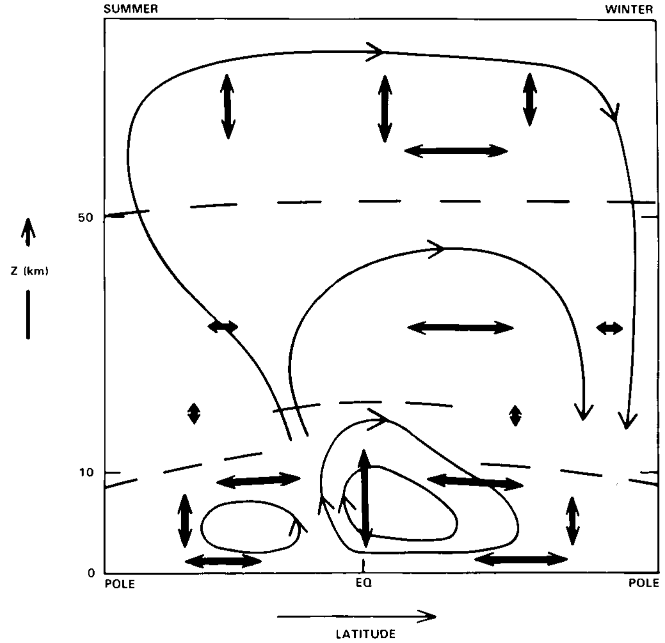
CHAPTER 6

DYNAMICAL PROCESSES



Panel Members

R.A. Plumb, Chairman

D.G. Andrews
M.A. Geller
W.L. Grose
A. O'Neill
M. Salby
R.A. Vincent

CHAPTER 6

DYNAMICAL PROCESSES

TABLE OF CONTENTS

6.0 INTRODUCTION

The distribution of ozone is maintained by combined radiative, chemical and dynamical processes. In particular, transport processes determine the movement of ozone precursor constituents such as N_2O and CH_4. Dynamics also influences the distribution of constituents such as NO_x, HO_x and ClX which determine ozone loss processes, and of course transport processes act on O_3 itself determining its distribution throughout much of the atmosphere. For these reasons, it is crucial that models that aim to simulate present day ozone correctly as well as those that aim to make quantitatively correct predictions of future ozone distributions include quantitatively correct transport. This, in turn, requires that the atmospheric dynamics responsible for effecting this transport be modeled in a quantitatively correct manner.

Transport in the atmosphere can be effected by dynamical processes acting on all spatial and temporal scales. For instance, the Brewer-Dobson circulation is a global scale circulation consisting of rising motion in the tropics and descending motion at high latitudes. In this chapter, it is emphasized that this type of global-scale pattern of rising and descending motions owes its existence to the presence of asymmetric eddy motions. These can be large-scale eddies (e.g., baroclinic waves and planetary waves) or small-scale eddy motions (gravity waves and turbulence). The fact that the existence of the global scale overturning circulations is so intimately connected to eddy transport processes has important implications for the formulation of transport models.

The last comprehensive discussion of the state of research of the ozone problem took place four years ago [WMO, 1981]. Several notable advances in our understanding of dynamics and transport have been made during these four years. These have included new observational analyses, improvements in our modeling capabilities, and an enhanced theoretical basis for our understanding of transport. Brief discussion of some of these is given in this introductory section with a more complete discussion given in the body of the chapter.

On the observational side, there have been analyses of new satellite data. For instance, Salby *et al.* [1984] have used the high vertical resolution LIMS data to identify equatorial Kelvin waves in both the stratosphere and mesophere. Also, the new availability of observations of O_3, H_2O, NO_2 and HNO_3 from LIMS and of N_2O and CH_4 from SAMS has made it possible to use these observations to evaluate current ideas on the interplay between chemistry and transport. There have also been new analyses of more conventional nadir-viewing stratospheric satellite data. Perhaps, the analysis of this type that has had the greatest impact on our conceptual understanding was that of McIntyre and Palmer [1983] in which they showed observational evidence for Rossby wave breaking. Several multi-year climatological analyses have also appeared in the literature during this past four years. This has allowed the average annual cycle in Northern and Southern Hemisphere stratospheric climatological variables to be determined as well as the interannual variability in the climatology. Indeed, our knowledge of the circulation of the Southern Hemisphere stratosphere has improved dramatically in the recent years, principally as a result of global satellite coverage. Our understanding of this half of the globe, however, still lags behind that of the Northern Hemisphere.

Ground-based radar and lidar observations have also contributed significantly to our knowledge of gravity wave processes in the mesosphere. Vincent and Reid [1983] have provided measurements of vertical momentum fluxes that demonstrate consistency with the theoretical concepts of gravity-wave-induced drag on the mean zonal flow proposed by Lindzen [1981] and Matsuno [1982]. There have also been the beginnings of the assemblage of a radar-derived gravity wave climatology for the middle atmosphere. While little has yet appeared in the literature, these works have motivated researchers to become increasingly concerned about what role gravity waves might play in the stratosphere.

On the theoretical side, there has recently been a much greater appreciation of the balance between advective and dispersive transport processes [Mahlman, 1985; Plumb and Mahlman, 1986]. The latter work sets forth the framework from which a consistent specification of advection and dispersion can be made in two-dimensional photochemical models. A framework with which the one-dimensional eddy diffusion used by photochemical modellers can be reconciled with large-scale dynamics has also been given by Holton [1985] and Mahlman *et al.* [1985].

Along with the developing appreciation of the climatology of the middle atmosphere has come the realization that the traditional conceptual separation of the three-dimensional structure into zonal mean and eddy components may have severe limitations. This is especially true during Northern winter, when the polar vortex may be shifted well off the pole. In the absence of any theoretical framework to simplify the transport problem, the full implications of this three-dimensionality can only be addressed via direct analysis of the observed circulation and constituent distributions or in three-dimensional models.

There has been a great deal of activity in middle atmosphere modeling over the last four years. Among other things, this has led to an increasing realization that, despite the large scale of flow patterns typical of stratospheric maps, general circulation models of the middle atmosphere need to use horizontal resolutions that were previously thought to be necessary only in the troposphere. There has been a slow, but appreciable, advancement in the inclusion of transport and photochemical processes in three-dimensional models. A continuing problem with middle atmosphere general circulation models is their pathology in producing excessively cold polar night temperatures with associated excessively strong westerlies. There remains some uncertainty over the role of improved parameterization of radiative processes versus inadequate dynamics in giving rise to this discrepancy. Even if the source of this problem is taken to be inadequacies in dynamics, it is unclear what role is played in this by gravity waves as opposed to large-scale waves.

Such shortcomings place limitations on the use of current models for assessment studies. However a more serious practical limitation is that inclusion of complex, realistic photochemistry into such models remains prohibitively expensive (except for short integration times), even for the fastest present day computers. It seems we must await a future generation of computers, or a theoretical advance in the simplification of three-dimensional transport, before three-dimensional assessments become a reality.

The remainder of this chapter is organized as follows. Section 6.1 deals with the observed structure of the middle atmosphere. In this section, there is a brief discussion of the techniques used to observe the middle atmosphere. This is followed by discussions of the zonally-averaged and eddy structure of the middle atmosphere; in particular, our rapidly expanding appreciation of transient motions is discussed in some detail. Discussions of observed middle atmosphere seasonal variations and interannual variability follow. Section 6.2 contains a discussion of our theoretical understanding of the middle atmosphere circulation. The point is made very strongly that our present understanding of middle atmosphere dynamics tells us a great deal about middle atmosphere transport processes. In particular, the relationship between the residual mean circulation and eddy processes is underscored. The present status of general circulation models of the middle atmosphere is discussed in Section 6.3. Section 6.4 reviews some recent satellite-based observational studies of stratospheric transport processes, as revealed by the behavior of constituents and of potential vorticity (a dynamical tracer). The theory of global transport and its representation in transport models is discussed in Section 6.5. Finally, Sections 6.6 and 6.7 contain some comments on future directions in this field of research and a summary of the main points of the chapter.

6.1 CLIMATOLOGICAL MEANS AND VARIABILITY OF THE MIDDLE ATMOSPHERE

6.1.1 Introduction

Climatologies of the middle atmosphere have been constructed from radiosonde/rocket soundings and other surface-based information. The set of observations due to Groves [1970] was adopted as the CIRA [1972] standard atmosphere. Few radiosondes reach levels above the lower stratosphere and rocket data are very sparse. It is therefore impossible to obtain a complete global picture of the structure and dynamics of the middle atmosphere from these types of observations alone. With the launch of satellite-borne radiometers, it became possible to map the temperature distribution in three dimensions and to follow its changes from day to day. From the thermal wind relation between the temperature and motion fields, the wind field in the extratropics can be derived to a good approximation if the height of some base pressure level in the atmosphere is available, and a wide range of dynamical studies can then be conducted. One major advantage of satellite measurements is their global coverage. Note however that the process of building up the height/wind analysis from some base level means that errors in the base height field (and there remain substantial gaps in ground-based observations, especially over the Southern oceans) permeate the entire analysis. Further major advantages of satellite measurements are their approximate spatial uniformity and the fact that they are made with a single instrument. These features are particularly important for dynamical studies as relevant diagnostic quantities can involve a high order of spatial differentiation, and global analyses based on readings of uniform reliability are consequently vital. Recent climatological studies of the middle atmosphere which have employed satellite data are those of Labitzke and Barnett [1979], McGregor and Chapman [1979], Hamilton [1982b], Geller *et al.* [1983, 1984], and Hirota *et al.* [1983a]. A new interim CIRA atmosphere is to be published in the Handbook for MAP, Vol. 16.

Confidence in the reliability of satellite-derived data has been increased through comparison of quantities derived from satellite data with equivalent quantities obtained from conventional meteorological analyses and/or other, independent, satellite data. Smith [1982] compared wind fields, wave structures and other derived quantities using SCR and LIMS satellite data and NMC analyses based on conventional (radiosonde) data below, and satellite data above, 10mb. The comparisons showed good qualitative agreement, although SCR and LIMS data could not be compared directly since the instruments were not operating at the same time. Kohri [1981] had earlier performed similar studies using a single month (Dec.1975) of LRIR data.

A somewhat more stringent and systematic attempt at intercomparison was made by the PMP-1 working group using SAMS and LIMS data from the Nimbus 7 satellite, SSU data from TIROS-N and conventional analyses (mostly based on radiosonde data) from NMC, ECMWF and the Free University of Berlin. The study focused on 8 individual days, including quiet and disturbed times; results are discussed in Rodgers [1984]. Overall the study found good qualitative agreement and, in some cases, good quantitative agreement between the various data sources. However, differences in horizontal and/or vertical gradients can produce marked differences in derived quantities; significant differences were apparent for momentum flux. A second study by this working group [Rodgers and Grose, 1985] has focused on monthly mean comparisons for periods between 1979-81 using the same data sources. Conclusions were similar to those of the earlier study.

Noticeably absent to date is a detailed comparison focusing on the Southern Hemisphere. This situation is perhaps a result of the sparseness of radiosonde data for the Southern Hemisphere and the fact that many of the sampling strategies for remote measurements from satellites have focused on the dynamically more active Northern Hemisphere.

Satellite measurements do not give readings at a point but are averages over a volume of the atmosphere [see Houghton *et al.*, 1984, Chapter 6]. Moreover, they are made asynoptically: measurements at different points are taken at different times, so that some form of time interpolation is also implied. During periods when the circulation is highly contorted important features may not be adequately resolved and other measuring techniques with higher spatial and temporal resolution then provide useful supplementary information. Two such techniques are lidar and radar which give good vertical and temporal resolution from ground sites, though clearly not global coverage.

Ground-based techniques offer refined vertical and temporal resolution, which makes them well suited to studying small-scale phenomena such as gravity waves. Their value in observing large-scale processes is limited by the localness of the measurements. An important development has been that of the MST (Mesosphere-Stratosphere-Troposphere) radar technique introduced by Woodman and Guillen [1974]. Radars of this type operating at VHF (30-300 MHz) and UHF (0.3-3 GHz) have the capability of measuring winds, waves and turbulence parameters with time and vertical resolutions of the order of 1 to 2 min and 50 to 300 m, respectively. Echoes are obtained from refractive index irregularities caused by temperature and density fluctuations in the troposphere and stratosphere and by fluctuations in electron density in the mesosphere. The Doppler shift of the echoes gives the line-of-sight velocity. Three radar beams enable the measurement of the complete velocity vector, as these radars have the unique ability to measure vertical velocities with reasonable accuracy. Depending on the radar sensitivity, it is possible to make almost continuous wind soundings at heights up to 25 to 30 km. In the mesosphere, echo occurrence is, however, often intermittent with diurnal, seasonal and geographic changes being found. Because of their excellent temporal and spatial resolutions MST radars have provided invaluable information about short period gravity waves and turbulent motions. However, long term studies of the dynamics of the mesosphere have so far been restricted to a few sites. Rottger [1980] and Gage and Balsley [1984] have provided recent reviews of MST techniques and their capabilities.

Partial reflection (PR) radars which operate at frequencies near 2 MHz share with the MST technique the ability to measure the structure of tides, gravity waves and the time-mean flow in the mesosphere. Although their time and height resolutions are moderate (a few min and a few km) PR radars have proved very reliable and, at some stations, almost continuous observations have been made over several years, giving insight into the morphology of mesospheric waves and winds [Vincent, 1984b].

Lidars which use Rayleigh scattering from atmospheric molecules to measure neutral density and temperature constitute an important new development since they provide information about the wave parameters which is complementary to that provided by radars [Chanin and Hauchecorne, 1981]. Time and height resolutions of 15 min to 1 hr and 100 m to 1 km can be achieved and since lidars can observe much of the stratosphere and mesosphere, they can study the 'gap' region between 30 and 55 km, which is the most difficult region to investigate with MST radars. To date, most information about small-scale dynamics in this range has come from infrequent rocket and balloon soundings.

This section on climatology is an account of the global structure of the large-scale circulation of the middle atmosphere that has been inferred from satellite measurements, and includes a description of the small-scale wave and turbulence processes determined by radar. The section begins with a discussion of the basic state of the middle atmosphere as represented by the zonal mean structure of temperature and winds at the solstices (Section 6.1.2). This zonal mean structure is, of course, an incomplete description of the three-dimensional state. Even on a monthly-averaged basis, the wintertime stratospheric wind field may depart substantially from zonally uniform flow, because of the presence of large-amplitude quasi-stationary waves. While it has become almost conventional to describe the climatology of stratospheric waves in terms of these time-averaged components, day-to-day variations are in fact substantial. Estimates of the ratio of variance in transient wave components to that in the time-mean waves range from 1 to 1 [Tomatsu, 1979] to 3 to 1 [Geller *et al.*, 1984; Hirota*et al.*, 1983a]. Consequently any meaningful description of climatology must account for fluctuations about the time-averaged state of the stratosphere. The emergence of continuous global satellite observations in recent years has greatly facilitated our knowledge of such features. Understanding has been advanced by these measurements for both transient extratropical disturbances and for equatorial wave modes. Many of these features were previously unobserved; indeed the identification of some of these modes has been a major success of limb-viewing satellite measurements.

The extratropical planetary wave field may be characterised, at least qualitatively, in terms of two basic contributions to frequency spectra of individual wavenumbers. At low frequencies, variance has the form of a red continuum, falling off more or less systematically with decreasing period, representing the baroclinic, ''quasi-stationary'' waves, which are largely confined to midlatitudes of the winter hemisphere. These disturbances propagate upwards through the stratosphere and play a central role in coupling the stratosphere with the troposphere. They are ''quasi-stationary'' in the sense that they fluctuate, in amplitude and phase, about climatologically preferred values. Their climatological structures are presented as monthly mean waves in Section 6.1.3. It must be emphasised, however, that the departures of the actual wave field from these means can be dramatic, especially during high-latitude warming events (see Section 6.1.7).

Adjacent to and superimposed on this red continuum are discrete spectral peaks appearing at westward frequencies for the smallest zonal wavenumbers. As will be discussed in Section 6.1.4, several of these discrete peaks may be manifestations of planetary normal modes. These waves are of global extent (at least at the lowest levels) and their vertical structure is for the most part barotropic.

Most of the wave activity of the middle atmosphere is believed to originate in the troposphere. Forcing of the large-scale, quasi-stationary waves is ultimately provided by the orographic effects of large mountain ranges and the thermal effect of longitudinal variations in diabatic heating. However the tropospheric planetary wave field incorporates a substantial transient component. While the nature of these disturbances has not been addressed in stratospheric studies, such considerations have been examined extensively in tropospheric investigations. A number of these [Eliasen and Machenhauer, 1965; 1969; Ahlquist, 1982; Lindzen *et al.*, 1984] indicate that a sizable fraction of the unsteady planetary wave activity at periods shorter than two weeks is associated with the planetary normal modes. Some of these studies [also Hirooka and Hirota, 1985] suggest that during sporadic amplifications, these transient components in combination may attain amplitudes as large as the stationary components. It is important that the relative contribution from these two elements of planetary wave activity be more fully understood.

A distinct set of wave motions is observed in the tropics. These ''equatorial waves'' are of large zonal scale, but confined in latitude about the equator. They have large horizontal phase speeds and short vertical wavelengths, which precluded their observation from satellites until the advent of limb-viewing instruments. These waves, which play a fundamental role in the momentum budget of the tropical middle atmosphere, are described in Section 6.1.5.

Gravity waves and tides are of relatively small amplitude in the lower stratosphere. Because of their high frequency and rapid vertical propagation they are, in the main, attenuated relatively little as they propagate upwards and therefore they attain large amplitudes in the mesosphere, where they become dominant influences on the mean circulation. Section 6.1.6 discusses our current knowledge of these motions and of the turbulence which they generate in the mesosphere.

The final two sections describe the variability of the circulation. The account of the seasonal cycle in Section 6.1.7 draws attention to the extreme departures from time-averaged fields that can occur in the middle atmosphere. On longer time scales, it has come to be recognised that the year-to-year variability of the circulation is more marked in the stratosphere than in the troposphere. Therefore, the circulation in any one year or small group of years should not be taken as fully representative; some findings on the inter-annual variability of the middle atmosphere are presented in Section 6.1.8.

6.1.2 Zonally Averaged Structure of Wind and Temperature

The zonal mean temperatures and zonal winds of the stratosphere and mesosphere for January and July are shown in Figures 6-1. and 6-2. Except at high latitudes in the lower stratosphere during winter, the temperature increases with height in the stratosphere (due to ozone heating) and decreases with height in the mesosphere. In the middle and upper stratosphere, the temperature decreases monotonically from the summer to the winter pole, whereas the temperature gradient has the opposite sign in the mesosphere.

In Section 6.2, below, this climatological temperature structure will be compared with that predicted on the basis of a radiative-convective equilibrium calculation. Some of the differences are substantial, with the winter pole as much as 100 K warmer than the radiative-convective equilibrium value. As will be discussed in Section 6.2, these differences are indicative of dynamical influences.

There are pronounced differences between hemispheres during winter in the stratosphere and lower mesosphere at high latitudes, which reflect different levels of dynamical activity. In the northern hemisphere, the polar temperatures are warmer through most of the stratosphere and colder near the stratopause than they are in the southern hemisphere. During summer, the stratopause in the southern hemisphere is about 5 K warmer than in the northern hemisphere, probably because the Earth is closer to the Sun in January than it is in July.

Zonal winds in winter and summer are westerly and easterly, respectively, with maxima located at mid latitudes in the lower mesosphere. Easterlies occupy equatorial latitudes in the stratosphere at the solstices, except in the lower stratosphere during the westerly phase of the quasi-biennial oscillation (Section 6.1.8). The middle and upper mesosphere are marked by a rapid decrease with height of both the westerly and easterly winds. This is thought to be due to the drag imposed by the dissipation of gravity waves, which propagate upwards from the troposphere [e.g., Lindzen, 1981]. Another major feature of the wind distribution is that zonal winds are typically much stronger during winter in the southern hemisphere than they are in the winter northern hemisphere. This is consistent (through thermal wind balance) with the colder polar temperatures in the southern hemisphere.

The zonal-mean meridional velocity $\bar{v}$ in the middle atmosphere can only be inferred indirectly from satellite measurements of radiance. In the mesosphere, however, long term-radar measurements have been used to obtain $\bar{v}$. Figure 6-3 shows an example of the annual changes of the meridional and zonal flow in the mesosphere observed with a partial reflection radar. It illustrates the variability due to the passage of large-scale waves but overall the mean meridional flow is from the summer to winter pole, in agreement with theoretical expectations (see 6.2.4). The peak magnitudes are between 5 and 20 ms^{-1} and occur at the mesopause [Nastrom *et al.*, 1982; Vincent, 1984b]. Mean vertical velocities, $\bar{w}$, are extremely difficult to measure because of their small magnitudes. However, Nastrom *et al.* [1985] found reasonable agreement in the lower atmosphere between MST radar measurements of $\bar{w}$ and values computed by other

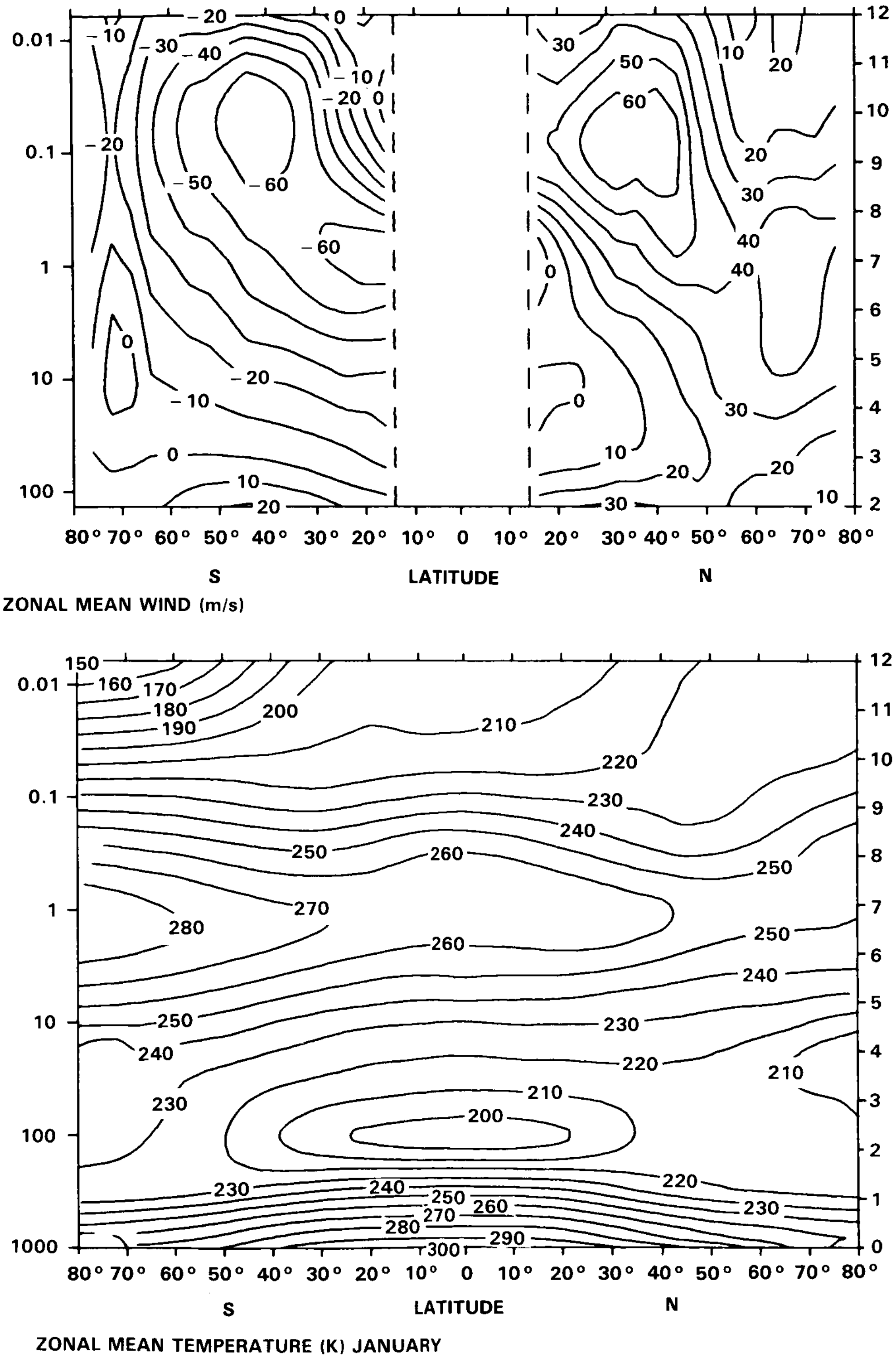


Figure 6-1. Cross sections [pressure (mbar)-latitude] of zonal mean geostrophic wind (ms^{-1}) and zonal mean temperature (K) for the average over 5 years of the monthly means for January. The data are from the combined SCR/PMR retrieval made at the University of Oxford for the period January 1973 to December 1974 and July 1975 to June 1978. (Supplied by J.J. Barnett and M. Corney).

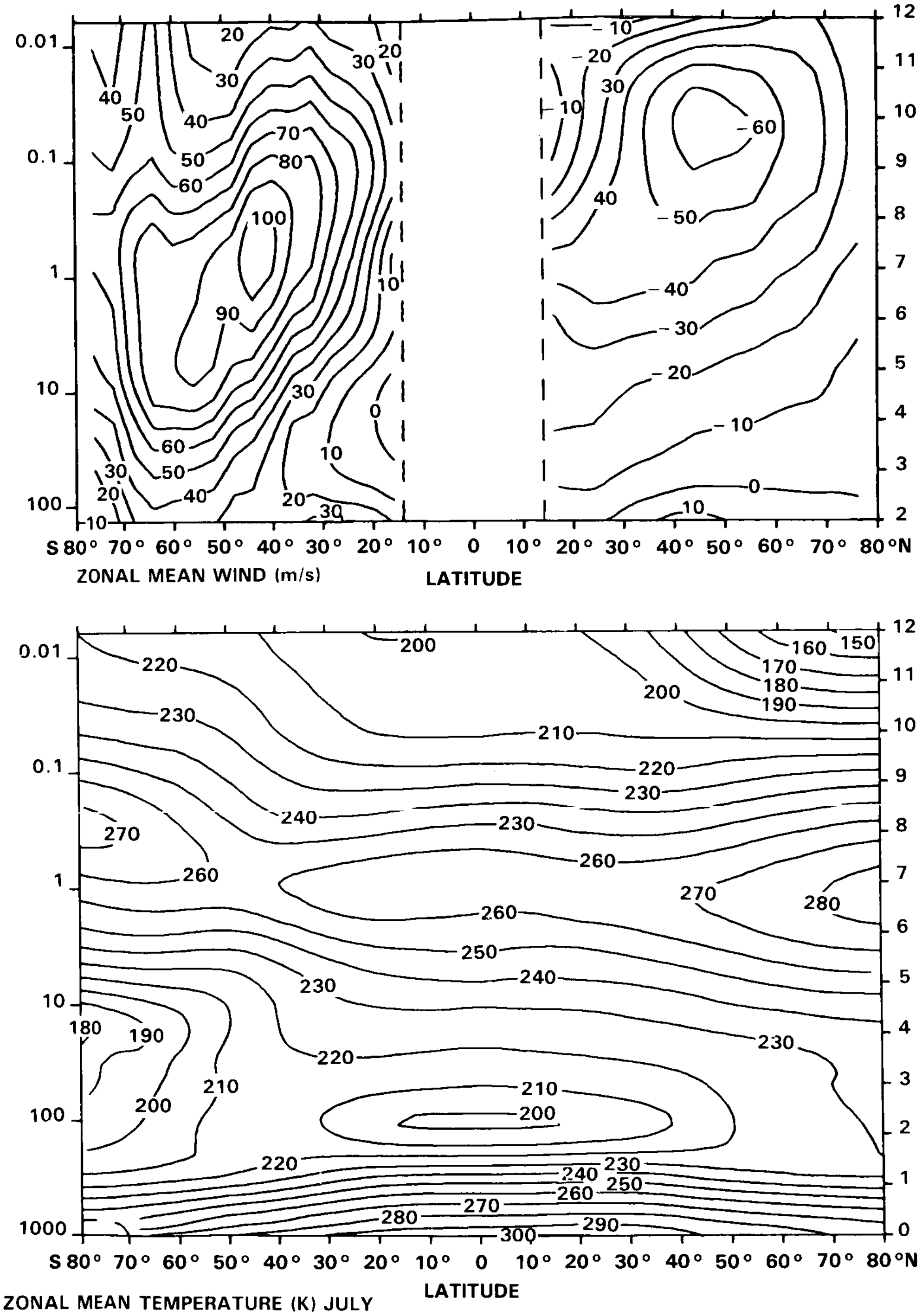


Figure 6-2. As Figure 6-1 but for July.

means. In the mesosphere, Balsley and Riddle [1984] measured long-term velocities of about 25 cm s^{-1} with the Poker flat [65 °N] MST radar which are much larger in magnitude and opposite in sign to the values expected from the observed $\bar{v}$ and temperature structure. Theory predicts a rising circulation over the summer pole and sinking motion over the winter pole (Section 6.2.4).

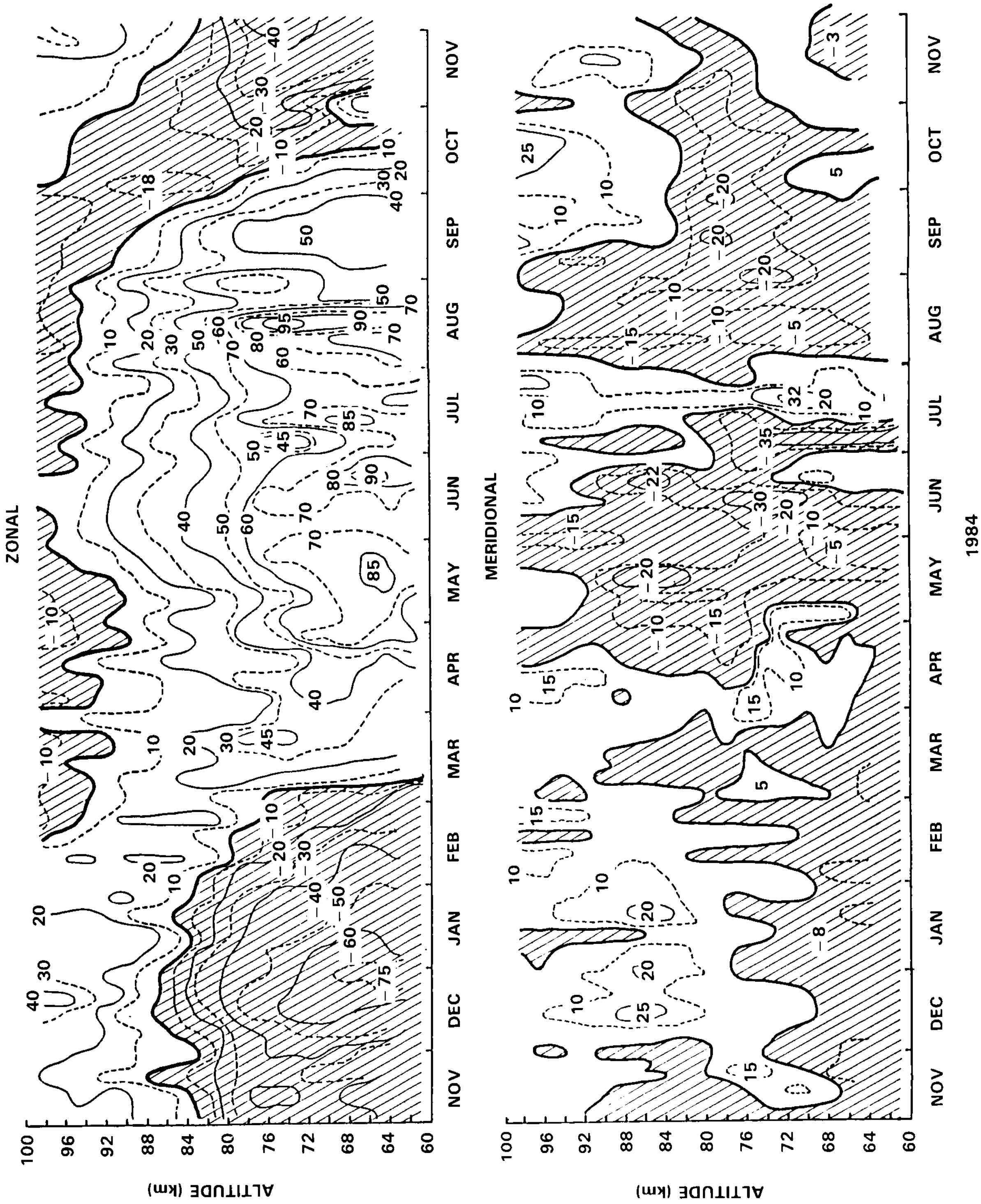


Figure 6-3. Zonal and meridional winds (ms^{-1}) in the mesosphere measured with a PR radar at Adelaide (35°S, 128°E). Shaded areas indicate regions of westward and southward flow.

6.1.3 Monthly-Mean Wave Structure

Departures of the circulation of the middle atmosphere from zonal symmetry are conventionally summarized by presenting the structure of harmonic waves around latitude circles. Whereas any zonally varying field can be represented mathematically as a Fourier spectrum of waves, such descriptions are most valuable when the field contains only the gravest components, e.g., wavenumbers 1 and 2, as is typically the case for the stratospheric height field. Such behavior is consistent with the planetary wave theory of Charney and Drazin [1961] who showed that only the longest waves generated in the troposphere can penetrate well into the middle atmosphere in winter. Recent studies, however, caution against an overreliance on linear theory when interpreting the complete behavior of the wave components [McIntyre and Palmer, 1983, 1984; Clough *et al.*, 1985].

The amplitudes and phases of wavenumbers 1 and 2 are shown in Figure 6-4 and 6-5. They are computed for the solstices from the monthly-mean geopotential height field. Maximum amplitudes are at high latitudes near the stratopause and decay with height in the mesosphere. This decrease may be due to a number of factors: dissipation of disturbances is more rapid in the mesosphere than in the stratosphere [Dickinson, 1973]; the vertical penetration of disturbances may be limited by the wind distribution resulting in equatorward refraction [Karoly and Hoskins, 1982]; large-scale disturbances may be dissipated to some extent in the stratosphere by the generation of small scales of motion during 'wave breaking' [McIntyre and Palmer, 1983,1984]. In summer, monthly-mean amplitudes are very small since large-scale quasi-stationary disturbances are confined to the troposphere by easterly winds in the stratosphere [Charney and Drazin, 1961]. In both hemispheres during winter, the amplitudes of wave 2 are less than those of wave 1. Both waves exhibit a westward tilt with height at midlatitudes in the stratosphere and less tilt in the mesosphere. The ridges and troughs associated with the waves are aligned from the south-west to the north-east in the northern hemisphere, and from the south-east to the north-west in the southern hemisphere. The tilt with height and the trough/ridge alignment correspond to poleward eddy momentum and heat fluxes, and upward and equatorward Eliassen-Palm fluxes (see Section 6.2).

There are inter-hemispheric differences in the amplitudes of the monthly-mean waves that can be associated with the differences in zonal mean temperatures and winds mentioned earlier (see Section 6.2). The northern hemisphere winter stratosphere is more disturbed than is that of the southern hemisphere, reflecting the more asymmetric circulation in the the northern hemisphere winter troposphere. As a consequence of this difference, theoretical arguments (Section 6.2) predict warmer polar temperatures and weaker zonal winds for the northern hemisphere during winter than during the southern winter. In the stratosphere, there is a stronger meridional gradient of the phase of the waves in the southern hemisphere (Eliassen-Palm fluxes tilt more equatorwards). This is consistent with the prediction of linear theory that waves will be more strongly refracted equatorwards in the broader and stronger westerly jet of the southern hemisphere.

The time-averaged waves (principally waves 1 and 2) combine with the zonal mean flow in the middle stratosphere to produce the monthly-mean winter circulation shown in Figures 6-6 and 6-7. The main features of the circulation are of a much larger scale than in the troposphere. In the northern hemisphere (Figure 6-6), a persistent feature near 180°E is the so-called Aleutian High. It is an anticyclonic circulation containing air drawn polewards from lower latitudes [Clough *et al.* 1985]. Strong fluctuations in its intensity occur throughout winter, particularly during sudden warmings. The circulation in the southern hemisphere (Figure 6-7) is far less disturbed and the pole is correspondingly colder. There is, however, a smaller interhemispheric difference in the temperature of the coldest air near the center of the westerly vortex.

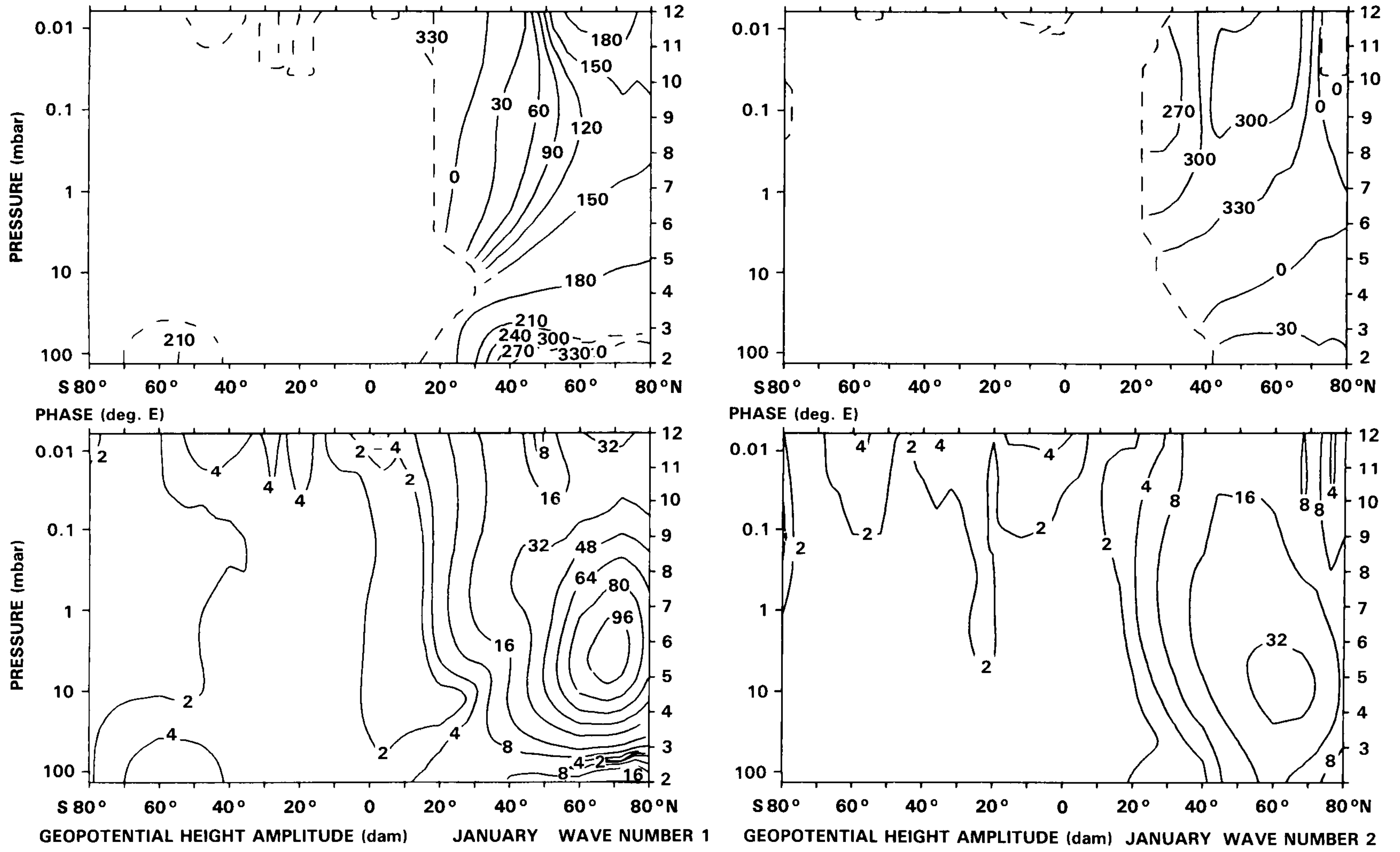


Figure 6-4. Cross sections of the amplitudes (dam) and phases (degrees east) of geopotential height waves 1 and 2 computed for January using the same data used for Figures 6-1 and 6-2. Scale on left is pressure (mbar). Courtesy of J.J. Barnett and M. Corney.

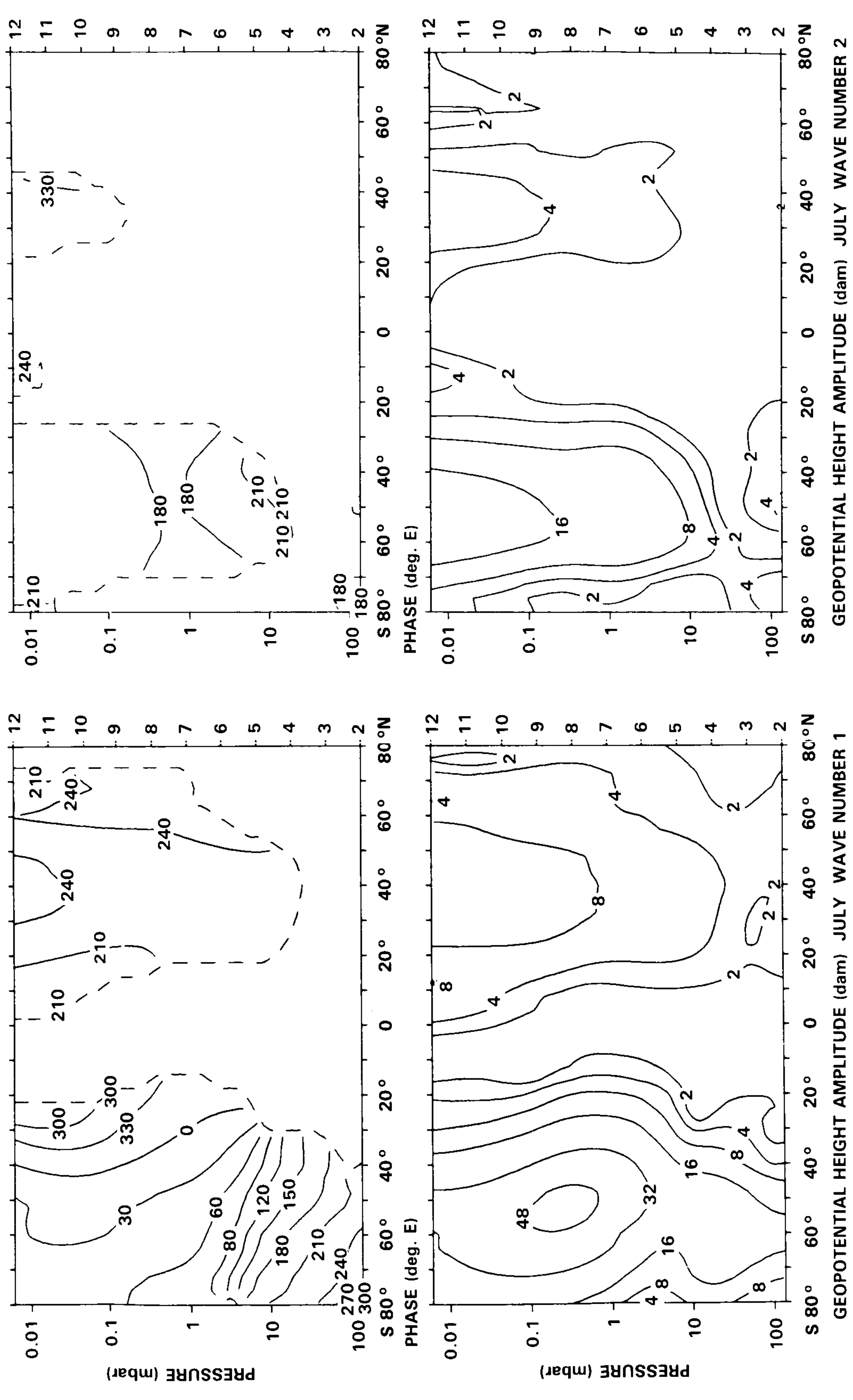


Figure 6-5. As Figure 6-4 but for July.

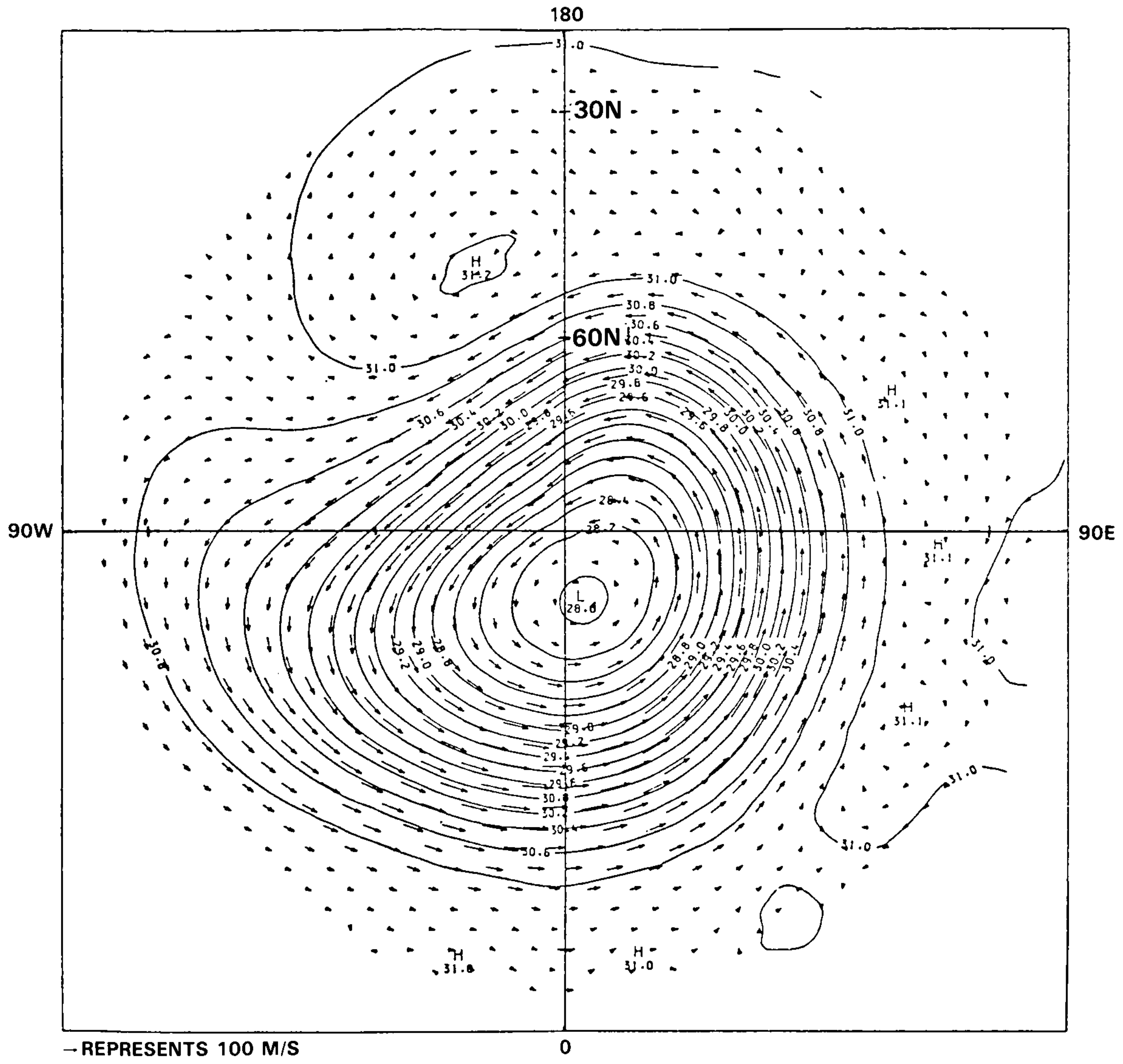


Figure 6-6. Polar stereographic map at 10 mb of monthly mean geopotential height (km) and geostrophic winds (m s^{-1}) for the northern hemisphere for January 1981. Data obtained from a stratospheric sounding unit on the satellite NOAA-6. Analysis made by the Middle Atmosphere Group, Meteorological Office, U.K.

6.1.4 Extratropical Transients

It was noted earlier that day-to-day departures of the actual stratospheric circulation from the monthly-mean picture can be substantial. Some of the variance in these transient components may be associated with planetary normal modes. Unlike the forced, quasi-stationary components, atmospheric normal modes are predicted to be rather independent of forcing details. Their structures and the frequencies at which they appear are consequently quite robust [Geisler and Dickinson, 1976; Salby 1981a,b]. This property, combined with the availability of global satellite measurements, has now permitted the identification of a number of these features.

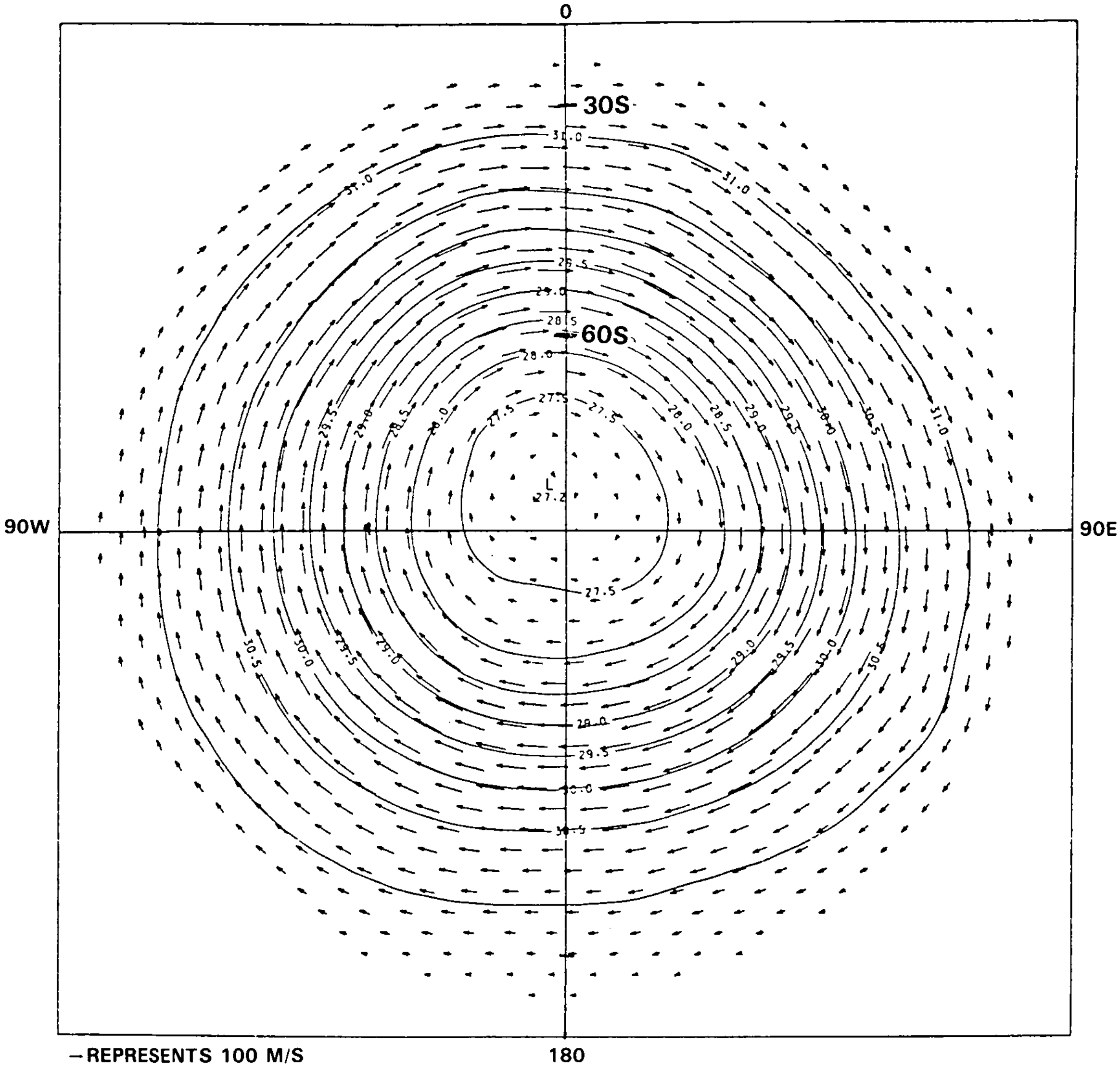


Figure 6-7. As Figure 6-6 but for the southern hemisphere for July 1981.

The first normal mode to be convincingly identified and the most widely documented is the 5-day wave [Madden and Julian, 1972b; 1973]. It has a wavenumber 1 global structure which is symmetric about the equator and propagates westward around the earth once in 5 days. Evidence for the disturbance has emerged from meteorological analyses [Madden, 1978], global satellite temperature retrievals [Rodgers, 1976a; Hirota and Hirooka, 1984], and from radar wind measurements [Salby and Roper, 1980; Hirota *et al.*, 1983b]. The last study indicated a sizable perturbation of the zonal mean flow by this component, particularly in the summer easterlies where quasi-stationary wave activity is absent.

With the increased availability of global satellite measurements, additional normal modes have been identified. Figure 6-8 shows the wavenumber 2 analogue of the 5-day wave, the 4-day wave at the stratopause, as derived from Tiros-N SSU by Hirota and Hirooka [1984]. The satellite observations have provided a fairly compelling picture of a global disturbance, nearly symmetric between the hemispheres. Higher

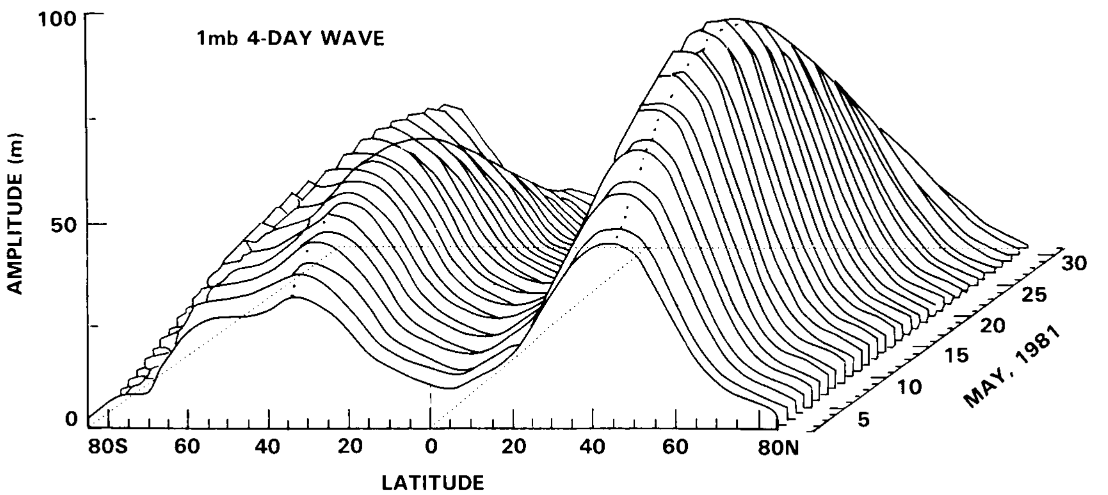


Figure 6-8. Structure and evolution of the 4-day wave at stratopause level, observed by Tiros-N SSU and NOAA-A HIRS. Corresponds to wavenumber 2 at 1 mb, bandpassed between 3.8-4.5 days over May 1981. Systematic retrogression is simultaneously observed. [After Hirota and Hirooka, 1984.]

degree normal modes have also emerged in Nimbus-5 SCR, Nimbus-6 PMR, Tiros-N SSU, NOAA-A HIRS data [Chapman and Peckham, 1980; Venne, 1984; Hirooka and Hirota, 1985]. Figure 6-9 shows the antisymmetric 10-day wave of wavenumber 1 at stratopause derived from Tiros-N SSU by Hirooka and Hirota [1985]. These features agree rather well with the theoretical properties of normal modes.

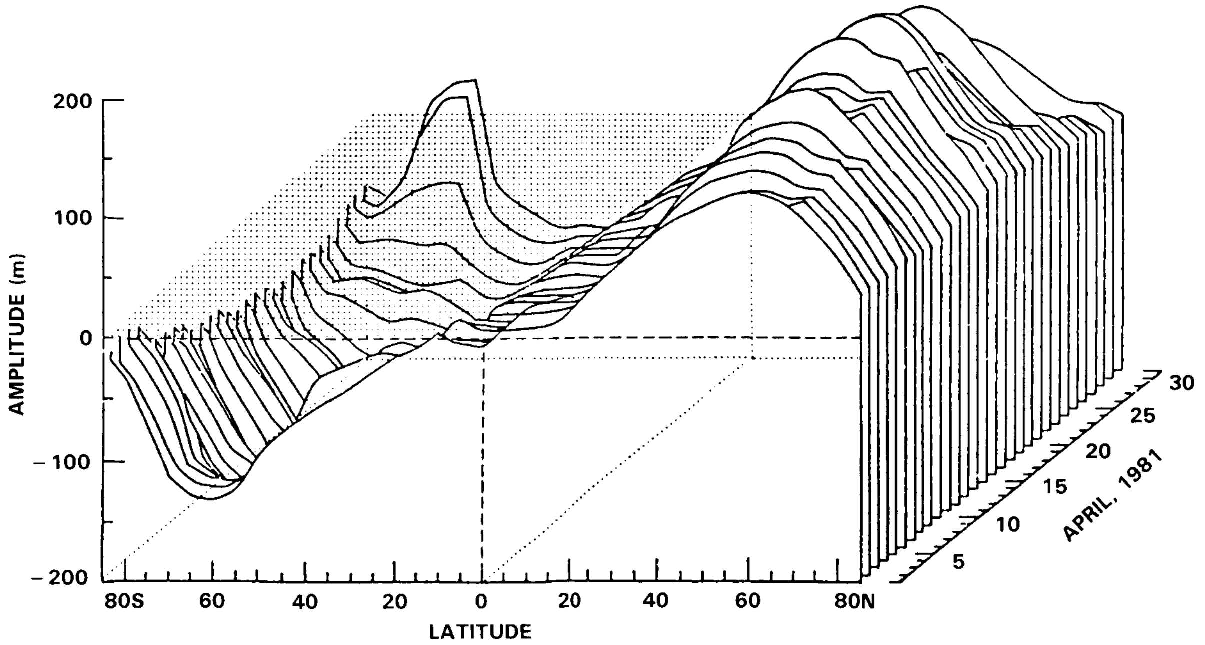


Figure 6-9. Structure and evolution of the 10-day wave at stratopause level, observed by Tiros-N SSU. Wavenumber 1 geopotential at 1 mb band passed about 9.2 days during April 1981. [Adapted from Hirooka and Hirota, 1985.]

Higher degree modes have larger amplitudes than the graver modes, which may be explained on the basis of their longer periods [Salby, 1984a]. Their structures, however, become markedly asymmetric in the solstice stratosphere as a result of their exclusion from the summer hemisphere by strong easterlies. Hirooka and Hirota [1985] have presented evidence of the amplification of slower modes prior to disturbed conditions in the stratosphere. In some respects this association with stratospheric warmings parallels the suggestion of Lindzen *et al.* [1984] that such components may play a role in blocking phenomena in the troposphere. However, considerably more study will be needed to establish such relationships.

One transient feature, which has attracted considerable attention, is the 16-day wave [Miller, 1974; Madden, 1978]. This disturbance is most evident in Northern Hemisphere winter; however it has been documented by both observations and GCM integrations in all seasons [Speth and Madden, 1983; Hayashi and Golder, 1983; Hirooka and Hirota, 1985]. Northern Hemisphere statistics [Madden, 1978; 1983] indicate a structure and evolution quite similar to the second symmetric normal mode of wavenumber 1. At least some studies have been able to document a global modal structure. Madden and Labitzke [1981] have identified the disturbed period of January 1979 as a manifestation of this phenomenon; this conclusion is supported by Lindzen *et al.'s* [1984] analysis of tropospheric behavior. In addition to its sizable amplitude, the 16-day wave is notable for being a recurrent climatological feature. It populates over half of an eight-winter ensemble of Northern Hemisphere winter dates [Madden, 1983]. Interference between this traveling wave and the stationary forced wave has been suggested as being responsible for the quasi-regular fluctuations which have been observed in eddy transports at stratospheric levels [Hirota and Sato, 1969; Madden, 1975] and upward migration of wave amplitude vacillations [Madden, 1977]. The same mechanism has been offered in connection with the switching of the EP flux vector prior to stratospheric warmings [Palmer, 1981a]. Quiroz [1979b] proposed a similar interference process, but involving different wavenumbers.

In wavenumber 1, the observed vacillations take the form of pulsations in wave amplitude, spaced 2-3 weeks apart. The fluctuations in eddy amplitude and attending transports which occur at various levels are not in phase, but rather tend to migrate upwards. Madden [1983] has recently shown that the observed behavior of such eddy transport vacillations is captured reasonably well by a simple barotropic normal mode propagating across the observed westward tilting stationary wave. Figure 6-10 shows the phase lag of the observed vacillation in heat and momentum fluxes as functions of height. The same quantities, shown for the simple model (solid line) are in reasonable agreement. Hirooka and Hirota [1985] have demonstrated with global TIROS-N SSU data that such fluctuations can induce vacillations in the zonal-mean circulation. Similar results were obtained from numerical calculations with a baroclinic beta-plane model [Garcia and Geisler, 1981]. From the synoptic viewpoint such oscillations in wavenumber 1 and their response in the zonal-mean flow correspond to a wobbling and perhaps translation of the vortex about the pole. Leovy *et al.* [1985] have reported such behavior in connection with the scavenging of ozone-rich air in low latitudes by the vortex when it is displaced off the pole. Once material is entrained into the perturbed flow configuration, a substantial transport to high latitudes is facilitated.

Conceptually such vacillations in eddy transports may be viewed as the traveling wave modulating the stationary wave fluxes. Its effect is to transform a steady uniform field of eddy fluxes into one which is localised into "capsules" which migrate upwards [Salby and Garcia, 1985]. Observations at a fixed level thus indicate a succession of pulses in wave activity and transport. At a particular point, the EP flux has a transient component which orbits about the time-mean vector (Figure 6-11). During the vacillation, eddy fluxes increase and decrease and may reverse direction, e.g., equatorward switching to poleward. Such behavior promotes mean flow deceleration through enhanced EP flux convergence which follows from the poleward convergence of meridians and focusing of wave activity [O'Neill and Youngblut, 1982].

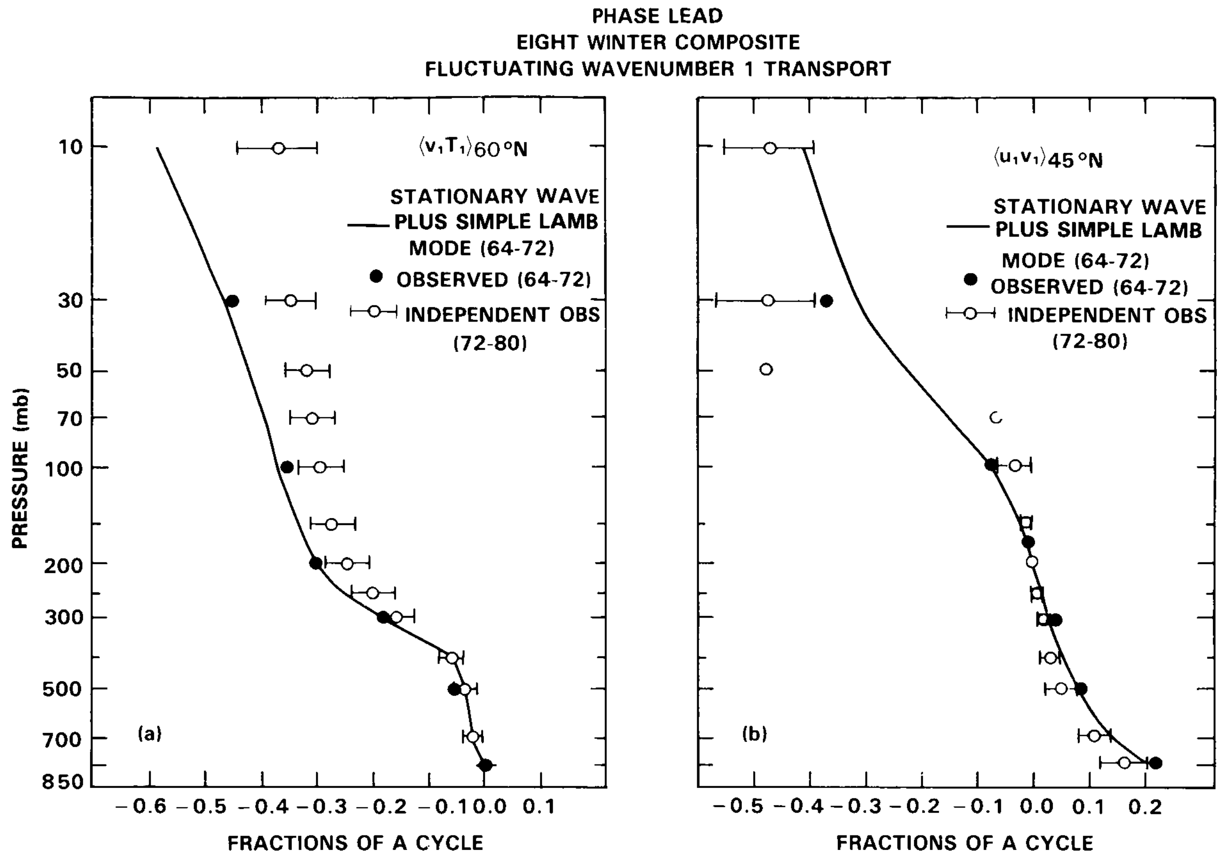


Figure 6-10. Vacillations in eddy transports induced by interference between traveling and stationary waves. (a) phase lead of heat flux vacillation at 60°N; (b) phase lead of momentum flux vacillation at 45°N. Vacillation resulting from interference between observed 8-winter (1964-1972) composite 16-day wave and stationary wave (closed circle). Vacillation resulting from interference between simple Lamb mode and observed (1964-1972) stationary wave (solid line). Vacillation resulting from interference between *independent* 8-winter (1972-1980) 16-day wave derived from cross spectral analysis (1/23-1/12 cpd) and previous stationary wave (open circle). [Adapted from Madden, 1983].

Since this process is responsible for local and instantaneous wave amplifications, it may lead to strong nonlinearity in the form of wave breaking [McIntyre and Palmer, 1984] at upper levels [Salby, 1984b].

Although the extratropical stratosphere is largely believed to be dynamically stable, it is important to recognize that such conditions may be violated locally. Polar regions where the stabilizing influence of the planetary vorticity gradient is weak are particularly prone to such behavior. In such regions, the potential vorticity gradient will always be susceptible to being reversed by curvature in the local flow field and thus satisfy the conditions for instability [Charney and Stern, 1962]. A disturbance which may fall into this category is the polar eastward moving 4-day wave, discussed by Venne and Stanford [1982] in Nimbus-4 and Nimbus-5 SCR data. Prata [1984] has recently documented the wavenumber 2 component of this disturbance with Nimbus-5 SCR and Nimbus-6 PMR data. Calculations performed in realistic zonal shear [Hartmann, 1983] yield unstable polar modes which are not dissimilar to these observations.

Another disturbance possibly arising out of instability is the eastward travelling wavenumber 2 anomaly of the Southern Hemisphere [Harwood, 1975; Hartmann, 1976a], as suggested by Hartmann [1984]. The

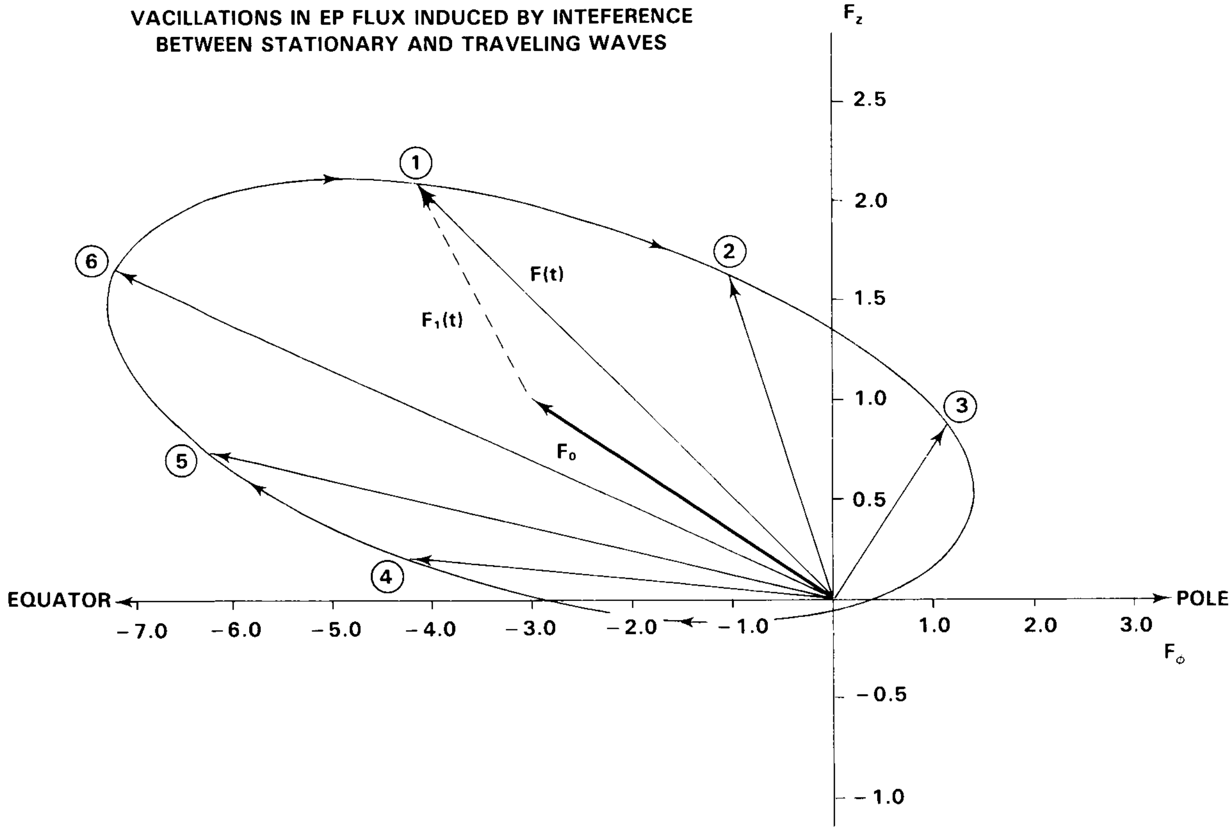


Figure 6-11. Modulation of EP flux vector by barotropic traveling wave migrating over a westward tilting stationary wave. Transient component orbits about time-mean vector. Both instantaneous EP flux components are modulated to zero and double their time-mean values, and reverse direction somewhere during the vacillation. Equatorward propagation switching poleward; upward flux driven between zero and double its time-mean value. [After Salby and Garcia, 1985].

2-day wave, a westward propagating wavenumber 3 disturbance [Muller, 1972; Rodgers and Prata, 1981] has likewise been suggested as arising out of baroclinic instability [Plumb, 1983a], although a normal mode explanation has also been proposed [Salby 1981c]. Velocity amplitudes in excess of 50 m s^{-1} have been observed in the Southern Hemisphere [Craig *et al.*, 1980]; a wave of this magnitude would be expected to induce closed streamlines and therefore to be highly nonlinear, perhaps playing an important role in large-scale transport (Craig *et al.* [1985]; see Section 6.5).

We have concerned ourselves here chiefly with disturbances which are most evident in the stratosphere. However tropospheric transients may be equally important if they induce transport near the tropopause. For example the pentagonal wave [Salby, 1982a; Hamilton, 1983b; Randal and Stanford, 1983], a regularly propagating feature of the Southern Hemisphere troposphere, was shown to induce sizable convergences of total ozone column abundances [Schoeberl and Krueger, 1983].

6.1.5 Equatorial Waves

Since their identification in radiosonde observations [Wallace and Kousky, 1968], Kelvin waves have been recognised as an important ingredient of the momentum budget of the tropical middle atmosphere (see Section 6.2.7). Another equatorial mode, the westward-propagating Rossby gravity wave, was diagnosed in radiosonde observations by Maruyama [1969]. Both of these classical waves are confined to the tropics and propagate vertically with wavelengths of 10 km or less. They are presumably excited by convective heating in the tropical convergence zone [Hayashi, 1970; Holton, 1972; Murakami, 1972; Hayashi, 1974; Hayashi and Golder, 1978]. While such heating is indicative of a broad-band process, Holton [1973] demonstrated a natural discrimination of Kelvin waves in the range of periods 10-20 days [phase speeds of 20-40 ms^{-1}] as is observed in the lower stratosphere. Chang [1976], Hayashi [1976], and Itoh [1977] have shown that the vertical scale of diabatic forcing favors vertical wavelengths twice the depth of the heating, more or less consistent with early observations of equatorial waves.

In addition to these waves, there are considerably faster eastward-propagating disturbances which have recently been recorded at upper stratospheric levels. These were first identified in rocketsonde data by Hirota [1978], although evidence of their existence dates back to Maruyama [1969], and may also be found in Holton [1973] and in Zangvil and Yanai [1980]. They correspond to Kelvin waves with periods of 5-10 days, phase speeds of 50-70 ms^{-1} (twice that of the Wallace and Kousky variety) and vertical wavelengths of about 20 km. Their presence is clearly seen in time-height sections of temperature and zonal wind over Ascension Island (Figure 6-12). Downward phase migration corresponds to upward group propagation. Evidence for the faster, longer vertical wavelength disturbances has also been derived from Nimbus-5 SCR data [Hirota, 1979]. Because of their higher frequency, these fast Kelvin waves are better able to traverse the strong radiative damping of the stratosphere [Dunkerton, 1979] and hence are more likely to play a role in the semi-annual oscillation at upper levels (Hirota, [1980]; see section 6.2.7).

Historically, satellite observation of equatorial waves has not met with great success, largely because of the deep vertical weighting functions and coarse resolution inherent to nadir viewing instruments. However recent limb-viewing configurations, with their narrower weighting functions, have provided an unprecedented opportunity to observe such disturbances. Tropical temperature fields derived with Nimbus-7 LIMS exhibit pronounced eastward variance over the entire stratosphere [Salby *et al.*, 1984]. Many aspects of this

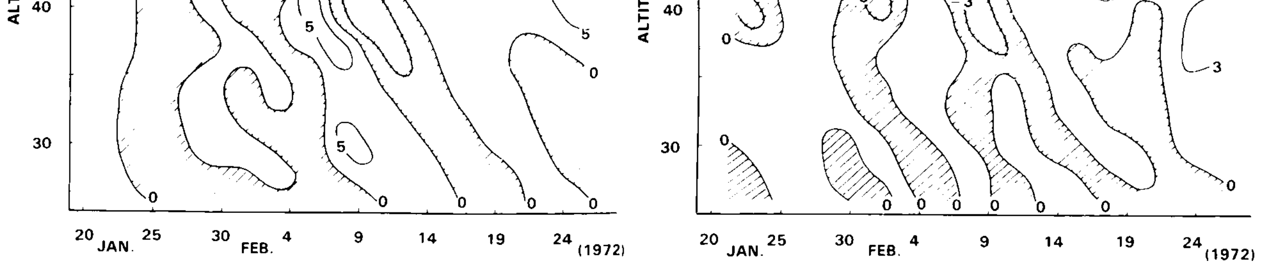


Figure 6-12. Signatures of two-day differenced (a) zonal wind, and (b) temperature over Ascension Island (8°S, 14°W), derived from rocketsonde measurements. [After Hirota, 1978].

wave activity are in accord with dispersion characteristics of equatorial Kelvin waves, although not solely the Wallace and Kousky [1968] variety or even the Hirota [1978] variety. Figure 6-13 shows wavenumber 1 temperature power as a function of latitude for three stratospheric levels. At each of these, unsteady behavior is dominated over the tropics by eastward propagating variance which is localized in particular bands of frequency. Each of these features is nearly symmetric about the equator. The slower of the two is evident at all three levels. It lies in the period-range 6.7-8.6 days, corresponding to phase speeds of 55-70 ms^{-1} and Hirota's variety of Kelvin waves. In the uppermost level shown, a second, apparently distinct, eastward feature emerges in the band 3.5-4.0 days, corresponding to phase speeds of 115-135 ms^{-1}, considerably faster than either the Wallace and Kousky or the Hirota variety.

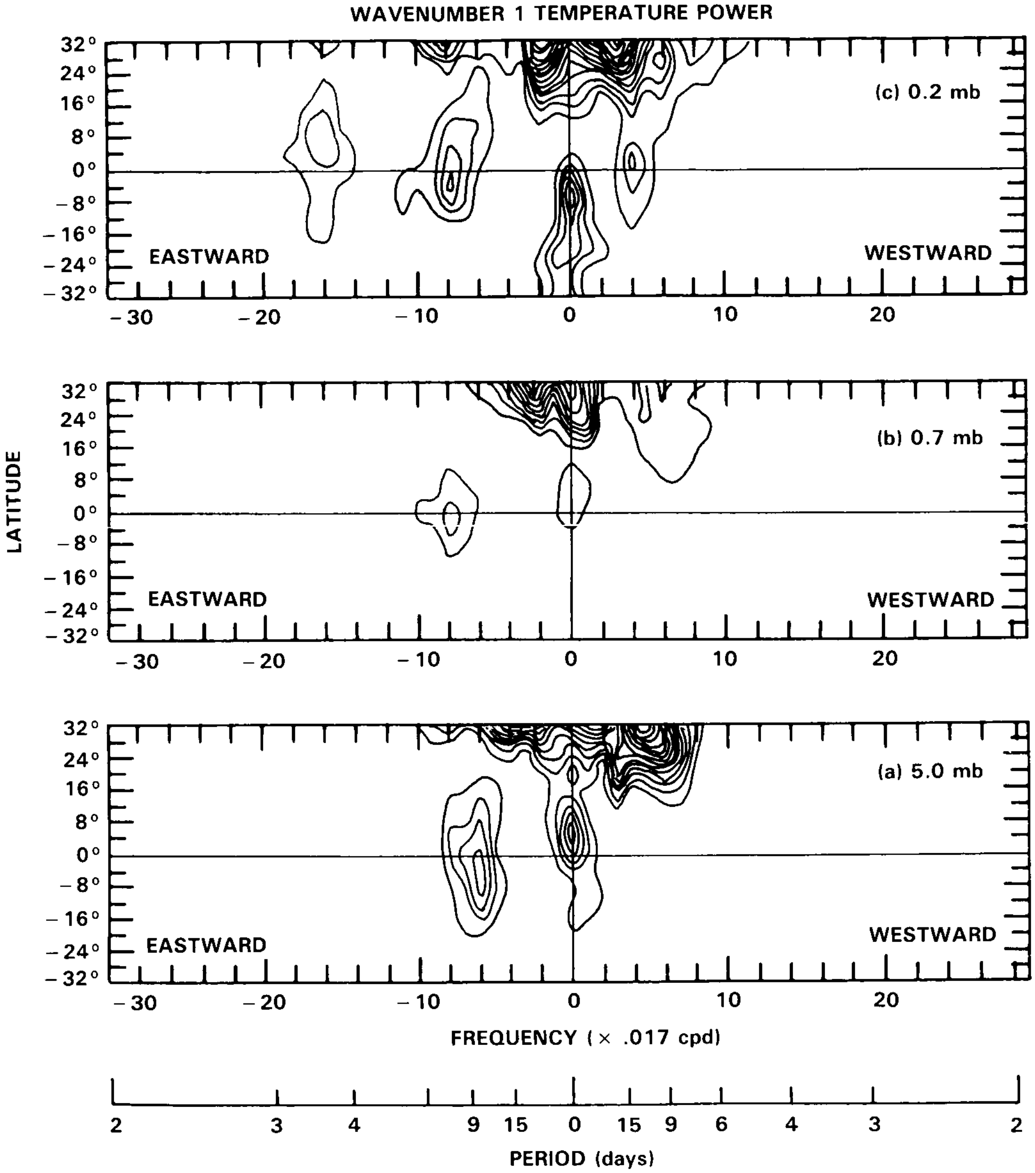


Figure 6-13. Temperature power for wavenumber 1 as a function of frequency and latitude at (a) 5.0 mb, contour increment = 0.2 K^2; (b) 0.7 mb, contour increment = 0.1 K^2; (c) 0.2 mb, contour increment = 0.04 K^2. [After Salby *et al.,* 1984].

Similar features are evident for wavenumber 2. The slowest has phase speeds of 30-40 ms^{-1} commensurate with the variety first reported by Wallace and Kousky [1968]. Its latitude-height structure, decomposed into components symmetric and antisymmetric about the equator (Figure 6-14), has virtually all of the variance captured by the symmetric contribution and maximizes on the equator, consistent with the behavior of Kelvin waves. The observed downward phase migration is consistent with upward wave activity propagation. Vertical wavelengths vary from 7-13 km, in agreement with the observed phase speed and the dispersion relationship for Kelvin waves. By comparison, the ultra-fast feature has a much deeper vertical wavelength, 41 km, and broader structure. However, its behavior is also consistent with the dispersion characteristics of Kelvin waves. The vertical growth of this mode is unabated through the highest level monitored by LIMS, 0.05 mb. This characteristic led Dunkerton [1982b] to speculate upon its role in the momentum budget of the mesosphere. Whereas these recent satellite observations have permitted a rather detailed description of Kelvin wave behavior in the stratosphere, Rossby-gravity waves have yet to be detected in satellite retrievals. Whether they are simply not present at these levels or they remain unresolved in limb-viewing measurements is unclear.

Features strikingly similar to the Kelvin waves derived from satellite retrievals have been diagnosed in GCM integrations (Figure 6-15) by Hayashi *et al.* [1984]. Maxima in each of the three ranges of phase speed appear in tropical temperature spectra. Lateral and vertical structures are similar to those shown from LIMS. In addition the barotropic 16-day wave is prominent at westward frequencies. It has a symmetric global structure similar to that of the classical normal mode [Hayashi and Golder, 1983].

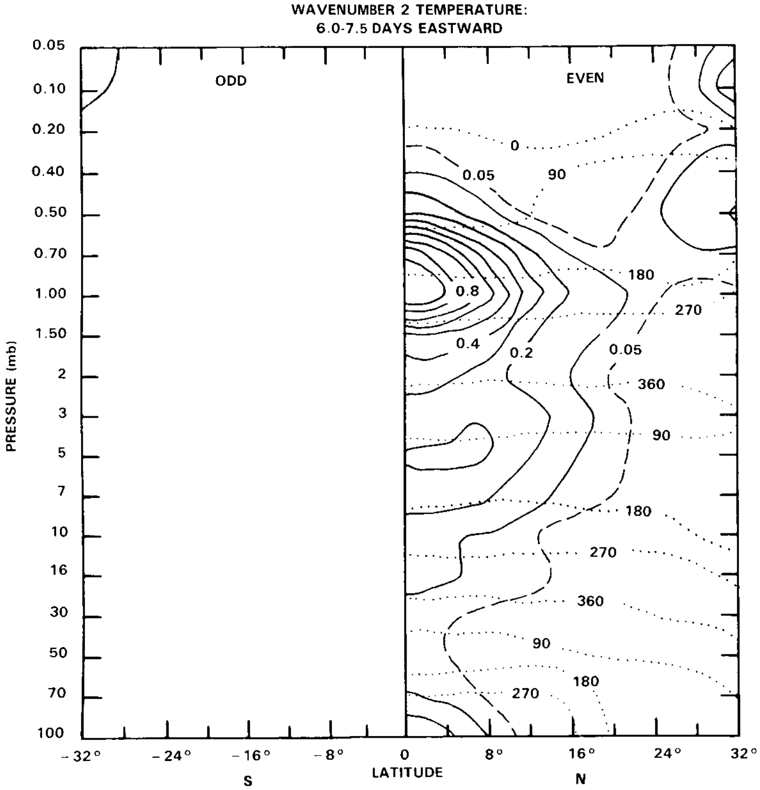


Figure 6-14. Wavenumber 2 temperature power (solid) and phase (dotted) corresponding to eastward periods between 6.0-7.5 days, as functions of latitude and pressure. Decomposed into latitudinal components symmetric and antisymmetric about the equator. Symmetric components displayed at positive latitudes, antisymmetric components at negative latitudes. Power in K^2. [After Salby *et al.,* 1984].

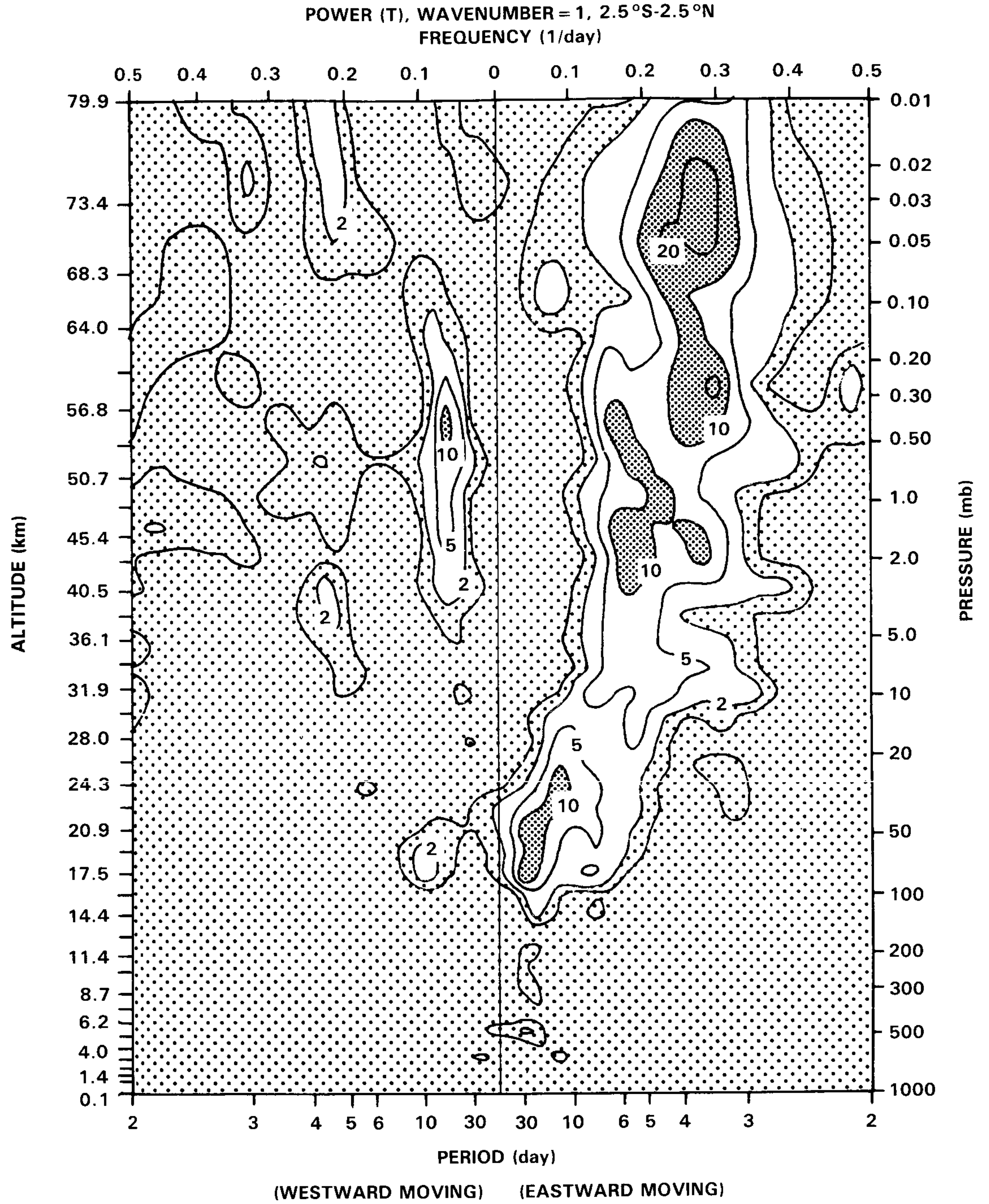


Figure 6-15. Temperature power spectral density for wavenumber 1, as a function of frequency and altitude over the equator, derived from 40-level GCM integration in annual-mean conditions. [After Hayashi *et al.,* 1984].

In combination, these Kelvin disturbances can lead to temperature perturbations of several degrees. From photochemical considerations, it follows that similar perturbations should develop in ozone at upper levels. If such features can be identified, they might serve to validate temperatures derived from UV constituent measurements at upper levels where signal/noise in IR measurements deteriorates.

6.1.6 Tides, Gravity Waves and Turbulence in the Middle Atmosphere

Tides, and gravity waves in particular, are believed to play an important role in determining the large-scale circulation and temperature structure of the middle atmosphere (see Section 6.2) as well as being the predominant source of small-scale turbulence in this region. Solar tides are thermally forced in the troposphere through infrared absorption by water vapor and in the stratosphere through UV absorption by ozone. Although the broad details of the forcing are well understood there are a number of aspects which require further study [Forbes, 1984]. Gravity wave sources are less well known, but the major sources are thought to be in the lower atmosphere and to include topographic forcing, frontal disturbances, convective activity, geostrophic adjustment and shear instabilities. As the waves propagate upwards, their exponential growth, in the absence of reflection and dissipation, leads eventually to wave breakdown and the generation of turbulence.

Much of the early information about gravity waves in the middle atmosphere came from rocket studies which, although somewhat limited because of their infrequency, provided evidence for seasonal and geographic variations in wave activity. In the mesosphere, Theon *et al.* [1967] found a large seasonal variation in wave activity at high latitudes with the maximum activity occurring in winter. At midlatitudes there is less evidence of seasonal changes. Hirota [1984], in a comprehensive study of the climatology of waves in the 20 to 60 km height range made with data taken at the Meteorological Rocket Network stations in the northern hemisphere, found similar seasonal changes as well as evidence for a semiannual cycle at low latitudes (Figure 6-16). High resolution rocket investigations have provided information about the gravity wave structures, although, of course, rocket soundings are too infrequent to resolve their behavior in time. For example, the vertical wavelength of the waves tends to increase as a function of height, with a minimum of about 1.5 km near 60 km increasing to about 3 km near 100 km [Philbrick, 1981]. Both rocket and radar studies suggest a dominant vertical wavelength of about 10 to 15 km in the mesosphere. From these values an intrinsic phase speed $|c-\bar{u}| = N/k_z$ of about 30 ms^{-1} may be inferred [Fritts, 1984] where N is the buoyancy frequency and k_z is the vertical wavenumber. There are only a few indirect estimates of phase velocity in the middle atmosphere but what measurements there are give values in the range 20 to 100 ms^{-1} for mesospheric waves [Vincent and Reid, 1983; Meek *et al.*, 1985a; Vincent, 1985].

Radars are now providing a large amount of information about gravity waves in the mesosphere, albeit with a somewhat limited geographical coverage. Morphologies of wave activity are now becoming available [Meek *et al.*, 1985b]. Power spectral studies of the wave motions show that, averaged over long periods, the energy densities observed at widely separated locations agree well in form and amplitude. Typically the wave energy decreases with frequency f, as f^{-k}, with $k \sim 1.5$-2.0 [Balsley and Carter, 1982; Vincent, 1984a; Meek *et al.*, 1985b]. The rms amplitudes for each wind component are in the range 15 to 20 ms^{-1} so that the total perturbation velocity is approximately 25 to 30 ms^{-1}; rms amplitudes of vertical motions are much smaller than this, being about 1 to 2 ms^{-1}. Most studies show that there is usual ly little, if any, amplitude growth with height so that the energy density $\varrho_0 \overline{v'^2}$ decays as $\exp(-z/h_0)$, where ϱ_0 is the neutral density and $\overline{v'^2}$ is the mean square perturbation velocity. The decay scale h_0, may change with season [Manson *et al.*, 1981] but is in the range 5 to 12 km. The comparable values of $|v'|$ and $|c-\bar{u}|$ as well as the energy decay with height are often taken as indirect evidence of wave saturation in the mesosphere [Fritts, 1984] although wind shears can also cause variations in wave amplitude.

With the data now available, estimates can be made of a number of important wave and turbulence parameters in the mesosphere. From the energy distribution as a function of frequency, Vincent [1984a] found seasonally averaged energy densities of about 5 to $10 \times 10^{-3} Jm^{-3}$ and vertical fluxes of

C - 4

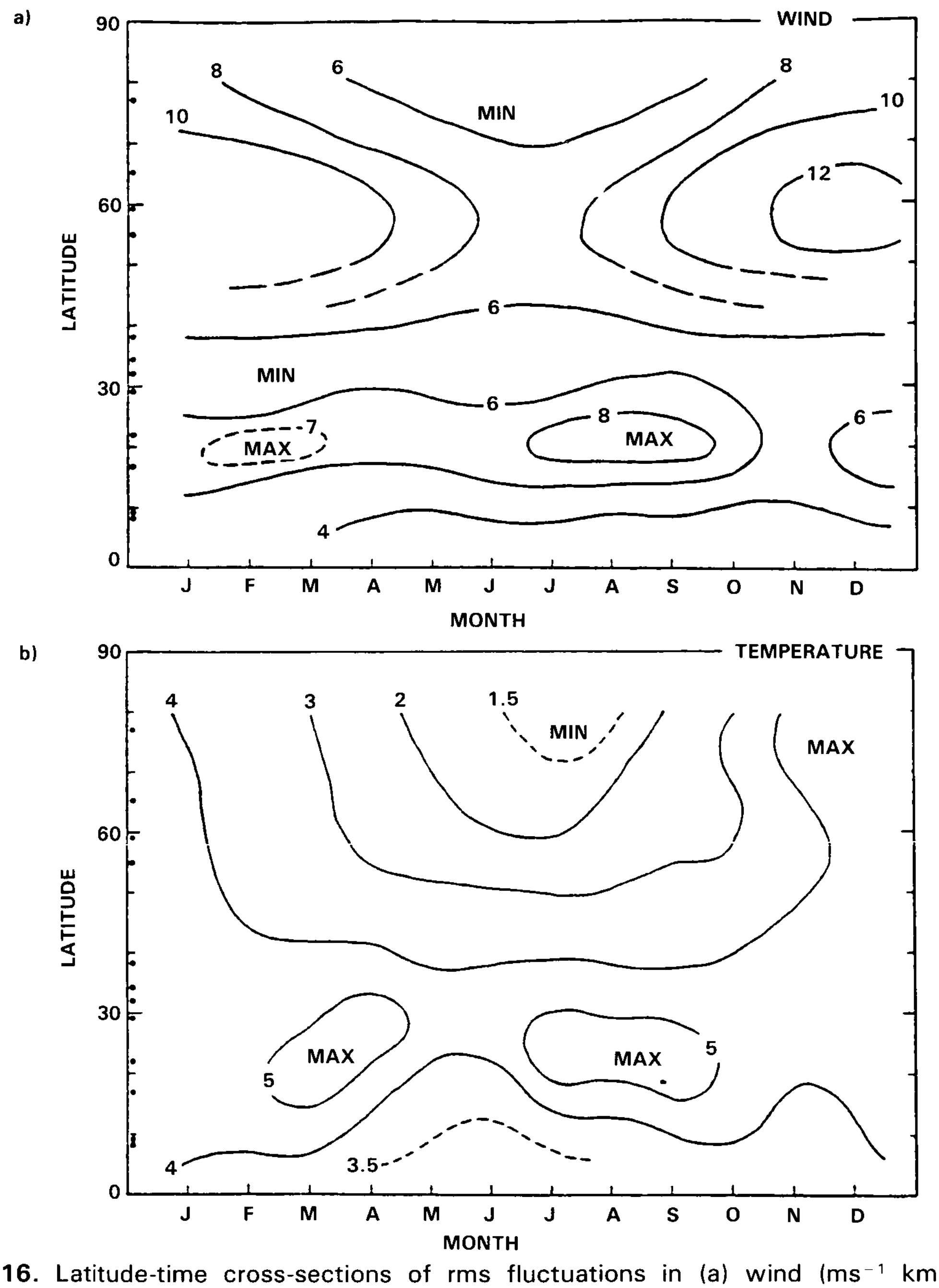


Figure 6-16. Latitude-time cross-sections of rms fluctuations in (a) wind (ms^{-1} km^{-2}) and (b) temperature (K km^{-2}) [After Hirota, 1984].

$\lesssim 10^{-2}$Wm^{-2}. The rate of decay of wave energy has been used to determine energy dissipation rates and diffusion coefficients by the method of Hines [1965]. Typical values for the 80 to 100 km height range are $\epsilon \cong 0.01$ to 0.2 Wkg^{-1} for energy dissipation rates and D $\cong$ 100 to 500 m^{2}s^{-1} for the eddy diffusion coefficients.

More direct evidence for wave saturation and measurements of the associated momentum flux convergence came with the development of a technique in which the momentum fluxes could be measured directly [Vincent and Reid, 1983]. The method uses two coplanar radar beams which point at equal and opposite angles to the zenith and the vertical flux of horizontal momentum is given by the difference between the mean square Doppler velocities observed along the two beams. So far, measurements have been made only at Adelaide but observations tend to support theoretical predictions. Figure 6-17 shows $\overline{u'w'}$ and the inferred flux convergence measured at Adelaide in May 1981 [Vincent and Reid, 1983] and these and other observations show that, averaged over several days, values of $[\overline{u'w'}]$ in the mesosphere are in the range of 1 to 5 m^2s^{-2} and the momentum flux $\varrho_0 \overline{u'w'}$ decreases approximately exponentially with increasing height, indicating a flux convergence. The wave drag is:

$$G = -\varrho_0^{-1}\frac{d}{dz}\left[\varrho_0\overline{u'w'}\right]$$

Thus, on the basis of these measurements, G is found to be of the order of 20 to 80 m s^{-1} day^{-1} and is in the correct sense to balance the Coriolis torques induced by the observed mean meridional flow (see Section 6.2.3). More indirect estimates of wave drag also support these findings [Meek *et al.*, 1985b].

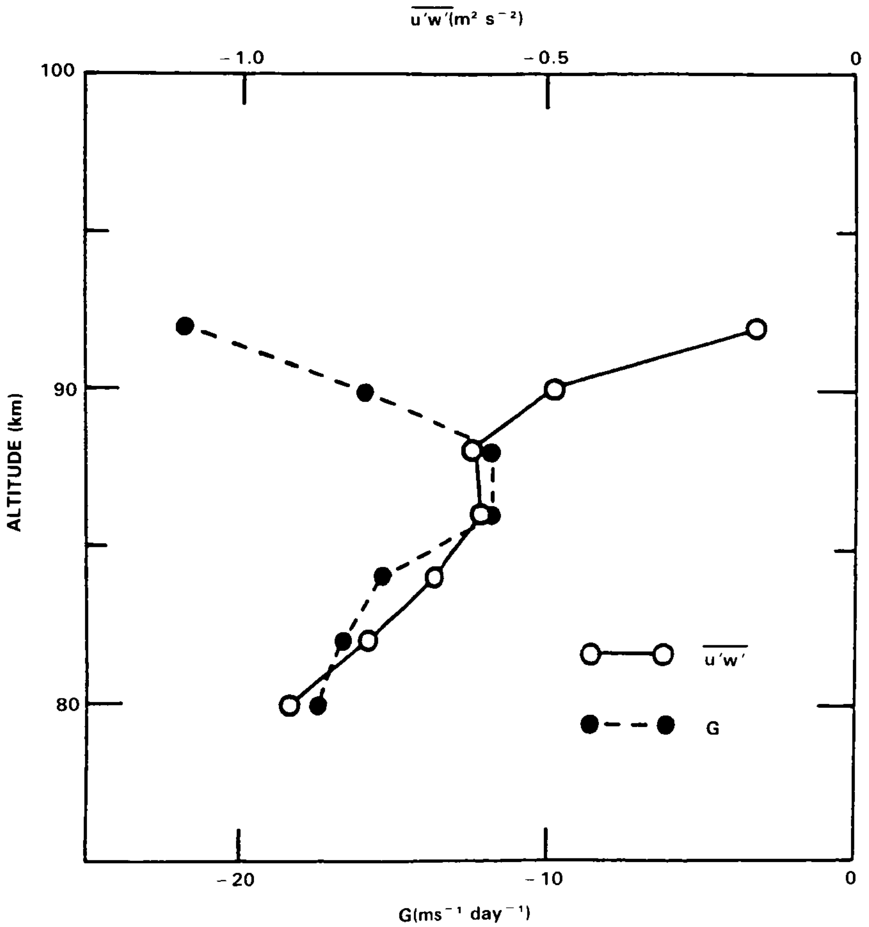


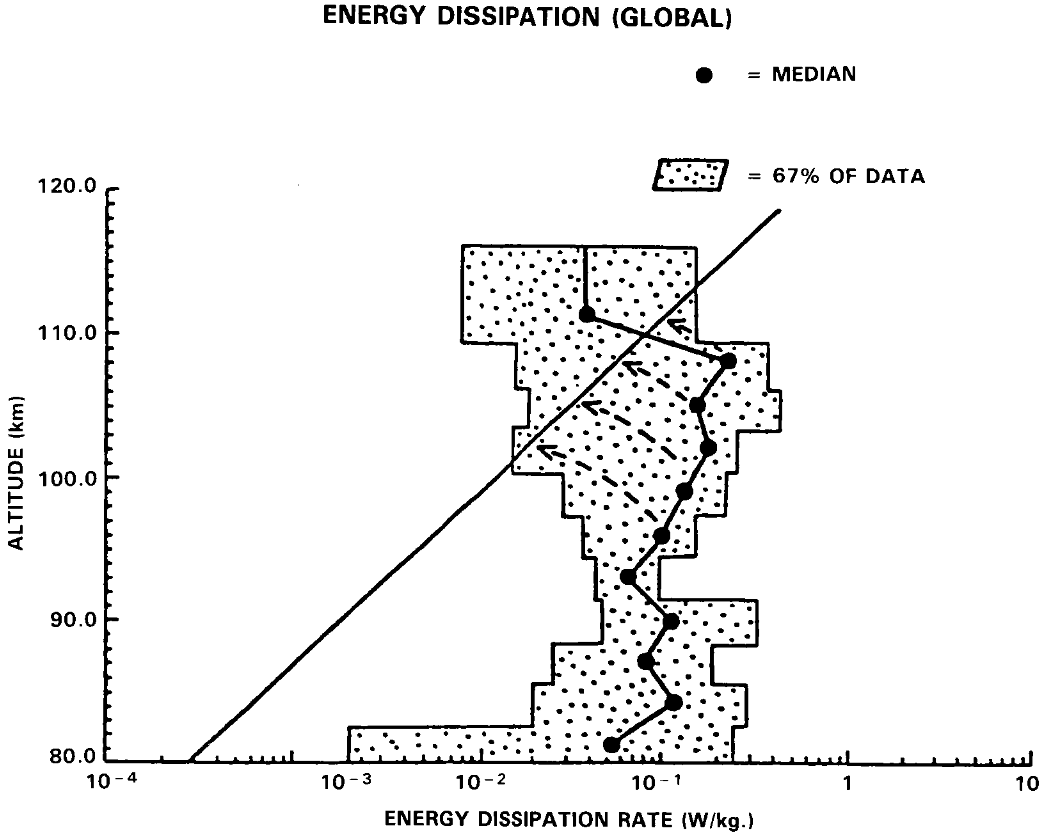
Figure 6-17. Height profiles of $\overline{u'w'}$ and zonal drag, G, observed at Adelaide in May 1981 [After Vincent and Reid, 1983].

However, there can be significant variability of $\overline{u'w'}$ over time scales ranging from hours to days and (presumably) in space so that extended measurements at more locations are required to understand fully the role gravity waves play in the momentum budget of the mesosphere. One interesting and potentially important feature of the measurements is that high frequency waves (periods less than 1 hour) appear to carry a substantial fraction of the energy and momentum fluxes. This is because their large vertical group velocities more than compensate for their smaller amplitudes in the lower atmosphere relative to the long period waves. Short-term variations in wave energy and the associated fluctuations in mesospheric heating have been investigated by Clark and Morone [1981].

Variability or intermittency in wave activity in the mesosphere is also evident in observations of small-scale turbulent motions. High resolution MST radar measurements show that the scattering regions appear as blobs, sheets and layers, with vertical thicknesses ranging from 150 m or less up to a few km [Rottger, 1980]. Horizontal scales are estimated to extend from 1 km up to hundreds of km. Some of the layers are observed to descend with time, which suggests the turbulent structures are associated with dissipating gravity waves. By measuring the Doppler spectral widths of the echoes it is possible, with care, to estimate the turbulent intensity directly and so infer eddy dissipation rates and diffusivities [Sato and Woodman, 1982; Hocking, 1983]. Applications of this technique to the mesosphere give values comparable to those inferred from gravity wave decay rates (viz. $\epsilon \cong 0.01$ to 0.2 Wkg^{-1} and D $\cong$ 100 to 500 m^{2}s^{-2}). Figure 6-18 shows vertical profiles of 'global averages' of ϵ and D in the upper mesosphere constructed from all available data [Hocking, 1985]. To date, there are insufficient data to deduce a seasonal and geographic climatology of turbulence in the mesosphere. It is possible that satellite measurements of ozone densities at the 80 km level could be used to monitor indirectly the amount of turbulence and wave activity. Thomas *et al.* [1984b] found equinoctial maxima in ozone concentrations at 80 km, which they attributed to reductions in the turbulent transport from lower levels of water vapor, probably the most active agent for ozone destruction in the mesosphere. The reduction in turbulence at the equinoxes is presumed to be due to the attenuation of gravity waves propagating through the weak zonal winds in the middle atmosphere and a corresponding reduction in wave breaking in the mesosphere. Some support for these ideas has come from studies of Meek *et al.* [1985b] who found an attenuation of wave activity in the mesosphere at the time of the equinoctial reversal of the zonal circulation at Saskatoon.

High-resolution radar and balloon studies show that turbulent processes in the lower stratosphere are also intermittent in space and time with the turbulence confined to thin layers. Layer thicknesses range from a few tens to a few hundred metres and horizontal scales are inferred to range from a few kilometres up to a few hundred kilometres in extent. The association of the turbulent layers with regions of high wind shear caused by the wave-like features of the wind profile is especially evident. The persistence of these features and their slow downward movement with time strongly indicates that much of the stratospheric turbulence is generated by dynamically unstable inertio-gravity waves [Barat, 1982; Sato and Woodman, 1982]. Estimates of mean dissipation rates vary between 5.10^{-6} and 10^{-4} W kg^{-1} with corresponding mean diffusivities ranging between 0.01-0.2 m^2 s^{-1}. Aircraft measurements show highest turbulence activity over the mountains [Lilly *et al.*, 1974]. Radar estimates of diffusivity in the stratosphere (0.2 m^2 s^{-1}) are about an order of magnitude larger than those inferred from aircraft observations [Lilly *et al.*, 1974]. The reasons for this difference are not yet resolved. As will become apparent below (Section 6.5), however, even these larger values are probably unimportant for large-scale transport.

Long-term radar measurements of horizontal wave motions show that, similar to the situation in the mesosphere, the energy is distributed with a $f^{-5/3}$ power law [Balsley and Carter, 1982]. In contrast to what is often assumed in theory, these and other studies are starting to indicate that the wave energy density does not remain constant with increasing height in the stratosphere but rather shows an exponential



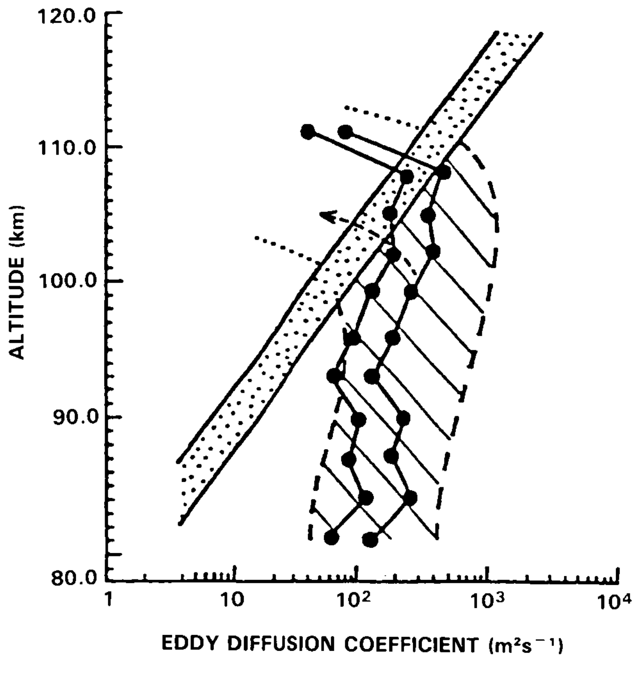


Figure 6-18. Profiles of global eddy dissipation rates and diffusion coefficients as function of height above 80 km [After Hocking, 1985].

decay, again similar to the mesospheric situation. While there are a number of reasons (including reflection and the effect of the background wind) why the wave amplitudes may not grow with height, saturation processes are possibly important, which suggests that gravity waves may contribute to the momentum budget of the stratosphere. Massman [1981], using southern hemisphere constant pressure balloon measurements, found large upward fluxes of wave energy and momentum in the upper troposphere but significantly smaller values in the stratosphere, which indicates flux convergence across the tropopause. The convergence of the vertical momentum flux, which is approximately equal to the Eliassen-Palm flux for small-scale gravity waves (see Section 6.2), suggests destruction of wave activity in the tropopause region. The maximum fluxes were near the Andes. Schoeberl [1985] has shown theoretically that topographically forced waves may superpose in the stratosphere to form unstable regions.

The amplitude of the solar diurnal and semidiurnal tides in the stratosphere are small (< 1 - 5 ms^{-1}) and they do not appear to play a significant role in the dynamics of this region. In the mesosphere, however, tidal motions can be large and, at latitudes of less than 30°, the migrating diurnal tide is probably an important source of heat, momentum and turbulence. Theory predicts that the fundamental (1,1) diurnal mode will become convectively unstable in the equatorial mesosphere [Lindzen, 1968] and because of its short vertical wavelength (25-30 km) this mode is strongly damped by turbulent processes in the mesosphere at higher latitudes. Observations show that the wave amplitudes maximize near 90 km altitude. Calculations suggest that the dissipating tide generates easterly winds of up to 60 ms^{-1} in the equatorial upper mesosphere and westerly winds of 35 ms^{-1} at 30° latitude [Miyahara, 1984]. Estimates put the upward energy flux into the mesosphere at about 1.5 Wm^{-2} and the most recent calculations indicate that the (1,1) mode generates large heating rates near 80 km [Groves and Forbes, 1984], which could be significant on a global scale. Finally, it is noted that the tides can affect the concentration of minor constituents such as ozone in the mesosphere not only by direct transport but also by their reversible temperature fluctuations, which can produce significant variations in the photochemistry [Forbes, 1984].

6.1.7 The Seasonal Cycle

The march of solar heating drives a seasonal cycle in the middle atmosphere; in the stratosphere, cold polar temperatures and a strong westerly vortex in winter are replaced in summer by warm polar temperatures and a weak easterly vortex. Because of dynamical disturbances, however, the seasonal cycle does not proceed uniformly.

Figure 6-19 shows the annual march of zonal mean radiance near the north and south poles in the upper stratosphere. Maximum temperatures (proportional to the radiances) are achieved in mid-summer and minimum in mid-winter. Winter temperatures are more variable in the Northern Hemisphere because of stratospheric warmings. The seasonal increase of temperatures in the Southern Hemisphere is highly oscillatory from midwinter to summer, whereas in the Northern Hemisphere temperatures increase gradually in spring after abrupt changes in winter. After mid-winter, upper stratospheric temperatures in the polar region are higher in the southern than in the Northern Hemisphere, except during major warmings in the Northern Hemisphere [Barnett, 1974]. The meridional temperature gradient reverses at high latitudes in the upper stratosphere of the Southern Hemisphere as a warm pool forms over the pole after midwinter [Hartmann, 1976a; Hirota *et al.*, 1983]. These changes are related to the observed poleward and downward movement of the zonal mean jet in the Southern Hemisphere during late winter [Harwood, 1975; Hartmann, 1976a; Leovy and Webster, 1976].

Figure 6-20 shows the evolution of the zonal mean wind in the upper stratosphere. In the Northern Hemisphere, the zonal mean westerlies are strongly disrupted by disturbances (e.g. during January-February

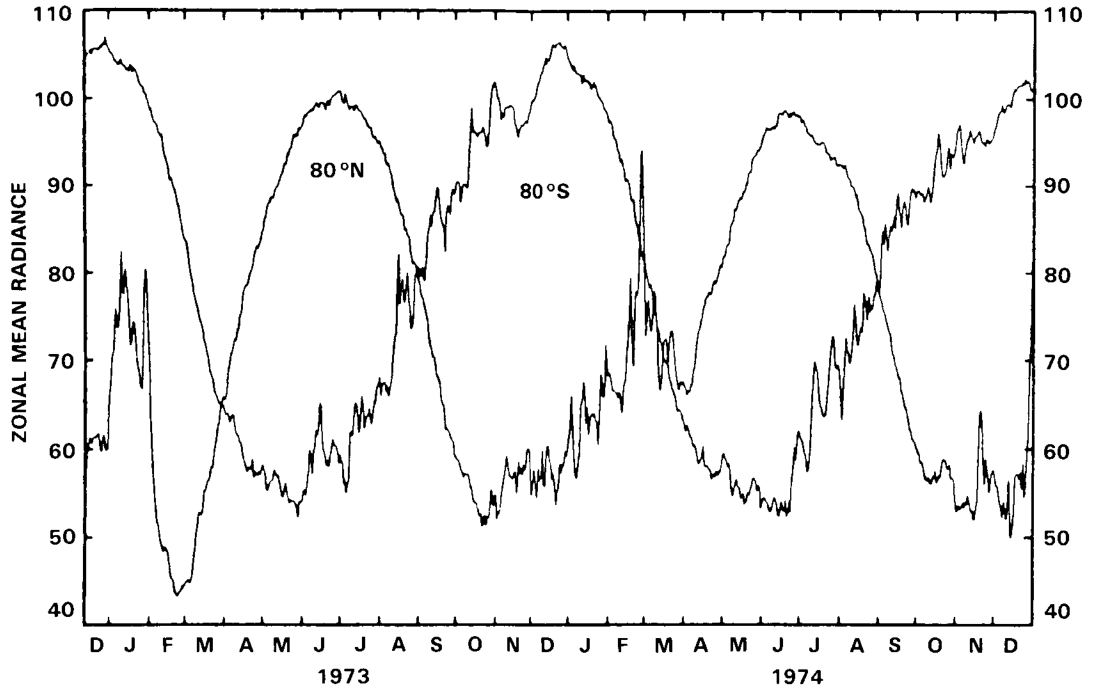


Figure 6-19. Annual march of zonal mean radiance observed by the Nimbus 5 SCR channel B12 for 80 °N and 80 °S. Units are mW m^{-2} ster^{-1} (cm^{-1})$^{-1}$. Taken from Hirota *et al.* [1983a].

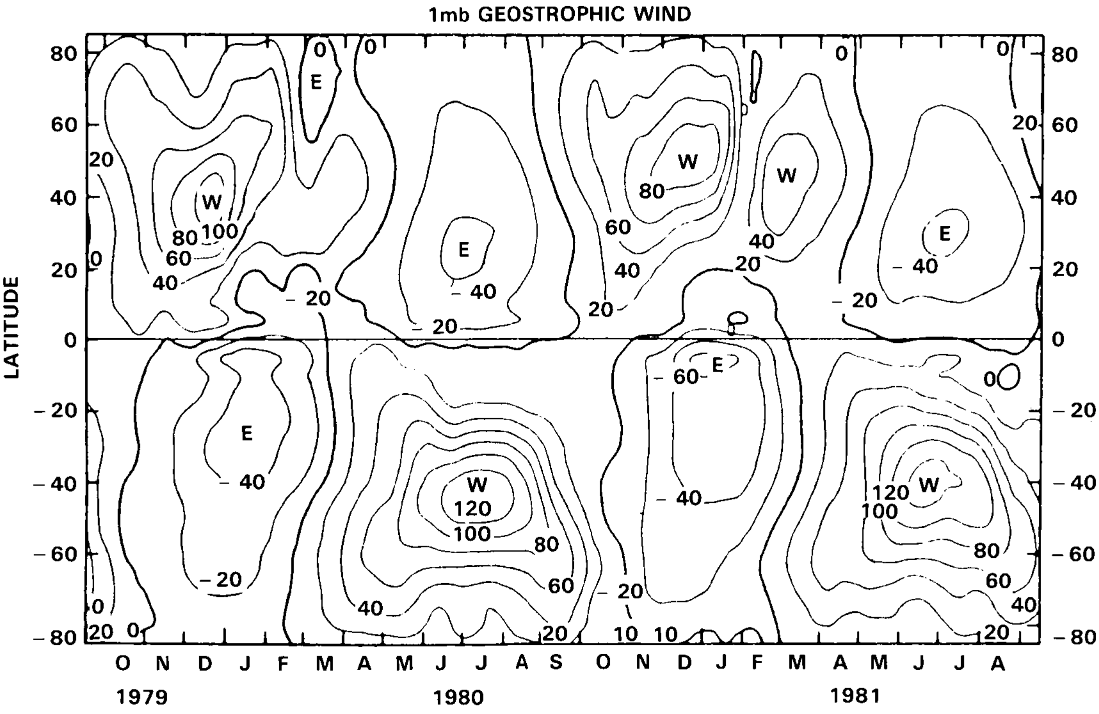


Figure 6-20. Latitude-time section on the zonal mean geostrophic wind at the 1 mb level estimated from the 20-day average height field observed by the Tiros N SSU. Units are ms^{-1}. Positive values denote westerly winds. Taken from Hirota *et al.* [1983a].

1981). Should these disturbances occur in late winter, the westerly circulation may never recover before summer easterlies are introduced by the evolving radiation field. These 'final warmings' occur earlier in the seasonal cycle in the Northern Hemisphere than in the Southern Hemisphere where the winds evolve more regularly. The zonal winds in the Southern Hemisphere are strongest at high latitudes in early winter, at mid latitudes in mid-winter and then, as the zonal mean jet moves polewards and downwards, at high latitudes again in late winter.

In equatorial regions the annual cycle is weak and the seasonal cycle in the upper stratosphere becomes dominated by a semiannual oscillation (SAO). The SAO was originally discovered by Reed [1965, 1966] and further observational evidence has been presented by Belmont and Dartt [1973], Hirota [1978, 1980] and Hamilton [1982a]. While a substantial semiannual component exists at middle and high latitudes, this appears to be largely a reflection of the non-sinusoidal character of the seasonal cycle there; in the tropics, however, a distinct maximum in the zonal wind SAO occurs, reaching amplitudes of about 30 ms^{-1} near the stratopause, with maximum westerlies just after equinox (Figure 6-21). Easterly zonal winds penetrate from the summer hemisphere into the winter hemisphere twice a year; the easterlies spread further across the equator during the southern hemisphere summer. As is evident from Figure 6-21, there is a second maximum of the SAO near the mesopause, out of phase with that in the upper stratosphere [Hirota, 1980; Hamilton, 1982a].

Figure 6-22 shows the evolution in the upper stratosphere of the amplitudes of waves 1 and 2 computed from the monthly mean fields of geopotential height. The distinct seasonal cycle of wave amplitudes is linked to that of the zonal mean winds. Large-scale disturbances in the troposphere can only penetrate into the stratosphere during winter when winds are westerly (although the seasonal cycle in planetary wave activity may also reflect changes in tropospheric forcing). Except for wave 1 in the northern hemisphere winter, the position of the maximum wave amplitudes is close to the latitude of strongest zonal mean westerlies (marked by broken lines on the figures). The maximum amplitude of wave 1 in the Northern Hemisphere winter is at higher latitudes. While amplitudes tend to be at their maximum in the Northern Hemisphere in mid-winter, wave 1 amplitudes (and to a lesser extent wave 2 amplitudes) show a relative minimum in mid-winter in the Southern Hemisphere. The poleward and downward movement of the jet in the Southern Hemisphere in spring is associated with an enhancement of wave 1 [Hartmann *et al.*, 1984; Shiotani and Hirota, 1985]. Transient waves at extra-tropical latitudes show similar seasonal variations in amplitude to those of the time-averaged waves, with largest amplitudes in winter and smaller amplitudes in summer.

The most spectacular departures from a regular seasonal cycle in the middle atmosphere occur near the winter pole during sudden warmings. Wave amplitudes can double and temperatures can rise locally by 80 K or more in a week or so. This dramatic phenomenon, which occurs sporadically in the winter stratosphere, is associated, in extreme cases, with a reversal of the zonal-mean westerly flow. The picture of a zonally symmetric flow which undergoes a deceleration and possibly a reversal in direction may, in some respects, be misleading. Such phenomena may be manifested in the total flow field (i.e., zonal mean plus waves) by a displacement of the vortex off the pole in the case of a "wave 1" warming [e.g., Palmer and Hsu, 1983], or by a split in the vortex in the case of a "wave 2" warming [e.g., Palmer, 1981a]. A recent example of this latter type of warming is that of Figure 6-23, which shows two cyclonic vortices separated by a warm anticyclone over the pole. The atmospheric disturbance can often be detected at the mesopause and above [e.g., Muller *et al.*, 1985]. Warmings are believed to be triggered by changes in the large-scale circulation in the troposphere, although the precise nature of these changes has not been adequately documented.

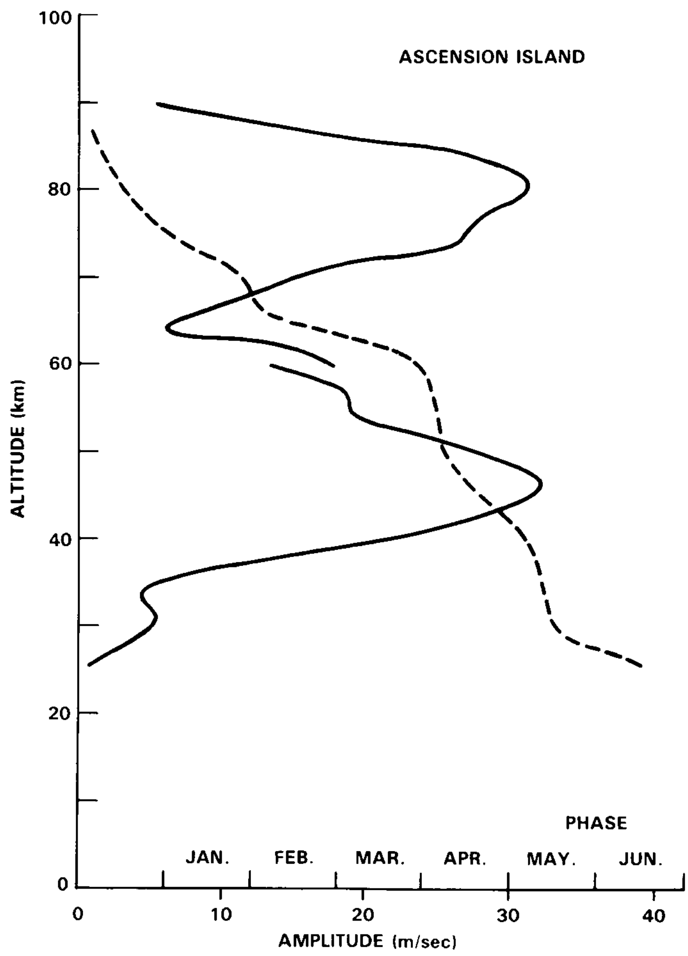


Figure 6-21. Vertical distribution of the amplitude (solid) and phase (time of the maximum westerly component [dashed]) of the semiannual cycle of zonal wind at Ascension Island. [After Hirota, 1978].

During warmings there is a large exchange of material between high and low latitudes. Not all this exchange will be permanent, but strong warmings may effect considerable poleward transport of trace chemical species. The poleward advection of air from low latitudes during warmings is indicated by maps of Ertel's potential vorticity on isentropic surfaces (see Section 6.2). (Over short periods, contours of potential vorticity on such surfaces are approximately material lines.) An example for an early winter warming in the Northern Hemisphere is shown in Figure 6-24. The area of low potential vorticity was drawn around the westerly vortex into the polar cap from low latitudes. McIntyre and Palmer [1983, 1984] have applied the term "wave breaking" to extreme and irreversible buckling of potential vorticity contours such as that depicted in Figure 6-24.

Major warmings involving the replacement of polar westerlies by easterlies do not occur in mid-winter in the Southern Hemisphere. This is probably because the circulation in the troposphere does not contain such persistent large-scale disturbances as it does in the northern hemisphere. In late winter, however, the final stratospheric warming is similar to the wave 1 type of minor warming that occurs in the northern hemisphere, when the zonally averaged jet moves polewards and downwards.

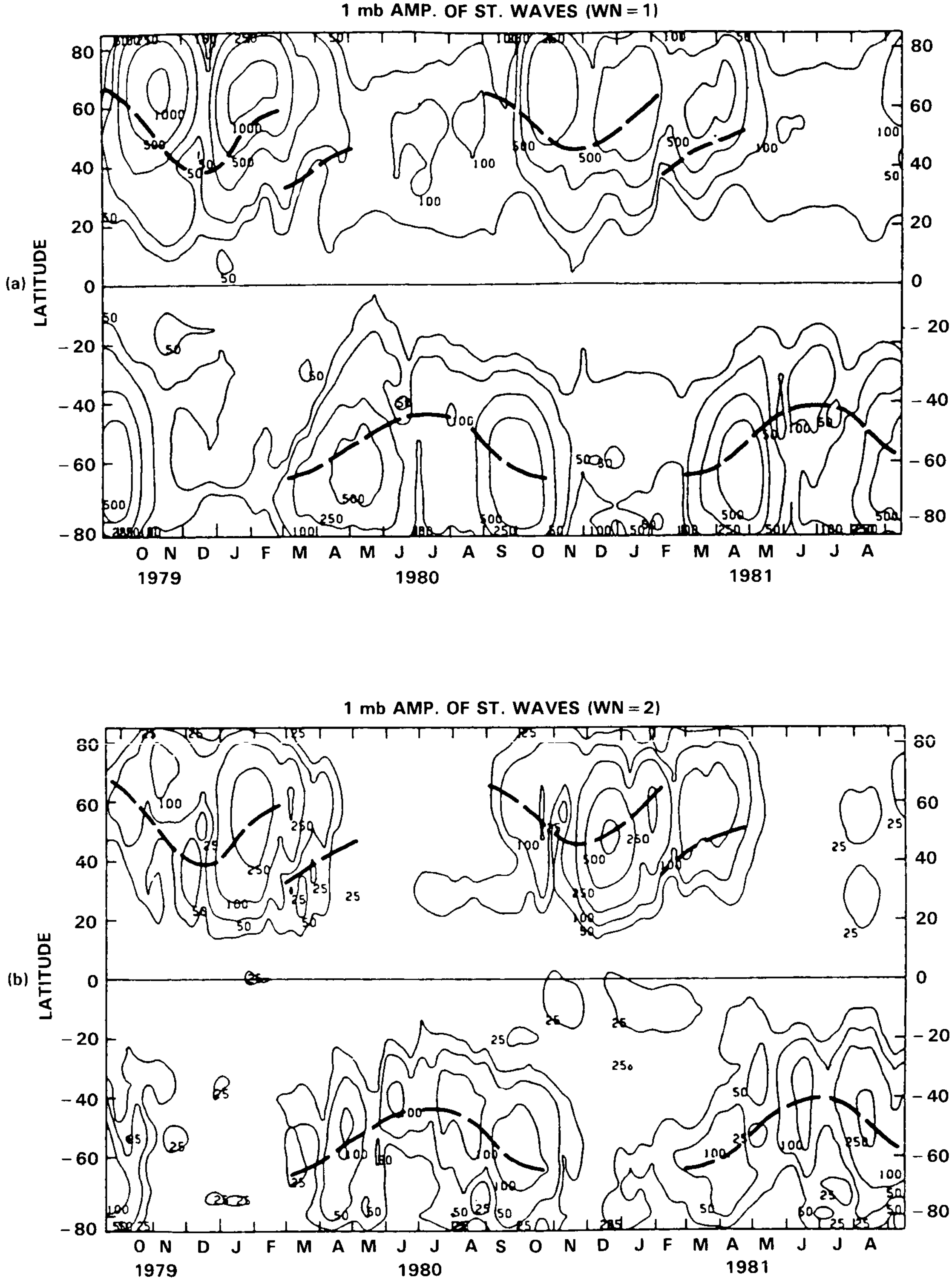


Figure 6-22. Latitude-time section of amplitude of quasi-stationary (a) wavenumber 1, and (b) wavenumber 2. Units are metres. Heavy broken lines denote axes of maximum westerlies. [After Hirota *et al.*, 1983].

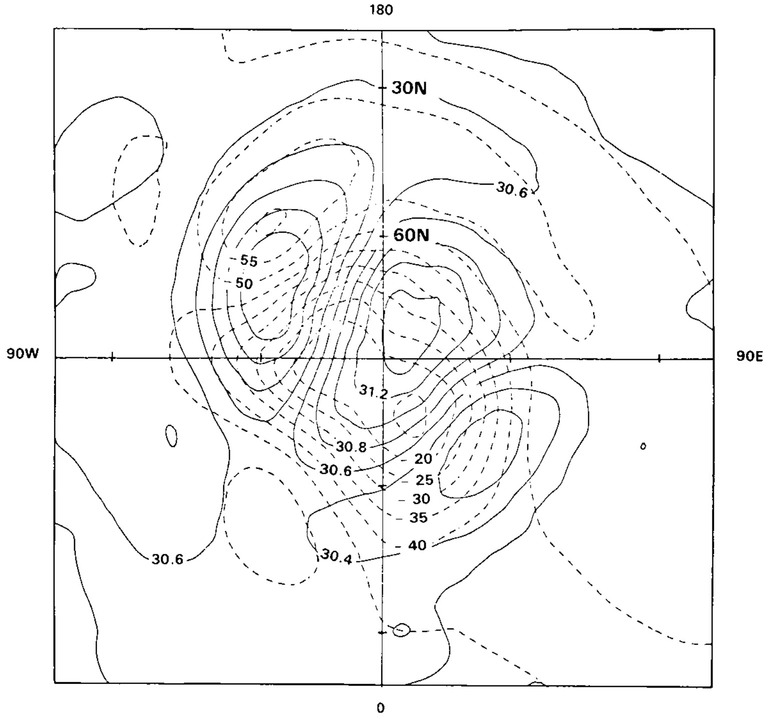


Figure 6-23. Polar sterographic maps at 10 mb of geopotential height (km solid lines) and temperature (K dashed lines) on 2 January 1985 at the height of a major stratospheric warming. Data obtained from a stratospheric sounding on the satellite NOAA-7. [Analysis made by the Middle Atmosphere Group, Meteorological Office, U.K.].

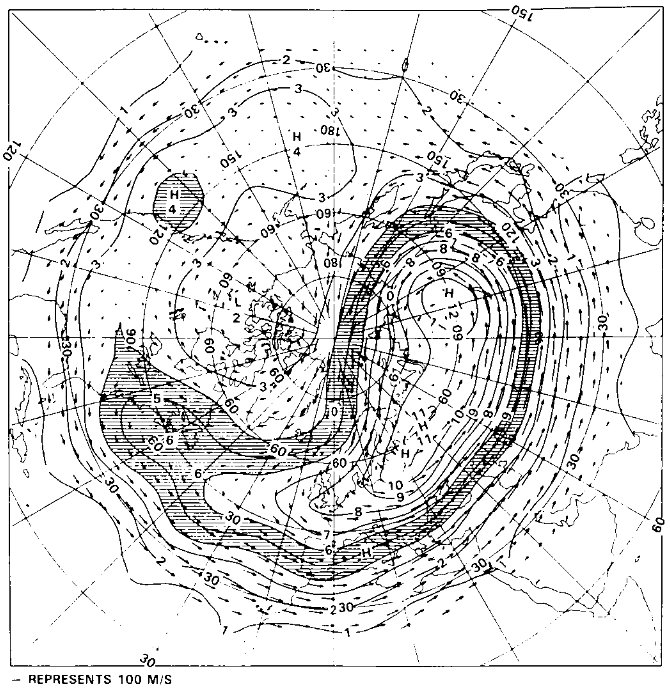


Figure 6-24. Ertel's potential vorticity and geostrophic winds evaluated on the 850 K isentropic surface near 10 mb for 4 December 1981. Units are km^2 kg^{-1} s^{-1} × 10^{-4}. [After Clough *et al.,* 1985].

6.1.8 Inter-Annual Variability

The circulation of the middle atmosphere exhibits strong interannual variability in winter, which can be observed up to the mesopause. Labitzke and Naujokat [1983] have used the long record of radiosonde data (nearly 30 years) to study the inter-annual variability of the lower stratosphere. Figure 6-25 is a frequency distribution of the temperature at 30 mb for both poles. During summer when the prevailing winds are from the east, tropospheric disturbances do not penetrate far into the stratosphere. Therefore, conditions are quiet and the inter-annual variability is small, particularly at mid latitudes. In the northern winter and spring, the frequency distribution is broad. Labitzke [1982] has noted that the circulation is dominated by different zonal harmonic waves in different years: ''disturbed'' winters have a pronounced wave 1 pattern, often leading to major mid winter warmings and thus a warm polar region; and undisturbed winters with a pronounced wave 2 and a very cold polar region (vortex more symmetrically placed with respect to the pole), and only minor warmings in the stratosphere.

In the middle stratosphere, the variability is much smaller during the southern midwinters because large-scale waves are weaker than in the northern hemisphere. Wave amplitudes increase in spring (Figure 6-22) and are associated with increased variability of polar temperatures in October and November.

Geller *et al.* [1984] have noted significant inter-annual variability in the monthly mean zonal winds in the northern hemisphere. This is shown for four winters in Figure 6-26. Despite the variability, however, the same general pattern emerges during each of the winters as winter proceeds from December to February.

Even a rather superficial inspection of the circulation statistics of the equatorial middle atmosphere reveals characteristics, particularly long-period variations, which are profoundly different from those of the mid-latitude circulation. The variability of the latter is dominated by the seasonal cycle; in tropical latitudes (within about 15° of the equator) the annual cycle disappears, to be replaced by a quasi-biennial cycle in the lower stratosphere and, as has already been noted, a semiannual cycle in the upper stratosphere and mesosphere.

Between about 20 and 35 km altitude the monthly-mean zonal winds at tropical stations are seen to exhibit strong quasi-cyclic behavior (see Figure 6-27); the winds alternate between easterlies and westerlies of 20-30 ms^{-1} with a period which ranges between 23 and 34 months (with a mean value of about 28 months). This phenomenon, which was first reported by Reed [1960] and Veryard and Ebdon [1961], has become known as the ''quasi-biennial oscillation'' [QBO]. Further description of its observed properties are to be found in Reed [1965], Wallace [1973], Coy [1979, 1980], Hamilton [1984], Dunkerton and Delisi [1985]. The wind oscillation is observed to be remarkably symmetric in longitude and dominated by the zonal component, so that it takes the form of alternating easterly and westerly jets over the equator; these jets descend slowly, successively being replaced from above by one of the opposite sign (Figure 6-28). The jets are confined to within about 15° of the equator [Wallace, 1973; Dunkerton and Delisi, 1984] although there is some evidence for a signal at higher latitudes, and perhaps in the planetary wave component of the winter hemisphere as well as in the zonal flow [Holton and Tan, 1980, 1982; Labitzke, 1982]. During the westerly phase of the QBO, the polar region tends to be colder and wave 2 amplitudes larger with less likelihood of a major warming; during the easterly phase, wave 1 amplitudes tend to be emphasized, leading to warmer polar temperatures and sometimes major warmings. This inter-annual variability in the middle latitudes may also be linked to the tropospheric Southern Oscillation [van Loon *et al.*, 1981], which in turn is thought to be connected with feedbacks between the atmosphere and local anomalies in sea surface temperature [e.g. Gill, 1982].

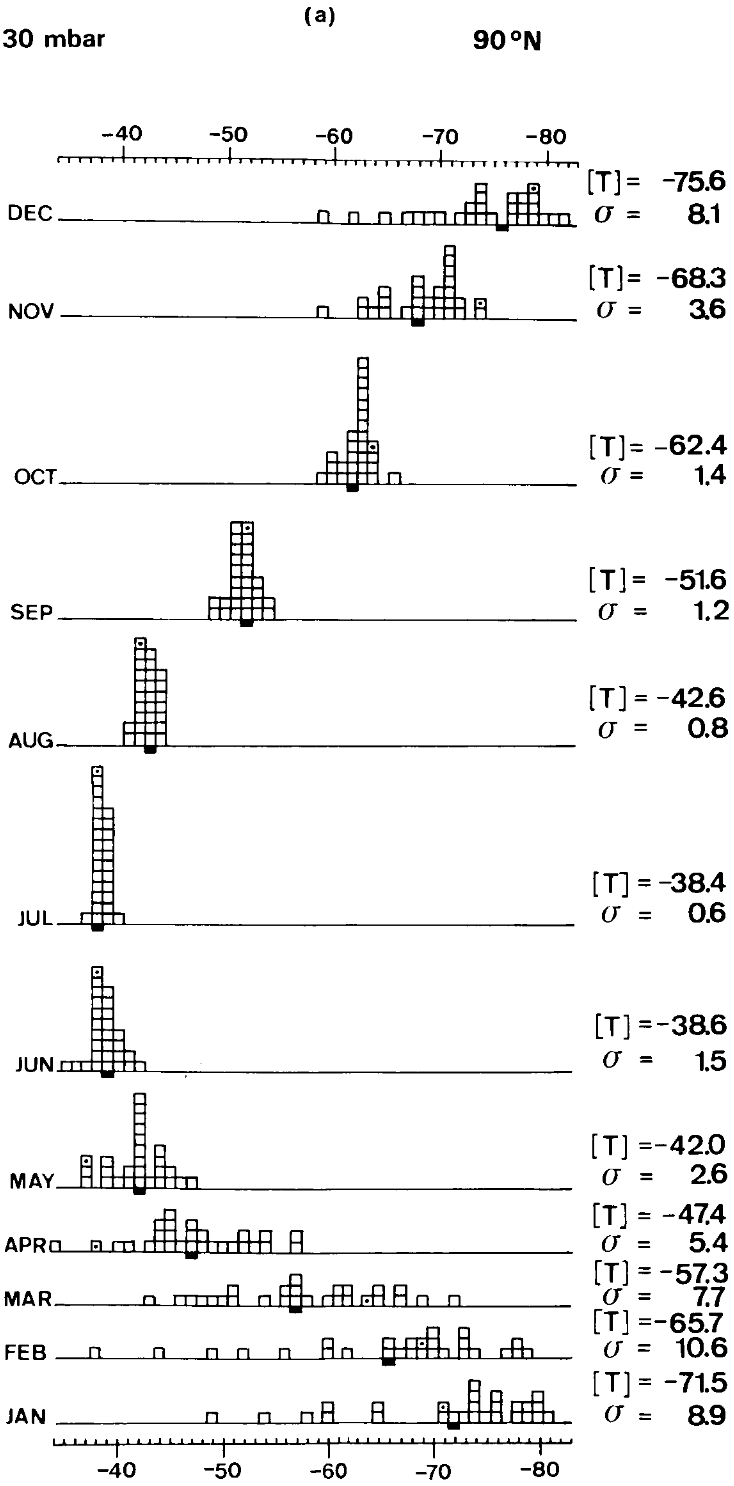


Figure 6-25. Frequency distribution of the monthly mean 30 mb temperatures (°C). The interval is 1 °C. The long-term average [T] is given on the right hand side of the picture together with the standard deviation σ, and [T] is also marked as a black box in the frequency distribution. (a) is for the North Pole using radiosonde data for the period July 1955-Dec. 1982,

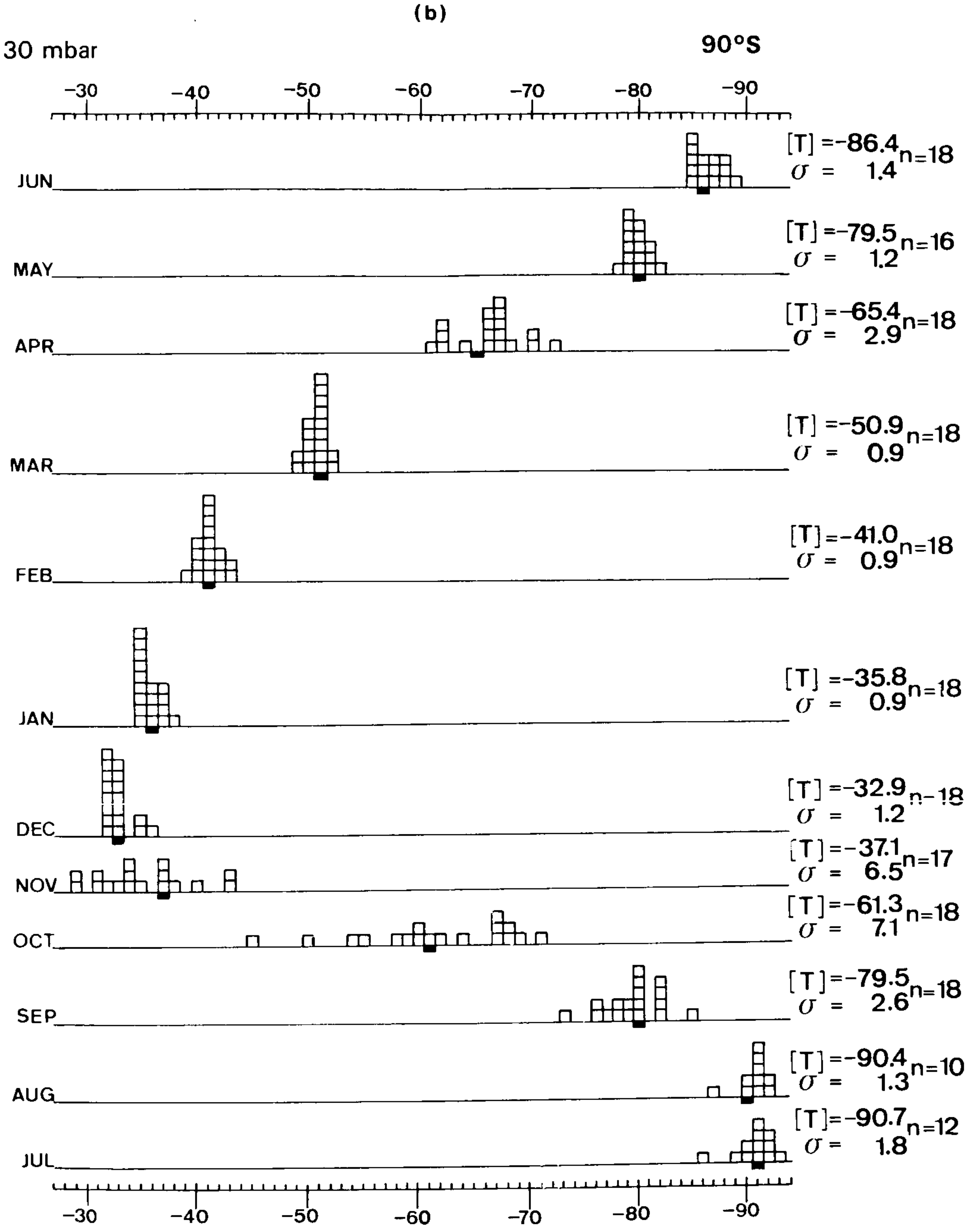


Figure 6-25. (b) is for the South Pole for the period 1961-1978 using, for each month, data for the number of years specified by n. [After Naujokat, 1981, and Labitzke and Naujokat, 1983].

Consistent with thermal wind balance, given the observed zonal wind structure, a temperature QBO (amplitude 1-2 K) is also observed [Reed 1962, 1964; Newell *et al.*, 1974; Angell and Korshover, 1978b]. Like the wind oscillation, the thermal component appears to be primarily confined to the tropics. However, a QBO in total ozone, first identified by Funk and Garnham [1962] and recently clarified by Hasebe [1983, 1984] in an analysis of Nimbus 4 BUV and ground-based observations, is more prominent in middle latitudes, especially in the Southern Hemisphere (Figure 6-29).

ORIGINAL PAGE IS
OF POOR QUALITY

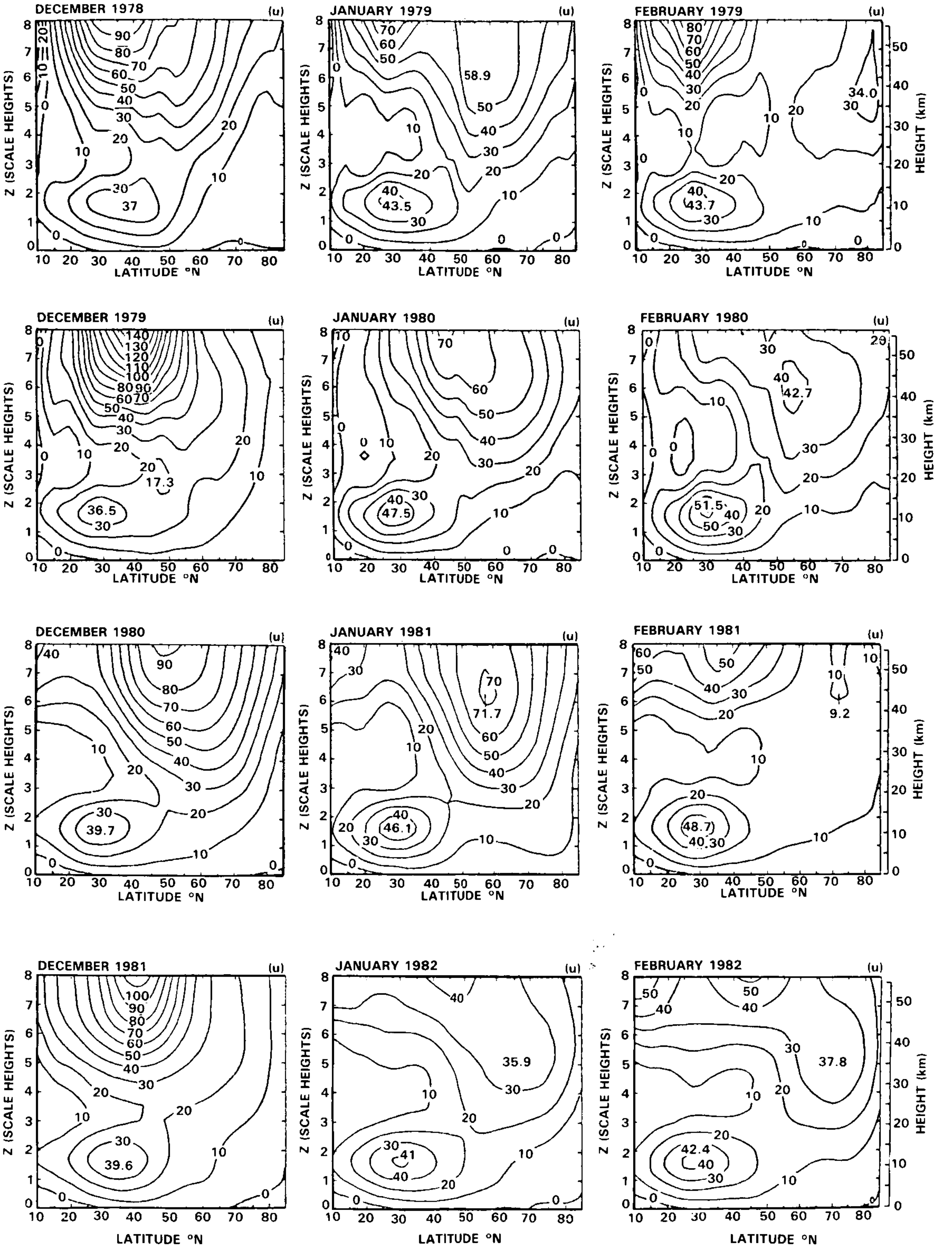


Figure 6-26. Northern hemisphere, monthly mean, zonally averaged geostrophic winds (ms^{-1}) for the months of December, January and February for the winters 1978-79 through 1981-82. Regions of easterlies are shaded. [After Geller *et al.,* 1984).

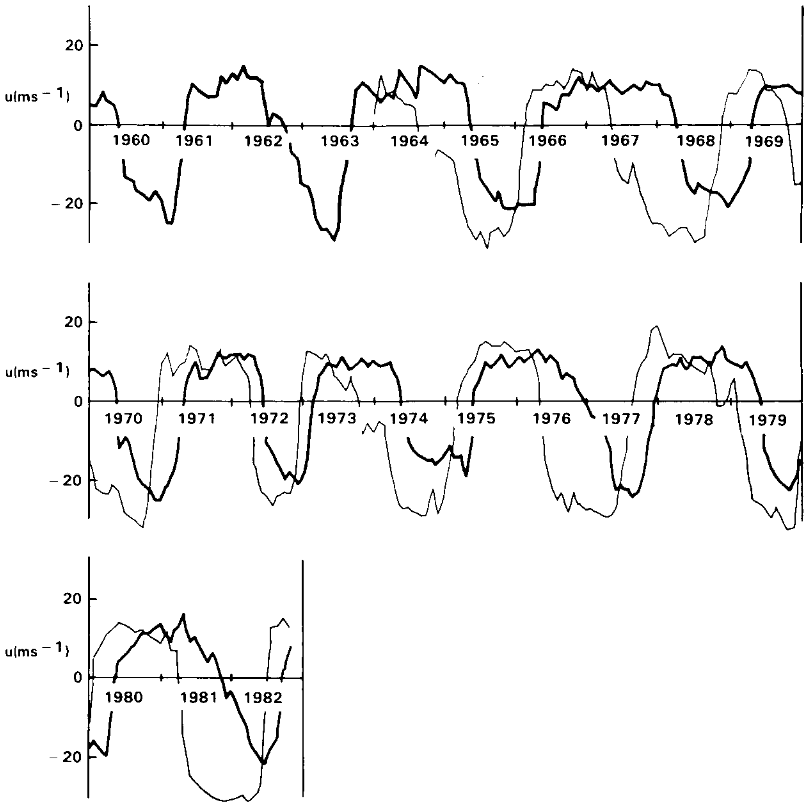


Figure 6-27. Monthly mean zonal winds at Singapore (1°20'N) at 50 mb (thick line) and 30 mb (thin line). Westerly wind is positive. [After Plumb, 1984].

An implication of the inter-annual variability of the middle atmosphere is that observations covering a single winter or even a small group of winters may be unrepresentative of the circulation of the middle atmosphere in a longer term mean. This should be borne in mind when verifying numerical simulations of the circulation or of the transport of trace species; this point will be addressed further in Section 6.6.

6.2 THEORY OF THE CIRCULATION OF THE MIDDLE ATMOSPHERE

6.2.1 Introduction

This section discusses the dynamical processes which maintain the observed large-scale circulation of the middle atmosphere in a state that is often far from radiative-photochemical equilibrium. It mostly concentrates on the maintenance of the climatological *zonal-mean* state, although some dynamical aspects of the climatological, non-zonally-averaged state are mentioned. The dynamics of stratospheric sudden warmings and long-period oscillations in the tropical middle ǎtmosphere are also discussed briefly. (The adjective "climatological" here refers to those features which vary slowly from month to month during the annual cycle and recur regularly from year to year. They can be isolated, for example, by taking time averages over the individual calendar months in a data record extending over many years.)

To investigate the role played by dynamical processes in producing the observed middle atmospheric circulation it is first useful to consider what form the circulation would take if dynamical processes were

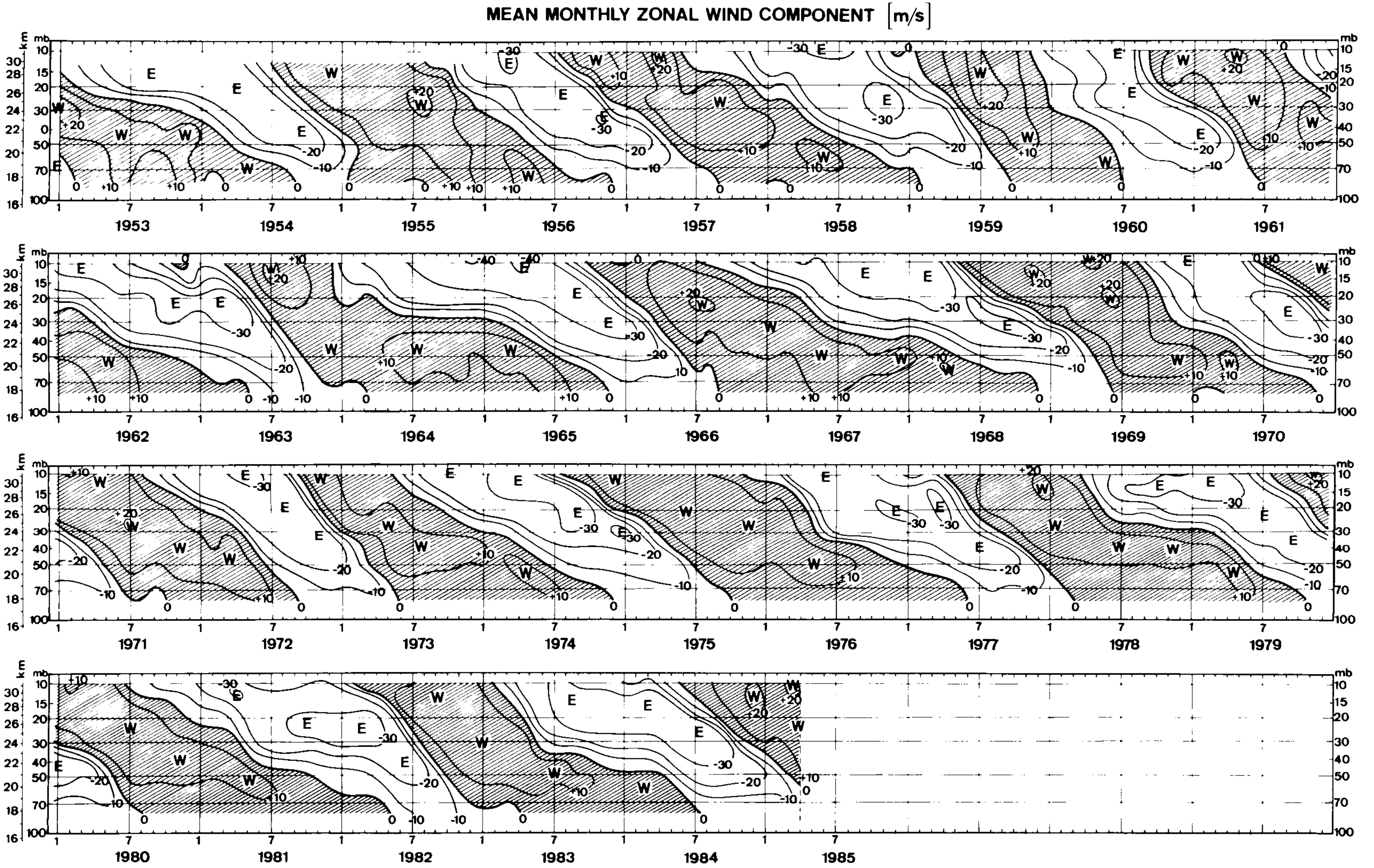


Figure 6-28. Time-height cross section of mean monthly zonal winds (m/s) at equatorial stations, calculated from all available daily values:

Jan 1953 - Aug 1967 Canton Island, 3°S/172°W
Sep 1967 - Dec 1975 Gan/Maledive Islands, 1°S/73°E
Jan 1976 - Apr 1985 Singapore, 1°N/104°E (After Naujokat, 1985).

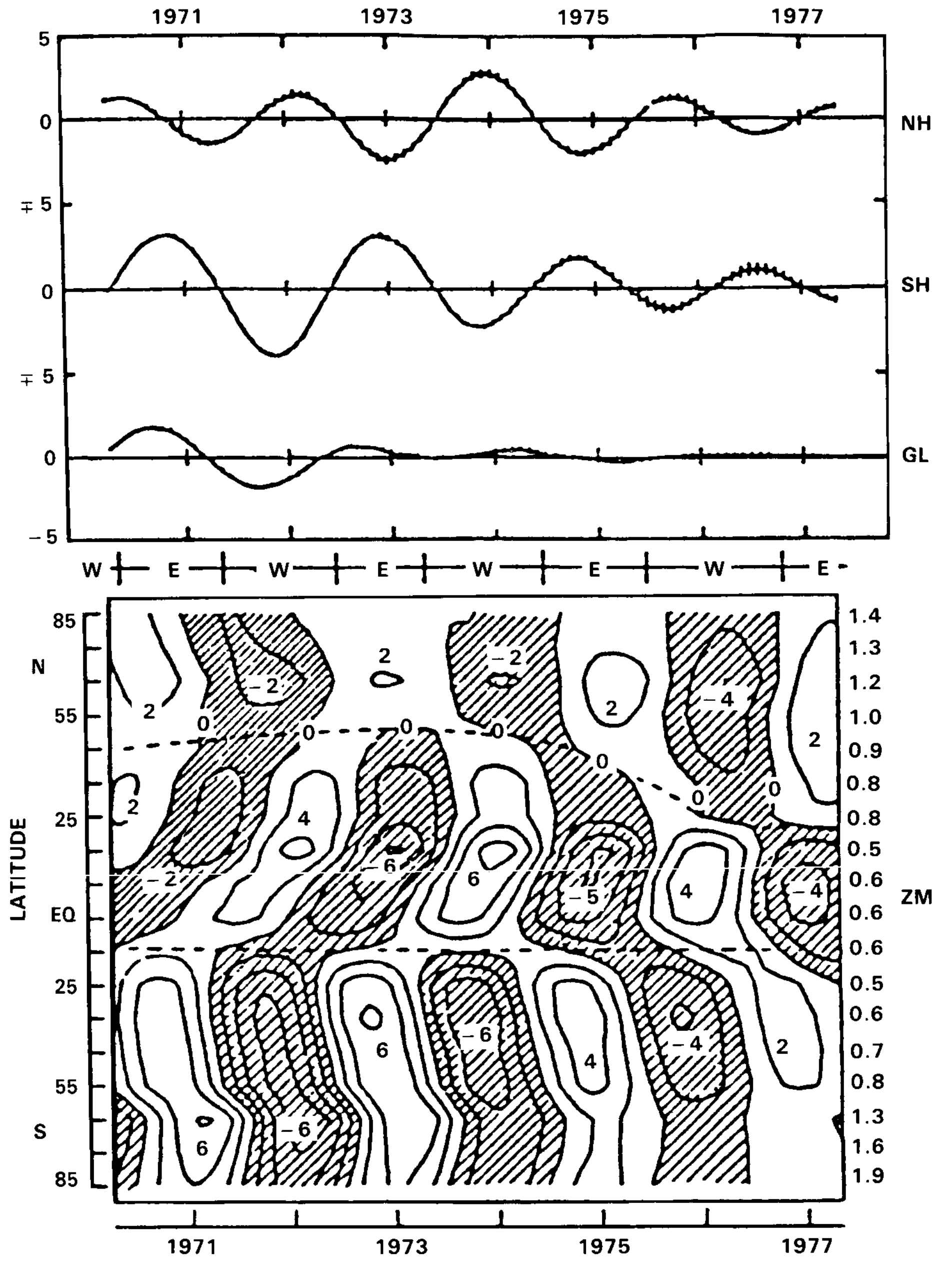


Figure 6-29. Quasi-biennial oscillation of total ozone (matm-cm) in the mean values of (NH) Northern Hemisphere, (SH) Southern Hemisphere, and (GL) globe, and (ZM) zonal mean values. The isopleths in ZM are drawn with the interval of 2 matm-cm, and the shaded areas correspond to negative deviations. The tick mark is January of the given year. [After Hasebe, 1983].

absent (except for some representation of convection, and perhaps baroclinic wave activity, in the troposphere). The temperature field associated with such a circulation can be calculated from a radiative-photochemical model of the stratosphere and mesosphere, together with a radiative-convective model of the troposphere. An example for near-solstice conditions is shown in Figure 6-30, from Fels and Schwarzkopf [1985] [reported in Mahlman and Umscheid, 1984], whose model is time-marched through an annual cycle. The figure shows strong latitude and height variations of the resulting zonally-symmetric tem-

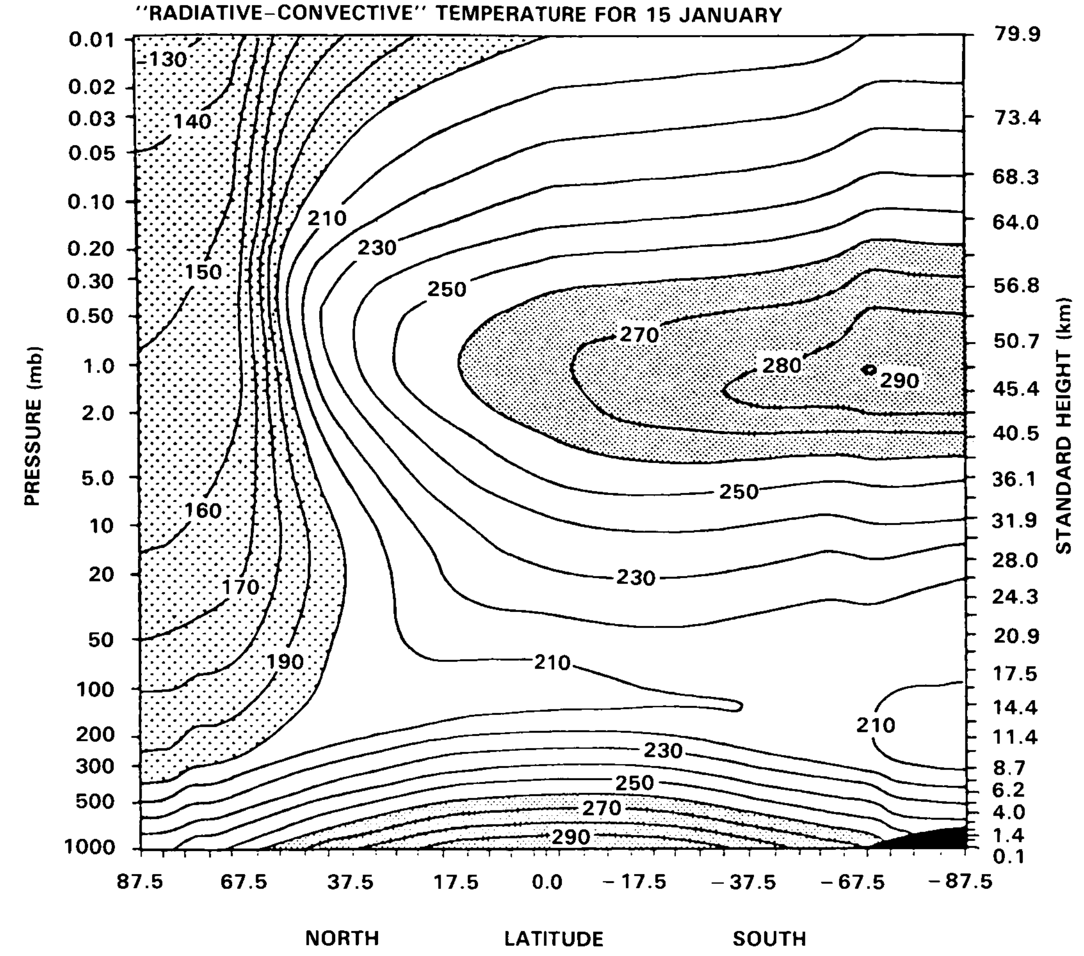


Figure 6-30. Time-dependent "radiatively-determined" temperature T_r for 15 January 1982 from the calculation of Fels and Schwarzkopf (1985). The surface temperatures are prescribed at their seasonally-varying observed values. Cloudiness, and ozone below 35 km, are prescribed at annual-mean values, as in Fels *et al.* (1980); ozone above 35 km is allowed to "float", in response to temperature variations, towards a crude photochemical equilibrium. Details of the water vapor prescription are relatively standard and are described in Fels and Schwarzkopf (1985). [From Mahlman and Umscheid, 1984].

perature T_r, with a maximum of about 280 K at the summer stratopause and temperatures below 180 K throughout the middle atmosphere at the winter pole, decreasing to 130 K at the winter mesospause. The temperature field T_r calculated in this way will be referred to below as the "radiatively-determined temperature". A comparison of this calculated temperature field with a typical observed January-mean zonally-averaged temperature (Figure 6-1) shows some overall similarities, but also some strikingly different features. In particular, the observed north polar night is much warmer (by about 20 K in the in the lower stratosphere, increasing to about 100 K in the mesosphere) than the corresponding T_r, while the observed equatorial tropopause is a little colder than T_r and the lower stratosphere in winter mid-latitudes is a little warmer. The July-mean zonally-averaged temperature (Figure 6-2) is roughly a mirror image of January, except that the southern winter polar midstratosphere, at about 180 K, is only just above the radiatively-determined value.

It is also of some interest to consider what zonal winds would be associated with a radiatively-determined temperature field like that of Figure 6-30. From thermal-wind balancc, assuming zero wind (or the observed zonal mean winds of a few ms^{-1}) at the ground, one would calculate extremely strong westerlies in the winter polar night and quite strong easterlies in the summer hemisphere (see Figure 6-31). In both hemispheres the magnitude of the winds would increase with height throughout the stratosphere and mesosphere. The observed zonal mean winds for January are close to being in thermal wind balance with the zonal-mean temperatures (at least in extratropical regions), and hence show a rather more moderate growth with height, peaking near 60 km and decreasing to small values at the mesopause (Figure 6-1). In July the Southern Hemisphere winter winds are stronger than their Northern Hemisphere winter counterparts of January.

Although some of the differences between the observations and the time-marched radiative-photochemical-convective calculations may be due to deficiencies in the radiative aspects of the model, by far the most important cause of these differences is the presence of *in situ* dynamical processes, which are deliberately ignored in this particular model. The extra heating or cooling that must be provided by the presence of dynamical effects is often called the "dynamical heating" [e.g. Fels *et al.*, 1980]. Some of the dynamical processes contributing to this heating would occur in a middle atmosphere whose circulation was zonally-symmetric; however, simple arguments suggest that the dynamical processes that are most important in accounting for departures from T_r (in the extra-tropics, at least) are associated with deviations from zonal symmetry - the "eddies" or "waves". The details of how eddies can accomplish this task are discussed in Sections 6.2.2-6.2.4. (A basic caveat to be noted here is that, although this definition of an eddy or wave is convenient mathematically, it may not always be the most suitable from a physical point of view).

6.2.2 Some Simple Zonally-Averaged Models of the Middle Atmosphere

To gain insight into the ways in which dynamical processes can lead to departures of the zonal-mean temperature from the temperature T_r of a hypothetical atmosphere controlled by radiative, photochemical and convective effects, it is helpful to begin by considering some rather simple models of the extratropical middle atmosphere. A convenient starting-point is the quasigeostrophic set of zonal-mean equations on a mid-latitude beta-plane, in log pressure coordinates. The "transformed Eulerian-mean" versions of these equations [e.g., Edmon *et al.*, 1980; Dunkerton *et al.*, 1981] can be written

$$\frac{\partial \bar{u}}{\partial t} - f_o \bar{v}_* = \varrho_o^{-1} \underset{\sim}{\nabla} \cdot \underset{\sim}{F} + \bar{X} \equiv G, \qquad (1a)$$

$$\frac{\partial \bar{T}}{\partial t} + N^2 H R^{-1} \bar{w}_* = \bar{J}/c_p, \qquad (1b)$$

$$\frac{\partial \bar{v}_*}{\partial y} + \frac{1}{\varrho_o} \frac{\partial}{\partial z} [\varrho_o \bar{w}_*] = 0, \qquad (1c)$$

$$f_o \frac{\partial \bar{u}}{\partial z} + RH^{-1} \frac{\partial \bar{T}}{\partial y} = 0. \qquad (1d)$$

Here $z = -H \ln [p/p_0]$, where p is pressure, p_0 a constant reference pressure and H a representative pressure scale height (typically about 7 km, in which case z is approximately equal to geometric height throughout the middle atmosphere). x and y denote eastward and northward distance, respectively, from some mid-latitude origin, R is the gas constant for dry air and c_p the specific heat at constant pressure. f_0 is a mean

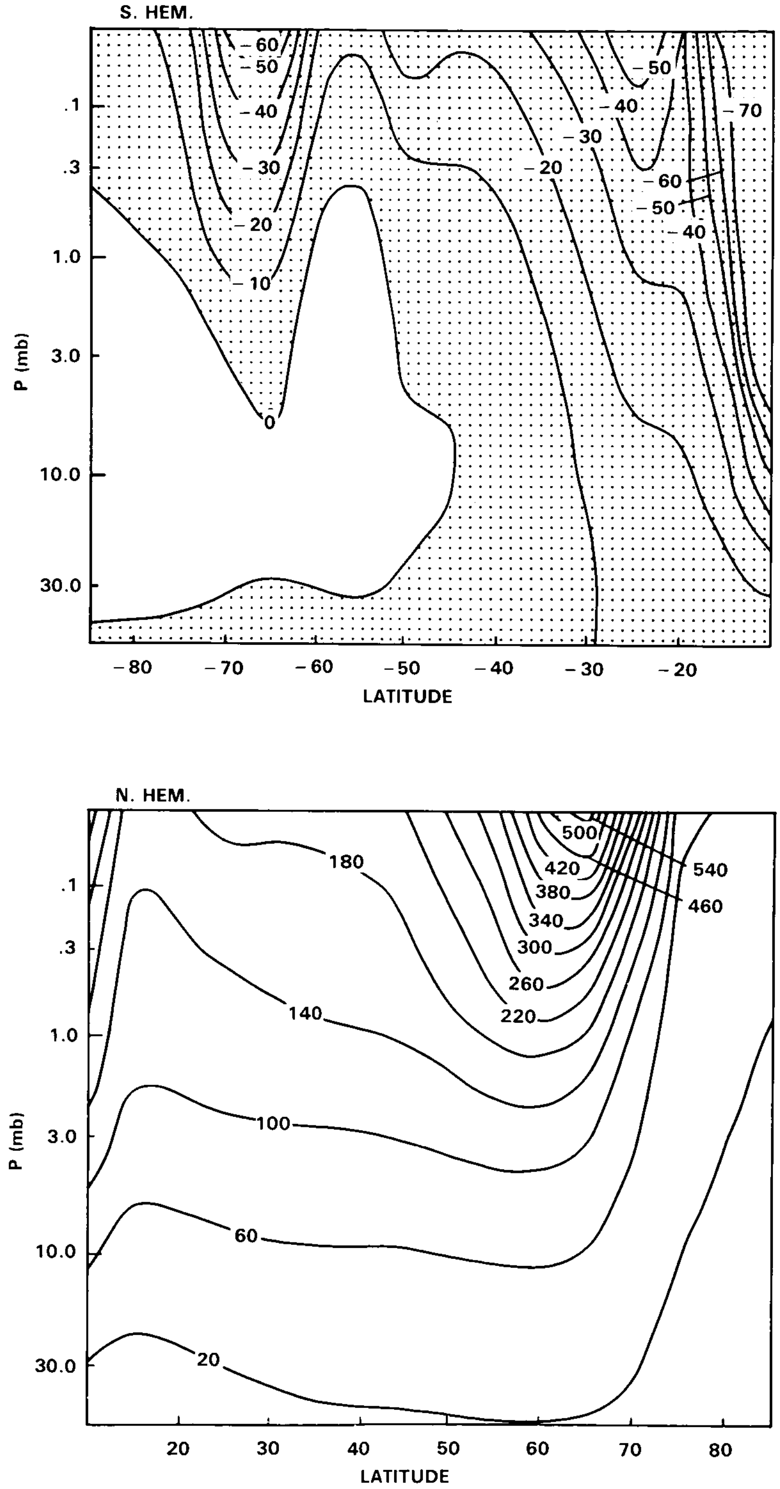


Figure 6-31. Geostrophic winds U (θ,P) calculated from the January 15 temperatures of Figure 6-30. The value of U (θ, 50 mb) is taken from Oort and Rasmussen, and the thermal wind equation integrated upward from 50 mb. The contours have been modestly handsmoothed. [After Fels and Schwarzkopf, 1985].

Coriolis parameter, $\bar{u}$ is the zonal mean wind and $\bar{T}$ the zonal-mean temperature. $(\bar{v}_*, \bar{w}_*)$ is the "residual mean meridional circulation", defined by:

$$\bar{v}_* \equiv \bar{v} - \frac{1}{\varrho_0}\frac{\partial}{\partial z}\left[\frac{\varrho_0 \overline{v'T'}}{N^2HR^{-1}}\right], \quad \bar{w}_* \equiv \bar{w} + \frac{\partial}{\partial y}\left[\frac{\overline{v'T'}}{N^2HR^{-1}}\right] \tag{2}$$

where $(\bar{v}, \bar{w})$ is the zonal-mean meridional circulation and $\overline{v'T'}$ is the northward eddy temperature flux. $\varrho_0[z]$ is a basic density, proportional to p, and $N^2HR^{-1} = dT_0/dz + \varkappa T_0/H$, where $T_0[z]$ is a reference temperature profile and $\varkappa = R/c_p$.

The eddy forcing of the mean flow is represented in (1) by an effective force, equal to $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$, acting on the mean zonal flow. The vector $\underset{\sim}{F}$ is called the Eliassen-Palm [EP] flux [Eliassen and Palm, 1961]; its full definition in log-pressure coordinates is given, for example, by Dunkerton *et al.*, [1981, eqs (A1)]. For quasigeostrophic, planetary-scale waves, it reduces to:

$$\underset{\sim}{F} = (F_y, F_z) = [-\varrho_0\, \overline{u'v'},\ \varrho_0 f_0\, \overline{v'T'}/N^2HR^{-1}] \tag{3a}$$

where $\varrho_0 \overline{u'v'}$ is the northward eddy flux of zonal momentum, while for small-scale, vertically-propagating internal gravity waves it is given by:

$$\underset{\sim}{F} = [0,\ -\ \varrho_0 \overline{u'w'}] \tag{3b}$$

where $\varrho_0 \overline{u'w'}$ is the vertical eddy flux of zonal momentum. The quantity $\bar{X}$ represents all other contributions to the zonal force per unit mass acting on the mean flow. $\bar{J}$ represents mean diabatic effects and, if small-scale or molecular diffusion of heat is negligible, equals the zonal-mean net radiative heating rate.

One advantage of the transformed set (1) over the standard "Eulerian-mean" equations [e.g., eqs (2) of Dickinson, 1969] is that no "eddy heating" terms appear in the mean thermodynamic equation (1b), under quasigeostrophic scaling. The only (large-scale) "eddy forcing" is then $\varrho_0^{-1}\, \underset{\sim}{\nabla} \cdot \underset{\sim}{F}$, in (1a). Another advantage is that $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$, unlike the "eddy-forcing" terms in the standard equations, can be related directly to certain rather general physical properties of the eddies, as will be discussed in Section 6.2.3.

Consider now a hypothetical steady-state atmosphere in which the seasonal cycle is absent; with time derivatives set to zero, (1a,b) give:

$$-f_0 \bar{v}_* = G, \quad N^2HR^{-1}\bar{w}_* = \bar{J}/c_p, \tag{4}$$

and substitution into (1c) yields

$$-f_0^{-1}\frac{\partial G}{\partial y} + \frac{1}{\varrho_0}\frac{\partial}{\partial z}\frac{\varrho_0 \bar{J} \varkappa}{N^2H} = 0. \tag{5}$$

This shows how the net heating $\bar{J}$ must be related to $G \equiv \varrho_0^{-1}\, \underset{\sim}{\nabla} \cdot \underset{\sim}{F} + \bar{X}$ in this hypothetical state. If $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$ and $\bar{X}$ both vanish then $\bar{J}$ must also vanish [provided that $\bar{w}_* = 0$ at the lower boundary]; the atmosphere is then in radiative equilibrium (if small-scale thermal diffusion is negligible), with the long-wave cooling balancing the solar heating everywhere. Then $\bar{T} = T_r(y,z)$, say, $u = u_r(y,z)$ [where $f_0\ \partial u_r/\partial z + RH^{-1}\partial T_r/\partial y = 0$ by (1d)] and $\bar{v}_* = \bar{w}_* = 0$ by (4), i.e., the residual mean meridional circulation vanishes.

Suppose next that G = 0 still, but that time-dependence is allowed by letting the solar heating take on an annual variation $\overline{J}_s(y,z,t)$. Further progress is aided by a parameterization of $\overline{J}$ in terms of $\overline{T}$; as a simple example consider the Newtonian cooling form

$$\frac{\overline{J}}{c_p} = -\frac{(\overline{T} - T_r)}{\tau_r(z)} \tag{6}$$

where $T_r(y,z,t)$ is the temperature calculated from a time dependent radiative-photochemical model (such as that from which Figure 6-30 was obtained) with specified solar heating $\overline{J}_s(y,z,t)$, and $\tau_r(z)$ is a radiative relaxation time. This parameterization is not expected to be quantitatively accurate for large departures of $\overline{T}$ from T_r; however, it does contain the important physical feature of relating the net heating to departures from a radiatively-determined T_r. From Equations (1) and (6) it can then be shown that:

$$\left[\frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial z}\left(\varrho_o^{-1}\frac{\partial}{\partial z}\varrho_o\epsilon\right)\right]\frac{\partial\overline{T}}{\partial t} + \frac{\partial}{\partial z}\left[\varrho_o^{-1}\frac{\partial}{\partial z}\left(\frac{\varrho_o\epsilon}{\tau_r}[\overline{T}\text{-}T_r]\right)\right] = 0, \tag{7}$$

where $\epsilon(z) \equiv f_0^2/N^2(z)$. In this equation the term in T_r (related to the solar heating $\overline{J}_s$) provides the *forcing,* while T represents the *response.* In general, $\overline{T}$ will follow $T_r(y,z,t)$, but will be somewhat lagged in time and somewhat differently distributed in space; the zonally-symmetric dynamics provides a kind of "inertia". Since $\overline{T} \neq T_r$ in general, $\overline{J} \neq 0$ by (6) and $(\overline{v}_*,\overline{w}_*) \neq 0$ by Equations (1a,b,c,d). It should be emphasized that the non-vanishing of the *net* heating $\overline{J}$ is essentially due to the presence of this "dynamical inertia", and cannot be regarded as imposed by by external agencies. To put it another way, although the solar heating has been specified in advance in this model, the long-wave cooling must be determined as part of the solution.

Simple order-of-magnitude arguments, assuming height scales of order H and horizontal scales of order L, where $f_0{}^2L^2 \sim H^2N^2$ (this fails near the equator, where $f_0{}^2L^2 << N^2H^2$, and where quasigeostrophic theory generally breaks down in any case) show that:

$$\left[\frac{\partial\overline{T}}{\partial t}\right] \sim \left[\frac{\overline{T} - T_r}{\tau_r}\right]$$

approximately, in this model. If $\Delta\overline{T}(y,z)$ is the maximum annual variation of $\overline{T}$ in the model and τ is a seasonal timescale (say 3 months), we have $\partial\overline{T}/\partial t \sim \Delta\overline{T}/\tau$. However, typical radiative relaxation times are mostly less than about 20 days in the middle atmosphere, so that $\tau_r << \tau$, whence

$$\overline{T} - T_r \sim \frac{\tau_r}{\tau}\Delta\overline{T} << \Delta\overline{T}\ , \tag{8}$$

and departures of $\overline{T}$ from the radiatively-determined value $T_r(y,z,t)$ are much less in this model than the actual annual swing $\Delta\overline{T}$. The model therefore predicts extratropical temperatures $\overline{T}(y,z,t)$ that are always close to the temperatures $T_r(y,z,t)$ determined from radiative-convective considerations. (Note that this conclusion does not depend on the details of the parameterization (6)). It is therefore clear that this simple model fails to predict the observed *large* departures of $\overline{T}$ from T_r in certain parts of the extratropical middle atmosphere (e.g., the polar night and the summer upper mesosphere). Additional effects must be included if a basic understanding of the annual variations of the temperatures structure of these regions is to be obtained.

Now, if $G \equiv \varrho_0{}^{-1}\underset{\sim}{\nabla}\bullet\underset{\sim}{F} + \overline{X}$ is retained, the set (1), with parameterization (6) yields:

$$\underbrace{\left[\frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial z}\left[\varrho_0{}^{-1}\frac{\partial}{\partial z}\varrho_0\epsilon\right]\right]\frac{\partial\overline{T}}{\partial t}}_{[1]} + \underbrace{\frac{\partial}{\partial z}\left[\varrho_0{}^{-1}\frac{\partial}{\partial z}\left[\frac{\varrho_0\epsilon}{\tau_r}[\overline{T} - T_r]\right]\right]}_{[2]}$$

$$+ \underbrace{\left[\frac{f_0 H}{R}\frac{\partial^2 G}{\partial y\partial z}\right]}_{[3]} = 0. \tag{9}$$

Suppose that G varies on a timescale τ_w : except for rapid events like sudden warmings we have $\tau_w \sim \tau >> \tau_r$. (If rapid events of this kind are present, they can be removed by averaging over a time $\tau_w = 0(\tau)$: see Andrews *et al.*, [1983].) A scaling gives the ratio of terms as:

$$[1] : [2] : [3] = \frac{\Delta\overline{T}}{\tau} : \frac{\overline{T} - T_r}{\tau_r} : \frac{f_0 L\Delta G}{R} \tag{10}$$

if $f_0^2L^2 \sim N^2H^2$ (thus excluding equatorial regions again) and ΔG is the variation in G over time τ_w. In the polar night stratosphere and in the upper mesosphere it is found that $(\overline{T}\text{-}T_r)/\tau_r >> \Delta\overline{T}/\tau$ in general, so here the effects represented by G must be large enough to give a balance between terms [2] and [3]. The term in $\partial\overline{T}/\partial t$ in (9) is therefore small in these regions: equivalently the time derivatives in (1a) and (1b) are small, so that (according to this model) the balances expressed by (4) and (5) hold approximately, at each t, in those regions where $\overline{T}$ exhibits large departures from T_r.

This simple model suggests that dynamical effects which contribute to the mean zonal force per unit mass $G \equiv \varrho_0{}^{-1}\underset{\sim}{\nabla}\bullet\underset{\sim}{F} + \overline{X}$ may be responsible for maintaining the large departures of $\overline{T}$ from T_r that are observed in parts of the middle atmosphere. Conversely, it also suggests that those regions - such as parts of the midlatitude lower stratosphere and the summer stratopause - which are observed to be close to radiative equilibrium [e.g. Houghton, 1978; Wehrbein and Leovy, 1982] may be in that state because of the absence of dynamical effects that can produce a significant G.

6.2.3 Aspects of Wave, Mean-Flow Interaction Theory

We now discuss the dynamical processes that are likely to contribute importantly to the forcing term $G \equiv \varrho_0{}^{-1}\underset{\sim}{\nabla}\bullet\underset{\sim}{F} + \overline{X}$. First consider $\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$: it was mentioned in Section 6.2.2 that one advantage of the transformed Eulerian-mean equations (1) is that the "eddy-forcing" term $\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$ depends on certain physical properties of the eddies. This dependence is expressed by the "generalized Eliassen-Palm Theorem" (Andrews and McIntyre, 1976, 1978a; Boyd, 1976) which, for eddies on a flow that is basically zonal, takes the form:

$$\frac{\partial A}{\partial t} + \underset{\sim}{\nabla}\bullet\underset{\sim}{F} = D + \left[\begin{matrix}\text{terms that are cubic}\\ \text{in wave amplitude}\end{matrix}\right] \tag{11}$$

which is of a standard form expressing the conservation of some quantity, here a measure of wave activity, A. A and D, like $\underset{\sim}{F}$, are zonal-mean quadratic functions of eddy quantities, while the term that is cubic in eddy amplitude is negligible for waves whose amplitudes are small enough for linear theory to be valid. D involves wave dissipation or forcing, and thus vanishes if the waves are conservative; $\partial A/\partial t$ vanishes if the waves are steady. Thus (11) states that $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$ depends on wave transience, nonconservative wave effects and wave nonlinearity. If all of these are absent then $\underset{\sim}{\nabla} \cdot \underset{\sim}{F} = 0$. Conversely, if waves are strongly transient, nonconservative or nonlinear (or any combination of these), we can anticipate large values of $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$ and perhaps large departures of $\overline{T}$ from T_r. It should be noted that A generally involves Lagrangian quantities like the northward particle displacement: "transience" thus means "Lagrangian transience" in this context (see Section 6.5.2). Important contributions to $\overline{X}$ are likely to be produced for example by turbulent mixing associated with the "breaking" of large-amplitude gravity waves (see Section 6.2.4).

6.2.4 Implications of the Theory

The theory sketched in the two previous sections, involving the use of a quasigeostrophic beta-plane model for the mean flow and restrictions to small amplitude waves in Equation (11), is clearly too crude for accurate quantitative comparisons with observations. (Some of these constraints can in fact be relaxed: for instance, similar ideas can be formulated for the primitive equations on the sphere - in the extratropics - by using isentropic coordinates.) However, perhaps a more important use of the theory is to provide qualitative insights into the physical mechanisms which maintain the zonal-mean climatological state of the middle atmosphere. It suggests how eddy motions on various scales can keep certain parts of the middle atmosphere far from the state predicted by radiative-convective models, and throws light on the physical eddy processes that may be responsible for this. It thus helps to clarify the role of the eddies in the maintenance of the climatological mean state. For instance, if in a climatological average certain eddies are found to exhibit a significant $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$, then, under the scaling described at the end of Section 6.2.2, the corresponding eddy forcing $G = \varrho_0^{-1} \underset{\sim}{\nabla} \cdot \underset{\sim}{F}$ is responsible for driving a climatological residual circulation $(\overline{v}_*, \overline{w}_*)$ and thence inducing a net radiative heating $\overline{J}$, by preventing $\overline{T}$ from relaxing to T_r [see Kurzeja, 1981; Plumb, 1982; Apruzese *et al.*, 1982]. (This point was originally made by Dickinson [1969] using the northward eddy flux of quasigeostrophic potential vorticity $\overline{v'Q'}$ as a measure of eddy forcing. Under quasigeostrophic scaling this flux equals $\varrho_0^{-1} \underset{\sim}{\nabla} \cdot \underset{\sim}{F}$: cf. Section 6.5.3). In general, therefore, it is clearly misleading to regard the residual circulation as a "diabatic circulation", *driven* by an externally-imposed net radiative heating $\overline{J}$ [Kurzeja *et al.*, 1984].

Of course, not all situations in the middle atmosphere will be as simple as this: for example, one cannot generally regard the eddy structure, and therefore $\underset{\sim}{\nabla} \cdot \underset{\sim}{F}$, as independent of the mean flow, nor, indeed, of radiative processes, since these contribute to the nonconservative dissipation of the waves. However, the case just described does caution against too naive an application of names like "eddy transport" and "mean transport" to the various terms in the mean equations of motion. Not only may such terms differ between one particular formulation and another - e.g., between the transformed set (1) and the standard Eulerian-mean set or the Lagrangian-mean set (Andrews and McIntyre, [1978b]; see also Plumb [1983b], for a slightly different example) - but they may none of them be entirely consistent with results of thought-experiments or model calculations in which eddies are artificially excluded from the middle atmosphere [e.g., Andrews *et al.*, 1983].

These insights have already proved useful in the interpretation of tracer transport [e.g., Mahlman *et al.*, 1981, 1984] and the behavior of numerical models of the middle atmosphere (see Section 6.3). They should also help in the future investigation of the observed middle atmosphere, not only by providing a basic theoretical framework for interpretation but also by indicating specific aspects of wave motions, in specific regions of the middle atmosphere, which may require particular attention. These aspects will now be considered in more detail.

The first question to ask is what kinds of eddy or wave motion could be responsible for the large observed climatological departures of $\overline{T}$ from T_r in thc polar night stratosphere (especially in the northern hemisphere) and in the upper mesosphere.

Taking first the polar winter stratosphere, obvious candidates are the planetary-scale wave disturbances, including the quasi-stationary planetary waves that are familiar features of the northern winter stratosphere but are of smaller amplitude in the southern winter (Section 6.1.3). The theory sketched above indicates that the effectiveness of the waves in driving $\overline{T}$ away from T_r depends mainly on the Eliassen-Palm flux divergence $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$, and that $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$ in turn depends on the waves being transient, nonconservative or nonlinear. One class of event in which all three of these processes are probably significant, namely the "breaking" of planetary waves, has recently been identified from satellite-derived measurements of the stratosphere [McIntyre and Palmer, 1983, 1984; Clough *et al.*, 1985]. This phenomenon involves the rapid, irreversible deformation of otherwise wavy material contours, and is most clearly depicted in terms of maps on isentropic surfaces of Ertel's potential vorticity P ($\equiv \varrho^{-1}\underset{\sim}{\zeta}_a\cdot\nabla\theta$, where ϱ is density, $\underset{\sim}{\zeta}_a$ is absolute vorticity and θ is potential temperature or entropy). Like θ, P is approximately conserved following fluid particles (it is exactly conserved for adiabatic, frictionless flow), and the maps therefore give a picture (which is somewhat blurred, owing to the finite resolution of satellite measurements) of the behavior of distorting, quasi-material lines of fluid particles in the stratosphere. During "breaking" planetary-wave events, long "tongues" of potential vorticity are strung out on isentropic surfaces in an apparently irreversible manner (see Figure 6-24), and significant mixing of potential vorticity may then occur. Several key aspects of this process have yet to be clarified, however, including the role of radiative effects and the extent to which the mixing is irreversible [Clough *et al.*, 1985].

If such planetary-wave breaking events are common in the northern winter stratosphere, they could well lead to systematically large climatological contributions to $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$ there. This contribution is in addition to that expected in the absence of wave breaking from the effects of wave dissipation through radiative damping, which may be particularly important in the upper stratosphere where the damping rates become large. Further research is needed to understand the relative contribution of these effects and to determine whether their net contribution to $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$ is large enough to account for the observed departures of the climatological temperatures from radiative equilibrium (see Section 6.4.2). However, some support for the application of the simple theory given here comes from the general circulation model results of Mahlman and Umscheid [1984] (in which the model's polar night stratosphere being too close to T_r may be associated with over-weak planetary wave amplitudes: cf. Section 6.3.3). Also relevant is the observation that the southern hemisphere winter stratosphere is closer to the radiatively-determined state than is the northern winter stratosphere: this may be due to the weaker southern winter planetary wave amplitudes mentioned above.

Consider now the departure of $\overline{T}$ from T_r in the upper mesosphere. In the summer hemisphere, at least, these cannot be due to quasi-stationary planetary waves, since such waves are essentially absent there. Following a suggestion of Houghton [1978], it is now generally believed that gravity waves provide the major part of the forcing in the upper mesosphere. In the absence of significant dissipation or reflection, the velocity and temperature amplitudes of such waves grow roughly exponentially with height as they propagate upwards from the lower atmosphere. Eventually nonlinear effects become important, leading to wave "breaking", as signalled by the overturning of isentropic surfaces, and hence turbulence, small-scale mixing and dissipation; as a result the wave growth with altitude is halted [Hodges, 1967]. This "saturation" process leads, as seen in Section 6.1.6, to a vertical momentum flux convergence, and thus a contribution to $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$ (by (3b)), which in fact tends to accelerate the mean flow towards the phase speed of the waves (Figure 6-32). The turbulence also causes diffusion of mean momentum and thus contributes to $\overline{X}$; diffusion of heat and chemical constituents is likely to occur in the same way (see Section 6.5).

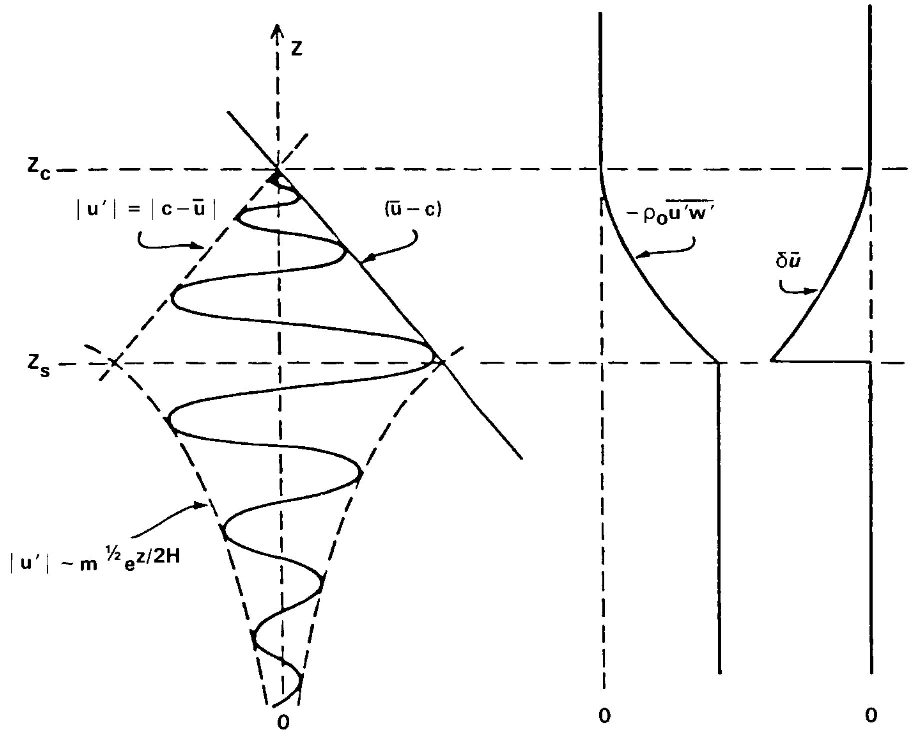


Figure 6-32. Schematic of the growth with height and saturation of a gravity wave due to convective instability. Wave damping produces both a divergence of the vertical flux of horizontal momentum and an acceleration of the mean flow toward the phase speed of the wave. Deceleration and diffusion cease above the critical level ($z = z_c$) in the linear theory. [After Fritts, 1984].

A simple model of this process was suggested by Lindzen [1981], who estimated the range of zonal phase speeds c of gravity waves that might be expected to propagate through the mean zonal wind, avoiding absorption at critical levels where $\bar{u} = c$, and reach the mesosphere (Figure 6-33). He then used a linear theory to estimate the altitude at which these waves would break and to calculate $\underset{\sim}{\nabla}\cdot\underset{\sim}{F}$ and the diffusion associated with the breaking waves. He found that $G = \varrho_0^{-1}\underset{\sim}{\nabla}\cdot\underset{\sim}{F} + X$ could equal several tens of metres per second per day above a "breaking level" in the middle or upper mesosphere. This is of the right order of magnitude to account for the observed departures of the upper mesosphere from its radiatively determined state.

Further work in this area, and application to simple mechanistic models, includes the papers by Matsuno [1982], Holton [1982, 1983], Dunkerton [1982a,b], Weinstock [1982], Schoeberl *et al.* [1983], Holton, and Zhu [1984] and Miyahara [1984]. The mechanistic models have been fairly sucessful in simulating the observed mean temperature and zonal wind in the upper mesosphere: mean north-south winds also seem to agree reasonably well with the limited number of observations (Section 6.1). The possible importance of the refraction of small-scale waves by planetary waves has also been considered [Dunkerton and Butchart, 1984] as has the possible damping or forcing of planetary waves by gravity waves [Miyahara, 1985; Schoeberl and Strobel, 1984; Holton, 1984].

The development of reliable parameterizations of the mean momentum forcing G due to mesospheric gravity waves will require an improved understanding of the mechanics of the breaking process and much more observational information on the global morphology of gravity waves in the middle atmosphere. Other wave motions which may contribute to the maintenance of climatological mean departures of $\overline{T}$

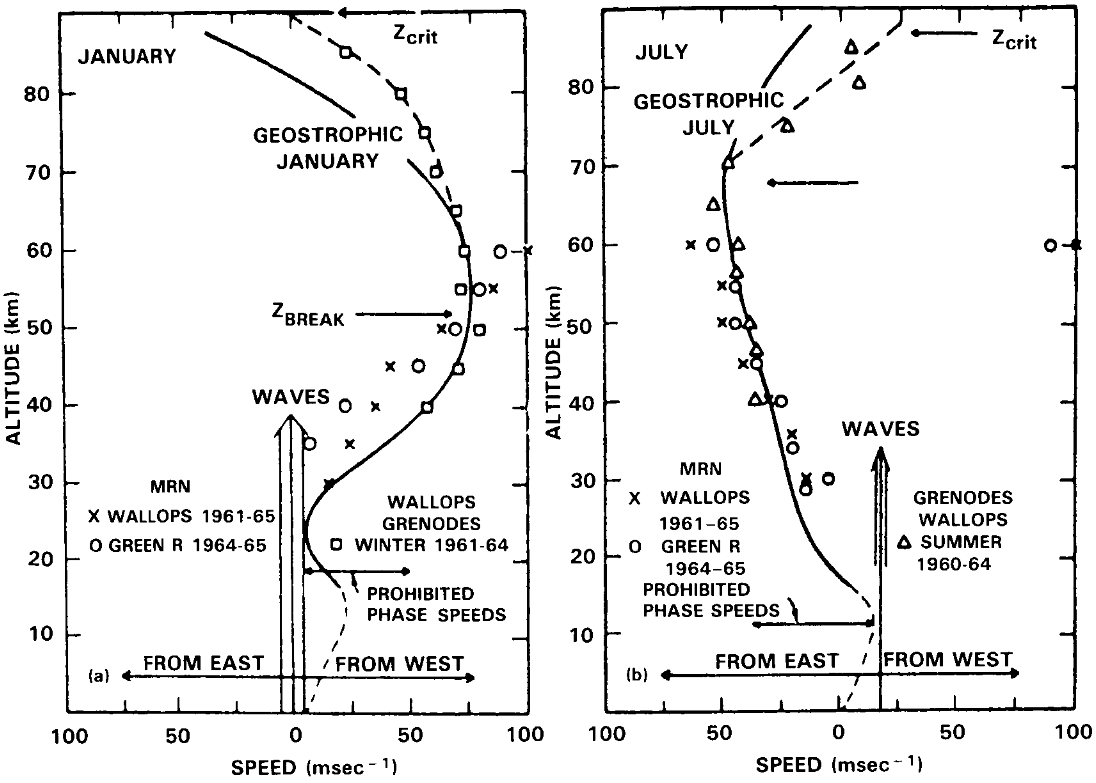


Figure 6-33. Profiles of the zonal wind as a function of height at mid-latitudes for winter and summer and the permitted and prohibited phase speeds for tropospheric gravity waves reaching the mesosphere. [After Lindzen, 1981].

from T_r include atmospheric tides, which may also break in the mesosphere and could contribute to G below and above their breaking altitudes [Lindzen, 1981; Hamilton, 1981a; Miyahara, 1984]. Gravity waves may perhaps break under some circumstances in the stratosphere and lower mesosphere, and may make some significant contributions to G at these levels [Hamilton, 1983a].

6.2.5 Stratospheric Sudden Warmings

The stratospheric sudden warming is a spectacular phenomenon which is observed to occur in certain northern hemisphere winters; a "major" warming occurs roughly every second year. It is manifested by a breakdown and (often) reversal of the basic zonal-mean polar westerly vortex, accompanied by a rapid rise in temperature in the stratospheric polar cap. Minor warmings of lesser amplitude occur more frequently in both hemispheres during their respective winters. In a sense, the major warming is just a large-amplitude example of a fairly frequent event. The advent of satellite measurements and the development of numerical models of the middle atmosphere have contributed enormously to our knowledge of sudden warmings. Recent reviews of the phenomenon include those by Labitzke [1981] (dealing mostly with observations) and McIntyre [1982] (dealing mostly with theory).

There is now little doubt that sudden warmings are intimately linked with the propagation from the troposphere into the stratosphere of some form of large-amplitude planetary-wave disturbance (indeed, there appears to be an association between major warmings and large-amplitude blocking events in the troposphere). This kind of dynamical mechanism was originally proposed by Matsuno [1971], although some of the details of his hypothesis have since required modification. A standard procedure for investigating observed and modeled sudden warmings has been to split each variable into a zonal-mean and an "eddy" part, and to study the resulting wave, mean-flow interaction [e.g., Dunkerton *et al.*, 1981; Palmer, 1981a,b; Butchart *et al.*, 1982; O'Neill and Youngblut, 1982; Simmons and Struefing, 1983]. In this case, theory of the type introduced in Sections 6.2.2 and 6.2.3 is helpful: thus the EP vector $\underset{\sim}{F}$ can be regarded as giving an indication of the direction of propagation of the planetary waves through the existing mean flow structure. Under normal climatological conditions a cross-section showing the direction of $\underset{\sim}{F}$ at various points in the meridional plane [Hamilton, 1982b; Geller *et al.*, 1983] suggests a general propagation of planetary waves from the mid- and high-latitude troposphere up into the stratosphere, followed by a tendency for such waves to propagate equatorwards (see Figure 6-34).

There is evidence, however, that this tendency reverses prior to sudden warmings, with $\underset{\sim}{F}$ being refracted poleward, causing some focusing of the waves into the high-altitude polar cap (see Figure 6-35). This may be due to refraction of the waves by a changing mean flow structure, although similar effects can be produced independent of mean flow changes by interference between stationary and transient waves (see Section 6.1.4). Whatever the cause, there are good theoretical grounds to believe that this "focusing" is an important precursor to the subsequent polar warming.

The effect of the mean flow of such focusing can be investigated with Equations (1a,b,c,d). In spherical coordinates, (1a) is replaced by:

$$\frac{\partial \bar{u}}{\partial t} - 2\Omega \sin \phi \, \bar{v}_* = \frac{1}{\varrho_0 a \cos\phi} \underset{\sim}{\nabla} \cdot \underset{\sim}{F} \tag{12}$$

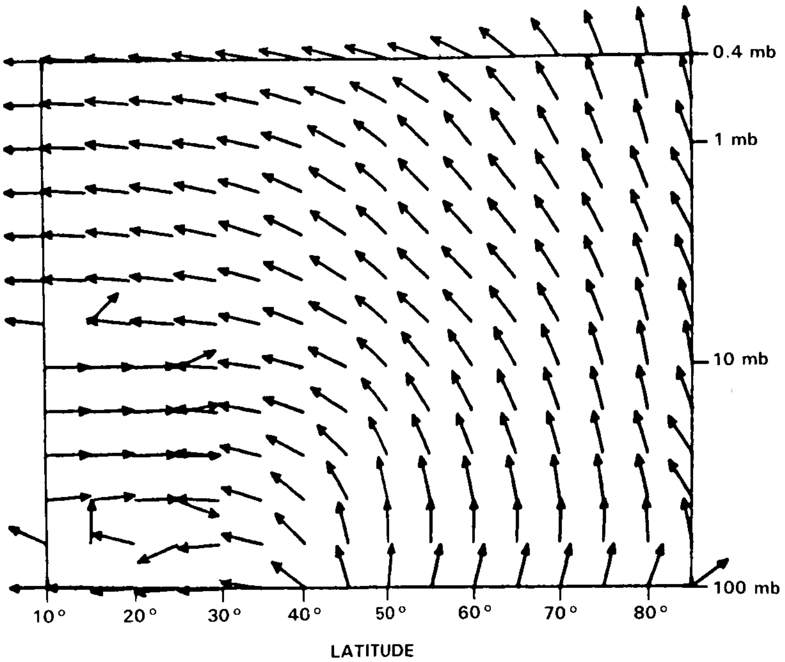


Figure 6-34. The climatological January-mean directions of the geostrophic Eliassen-Palm flux $\underset{\sim}{F}$, defined by equation (3), at various latitudes and heights in the northern hemisphere, based on four years of stratospheric data. [From Hamilton, 1982b].

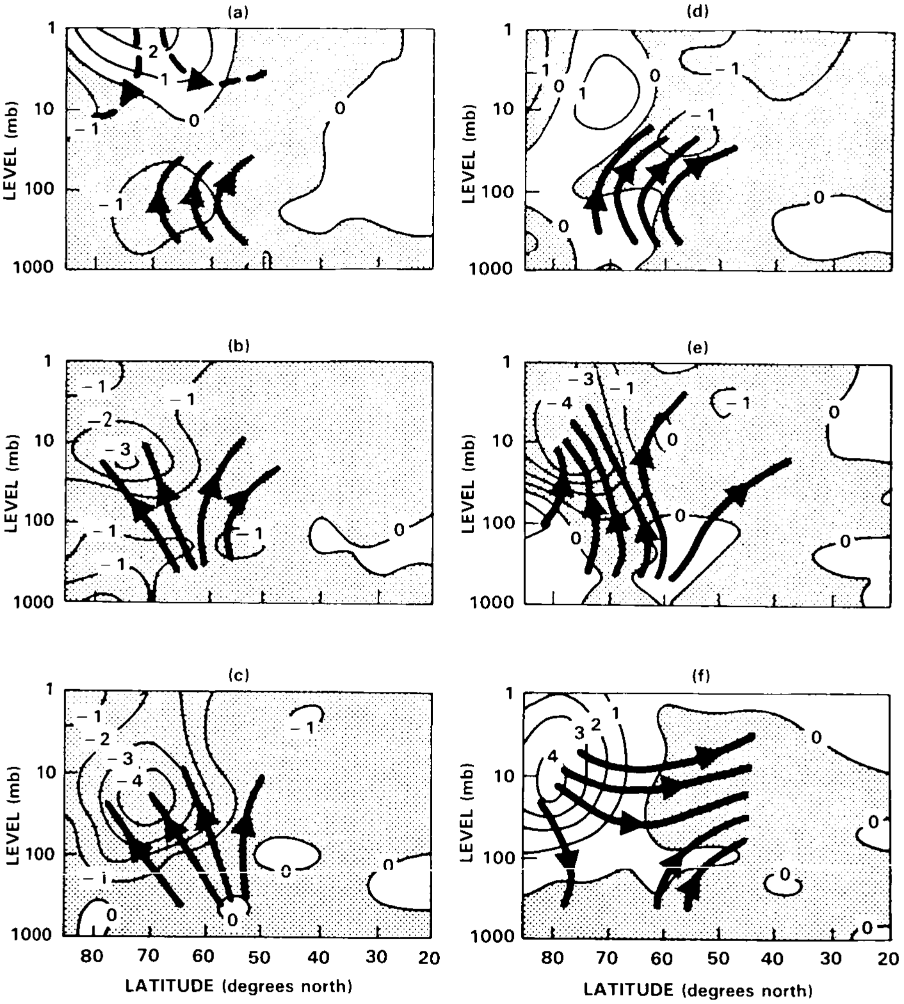


Figure 6-35. "Integral curves" giving the local direction of $\underset{\sim}{F}$, and contours of $(\varrho_0 a\cos\phi)^{-1}\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$ labelled in units of 10^{-4} ms^{-2} (negative values stippled) for several days in February 1979; (a) 17th, (b) 19th, (c) 21st, (d) 23rd, (e) 26th, (f) 28th. (Dashed integral curves are dominated by zonal wavenumber 1 disturbances; full integral curves are dominated by wavenumber 2). [From Palmer, 1981a).

(neglecting $\overline{X}$), where ϕ is latitude and Ω and a are the earth's rotation rate and radius, respectively. If waves are transient, nonconservative or nonlinear, then significant values of $\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$ are expected, by Equation (11). The "transience" contribution $(-\partial A/\partial t)$ to $\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$ (and also possibly the dissipation contribution D) will be negative as waves first penetrate into the high polar cap and, while there is no simple way of estimating the nonlinear contribution, model calculations and observational data show that the net result of all these processes is to produce a negative $\underset{\sim}{\nabla}\bullet\underset{\sim}{F}$ there (see Figure 6-35). The effect of this negative forcing on the right of (12) will be enhanced by the small values of ϱ_0 and $\cos\phi$ in the high-altitude polar regions. Consideration of a full set of equations analogous to (1) shows that this large negative zonal force leads to a rapid deceleration $(\partial\overline{u}/\partial t < 0)$, despite the mitigating effect of the Coriolis term $2\Omega\sin\phi\overline{v}_*$. By thermal-wind balance this deceleration is associated with a rapid rise in temperature, as observed in the sudden warming. In terms of the thermodynamic equation (1b) we find $\partial\overline{T}/\partial t = -N^2H\overline{w}_*/R > 0$, with $\overline{J}$ negligible on the timescale of the warming: thus the temperature increase is associated with a residual mean descent $(\overline{w}_* < 0)$, although the Eulerian-mean vertical velocity $\overline{w}$ is often found to be positive. It is observed that air parcels are descending, on average, too [Mahlman, 1969b; Dunkerton *et al.*, 1981]. (But note that $\overline{w}_*$ is not generally equal to the Lagrangian-mean vertical velocity during transient wave events).

This description of sudden warmings leaves a number of questions to be answered. First, how important are changes in the mean flow structure in favoring focusing of planetary waves into the polar cap, and how do these changes arise? Such ''preconditioning'' appears in some (but not all) cases to be due to an earlier wave event. Second, what is the cause of the anomalous amplification of planetary wave activity? It may be a result of as yet poorly understood processes in the troposphere; other possibilities are near-resonance of the stationary waves [Tung and Lindzen, 1979], which may even be self-induced by nonlinear ''self-tuning'' [Plumb, 1981], or constructive interference between stationary and traveling wave components.

Very recent studies suggest that for some planetary-wave events, especially those of large amplitude, the separation into zonal mean and eddy parts may be an unnecessarily complicated way of viewing the dynamics and may perhaps give misleading impressions of causality [see Clough *et al.*, 1985]. As an alternative, the use of isentropic potential vorticity maps (Hoskins *et al.*, 1985; see Sections 6.2.4 and 6.4.4) may provide a simpler method of analysing sudden warmings and other transient phenomena involving large departures from a zonally-symmetric state. However, a body of theory, comparable to that described above for the ''zonal-mean, eddy'' separation, and capable of showing how to analyse such maps in a quantitative way, has yet to be developed.

6.2.6 The Non-Zonally-Averaged Climatological State

In Sections 6.2.2-6.2.4 the maintenance of the climatological *zonal-mean* state of the middle atmosphere was discussed. A rather different process will be briefly considered in this section, namely the control of the climatological state in which the zonal mean is not taken. This question needs to be looked at in a different way, since the time-mean state now contains stationary (or quasi-stationary) planetary-wave deviations from zonal symmetry. These ''stationary eddies'' were included among those wave motions which were invoked to explain departures of the zonal-mean climatology from a radiatively-determined state.

The discussion of the maintenance of a zonally-asymmetric time-mean state is hampered by the lack, at present, of a comprehensive theoretical framework for the dynamical evaluation and interpretation of the processes involved. As in the zonal-mean case, one must not give too much weight to naive physical interpretations of the various terms in the time-averaged equations of motion. However, some general qualitative remarks can be made. In the first place, monthly-mean climatological data would presumably be zonally-symmetric in an idealized atmosphere whose lower boundary contained no zonal asymmetries, since any traveling-wave structures would tend to be removed by the climatological time-averaging and the phases of any disturbances that happened to be stationary would tend to be randomly distributed in longitude. It thus follows that any zonal asymmetries in the climatological fields will be linked in some way with zonal asymmetries in the earth's surface.

It is now generally accepted that the observed stationary waves in the *troposphere* are ultimately associated with orography and with thermal aspects of the land-sea difference [e.g., Wallace, 1983; Held, 1983; Donner and Kuo, 1984]. Since stationary disturbances of the longest wavelengths can propagate from the troposphere into the stratosphere in the winter hemisphere [Charney and Drazin, 1961], it is to be expected that some, at least, of the climatological zonal asymmetries in the middle atmosphere can be attributed to the orographic and thermal forcing in the troposphere. This expectation is partly confirmed by studies with numerical models of linear stationary waves in climatological zonal-mean states. Some of these [e.g., Matsuno, 1970; Schoeberl and Geller, 1977] impose as a lower boundary condition an observed monthly-mean geopotential height field in the middle or upper troposphere (without enquiring into the precise cause of this field) and compute the linearized response of the stratosphere and mesosphere. These models have achieved a fair degree of success in simulating observed monthly-mean zonally-asymmetric fields in the middle atmosphere. Others [e.g., Lin, 1982; Alpert *et al.*, 1983] include details of the tropospheric forcing as well; in this case it appears to be more difficult to achieve a satisfactory simulation of stationary waves in the middle atmosphere. This reflects our poor understanding of the forcing of the stationary waves. One possibly important failing of such linear models may be the neglect of the time-averaged nonlinear effects of ''transient eddies'' (departures from the monthly-mean fields) on the stationary eddies. Given some basic flow that is zonally asymmetric (perhaps because of the influence of tropospheric thermal or orographic forcing) these transient eddies may help to enhance, or alternatively diminish, the zonal asymmetry. Recent studies [e.g., Hoskins, 1983] suggest that these effects may be important in the troposphere; whether they also take place to a significant degree in the middle atmosphere has not yet been investigated.

6.2.7 Theory of the Low-Latitude Zonal-Mean Circulation

It was mentioned in Section 6.1 that zonal-mean circulation of the equatorial middle atmosphere is dominated by long-period variations, namely the quasi-biennial oscillation (QBO) and semi-annual oscillation (SAO), rather than the annual cycle that predominates in extra-tropical regions. What makes the dynamics of these low-latitude zonal circulations so intriguing and so different from those of high latitudes is that while the latter are qualitatively similar to those expected on the basis of a simple atmospheric response to seasonal thermal driving, albeit with sometimes dramatic modulations induced by tropospherically-forced eddy activity, the former appear to be dependent on eddy transport in a more fundamental way. A major clue that this is so comes from the observation of strong equatorial westerlies during the westerly phases of both the QBO and the SAO. Since inviscid zonally symmetric circulations conserve absolute angular momentum, they would inevitably tend to drive easterlies at the equator by bringing in air of relatively low absolute angular momentum from higher latitudes; therefore zonally asymmetric eddies are required to achieve the implied angular momentum transport. The major eddy motions observed in the tropical middle atmosphere and described in Section 6.1.5 are depicted schematically in Figure 6-36. Of these, only the equatorial Kelvin waves appear capable of transporting westerly momentum into the equatorial stratosphere and it therefore seems clear these motions must be responsible for the westerly phase of the QBO and of the stratopause SAO.

In fact, it is now believed that *both* westerly and easterly phases of the QBO are wave-driven. This belief stems from a suggestion of Lindzen and Holton [1968], updated by Holton and Lindzen [1972], that the QBO results from an interplay between westerly forcing by slow equatorial Kelvin waves and easterly forcing by mixed Rossby-gravity waves, both of which dissipate (and therefore interact with the mean flow) primarily in the lower stratosphere. Their theory, which was discussed further by Plumb [1977] and is reviewed in Plumb [1984], predicts well the major features of the observed time/height structure of the QBO (Figure 6-37; cf. Figure 6-28) and was further supported by the generation of an analogous oscillation in a laboratory experiment of Plumb and McEwan [1978] and by an analysis of the observed momentum budget of the equatorial lower stratosphere [Lindzen and Tsay, 1975]. Therefore the Holton-Lindzen theory has become widely accepted; a few modifications have been proposed, perhaps the most substantial of which is the suggestion of Dunkerton [1983a] (see also Dickinson [1968] and Andrews and McIntyre [1976]) that lateral momentum transport by the quasi-stationary planetary waves of the winter hemisphere may contribute substantially to the driving of the easterly regime.

The thermal structure of the QBO and, correspondingly, the associated (Lagrangian-mean) meridional circulation were investigated in a two-dimensional model by Plumb and Bell [1982]. Their results, depicted

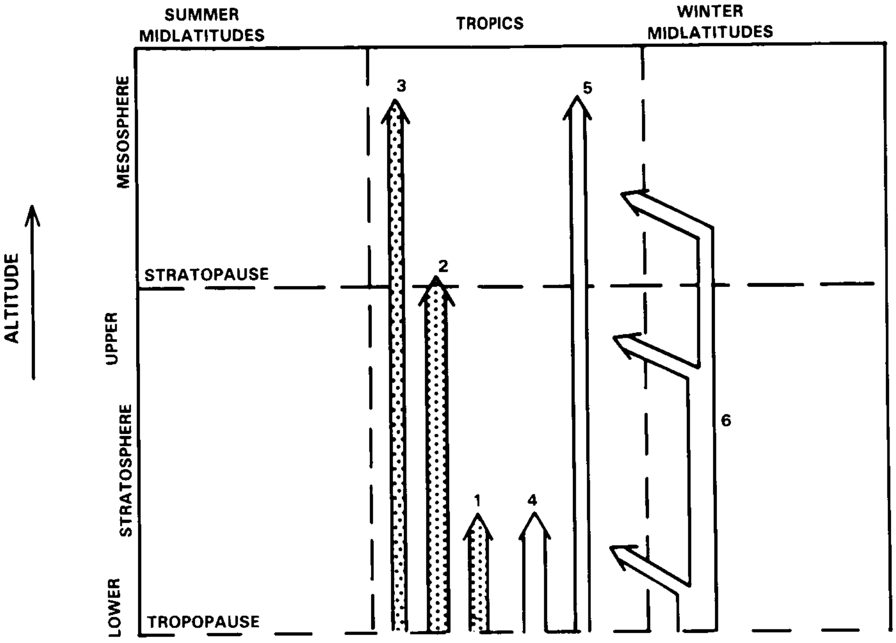


Figure 6-36. Schematic illustration of the propagation of wave motions into the tropical stratosphere and mesosphere. Arrows terminate where waves dissipate and interact with the mean flow. Stippled arrows (open arrows) represent waves which produce an effective westerly (easterly) force on the flow in these regions. 1. Slow Kelvin waves, 2. Fast Kelvin waves, 3. Ultra-fast Kelvin waves and internal gravity waves, 4. Mixed Rossby-gravity waves, 5. Internal gravity waves and tides, 6. Quasi-stationary planetary waves.

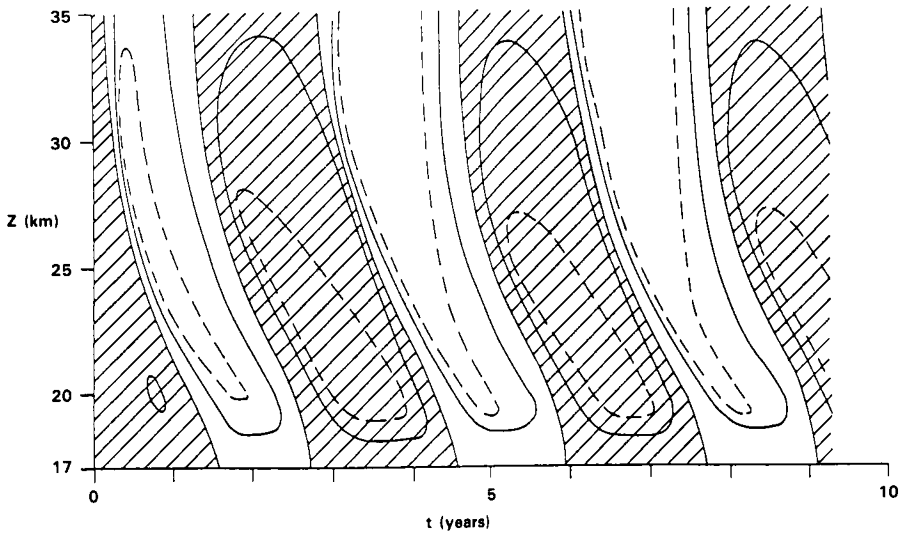


Figure 6-37. Theoretical evolution of mean zonal flow ū at the equator according to the Holton-Lindzen model. Solid contours are at intervals of 15 ms^{-1}, dashed contours represent ± 22.5 ms^{-1}. Westerlies are shaded. [After Plumb, 1977].

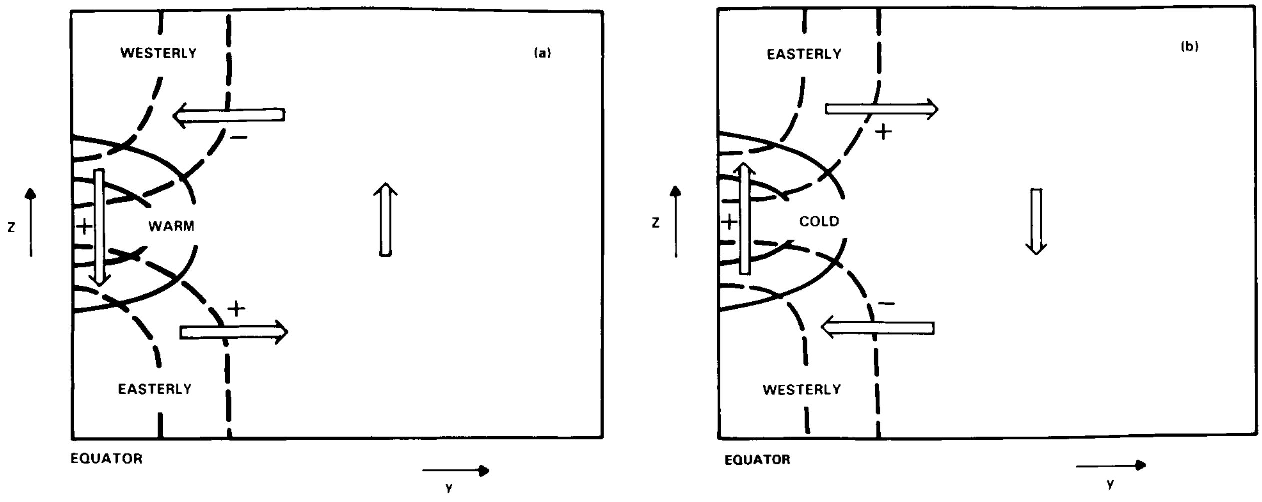


Figure 6-38. Schematic representation of the mean meridional circulation driven by an equatorial thermal anomaly, and the consequent acceleration of the mean zonal wind. Solid contours: Potential isotherms. Dashed contours: Isopleths of zonal velocity. ± Sign of zonal acceleration. (a) Warm anomaly, (b) Cold anomaly. [After Plumb and Bell, 1982].

schematically in Figure 6-38, show the meridional circulation to be an important determinant of the structure of the zonal wind QBO. Advection by this meridional circulation is also likely to contribute to trace species transport in low latitudes; Hasebe [1984] and Ling and London [1985] have discussed this in the context of the QBO in total ozone.

The driving of the SAO is in some ways similar to that of the QBO. The westerly SAO regime at the stratopause again appears to be driven by Kelvin waves, as suggested by Holton [1975], Hirota [1978] and Dunkerton [1979]; the fast Kelvin waves described in Section 6.1.5 propagate to these levels where they are observed to be of sufficiently large amplitude to account for the observed westerly accelerations [Coy and Hitchman, 1984]. Unlike the QBO, however, the easterly phase of the stratopause SAO appears to be driven not by equatorial waves but by the effects of the seasonal meridional circulation [Holton and Wehrbein, 1980b; Mahlman and Sinclair, 1980; Takahashi, 1984] or planetary waves propagating from the winter hemisphere [Hopkins, 1975; Dunkerton, 1979]. Indeed, on the basis of a successful simulation of the stratopause SAO in the GFDL "SKYHI" model, Mahlman and Umscheid [1984] concluded that both effects are important in producing the observed structure. Note that the periodicity of the SAO is thus externally imposed by the semiannual variability of mean angular momentum advection and planetary wave circulations in the tropics, whereas that of the QBO is internal to the dynamics of the wave mean-flow interaction [Plumb, 1977].

The mesopause SAO has been explained by Dunkerton [1982b] as a secondary phenomenon, produced as a "shadow" of the stratopause oscillation. He suggested that selective absorption of vertically-propagating internal gravity waves through the stratopause region so modulates the spectrum of waves propagating to higher levels as to force an oscillation of opposite sign to that at the stratopause. As yet there is no direct evidence to support this prediction.

6.3 MIDDLE ATMOSPHERE GENERAL CIRCULATION MODELS

6.3.1. Introduction

Quantitative assessment of our understanding of the middle atmosphere circulation relies on numerical models of varying degrees of complexity. These range from highly simplified models such as radiative-convective models (which neglect dynamics) through so-called ''mechanistic'' dynamical models (which neglect radiative effects, or represent them in a very crude fashion) to general circulation models (GCMs) which are defined here, following Mahlman and Umscheid [1984], as models which aim to include all major physical processes in a self-consistent manner. However, at present no models fully meet this criterion. While most GCMs use similar radiation schemes, for example, the manner in which momentum transfer by small-scale motions is parameterized ranges from very crude forms such as Rayleigh friction to more sophisticated schemes which are nonetheless highly simplified representations of a very complex process. Mahlman and Umscheid [1984] have pointed out that the important differences between GCMs and ''mechanistic'' models of the middle atmosphere relate to the inclusion of realistic radiative transfer (i.e., the radiative transfer should agree well with modern line-by-line calculations), a self-determined troposphere (including moist convection effects) and, perhaps, a ''sufficient'' model spatial resolution. Following the discussion of Section 6.2, it is clear that one of the major scientific challenges is the explanation of the maintenance of the observed large departures from radiative equilibrium in the mesosphere and in the winter stratosphere. It is therefore of particular importance that radiative processes which, in the absence of dynamical effects, would drive the system back to radiative equilibrium be represented as accurately as possible.

Given the fact that GCMs are typically much more expensive to run than are ''mechanistic'' models, it is important to note for which types of middle atmosphere scientific problems mechanistic models are appropriate and for which classes of problems GCMs are appropriate. Mechanistic models can be used to test specific physical hypotheses. For example, Matsuno's [1971] classic paper used such a model to illustrate the role of stationary planetary waves in giving rise to sudden stratospheric warmings. Such mechanistic studies focus on limited aspects of a problem and leave other aspects unaddressed. For instance, in Matsuno's model, the question of what gives rise to the tropospheric planetary wave activity or the basic state zonal wind in the first place is not addressed. However, the role of radiation in preventing such warmings from taking place in the real atmosphere is severely underestimated in such models. GCM experiments can be used to address these points.

One difficulty with the use of GCMs is associated with their large amount of output. The volume that emerges from a GCM integration is comparable to that of an atmospheric data set (and can even be greater depending on the frequency at which it is retained). This makes diagnostic studies of GCM behavior a formidable task. Moreover, the multiplicity of physical effects in a GCM, compared with a mechanistic model, makes GCM results sometimes difficult to interpret. Given these circumstances, both GCMs and mechanistic models need to be used cooperatively in middle atmosphere studies.

GCMs are very useful in studying the coupling of middle atmosphere dynamics with radiative and chemical processes. Examples are studies of the coupling of middle atmosphere dynamics with radiation by Ramanathan *et al.* [1983] and of the coupling of dynamics with chemical processes by Mahlman *et al.* [1980].

Tropospheric GCMs have also been used to examine the sensitivity of tropospheric climate and/or weather forecasting to inclusion of the middle atmosphere. Simmons and Struefing [1983] have looked at the sensitivity of tropospheric weather forecasts to extending the top of the European Centre for Medium-Range Weather Forecasting model from 50 to 10 mb and using a hybrid vertical coordinate system instead of sigma coordinates. Mechoso *et al.* [1982] used the UCLA GCM to study the sensitivity of numerical forecasts to moving the top level of the model from the lower stratosphere to the stratopause.

Middle atmosphere GCMs can also be used to understand the limitations of more simplified models or to develop more proper parameterizations for such models. For instance, Mahlman [1975] and Tuck [1979] have used tracer transport experiments with GCMs to study some of the shortcomings of simplified transport formulations commonly used in one- and two-dimensional photochemical models. Plumb and Mahlman [1985] have used tracer transport experiments with GCMs to establish the formulation of transport for two-dimensional photochemical models on a firmer physical basis.

GCM results can also be used as proxy for atmospheric data to examine the representativeness of present observational networks as well as to look at the impact of proposed future systems. Moxim and Mahlman [1980] have used the results of GCM transport studies to look at the representativeness of globally averaged ozone amounts that are inferred from the ground-based ozone network.

Middle atmosphere GCMs are just beginning to be used in the analysis of satellite data. Such forecast-analysis methods are already used extensively for tropospheric studies [e.g., McPherson *et al.*, 1979]. In this system, data are analyzed, a forecast is run, and the results of the forecast are used to help in the analysis at a later time. The cycle is then repeated. This type of analysis yields a complete data set, constrained by all available observations, that is dynamically and energetically consistent with the GCM formulation. Stratospheric forecast-analysis methods are now beginning to be applied to stratospheric satellite data by the British Meteorological Office and by groups in the United States.

6.3.2. Current Status of Middle Atmosphere GCMs

GCMs have modeled several aspects of observed middle atmosphere structure. In the following, we will look into some of the contributions of GCMs to our understanding of middle atmospheric behavior. We will also look at some model shortcomings. In this discussion, we shall confine attention to model studies for which the upper model boundary is at the stratopause or above.

a) Climatology

Several recent middle atmosphere GCMs have appeared recently in the published literature. These include the model results presented by Schlesinger and Mintz [1979], Hunt [1981], O'Neill *et al.* [1982], Mahlman and Umscheid [1984], and Rind *et al.* [1985]. A typical deficiency that appears in most middle atmosphere GCM results is that of excessive winter westerly and summer easterly zonal winds together with excessively cold winter polar temperatures as required by thermal wind balance. Figure 6-39 gives examples of these results from the works of Schlesinger and Mintz [1979], Hunt [1981], Mahlman and Umscheid [1984], and Rind *et al.* [1985]. All four of these results are for the month of January although Schlesinger and Mintz' and Hunt's result are for perpetual January integrations while the Mahlman and Umscheid and Rind *et al.* results are from annual cycle integrations. The four results shown in Figure 6-39 all use different formulations for radiative transfer as well as different parameterizations for the diffusion and drag resulting from subgrid scale processes. Nevertheless, certain features are found to be common to all of these results. All four simulations give middle atmosphere winter hemisphere westerlies and summer hemisphere easterlies as are observed. However, all have winter westerlies too strong compared to

ORIGINAL PAGE IS
OF POOR QUALITY

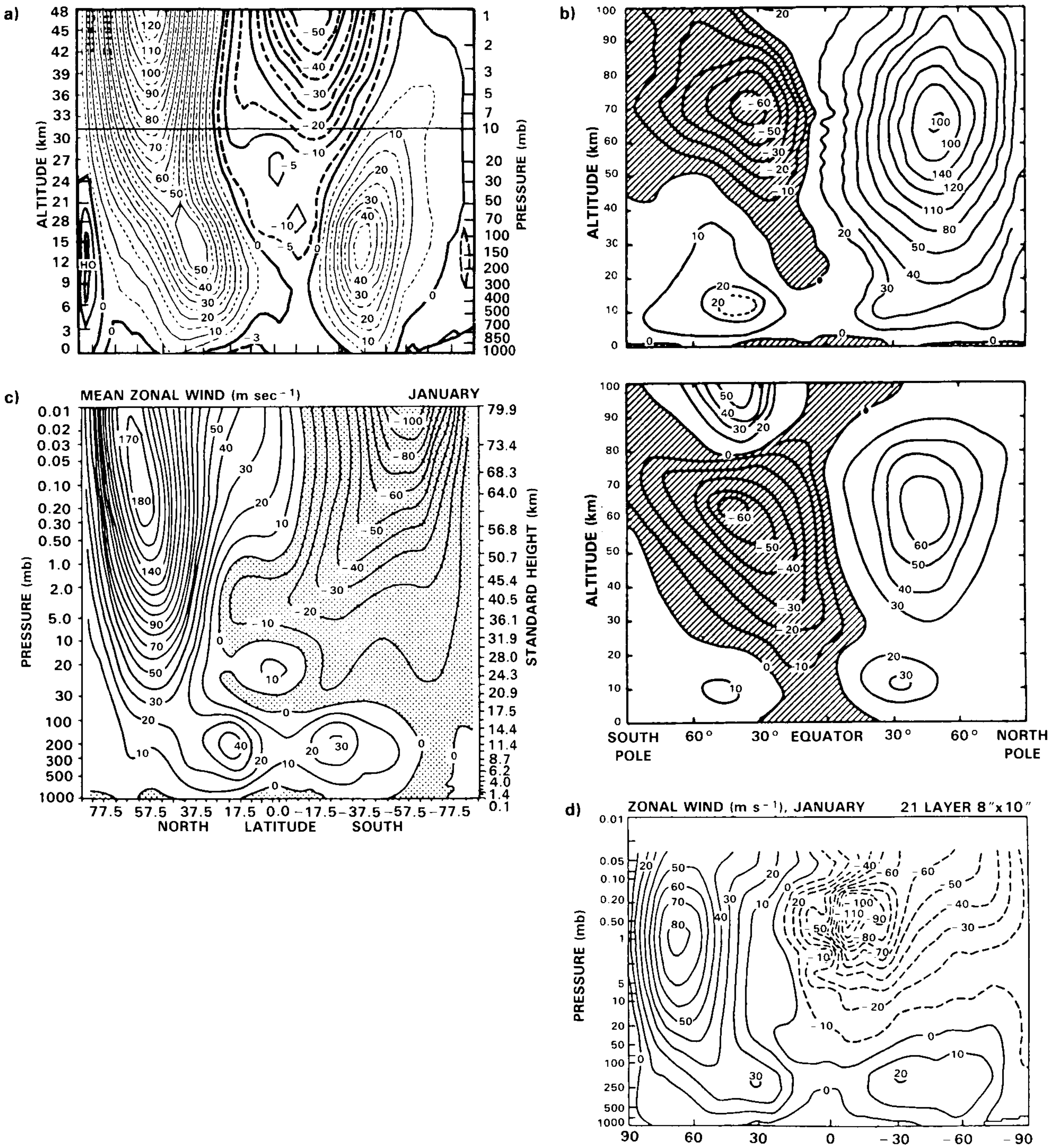


Figure 6-39. (a) Mean zonal wind in m/s as modeled with perpetual January insolation by Schlesinger and Mintz [1979]

(b) Mean zonal wind in m/s as modeled with perpetual January insolation by Hunt [1981] in top panel and "observed" mean zonal wind for Northern Hemisphere January and July conditions from Newell [1968] in bottom panel

(c) Mean zonal wind for January from the model of Mahlman and Umscheid [1984]

(d) Mean zonal wind for January from the model of Rind *et al.* [1985].

observations. One way to see this is to compare the maximum westerly wind values at 10 mb (about 30 km) in the three simulation results with observations. In the results of Schlesinger and Mintz, Hunt, and Mahlman and Umscheid, the maximum westerly wind at 10 mb is about 70 ms^{-1}. Rind *et al.*'s maximum westerly wind at 10 mb is about 60 ms^{-1}. For comparison, we see that the "observations" show maximum westerly wind values at 10 mb of about 20 ms^{-1}. (More recent analysis of a four year data set by Geller *et al.* [1984] indicates maximum mean January winds at 10 mb in the Nothern Hemisphere ranging from about 30 to 50 ms^{-1}).

We see then that all middle atmospheric GCMs (given our definition that GCMs must include proper radiative transfer calculations) give excessive winter westerlies. Mahlman and Umscheid [1984] discussed extensively this aspect of the behavior of the GFDL SKYHI model and attributed this problem to a combination of underestimation of the upward flux of planetary waves out of the troposphere and the neglect of the effects of small-scale gravity waves in their model.

b) Perturbation Studies

Middle atmospheric GCMs have also been used to test for their response to imposed perturbations. Fels *et al.* [1980] investigated the effects of doubling CO_2 and halving O_3 in a low resolution model extending up to about 80 km with annual average insolation and boundary conditions. Figure 6-40 shows their modeled zonally averaged temperatures for the control (unperturbed) case and for the doubled CO_2 and halved O_3 cases. They found for the doubled CO_2 case that the resulting middle atmosphere cooling was quite independent of latitude; little change was found therefore in the mean zonal wind. The halved O_3 case gave much more latitudinal structure in the cooling and consequently a much greater effect on the mean zonal flow, mostly above 30 km. No significant changes in tropospheric planetary wave structure were found in either of the perturbation experiments. This last result is consistent with that found in studies of the effects of middle atmosphere changes on tropospheric planetary waves using simpler mechanistic models [Schoeberl and Strobel, 1978a; and Geller and Alpert, 1980]. One result that did emerge from the Fels *et al.* study that would not have been easy to obtain with a mechanistic model was the sensitivity of the tropical tropopause temperature to the imposed O_3 changes. The strong cooling found at the tropical tropopause in the halved O_3 experiment should result in much less stratospheric water vapor by increasing the freezing out of water in the rising branch of the Hadley circulation.

c) Comparison with Simpler Models

Another application of middle atmosphere GCMs is in comparing the results of full GCMs with more simplified models for the purposes of seeing the limitations of the simpler models and improving their parameterizations. Two examples of this type of application of middle atmospheric GCMs are the works of Fels *et al.* [1980] and Plumb and Mahlman [1985]. Fels *et al.* [1980] used the GFDL SKYHI model to explore the applicability of the simpler Radiative-Convective-Equilibrium (RCE) and Fixed-Dynamical-Heating (FDH) models to the doubled CO_2 and halved O_3 experiments that were discussed in the previous section.

The RCE model is one in which the solar heating is taken to balance the long wave cooling at each point in the model except where the computed temperature lapse rate exceeds some critical value, in which case the lapse rate is fixed at this value; the surface temperature is usually calculated self-consistently. Fels *et al.* fixed the surface temperatures, however, so that the response to the imposed perturbation is purely radiative above the convective zone and purely dynamical below.

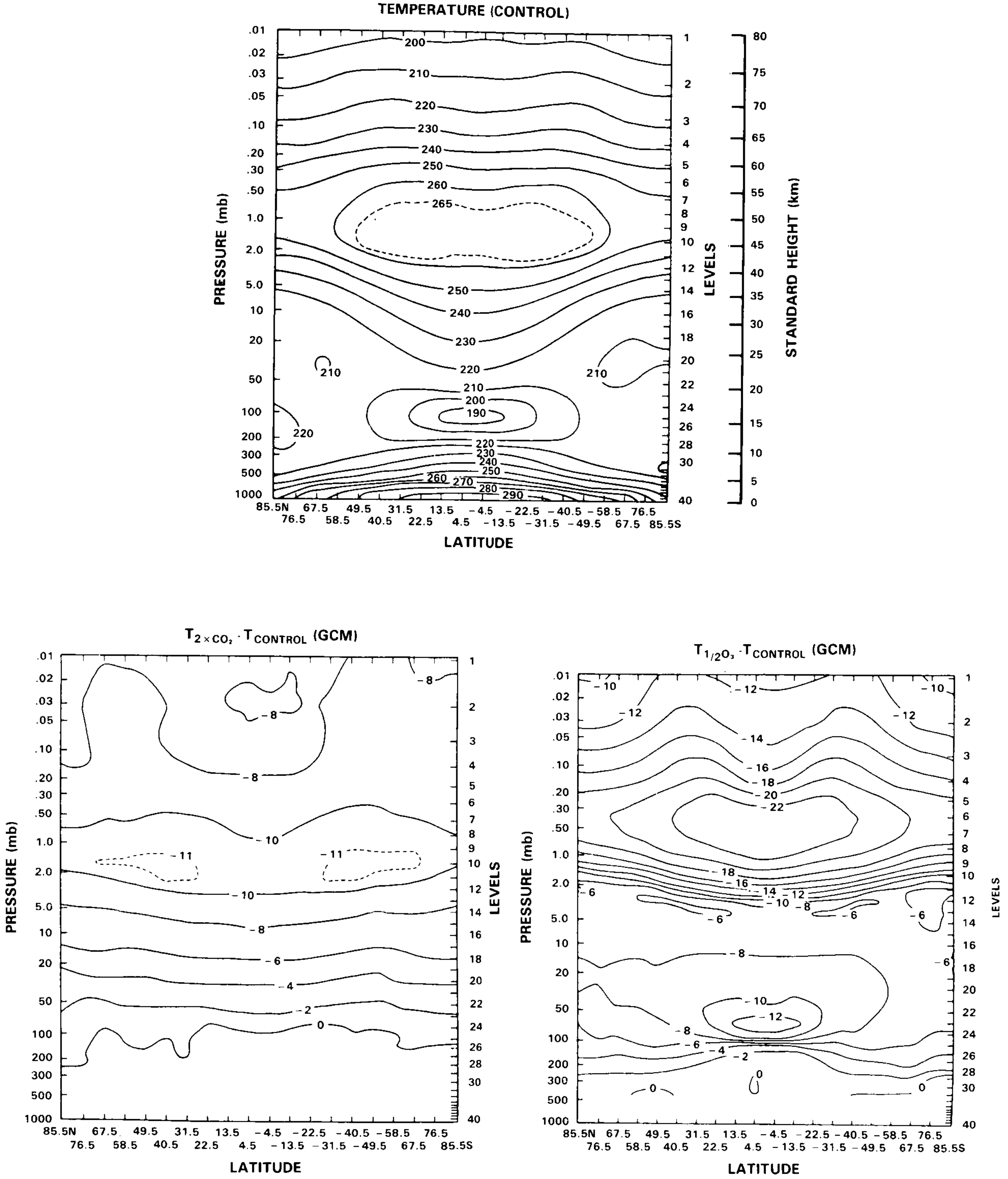


Figure 6-40. (Top) Modeled zonally averaged temperature in degress Kelvin; (lower left) difference between modeled zonally averaged temperatures and those with uniformly doubled CO_2 concentrations; (lower right) difference between modeled zonally averaged temperatures and those with uniformly halved O_3 concentrations. All are general circulation model results from Fels *et al.* [1980].

The FDH model is one in which the dynamical heating is taken to remain unchanged when the distribution of radiatively active constituents is perturbed. Thus, given Fels *et al.* [1980] GCM results for annual averaged insolation, the equilibrium solution is one where the sum of the solar heating, long wave cooling, and dynamical heating rates are everywhere known and the distribution of dynamical heating rates is easily determined. In performing the perturbation experiments, the new FDH temperature distribution is one for which the new local net radiative heating rate plus the unchanged dynamical heating rate is everywhere zero.

Figure 6-41 shows the RCE and FDH modeled temperature differences for the halved O_3 perturbation experiment. This is to be compared with the full GCM results for this case that were shown in Figure 6-40. From the comparison, we see that above 35 km there is qualitative similarity among all three model results with maximum cooling of 22-25 K near the tropical stratopause; however, larger horizontal gradients in the amount of cooling are seen in the FDH results. Below about 25 km, there are substantial differences among the three results that can be attributed, in part, to the fixed lapse rate constraint in the troposphere and to the different "control" temperature distribution of the RCE model [see Fels *et al.*, 1980]. There is a qualitative difference between the FDH and GCM results in the tropics between about 55 and 75 km where the dynamical cooling has apparently changed in the GCM case. There are also significant differences in the cooling results just above the tropical tropopause among the GCM, RCE, and FDH results. Fels *et al.* point out that this region is particularly sensitive to very small changes in the heating or cooling rates. Quite similar results are obtained in all of the three models (GCM, RCE, and FDH) for the doubled CO_2 case (not shown here). The only very significant difference is in the shape of the perturbation response in the lower stratosphere in the RCE model compared with those of the GCM and FDH models. This is due to the fixed lapse rate constraint in the RCE case. The fact that the FDH model does so well in the doubled CO_2 case is consistent with the rather flat response in the GCM perturbation experiment with latitude. Such a flat response should not produce greatly altered dynamics.

Thus, the Fels *et al.* [1980] investigation showed that for perturbation experiments involving changes of concentrations of long wave radiatively active constituents such as CO_2, which give a flat distribution of temperature change with latitude, the FDH, and, in fact, the RCE models give fairly good simulations of temperature change. In cases where the altered constituent is a strong solar radiation absorber such as ozone, the GCM gives a temperature response with more latitudinal structure and more altered dynamics. Even in this case, however, the FDH model does well except in the tropical lower stratosphere and mesosphere. Investigations of this type show the utility of middle atmosphere GCMs in seeing the limitations of simpler models.

Plumb and Mahlman [1986] have used the GFDL tracer model (with its top at 10 mb) to investigate better ways in which to parameterize transport in two-dimensional photochemical models. They have derived the two-dimensional transport tensor by performing two transport experiments that allow them to solve for the four components of the "diffusion" tensor. They do this by calculating the zonally-averaged meridional and vertical tracer fluxes in the GCM. This, together with the calculated meridional and vertical gradients of the computed zonally averaged distributions allows solution for the transport tensor. Plumb and Mahlman have used this transport tensor in a two-dimensional model to show that they can mimic reasonably well the zonal average of the three-dimensional transport of the GCM. Thus, they demonstrate an internally consistent manner in which a transport formulation can be developed for two-dimensional models to give equivalent transport to a GCM in a zonally averaged sense. Plumb and Mahlman's GCM study underscores the importance of including consistent advection and diffusion in two-dimensional transport models (see Section 6.5).

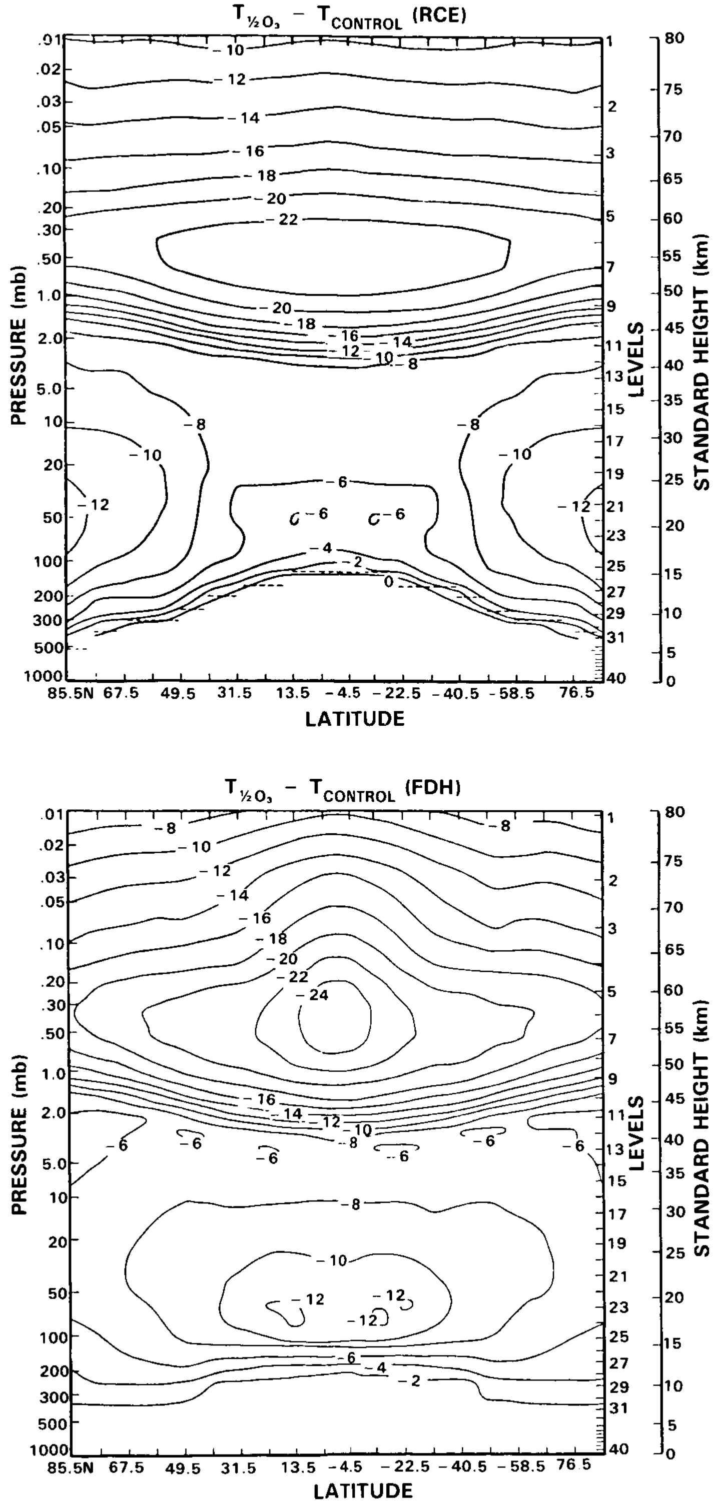


Figure 6-41. Difference in zonally averaged temperatures (in degrees Kelvin) between halved O_3 case and control case using a radiative-convective-equilibrium (RCE) model (top) and fixed-dynamical-heating (FDH) model (bottom). From Fels *et al.* [1980].

d) Transport and Photochemistry Studies

Middle atmosphere GCMs have been very useful for studies of transport and photochemistry. Two very different types of studies have been accomplished by the use of marked parcels and by tracer transport formulations, some of which have involved photochemistry while others have not. Examples of middle atmosphere transport studies using marked parcels are those of Kida [1977, 1983a,b], and Hsu [1980]. Examples of transport studies without photochemistry include Mahlman and Moxim [1978] while those with some degree of photochemistry include Cunnold *et al.* [1975, 1980], Mahlman *et al.* [1980], Levy *et al.* [1979], and Golombek [1982].

As examples of these two types of investigations, we will briefly discuss the work of Kida [1983a,b] and Mahlman *et al.* [1980]. Kida used a 12-level hemispheric GCM extending from the ground to 1 mb. Its horizontal grid was 3 degrees of longitude by 2.5 degrees of latitude. It had no topography but did include a parameterization of thermal forcing of planetary waves. Kida examined the very long term motion of air parcels into and out of the stratosphere by performing trajectory analyses on a large number of marked air parcels. In this work, he defined the age of a stratospheric air parcel as the length of time elapsed since the parcel first entered the stratosphere. Figure 6-42 shows parcel age spectra for 5 degree latitude by 1 km altitude domains for three separate latitude bands: 5-10 degrees (tropics), 45-50 degrees

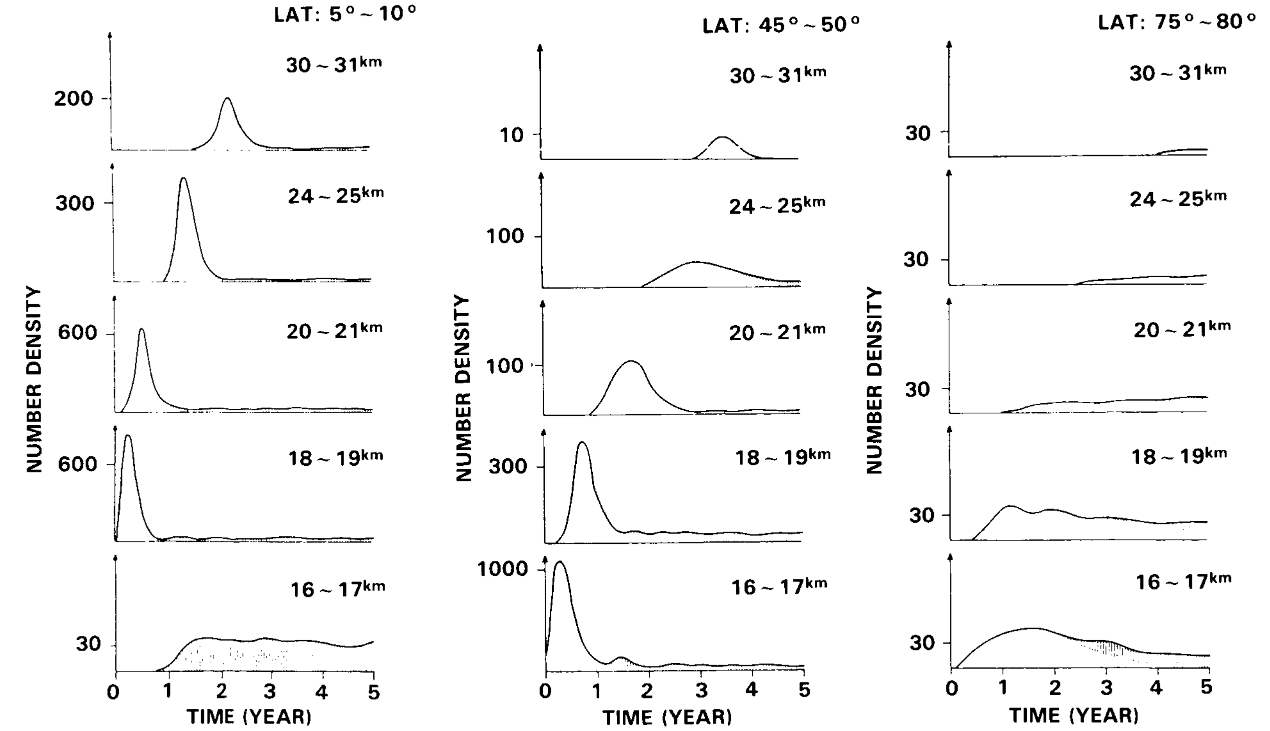


Figure 6-42. The age spectrum of air parcels whose initial location was just below the tropical tropopause for selected domains in the lower stratosphere. Left – 5-10 degrees; middle – 45-50 degrees; right – 75-80 degrees. Each domain covers 5 degrees in latitude and 1 km in altitude. Shaded portions are due to parcels that have once entered the troposphere and, after a long time, re-entered the stratosphere or remained in the troposphere. From Kida [1983b].

(mid-latitudes), and 75-80 degrees (polar latitudes). These are given for five altitude ranges. At the start of this experiment, the parcels are all located just beneath the tropical tropopause. In the tropics, the tropopause altitude is about 17 km, so we may interpret the age spectra as showing strongly peaked distributions above the tropopause with the peak in age spectra occurring at later times at increasing altitude consistent with the speed of the rising motion in the Hadley circulation of ~ 5 km/year. Below the tropopause, the distribution is flat indicating that after about a year the marked parcels start reentering the troposphere and build up to a steady state number density. At middle latitudes, one sees less sharp stratospheric distribution peaks than was the case for the tropics. The broader peak in the age spectra at longer times with increasing altitude indicates that it has taken longer for stratospheric air parcels to reach these higher altitudes and that the trajectories taken are more diverse. Note that all five altitudes are in the stratosphere at middle latitudes. Finally, the high latitudes have very flat distributions indicating that it takes a long time for ''new'' stratospheric air parcels to reach the polar stratosphere, and that the trajectories followed by these polar stratospheric air parcels were very diverse.

Another very different use of a middle atmospheric GCM in investigating transport and photochemistry is that of Mahlman *et al.* [1980] who used the GFDL 11-level tracer model [see Manabe and Mahlman, 1976 and Mahlman and Moxim, 1978] for two idealized ozone experiments. The first of these experiments, the Stratified Tracer Experiment, specified instant relaxation of the ozone concentration at the top level (10 mb) to 7.5 ppmv. Ozone was then treated as inert tracer at all levels below this top level until it was removed in the lower troposphere. In the second experiment, the Simple Ozone Experiment, a simplified ozone photochemistry is used at the top model level, and again ozone is taken to be inert at lower levels and is removed in the lower troposphere. Mahlman *et al.* [1980] found that both experiments gave remarkably similar results. For example, Figure 6-43 shows the zonal-mean ozone mixing ratio [ppmv] from the fourth year of both experiments. These results suggest that the details of middle stratosphere ozone chemistry exert very little influence on the distribution of ozone in the lower stratosphere compared with the influence of transport processes.

e) Modeling of Stratospheric Warmings

Two types of GCM studies of stratospheric warmings have been undertaken. These are the analysis of stratospheric warmings that have spontaneously arisen in the course of climatological runs with GCMs and GCM forecasting of observed stratospheric warming events. The first spontaneous appearance of a stratospheric warming in a GCM was reported by Newson [1974] using an early version of the British Meteorological Office stratospheric model. Similar results were found with the NASA/Langley quasi-geostrophic stratospheric model by Haggard and Grose [1981]. Neither of these models meets our present definition of a GCM since the model used by Newson used Newtonian cooling in place of an infrared radiative transfer treatment, and Haggard and Grose's model was quasi-geostrophic and contained a parameterized troposphere. Nonetheless, these were two of the earliest reports of stratospheric warming events arising spontaneously in stratospheric circulation models. Mahlman and Umscheid [1984] have recently reported the spontaneous appearance of a sudden warming-type event in the GFDL SKYHI GCM. So far, middle atmosphere GCMs have not produced simulations of stratospheric warming events as intense as those observed in the actual atmosphere.

The pioneering attempt at forecasting stratospheric warming events with a GCM was by Miyakoda *et al.* [1970]. This effort showed some success but failed to produce a true warming event. More recent attempts by Simmons and Struefing [1983], Mechoso *et al.* [1985], and Geller *et al.* [1985] have shown greater success.

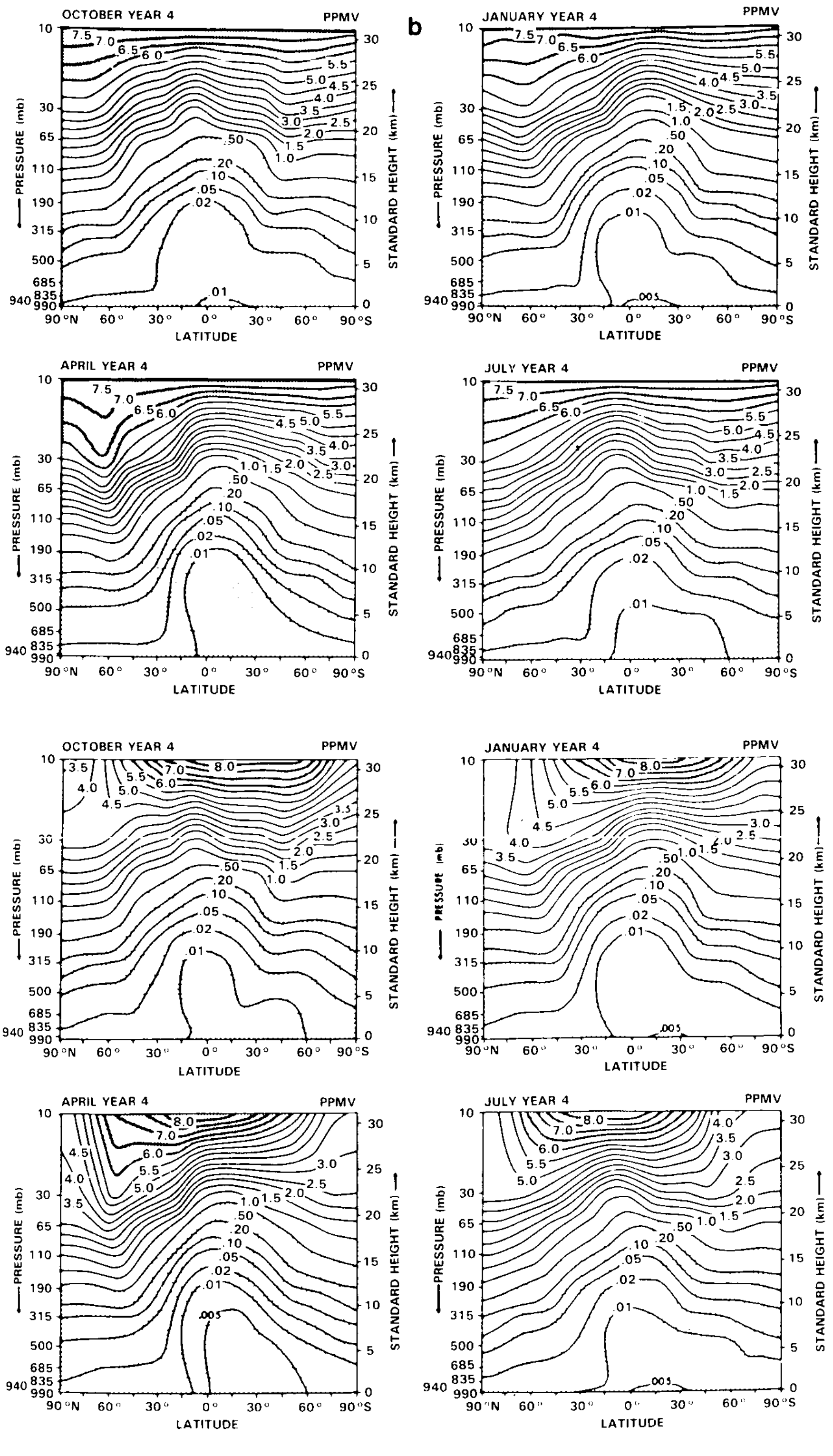


Figure 6-43. Zonal mean mixing ratio (ppmv) for selected months from the fourth year of the Stratified Tracer Experiment (top four panels) and the Simple Ozone Experiment (bottom four panels) of Mahlman *et al.* [1980].

In summary then, troposphere-middle atmosphere GCMs have had a number of successes in simulating some middle atmospheric phenomena. They reproduce many of the observed features of the middle atmosphere [see NASA, 1979, pages 79-91]. Some of the successfully simulated features are the following: the reversed meridional temperature gradient in the lower stratosphere [Smagorinsky *et al.*, 1965]; scale filtering of disturbances with increasing altitude [Manabe and Hunt, 1968]; equatorial tropopause structure [Manabe and Mahlman, 1976]; midlatitude warm belt [Manabe and Mahlman, 1976]; interhemispheric asymmetries [Manabe and Mahlman, 1976]; stratospheric tracer structure [Hunt and Manabe, 1968]; seasonal variation of lower stratospheric temperature [Manabe and Mahlman, 1976]; cancellation between mean cell and eddies [Smagorinsky *et al.*, 1965]; gross features of stratospheric-tropospheric mass exchange [Mahlman and Moxim, 1978]; summertime easterly flow [Manabe and Mahlman, 1976]; identification of Kelvin and mixed Rossby-gravity wave modes in the lower stratosphere [Hayashi, 1974; Tsay, 1974]; phase relationships between ozone and pressure perturbations [Schlesinger and Mintz, 1979]; simulation of the semi-annual oscillation [Mahlman and Sinclair, 1980]; identification of upper stratospheric and mesospheric Kelvin waves [Hayashi *et al.*, 1984]; identification of the effects of gravity waves on the mean zonal flow and planetary waves [Miyahara, *et al.*, 1985]; spontaneous stratospheric warmings [Newson, 1974]; and successful forecasting of stratospheric warmings up to the lower stratosphere [Simmons and Struefing, 1983]. References given above are for the first apparent publication reporting each achievement.

6.3.3. GCM Deficiencies

a) Cold Pole

As discussed in the previous section, all middle atmosphere GCMs (using our definition that they must contain state-of-the-art radiation calculations and a self-determined troposphere) suffer from producing a winter polar night stratosphere that is too cold. Along with this, by the thermal wind relation, the winter westerlies are too strong. This problem typically exists in middle atmosphere GCMs from the tropopause to the mesopause. There is some uncertainty about the cause for this, although it appears that the causes in the lower and upper stratosphere may differ. As discussed in Section 6.2, the climatological state of the winter stratosphere arises from a balance between dynamical effects, driving the system away from radiative equilibrium, and the restoring influence of radiation. In the lower stratosphere, where both effects are weak, the balance is a very sensitive one and the results of Ramanathan *et al.* [1983] suggest that an accurate representation of radiation is crucial to a proper simulation of this region. This conclusion is confirmed by the results of Mahlman and Umscheid [1984]. In the middle and upper stratosphere, however, the balance is less sensitive to small changes in radiation and it seems most likely that the models are in some way underestimating dynamical heat transport. This deficiency may arise from deficient planetary-wave transports or from the failure of current models to represent small-scale motions such as internal gravity waves. The importance of gravity wave transports in maintaining the climatological balance of the stratosphere is currently not well understood.

b) Interannual Variability

Typically, stratospheric GCMs exhibit considerably less interannual variability than does the actual atmosphere [e.g., Geller *et al.*, 1984]. This is almost certainly a planetary wave, mean flow interaction problem. Mahlman [personal communication] has noted that the GFDL SKYHI model at 5 ° resolution shows deficient tropospheric disturbances on all scales (both standing and transient); the representation is improved at 1 ° resolution. Deficient planetary waves will lead to underestimation of the planetary wave effects on the mean flow.

c) Quasi-biennial Oscillation

To date, no GCM has successfully simulated the quasi-biennial oscillation of the equatorial stratosphere. This is to be compared with the relatively successful simulation of the equatorial semi-annual oscillation that was discussed above. The reason for this is probably (see Section 6.2.7) that the equatorial wave modes responsible for the quasi-biennial oscillation have shorter vertical wavelengths than those responsible for the semi-annual oscillation, and the vertical grid spacing in current GCMs cannot resolve these shorter wavelengths. It should be pointed out also that most middle atmosphere GCMs are probably too dissipative to support a quasi-biennial oscillation [see Plumb, 1984].

d) Tropical Tropopause and Stratospheric Water Vapor

Several middle atmosphere GCMs with relatively high vertical resolution have simulated a tropical tropopause that is too cold by ~ 3-10 K. Lower vertical resolution models with incomplete radiative transfer schemes have sometimes been too warm. Since stratospheric water vapor is very drastically sensitive to the equatorial tropopause temperature, models that attempt to calculate stratospheric water vapor get too high or too low values depending on their bias in the model tropical tropopause temperatures. Also, numerical simulation of water vapor advection is very difficult due to its extreme vertical gradients in the upper troposphere.

e) Inclusion of Chemistry

It is an expensive proposition to include relatively complete chemistry schemes in middle atmosphere GCMs. This is due to several factors. One is the number of extra prognostic equations that must be included. A GCM without chemistry calculates four prognostic variables: the two horizontal wind components, temperature, and water vapor. Without treating chemistry by chemical families, one has to calculate about fifteen constituent prognostic equations for transported species. Using chemical families might only imply doubling to tripling the number of prognostic variables. Thus, the number of prognostic equations is greater by about a factor between two and five when chemistry is included.

Furthermore, several of the previously discussed deficiences of middle atmosphere GCMs in calculating the proper dynamical structure of the middle atmosphere will lead to deficiences in calculating proper constituent distributions. For instance, if the GCM-calculated tropical tropopause temperature is too cold, the dryness of the air entering the stratosphere will be increased. This will reduce production of the OH radical and thence, ultimately, affect the modeled ozone chemistry. Also, if the "cold winter pole bias" is due to inadequate dynamical heating then the diabatic circulation, and by inference, the transport circulation (see Section 6.5), and the nonadvective eddy transport will be too weak.

Thus we see that some of the problems in calculating proper middle atmosphere structure with GCMs are expected to lead to difficulties in simulating the interactive chemistry of the stratosphere. In particular, if the GCM has difficulty in simulating the NO_x and HO_x distributions due to dynamical and related deficiencies, then problems must be expected in the representation of ozone, even if by some miracle the model chemistry were to be perfect.

6.3.4. Future Directions

a) Model formulations

We have touched on a variety of GCM applications for both theoretical and observational problems. Differences in model formulations may lead to discrepancies between results obtained from different models.

It is important to understand how the characteristics of a GCM influence its behavior and, in particular, to distinguish those aspects of GCM results which are robust from those which are influenced by details of the model formulation.

A number of features distinguish one GCM from another. These include the numerical scheme (gridpoint or spectral), vertical and horizontal resolution, the radiation scheme, parameterization of sub-grid-scale mixing, convection and gravity wave transports and, for transport applications, the means by which conservation of tracer amount is ensured and occurrence of negative mixing ratios is avoided. The level of sophistication of the parameterization schemes used in GCMs varies greatly and their impact on model performance is not well understood.

b) Comparison of GCMs with Observations

Given the problems that middle atmosphere GCMs have had in reproducing some of the most basic features of the observed middle atmosphere (i.e., zonally averaged temperatures and winds) and, until recently, the scarcity of analyses of the middle atmosphere general circulation, middle atmosphere GCM comparison with data has largely been on the basis of morphology comparisons. More sophisticated comparisons are now possible. For instance, GCM diabatic circulations can be compared with that derived from atmospheric data. There are also methods that can be used to compare the effective diffusion in GCMs and the observed atmosphere. In a two-dimensional sense, these two quantities (the diabatic circulation and the effective diffusion) are the two quantities that should be compared if a GCM is to properly transport species [see Mahlman *et al.* 1984, for example].

c) Horizontal and Vertical Resolution

There has been a persistent belief in middle atmosphere modeling that lesser horizontal resolution is required in simulating the middle atmosphere than for the troposphere. Experience is showing otherwise. Both the GFDL experience and the British Meteorological Office experience have indicated the necessity for using higher horizontal resolution than was previously thought to be needed in middle atmosphere GCMs. Mahlman's research group at GFDL has found that their GCM dynamics become progressively more active as they go to finer and finer horizontal resolutions. This leads to progressively better simulations as horizontal resolutions are increased. O'Neill's group at the U.K. Meteorological Office has found in stratospheric warming simulations that higher horizontal resolutions are required during circumstances of more active dynamics.

d) Gridpoint versus Spectral Models

A number of modeling deficiencies have been attributed to fundamental differences between finite difference and spectral transform models. Each modeling architecture has its advantages. The spectral transform method computes horizontal derivatives exactly whereas the finite difference methods used in gridpoint models have their associated errors. (Of course, spectral models have their own truncation errors, entering the calculation in a different way). Spectral transform models are set up for efficient implementation of semi-implicit time differencing which make them more efficient to run. On the other hand, the semi-implicit time differencing makes the model's gravity waves propagate more slowly. This may imply that gravity wave effects must be completely parameterized rather than treated explicitly in these models. A true comparison between middle atmosphere spectral and gridpoint GCMs at the present time is impeded by the different manner in which these two modeling communities have treated dissipative processes. One area that needs to be explored in gridpoint models, however, is the effect of polar filtering on GCM middle atmosphere dynamics. This, potentially, could be an area where spectral models would have a distinct advantage over gridpoint models.

e) Gravity Wave and Turbulence Effects

Perhaps the most important area in which progress is needed in modeling the middle atmosphere is in understanding the proper methods to include the effects of gravity waves and turbulence on the large-scale flow. The pioneering works of Lindzen [1981] and Matsuno [1982] have shown that the effects of gravity waves must be included for proper simulation of the mesosphere. Hunt [1985] has included Lindzen's [1981] parameterization for gravity wave breaking in a GCM that extends upward to 100 km. He has also explicitly included the effects of the diurnal tide in this model. He finds that inclusion of these effects improves his model results significantly. Kida [1985] has constructed a simplified GCM-type model which was meant to study the effects of explicitly simulating gravity wave effects on the middle atmosphere. This model extended from 15 km to 135 km and from the South Pole to the North Pole but extends only ten degrees in longitude (with periodic boundary conditions). His gridspacing was chosen to be able to simulate eddy motions with zonal length scales down to 100 km. He specified a random forcing of gravity waves at the model's lower boundary. These gravity waves will break higher up in the model by virtue of their exponential growth together with the GCM's treatment of convective adjustment. Kida's results show that explicitly modeling the effects of gravity waves in his model gives a simulation that, qualitatively, at least, agrees with middle atmosphere observations. Miyahara *et al.* [1985] have carried out a study of the role of gravity waves in the high-resolution GFDL SKYHI GCM and have also shown the very important role of the model's gravity waves in the mesosphere.

One of the most significant questions in middle atmospheric dynamics at the present time is whether or not gravity waves play an important role in the stratosphere. Tenenbaum [1982] has pointed out the nearly omnipresent problem of GCMs producing too little negative shear on the topside of the subtropical jet stream. Schoeberl [1985] has carried out an idealized study of the linear gravity wave spectrum that is produced by airflow over topography. He found that the superposition of several gravity waves produced regions of shear instability in the lower stratosphere, but that gravity wave instability in the mesosphere usually was the result of a single wave breaking. Thus, parameterization schemes for gravity waves in the stratosphere may have to be more complex than those currently used for the mesosphere.

Observational analyses of gravity waves and turbulence are needed to understand more about the sources of gravity waves, their climatology, and their effects on large-scale middle atmospheric flow.

e) Radiative Transfer

All of the previous discussion assumes that present day treatments of radiative transfer in the middle atmosphere are sufficient for inclusion into GCMs. If our ability to calculate heating rates in the middle atmosphere is found to be deficient, this would affect all of our perceptions about general circulation modeling of the middle atmosphere. Our understanding of these radiative processes has been discussed in detail in Chapter 7.

Middle atmosphere GCMs that extend sufficiently far upward (above 70 km) must include the effects of the breakdown of local thermodynamic equilibrium. That is to say collisions at this altitude become insufficiently frequent to fully populate the Boltzmann distribution so that the formulation for infrared transfer must be altered [see Dickinson, 1984].

6.3.5. Summary

In summary then, present day middle atmosphere GCMs have shown considerable success in modeling certain aspects of the observed circulation. They also have severe problems (e.g., the cold winter pole bias). These problems are sufficiently great that GCM transport studies, while probably being representative of actual atmospheric processes, cannot be taken to give quantitatively correct values for atmospheric transport. It is desirable that middle atmosphere GCMs be developed to the point where they can be taken to be quantitatively correct atmospheric surrogates since data quality may never be sufficient for such transport studies to be carried out without model intervention.

6.4 OBSERVATIONS OF TRANSPORT PROCESSES

6.4.1 Introduction

Observations which reveal information about the processes responsible for the transport of constituents in the middle atmosphere must necessarily provide the foundation upon which a sound theoretical description of constituent transport can be developed. Historically, it has been the process of seeking agreement between observation and theory that has led to advances in our understanding. Quite often, observations have dictated major revisions in existing theories. Significantly, the earliest example of the latter process with respect to constituents in the middle atmosphere relates to ozone. Measurement of total ozone column by Dobson *et al.* [1929] indicated a distribution that contradicted the photochemical theories developed by Chapman [1930] and subsequent investigators. It soon became apparent that poleward and downward transport of ozone from the photochemical source region in the high equatorial stratosphere was required for consistency with the observations. There followed some 40 years during which the interplay between observations and theory greatly increased our understanding of transport phenomena. Mahlman *et al.*[1984] provide an extensive discussion of this period.

In the 1970's the advent of satellite observations dramatically enhanced our knowledge of the atmospheric circulation, thermal structure, and constituent distributions by providing near-global measurements of temperature and species concentrations continuously over long periods. The impact of these measurements on our understanding of the structure and climatology has been discussed in Section 6.1; here we specifically address observations (both direct and indirect) of processes which determine the transport of dynamic and thermodynamic quantities (momentum, heat and potential vorticity) and of trace constituents.

6.4.2 Intercomparisons of Derived Quantities

Many recent studies of stratospheric transport processes and their impact on the structure and dynamics of the stratosphere rely on the determination of potential vorticity and eddy fluxes of potential vorticity (or, equivalently, the divergence of Eliassen-Palm fluxes; cf. Section 6.2) from analyses based on satellite data. Given the degree of differentiation in the vertical and horizontal required to derive potential vorticity from temperature retrievals, it is pertinent to question the accuracy of these determinations. Since errors in potential vorticity depend in a complex way on the magnitude and structure of the errors in temperature and geopotential analyses, the simplest avenue for the assessment of the reliability of potential vorticity calculations (or of the EP flux divergence) is via comparison of results from different data sources.

Grose [1984] reported excellent agreement for Ertel's potential vorticity (EPV) derived from LIMS data with that derived from SSU data reported in McIntyre and Palmer [1983, 1984]. A comparison between SSU and LIMS results for January 27, 1979, is presented in Figure 6-44. Additional comparisons for the entire period January and February 1979 exhibit equally good agreement. The differences between the two sets of results are largely due to the differences in resolution of the nadir viewing SSU [Pick and Brownscombe, 1981] and the limb viewing LIMS [Gille and Russell, 1984].

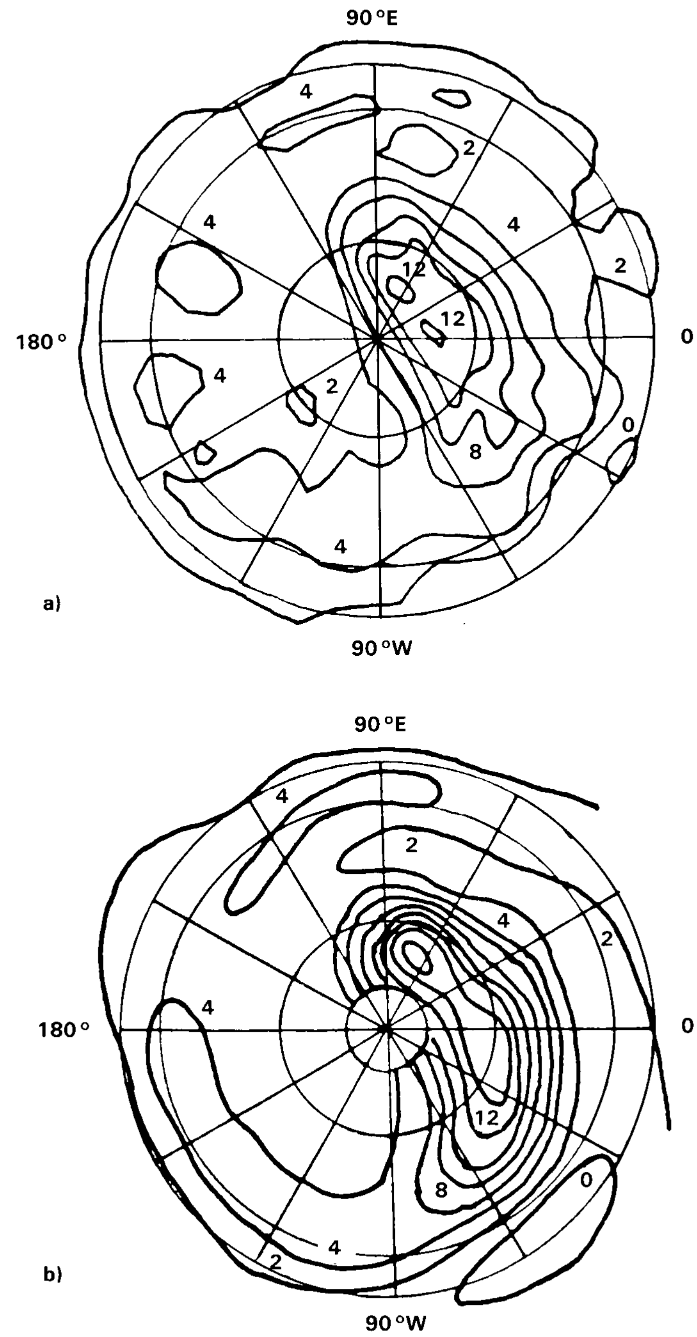


Figure 6-44. Ertel's potential vorticity (K m^{-1} s^{-1}) on the 850 K isentropic surface (average pressure about 10mb) for January 27, 1979. (a) SSU data [from McIntyre and Palmer, 1984], (b) LIMS data [from Grose, 1984].

Another intercomparison study for the Northern Hemisphere winter stratosphere using satellite and radiosonde/rocketsonde measurements was carried out by Miles and Chapman [1984]. This study utilized Nimbus 5 SCR data and conventional data from NMC and Berlin to derive zonal mean winds, time-mean wave structure and Eliassen-Palm cross sections for the period December 1973-February 1974. The results exhibited good qualitative agreement with the exception of the Eliassen-Palm flux divergence in the stratosphere. The disagreement between the EP flux divergences for the different sources could be traced to differences in the horizontal eddy momentum flux distributions. Although all three data sources were found to give somewhat different results, the SCR results are more compatible with those from Berlin data. Surprisingly, the NMC results are quite different from the Berlin results. One factor which could explain the NMC-Berlin differences is the incorporation of vector wind reports in the Berlin analysis although several other factors could be responsible. These results demonstrate that useful derived quantities can be obtained from satellite data, but emphasise the difficulty of accurate determination of quantities requiring several orders of differentiation.

6.4.3 Studies of Wave-Mean Flow Interaction and Stratospheric Warmings

(a) Heat and momentum budgets and the mean circulation

Hartmann [1976b] utilized Nimbus 5 Selective Chopper Radiometer (SCR) data to examine the dynamical climatology of the winter stratosphere in the Southern Hemisphere for 1973. This study was significant in that derived quantities were used in conservation equations to study zonal mean budgets of heat, momentum, and energy. The mean meridional velocity was independently inferred from both heat and momentum balance considerations. For the lower and middle stratosphere, the two results were qualitatively similar and the differences were most probably a result of errors in the data or errors introduced from the approximations used in the analysis. However, in the upper stratosphere the results were qualitatively dissimilar. One implication of these results is that momentum dissipation on scales unresolved by the satellite instrument might be important in the upper stratosphere. Crane *et al.* [1980] came to similar conclusions on the basis of an analysis of heat and momentum budgets for the stratosphere and mesosphere using Nimbus 6 Pressure Modulator Radiometer (PMR) data.

Hamilton [1983a] utilized NMC data from four Northern Hemisphere winters and evaluated the terms in the zonally averaged heat and momentum budgets. Vertical velocity was inferred from the thermodynamic equation and then utilized in the mass continuity equation to evaluate the meridional velocity. The terms in the momentum equation were then evaluated with the residual needed to balance the equation interpreted as the momentum dissipation by unresolved waves (such as gravity waves). These results suggested that an easterly acceleration was required to balance the momentum budget in the region near the stratopause.

Smith and Lyjak [1985] performed a similar study, but utilized LIMS data extending into the mesosphere. The results were broadly consistent with those of Hamilton for Northern Hemisphere winter, indicating a requirement for an easterly acceleration to account for the evaluated residual momentum deficit. Their results also demonstrated a requirement for a westerly acceleration in the springtime Northern Hemisphere. The momentum deficits were then utilized to evaluate an equivalent Rayleigh friction coefficient. The calculated values, however, were substantially larger than the values typically utilized in numerical circulation models.

The results of Hamilton [1983a] and Smith and Lyjak [1985] are qualitatively consistent with current theories for momentum dissipation from breaking gravity waves, although they suggest a possibly significant role for gravity wave stresses in the upper stratosphere as well as in the mesosphere. However, the quantitative estimates of this effect are uncertain because errors inherent in the data set, as well as those introduced by the approximations utilized in the analysis, are also incorporated into the evaluated momentum residual term.

(b) Stratospheric warmings in the Northern Hemisphere

Numerous investigators have utilized satellite data to investigate stratospheric warming phenomena. Palmer [1981a,b] studied the stratospheric major warmings of 1979 and 1980 employing SSU data and the transformed Eulerian-mean formulation of Andrews and McIntyre [1976, 1978a]. Palmer diagnosed the thermal and momentum budgets of the stratosphere during the warming events. Calculated Eliassen-Palm cross sections were used to study the evolving wave-mean flow interactions. Based upon the results of these analyses, Palmer speculated upon the idea of preconditioning of the stratosphere with formation of a high-latitude jet core prior to a major warming. This study generally supports the concept advanced by Kanzawa [1980, 1982, 1984] (and foreshadowed by Quiroz *et al.* [1975]) of a minor warming preconditioning the polar latitudes for a subsequent major warming.

Additional studies of the 1979 warming analyzed by Palmer [1981a,b] have been performed by Gille and Lyjak [1984], Gille *et al.* [1983], and Grose [1984] using LIMS data. The results are consistent with those of Palmer. Minor differences between the various results are attributable to the different data sets (SSU and LIMS) and differences in the analyses.

O'Neill and Youngblut [1982] used NMC data to perform an analysis of the 1976/1977 warmings. They also adopted the transformed Eulerian-mean formulation and Eliassen-Palm cross sections as a diagnostic in their analysis. Ray tracing was used in conjunction with quasi-geostrophic refractive index to study the propagation of wave disturbances. Their analysis concluded that a strong jet at high latitudes favored focusing of wave activity into the polar region with subsequent deceleration of the zonal jet; this process was discussed in Section 6.2.

(c) Southern Hemisphere studies

Relatively fewer diagnostic studies of the Southern Hemisphere have been conducted using satellite data. One inhibiting factor is the relatively poorer upper tropospheric base level analyses that must be used to build up the geopotential heights with the satellite temperature data. Larger ocean areas and fewer radiosonde stations (in comparison to the Northern Hemisphere) make accurate base level analyses for the Southern Hemisphere more difficult. Hartmann *et al.* [1984] utilized SSU data (NMC analysis) to study wave-mean flow interaction for the Southern Hemisphere winter of 1979. Their results showed evidence for a dipole structure in the Eliassen-Palm flux divergence centered near 65 °S corresponding to acceleration of the mean-flow, perhaps suggestive of a local source of wave activity. The origin of this source is unclear, although Hartmann *et al.* suggest barotropically unstable modes as the cause.

Yamazaki and Mechoso [1985] investigated the Southern Hemisphere final warming, when the winter westerlies give way to summer easterlies, of 1979. The evolution of the flow during this period was mostly a gradual process, suggestive of radiative control, but with some contribution from intermittent planetary wave events.

The evolution of a stratospheric minor warming during Southern Hemisphere mid-winter was examined by Al-Ajmi *et al.* [1985] using SCR data. This study highlighted some fundamental differences between

Southern Hemisphere warmings and their counterparts in the Northern Hemisphere. This warming was confined to higher levels and lower latitudes than typical for Northern Hemisphere warmings. The jet shifted poleward and downward. The calculated Eliassen-Palm cross sections exhibit the dipole structure of high latitude divergence and low latitude convergence similar to that noted by the Hartmann *et al.* [1984] study. The deceleration of the zonal mean wind associated with the region of convergence was comparable with that of Northern Hemisphere major warmings. Isentropic maps of potential vorticity suggest low latitude air being irreversibly mixed into middle latitudes.

6.4.4. Studies of Transport Processes

(a) Potential vorticity

Hartmann [1976b] studied the zonal mean budget of potential vorticity for the Southern Hemisphere winter of 1973 utilizing SCR, ITPR, and NEMS data from the Nimbus 5 satellite. As noted in Section 6.2, Ertel's potential vorticity is a conserved tracer under the assumptions of adiabatic, frictionless flow [Ertel, 1942]. In the lower to middle stratosphere, both ozone and potential vorticity are quasi-conserved for a few days, at least. Therefore, potential vorticity serves as a proxy for studying the transport of quasi-conserved trace species in this region of the atmosphere. Hartmann concluded in this study that the eddy transport of potential vorticity was consistent with observations of ozone transport and calculations conducted with general circulation models.

McIntyre and Palmer [1983, 1984] utilized maps of Ertel's potential vorticity on isentropic surfaces in the middle stratosphere to study transport and mixing in the winter stratosphere during the period of January-February 1979. The data used in this study were from the SSU instrument on the Tiros-N satellite. In contrast to the zonal mean potential vorticity calculations of Hartmann [1976b], these hemispheric maps vividly depicted the large-scale transport processes occurring during a minor and major warming in this period. McIntyre and Palmer utilized these maps of potential vorticity to show evidence of "wave breaking" or irreversible deformation of the isentropic potential vorticity contours. They suggested that the breaking waves act to erode the polar vortex (a region of high potential vorticity with strong meridional gradients) and produce a mixed region in middle latitudes (or "surf zone") with relatively weak gradients. These results suggest that isentropic potential vorticity maps can provide extremely valuable insight and further understanding of transport processes.

Hoskins *et al.* [1985] have advocated mapping of potential vorticity distributions as a powerful diagnostic technique for studies of dynamics and transport in both stratosphere and troposphere. The power of potential vorticity as a dynamical diagnostic stems not only from the conservation property, but also from an "invertibility" principle which states that knowledge of the isentropic potential vorticity fields with suitable boundary information is sufficient to determine all of the dynamical variables of interest.

The concept of using isentropic mapping potential vorticity as a diagnostic has received impetus from the study of Clough *et al.* [1985], who produced potential vorticity maps from SSU data to study the evolution of a Canadian warming in Dec. 1981. Butchart and Remsberg [1986] utilized LIMS data and developed the concept of "area diagnostics" (time evolution of the area contained within a contour of either constant potential vorticity or species mixing ratio on an isentropic surface) advanced by McIntyre and Palmer [1983, 1984]. They demonstrate the use of the area diagnostics as a means of delineating the mixed or surf zone from the polar vortex, and for detecting when nonconservative effects such as diabatic heating or photochemistry become important.

(b) Constituents

Leovy *et al.* [1985] studied the transport of ozone in the stratosphere using LIMS data for the period Oct. 1978 to May 1979. These results were supportive of the wave breaking hypothesis of McIntyre and Palmer [1983, 1984]. Leovy *et al.* noted that the ozone displayed a strong negative correlation with isentropic potential vorticity maps. Ozone mixing ratio from LIMS on the 850 K isentropic surface of January 27, 1979, is presented in Figure 6-45 for comparison with the potential vorticity maps in Figure 6-44. Both ozone and EPV maps display the signature of a "breaking wave". The core of high EPV (low ozone) is being drawn out clockwise around an intensifying Aleutian anticyclone. Concomitantly, a tongue of lower EPV (higher ozone) is being drawn from low latitudes poleward in the region of confluence between the anticyclone and the polar low. This result is to be anticipated since both ozone and potential vorticity are quasi-conserved over times of a few days in the lower to middle stratosphere. Leovy *et al.* suggested that observed differences in the ozone and potential vorticity arose from differing nonconservative processes affecting the two tracers and from their different gradients. The results of the analysis provide support for the idea of the ozone hole associated with the polar vortex being filled as the episodes of wave breaking produce irreversible mixing of tongues of higher ozone mixing ratio from the tropics into high latitudes. The contribution of this process to the seasonal evolution of ozone needs further study.

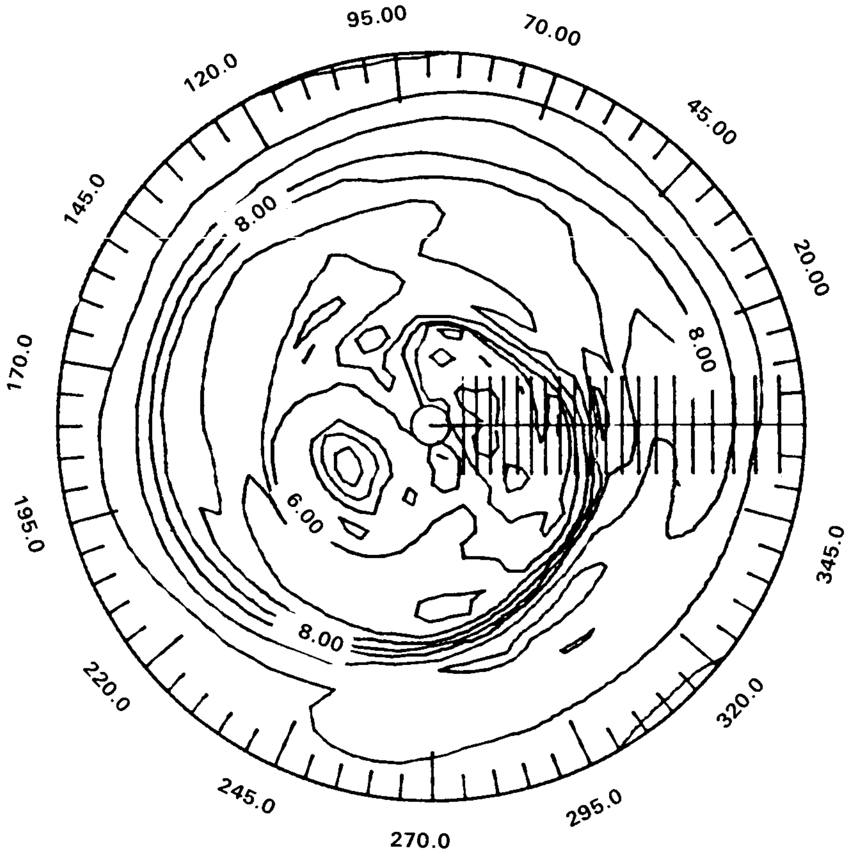


Figure 6-45. Ozone mixing ratio (ppmv) on the 850 K isentropic surface, (~10 mbar), January 26, 1979. LIMS data [from Grose and Russell, 1985].

Grose [1984] and Grose and Russell [1985] also point out the correspondence between potential vorticity and quasi-conserved species. In particular, water vapor shows a strong correlation with potential vorticity as shown in Figure 6-46. The results suggest that fields of isentropic potential vorticity and quasi-conserved species in combination have the potential for providing a wealth of information on dynamical, chemical and radiative processes.

Calculation of parcel trajectories using satellite data has great potential for studying the coupling of radiative, chemical, and dynamical processes. For example, Austin and Tuck [1985] and Austin [1985] have used SSU and LIMS data to calculate parcel trajectories in the stratosphere. Solomon and Garcia [1983b] and Callis *et al.* [1983b] illustrated the use of trajectory analysis to show the coupling between dynamics and photochemistry responsible for producing the "Noxon cliff" or strong latitudinal gradient in column NO_2 observed during displacement of the polar vortex during so-called "wave 1 events".

6.4.5. Conclusions

The recent availability of satellite data of sufficient quality to derive dynamical quantities has stimulated many diagnostic studies. The uncertainties associated with derived quantities remain to be evaluated. However, there is little doubt that they provide much useful qualitative insight into dynamics and transport processes. The use of isentropic maps of potential vorticity as a diagnostic is only beginning to be

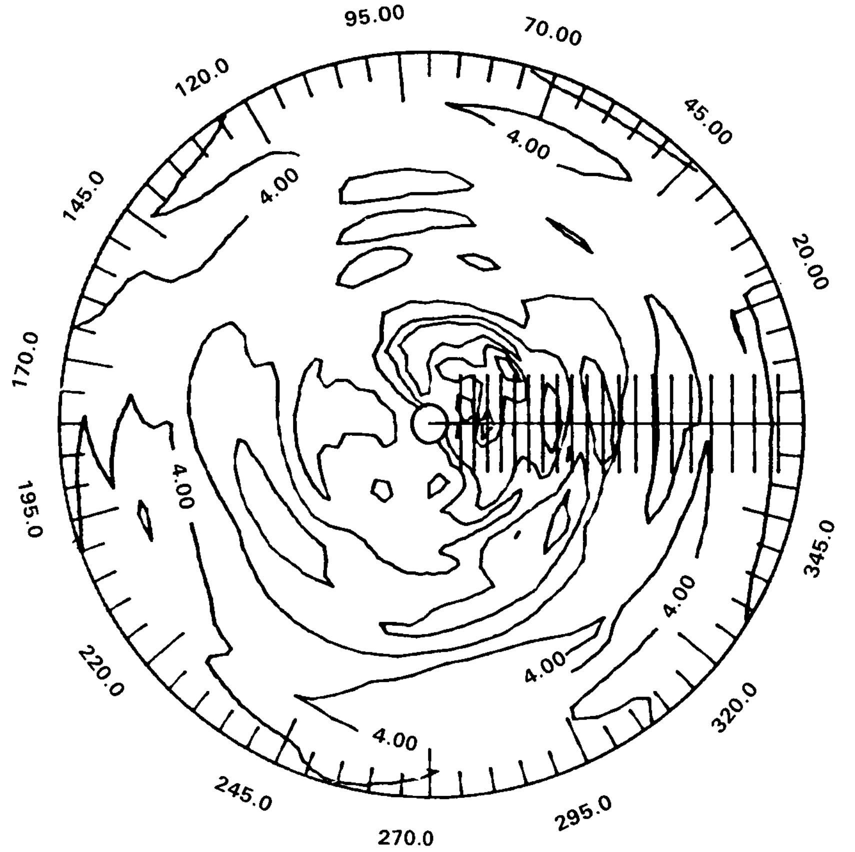


Figure 6-46. Water vapor mixing ratio (ppmv) on the 850 K isentropic surface (~ 10 mbar), January 27, 1979. LIMS data [from Grose and Russell, 1985].

explored, but promises to greatly expand our understanding of fundamental processes, particularly when used in conjunction with fields of trace chemicals measured simultaneously.

6.5 THEORY OF TRANSPORT PROCESSES

6.5.1 Introduction

The problem of transport of a conserved tracer - the amount of which is constant with time within a given material parcel of air - is in principle simply a matter of tracking parcel movements. Over short periods of time, this may be a useful approach - and indeed has proved to be so in some such cases [e.g., Allam and Tuck, 1984b; Austin and Tuck, 1985]. However, given the complexity of the atmospheric trajectories, this is not in itself a practical avenue for understanding the global transport of atmospheric constituents. Nevertheless this simple property of *Lagrangian* conservation (i.e., following material elements) is the root of all transport processes and must not be hidden in any analysis procedure if we wish to retain insight into global transport mechanisms.

The global transport problem has for twenty years or more been reduced to one of manageable (though still difficult) proportions by limiting the aim to one of explaining the behavior of zonally-averaged atmospheric constituent structures on long time scales (say, a month and longer). The restriction to zonal averages greatly reduces the number of degrees of freedom in the problem, without a profound loss of information, since the strong quasi-zonal flow in most regions of the atmosphere ensures that much of the seasonal-mean structure of long-lived constituents is in the meridional (latitude-height) plane. In fact it is only relatively recently, with the advent of satellite monitoring of atmospheric constituents, that our knowledge of their structure has been comprehensive enough to demand anything more sophisticated than a zonal-mean conceptual framework. As noted in Sections 6.1 and 6.4, it is now recognised that even the monthly-mean flow in the winter stratosphere may be highly asymmetric in the zonal direction and therefore that the zonally-averaged approach may have severe limitations. These limitations and possible avenues for avoiding the restriction to zonal averages will be addressed later in this section. However at the present time the zonal-mean formulation forms the basis of assessment modeling (Chapters 12 and 13) and we therefore begin this Section with a detailed discussion of recent developments in zonal-mean transport theory.

6.5.2 Zonally-Averaged (2D) Formulations

A quasi-conserved tracer of mixing ratio q satisfies a conservation relation of the form:

$$\frac{dq}{dt} = S \tag{13}$$

where d/dt is the derivative following the flow and S represents sources and/or sinks of q. For a quasi-conserved quantity, S is normally small although it is important to recognize that the transport processes themselves may ensure that S becomes large, even though a naive scaling analysis may indicate otherwise. (For example, the importance of local diffusion of q may be greatly intensified as a result of the shearing of q gradients by the flow and the consequent cascade to small scales). The traditional zonal mean budget equation for $\bar{q}$ then follows from the zonal mean of (13), viz.,

$$\left[\frac{\partial}{\partial t} + \underset{\sim}{\bar{u}} \cdot \underset{\sim}{\nabla}\right]\bar{q} + \frac{1}{\varrho_0}\underset{\sim}{\nabla} \cdot \varrho_0 \overline{\underset{\sim}{u}'q'} = \bar{S} \tag{14}$$

where ϱ_0 is the basic atmospheric density. Thus the zonal-mean transport of q is mathematically split into two components: advection of q by the mean meridional circulation $(\overline{v},\overline{w})$ and an "eddy flux" $\varrho_0\overline{\underline{u}'q'}$. Attempts to arrive at a simple description of transport on this basis, however, proved confusing (e.g., see the discussion of McIntyre [1980b] and Mahlman *et al.* [1984]). It is now recognized following the work of Andrews and McIntyre [1976, 1978b] that the simplicity of (14) is deceptive and that the mathematical separation of transport into "mean" and "eddy" components is an arbitrary procedure which may not (and in practice does not) yield the simplest picture of transport [McIntyre, 1980a]. The formulation (14) proves to be less straightforward than it might appear simply because this separation has masked the three-dimensional Lagrangian conservation properties of (13). These properties are more faithfully preserved in an alternative formulation, which is even simpler in form than (14), viz., the budget equation for the generalized Lagrangian-mean mixing ratio $\overline{q}^L$, as defined by Andrews and McIntyre [1978b], which is

$$\left[\frac{\partial}{\partial t} + \overline{\underset{\sim}{u}}^L \cdot \underline{\nabla}\right] \overline{q}^L = \overline{S}^L \tag{15}$$

[McIntyre, 1980a]. Thus transport of the Lagrangian-mean mixing ratio resides solely in the term representing advection by the Lagrangian-mean meridional circulation $\overline{\underset{\sim}{u}}^L = (\overline{v}^L, \overline{w}^L)$ and there are no explicit eddy transport terms (apart from any effects implicit in $\overline{S}^L$).

The reason for the greater simplicity of the generalized Lagrangian-mean approach is evident from Figure 6-47. The generalized Lagrangian-mean mixing ratio $\overline{q}^L$ at reference latitude and height (y_L, z_L) is, by definition, the average along a material tube of air whose center of mass in the latitude/height plane is at (y_L,z_L), as shown in Figure 6-47a. After a time δt (Figure 6-47b) this material tube is displaced by mean motions to a new reference (center of mass) position $(y_L+\overline{v}^L\delta t,\ z_L+\overline{w}^L\delta t)$ and is distorted by eddy motions. However, if $S = 0$, then every element in the tube conserves its value of q and therefore the Lagrangian mean $\overline{q}^L$ *evaluated at the new reference position* $(y_L+\overline{v}^L\delta t,\ z_L+\overline{w}^L\delta t)$ is identical to that at (y_L, z_L) at the outset, independent of the wavy distortions of the tube. Therefore, in a frame of reference moving with the center of mass of the tube (which by definition moves with the Lagrangian-mean velocity $\overline{\underset{\sim}{u}}^L$), $\overline{q}^L$ is constant if $S = 0$. This is the essence of Equation (15).

Despite the simplicity of this description - which has led to profound conceptual advances in atmospheric transport theory - its practical application is beset by many problems [McIntyre, 1980b] not the least of which is the determination of the Lagrangian-mean flow $(\overline{v}^L, \overline{w}^L)$ itself. Dunkerton [1978] argued that a reasonable estimate of this circulation could be obtained via the diabatic circulation (see Table 6-1, below). Thus Dunkerton obtained a picture of the circulation of the middle atmosphere (Figure 6-48a) which differs substantially from the Eulerian-mean picture - cf. Figure 6-48b. The implied transport characteristics of the flow depicted in Figure 6-48a are, unlike those which would be inferred from Figure 6-48b, consistent with the global structure of trace constituent distributions (and, indeed, flow patterns similar to 6-48a had earlier been inferred from observed tracer distributions [Brewer 1949, Dobson, 1956]). For one thing the upward motion into the lower stratosphere occurs solely in low latitudes, as has been inferred from the observed dryness of the stratosphere (see Chapter 5). Further, the poleward/downward flow in the winter hemisphere is consistent with the observed structure of long-lived stratospheric tracers, including ozone and potential temperature, the circulation transporting ozone from the tropical middle and upper stratosphere to lower levels in the winter high latitudes and the subsidence at the winter high latitudes maintaining the temperature above its radiative equilibrium, as observed.

While this conceptual picture of large-scale transport processes represents a considerable advance over the traditional viewpoint embodied in Equation (14) it is not as such a complete description. The main

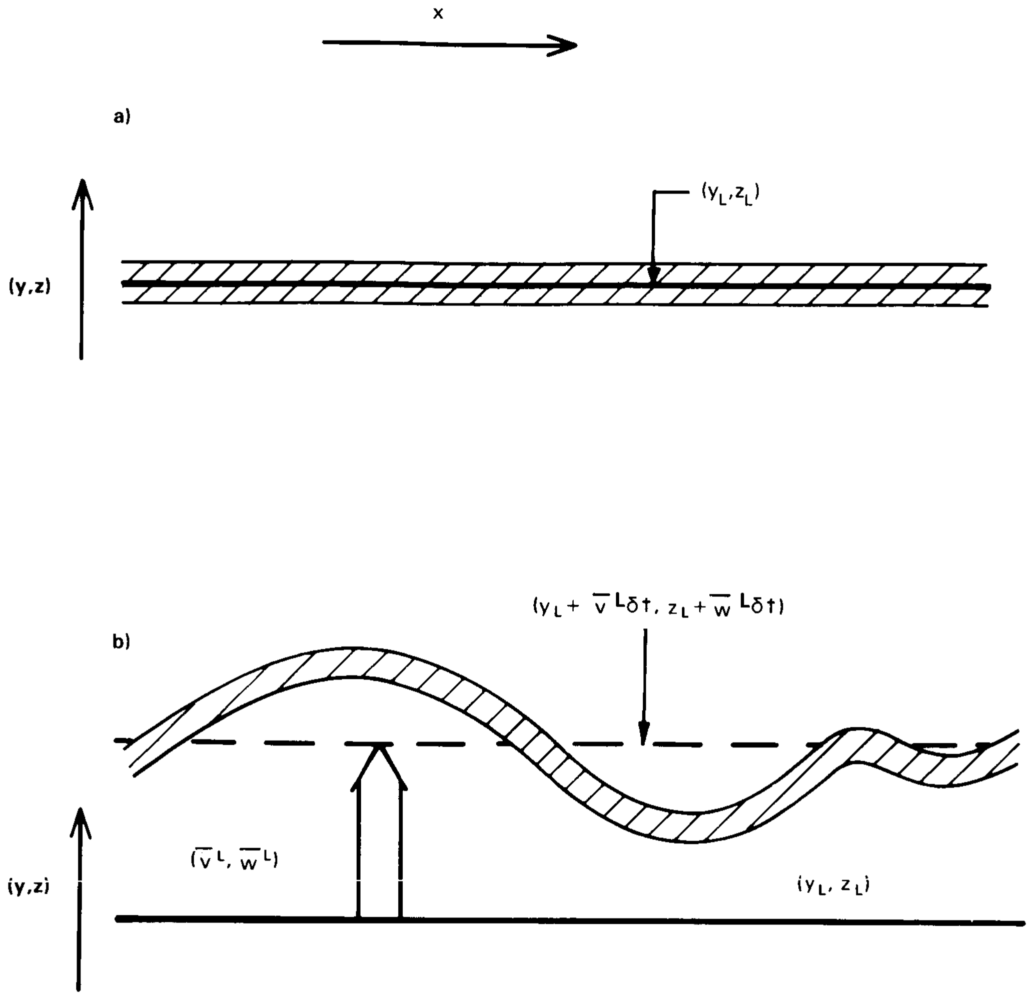

x
a)
(y_L, z_L)
(y,z)
$(y_L + \overline{v}^L\delta t,\ z_L + \overline{w}^L\delta t)$
b)
$(\overline{v}^L, \overline{w}^L)$
(y,z)
(y_L, z_L)

Figure 6-47. Schematic illustration of the conservation of the generalized Lagrangian-mean mixing ratio of a conserved tracer. The hatched region is a material tube of air which moves from its location shown in (a) at some time to that shown in (b) a time δt later. See text for discussion.

reason for this is that (15) is a prediction equation for the Lagrangian-mean mixing ratio $\overline{q}^L$ whereas what is usually required is a description of the evolution of Eulerian measures of q, which may be quite different from the evolution of $\overline{q}^L$. Consider, for example, a wave breaking event as depicted schematically in Figure 6-49, in which, for simplicity of argument, it is assumed that $\overline{v}^L$ and $\overline{w}^L$ are zero. The mixing ratio isopleths of a conserved tracer q (it is further assumed that $S = 0$) are initially aligned zonally, with, say high values to the south. A transient breaking wave event distorts these isopleths to such an extent that the reference material curve C (on which $q = q_c$ is constant) is fractured leaving, after the passage of the event, an isolated pool of high q to the north of its initial latitude and a pool of low q to the south. This irreversible dispersion of material contours clearly achieves a very real (Eulerian) transport in latitude which is not explicitly revealed by the Lagrangian-mean budget, since (15) tells us, under the present assumptions that $\partial\overline{q}^L/\partial t = 0$. The point is, of course, that transport has been achieved, not by any change in $\overline{q}^L$, *but by an irreversible deformation of the contour C along with $\overline{q}^L$ is determined.* Mathematically, this transport is implicit in (15) only through the mapping $\overline{q}^L \rightarrow \overline{q}$ which must be performed in order apply (15) to practical problems [Plumb, 1979; McIntyre, 1980b].

Table 6-1. Measures of the mean meridional circulation

(i)	Eulerian mean circulation $(\bar{v},\bar{w})$	Conventional, Eulerian, zonal average
(ii)	Generalized Lagrangian-mean circulation $(\bar{v}^L,\bar{w}^L)$	Zonal average along (wavy) material lines. Velocity of center of mass of material tubes of fluid. The only circulation on this list which is not, in general, nondivergent.
(iii)	Residual circulation $(\bar{v}_*,\bar{w}_*)$	Defined by Equation (2). The mean circulation of ''transformed Eulerian-mean'' theory. This formulation greatly simplifies the quasigeostrophic zonal-mean budget equations for heat and momentum.
(iv)	Diabatic circulation $(\bar{v}_D,\bar{w}_D)$	The mean circulation in isentropic coordinates. Equals (iii) for quasigeostrophic flow if mean isentropes are stationary.
(v)	Transport circulation (V_T,W_T)	Defined by Equation (17). The ''advective mass flux'' of Kida [1983a]. The non-diffusive component of transport in the formulation (16). Equals (ii) if eddy-induced dispersion is spatially homogeneous; equals (iii) for small-amplitude, adiabatic eddies.

Another serious problem with the practical application of (15) has been brought to light by determination of the Lagrangian-mean circulation in the stratosphere of numerical models [Kida, 1977, 1983a; Plumb and Mahlman, 1986]. Andrews and McIntyre [1978b] pointed out that $\underset{\sim}{\bar{u}}^{L}$ is not a nondivergent velocity (i.e., it does not satisfy the usual continuity equation) and discussed the reasons for this. Now, our concept of an advective process - one that displaces the center of mass of a tracer distribution without, to first over, affecting the spread of the distribution about the center of mass (unlike dispersive processes such as that discussed above in the context of Figure 6-49) is implicitly based on an assumption that the advecting velocity is nondivergent. However, Kida [1983a] and Plumb and Mahlman [1986] found the Lagrangian-mean meridional circulation in the numerical models they investigated to include a large divergent component, so much so that the Lagrangian-mean flow, shown in Figure 6-50, bears limited resemblance to Figure 6-48a.

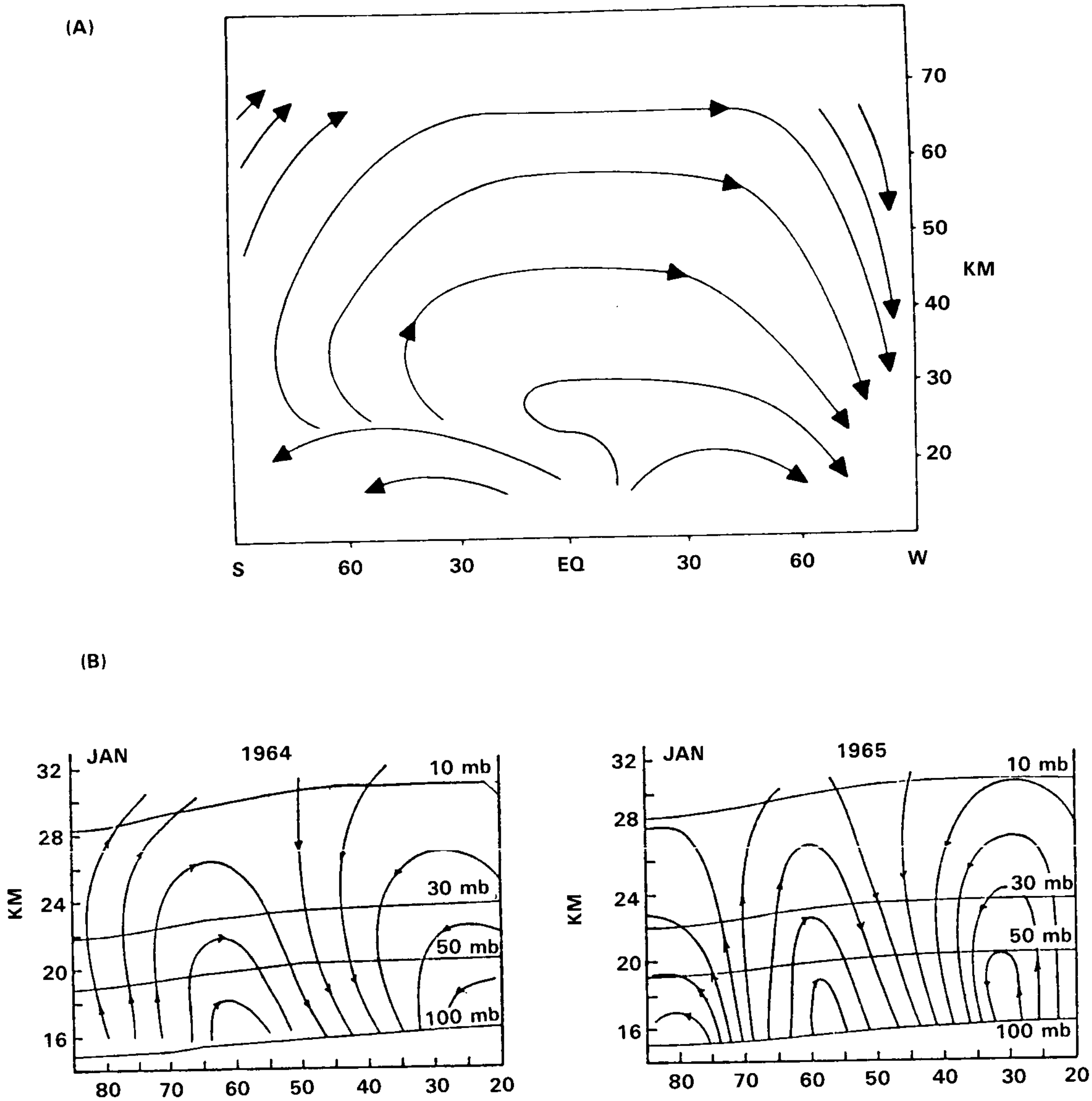


Figure 6-48. (a) Streamlines (schematic) of the diabatic circulation of the middle atmosphere at the solstices. 'S' and 'W' denote summer and winter pole, respectively. [After Dunkerton, 1978]. (b) Eulerian-mean meridional circulation (schematic) of the Northern Hemisphere winter stratosphere [After Vincent, 1968].

Plumb [1979] showed that, for small-amplitude eddies, the zonal-mean constituent budget equation could be written:

$$\frac{\partial \bar{q}}{\partial t} + V_T \frac{\partial \bar{q}}{\partial y} + W_T \frac{\partial \bar{q}}{\partial z} = \frac{\partial}{\partial y}\left[K_{yy}\frac{\partial \bar{q}}{\partial y} + K_{yz}\frac{\partial \bar{q}}{\partial z}\right] + \frac{1}{\varrho_\alpha}\frac{\partial}{\partial z}\left[\varrho_o K_{zy}\frac{\partial \bar{q}}{\partial y} + \varrho_o K_{zz}\frac{\partial q}{\partial z}\right] + \bar{S} \qquad (16)$$

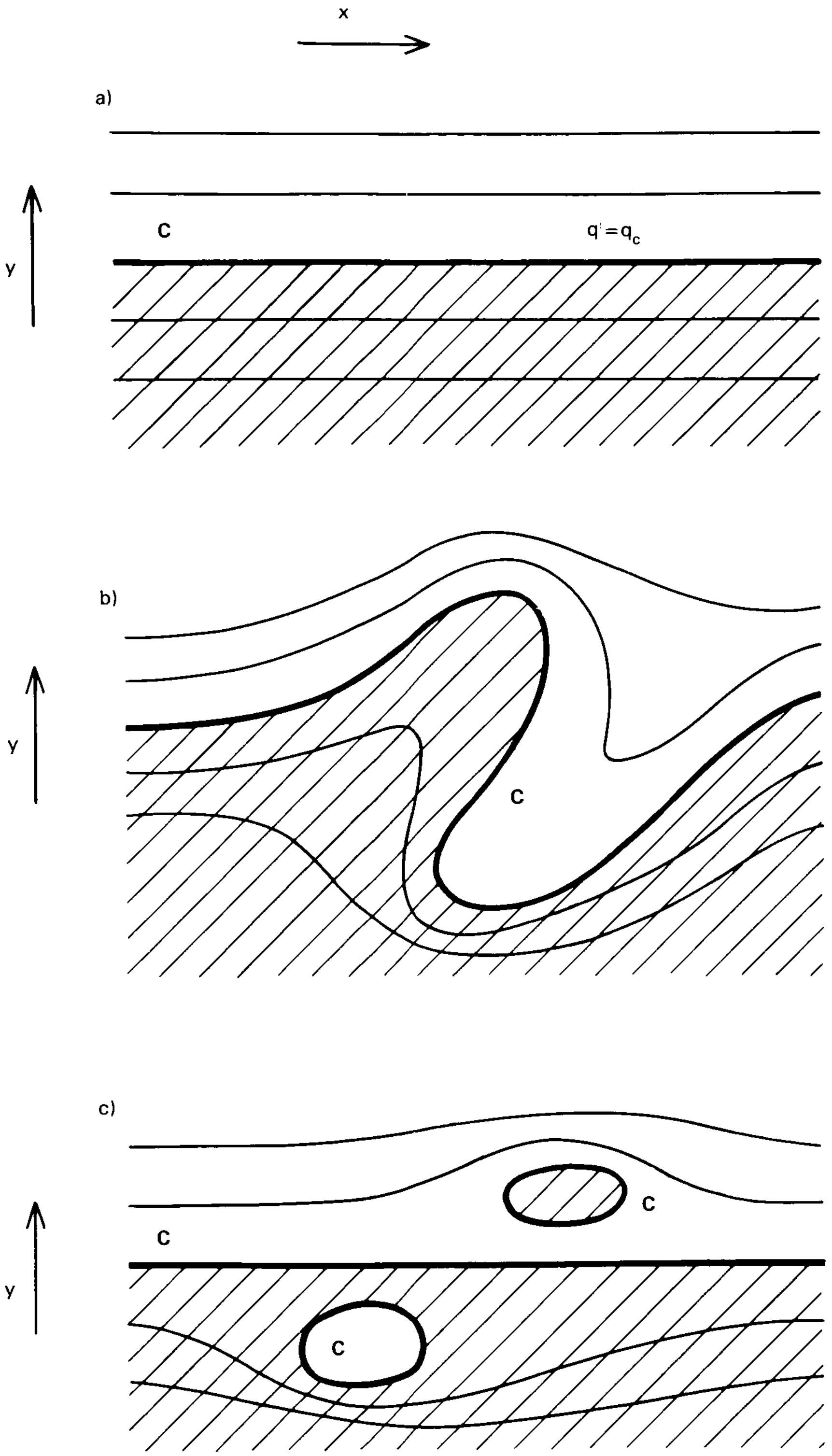


Figure 6-49. Schematic illustration of the irreversible distortion of material lines (a) before, (b) during and (c) after a breaking wave event. The heavy curve C is an isoline of a tracer $q = q_c$. Regions of $q > q_c$ are shaded. See text for discussion.

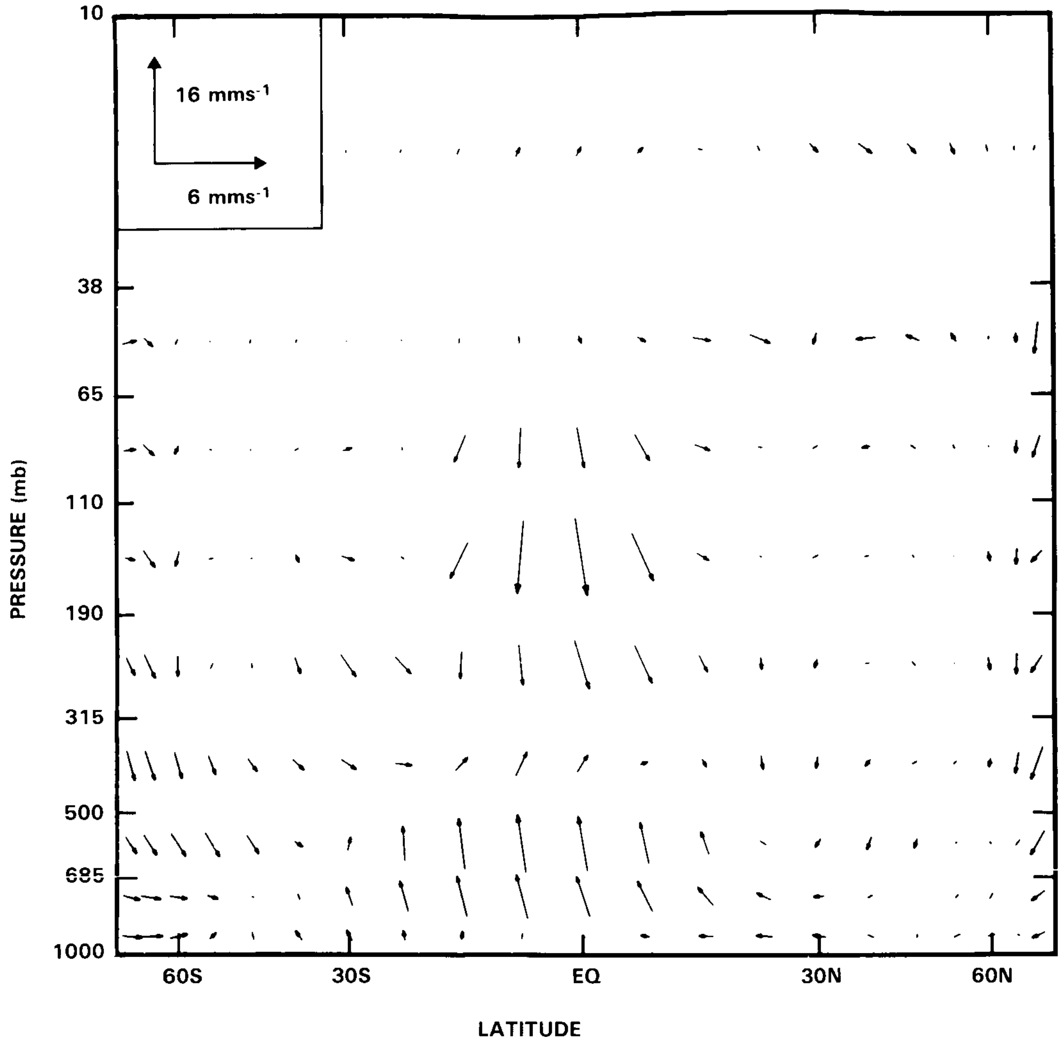


Figure 6-50. Lagrangian-mean circulation ($\overline{v^L}, \overline{w^L}$) in the GFDL general circulation/tracer model [Mahlman and Moxim, 1978] as determined by Plumb and Mahlman [1986].

where (V_T,W_T) is a nondivergent meridional circulation defined by

$$\begin{bmatrix} V_T \\ \\ W_T \end{bmatrix} = \begin{bmatrix} \bar{v} + \dfrac{\partial L}{\partial y} \\ \\ \bar{w} - \dfrac{1}{\varrho_0}\dfrac{\partial}{\partial z}(\varrho_0 L) \end{bmatrix} \qquad (17)$$

with $L = \frac{1}{2}(\overline{v'\zeta} - \overline{w'\eta})$, and **K** is a diffusivity tensor which, for conserved tracers, is defined by

$$\begin{bmatrix} K_{yy} & K_{yz} \\ \\ K_{yz} & K_{zz} \end{bmatrix} \begin{matrix} = \\ \\ = \end{matrix} \begin{bmatrix} \dfrac{1}{2}\dfrac{\partial\overline{\eta^2}}{\partial t} & \dfrac{1}{2}\dfrac{\partial\overline{(\eta\zeta)}}{\partial t} \\ \\ \dfrac{1}{2}\dfrac{\partial\overline{(\eta\zeta)}}{\partial t} & \dfrac{1}{2}\dfrac{\partial\overline{\zeta^2}}{\partial t} \end{bmatrix} \qquad (18)$$

where (η, ζ) are the eddy displacements in the (y,z) directions as defined by Andrews and McIntyre [1978b]. These terms in (18) describe not only the effects of dispersion relative to the Lagrangian-mean flow as discussed above, but also the transport by the divergent part of the Lagrangian-mean circulation. Then the advection by the nondivergent "transport circulation" (V_T, W_T) is truly advective, in the normal sense.

Since the pioneering work of Andrews and McIntyre [1976, 1978b] brought about a breaking away from the conventional definition of the "mean" circulation as the basis of transport formulations, several new definitions of "mean circulation" have appeared, of which the transport circulation is one of four. Each of these measures of the mean meridional circulation has its own particular application. There has, however, been a confusing tendency in the recent literature of transport modeling to overlook the differences between some of these. The different circulations are listed in Table 6-1, in which some of their interrelationships are noted. The transport circulation differs from the Lagrangian-mean velocity unless the diffusivity **K** is spatially homogeneous [Andrews and McIntyre, 1978b; Plumb, 1979; Matsuno, 1980; Kida, 1983a]. While (V_T,W_T) may be defined by Equation (17), this is not a practical avenue for its determination. Kida [1983a] refers to this velocity as the "advective mass flux" and determined it from a numerical model as that part of the circulation which is asymmetric with respect to a time-reversal, while Plumb and Mahlman [1986] derived it by inverting a flux-gradient relation for trace constituent fluxes; their results are illustrated in Figure 6-51. This circulation is very similar to that obtained by Dunkerton [1978] as shown in Figure 6-49; reasons for this will be discussed below.

The same formalism can be used to help understand the variability of trace constituents. The perturbation concentration of a conserved tracer is related to the eddy displacements through

$$q' = -\eta \frac{\partial \bar{q}}{\partial y} - \zeta \frac{\partial \bar{q}}{\partial z}$$

If $\gamma = -(\partial \bar{q}/\partial y)/(\partial \bar{q}/\partial z)$ is the slope of the mean isopleths of q, then

$$q' = -\frac{\partial \bar{q}}{\partial z}(\zeta - \gamma\eta).$$

Therefore the variance of q around a latitude circle is

$$\overline{q'^2} = \left[\frac{\partial \bar{q}}{\partial z}\right]^2 \overline{(\zeta - \gamma\eta)^2}. \qquad (19)$$

Ehhalt *et al.* [1983] defined an "equivalent displacement height"

$$\Delta = (\overline{q'^2})^{1/2} \Big/ \left|\frac{\partial \bar{q}}{\partial z}\right| \qquad (20)$$

and found this quantity to be essentially the same for a range of long-lived constituents in the lower stratosphere. (Actually the variance used in (20) by Ehhalt *et al.* was the variance in time at a fixed point). This result can be understood from (19) under two conditions: that the constituents are conserved on the time scale of the eddies (so that (19) is valid) and that the mean isopleth slope of all long-lived constituents is the same. The latter condition has been argued and discussed by Mahlman [1985] and Mahlman *et al.*

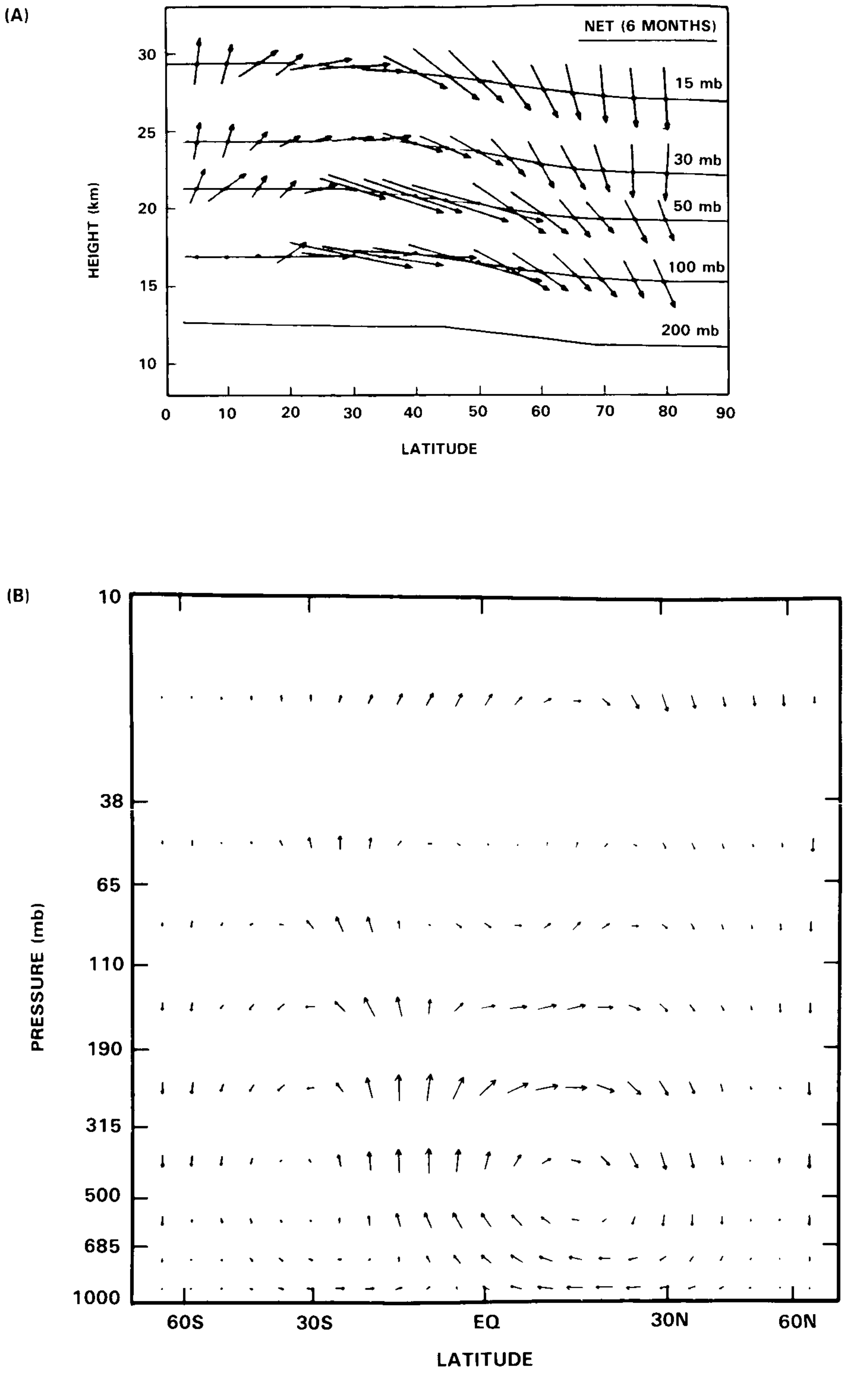


Figure 6-51. Model-determined transport circulation (V_T,W_T) for Northern winter according to (a) Kida [1983a]; (b) Plumb and Mahlman [1986].

[1985]. Under these circumstances, we can equate Δ^2 in (20) with $\overline{(\zeta-\gamma\eta)^2}$ in (19); Δ is just the eddy displacement normal to the mean isopleths.

For small-amplitude eddies, then, (16) describes the transport of exactly conserved constituents in terms of advection by the transport circulation (V_T,W_T) and diffusion with a diffusivity defined by (17). This approach has been discussed further by Matsuno [1980] and Danielsen [1981]. Note that, as described thus far, these diffusivities are purely *kinematic*, i.e., they depend solely on the properties of the flow field and are independent of the constituent field. However, for non-conserved constituents, this is no longer the case. The reason for this is that additional transport effects arise - the so-called "chemical eddy" terms - which depend on the nonconservative effects. For a weak sink of the form $s = -\lambda(q-q_o)$, Plumb [1979] showed that (16) needs to be modified by the inclusion of an additional diffusivity $\mathbf{K}^{(c)}$ where:

$$\begin{bmatrix} K_{yy}^{(c)} & K_{yz}^{(c)} \\ K_{zy}^{(c)} & K_{zz}^{(c)} \end{bmatrix} = \begin{bmatrix} \lambda\overline{\eta^2} & \lambda\overline{\eta\zeta} \\ \lambda\overline{\eta\zeta} & \lambda\overline{\eta^2} \end{bmatrix} \qquad (21)$$

Similar, but more general, expressions are given by Matsuno [1980], Pyle and Rogers [1980b] and Tung [1984]. The reason for this additional effect, which is implicit in the Lagrangian-mean budget of (15) via the term $\overline{S}^L$, can be understood with reference to Figure 6-52. Consider a steady, laminar, non-dispersing flow (in which the kinematic tracer dispersion effects discussed above do not occur) for which the streamline is a trajectory. Consider now the mixing ratio q of a parcel moving along this trajectory; it is assumed that q has a background gradient such that high values of q are to the south. If q were exactly conserved then any parcel moving along the trajectory C would move northward at A and then return southward at B with exactly the same value of q. Thus there is no net northward transport. If q is allowed to relax toward its local mean value, however, then in that part of the trajectory between A and B, where q is greater than the local mean value (since the parcel is northward of its mean position), q decreases and therefore the parcel returns southward at B with a smaller mixing ratio than it took northward at A. Therefore there is a net northward (downgradient) flux of q induced by the nonconservative term. Since this effect is dependent on the chemistry of the constituent (or, for the quasi-conservative quantities entropy and potential vorticity, effects such as radiation) then in principle the transport properties differ from one constituent to another. As Pyle and Rogers [1980b] showed, the situation is even more complicated

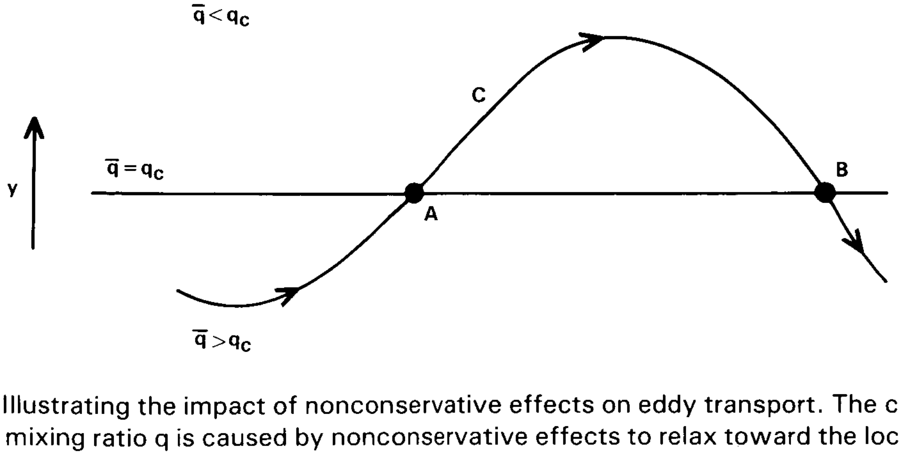


Figure 6-52. Illustrating the impact of nonconservative effects on eddy transport. The curve C is a trajectory. If the mixing ratio q is caused by nonconservative effects to relax toward the local mean value, then parcels move northward at A with a greater value of q than that with which they return southward at B, thus achieving a net northward transport of q. See text for discussion.

if the constituents interact with one another, since the transport properties of each constituent then depend on the structure of the others; they showed, however, that this additional complication may be avoided by restricting attention to the transport of families of interacting constituents. Further, as Tung [1984] has noted, within the range of validity of the simplified expression (21), the appropriate values of $\mathbf{K}^{(c)}$ may be easily determined for any constituent given the kinematic statistics $\overline{\eta^2}$,$\overline{\eta\zeta}$ and $\overline{\zeta^2}$ (which are dependent only on the flow characteristics) and the relaxation rate coefficient λ for the particular constituent.

Comparing (18) and (21), it is clear that the relative contribution of kinematic dispersion and nonconservative effects to constituent diffusion depends on the ratio of timescales on which these processes act. The transport of all constituents whose chemistry is much slower than the parcel dispersion rate is governed by the same kinematic effects; those constituents with faster chemistry must be treated differently. An indication of the actual value of the critical time scales in the middle atmosphere will be discussed below.

Together, (18) and (21) reveal explicitly the now well-known dependence of transport on eddy transience and nonconservative effects. If the eddies are steady [in the Lagrangian sense demanded by the vanishing of (18)] and the constituent is conserved ($\lambda=0$) then $\mathbf{K}=0$; if, further, the transport circulation vanishes (which, as will be discussed below, depends on similar conditions on potential vorticity transport) then there is no transport of the constituent at all. This corollary of the celebrated "non-acceleration" theorem of Andrews and McIntyre [1976, 1978a] and Boyd [1976] has been called the "non-transport" theorem by Mahlman *et al.* [1980]. The power of this theorem stems not so much from the fulfillment of these conditions (since they are never exactly satisfied in real situations) as from the fact that it highlights those processes which are important (viz. transience and nonconservative effects) to constituent transport. It cannot be over-emphasized that the transience as defined by (17) is *Lagrangian* transience (i.e. the time derivative of Lagrangian displacement statistics) which is not necessarily related in a simple way to the properties of Eulerian statistics such as velocity variances. This distinction between Lagrangian and Eulerian behavior has on some occasions been overlooked in recent literature on stratospheric transport, even though it has long been recognized as fundamental in classical turbulent transport theory. For example, the integrated transience $\{\int(\partial\alpha^2/\partial t)\bullet dt\}$ over a wave pulse (whether a temporary event or an entire winter season) necessarily vanishes if α is an Eulerian quantity which is sufficiently small before and after the event. However the same is clearly not true of Lagrangian statistics as exemplified by the hypothetical example of Figure 6-49 where a wave event leaves a *permanent* distortion of material lines. A related point is that the structure of Eulerian eddy statistics is not in general a good guide to the structure of transport processes.

6.5.3 Gross Characteristics of Atmospheric Transport

Formulations of the form of (16) have now been used for two decades as the basis of two-dimensional transport modeling, since the pioneering work of Reed and German [1965]. The parameterization problem thus posed will be specifically addressed in Chapter 12; however a discussion of the properties of atmospheric transport as represented by such formulations is in order here. The major qualitative question that arises is whether, in practice, the meridional transport of atmospheric constituents is primarily advective or diffusive in character (or both). Reed and German [1965] made assumptions which led to the conclusion that the eddy fluxes are diffusive. More recently, however, it has been shown that these fluxes are, in the midlatitude stratosphere, better represented by an effective advection [Clark and Rogers, 1978; Plumb, 1979, Matsuno, 1980; Pyle and Rogers, 1980b; Danielsen, 1981] which to a first approximation achieves the now well-known cancellation between mean and eddy transport in the traditional formulation of (14) [Hunt and Manabe, 1968; Mahlman and Moxim, 1978] and which is thus a manifestation of the "non-transport" theorem. These developments have led to suggestions that midlatitude stratospheric transport is primarily advective in character [Holton, 1981; Garcia and Solomon, 1983; Guthrie *et al.*, 1984; Solomon and Garcia, 1984b; Tung, 1984]. However, one needs to be careful here; as already emphasised in Section 6.2, the mean circulation - however one defines it - can never been regarded as independent of the eddies. In fact if we assume, following Plumb and Mahlman [1986], that the eddy forcing G discussed in Section 6.2 is dominated by quasigeostrophic eddy transport (a reasonable assumption for the midlatitude stratosphere) then, since $\varrho_o^{-1}\, \underset{\sim}{\nabla} \cdot \underset{\sim}{F} = \overline{v'Q'}$ for such motions [Edmon *et al.* 1980] where Q is the quasigeostrophic potential vorticity, the steady state balance of momentum in Equation (4) becomes

$$-f_o\, \overline{v}_* = \overline{v'Q'} \tag{22}$$

If it can be assumed that Q is sufficiently well conserved, its transport characteristics are the same as those of any other conserved tracer and then $\overline{v'Q'} \cong -\, K_{yy}\, \partial\overline{Q}/\partial y$ (other terms being negligible for quasigeostrophic flow). Therefore the steady momentum budget becomes

$$f_o\overline{v}_* = K_{yy}\, \frac{\partial \overline{Q}}{\partial y} \tag{23}$$

What (23) expresses is that for a steady circulation (e.g., in solstice conditions, at least) the dynamical effects of *advection* by the residual circulation and *diffusion* of potential vorticity must balance. Plumb and Mahlman [1986] used this result to argue that, globally [if not locally], the effects of advection and diffusion of long-lived tracers as expressed by (16) must be formally comparable, *provided* Q is a good enough tracer for K_{yy} in (20) to be approximated by the kinematic value appropriate to longlived tracers. A more direct practical indication of the importance of both influences is the accumulation of evidence from observations and numerical model studies that very long-lived stratospheric tracers appear to exhibit a "slope equilibrium" whereby their isopleths of constant zonally-averaged mixing ratio have almost the same shape (e.g., Figure 6-43); Mahlman [1985] and Mahlman *et al.* [1985] explain this characteristic as a balance between the steepening effects of advection and the slope-flattening effects of quasi-horizontal diffusion - certainly advection alone by a circulation such as Figure 6-48a or 6-51 would be expected to produce much steeper slopes, especially at high latitudes.

The argument that advective and diffusive transport must be formally comparable rests on the assumptions of quasigeostrophy and potential vorticity conservation [so that (22) and (23) respectively are valid]. If planetary waves are the primary vehicles for stratospheric transport (which seems very likely, notwithstanding the possible contribution from gravity waves noted elsewhere in this chapter) then the first assumption is justifiable. The second, however, may be suspect where radiative effects contribute significantly to wave dissipation and hence to $\overline{v'Q'}$. Indeed, in theory one can conceive of situations of a non-dispersing wave ($K_{yy} = 0$) which exhibits a nonzero potential vorticity flux by virtue of radiative dissipation; then, by (22), the wave would drive a nonzero residual circulation without any accompanying diffusion of conserved constituents. In general, diffusive transport will be comparable with advective transport whenever dispersion is a significant contributor to potential vorticity transport. This appears to be a safe assumption in the lower stratosphere where radiative time scales are long (20 days or so) and where the mixing processes described in Section 6.4 are quite intense (during winter - all transport processes are weak in the summer stratosphere). The situation is less clear in the upper stratosphere, where radiative time scales are only a few days. However planetary wave mixing is also intense in the winter upper stratosphere and so, even here, dispersive effects may be a major component of potential vorticity transport. It is clearly conceptually important to establish whether this is so.

In the mesosphere, where gravity waves become the dominant vehicle for momentum transport, the quasigeostrophic assumption breaks down. However, as will be discussed below, estimates of diffusivities in the mesosphere imply that here also diffusion (in this case, in the vertical) is a major component of constituent transport.

In order to represent transport processes satisfactorily, it is necessary to know their structure as well as their intensity. McIntyre and Palmer [1983, 1984] noted that quasi-horizontal mixing events in the northern winter stratosphere, revealed in potential vorticity maps of Figures 6-24 and 6-44, occur primarily in a mid-latitude "surf zone", associated with the deformation of material lines, e.g., in the vicinity of the Aleutian anticyclone. Studies of tracer diffusion in numerical models [Kohno, 1984; Mahlman, 1985; Plumb and Mahlman, 1986] have also revealed such a zone of maximum mixing in this region; the morphology of the process, discussed by Mahlman [1985] and illustrated in Figure 6-53 is, not surprisingly, similar to that revealed by McIntyre and Palmer's potential vorticity maps. An example of Plumb and Mahlman's results for the horizontal diffusion coefficient in the GFDL model is shown in Figure 6-54. The result that the diffusion is weak in high latitudes (where Eulerian measures of planetary wave amplitude such as geopotential height or northward velocity maximize) and strong in the subtropics (where Eulerian amplitudes are weak) illustrates the point made earlier about the importance of distinguishing between Eulerian and Lagrangian eddy statistics. The dispersion of material lines tends to be strongest where the mean zonal winds are weak; in fact, for small amplitude stationary waves, the northward displacement amplitude $|\eta|$ is related to the geopotential amplitude $|\phi'|$ by $|\eta| \cong |\phi'|/(f\bar{u})$. Therefore, although $|\phi'|$ maximizes at middle-to-high latitudes where the zonal winds are strong (cf. Figures 6-4 and 6-5), $|\phi|$ maximizes in the subtropics where $\bar{u}$ is weak. In fact the location of the $\bar{u} = 0$ surface, which is indicated on Figure 6-54, seems largely to determine the locations of strong horizontal diffusion in the GFDL model. This suggests that the transport processes are dominated by the quasi-stationary waves, although in reality the transports will be strongly modulated by the transients discussed in Section 6.1.4 (cf. Figure 6-11).

In both the studies of Kohno [1984] and Plumb and Mahlman [1986] the mixing was found to be weak in high latitudes. However, as discussed in Section 6.3, general circulation models apparently underrepresent high-latitude dynamical activity in the stratosphere; therefore it seems likely that transport is underestimated here. In a mechanistic model of a high-latitude warming event, Hsu [1980] demonstrated strong high-latitude dispersion of air parcels; synoptic potential vorticity and ozone maps at the time of such events appear to confirm substantial latitudinal transport at these times.

Of course, planetary wave mixing does not take place purely in the horizontal plane. Mahlman *et al.* [1981] and Tung [1982, 1984] suggested that, for the almost adiabatic conditions typical of much of the stratosphere on the time scales of interest, the mixing should occur along the isentropic surfaces. Mahlman [1985], however, has noted some caveats in this argument; Plumb and Mahlman [1986] in fact found the mixing to occur along directions a little steeper than the mean isentropes. Since the residual and transport circulations are equal if the mixing is isentropic [Holton, 1981; Plumb and Mahlman, 1986], this result explains the similarity between these circulations evident from Figures 6-48a and 6-51.

A schematic summary of our current view of zonally-averaged transport phenomena in the troposphere, stratosphere and mesosphere is shown in Figure 6-55. This figure is based on that of Kida [1983b] with modifications incorporating results of Plumb and Mahlman [1986] and the characteristics of mesospheric advection and diffusion described in Section 6.2. We shall here summarize the major characteristics of transport in the troposphere, stratosphere and mesosphere as we currently understand them; the important and complex issue of tropospherestratosphere exchange is considered separately in Chapter 5.

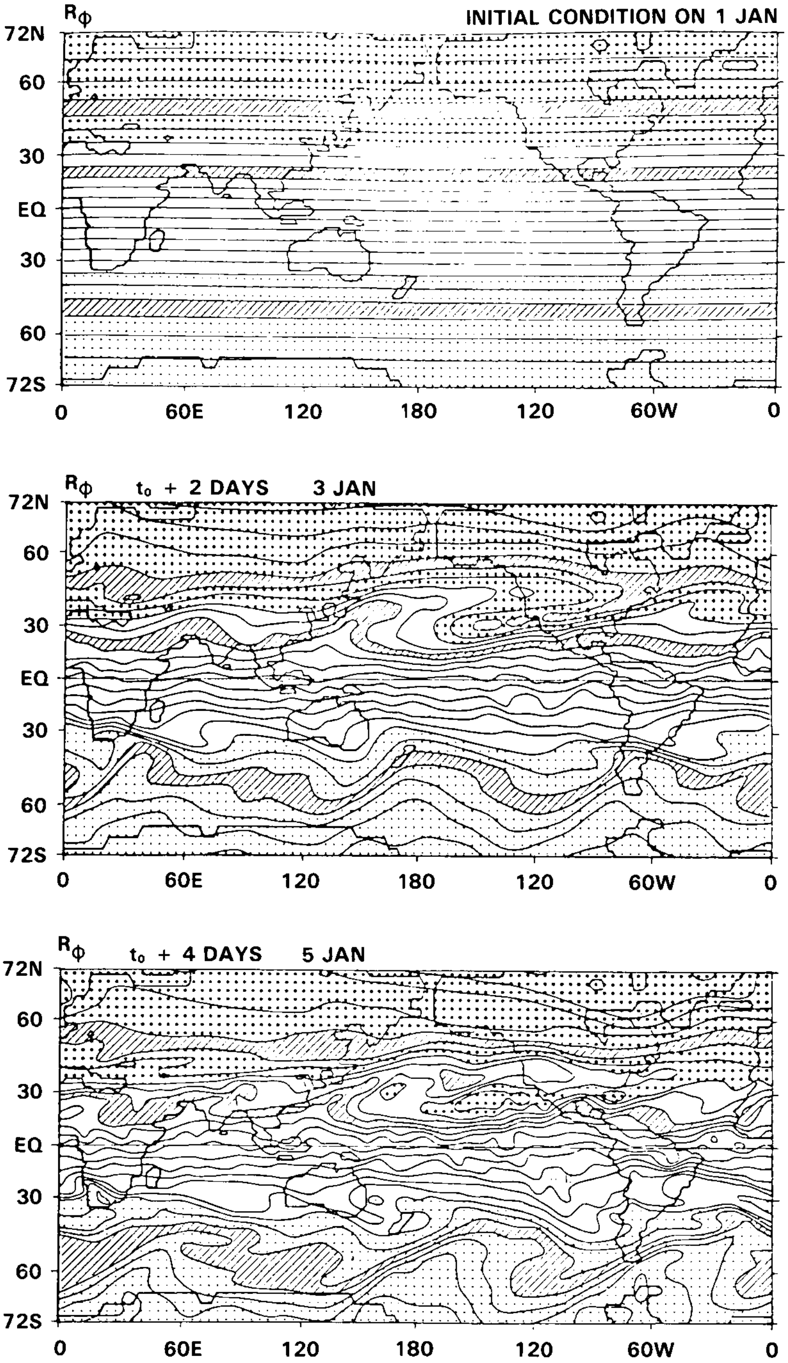

R_ϕ
INITIAL CONDITION ON 1 JAN
R_ϕ t_0 + 2 DAYS 3 JAN
R_ϕ t_0 + 4 DAYS 5 JAN
72N
60
30
EQ
30
60
72S
0
60E
120
180
120
60W
0

Figure 6-53. Isopleths of modeled evolution of mixing ratio on the 480 K isentropic surface of a conserved tracer initially (1 Jan) stratified uniformly in latitude. [After Mahlman, 1985].

Tropospheric transport is the most complex, with advection by the Hadley circulation, quasi-horizontal mixing associated with planetary and synoptic eddies and vertical convective mixing all significant contributors to large-scale transport. On the whole, transport time scales are relatively short - for example, Plumb and Mahlman's [1986] estimates for K_{zz} give time scales for vertical diffusion over the tropospheric

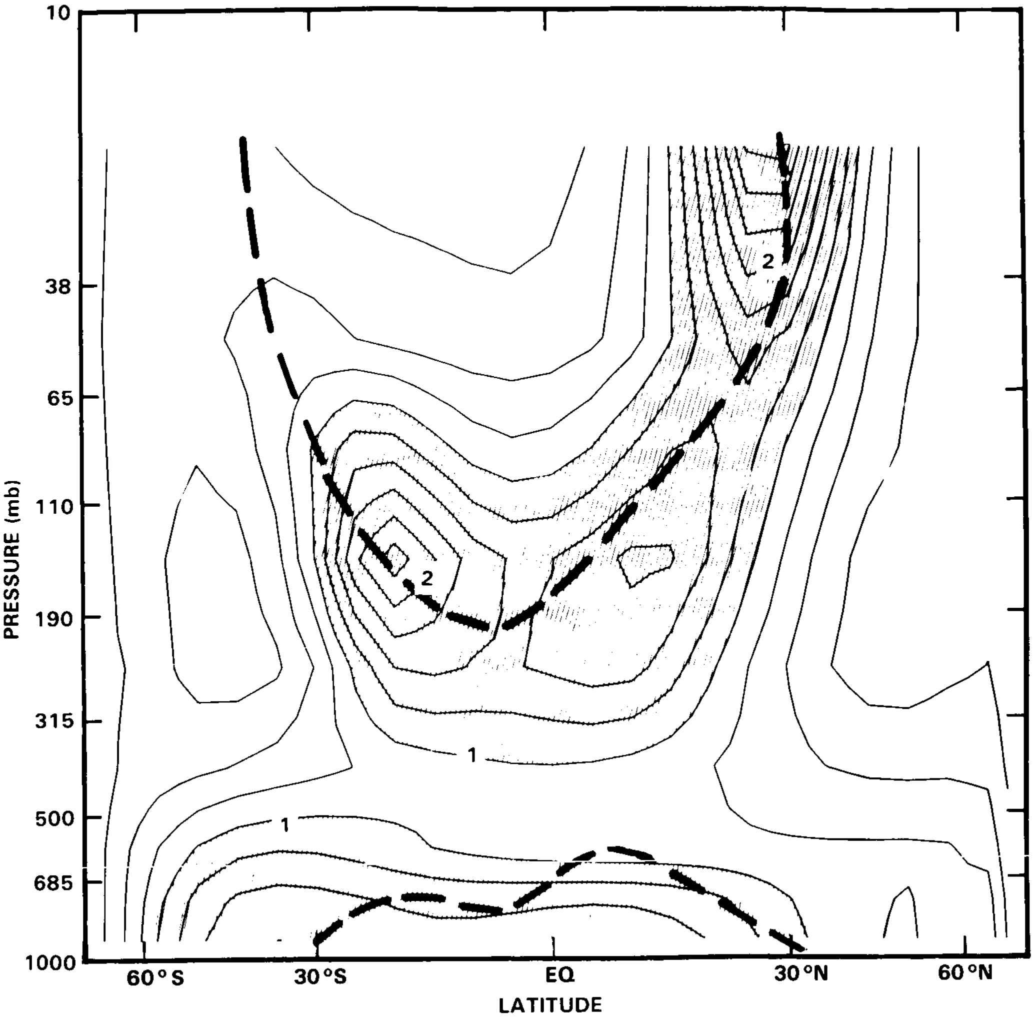


Figure 6-54. Calculated horizontal diffusivites (10^6 m^2 s^{-1}) for the GFDL general circulation/tracer model in January [after Plumb and Mahlman, 1986]. The dashed curve is the line of zero zonal-mean wind.

depth ranging from 10-30 days in the tropics to 30-100 days at middle and high latitudes - so that long-lived tracers with weak local sources and sinks will be well-mixed through the troposphere.

The stratospheric situation seems simpler, with transport largely dictated by the quasi-stationary eddies (via direct quasi-horizontal mixing and indirectly via the wave-driven meridional circulation). Vertical diffusion is relatively unimportant here. In Section 6.2 it was estimated that the contribution to K_{zz} from small-scale turbulence is at most 0.2 m^2 s^{-1}, and Kida [1983a] and Plumb and Mahlman [1986] found similar values for large-scale transport in their model studies; these values give vertical mixing times over one scale height of a few years or more, too long to be competitive with the processes noted above. (An independent upper limit on the vertical diffusion rate for momentum in the equatorial lower stratosphere comes from the existence of the QBO, which, as discussed by Plumb [1984], requires the momentum mixing timescale to exceed about 2 years). Kida [1983a] noted that the relative importance of advection and quasi-horizontal diffusion for any particular tracer depends on the meridional slope of the tracer isopleths; for slopes characteristic of long-lived tracers, both he and Plumb and Mahlman [1986] estimated the two

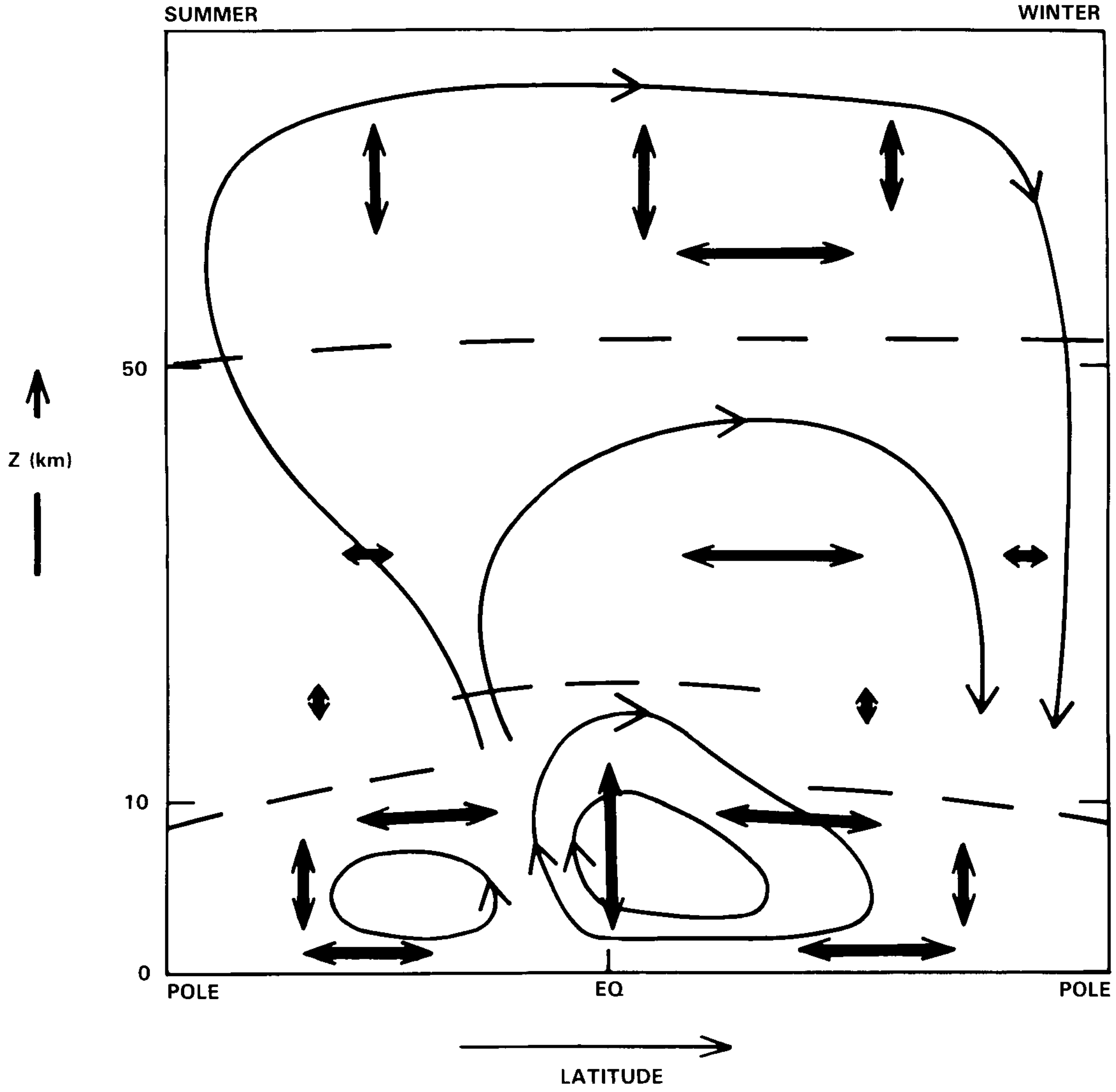


Figure 6-55. Schematic illustration of zonally-averaged transport processes up to the mesopause. Single arrows: mean circulation; double arrows: quasi-horizontal and vertical diffusion. See text for discussion.

to be of comparable importance. (Indeed, as noted above, it seems to be this balance that determines the isopleth slopes). Kida estimated a typical global value in the lower stratosphere of $K_{yy} \cong 3 \times 10^5$ m^2 s^{-1}, although, since his model stratosphere is dynamically less active than the observed stratosphere, this is likely to be an underestimate. Plumb and Mahlman's results, shown in Figure 6-54, are similar to this value in the summer hemisphere and in winter middle and high latitudes but much larger (up to 2×10^6 m^2 s^{-1}) in the "surf-zone" of the winter subtropics.

In the mesosphere, planetary wave amplitudes decrease and control of transport processes appears to be dominated by gravity waves, through vertical mixing and the strong wave-driven pole-to-pole circulation. These processes are very rapid, with a characteristic hemispheric advection time scale, given $V_T \cong 10$ m s^{-1}, of about 10 days and a similiar time scale for vertical mixing over a scale height, given $K_{zz} \cong 10^2$ m^2 s^{-1} (comparable with the values estimated in Section 6.1 for momentum mixing). However, one should note the suggestion of Chao and Schoeberl [1984] and Fritts and Dunkerton [1985] that the diffusivity associated with breaking gravity waves for heat and, by implication, long-lived constituents, may be much smaller than that for momentum. It might also be mentioned in passing that quasi-horizontal mixing processes may not be altogether negligible; in particular Craig *et al.* [1985] have noted the probable impact on tracer distributions of summertime pulses of "two-day wave" activity.

6.5.4 Some Outstanding Issues in 2D Modeling

Following the discussion in Section 6.5.2 it would be fair to conclude that a profound change has taken place over the past ten years in our conceptual formulation and understanding of zonally-averaged transport problems. More recently, this has begun to impact on the practice of 2-D modeling, in particular in models using the residual circulation, rather than the Eulerian mean circulation, as a basis [e.g., Holton, 1981; Garcia and Solomon, 1983; Ko *et al.*, 1985; Rogers and Pyle, 1984]. This circulation should be a good approximation to the transport circulation in the stratosphere, though less so in the troposphere [Plumb and Mahlman, 1986]. More recently still, we have seen objective assessments of transport processes and their magnitudes in numerical models and, as discussed in preceding sections of this chapter (particularly 6.4.4), we are beginning to understand the morphology of large-scale transport processes. Together, these advances are for the first time permitting real insight into the dynamics of these phenomena and have, amongst other things, led us to appreciate the practical importance of kinematic transport effects (i.e., large-scale turbulence), which in some recent literature has been played down in favor of the so-called "chemical eddy" terms.

While some of these GCM-based parameterizations are being applied in 2-D models [Enting, 1985; Pitari and Visconti, 1985; Plumb and Mahlman, 1986] it would seem desirable (in view of the shortcomings of GCMs discussed in Section 6.3) to obtain similarly objective representations of the actual atmospheric circulation. Deriving the full set of transport coefficients from first principles requires accurate information on horizontal *and* vertical winds or constituent fluxes - hence the dependence hitherto on GCM data, since accurate observations of vertical winds or fluxes on a global scale are just not available (nor are they likely to be in the forseeable future). However, some progress may be attainable on the basis of assumptions about the characteristics of eddy transport. The most promising such assumption is that the eddy motions are almost adiabatic [Mahlman *et al.*, 1981; Tung, 1982,1984]. Under these conditions, the transport circulation may be represented by the residual circulation [Holton, 1981; Plumb and Mahlman, 1986] - which can, at least at the solstices, be approximated by the diabatic circulation, which in turn may be determined from a radiation calculation - while the diffusion tensor collapses to a scalar diffusivity representing mixing along the isentropes, whose determination then only requires quasi-horizontal (isentropic) information on the flow or constituent fluxes. Thus the problem of deriving the transport coefficients from atmospheric data becomes practicable - though still difficult. A preliminary effort in this direction has been made by Newman *et al.* [1986] using satellite-based analyses to derive potential vorticity fluxes and thence to estimate the quasi-horizontal diffusivities. However, there are some problems with the calculation of potential vorticity fluxes in the stratosphere (this flux is proportional to the divergence of the EP flux; intercomparisons of this quantity were discussed in Section 6.4.2), and, at least in the upper stratosphere, potential vorticity may not be representative of long-lived tracers. Perhaps studies of long-lived trace constituents (satellite observations of which are now becoming available as discussed in Section 6.4) will provide alternative avenues.

These procedures - and, indeed, most of the body of the theory outlined in Section 6.5.2 - rely on a complete knowledge of a *given* (real or model) atmospheric circulation. It must be emphasized, of course, because of the relation between the advective and diffusive components and of the dependence of equilibrium tracer isopleth slopes on the balance between the two, that the two should be represented in a mutually consistent way. There appear to be two ways of achieving this; the first is simply to calculate the two from the same circulation data (e.g., the GCM-based calculations discussed above). A second approach is, given the diffusivities **K**, to calculate diagnostically the residual circulation corresponding to this eddy transport (via Equation (23) or via the coupled heat and momentum budgets as in, for example, the model of Garcia and Solomon [1983]). The latter approach, however, implicitly invokes the assumption that entropy and/or potential vorticity may be regarded as long-lived tracers and, as noted above, this assumption may be suspect above the middle stratosphere.

Another, less direct, way of estimating the diffusivities from the observed atmospheric behaviour is to tune 2-D models to reproduce the observed distributions of long-lived constituents [e.g., Ko *et al.*, 1985]. However, this procedure is not suitable for revealing more than the gross magnitudes of the diffusivities. Moreover, if the mean circulation used in the model is in error, similar errors will be produced in the calculated diffusivities, in order to give the "correct" balanced state.

The detailed formulation of 2-D models currently in use for assessment purposes is discussed in Chapter 12 of this document and a comparison of model assessments is presented in Chapter 13. Differences between the various models - in the model chemistries as well as transport formulations - are manifested in such comparisons; it is not easy to determine, however, to what extent the differences in representation of transport are responsible for differences in results. From a dynamical point of view, it would be desirable to compare the behavior of models with identical chemistries. The easiest way of achieving this is to bypass the model chemistries altogether by running an experiment with an inert tracer, spreading out from an initially localized source. An ideal case would be the "instantaneous midlatitude source" experiment of Mahlman and Moxim [1978] (cf. Plumb and Mahlman [1986]; not only would this allow comparison of the transport characteristics of the 2-D models with a 3-D model, but also with the observed behavior of the radioactive debris from the atmospheric bomb tests in the early 1960s. Such a comparison would highlight any major inadequacies in model transport formulations.

For practical as well as theoretical reasons, transport parameterizations are usually designed to represent monthly or seasonally averaged transport. Thus such representations are incapable of modeling transport variability on time scales less than about a month or longer than the annual cycle. We have seen in Section 6.1 that wave amplitudes and fluxes can fluctuate markedly on time scales as short as 1-2 weeks, especially (but not only) during winter warming events. In principle - given the required circulation data - the transport coefficients could be determined day by day to enable this variability to be incorporated into 2-D models. Plumb and Mahlman [1986] have argued, however, that 2-D models are not well posed on these time scales, and it may therefore be necessary to rely on three-dimensional studies [e.g. Rose and Brasseur, 1985].

On interannual time scales, 2-D models are in principle better suited to representing transport variability. The approach has been applied, with some success, to modeling the QBO in ozone, driven by the QBO in tropical winds and temperatures [Hasebe, 1984; Ling and London, 1985]. The more general question of the impact of year to year variations in the circulation of the winter middle atmosphere is more difficult to address. The problem is not so much theoretical - given an adequate (real or model-generated) multiyear data set, one could in principle generate transport coefficients month by month or season by season - as philosophical, viz., to define what it is that one is trying to model (a particular year or a "typical" year?). This issue, which is not confined to 2-D transport models, becomes particularly acute for model validation.

We are now beginning to appreciate the large interannual variability of wind and temperature fields in the winter stratosphere [Labitzke and Naujokat, 1983; Geller *et al.*, 1984; see Section 6.1.8]. The impact of this variability on constituent transport needs to be addressed; since the wind and temperature variability is presumably indicative of variability of the transport processes, this impact is likely to be major. A related question, since the frequency and intensity of warming events are variable from year to year. is how much of the "average" wintertime transport is associated with such events. If they account for a substantial fraction of the net transport, then the adequacy of current models (including most GCMs) is open to question.

A further issue arises in long-term assessment studies. If a model is being used to predict changes in the distribution of radiatively active constituents such as ozone, should the model be "interactive" in the sense that some attempt be made to allow the model to respond to the changing climate? This question can be answered in part from the discussion of Section 6.2. From this discussion and the results of Fels *et al.* [1980] described in Section 6.3, we have seen that both the distribution of mean diabatic heating $\bar{J}$ and the residual circulation $\bar{w}_*$ are, to a good approximation, outside the tropics, determined by the eddy transport. If the latter is fixed, then so are $\bar{J}$ and $\bar{w}_*$. Therefore, they are - to a first approximation - unaffected by changes in radiatively active constituents, as the results of Fels *et al.* [1980] clearly demonstrate. The mean temperature will, in general, change in order to maintain $\bar{J}$ as will the mean zonal wind, to maintain thermal wind balance with the temperature distribution. (A detailed assessment of the impact of changing trace gas distributions on climate is given in Chapter 15). It may well be important to allow these temperature changes to impact on model chemistry. However, changes in *transport* can only occur if the mean temperature and zonal wind distributions change so much as to impact significantly on the eddy motions themselves (via altered propagation or dissipation characteristics) and, of course, that is a three-dimensional problem. Therefore it is futile in a 2-D model to attempt to predict changes in the residual circulation, for example, rather than specifying it at the outset, simply because 2-D models are by their nature incapable of addressing situations where such effects are important.

6.5.5 Three-Dimensional Modeling

The recent advances in the theory and practice of 2-D transport modeling which have been reviewed above have brought the zonally-averaged model to the level of a sophisticated assessment tool. At the same time, however, fundamental limitations of the zonally averaged approach have become increasingly apparent. We must look to three-dimensional models to overcome these limitations. As we have seen in Section 6.3.3, however, even three-dimensional models are not without their limitations, particularly in regard to their deficient climatology in the winter stratosphere and the prohibitive cost of including full chemistry for extended integrations.

On short time scales, the use of such models for data assimilation in forecast mode offers an exciting avenue for exploiting the detailed, global observations of constituent distributions now available from satellites and described in Section 6.4. Thus, we can expect to learn much about mixing events on these time scales and, by implication, of long-term transport since this is to a large degree the aggregate of these events.

For seasonal time scales and longer, three-dimensional models with simplified chemistry have taught us much about large-scale transport processes, as well as providing data bases from which to derive parameterized transport formulations for 2-D models. In order for the models to reach their full usefulness, however, the deficiencies in GCM climatologies noted in Section 6.3.3 will need to be overcome.

The use of three-dimensional models (with realistic chemistry) for long-term assessment experiments appears to be out of the question in the forseeable future and there is therefore a continuing need for zonally-averaged models, despite their shortcomings. It is possible that some of these shortcomings could be overcome through the development of three-dimensional models that seek not to represent the transport associated with every eddy, in the way a GCM does, but rather to represent the aggregate effects of eddy transport in a similar way to a 2-D model, but relative to a slowly evolving, nonzonal basic state. This approach, which is similar to that suggested by McIntyre and Palmer [1983], would involve models with lower spatial and temporal resolution than GCMs and which would therefore be capable of incorporating more complex chemistry and of being integrated for longer times. However, developments in transport theory will be needed. While some progress is being made in the understanding of three-dimensional processes in tropospheric dynamics [e.g., Hoskins, 1983] we are still a long way from a simple understanding of transport relative to spatially nonuniform basic states.

6.6 FUTURE NEEDS

6.6.1 Satellite Observation

With the emergence of continuous global satellite measurements our picture of dynamical and transport processes in the middle atmosphere has sharpened considerably. They have made possible the identification of a number of transient waves hitherto unresolved, and have vividly exposed the complexity of transport as the westerly vortex evolves during disturbed conditions. Our knowledge of the morphology of a number of chemical constituents has also been advanced by these observations. Unfortunately with this improved focus on stratospheric behavior so too has emerged a recognition of the limitations posed by current observations and the need for even more sophisticated measurements. Perhaps the most striking demand of this sort stems from the possibility of planetary wave breaking, which may lead to a significant cascade to smaller scales. An irregular distribution of some quantity such as potential vorticity implies not only a high degree of spatial variance but also temporal variability through the advection of parcels by the flow field. Dynamical features such as the fast equatorial waves (Section 6.1.5) also place constraints on the quality of observations required. Such considerations appear to be increasingly significant at upper levels due to the tendency for the wave spectrum to be dominated by higher frequencies and the increasing amplitudes of gravity waves and tidal oscillations.

One of the chief difficulties in applying satellite data is dealing with their asynoptic nature. A ''transience error'' arises from the fact that measurements are not made simultaneously [Hartmann, 1976a], but rather are taken at different locations at different times. Their analysis therefore presumes some form of space-time interpolation. Several methods of estimating synoptic fields from asynoptic data have been employed [Rodgers, 1976b; Hirota, 1976]. These range in sophistication from presuming simultaneity over a day of measurements to four-dimensional assimilation in numerical forecast models.

The latter approach, although considerably more involved, has some advantages, such as the the ability to provide the ageostrophic velocity components. Whether these and higher order quantities are more of a reflection of the data or of the model, however, is unclear. Alternative approaches of deriving such quantities directly from temperature retrievals have also been proposed [Salby, 1982b]. Recently it has been shown [Salby, 1982b] that synoptic maps may be recovered uniquely from asynoptic data provided that the observed field is adequately sampled. The temporal scales that can be resolved in asynoptic measurements depend on whether single or combined node (ascending and/or descending) data are used. In the former case, frequencies up to 0.5 cpd can in principle be recovered, while in the latter synoptic irregularities at middle and high latitudes, intrinsic to asynoptic data, are explicitly accounted for [Salby, 1982a]. By taking into account these sampling irregularities of combined node data, Prata [1984] was able to obtain the wavenumber 2 component of the 4-day wave, previously unobserved. The aforementioned issues determine how frequently it is sensible to map asynoptic data.

Considerations of asynoptic sampling and resolution are more general than applying solely to dynamical quantities. They pertain equally to unsteady disturbances in any field, e.g., distributions of chemical species. The question of resolution then reduces to how much of the spatial and temporal variability is captured by the sampling. For example, localised features in potential vorticity or a constituent will not be resolved if their dimensions are comparable to the spacing of the data or if the feature moves appreciably over the time it takes the satellite to circle the globe.

Another important example arises in connection with diurnal variations, which may accompany tidal oscillations or a photochemically active species. Such features are not resolved by observations from a single sun-synchronous satellite because the orbit drifts westward at the same rate as the feature and thus views the same relative point on the feature with each latitude crossing. Global observation of such phenomena will require measurements from multiple satellites.

Many quantities central to dynamical and hence transport considerations, e.g., motion fields, potential vorticity, Eliassen-Palm flux divergence, are not measured directly but rather must be derived from observed temperature behavior. Such quantities are higher order in that they involve one or more derivatives of observed fields. The availability of direct velocity measurements such as will be made on UARS should be of great value in obtaining such quantities. Differentiation has the effect of increasing the spatial variability of the fields and eroding the signal to noise ratio, perhaps below useful values. Of course, fields such as potential vorticity are themselves of inherently richer structure than lower order quantities, making it difficult to distinguish legitimate features from observational error. Multiple satellite observations and scanning radiometers may be valuable in alleviating this problem. However the assimilation of such data will require that spatial and temporal irregularities of the combined sampling be accounted for, in order that the full information content of the data be recovered.

There also exists a need for refined vertical resolution. Rocketsonde and radiosonde measurements suggest that dramatic changes in temperature can occur within quite a thin layer of the stratosphere during sudden warmings. In order to capture such behavior in global analyses, it may be necessary to make use of different observing systems, both satellite- and ground-based. Rocketsonde measurements have historically been invaluable in validating remote temperature retrievals. Their waning in recent years presents a problem for future missions. As was noted above, one of the prime virtues of satellite observations is their homogeneity. The importance of long-term continuity with regard to the monitoring of trends and expanding our understanding of the complex realm of stratospheric behavior cannot be overstated. It is hoped that further remote sensing missions will be planned to succeed UARS following 1989.

Despite some recent examples to the contrary, most studies of the middle atmospheric circulation are still focused on midlatitudes of the Northern Hemisphere. What we have learned of the climatology and transient behavior of the Southern Hemisphere has confirmed and indeed strengthened the impression of significant differences between the two hemispheres during their respective winters, and a fuller exploitation of the contrast between these two regions will surely help to further our understanding of both. Southern Hemisphere analyses are currently undermined to some extent by the relatively poor quality of low level analyses upon which to build; this is another area where direct wind measurements (such as will be provided by UARS) will prove invaluable.

The middle atmosphere tropics have received even less attention than the Southern midlatitudes. To some extent this is a result of the weak temperature structure in low latitudes; again, direct wind measurements will be needed before much of the tropical circulation is visible in satellite observations. Equatorial wave motions, with their small vertical scales, have been revealed by limb-viewing instruments and there is a continuing need for such observations.

6.6.2 Ground Based Studies

The use of radars and more recently lidars has led to a significant improvement in our understanding of wave and turbulence processes in the middle atmosphere. Radars offer the capabilities of measuring important time-mean quantities such as $\overline{u}$, $\overline{v}$ and $\overline{w}$ as well as the eddy flux terms $\overline{u'^2}$, $\overline{u'w'}$ etc., while lidars give information on $\overline{\varrho}$, $\overline{T}$, T' etc. Although the usefulness in studying large-scale phenomena is limited by the local nature of radar measurements, they offer the advantage over satellite retrievals that they measure the velocity field directly. This feature makes them particularly attractive in the tropics where geostrophy breaks down. There they may serve as independent verification of velocities derived from remotely monitored temperatures.

If the potential of ground-based techniques is to be fully realised, then a number of deficiencies need to be rectified. At present many radars operate on an irregular basis and it is desirable that their operation be made continuous to make long-term measurements throughout as much of the middle atmosphere as possible. In particular, the possible influence of gravity waves on the stratospheric circulation (see Section 6.2.4) needs to be investigated. Coordination of observations is also important; this is an issue being partially addressed by projects organized under the auspices of the Middle Atmosphere Program.

At present most radars are located in continental areas where the effect of topographically forced waves may be significant. In order to assess the global role of gravity waves it is therefore important that radar measurements be made in oceanic regions. Observations in equatorial regions are almost non-existent and the establishment of radars near the equator is crucial for the study of gravity and tropical waves. The predicted breakdown of the diurnal tide in the equatorial mesosphere has yet to be investigated experimentally.

The construction of radar networks will help with the identification of wave sources and with the measurement of important wave parameters such as horizontal phase velocities and wavelengths. For example, a valuable application would be the construction of an array of sounders along the equator downfield from a convective center such as Indonesia. Cross-spectral analyses could then be performed on the combined array of velocity profiles to derive structural and temporal behavior of the motion field. This would facilitate the analysis of particular wavenumber-frequency bands so that individual components (e.g., Kelvin waves) could be discriminated for.

The widespread use of lidar investigations of density and temperature is to be encouraged. As well as providing complementary information to the radars, lidars give information in the 30 to 60 km height range (the so-called 'gap'). By colocating radars and lidars it will be possible for the first time to measure heat fluxes, albeit locally, in the mesosphere.

Finally, it is noted that the more widespread use of lidar measurements of temperature and density may help to compensate for the decreasing number of rocket measurements due to the reduction in the meteorological rocket network. This would be especially the case if lidar techniques can be extended to wind measurements.

6.6.3 Dynamical Theory

Section 6.2 has discussed the importance of eddy motions – especially planetary waves in the winter stratosphere and gravity waves in the mesosphere – in maintaining the climatological state of the middle atmosphere. It has also been emphasized that it is misleading to regard the residual circulation as independent of the eddy forcing or as a ''diabatic circulation'' *driven* by an externally-imposed net radiative heating. Some general questions in this area which will require particular attention in the near future include the following:

(a) Planetary-Wave Breaking

McIntyre and Palmer [1983, 1984] have suggested that breaking planetary waves (Section 6.2.4) may bring about substantial mixing of potential vorticity in the stratosphere and may thereby be responsible for the region of weak potential vorticity gradient (which they call the ''surf zone'') often observed to surround the main cyclonic vortex in the northern winter mid-stratosphere. A combination of theoretical and observational studies will be required to confirm this mixing hypothesis and to assess the impact of diabatic processes on the breaking-wave phenomenon. Indeed, improved understanding of the relative roles of mixing and diabatic processes in net potential vorticity transport is important not only to our understanding of the dynamical structure of the middle atmosphere, but also to our conceptual picture of stratospheric transport. The feedback effects of the wave-induced mean flow changes on the wave propagation characteristics also need further study. For example it has been speculated [McIntyre, 1982; McIntyre and Palmer, 1983, 1984] that the weakened potential vorticity gradients in the subtropics may inhibit the meridional propagation of planetary waves and thus form ''resonant cavities'' which may enhance the local growth of the wave amplitudes. A related requirement is for a careful combination of theoretical and observational studies of sudden warmings and a clearer understanding of the relationship between the ''zonal-mean, eddy'' approach and the ''synoptic map'' approach mentioned in Section 6.2.5.

(b) Gravity Wave Drag

An increased understanding of the basic mechanics of gravity-wave breaking is needed for the development of improved parameterizations of the effects of such breaking on the zonal-mean flow and on planetary waves. For example, more account may need to be taken of the effects on the parameterizations of including a broad-band spectrum of gravity waves, rather than the small discrete sets of wavenumbers and phase speeds that have mostly been used hitherto. The relative magnitudes of the gravity-wave induced diffusivities of momentum, heat and constituents [Chao and Schoeberl, 1984; Fritts and Dunkerton, 1985] also need to be investigated.

It is particularly important that the role of gravity waves in the stratosphere be better understood. The ''cold pole'' problem of middle atmosphere GCMs is suggestive of weak wave drag in these models, and one process that could be lacking in these models is gravity wave drag.

(c) General Theory

There is a need for a basic theoretical framework, perhaps analogous to the ''zonal-mean, eddy'' framework, but more suitable for the organization and interpretation of data concerning strongly zonally-asymmetric planetary-scale disturbances in the middle atmosphere.

6.6.4 General Circulation Models

While there have been several notable successes that have been associated with middle atmosphere GCMs, considerable caution should be exercised in carrying over GCM-derived results to the actual atmosphere. This is so because there are several well known deficiencies in middle atmosphere GCM simulations. These include the cold winter pole/excessive westerlies problem (and the associated problem that middle atmosphere GCMs do not produce stratospheric warming episodes of sufficient intensity). From these deficiencies, we conclude that existing middle atmosphere GCMs are probably deficient in eddy transport effects and in the strength of their residual mean circulation. Thus, we expect that 2-D transport parameters calculated from GCMs will be too small compared with those that are representative of the actual atmosphere. Furthermore, since the derivation of those transport parameters formally depends on the assumptions of small amplitude theory, it is possible that 2-D representations of 3-D dynamical transport may be less applicable and work less well in the real atmosphere than in GCMs.

The wintertime "cold pole" problem is perhaps the most serious deficiency of current middle atmosphere GCMs. As we have seen, dynamical theory ascribes this problem to an under-representation in the models of wave drag on the flow. This could be a result of the forcing of planetary waves in the model being too weak because of inadequate representation of orography or an inadequate parameterization of convective heating in the troposphere, or because of the importance in the real world of other wave motions which are not adequately resolved in the models. Indeed, the underlying problem - one which the shortcomings of GCMs may be helping to solve - is our incomplete understanding of the momentum budget of the middle atmosphere. As discussed in Section 6.4, several studies have highlighted the difficulty in balancing the momentum budget from analyses of the observed circulation. It is not clear whether these results indicate data inaccuracies on the large scale or whether small-scale, unresolved motions are making a significant contribution. The crucial role of gravity wave drag in the mesosphere is now appreciated and it is essential that its role, if any, in the stratosphere be fully clarified. Improved understanding of this process is a prerequisite to improved parameterization in models.

As mentioned previously, the application of GCM-derived parameters to the actual atmosphere is limited by the fact that middle atmosphere GCMs show some very significant modeling deficiencies. There is an alternative to using a GCM for middle atmosphere studies. This is the forecast/analysis sense in which available data is continually inserted into the GCM, which is then used as an analysis tool to produce regular gridded analyses of both observed and unobserved variables. That is to say, the GCM governing equations are used to both provide both continuity in the observables and to derive such unobservables as the ageostrophic wind components. Thus, using the output from GCM forecast/analysis will give the balanced dynamic fields that are needed for transport studies but are consistent with observations. It must not be overlooked, however, that results of such procedures will be model-dependent and it is important that their robustness be assessed.

Several groups in the world are now attempting to perform transport-chemistry studies using GCMs. Most of this work has been aimed at understanding the dynamical and chemical processes that maintain observed species distributions. Complex chemistry schemes have not been used in GCMs for several reasons. One is the great expense and complexity in doing so. Another is related to the problems that GCMs have in reproducing the observed atmospheric structure. The use of GCMs to forecast multidecade ozone scenarios is not envisaged in the near future.

6.6.5 Transport Theory and Modeling

Theoretical developments over the past decade have *in principle* provided a stronger theoretical basis for the parameterization of eddy transport of trace constituents in 2-D (zonally averaged) models, although it must be emphasized that the approach rests on the basic assumption that departures from zonal symmetry are small. One important lesson to come out of the basic theoretical considerations discussed in Sections 6.2 and 6.5 is the interdependence of advective and diffusive transport processes, and the fact that these should therefore be represented in transport models in a mutually consistent way. It is also now recognised that mixing processes are spatially inhomogeneous - most dramatically illustrated by the stratospheric ''surf zone'' - and that models using latitudinally constant diffusivities may not simulate correct latitudinal constituent structures.

The practical problem of determination of appropriate transport coefficients for a given flow climatology is still not entirely satisfactory. Those derived from general circulation models have the attribute of being self-consistent but of course these can reflect the properties of the real atmosphere no better than the models themselves, and we have seen that at the present state of the art these have serious shortcomings, especially in the representation of dynamical activity in the stratosphere. It is possible to estimate the residual (ageostrophic) circulation diagnostically from the heat budget as done (implicitly) by Murgatroyd and Singleton [1961] and indeed, their results are currently used in some 2-D models [e.g., Miller *et al.*, 1981; Guthrie *et al.*, 1984]; given the advances in our knowledge of the climatology of the middle atmosphere since this calculation, it would seem desirable to update this calculation.

Estimation of the diffusivities from atmospheric circulation data is more difficult since reliable ageostrophic wind data is not generally available. Newman *et al.* [1986] have applied Plumb and Mahlman's [1986] technique to estimate K_{yy} via inversion of a flux-gradient relation for quasi-geostrophic potential vorticity, although this approach breaks down in low latitudes. With the rapidly improving coverage of trace constituent distributions, it may become practicable to estimate gross mixing rates from observed global fields, or by tuning 2-D models to match observations [e.g., Ko *et al.*, 1984] although the success of this approach depends on having a good model of the transport circulation (since the modeled constituent fields will represent a balance between advection and diffusion).

Of particular interest, since GCM approaches have been unable to address the matter, is an assessment of transport characteristics during warming events in winter high latitudes. Indeed a basic question is what proportion of the net wintertime transport is associated with these events. Ozone records and, indeed, the observed temperature increases suggest that the effects are significant. Therefore incorporation of such phenomena into transport models appears desirable. However, given the extreme convolution of streamlines during such events (Figure 6-23), one must question the ability of 2-D models to represent them to any useful degree; even if such models can predict the zonal mean adequately, the relevance of zonal mean predictions in such cases, at least on the time scale of a single event, is not clear.

Perhaps to a lesser extent the same is true throughout the winter and it is to be hoped that advances in 3-D modeling, either via full GCM treatments or in low resolution models, will enhance our ability to model more meaningfully the structure of trace constituents, especially in the winter stratosphere. We can also look forward to an increased understanding of 3-D transport processes via analyses of the behavior of global constituent fields from observations and in three-dimensional numerical models run in forecast mode.

6.7 SUMMARY

6.7.1 Observations of the Middle Atmosphere

The advent of global satellite monitoring of the middle atmosphere has had a profound impact on our appreciation of the structure and dynamics of the region. In recent years we have seen the emergence of multiyear, global, satellite-based climatologies of the stratosphere and mesosphere and, correspondingly, recognition of the substantial interannual variability of the circulation. The development of a Southern Hemisphere climatology has, of course, been very much dependent on satellite observations and is of particular importance, as the contrast between the two hemispheres (especially in winter) is substantial and presents a challenge to dynamical theories. The observed differences between the two hemispheres have provided valuable input to theoretical developments, but have yet to be exploited to the full, and our understanding of the Southern Hemisphere middle atmosphere still lags behind that of the Northern Hemisphere.

Analyses of transient events have also advanced remarkably, partly because of improved data coverage in time and space and partly because of the application of more sophisticated analysis procedures. Thus a number of transient planetary wave modes have been identified, several of which can be associated with theoretically-predicted normal mode oscillations. The improved vertical resolution of limb-viewing instruments has permitted, for the first time from satellites, identification of equatorial waves in the stratosphere and mesosphere and, indeed, of ''ultrafast'' Kelvin waves, which were previously unobserved in the atmosphere (but which had been found in a general circulation model). As yet, however, there is no satellite-based observation of the mixed Rossby-gravity waves, which have been postulated to provide the driving for the easterly phase of the Quasi-Biennial Oscillation (although it has been suggested that this driving could be provided by other means).

Sudden warming events continue to be a source of dynamical interest. Although progressively more events are being monitored by satellite, much of our understanding is based on analysis of the event of February 1979. While it is understandable that this well-observed event (it occurred during the FGGE year) should receive considerable attention, there are dangers in ascribing too much significance to a single case. However other occurrences are also being studied, including the final warming in the Southern Hemisphere. The improved data base, together with theoretical developments, have led to an improving description of warmings but we are still far short of a complete understanding of the dynamical processes involved.

Diagnostic techniques have advanced in parallel with improving data coverage. In particular, evaluation of the Eliassen-Palm flux (both in observed data sets and in numerical models) has proved an illuminating means of elucidating the meridional propagation of stratospheric planetary waves and, through transformed Eulerian mean theory, the interaction of these waves with the zonal mean flow. Thus, the switching of wave propagation from equatorward to poleward prior to sudden warmings, with the associated tendency to deceleration of the high latitude westerlies, has been claimed to be an important precursor of such events. However, it is becoming increasingly apparent that the stratospheric flow is highly three-dimensional (especially during warming events) and that the ''zonal-mean, eddy'' separation inherent in such approaches may not always be appropriate. One more generally applicable technique which has aroused considerable recent interest is that of mapping Ertel's potential vorticity (EPV) on isentropic surfaces. EPV is a function of dynamical variables (importantly, most dynamical quantities of interest can be recovered from the EPV distribution) which acts as a tracer for adiabatic, frictionless flow. This fact has been known to theoreticians for many years, but it is only in the last few years that it has come to be applied to the large-scale stratospheric circulation. Despite the demands this technique places on data quality (the second horizontal derivative of satellite radiance data is required) its use has led to the identification of the "breaking" of planetary waves in the subtropical winter stratosphere, where the deformation field associated with the wave motion acts to distort, perhaps irreversibly, the material surfaces mapped by the EPV contours.

This new insight into large-scale stratospheric transport has been complemented by the recent availability of satellite observations of the global distributions of a number of trace constituents. Features similar to those evident in EPV maps have been observed in the distribution of ozone and water vapor. In this respect, as in others, exploitation of the wealth of new material which has become available with these new observations promises to enhance profoundly our understanding of the distribution of middle atmosphere constituents and of the transport processes which influence them.

Not all motions of dynamical interest in the middle atmosphere have been, nor are they likely to be, observed by satellite-borne instruments. Ground-based radar and, to a lesser extent, rocket and lidar measurements provide the only means of observing gravity waves in the region, as well as yielding other information currently unattainable from satellites such as ageostrophic wind data. Gravity wave observations are of particular importance, in view of their role in driving the mesospheric circulation. This role has recently been quantified (in support of theoretical predictions) following the development of a technique to measure gravity wave momentum fluxes using a twin-beam radar. In this application, as in many other respects, interpretation of the results of such experiments is clouded by the question of the global relevance of observations made at a single site. While a global synoptic network of ground-based observing facilities is inconceivable, it is desirable that a more representative coverage be achieved.

Ground-based observations play another crucial role, viz., as "ground truth" for calibration of satellite measurements. This aspect is important in maintaining continuity between successive satellite instruments, and especially so if satellites are to be used to detect long-term trends such as those which may occur in response to trends in constituent concentrations. It is therefore essential that ground-based networks, such as they are, be maintained and even augmented (thus reversing recent trends).

Satellite observations have become such an important source of information on the circulation of the middle atmosphere that it goes without saying that their continuation is crucial to the progress of the subject. Continuity of monitoring is particularly important; ideally, this requires not only that each monitoring satellite be replaced before the end of its useful life, but also that both old and new instruments operate simultaneously for some time, in order that intercalibrations may be carried out. Most of our routine data has come, and will continue to come, from nadir-viewing radiometers. However, the value of limb-viewing instruments has been noted above and elsewhere in this chapter; it is highly desirable that such instruments should operate on a continuous basis. Finally, the potential value of direct wind velocity measurements cannot be overstressed. Such data would be profoundly valuable, particularly in the tropics, where the temperature signal is weak and the derivation of wind from temperature is unsound, and in the Southern Hemisphere, where the low-level analysis required to build up the wind field from temperature retrievals is of poor quality.

6.7.2 General Circulation Modeling of the Middle Atmosphere

The application of general circulation models (GCMs) to the middle atmosphere has so far met with mixed success. Many key aspects of the observed dynamical behavior of the region have been reproduced successfully (at least qualitatively) in these models and their use in both climatological (long-term) and forecast (short term) integrations, in transport experiments and as experimental tools for the conduct of perturbation experiments has contributed greatly to recent advances in our understanding of the middle atmosphere circulation. However, GCM results still exhibit serious deficiencies. The most serious of these concerns the climatology of the winter stratosphere, where the models consistently predict a polar night that is much colder (and therefore closer to radiative equilibrium) than that observed. While it has been shown that this problem can be alleviated in the lower stratosphere through improved representations of radiation, it seems clear that the more serious and more robust error in the middle stratosphere and above reflects inadequacies in the representation of dynamical transports. Other shortcomings of current GCMs include a failure to generate a Quasi-Biennial Oscillation in the equatorial stratosphere, errors in predicted tropical tropopause temperatures and an inability to resolve internal gravity waves. Since these latter motions are a crucial component of the mesospheric circulation, the fact that they must be parameterized in GCMs is rather unsatisfactory.

Successes of GCM simulations have included several examples of forecasts of stratospheric warmings. While these forecasts are only useful for a limited period (because of climate drift associated with the models' "cold pole" problem) their success has presented us with an additional tool for investigations of such phenomena. GCMs have also been used as vehicles for the conduct of perturbation experiments, for the investigation of large-scale transport processes and, in the absence of an adequate observational data set, as a source of data for the derivation of transport coefficients required by two-dimensional transport models. While what is learned from these exercises is very much model-dependent, there are good reasons to believe that the insight thereby gained is meaningful. It should also be noted that even the failures of GCM simulations have been of considerable benefit in guiding research priorities. As the GCM is the most complete synthesis of our quantitative understanding, the failure of such a model to represent correctly the observed atmospheric state (in circumstances where factors such as resolution are considered satisfactory) is an indication of the inadequacy of that understanding. Thus, identification of the causes of GCM errors is of considerable scientific value; for example, the "cold pole" problem has led us to question our understanding of planetary wave generation processes and the role of gravity waves in the stratospheric momentum budget.

Despite the availability of ever-faster computers, we have still not reached the stage where it is possible to incorporate a comprehensive chemistry into middle atmosphere GCMs in order to undertake long-term assessment experiments. Moreover, the climatological inadequacies of current models would in any case limit the usefulness of such assessments. Nevertheless, GCM experiments with highly simplified chemistry have proved valuable in complementing other approaches in the investigation of global transport processes in the middle atmosphere. In the forseeable future, however, assessment studies will continue to rely on the use of simplified transport models.

6.7.3 Theory of Dynamics and Transport

Over the past eight years or so we have seen a considerable improvement in our conceptual picture of the dynamics of the middle atmosphere circulation and of the transport processes which maintain it. It has become recognized that the "Brewer-Dobson" picture of an equator-to-pole circulation in the lower stratosphere and a summer-pole-to-winter-pole circulation in the upper stratosphere and mesosphere is, in a dynamical and transport sense, a more meaningful description of the mean meridional circulation than that obtained by simple zonal averaging of the local meridional winds. The former was originally inferred from observed distributions of ozone and water vapor in the stratosphere and we now know that the transport of constituents such as these is not well depicted by the latter, Eulerian mean, representation but is described more simply in terms of what has here been called the "transport circulation" which, in the middle atmosphere at least, is similar to the now more familiar ''residual'' and ''diabatic'' circulations. The Brewer-Dobson circulation is therefore an estimate of the transport circulation, which is what we need to know for understanding the global transport of constituents, and which is turn is a close approximation to the residual circulation; the latter is the appropriate measure of the mean circulation for dynamical investigations via transformed Eulerian-mean theory.

A description of meridional exchange in the middle atmosphere solely in these terms, however, is incomplete. The existence of this meridional circulation has in the literature (and in textbooks) frequently been described as a simple response to diabatic heating. This view is, however, logically unsound since the departures of the atmosphere from radiative equilibrium which give rise to this heating must be maintained by dynamical effects, in fact by eddy momentum transport (as expressed by the divergence of the Eliassen-Palm flux). This theoretical result has been in the literature for many years now, but its significance has only recently become widely appreciated. For one thing, it offers a simple diagnostic approach to the determination of where, and to what degree, these eddy processes act; thus it can be inferred from the Brewer-Dobson circulation that such processes must be acting in the winter stratosphere and in the mesosphere.

It is now widely accepted that the mesospheric driving is provided largely by internal gravity waves, with perhaps some contribution from planetary waves in the winter hemisphere and from tides in the tropics. These waves, propagating upward from the troposphere, reach large amplitude in the mesosphere, where they may break and thus act as *in situ* forcing on the mean flow. Much theoretical effort has been directed recently at understanding this phenomenon and its parameterization in numerical models. However, apart from a measurement of the magnitude of the effect by radar supporting this general picture, these developments have proceeded largely unconstrained by experimental evidence and there remains a need for a deeper understanding of the properties of atmospheric gravity waves and of the breaking process.

In the winter stratosphere, quasi-stationary planetary waves are the most likely driving mechanism. It has until recently been supposed that this interaction must derive from the dissipation of these waves by radiative effects, although this view presented problems in the lower stratosphere where such effects are weak. However it is now realised that planetary wave breaking, a process that has been identified from the observed behavior of potential vorticity maps, may be an important factor (and perhaps the dominant one, at least in the lower stratosphere). Thus, strong eddy mixing processes in the ''surf zone'' of the winter subtropics may be a powerful influence on the dynamical structure and evolution of the winter stratosphere. However, a number of crucial issues need clarification, including the efficiency of this mixing and the relative contribution of radiative dissipation; the latter question is particularly important for our conceptual picture of constituent transport.

During high-latitude warming events the location of this planetary wave driving switches from the subtropics to the polar cap, apparently as a consequence of the focusing of waves into that region. The reasons for this focusing and for the characteristic amplification of planetary wave activity at such times are not well understood, although it has been speculated that both these effects may be a consequence of the changes in the dynamical structure of the stratosphere brought about by previous low-latitude mixing processes.

Another outstanding issue - one that is raised by the apparent failure of current GCMs to generate sufficient eddy driving in the winter stratosphere as well as observational indications of a deficit in the large-scale stratospheric momentum budget - is the role of gravity waves in driving the stratospheric circulation. At present there is little consensus amongst theoreticians on this issue and little or no relevant observational evidence. It is to be hoped that stratospheric observations of gravity waves (by MST radars, for example) will be forthcoming to help resolve this question.

Recognition of the existence of large-scale mixing (i.e. quasi-horizontal turbulence) in the stratosphere has had an impact on our conceptual picture of zonally-averaged constituent transport. Until recently, this picture comprised advection by the Brewer-Dobson circulation, with some diffusion superimposed on this for nonconservative consistuents (the so-called ''chemical eddy'' contribution). Large-scale mixing events will also impact (as an effective diffusion) on even conserved constituents. Indeed, if the eddy momentum transport is dominated by this mixing (rather than the effects of radiative dissipation), then it can be argued that advective and diffusive transport of constituents must be formally comparable, so that an adequate representation of both is necessary in two-dimensional transport models. More generally, recognition of the central role of eddy transports in driving the mean meridional circulation implies that such models are incapable of being ''interactive'' in the sense of predicting meridional circulation changes.