976). Use of excess carbon-14 data to calibrate models of stratospheric ozone depletion by supersonic transports, *J. Geophys. Res. 81*, 368.

Junge, C. E. (1962). Global ozone budget and exchange between stratosphere and troposphere, *Tellus 14*, 363.

Liu, S. C., R. J. Cicerone, T. M. Donahue, and W. L. Chameides (1976). Limitation of fertilizer induced ozone reduction by the long lifetime of the reservoir of fixed nitrogen, *Geophys. Res. Lett. 3*, 157.

Machta, L., K. Telegadas, and R. J. List (1970). The slope of surfaces of maximum tracer concentration in the lower stratosphere, *J. Geophys. Res. 75*, 2279.

Mahlman, J. D. (1973). Preliminary results from a three dimensional general circulation/tracer model, in *Proc. of the 2nd Conf. on CIAP*, Dept. of Transportation, Washington, D.C., pp. 321–337.

McConnell, J. C., and M. B. McElroy (1973). Odd nitrogen in the atmosphere, *J. Atmos. Sci. 30*, 1465.

McElroy, M. B., S. C. Wofsy, J. E. Penner, and J. C. McConnell (1974). Atmospheric ozone: possible impact of stratospheric aviation, *J. Atmos. Sci. 31*, 287.

McElroy, M. B. (1976). Chemical processes in the solar system: A kinetic perspective, *MTP International Review of Science*, D. R. Herschbach, ed., Butterworths, London.

Mitra, S. K. (1952). *The Upper Atmosphere*, The Asiatic Society, Calcutta, 713 pp.

Molina, M. J., and F. S. Rowland (1974). Stratospheric sink for chlorofluoromethanes—chlorine atom-catalyzed destruction of ozone, *Nature 249*, 810.

Nicolet, M. (1975). Stratospheric ozone: an introduction to its study, *Rev. Geophys. Space Phys. 13*, 593.

Pittock, A. B. (1974). Ozone climatology, trends and the monitoring problem, in *Proc. of the Internat. Conf. on Structure, Composition and General Circulation of the Upper and Lower Atmospheres and Possible Anthropogenic Perturbations*, Jan. 14–25, 1974, International Assoc. Meteorology and Atmospheric Physics, Toronto, Vol. 1, pp. 455–466.

Ruderman, M. A., and J. W. Chamberlain (1973). *Origin of the Sunspot Modulation of Ozone: Its Implications for Stratospheric NO Injection*, JSS-73-3, Inst. for Defense Analyses.

Sze, N. D., and M. B. McElroy (1975). Some problems in Venus' aeronomy, *Planet. Space Sci. 23*, 763.

Thomas, R. J. (1974). Total mixing ratios, *Planet. Space Sci. 22*, 175.

Wofsy, S. C. (1976). Interactions of CH_4 and CO in the earth's atmosphere, *Ann. Rev. Earth Planet. Sci. 4*, 441.

Wofsy, S. C., and M. B. McElroy, (1973). On vertical mixing in the upper stratosphere and lower mesosphere, *J. Geophys. Res. 78*, 2619.

Wofsy, S. C., M. B. McElroy, and N. D. Sze (1975). Freon consumption: implications for atmospheric ozone, *Science 187*, 535.