The vertical structure of stratospheric planetary waves and its variability: Theory and observations.

by

Nili Harnik

B.A., Tel-Aviv University
(1993)

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2000

Signature of Author .....
Department of Earth, Atmospheric, and Planetary Sciences
16 September, 1999

Certified by .....
Richard S. Lindzen
Alfred P. Sloan Professor of Meteorology
Thesis Supervisor

Accepted by .....
Ronald G. Prinn
Chairman, Department of Earth, Atmospheric, and Planetary Sciences

Lindzen

The vertical structure of stratospheric planetary waves and its variability: Theory and observations.

by

Nili Harnik

B.A., Tel-Aviv University
(1993)

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences
on 16 September, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Observations of the vertical structure of stratospheric planetary waves reveal a large variety of structures, and a variability both on seasonal and on daily time scales. The extent to which linear wave theory explains these structures and their time evolution at a given time or season is not well known. The sensitivity of linear wave models to details of the basic state and model damping, both of which are not determined from observations in great accuracy, makes it hard to determine why the observations deviate from modeled waves in any given case. In addition, the ability of the observations to resolve the vertical structure of planetary waves is not obvious, given the low vertical resolution of satellite retrievals. The goal of this thesis is to understand the sources of observed variability of vertical wave structure, in particular, to determine whether linear wave theory can explain this variability, and whether the observations are capable of resolving it.

We start by testing the ability of satellite retrievals to resolve the vertical structure of the waves. We calculate the radiances that a virtual satellite sitting at the top of our model atmosphere would see, and invert them to obtain retrieved temperature fields. The comparison to the model temperatures suggests that the retrievals are able to resolve their general features quite well, with a few exceptions. Above 1.5mb there is little observed information in the retrievals, and errors start growing above 5 mb. Also, small scale features are not resolvable, but most waves have large enough vertical wavelengths to be resolved. We also identify dynamic situations in the real atmosphere which are more prone to retrieval errors. These are mostly relevant to summer or to the breakup of the polar vortex, when the existence of critical surfaces may cause the waves to have sharp features.

The next part consists of understanding the relation between vertical wave structure and the wave propagation characteristics of the basic state in a series of linear wave models, both steady state and time dependent. We study the normal modes

on a one dimensional (vertical) troposphere-stratosphere system, using a framework of wave geometry, which allows us to generalize the results to many basic states. A large variety of vertical wave structures is found, similar to observed. This variety is due to the existence of stratospheric turning points. We extend these results to basic states that vary in latitude and height in a nonseparable way. The main problem is how to separate the wave propagation into the vertical and meridional directions. Our approach is diagnostic, where we calculate meridional and vertical wavenumbers from the steady state wave solution to a given basic state, and use them as a diagnostic of the basic state wave propagation characteristics. In particular, we are able to determine the location of turning surfaces for meridional and vertical propagation. By applying this wavenumber diagnostic to many model runs we show that the existence of a stratospheric waveguide renders the problem qualitatively one dimensional by determining the meridional wavenumber, regardless of the characteristics of the tropospheric forcing. In particular, the effects of damping and turning surfaces on the vertical structure are qualitatively as in the vertical propagation problem.

In a complementary study, we regard the waves as consisting of many wave activity packets that propagate from the troposphere through the stratosphere, until they dissipate. A technique that follows a wave packet on its journey through the stratosphere, while keeping track of variations in wave activity that are due to refraction of the waves, is introduced and applied to the model runs. This allows us to separate between the contributions to the wave activity budget of damping, refraction, and time variations in the source of wave activity. Also, we can estimate the time scale for vertical propagation through the stratosphere of specific wave events.

Finally, we use our diagnostics to study observed wave episodes. We show that the differences in vertical wave structure between middle and late winter episodes in the southern hemisphere can be explained as a linear response to the seasonal evolution of the basic state wave propagation characteristics. We also show that the occasional daily time scale variations of vertical wave structure within a given wave episode are qualitatively a linear response to time variations of the basic state wave propagation characteristics. Since the basic state variations are wave driven, the relevant theory is quasi-linear. Estimates of wave propagation time scales, obtained using our wave activity diagnostic, are also consistent with the theory. We take this is a qualitative assessment of the applicability of quasi-linear wave propagation theory on daily time scales, as well as an assessment of the observations of the waves and the basic state. The latter is not obvious since most of the relevant variations in the basic state occur above 5mb, where observations are less accurate.

Thesis Supervisor: Richard S. Lindzen
Title: Alfred P. Sloan Professor of Meteorology

To my grandmother, Ibi Harnik,
whom I love very much.

Acknowledgments

There are many people who have helped me through my years in MIT. I have been fortunate to receive support not only professionally but also morally through the encouragement and friendship of the many people who went through this process with me.

First and foremost, I would like to thank my advisor, Richard Lindzen. My discussions with Dick have been a source of inspiration throughout my time here, and his insight of the workings of nature and the atmosphere as well as his general approach to research will guide me in the future. Dick patiently led me, and stood by as my view clarified and focused, occasionally making extremely useful suggestions.

I would like to thank Peter Stone, my first advisor, for his guidance through my early days in MIT and for enabling me to pursue my own interests. I would also like to thank the rest of my committee members: Alan Plumb, Kerry Emanuel and Edmund Chang for their time, support and helpful comments. I especially thank Alan Plumb for many useful discussions, and for including me in some of his group meetings. I am obliged to all my professors and teachers in MIT whose courses I participated in and whose instruction served as the invaluable foundation of this thesis. In particular, I thank Ed Lorenz for guiding me in a fun and interesting reading course on Chaos.

The Chemistry and Dynamics Branch (code 916) at NASA GSFC have provided me with the observational data which made my research possible. I thank Paul Newman, Larry Coy and Eric Nash, who acquainted me with their data system and answered my numerous questions.

I have received additional assistance in the research which served the basis for the Satellite Retrievals Chapter. Thomas Kleespies from NESDIS supplied me with the OPTRAN code used for calculating the NOAA 14 HIRS and MSU transmittance and weighting functions. Larry McMillin from NESDIS explained how current operational retrievals are obtained. Philip Rosenkranz from the Research Lab of Electronics at MIT answered my questions regarding satellite retrieval techniques and assured me of the correctness of my methods. Laurie Rokke from the DAO at NASA/GSFC kindly provided helpful information about the SSU instrument and about the operational retrievals.

There are many whose dedicated work and warm manner made my years at MIT more pleasant. I am indebted to Jane McNabb, Mary Elliff, Tracey Stanelun, Stacey Frangos and the rest of the administrative staff at EAPS who assisted me over the years, cared and endeavored to make my life at MIT as hassle-free as possible. Joel Sloman spiced up many days by sharing some of his poetic and artistic thoughts.

Repetitive Strain Injury has made my long days at the computer somewhat uncomfortable. I am therefore grateful to Lodovica Illari who in addition to a wonderful lab experience took it upon herself to provide me and my fellow students with the appropriate computer setup. I also thank the computer support staff: Linda Meinke, Tom

Yates, and Michael Batchelder. I thank Linda especially for patiently and relentlessly helping me through numerous logistic computer setup problems.

I have been fortunate in fellow students at MIT, and over the years a few of them have become good and hopefully lifelong friends. I deeply thank: Gerard Roe, a true friend and angel, who pulled me out of the hardest moments with unyielding support and belief in my ability to get this thesis done. I can't thank him enough; Amy Solomon for long long chats in which, with a lovely excitement, she introduced me to aspects of American culture I did not really know before; Adam Sobel for keeping me excited about science through his own genuine excitement and insight which he passed on in many of our discussions, for providing me with the matrix solver which is the core of the two dimensional wave models I have used and showing me how to read some of the observational data, and for his and his wife Marit's warm friendship; and Constantine Giannitsis for hours of discussions about the stratosphere and more artistic topics, and for his good and frank criticism, especially during the last stages of my thesis.

I have been very fortunate to meet Gidon Eshel, whose wonderful friendship I have relied on many a time. I thank him for the fun conversations over lunch, for loads of good advice, and for introducing me to his sweet wife Laura. I thank Gavin Esler for some very helpful scientific discussions, for teaching me how to understand an Irish accent, and for fun hours in the darkroom. I thank Rebecca Morss for bravely letting me paint her casts, and for being the kind and impressive person that she is. My first officemates, Juno Hsu, Danny Kirk-Davidoff and Moto Nakamura made my earlier days at MIT easier and more enjoyable. I especially thank Danny for reading my thesis and giving comments, and for all the incredibly delicious dinners with his lovely family.

I also thank: Greg Lawson and Pablo Zurita for not leaving me alone here at the wee hours of the night and, along with Tie Yong Koh, for listening to practice talks; Veronique Bugnion for her wonderful fondues; Marja Bister for getting me started on GRADS and for the long pleasant hours of sharing a computer room; Jessica Neu for supplying me with needed stratospheric papers; Sarah Samuel for her red/purple hair; Eyal Heifetz for his encouragement and energetic readiness to help during the late stages of my thesis; Sudharshan Sathiyamoorthy for being my first MIT friend; and Michael Morgan, Francoise Robe, Chris Forest, Lars Schade and James Risbey for helping out in the old days when I just got here.

A special special thanks to Ed McCluney for his kindness and for heading the Student Art Association at MIT which provided a much needed place in which I can relax from the scientific life and let artistic expression flow. I also thank my teachers there, Thery Mislick, Graham Ramsay and Susan Anderson.

It was my good fortune that some of my best friends happened to be in Boston during these years. I thank Gaia Bernstein and Amit Solomon, Michali Barzuza, and Alon and Julie Yavnai-Criniere for all the fun stuff we did together.

Boaz deserves the most thanks of all, too much too express here in words.

Finally, the love and support of my dear and wonderful family has helped me greatly along the way. I especially thank my parents, Miki and Viki, who passed on to me their love of math and science, and my brothers, Danny and Roni, for their friendship and humor. I dedicate this thesis to my grandmother, Ibi Harnik, whom I love very much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and motivation

Planetary Rossby waves are the dominant mode of intra-seasonal variability in the stratosphere. They are believed to be responsible for most of the deviation of the zonal mean state from radiative equilibrium, and for the poleward transport of tracers. Understanding the evolution and structure of these waves, which are defined to be the large (planetary) scale deviations from the zonal mean state[1], is therefore essential for understanding the general circulation of the stratosphere. Most of the observations of the stratosphere are based on satellite retrievals of temperature. The temperature and geopotential height are the most directly observed wave fields[2] and the variation of their longitudinal orientation with height and latitude is indicative of a wave activity flux in these directions. Looking at consecutive maps of vertical structure of the wave is therefore extremely useful as a diagnostic. We will show that at times, waves undergo changes in their vertical structure on time scales of a few days and that often this is indicative of interesting dynamics (e.g. a deceleration of the mean flow, variation in the tropospheric forcing). The response of the wave field in these cases is also interesting because it is qualitatively linear. Surprisingly, the variability on daily time scales has not been discussed much in the literature as a daily evolution of vertical structure. Since variation in vertical structure often appears as a phase propagation at some altitudes, it has been discussed as an occasional phase propagation.

The goal of this thesis is to understand what the observations of vertical structure mean and to use them to gain insight into life cycles of a few observed waves. This involves first of all understanding what the satellite retrievals of vertical structure mean, given the current observation system with its low vertical resolution (chapters 2

[1]Small scale deviations from a zonal mean state are most likely gravity waves

[2]The satellite instruments measure layer mean temperatures. Geopotential height is calculated by adding the layer means to a surface height field, while temperature is calculated by interpolation.

and 3), and using simple models to understand what controls the vertical structure and its relation to higher order and more commonly used diagnostics like the EP flux (chapters 4, 5). We will also present a different way to look at wave fields, from the point of view of wave activity and its propagation and dissipation through the stratosphere (chapter 6). Finally, we will diagnose a few specific wave events in which the vertical structure varies on time scales of a few days, and where our understanding of the mechanisms that control vertical structure allow us to determine a causality between observed evolution of the waves and the basic state (chapter 7).

In this chapter we will introduce a few of the outstanding issues regarding stratospheric planetary waves that have motivated our work (section 1.2), along with a brief review of past studies (section 1.1) and some observations (1.3). We will restrict the discussion to planetary waves and their structure, and mention only briefly some aspects of the general circulation of the stratosphere and wave-mean flow interaction. We should note that the observational examples we will show are from the southern hemisphere, while much of the theory we present was developed specifically with the northern hemisphere in mind. This reflects the fact that the northern hemisphere, being more dynamically active, roused the interest of scientists earlier. While there are important differences between the hemispheres, the basic theory is the same, especially the linear parts of it. We also have reason to believe that our discussion and results are relevant at least to the relatively quiescent periods of the northern hemisphere, when waves behave more linearly.

1.1 Charney and Drazin's theory for vertical wave propagation

The first stratospheric maps that showed planetary waves were compiled from radiosonde and rocketsonde data which were gathered in the late fifties, mostly during the international geophysical year (IGY) (e.g. Boville, 1960 and references therein). Two of the striking features were large planetary scale perturbations during winter, with little variability in smaller scales, and the lack of such perturbations in summer. Charney and Drazin (1961) explained these two features in terms of vertical propagation of Rossby waves. Using a β-plane model of vertical wave propagation, they showed that planetary waves forced in the troposphere could propagate vertically only through westerlies that are weaker than a certain limit, which depends on the wavenumber. Only the largest waves (wavenumbers one and two) can propagate into the stratosphere for the climatological wind values. The easterly winds in summer

explain the absence of waves in that season. Charney and Drazin's theory was extended by Dickinson (1968b, 1969b), who included Newtonian damping and showed the importance of the meridional structure of the wind by introducing the notion of a wave guide in the stratosphere. Dickinson (1968b, 1969a) also showed that critical surfaces (where the wind equals the phase speed of the waves) will absorb wave activity, causing the wave guide to be leaky and the amplitude of a forced wave in such a waveguide to decrease as a result. Dickinson (1968a) pointed out that since the critical wind of the Charney and Drazin model depends on latitude, a spherical model is needed to study vertical propagation and obtain a Charney-Drazin critical velocity in the real atmosphere.

Observational evidence of a link between the troposphere and stratosphere was first shown by Boville (1960), who performed a Fourier analysis of perturbations on the 500mb and 25mb surfaces. Muench (1965) found evidence of a tropospheric growth followed by vertical propagation to the stratosphere by constructing pressure-time plots of wave one and two amplitudes from northern hemisphere radiosonde data and estimated a vertical propagation of about $6km/day$. A more comprehensive review of early observational studies can be found in Matsuno (1970), who was the first to attempt to simulate the observations by forcing a model with observed 500mb heights. One of the main contributions of Matsuno's study is the formulation of the quasi-geostrophic equations of vertical wave propagation on a sphere and the introduction of an index of refraction for Rossby wave propagation in the vertical-meridional plane. The basic state used was an idealized northern hemisphere winter jet, with a zero wind line in the tropics. The corresponding PV gradients had a ridge oriented along the jet maximum, acting as a wave guide[3]. Matsuno also suggested the possibility of a cavity forming, rather than a waveguide, where the perturbations are bounded from above by large winds, as suggested by Charney and Drazin (1961). The results of Matsuno were quite good for wave one, but not for wave two, which was considerably weaker than observed. One of the shortcomings of Matsuno's calculation in terms of simulating observed waves is the use of idealized zonal mean winds instead of the observed winds[4]. Matsuno explained this by noting that the observations include

[3]Matsuno (1970) correctly pointed out a mistake by Dickinson (1968b) who had the waves guided up regions of weak westerlies and not along the maximum winds. The source of Dickinson's error was in choosing a basic state that renders the problem separable but at the same time cancels the contribution of the meridional curvature to the PV gradients.

[4]Other shortcomings include the assumption of an isothermal atmosphere, with a constant Brunt Vaisala frequency and not accounting for thermal damping that is a function of height. Matsuno did use some damping, in the form of an imaginary phase speed (constant and equal Newtonian damping and Rayleigh friction coefficients) in order to get rid of the singularity at the zero wind line.

small scale features which are unreliable but may have a large influence on the results through a contribution to the PV gradients. This sensitivity of the response to details of the basic state may explain the discrepancies between his results and observations, however, it also makes it hard to generalize the results. In particular, the two-dimensionality of the problem (meridional and vertical directions) makes it hard to analyze the results in terms of vertical wave propagation as in Charney and Drazin (1961).

One of the major issues is the extent and manner in which wave amplitudes depend on the basic state of the stratosphere, both instantaneously, and in a climatological sense. A theoretical framework in which we can study the sensitivity of the system and identify the important parameters is essential. Studies that followed Charney and Drazin (e.g. Simmons, 1974; Schoeberl and Geller, 1977; Karoly and Hoskins, 1982) were concerned with finding a simplifying framework in which to study the waves and understand what controls their structure. This is also one of the goals of our study.

Simmons (1974) formulated a simplified problem in which the basic state wind was separable in latitude and height, and an approximate β plane was used in a way that renders the equations separable. This reduced the problem to one dimension, as in Charney and Drazin (1961). The wavenumber two response using zonal jets characteristic of observations was not deficient as in Matsuno (1970), suggesting that a linear model is capable of explaining observations.

Schoeberl and Geller (1977), using a spherical model with realistic winds, forced their model with observed geopotential height amplitudes and phases of waves 1-3 at 100mb. They used an approximate separable version of the equations to find meridional eigenfunctions which they called Fourier-Hough modes and analyzed the results in terms of their vertical propagation characteristics. They were able to explain the sensitivity of the model to increasing the zonal mean winds and to various damping profiles in terms of the response of individual modes. In order to separate the equation in the meridional and vertical directions, however, Schoeberl and Geller had to ignore the vertical shear term in the PV equation (equation 10 in their paper, equations 5.2, 4.3 here). Since shear in the stratosphere is often large, the neglected terms may be important.

Lin (1982) used a linear primitive-equation model to study the sensitivity of the response to variations in the structure of the zonal mean wind using the index of refraction as a diagnostic. Lin discussed the importance of a waveguide configuration for the stratospheric response, noting that it makes the location of maximum response less sensitive to the latitude of tropospheric forcing. Figure 1.1 shows an example of

one of Lin's runs. Shown are the basic state zonal mean wind and index of refraction squared[5] (n^2_{ref}), and the wave 1 geopotential height and EP fluxes. We see a ridge of n^2_{ref} in high latitudes and very large values in the topics, with a region of reduced but positive values in between. It is not clear whether this basic state is a wave guide in the sense that the waves are evanescent between the ridge and the tropics. The EP fluxes are not a good indication because they point out of the ridge (they are equatorward at 40-60°N, 20-60km). The ridge has to be large enough for a waveguide to form.

To determine whether a given basic state is a waveguide, we need to separate the meridional and vertical propagation characteristics. Karoly and Hoskins (1982, see also O'Neill and Youngblat, 1982) essentially did this by calculating wave rays for a given point source. They showed that some wave rays reflect back and forth in the meridional direction as a result of a waveguide, but eventually, due to the spherical geometry, they bend equatorwards. Whether they bend equatorwards right away or not depends on the basic state and on the initial ray propagation direction in the meridional-vertical plane. The relation between a given wave field structure and the wave rays is not simple because the wave is a superposition of many point sources, and once the wave reflects in the meridional direction it will superpose with itself. Also, ray tracing cannot incorporate damping and wave tunneling through negative n^2_{ref} regions.

In chapter 5 we will develop a framework in which to diagnose propagation in the meridional-vertical direction, and in particular, to diagnose the existence of a waveguide. Also, by making use of the special characteristics of a waveguide we determine the propagation characteristics in the vertical direction and compare the results with an equivalent one-dimensional model that we formulate. In chapter 6 we will discuss the relation between the EP flux and the index of refraction configuration, and the relation to Karoly and Hoskins' ray tracing.

Understanding what determines wave structures and what various diagnostics of the waves mean is essential for understanding many aspects of their behavior. In the rest of this chapter we will introduce a few outstanding questions regarding stratospheric planetary waves, both by describing past studies and by bringing examples from observations.

[5]Actually plotted is Matsuno's Q_o, which is the quasi-geostrophic index of refraction squared minus the zonal wavenumber term $\frac{s^2}{\cos^2\theta}$, where s is the integral wavenumber and θ is latitude.

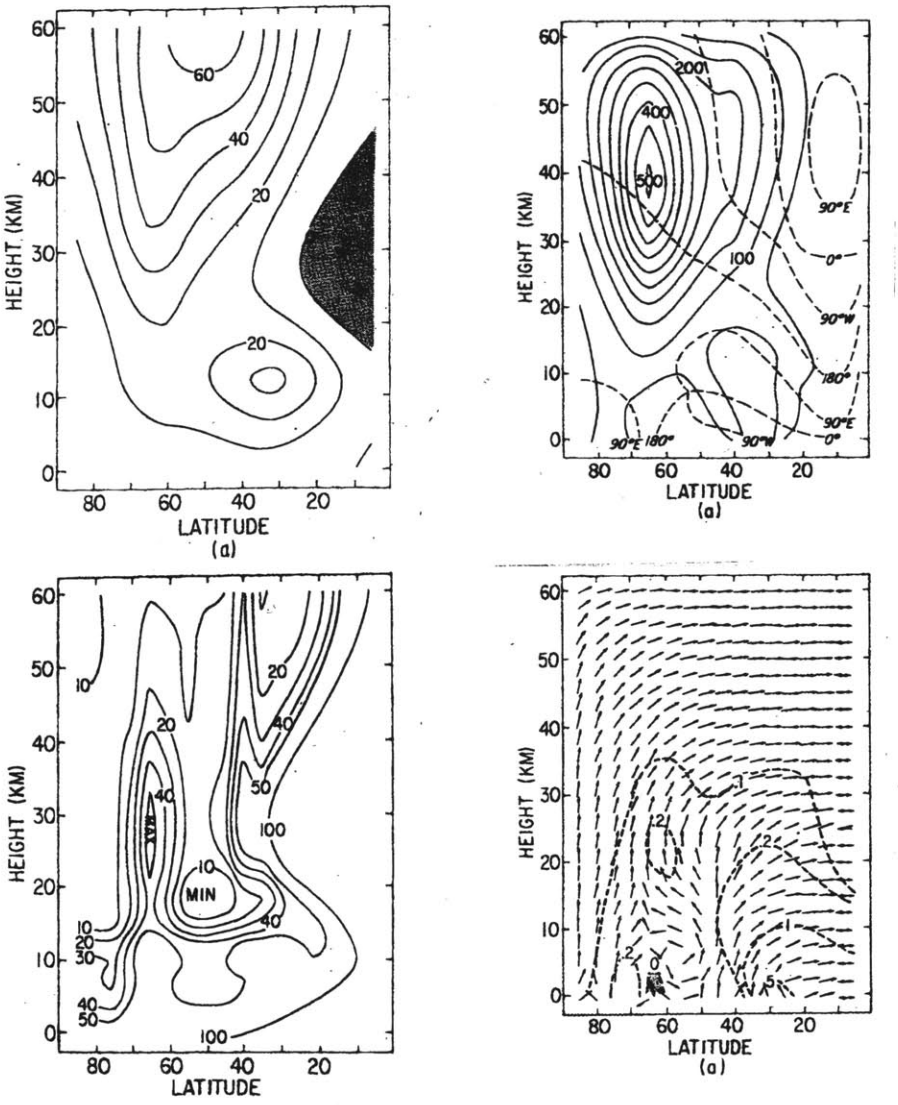

Figure 1.1: Results from one of the runs of Lin (1982). Top- Left: Zonal mean wind (m/sec). Right: Stationary wave 1 geopotential height amplitude(solid, meters) and phase (dashed). Bottom- Left: The index of refraction squared without the zonal wavenumber term (Matsuno's 1970 Q_o). Right: Eliassen-Palm Flux for wave 1 (arrows). Figure taken from Lin (1982).

1.2 The observed planetary waves- some resolved and unresolved questions

Remote sensing of temperatures, which started in the early 70's and became part of the current operational system in 1979, revealed new and interesting aspects of stratospheric planetary waves that were not observed before. Rather than continue to describe the various studies chronologically, we will discuss a few aspects of the observations that interest us and have motivated this study.

1.2.1 The seasonal cycle of stratospheric planetary waves

Charney and Drazin (1961) and the studies that followed (see previous section) showed the importance of the basic state wind for planetary wave propagation. While the lack of planetary waves in summer is easily explained by Charney and Drazin's theory, the climatological seasonal evolution of wave activity in winter is not so obvious. Hirota et al. (1983) studied the wave activity in both hemispheres for October, 1979-August, 1981, and found that while in the northern hemisphere planetary waves were large throughout the winter, in the southern hemisphere they were large in spring and fall and weak in June-August. Randel (1988, 1992), looking at eight and twelve years of data, showed that the southern mid-winter minimum in wave activity is a climatological feature, both of the stationary and the transient waves (figure 1.2).

Plumb (1989) suggested the Charney-Drazin criterion could explain the seasonal cycle. In the southern hemisphere, mid-winter winds are strong enough to block propagation of waves and cause a minimum of wave activity, while in the Northern hemisphere, the waves are large enough to decelerate the winds and allow propagation all winter long. Plumb demonstrated this mechanism using a one-dimensional β-plane model of wave-mean flow interactions that was forced with constant forcing at the surface and relaxed to a seasonally varying zonal mean wind (easterly in summer and westerly in winter). The model was essentially linear for a weak forcing, resulting in a mid-winter minimum due to the strong winds, and quasi-linear for large waves, resulting in a maximum of wave activity in mid-winter with much reduced winds. This behavior essentially supports the notion that the wave activity is determined according to Charney and Drazin's theory, however, there are limitations to using a one-dimensional model. Most notably, in the absence of meridional propagation of waves and a critical surface at the equator, the wave-mean flow interaction expected to be weaker. Also, there are no meridional gradients of the basic state, resulting in smaller PV gradients and correspondingly in reduced vertical propagation. Wirth

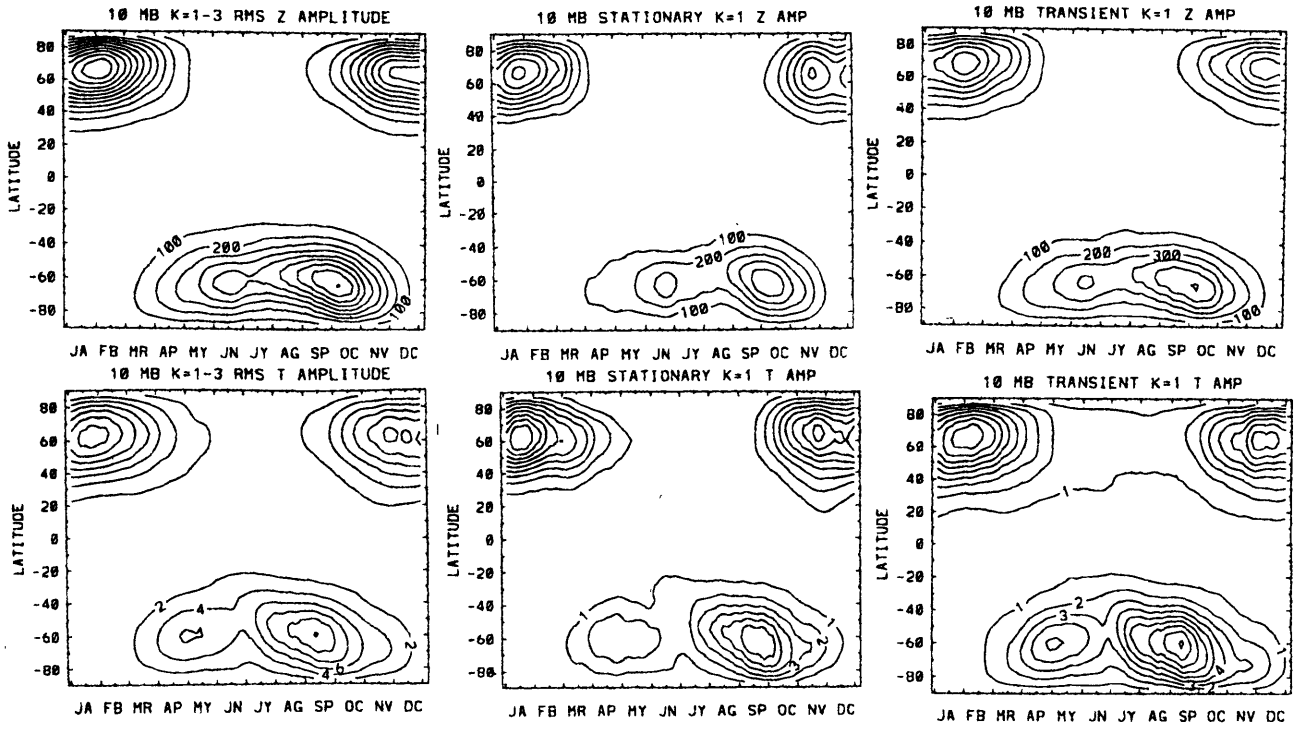

Figure 1.2: .
Geopotential height (top) and temperature (bottom) and ensemble rms zonal waves 1-3 amplitudes (left), ensemble wave 1 stationary (middle) and transient (right) amplitudes at 10mb. The ensemble average is over the years 1979-1986. Stationary waves are the 33 day running means and transients are the deviations from it. Non-overlapping 3-day means are plotted. Figure taken from Randel (1992).

(1991) forced a two-dimensional spherical model similar to Matsuno's (1970) using monthly mean basic states for the months April-October to see if the response will explain the seasonal cycle of planetary waves. The model did well in late winter but failed to reproduce the spring wave activity maximum. It is hard to conclude whether the Charney-Drazin criterion, and Plumb's results for one dimension, explain the response of Wirth's model. The main problem is that whether the waves are propagating or are evanescent in a given direction is no longer a *local* property of the basic state. Rather, the full wave solution needs to be obtained and the propagation characteristics *diagnosed.* In chapter 5 we will develop a diagnostic framework that makes use of the special waveguide characteristics of the basic state to allow us to generalize the Charney-Drazin criterion to the two dimensional case.

It is important to note that the seasonally averaged wave may not resemble the steady state response to the seasonal mean basic state because of the transience in the system. In the next section we will show that wave activity usually comes in episodes of a few weeks, rather than continuously. Also, in section 1.3 we will present some evidence that the climatological mid-winter minimum may be, at least partially, a result of less frequent wave events rather than weaker continuous waves.

1.2.2 Wave episodes and the time evolution of vertical structure

Another important aspect of the planetary waves is their time variability, both of amplitude and of phase. Hirota and Sato (1969) were the first to note that the quasi-stationary wave amplitudes, as well as the zonal mean flow are highly correlated and oscillate with a period of up to a few weeks. These oscillations are a common feature and have been observed by others (e.g. Hartmann, 1976 for the southern hemisphere, and Madden, 1975, for the northern hemisphere). Another form of transience is a phase propagation of the waves. Mechoso and Hartmann (1982) did a space-time spectral analysis of observations and found a large amount of variance concentrated in eastward and westward propagating modes. The westward waves, which were dominant in the northern hemisphere, were highly coherent between the troposphere and stratosphere, while the eastward waves were dominant in the southern hemisphere and exhibited no coherence between their tropospheric and stratospheric parts. Leovy and Webster (1976) pointed out the behavior of wave 2 in the southern hemisphere is unique, because it propagates eastward regularly, and over very long periods. Other waves are mostly quasi stationary, with occasional periods of propagation, often observed during wave growth. We will distinguish between the episodic propagation of quasi-stationary waves and the clear modal propagation of wave 2 (a coherent propagation throughout the stratosphere), and discuss the latter separately in section 1.2.3.

There have been a few approaches to explaining the variability of the waves. One kind of study has concentrated on the simultaneous existence of stationary and free traveling Rossby waves, which superpose to cause periodic oscillations in amplitude and wave fluxes, both in the troposphere and in the stratosphere (e.g. Salby and Garcia, 1987; Lindzen et al., 1982; Madden, 1983). There are a few uncertainties regarding this mechanism. In the southern hemisphere, where *eastward* phase propagation dominates, it is not clear that the modes exist at all. Westward propagating modes are external Lamb waves. Model studies have shown that their periods are insensitive to the basic state configuration and they have been observed extensively in

the troposphere-stratosphere (see the review by Salby, 1984, and references therein). Eastward propagating modes, on the other hand, can only exist as internal modes, which requires the existence of turning points. This would make these modes very sensitive to the zonal mean wind. Hirota (1971) and Garcia and Geisler (1981) showed that such modes can be excited in simple models as a result of time varying forcing. These results heavily depend on the simplicity of their basic states. Da Silva and Lindzen (1987, 1993) showed that *westward* propagating external modes were not easily excited in a baroclinic model by variations of the zonal mean wind flowing over topography, even though they were excited in a barotropic model when similar variations were specified. Given that it is hard to excite the external westward propagating modes in a baroclinic model, even though their frequency does not depend strongly on the zonal wind, it is hard to see how internal modes which are very sensitive to the basic state configuration can be excited.

A different point of view, which is closer to our current approach, is of the waves undergoing life cycles of vertical propagation followed by barotropic decay, as a response to episodic tropospheric forcing (Randel, 1987b, 1990; Randel et al., 1987)[6]. Two questions that need to be answered are what causes the variations in forcing in the troposphere, and what determines which tropospheric episodes propagate vertically. Nigam and Lindzen (1989) showed in a linear model that variations in the latitudinal structure of the zonal mean winds can result in large variations in the stationary wave amplitudes in the troposphere. Shiotani and Hirota (1985) suggested that variations in the strength of subtropical jet will modulate the amount of wave activity propagating up to the stratosphere, by attracting some of the wave activity. Shiotani and Hirota also suggested the lower stratospheric winds may act as a shutter to wave activity, thus affecting wave propagation. The problem with this mechanism is that the vertical wavelengths are large enough for waves to tunnel through regions of evanescence in the lower stratosphere (e.g. see Jacqmin and Lindzen, 1985[7], and results in chapters 4, 5). Also, this mechanism leaves open the question of the source of the variations of lower stratospheric winds.

In the current work, we will view the waves as evolving in episodes. The occasional

[6]Randel (1987b, 1990), as opposed to Mechoso and Hartmann (1982), found a substantial correlation between the troposphere and stratosphere, because he took into account a time lag due to finite vertical propagation times. Also, he noted that not all tropospheric episodes propagate into the stratosphere, which decreases the correlations.

[7]Jacqmin and Lindzen, 1985, tested the sensitivity of a linear primitive equation model stationary wave response to variations in the basic state. For reasons explained there, most of their basic states refractive indices were such that a region in the lower stratosphere was evanescent. The stratospheric responses on the various basic sates varied considerably, but the evanescent region was not crucial because the waves were able to tunnel through it.

phase propagation will then be explained as the result of changes in the vertical structure of the wave as it adjusts to its steady state (which is not always reached). Viewed in this way, the oscillations that arise when a source is switched on are due to the adjustment of the mode to steady state. For example, in a basic state that has turning points, the perturbation will propagate vertically, then reflect downward, adjusting its vertical structure in the mean time. This will also result in oscillations in wave fluxes. When viewed at one level, the changes in structure will appear as a burst of phase propagation, resulting in some power in propagating modes. In chapter 7 we will discuss a few observed episodes in this light. Randel et al. (1987) showed another kind of adjustment of the modes, by compositing a few wave episodes to study the life cycle of the waves. They found that the EP fluxes evolve from being vertical to tilting towards the equator. This sort of adjustment also results in phase changes at some latitudes and heights that may appear as a phase propagation. If the mechanisms for structure changes have a specific time scale involved, the time-space Fourier decomposition will have a peak (probably broad) around the characteristic frequency.

1.2.3 Eastward propagating wavenumber two in the southern hemisphere

One of the striking phenomena in the stratosphere is the eastward-propagating wavenumber 2 in the southern hemisphere (observed in early radiosonde and satellite data by Phillpot, 1969, Deland, 1973, Leovy and Webster, 1976, Harwood, 1975, Hartmann, 1976). Unlike other waves, wave 2 in the southern hemisphere is predominantly propagating, with a similar phase speed at all levels and latitudes in the stratosphere, suggesting it is a mode. Manney et al. (1991a) compiled a ten-year climatology of wave two in the southern hemisphere. As with the quasi-stationary waves, wave 2 appears in episodes of a few weeks. The period of the waves varies on interannual and intra-seasonal time scales, between 5-40 days ($3-23\frac{m}{sec}$ at 60°S). The meridional structure is a broad peak between 55-65°S, with maximum geopotential amplitudes typically between 400-800 meters at 10mb. The source of the waves is not yet clear. One possibility is that the waves are due to instability, either a barotropic instability on regions of negative meridional PV gradients which are occasionally observed on the poleward or equatorward flanks of the jet (Hartmann, 1983), or a stratospheric extension of a tropospheric baroclinic instability (Geisler and Dickinson, 1975; Geisler and Garcia, 1977; Hartmann, 1979; Straus, 1981; and Young and Houben, 1989). Another explanation is that the eastward propagating waves are forced by nonlinear wave-wave

interactions in the upper troposphere, essentially, by the baroclinic waves organizing into wave packets with a wave-2 envelope (Scinocca and Haynes, 1998, and references therein). All of these mechanisms have been shown to occur in models (see corresponding references mentioned above). Models exhibit two kinds of barotropically unstable modes. The first, occurring on regions of $\overline{q}_y < 0$ which are on the poleward flank of the jet, tend to peak at 70°S and to propagate eastward with a period of a few days. The corresponding momentum flux is outward from the jet. The second kind of barotropically unstable modes occurs on regions of $\overline{q}_y < 0$ that are equatorward of the jet. The perturbations peak in the middle of the jet and have equatorward momentum fluxes in the region of negative PV gradients. Observed waves, apart for rare occasions, have poleward momentum fluxes in mid-latitudes, pretty much ruling out the second kind of barotropically unstable modes (Hartmann, 1983, 1985). Also, the momentum fluxes of observed fast moving polar perturbations (the first kind of modes) point outward from the jet, which is also opposite the observed momentum fluxes in the southern hemisphere (Hartmann, 1983). The problem remains to distinguish observationally between the long wave baroclinic instability mechanism and the tropopause wave-wave interaction mechanism, because both essentially consist of a response of the stratosphere to tropospheric forcing, and because the tropospheric observations in the southern hemisphere are poor.

The main problem with the wave-wave interaction mechanism is to explain long lasting modes like the wave 2 episode in the fall of 1983, which showed a constant phase progression that lasted for over 50 days (Shiotani et al., 1990). The attraction of it is that it explains why phase propagation does not always extend into the troposphere (Manney et al., 1991a), and the lack of coherence between the tropospheric and stratospheric parts of the wave (Mechoso and Hartmann, 1982, see section 1.2.2[8])

In chapter 4 we will look at the baroclinic instability problem and discuss the possibility of internal stratospheric instability.

1.3 Some examples from the southern hemisphere winter of 1996

In this section we will present some of the observations that have motivated our study. Observations are from the southern hemisphere winter of 1996. For more

[8]As was mentioned in section 1.2.2 regarding quasi-stationary wave 1, Randel (1987b) found a large correlation between the troposphere and stratosphere, when a vertical propagation time lag was taken into account (for wave 2 as well as for wave 1).

details about the data source and quality, see chapter 2.

Figure 1.3 shows time-longitude plots of the temperature perturbation at 10 mb, 60S. Shown are the deviations from a zonal mean, and the wave 1 and 2 components. We see a few wave episodes, which are either dominantly wave 1 or dominantly wave 2. Wave 1 is mostly quasi stationary (early July, mid-July-August, September) and wave 2 eastward propagating (late July). Wave 1 propagates eastward for a short period of time around August 16th. In September, the perturbation is mostly of zonal wave 1 (one trough and one ridge), but wave 2 is quite large, causing the perturbation to concentrate in the western hemisphere. The coexistence of both wave 1 and 2 is quite common in late winter, when the vortex is weaker and in the process of breaking down.

In section 1.2.1 we discussed the seasonal cycle of wave activity. Randel's climatologies (1988, 1992) show there to be two maxima of wave activity in early and late winter, and a minimum in mid-winter (figure 1.2). Wave activity in 1996 does not behave like the climatology would suggest, because the largest waves are in mid-winter. Also, the temperature amplitude of the waves is much larger (maximum amplitude of 15-25°K, as opposed to a climatological maximum of 10-12°K). Since wave activity appears in episodes, this could mean that the climatological minimum is due to a less frequent occurrence of wave episodes, rather than to the existence of lower amplitude waves. It is important to distinguish between having a succession of large and relatively short wave episodes and having a smooth constant level of smaller amplitude waves, because only the latter can be said to be a steady state. This raises the question of the relevance of steady state models to explaining the wave structure.

To get an idea of the relevance of a steady state, we calculated the wave variance in the southern hemisphere over the years 1980-1998. The daily variance of temperature and geopotential amplitude is calculated for each latitude between 40-75°S, and the maximum value is plotted in figures 1.4 and 1.5. This is almost identical to the variance at 60°S, since that is where the waves generally peak. Note that the square of the variance, ($Var^2(f) = \overline{(f - \overline{f})^2}$, overline denotes a zonal mean) is half the square of the wave amplitude. We show the geopotential height data, which was also calculated by Randel (1992), for comparison with the temperature time series. We see that time variations in temperature tend to be larger, appearing more like episodes than a continuous wave event. Since temperature is a vertical derivative of geopotential height, it is more sensitive to variations in the vertical structure, and reveals such variability more clearly. We see that there are some years that have a mid-winter wave minimum in between two wave maxima, as the climatology does (1980, 1981, 1984, 1998), some years have just one peak of wave activity, and it is large

during mid-winter (1988, 1992, 1997), and some years look more like a succession of wave events (1989, 1996).

The vertical structure of the waves may vary from one wave episode to the other. Some of these structures are shown in figure 1.6, in longitude-height sections of wave 1 at 60°S. Also shown are the corresponding 5-day averages of the zonal mean wind. We see three different structures, along with some notable differences in the zonal mean wind. The August wave tilts westward with height, with the geopotential height increasing throughout the stratosphere. In September, on the other hand, the geopotential height peaks in mid-stratosphere, where the temperature has a node, and on the day shown the wave is vertical. The zonal mean wind also peaks in mid-stratosphere. A double peaked temperature structure is found in September of other years, sometimes also in wave 2. In June, the geopotential height has a node in mid-stratosphere, slightly below a broad peak in temperature. We will show in chapter 4 that these structures can be reproduced in a one-dimensional steady state model, and are a function of the basic state vertical wave propagation characteristics. We will further show in chapter 5 that the observed differences in vertical structure between waves in September and August are a result of the basic state changes that occur towards late winter.

In addition to a seasonal time scale variability of the vertical wave structure, we find variability on time scales of a few days. Figure 1.7 shows a succession of daily longitude-height sections of wave 1 temperature at 60°S in August. We see that the wave undergoes a change in structure over a time period of a few days. This accounts for the eastward propagation of wave 1 that is observed in this period (figure 1.3). It is often the case that eastward propagation is not the same at all levels, implying a change in the wave structure, rather than propagation of the wave. In chapter 7 we will show other instances of structure changes and discuss the reasons for the variations in each instance. The propagation of wave 2 is different because it occurs at all levels in the stratosphere, suggesting it is a propagating mode. Occasionally, however, wave 2 also undergoes variations in its vertical structure (for example, the sudden shift westward for a few days around September 23).

Finally, we need to worry about the quality of the data, because we are interested in vertical structures, and the resolution of satellite observations is quite poor in the vertical. As an example, figure 1.8 shows longitude-height cross-sections of the temperature perturbation and its wave 1 and 2 components for three consecutive days in June. There are small scale features that are most likely spurious (e.g. June 2nd at 1mb). Also, wave 2 shows discontinuous evolution with the amplitude being much smaller and the pattern 180° our of phase on June 2nd, compared to June 1 and 3.

This results from the total temperature perturbation at 10-2mb, 60-120°E growing from 2 to 8°K on June 2nd, and moving westward to 0-60°E while intensifying to 12°K on June 3rd. It is not clear whether this wave 2 evolution is real or whether it is within the observational uncertainty.

Table 1.1 shows some results from a comparison of co-located radiosondes and satellite based temperatures. Data shown is for September 1991 to August 1997, at

Year	Mean(°K) Jul-Aug (yearly)	STD(°K) Jul-Aug (yearly)	Min/Max yearly (°K)	No. of observations
1991*	(3.51)	(4.2)	-4.5/15.5	172
1992	5.6(2.81)	3.4(6.1)	-9.0/21.0	553
1993	5.2(2.34)	4.0(6.5)	-10.0/14.0	451
1994	5.6(3.05)	3.8(4.0)	-4.0/16.0	1024
1995	6.8(3.15)	3.7(3.7)	-8.0/15.0	864
1996	6.5(3.49)	5.5(3.7)	-10.0/13.5	1030
1997*	7.6(3.91)	6.2(4.1)	-7.5/14.5	722

Table 1.1: Statistics of satellite observations minus co-located radiosonde measurements, at 10 mb in the southern hemisphere, for the period September, 1991-August, 1997. Years marked by a * have observations for only part of the year. The minimum and maximum values are given after taking out values that are more than four standard deviations away from the mean.

10mb in the southern hemisphere. The satellite data, which is the same as we use in our observational studies, is independent of the radiosonde data (see chapter 2 for an explanation). The biases in July-August are in the range of 5-8°K and the standard deviation varies between 3-6°K. Yearly biases and standard deviations are smaller than the July-August values for most years, because the deviations tend to be smaller in summer. Also shown are the minimum and maximum differences between the radiosondes and the satellite observations, where deviations that are larger than four standard deviations are discarded. We see that the spread is very large. By looking at the errors of all observation stations on specific days (not shown), we can determine if the errors are a constant bias or whether they have a horizontal pattern. We find that at times there is a constant bias, meaning the wave patterns are not affected by the satellite errors but at times the horizontal structures of the satellite and the radiosonde measurements are very different. These results point out large errors in the satellite data, even from quite recent years. It is important to remember that the radiosondes can see small-scale, short-lived features, like gravity waves, which

the satellites aren't able to resolve. This can account for some of the spread but not for the biases. In the next two chapters we will discuss the observations and the various sources of error.

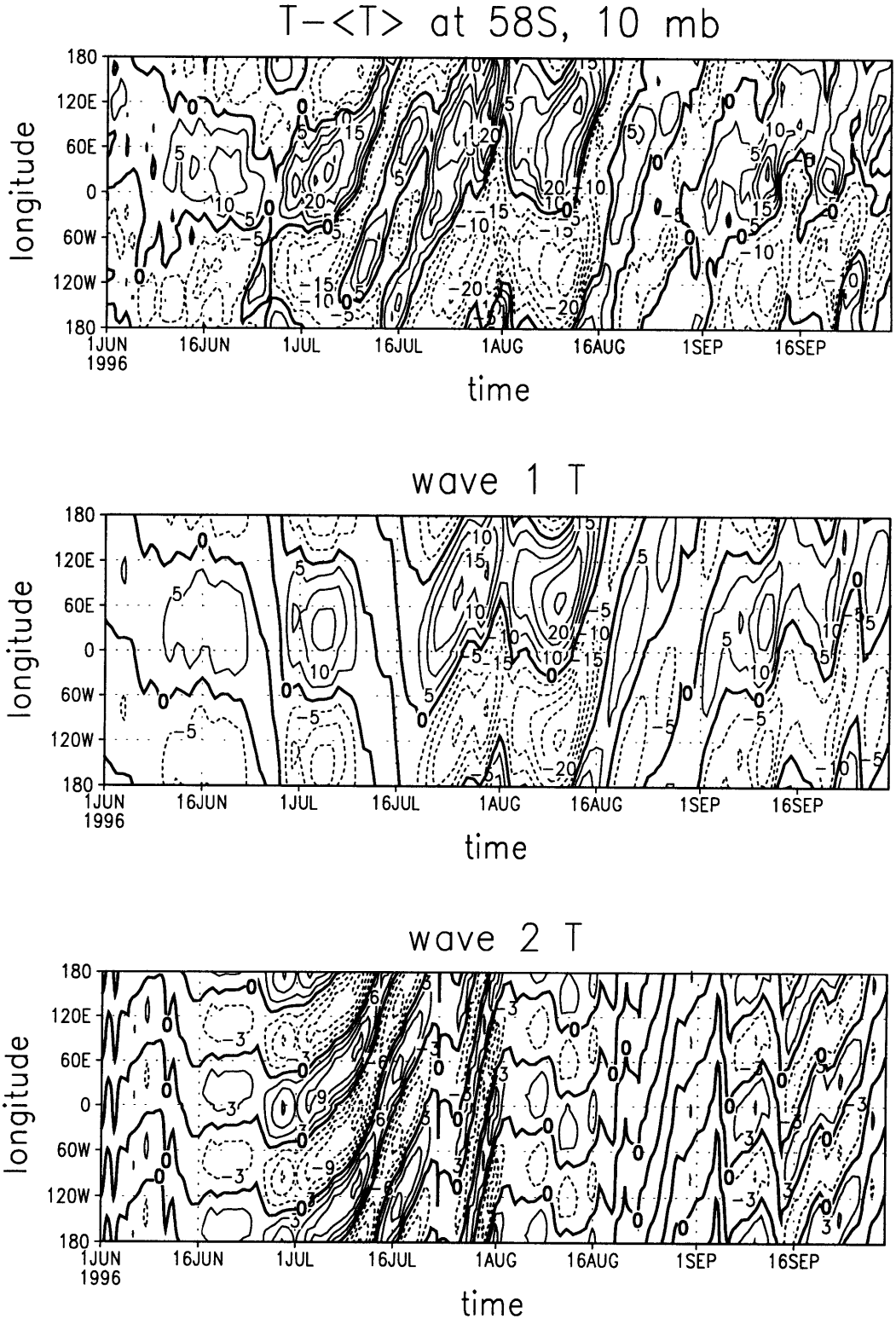

Figure 1.3: Longitude-time sections of temperature deviations from a zonal mean (top), and the wave 1 (middle) and 2 (bottom) components, at 10 mb, 58°S, for June 1st - September 30th, 1996. Contour interval is 5°K for the top two plots and 3°K for the bottom one. Zero line is thick and negative values dashed.

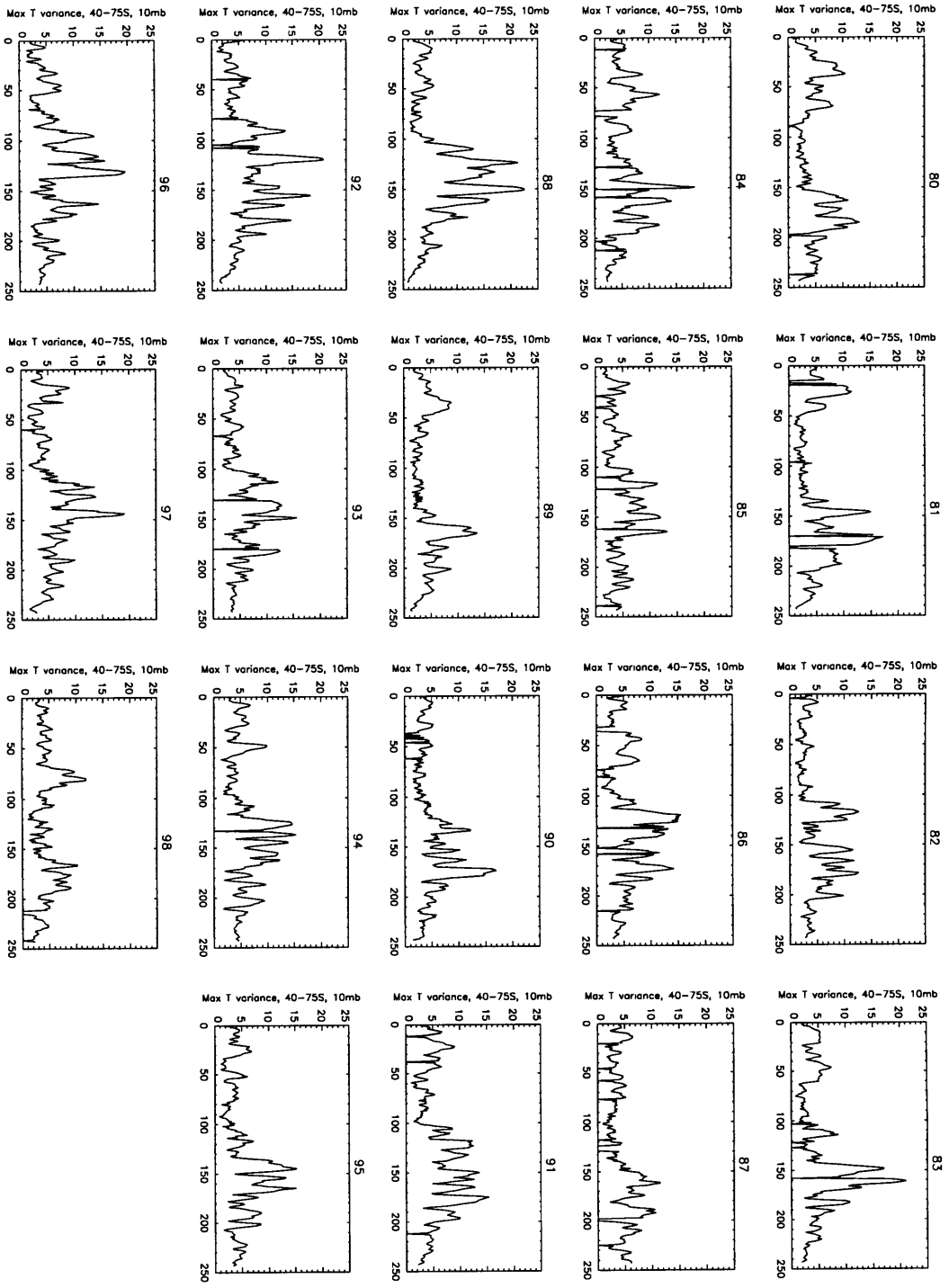

Figure 1.4: The maximum temperature variance at 10 mb in the latitude band of 40-75°S, for each of the years 1980-1998 (see text for details). Time series are daily values for April 1st-November 30th. Days are numbered consecutively (August is days 123-154). Years are marked at the top of each sub-plot.

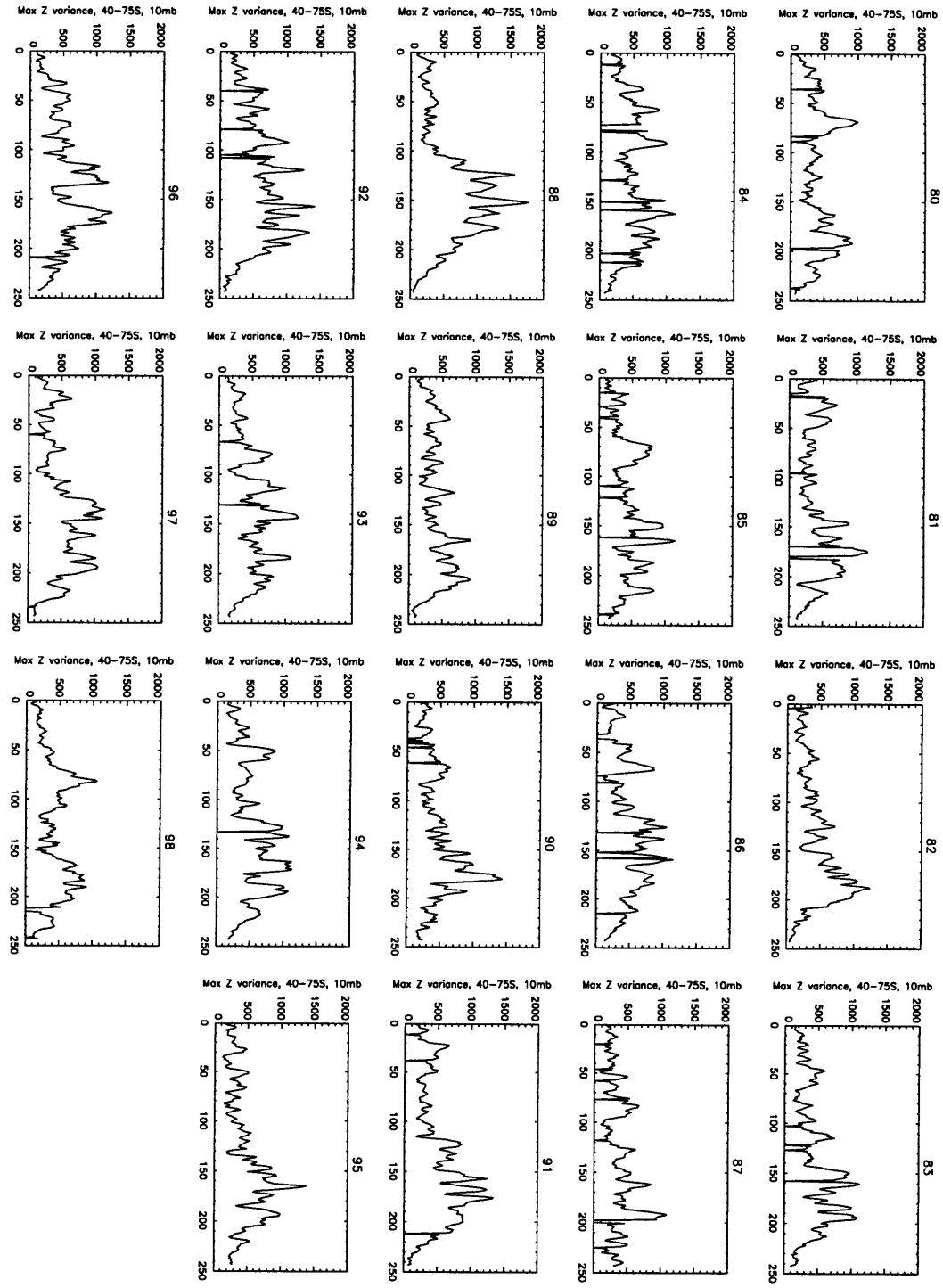

Figure 1.5: As in figure 1.4, for geopotential height.

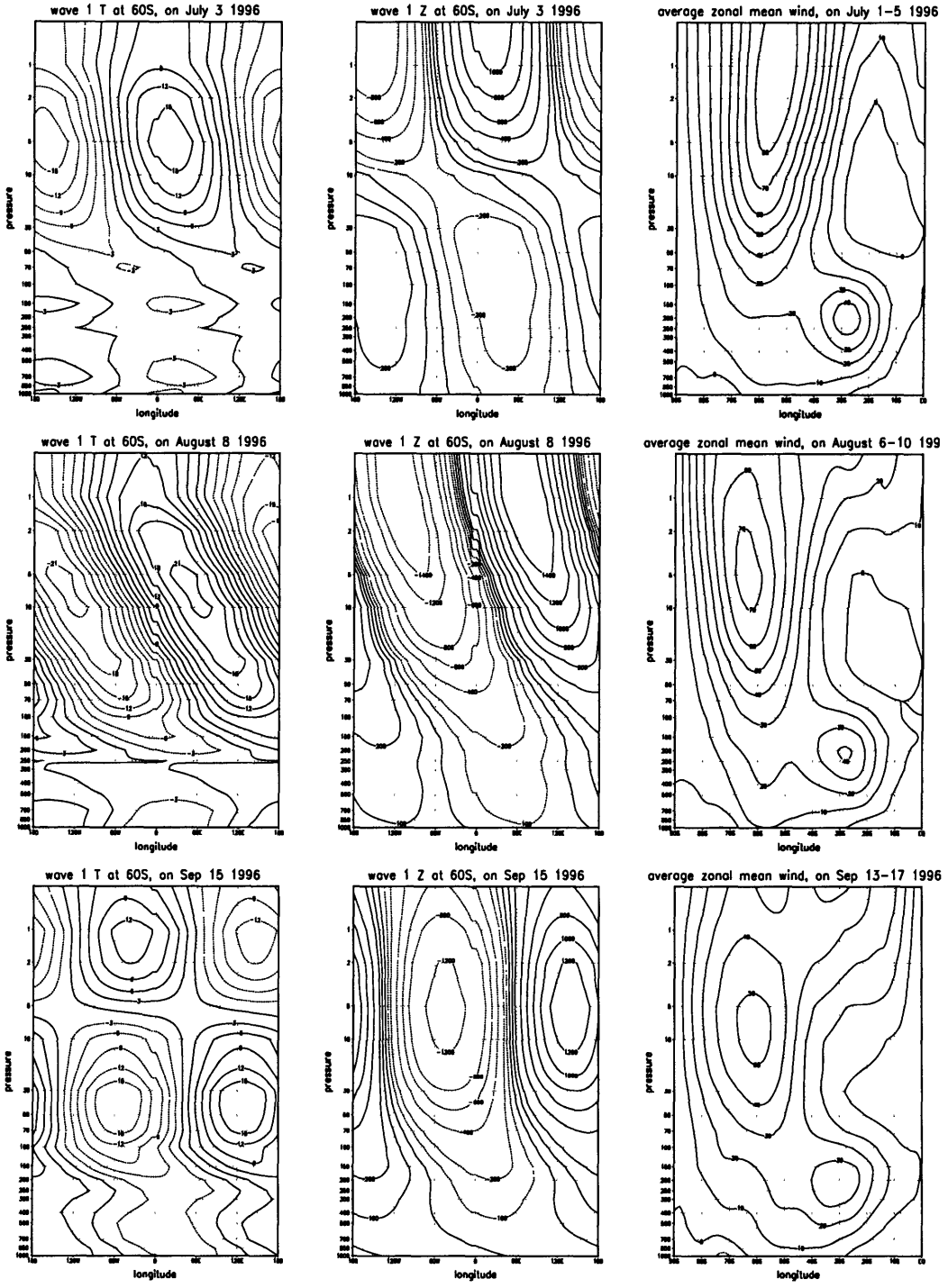

Figure 1.6: Longitude-height cross-sections at 60°S of wave 1 temperature (left) and geopotential height (middle column) for July 3rd (top), August 8th (middle row) and September 15th (bottom), 1996. On the right are latitude-height plots of the five day average of zonal mean wind, centered around the corresponding days. Contour intervals are 3°K for temperatures and 10m/sec winds. Geopotential height contours are at 0, ±100, ±200, ±400, ±600, ±800, ±1200, ±1400. Negative values are dashed.

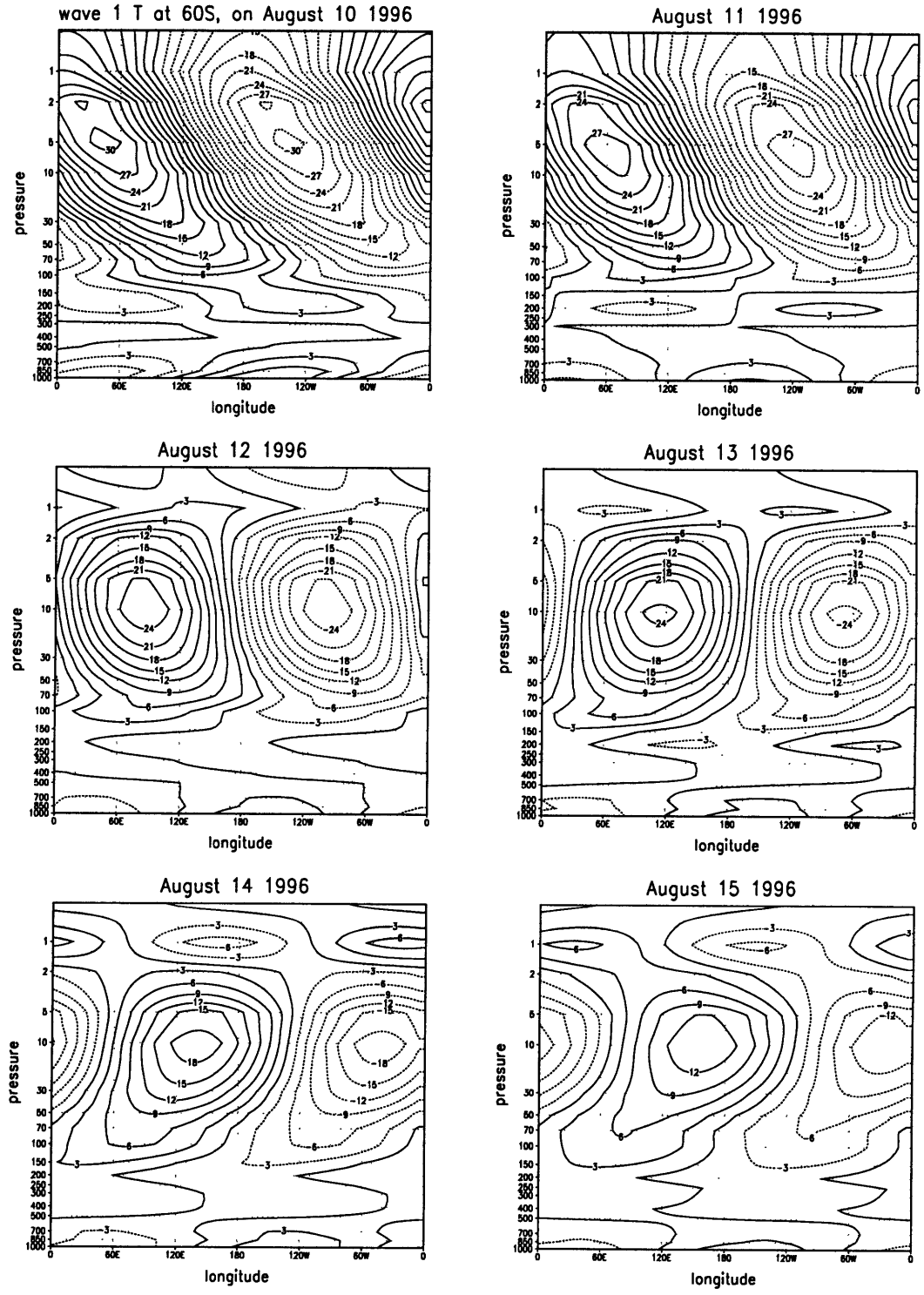

Figure 1.7: Six consecutive days (August 10-15, 1996) of wave 1 temperature longitude-height sections at 60°S. Dates are marked on top of each subplot. Contour interval is 3°K, negative values are dashed.

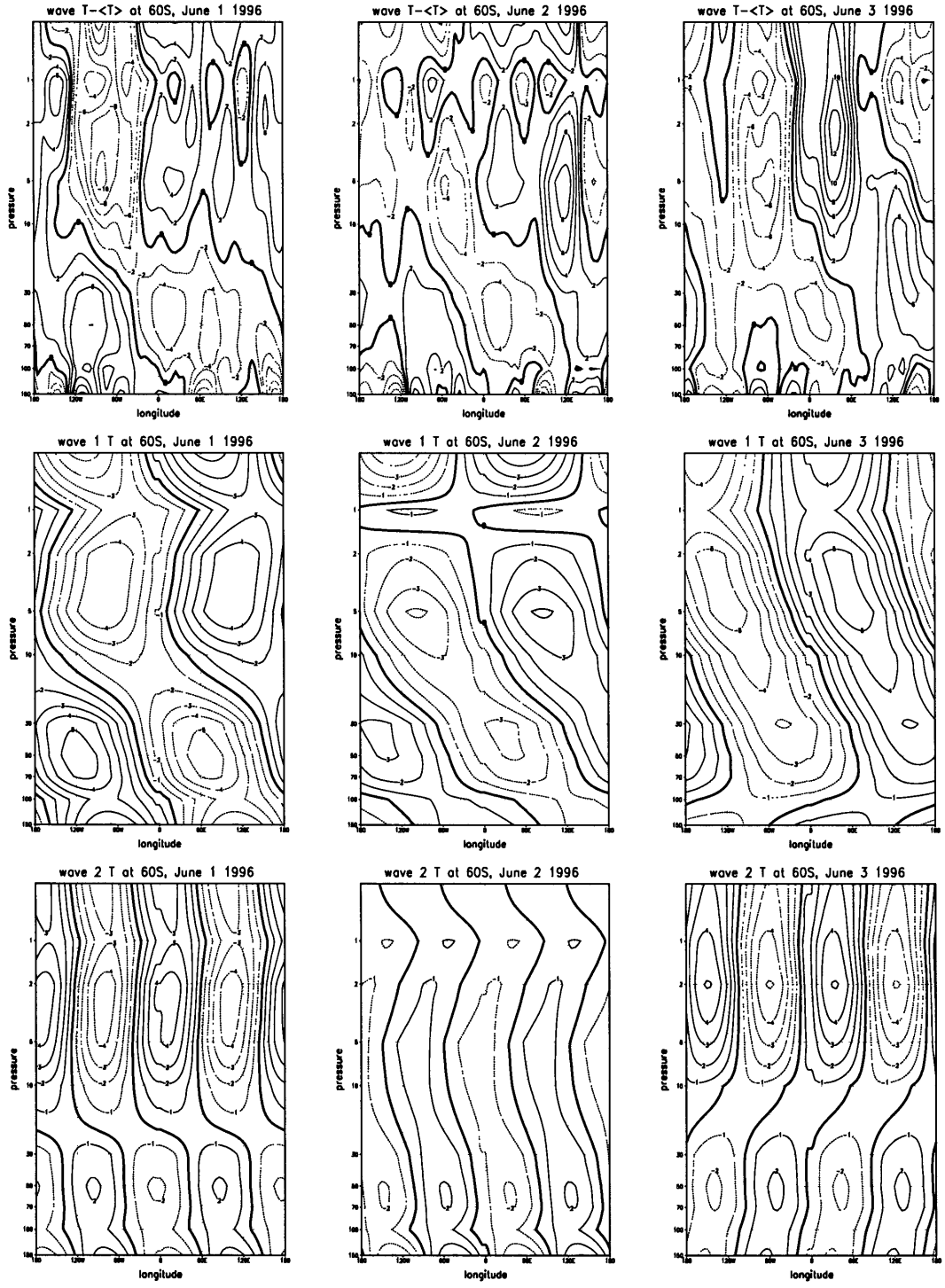

Figure 1.8: Longitude-height cross-sections of temperature deviations from a zonal mean (top), and the wave 1 (middle) and 2 (bottom) components, at 60°S for three consecutive days (left to right: June 1-3, 1996). Note that the bottom is at 150mb. Contour interval is 2°K for the top plot and 1°K for the bottom two. Zero line is thick and negative values dashed.

Chapter 2

The operational observations products

In 1.3 we saw that the satellite observations do not agree very well with co-located radiosonde observations. We also saw that the observations are very noisy and that the waves we are interested in occasionally show a large variability on short time scales. Since our study involves using observations, it is of utmost importance to understand their limitations. In this chapter we will describe the operational observational product we use, and discuss various sources of uncertainty and corresponding assessment studies. We will only discuss the operational product from nadir-viewing instruments on board the polar orbiting TIROS N and NOAA satellites. We will not include products from research and experimental satellites (e.g. UARS) which have provided valuable, but limited coverage data (limited either in time or space).

There are a few operational data products for the standard global observations of the stratosphere which use the same raw data (radiances), but differ either in the methods by which radiances are processed to get temperatures or in the methods of using the temperature fields to calculate other quantities. One of the major distinctions is between a stratospheric analysis product and an assimilation product. We use the term *analysis* for a product in which the satellite retrieved temperatures and geopotential heights are interpolated objectively onto a regular grid and then used to calculate other quantities assuming some balance. An *assimilation* product, on the other hand, combines the satellite retrievals with a numerical model, both to obtain the retrieved fields on a regular grid and to obtain other quantities. We use an analysis product for the current studies because we want to avoid the additional sources of uncertainty related to the model, and to have a product that is as simple as possible.

The stratospheric analysis product we use is compiled and distributed by the

NASA/GSFC Stratospheric Chemistry and Dynamics Branch. Temperature and geopotential height fields are provided on 18 levels (1000-0.4mb) by the NCEP Climate Prediction Center (CPC) for the stratosphere (70-0.4mb), and by the Global Data Assimilation System (GDAS) for the troposphere (1000-100mb). The reader is referred to McPherson et al. (1979) for a description of GDAS data. The data is regridded at NASA/GSFC onto a 5° longitude by 2° latitude grid.

The satellite data are constructed by retrieving layer mean temperatures from the radiances, and integrating to get geopotential heights, using the 100mb tropospheric analysis as a lower boundary condition. This boundary condition, referred to as a base level analysis, is a major source of error. The temperature and geopotential height fields are then interpolated on to a regular grid using a successive corrections method (Cressman, 1959), which consists of calculating corrections to an initial guess field (usually the previous day's analysis; Gelman and Nagatani, 1977). The same corrections method is used to add radiosonde data to the 70-10 mb northern hemisphere fields (Finger et al., 1965). Occasional radiosonde observations from above 10mb in the northern hemisphere and above 70 mb in the southern hemisphere provide a useful source of comparison to test the data (see section 1.3). Apart from the obvious errors involved in interpolating data from one grid to another, there are aliasing errors as a result of the asynoptic sampling of the satellite. Finally, winds, vorticity and Ertel's potential vorticity are calculated at GSFC from the geopotential heights using a balanced wind approximation (Randel, 1987a). We will now proceed to describe some of the stages described above in more detail, with an emphasis on the sources of uncertainties[1].

2.1 Retrieving temperatures

The current operational instrumentation package is the TIROS Operational Vertical Sounder (TOVS; Smith et al., 1979). It was first flown on the experimental TIROS N satellite, launched in 1978, and later on the NOAA6-NOAA14 operational satellites[2]. TOVS consists of three scanning radiometers; The High Resolution IR Spectrometer

[1]The operational analyses system has undergone many changes over the years. Since our work concentrates on specific wave events and not on interannual variability, we will concentrate on the current analysis, and the reader is referred to Randel (1992) and references therein.

[2]The first operational sounder was launched in 1972 on the NOAA2 satellite. It was succeeded by the next generation of sounders, launched in 1975 on NIMBUS 6. These earlier sounders provided the data for the first comprehensive observational studies of the stratosphere. Continuous operational stratospheric data is available since the Vertical Temperature Profiler Radiometer (VTPR) was launched, in September 1978. TOVS replaced the VTPR on October, 1979.

(HIRS), the Microwave Sounding Unit (MSU) and the Stratospheric Sounding Unit (SSU). These measure the radiances in a total of eight channels that peak in the stratosphere, only three of which peak at or above 10 mb (see figure 3.1 and section 3.2.2). The raw data from each of the three instruments is processed to produce radiances. This involves applying various corrections to the data (e.g. an antenna side lobe correction) and an extrapolation of the radiances from the SSU onto the HIRS scan spots because the former has a narrower scanning band than the latter[3]. These corrected radiances are further processed to obtain a set of spatially averaged clear-column radiances, by using cross calibration with MSU radiances to account for the contamination by clouds and water vapor. We will not discuss the errors involved in all these stages. They are described in detail in Smith et al. (1979), Kidwell (1986), and Kidder (1995). The final set of clear-column corrected radiances are inverted at the National Environmental Satellite Data and Information Services (NESDIS) to obtain the temperature profiles on 42 operational TOVS levels. The retrieval technique has evolved over the years, and is currently a Minimum-Variance method (see chapter 3 for a more detailed discussion). The inversion of a discrete set of radiances to obtain a vertical temperature profile is inherently ill posed, and it points out one of the major shortcomings of the data- its limited vertical resolution. The retrieved temperature profiles are integrated to give layer mean temperatures (layers above 14 km are: 200-100mb, 100-70mb, 70-50mb, 50-30mb, 30-10mb, 10-5 mb, 5-2 mb, 2-1mb, 1-0.4 mb). Operational temperature profiles are calculated on the operational analyses grid from these layer means using linear interpolation (the grid levels above 14 km: 100, 70, 50, 30, 10, 5, 2, 1 and 0.4 mb).

Studies that deal with the assessment of satellite retrievals usually compare the satellite observations to data from other sources, for example, different satellite instruments, radiosondes or rocketsondes. These studies usually compare data from specific days, or climatological fields, either of directly observed quantities or of diagnostics that are derived from them (e.g. Schmidlin, 1984, Barnett and Corney, 1984). Others compare the satellite retrievals of a specific event with observations that were taken at one time as part of a mission (e.g. Claud et al, 1998). Results vary, depending on

[3]This extrapolation process involves an unfortunate error (Laurie Rokke, personal communication, 1997). The two outer most spots of each scan line of the HIRS and MSU do not have a corresponding SSU spot. An extrapolation routine (the stratospheric mapper module) is used to extrapolate the SSU measurements to the two outer most spots of each scan line of the HIRS and MSU. This extrapolation routine does not work as it should and the SSU extrapolated data drifts off over time. At the time of speaking with L. Rokke (who is involved in developing an alternative retrieval system), the program was just being reinitialized every couple of weeks to get rid of the errors.

the instruments compared, on the years of data used or the specific events compared. Generally, the temperature differences between various instruments are around 5°K but some studies have reported a difference of 15-20°K in the upper stratosphere in early winter, while others have shown differences of only 2°K. Generally, time mean fields have fewer differences. Miles and O'Neill (1986) give an extensive reference list. An obvious limitation of all such studies is the existence of errors in all these measurements and the lack of a ground truth to compare them with.

A different approach, taken by Graves, (1986, see also Karoly and Graves, 1990) was to use a model to test the retrievals by applying them to the model generated fields. Graves used the SKYHI GCM as the truth, and used the operational routines used at the time by NESDIS, to retrieve temperature, geopotential heights, zonal winds and various other diagnostics. She also looked at how the model resolves wave statistics and wave amplitudes, but did not look systematically at wave structures. The general results show that *zonal mean* temperature can be retrieved to within 3°K in the stratosphere during periods when the evolution is dominated by large scale, slowly evolving dynamics. Amplitudes of planetary waves are off by about 20%, and sometimes there is also a phase lag. Higher order derived quantities are not captured very well by the retrievals, and the errors decrease if the fields are averaged either in time or zonally. In chapter 3 we use a similar but much simplified setup to examine the ability of the retrievals to resolve vertical wave structures.

2.2 Calculating geopotential heights: Errors due to base level analysis

Geopotential height is calculated by adding the operational layer mean temperatures to the 100mb tropospheric analysis. Errors in the base level analysis will, of course, affect the stratospheric analysis (see Trenberth and Olson, 1988, for an evaluation of tropospheric analyses). Karoly (1989) compared stratospheric circulation statistics calculated using one set of retrieved layer mean temperatures and a few different base level analyses. Errors in base level geopotential height were on the order of 100m, with the largest differences at high latitudes, over Antarctica. Not surprisingly, Karoly found errors to decrease with averaging, both of time and space, and to increase with the amount of spatial differencing. For example, daily zonal mean wind variations and the corresponding EP flux divergences had a similar sense but very different magnitudes (more than 50%) in the different analyses. Smoothing of the base level fields reduced some of the differences in highly differentiated fields like vorticity, but

large scale differences in the base level analyses were not removed. Geographically, differences were largest at high latitudes, especially above Antarctica. Other studies were conducted as part of the Middle Atmosphere Southern Hemisphere (MASH) project, with similar results (e.g. Grose and O'Neill, 1989).

2.3 Interpolation: Asynoptic sampling and aliasing

Unlike a radiosonde network, where all measurements are taken simultaneously at specified times, a satellite samples the domain in a continuous scan. Figure 2.1 shows part of the ground track typical of the NOAA satellites which operate in a near-polar, sun-synchronous orbit. At a given latitude, the satellite samples the whole circle twice

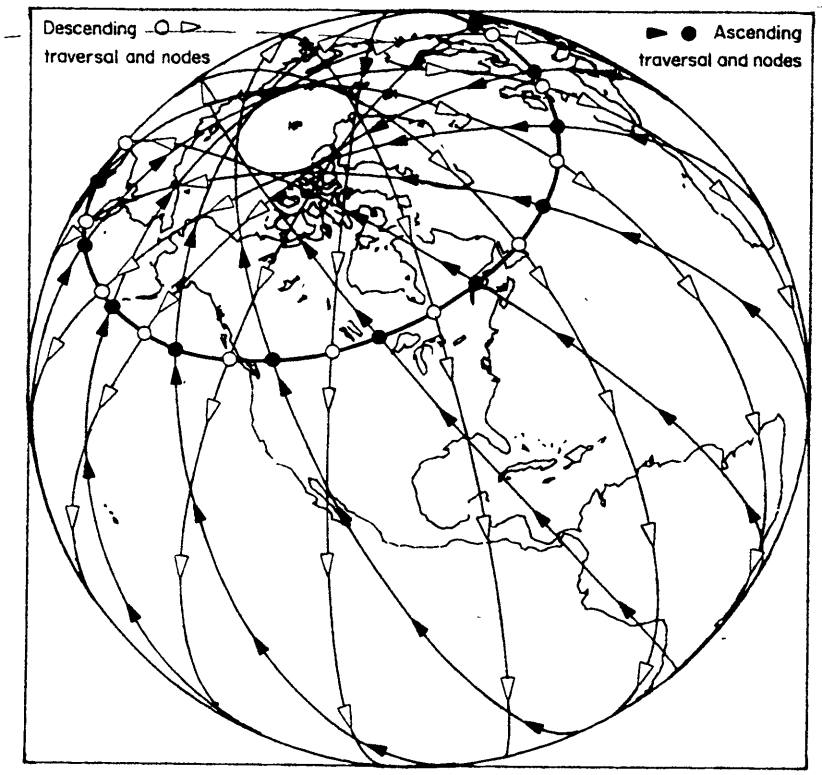

Figure 2.1: Trajectory of nadir observations viewed from a reference frame of the earth. Figure taken from Salby (1982a).

during a day, during the ascending and descending nodes of the orbit. There are about 14 cycles per day, meaning 28 samples in total. Salby (1982a) showed that for

a given latitude circle, asynoptic sampling is equivalent to a rotation of the frequency-wavenumber plane by an angle that depends on the drift-speed of the orbital plane relative to the earth's surface (see figure 2.2). Correspondingly, the region of aliasing is also rotated. Salby (1982b) derived a Fast Fourier Synoptic Mapping method[4] to account for this when doing a time-space spectral decomposition of the data as is generally done when analyzing the data for normal modes.

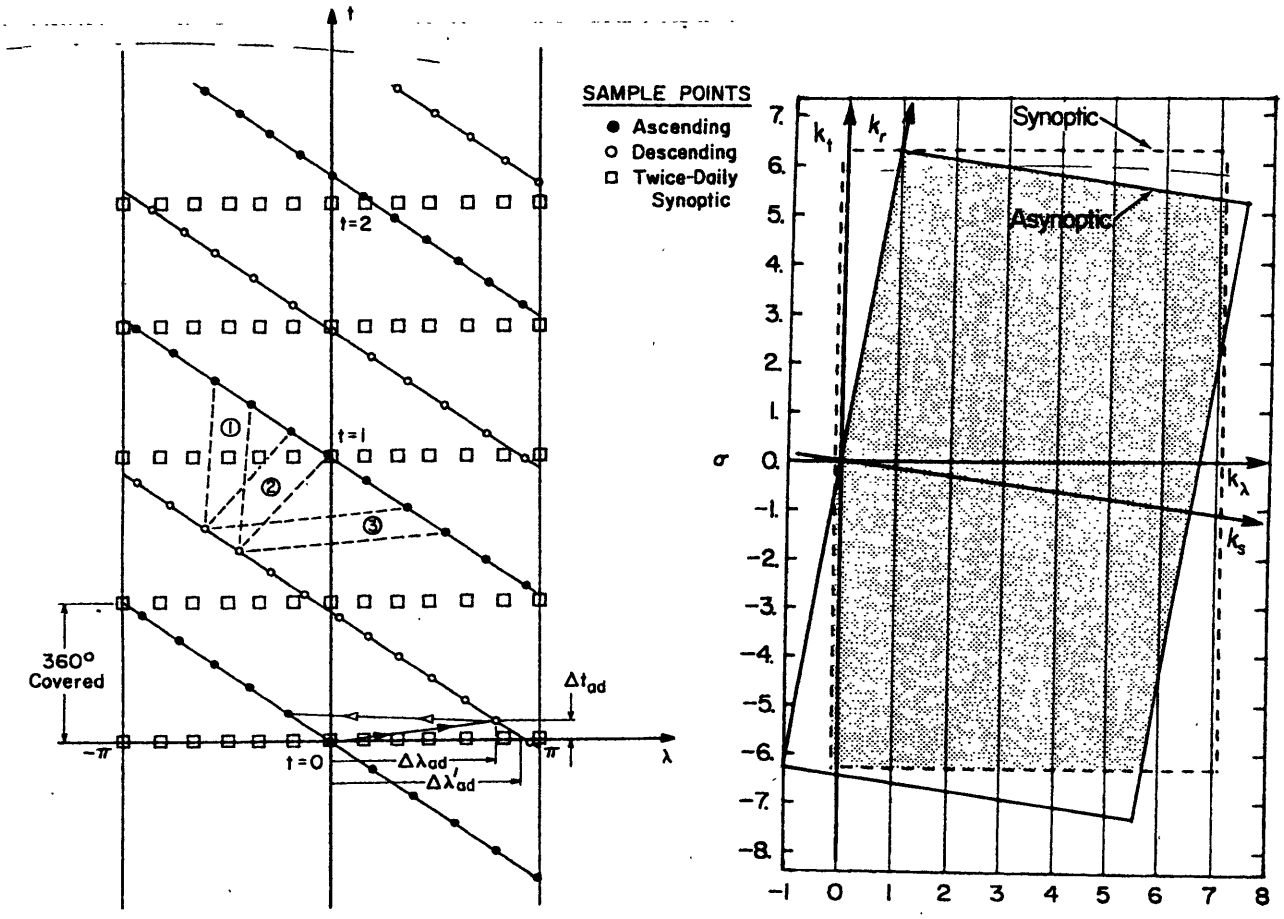

Figure 2.2: Left: The sampling pattern of observations on a latitude circle in the longitude-time plane. Shown are the ascending and descending nodes (full and empty circles, respectively), which are not equidistant, and the twice-daily synoptic pattern (squares). Right: The allowed wavenumber-frequency spectra for a twice-daily synoptic sampling (dashed) and a combined asynoptic sampling (solid). The shaded region corresponds to spectra resolvable in both types of observations. The frequency axis is $\sigma = \frac{2\pi}{days}$ with positive values for westward propagation, and the wavenumber axis is the integral wavenumber. Figures taken from Salby (1982a,b).

[4]The Fast Fourier Synoptic Mapping was applied by Lait and Stanford (1988a,b) to real data.

Observing a synoptic evolution is quite different, however. We need to worry about the effects of asynoptic sampling, when looking at the short time scale variations in wave structure (figure 1.7). The operational maps are created by combining all the data gathered in a 24 hour period as if it were gathered instantaneously, using a Cressman interpolation (Cressman, 1965), which essentially weights the observations around each grid point according to the distance from it. Graves (1986) studied the effects of asynoptic sampling by sampling the data of a GCM in the same way a polar orbiting satellite would, and interpolating it onto the operational grid using two methods. The first is essentially similar to the operational method and the second uses the Fast Fourier Synoptic Mapping (FFSM) scheme of Salby (1982b) to get rid of all aliased frequencies. The different methods did similarly well in simulating quiet periods. Interestingly, the simple 'instantaneous plotting' method did much better than the FFSM method during a sudden warming, with errors in zonal mean temperature of 3°K for the former (which were mostly due to the sampling of the model onto a lower resolution) and of as much as 20°K for the latter. The FFSM method threw out all the high frequencies which are naturally present in sudden changes. Graves did not specifically look at the vertical wave structures. Rather, she looked at zonal means and at maximum and minimum temperatures on a given latitude circle. In the following section we perform a simple sampling exercise to get a feel for the distortions involved in asynoptic sampling of vertical structure changes that occur on time scales of one to a few days.

2.3.1 Asynoptic sampling of a wave undergoing vertical structure changes

We specify analytically a very simple zonal wave 1 that is changing its vertical tilt in the course of a few days (figure 2.3). We use 28 grid points in the longitude direction, to facilitate sampling by a satellite that orbits the earth 14 times a day. We then sample the wave field as a satellite would, assuming the ascending nodes are exactly centered between descending nodes and coincident exactly with the grid points. Finally, we plot all the 'observations' taken in the course of one day on a single map. During times of rapid change, the alternation between ascending and descending nodes results in a 2-grid oscillation. When a 1-2-1 smoother is applied the fields are smooth, except for a jump at the longitude where the sampling day starts. The smoothing has an almost unnoticeable effect on the wave 1 component of the sampled field. We will show a pattern that is effectively moving eastward, because the sampling errors are expected to be larger than for a westward moving

wave. Since the satellite drifts westward relative to the earth, eastward phase speeds are essentially increased and westward phase speeds decreased. Figure 2.3 shows four consecutive days of a specified time-varying wave field, sampled at mid-day[5], along with the corresponding wave 1 component of the asynoptically sampled field and the difference between the two. The reduction in amplitude is of 30% for a wave that moves about 135° longitude in one day (i.e. a 2.67 day period, 86 m/sec at 60°S). To get an idea of observed phase variations, one of the fastest shifts observed in the southern winter of 1996 was of 90° longitude between August 11 and 12 at 1 mb (figure 1.7) which is smaller than the case tested (we need of course to take into account that these observations may be biased, but the calculation above suggests the distortions in phase are hardly noticeable even for the larger phase speed tested).

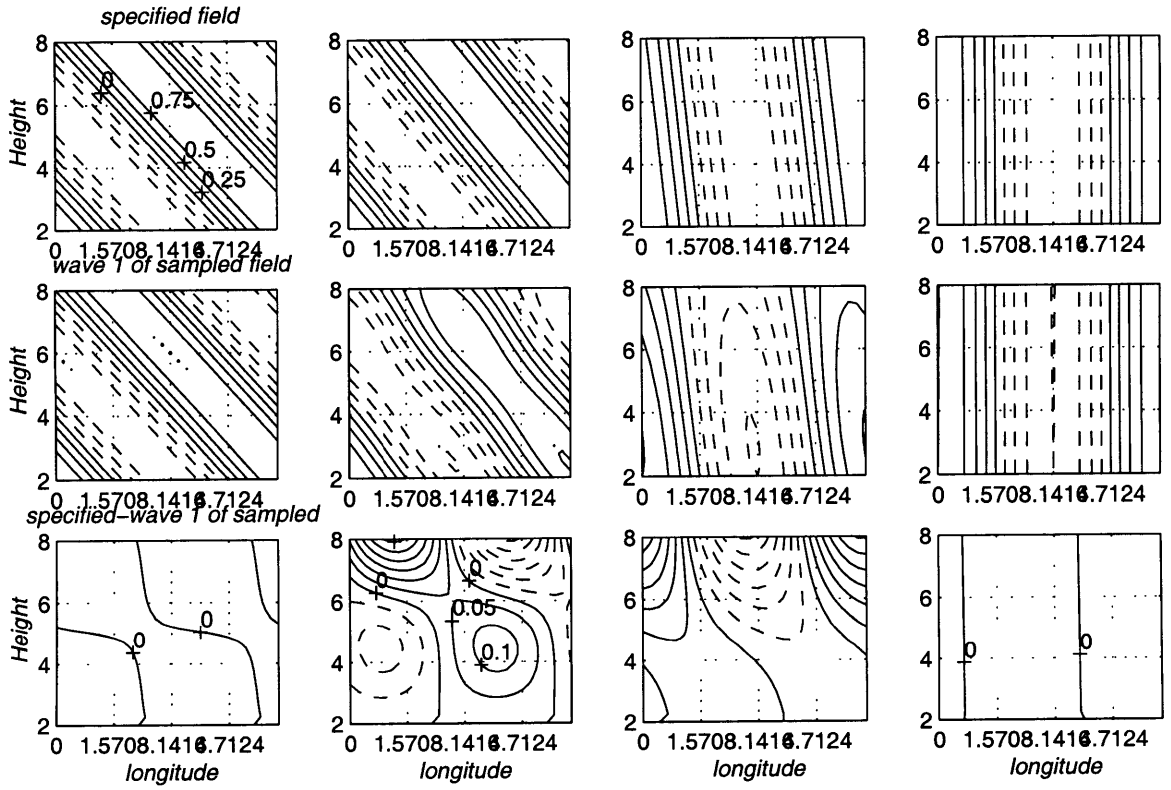

Figure 2.3: Longitude-height sections of four consecutive days (left to right) of a simple wave undergoing structure changes. Top: The specified field. Middle: The wave 1 component of the asynoptically sampled field. Bottom: The specified minus the wave 1 sampled field. Contour intervals: 0.25 top two rows; 0.05 bottom row; negative values are dashed. See text for details.

[5]We compare the asynoptic field sampled over the course of a day with the synoptic field at mid-day.

We have also tested variations in amplitude. We found maximum sampling errors of 5% for an increase with a time scale of 1 day, and 13% for a time scale of 0.5 days, which is much faster than growth expected in the stratosphere because the vertical group propagation times are at least one day, and the relevant instability growth rates are also much smaller (chapter 4). To sum up, distortions due to asynoptic sampling are noticeable only for rapid variations in vertical wave structure. The distortions are usually a decrease in amplitude, which increases with increasing phase speed. Distortions are slightly larger for eastward moving patterns. The general pattern, however, is captured by the asynoptic sampling. This effect may explain, at least partially, observed decreases in amplitude in the upper stratosphere that accompany structure changes (e.g. August 11-15, figure 1.7).

2.4 Winds and higher order diagnostics

Since there are no direct observations of winds in the stratosphere, they need to be inferred from the geopotential heights, by assuming some balance. Winds in the analysis product we are using are *balance winds* (see Randel, 1987a), which are calculated by dropping the time tendency and vertical advection terms from the zonal momentum primitive equations, and solving iteratively starting from geostrophy (as suggested by Gent and McWilliams, 1983). The largest errors are in the tropics, and in meridional winds and momentum fluxes in the upper stratosphere. The kind of balance used for calculating winds has a very large effect, especially on instantaneous winds and on higher order quantities which are derived using winds. For a thorough discussion, see Randel (1987a). Balance winds are a significant improvement over geostrophic winds, which were used in many earlier studies (e.g. Geller et al, 1983, 1984, Hartmann et al, 1984, Mechoso et al., 1985). Geostrophic zonal mean winds are generally too strong (Randel, 1987a), and the geostrophic EP flux divergence is an overestimate of as much as 100% (Robinson, 1986). This is especially important in light of the puzzling observation in most of these studies, of large regions of positive $\nabla \cdot \vec{F}$ in the upper middle and high-latitude stratosphere (implying a source of wave activity)[6].

The geopotential heights are also used to calculate higher order diagnostics like the EP fluxes and various quadratic wave fields. The uncertainties involved in these calculations, apart from those associated with calculating winds, are mostly due to

[6]It is important to note that the problem is not solved completely. Even with the use of balance winds, there is still a small region of positive $\nabla \cdot \vec{F}$, in the upper high latitude stratosphere that is not understood.

the low resolution of the observations, especially in the vertical. The uncertainties involved, both due to the assumption made in calculating winds, and due to the low resolution, have led us to look at temperature and geopotential height fields of the waves, which are the most directly observed quantities.

2.5 Summary and the relevance to our study

In this chapter we have discussed the process of putting together the operational observations product. There are a few stages to this process, which are obtaining the raw data (radiances), retrieving temperatures from the radiances, calculating geopotential heights from the temperatures, interpolating onto a regular grid, and calculating higher order quantities from the geopotential heights and temperatures. Each of these stages involves uncertainties and errors.

We are especially interested in how these uncertainties affect the observations of planetary waves, in particular their vertical structure. We have shown, by performing a simple exercise, that asynoptic sampling can distort the observations in periods of very rapid variations. The main effect is to decrease the amplitude of the perturbation. We expect this effect to be small for realistic structure variations and phase propagations. The base level analysis will not have a large effect on the *vertical structure* of the waves (it will have no effect on temperatures). The largest uncertainty comes from the coarse vertical resolution inherent to the temperature retrieval process. There have been theoretical studies of the accuracy of temperature retrievals. Most papers that suggest a new retrieval method also looks at the ability of the retrieval algorithm to reproduce an isolated temperature profile (e.g. Smith, 1970, Smith and Woolf, 1976). These studies, however, mostly emphasize the ability to simulate the vertical structure of the temperature field, and do not look generally at the ability to resolve specific dynamic phenomena. The ability of the retrievals to resolve wave structures, in particular, has never been looked at to our knowledge. It is important to note in this context, that a wave field is the deviation from the zonal mean profile, which varies substantially with height. A given satellite retrieval, therefore, may capture the vertical structure of the total temperature field sufficiently well, but still not resolve the vertical *wave* structure satisfactorily.

In chapter 3 we discuss the information content of the retrievals and deal with the ability to use the retrievals to study vertical structures. Our approach is to use a model to give us waves, and to check how well we can retrieve them. This is different from most observational assessment studies because we have a ground-truth to compare with. It is different from Graves' (1986) study, which took a similar

approach only using a GCM, because we are using a simple model and are looking specifically at the dynamical phenomena we are interested in. Graves, in her study did not look at vertical structures specifically.

Chapter 3

The 'Virtual Satellite' problem

3.1 Introduction

In the previous chapter we described the operational observational product and discussed the uncertainties involved. In this chapter we investigate the uncertainties of the temperature retrieval stage in much more detail, and look at the ability of satellite observations to resolve planetary waves.

As we saw in chapter 1, at times, observed waves exhibit a large degree of time variability on short time scales in their vertical structure (e.g. figures 1.7). The time variations are an interesting feature that we want to understand. It is very important, as part of our study, to have a good idea of the ability of observations to capture wave structures. It is conceivable that the errors in wave structure are large enough and random enough to appear in our data as short time scale variability in wave structure (as in figure 1.8). We want to know how real the phenomena we see are. The approach we will take in this study is to use a model to give us waves, and to check how well we can retrieve them[1]. In the following chapters we use a model of stratospheric planetary waves to study their linear dynamics. Details of the model are found in chapter 5 and appendix B. We use the same model here. Generally, we characterize waves quantitatively and qualitatively by the amplitude and phase structure. We will therefore concentrate on understanding how well the amplitude and phase of waves are resolved by the satellite observations.

[1]See section 2.1 for a discussion on previous approaches to assessing satellite retrievals, and how they differ from the present study.

3.1.1 Outline of experiment

Using a model of quasi geostrophic (QG) stratospheric waves, we obtain temperature fields that have stratospheric waves in them. For the model run, we specify a vertical profile of Brunt Vaisala frequency, a zonal mean wind as a basic state and a geopotential height perturbation at the bottom of our model as forcing. We then run the model to obtain a perturbation geopotential height field, from which we calculate a temperature field (see appendix A.1 for details).

We calculate the radiances a satellite sitting at the top of our model would see, and then take the radiance's and apply some inverse technique to get the retrieved temperature fields. This is done at every horizontal model grid point, to get a three dimensional retrieved wave temperature field. We use a few retrieval algorithms and look at a variety of waves, mostly to see what the retrieval algorithm does to wave structures (amplitude and phase of temperature).

3.2 The virtual satellite

3.2.1 The basic principles of remote sounding

Remote sensing of temperature is based on the fact that the radiation emitted at a given wavelength, from a gas in local thermodynamic equilibrium, is a function of the local temperature, through the Planck law, and of the concentration of the relevant emitting gases. Emission by gases with a well known concentration that is essentially constant with height (CO_2, O_2) is useful in determining temperature (Kaplan, 1959)

The relevant process is emission resulting from molecular vibrational-rotational transitions. Temperature soundings mostly utilize the fundamental vibrational transition bands. Since the emission is much stronger at the center of a band than at the edges of it, the atmosphere is more optically thick at these frequencies, and most of the radiation is emitted from a thin high layer. Correspondingly, the radiation from the sides of the band originates at lower levels. A satellite instrument measures radiance over a narrow frequency band that is much narrower than a vibration emission band. This allows it to sense different levels of the atmosphere, by measuring at different positions along the band.

We define an optical depth $\tau = \int_z^\infty k_\nu(z)\rho(z)dz$, where k_ν is the monochromatic extinction coefficient, ρ is the density of the absorbing/emitting gases and z is log pressure. The total monochromatic radiance at an optical depth τ, viewed by a sensor

looking downwards along the local vertical is:

$$I_\nu(\tau) = I_\nu(\tau_o)e^{-(\tau_o-\tau)} + \int_\tau^{\tau_o} B_\nu[T(\tau')]e^{-(\tau'-\tau)}d\tau' \tag{3.1}$$

where T is temperature, $B_\nu[T(\tau)]$ is the Planck function, and τ_o is the optical depth at the bottom level, in our case, the lowest model level, which is at 2 scale heights.

The total monochromatic radiance at the top of the atmosphere, viewed by a satellite looking directly down is obtained by choosing $z = \infty$. The equation becomes simpler if we define a monochromatic transmittance function and write the equation in log pressure coordinates (z) as follows:

$$\mathcal{T}_\nu(z) = e^{-\tau(z)} \tag{3.2}$$

$$I_\nu(\infty) = I_\nu(z_s)\mathcal{T}_\nu(z_s) + \int_{z_s}^{\infty} B_\nu[T(z)]\frac{\partial\tau}{\partial z}dz \tag{3.3}$$

where we have used the fact that $\mathcal{T}_\nu(\infty) = 1$ ($\tau(\infty) = 0$). z_s denotes the bottom boundary. $W_\nu(z) = \frac{\partial\tau}{\partial z}$ is called the weighting function. The shape of the weighting function indicates the levels that contribute most to the radiance at the given frequency ν.

3.2.2 The satellite instruments and transmittances

The current satellite that remotely senses temperature is NOAA 14, which is a polar orbiting satellite from the TIROS N series. The relevant instrumentation package is the TIROS Operational Vertical Sounder (TOVS), which consists of three scanning radiometers; The High Resolution IR Spectrometer (HIRS), the Microwave Sounding Unit (MSU) and the Stratospheric Sounding Unit (SSU).

The HIRS instrument utilizes the 15μm and 4.3μm CO_2 IR bands, which are used to sense temperature. In addition, it uses water vapor absorption in the 6.3μm band to sense moisture, and four window channels that sense surface temperature or detect clouds. There are 19 operational channels, 11 of which are used for temperature soundings, only four of which peak at or above the bottom of our model (2 scale heights).

One of the MSU's main purposes is to make temperature soundings in the presence of clouds. It utilizes an O_2 absorption band. The horizontal resolution is much coarser than the HIRS resolution because the wavelength is much longer. Only one channel peaks above 2 scale heights. We will not use this one channel for reasons that will become clear later on.

The SSU measures radiation near the center of the 15μm CO_2 band. It is used for temperature soundings in the stratosphere. It has three channels that peak at 1.5,5,15 mb. Since the measurements near the peak of the absorption band have to be very narrow in frequency space, and since there is little emission because the density at high levels is low, a special technique that uses pressurized CO_2 cells to filter the radiation is used. Taylor et al. (1972) describe this instrument in detail and also derive a simple equation for the weighting function:

$$W_\nu(p) = \frac{\mathcal{P}^2}{(1+\mathcal{P}^2)^{3/2}} \tag{3.4}$$

$$\mathcal{P}(p) = \frac{\sqrt{2}p}{p_{peak}} \tag{3.5}$$

p is the pressure, and p_{peak} is the pressure at which the weight function peaks. p_{peak} depends on the pressure in the CO_2 cell and on various instrument parameters. For a more detailed description of these instruments, see Smith et al. (1979), Kidder and Vonder Haar (1995) and Kidwell (1986).

Vibrational transitions are accompanied by finer rotational transitions, which are separated into three branches, P, Q, R, corresponding to jumps of -1, 0, 1 of the rotational quantum number. The satellite instrument response function is much wider than these rotational lines. The radiance measured by the satellite instrument is therefore an integral of equation 3.3 over frequency, weighted by the instrument response function. Since the integration is over many absorption lines, the absorption coefficient is highly variable and is generally done using approximate band models or other approximate radiation codes. The Planck function, however, varies slowly enough with frequency, hence the integration over frequency is applied only to the transmittance function. In the current study, a fast transmittance model called OPTRAN (McMillin et al., 1995) was used for the HIRS and MSU instrument channels. Equations 3.4 and 3.5 were used to calculate the SSU weighting functions. Figure 3.1 shows all the weight functions that peak at or above 14 km (the bottom of our model). The channels and the heights of the corresponding weighting functions are also listed in table 3.1. The solid lines correspond to SSU channels 1-3 and HIRS channels 1-3, which use the 15μm CO_2 band) and to HIRS channel 16, which uses the 4.3μm CO_2 band. The dashed line corresponds to MSU channel 4. All channels except the latter one are used in this study. The reasoning for not using the latter is as follows. The MSU channel peaks at around 70 mb, at a region that is well covered by other HIRS channels, and is relatively narrow. We are mainly interested in the middle and upper

stratosphere, at heights where there are no radiosonde observations and the MSU has very little influence there. We will show later (section 3.3.4) that since this channel overlaps two HIRS channels so much, it will not add much new information. We will also show later (section 3.3.3) that the retrieval becomes much simpler if we assume a single frequency for all channels. We therefore prefer to throw out the MSU channel, which has a very different frequency from the rest.

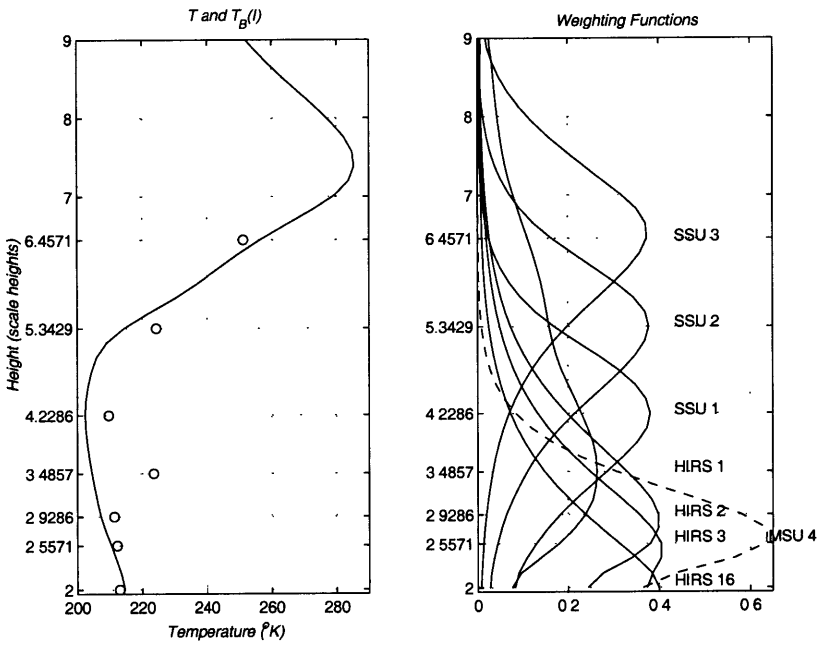

Figure 3.1: Left: Brightness temperatures (circles) corresponding to the radiances measured by each of the channels used in this study for the temperature profile shown (solid line), plotted at the peaks of the corresponding weighting functions. Right: The weighting functions that peak in our model domain. Solid lines are the channels used in our study (marked next to the peaks) and the dashed line is the MSU channel we drop. In all relevant figures, the grid lines below 7 scale heights are at the peaks of the weighting functions.

3.2.3 Calculating the radiances

Calculating the radiance that a satellite would measure at a given frequency, for a given temperature profile, is referred to as the forward problem. The transmittance functions are plugged into equation 3.3 to calculate a set of radiances at each horizontal model grid point. The radiances are then used as input for the temperature retrieval schemes. Once the weighting functions are known, all we need for calculating the radiances is the vertical profile of the Planck function, which is a function of the temperature profile and the channel frequency. In reality, the frequency is determined

by the satellite instrument, hence the correct one should be used to retrieve temperature. In our virtual satellite problem, we are free to specify any frequencies we want as long as we keep them the same in the forward and inverse problems. The ability to resolve temperature structures will not be affected by our choice of frequencies. The problem is much simpler to solve when we use the same frequency for all channels (see section 3.3.3), because we can solve the inverse problem for the Planck function directly.

Figure 3.1 shows the brightness temperatures (T_b), corresponding to the radiances that were calculated from a given temperature profile (shown) using equation 3.3, and plotted at the height of their corresponding weighting functions, which are shown on the right. We can see that the brightness temperatures correspond roughly to the mean temperature of a layer centered at the peak of the weighting function. It is clear that HIRS channel 1 is different from the rest because it is very deep. The brightness temperature deviates significantly from other channels, and doesn't follow the actual temperature profile like the rest of the channels do. This channel proves problematic in one of the retrieval methods we use (referred to later as Chahine's method), hence, we drop it when we use that method. Finally, this calculation was done using the SSU wavenumber of 668 cm^{-1} for all channels. When we repeat the calculation for the actual TOVS frequencies (table 3.1), we get almost identical brightness temperatures, except for the surface one which is a couple of degrees different. This supports our assumption that we are free to use a constant frequency as long as we are consistent in the forward and inverse calculations.

Table 3.1 lists the instrumental errors of the channels used in this study, in terms of a temperature error at a given mean scene temperature. These numbers are taken from the NOAA POD guide (1997). Also shown are the equivalent errors in radiance, using a wavenumber of 668 cm^{-1} for all channels. These errors will be used later in this study to test the sensitivity of our retrievals to noise (section 3.4.1). The true channel wavenumbers are listed for reader information only. They were not used in the study.

The bottom boundary

Generally, the retrieval problem is solved for the troposphere-stratosphere, and the bottom boundary is the surface. For many of the channels we are interested in, $\mathcal{T}_\nu(0) = 0$, hence the surface term does not contribute to equation 3.3. For channels that have a surface contribution, we use a gray or black body law:

Instrument and channel	Peak of weighting function (km)	ΔT(°K) at mean scene temperature	Mean scene temperature (°K)	ΔI $10^{-5}\frac{W/m^2}{Sr\ m}$	Instrument wavenumber (cm^{-1})
HIRS 16	14.0	0.31	230.0	0.32	2265.5
HIRS 3	17.9	0.55	220.0	0.51	689.22
HIRS 2	20.5	0.74	220.0	0.68	682.22
HIRS 1	24.4	2.77	235.0	2.96	668.51
SSU 1	29.6	0.25	273.0	0.36	668.00
SSU 2	37.4	0.50	273.0	0.72	668.00
SSU 3	45.2	1.25	273.0	1.79	668.00

Table 3.1: The channels used in this study. Error data is taken from the NOAA POD guide (1997).

$$I(0) = B[T(0)]\epsilon_\nu \tag{3.6}$$

where (ϵ_ν) is the surface emissivity (equals 1 in the IR domain, to a good approximation).

In our wave model, however, the bottom boundary is above the tropopause, at two scale heights (14km) and equation 3.6 does not necessarily hold. To simplify matters, we still use this relation. We can derive it if we assume the Planck function is linear with height in the troposphere, and the transmittance at the surface is negligible, which is true of all channels used in this study. The transmittance of 5 of the 7 channels used is less than 0.1 at our model bottom boundary, hence the exact value of ϵ_ν is important only for the 2 channels for which it is larger. We tested this by using various values of ϵ_ν, ranging from 0.5 to 1.5. The resultant temperature retrievals were almost identical, with the differences being at the surface. These were around 2°K for the Chahine retrieval (section 3.3.2), and much less for the Minimum Variance retrievals (section 3.3.3).

3.3 The Inverse problem

3.3.1 General outline and solvability

The inverse problem consists of solving the set of M equations, M being the number of satellite channels, to get a vertical profile of temperature:

$$I_i(\infty) = B_i[T(z_s)]\mathcal{T}_i(z_s) + \int_{z_s}^{\infty} B_i[T(z)]W(z)_i dz \qquad i = 1...M \tag{3.7}$$

The subscript i denotes the frequency of the i'th channel ν_i. The problem is inherently ill posed, since we are trying to solve for a continuous profile using a discrete set of measurements. We therefore need to settle for less. A possible way is to solve for the temperature at N discrete levels, where N has to be less than or at the most equal to M, along with some assumption about interpolation between the discrete levels. Physically, the satellite senses layer mean temperatures, and we can only obtain a temperature profile with low vertical resolution.

There is a great body of literature that deals with the inversion of satellite IR measurements to obtain temperatures (Rodgers, 1976 is an excellent review paper). There are two general kinds of papers. One deals with determining the inherent vertical resolution of the inverse solution (Mateer, 1965, Conrath, 1972 Backus and Gilbert, 1970, Rodgers, 1990). Such papers generally show there is a tradeoff between resolution and sensitivity to noise, where higher resolution also implies higher sensitivity to noise in the measurements. The second kind of study deals with finding practical inversion methods using various approaches. Some methods give a local temperature at M discrete levels (Chahine, 1970), some express the vertical temperature profile as a linear combination of M basis functions and find the relevant coefficients (Mateer, 1965), some use a purely statistical regression approach (Smith and Woolf, 1976), and some combine the inverse solution with a statistical solution, the way statistical measurements are combined, using error covariance matrices as weights (Rodgers, 1976, and references therein). The latter approach is used operationally. Rodgers (1976) in his review explains how all methods give essentially the same solution with the differences stemming from the differences in the additional information that is supplied in order to solve the inverse problem.

In the current study, we retrieve temperatures using two methods. The first is a nonlinear algorithm that solves for the temperature at six discrete levels. This method is used because it is independent of additional statistical information, hence it is objective. The second method retrieves a continuous profile by combining the retrieval with a statistical constraint. In order to simplify the solution we make the problem linear. This method is used because it is similar to the operational temperature retrievals (see appendix for more details on the operational retrieval). The combination of using these two different retrieval algorithms will separate the robust features from those that are method dependent, in particular, features that depend on the specifics of the statistical constraint.

3.3.2 Chahine's retrieval algorithm

Chahine (Chahine, 1968, Chahine 1970) developed an iterative method to solve equation 3.7, for a set of discrete temperature values at the peaks of the weighting functions. He bases his analysis on the fact that the weighting functions have a well defined peak, and most of the contribution to equation 3.7 comes from a narrow region in the vertical, and approximates the relation between the radiances of two different temperature profiles as follows:

$$\frac{I_i(\infty) - B_i[T(z_s)]\mathcal{T}_i(z_s)}{\tilde{I}_i(\infty) - B_i[\tilde{T}(z_s)]\tilde{\mathcal{T}}_i(z_s)} = \frac{B_i[T(z_i)]W_i(z_i)\Delta_i z}{B_i[\tilde{T}(z_i)]\tilde{W}_i(z_i)\tilde{\Delta}_i z} \tag{3.8}$$

The subscript i denotes the i'th channel, z_i is the height (log pressure) of the peak of the i'th weighting function, and $\Delta_i z$ is the effective width of the contributing region, defined as follows:

$$\Delta_i z = \frac{I_i(\infty) - B_i[T(z_s)]\mathcal{T}_i(z_s)}{B_i[T(z_i)]W_i(z_i)}$$

T and $\tilde{T}$ are two different temperature profiles, and $\sim$ denotes a function of $\tilde{T}$. If the contribution comes from a narrow enough region (i.e. $\Delta_i z$ is small enough) and the variation of the weighting functions with temperature is much smaller than the variation of the Planck function with temperature, and if the contribution to the radiance by the bottom boundary term is either negligible or dominant, we get the following approximation:

$$\frac{I_i(\infty)}{\tilde{I}_i(\infty)} \approx \frac{B_i[T(z_i)]}{B_i[\tilde{T}(z_i)]} \tag{3.9}$$

Equation 3.9 is the basis of the iteration method. We can start off with a given temperature profile, T^0 given at the peaks of the weighting functions, and use it to integrate equation 3.7 for each channel, to obtain a set of radiances I_i^0. We can then use the set of calculated radiances, the measured radiances ($\tilde{I}_i$), and the initial temperature profile, to obtain a new profile T^1 from equation 3.9. We can repeat this for every channel, till convergence of the calculated radiances towards the observed ones is reached, as follows:

$$T_i^{n+1} = \frac{2h\nu_i^3 c^2}{\ln\left(\frac{I_i^n}{\tilde{I}_i}(e^{hc\nu_i/KT_i^n} - 1) + 1\right)} \tag{3.10}$$

where we have used

$$B_\nu(T_i) = \frac{2h\nu_i^3 c^2}{e^{hc\nu_i/KT_i} - 1} \tag{3.11}$$

T_i^n and I_i^n are respectively the n'th iteration temperature and radiance corresponding to the i'th channel. h is the Planck constant, c the speed of light, K the Boltzmann constant. We repeat this until the radiances I_i^{n+1} that are calculated from all the T_i^{n+1}, using equation 3.7, converge to the observed radiances to within a specified limit[2].

We start the iteration from a constant temperature profile of $300°K$. Since this algorithm assumes the weighting functions are narrow, it is not surprising that HIRS channel 1, which has a very wide weighting function, degrades the retrieval. What we get looks similar to the profile shown in figure 3.1, where the temperature corresponding to HIRS channel 1 is much larger than the true profile. This channel is therefore not used in Chahine's algorithm.

The Chahine retrieval method, from the start, points out the inherent limitations of the satellite observations, namely, even if it were to do a perfect job by retrieving the exact temperature at the six levels, we are retrieving only six points. There is no information above 1.5 mb and the resolution is at best as good as the distance between the peaks of the weighting functions. This method is useful because it is nonlinear and can be used to test whether using one frequency for all channels affects our results. Comparisons between retrievals using the correct frequency and just one frequency shows the retrieved temperatures to be almost identical. We will therefore only show results from linear runs in this study. All runs use the SSU channels' wavenumber of 668 cm^{-1}.

This brings us to the next section, where we look at a retrieval method that gives a continuous profile, instead of a discrete one. The way this is achieved is by adding a climatology profile to the retrieval that effectively fills in the gaps between the points we get from Chahine's method. The fact that we get a smooth profile that extends much higher than 1.5 mb does not mean we can see more levels or higher up, or that we can resolve more, it just means we have added more information.

[2]In order to calculate the radiances from the set of six temperature points, we use a bicubic spline interpolation to interpolate the temperature onto a higher resolution grid, and integrate using a simple trapezoidal integration. This interpolation is the additional information we use in the Chahine retrieval.

3.3.3 The Minimum Variance method

In the following section we describe a linear retrieval algorithm, which combines the inverse solution with additional statistical information, to get a continuous profile. Our derivation follows Rodgers (1976), with slightly different notation. In general the problem is nonlinear, due to the nonlinearity of the Planck function. If, however, we use the same frequency for all channels, we make the problem linear because we can invert the radiances to obtain $B(T)$ directly, instead of T. We can then write equation 3.7 as a matrix equation by discretizing it onto N grid points in the vertical direction:

$$\mathbf{r} = \mathbf{K}\mathbf{b} \tag{3.12}$$

$\mathbf{r}$ and $\mathbf{b}$ are M and N dimensional vectors respectively representing the radiance measured by the different satellite channels and the discretized vertical profile of the Planck function:

$$\mathbf{r} = \begin{pmatrix} I_1 \\ . \\ . \\ . \\ I_M \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} B(z_1) \\ . \\ . \\ . \\ B_{(z_N)} \end{pmatrix}$$

$\mathbf{K}$ is an $M \times N$ matrix, that represents the discrete version of the integral in equation 3.3 as well as the bottom boundary transmittance. Mathematically, $\mathbf{K}$ can be inverted to solve for $\mathbf{b}$ only if $N \leq M$. Assuming we can, we define the matrix $\mathbf{D}$ such that:

$$\mathbf{b} = \mathbf{D}\mathbf{r} \tag{3.13}$$

therefore, $\mathbf{KD} = \mathbf{1}$

If we combine our inverse solution with an additional constraint, for example, a climatology (a mean profile and an error covariance matrix for the variance around this mean), we can get a retrieval that has a much higher resolution than M levels.

Let us call the part of the solution that is a direct inversion of the radiances the *inverse solution.* In the case where the number of measurements would equal or exceed the number of layers in our retrieval, this would correspond to $\mathbf{b} = \mathbf{Dr}$. Let us also call the profile with which we combine our inverse solution the *constraint profile*, and denote it by $\mathbf{b_o}$. The corresponding error covariance matrix is denoted by $\mathbf{S_o}$ [3].

[3]The error covariance matrix is calculated as follows. Supposing our climatology is a set of L profiles $\mathbf{x_l}$, where the subscript l denotes the measurement number. We take our constraint profile

The error covariance matrix for the inverse solution will be a function of the error covariance of the radiances, $\mathbf{S}_\epsilon$. Since the errors in the different satellite channels are uncorrelated, $\mathbf{S}_\epsilon$ is a diagonal matrix, with the square of the standard deviations as the diagonal values. The corresponding error covariance matrix of the inverse solution is:

$$\mathbf{S_r} = (\mathbf{K}^T \mathbf{S_\epsilon}^{-1} \mathbf{K})^{-1} \tag{3.14}$$

The inverse solution is combined with the constraint profile, by inversely weighting by the error covariance matrices, as follows:

$$\hat{\mathbf{b}} = (\mathbf{S_o}^{-1} + \mathbf{K}^T \mathbf{S_\epsilon}^{-1} \mathbf{K})^{-1} (\mathbf{S_o}^{-1} \mathbf{b_o} + \mathbf{K}^T \mathbf{S_\epsilon}^{-1} \mathbf{r}) \tag{3.15}$$

Note that equation 3.15 is equivalent to writing

$$\hat{\mathbf{b}} = (\mathbf{S_o}^{-1} + \mathbf{S_r}^{-1})^{-1} (\mathbf{S_o}^{-1} \mathbf{b_o} + \mathbf{S_r}^{-1} \mathbf{Dr}) \tag{3.16}$$

where $\mathbf{Dr}$ is the inverse solution $\mathbf{b}$, only 3.16 is singular when the matrix equation 3.12 is ill defined (when $N > M$), whereas equation 3.15 is not.

The covariance of the solution $\hat{\mathbf{b}}$ is:

$$\hat{\mathbf{S}} = (\mathbf{S_o}^{-1} + \mathbf{K}^T \mathbf{S_\epsilon}^{-1} \mathbf{K})^{-1} \tag{3.17}$$

Some matrix manipulation results in a different form of equation 3.15 that will prove useful later on, and involves fewer matrix inversions:

$$\hat{\mathbf{b}} = \mathbf{b_o} + \mathbf{S_o} \mathbf{K}^T (\mathbf{K} \mathbf{S_o} \mathbf{K}^T + \mathbf{S_\epsilon})^{-1} (\mathbf{r} - \mathbf{K} \mathbf{b_o}) \tag{3.18}$$

This equation constitutes the second retrieval method used in this study, along with Chahine's method. The nonlinear variation of this method can be constructed for nonlinear problems by linearizing the equations around some profile. Such an iterative procedure that starts with the constraint profile $\mathbf{b_o}$, and converges to equation 3.18 is used by most operational centers that retrieve temperature from radiances (see Rodgers, 1976, equation 99).

($\mathbf{b_o}$) to be the mean of this sample and the error covariance matrix (*error* because we treat the climatology as a measurement with deviations around it) is defined as follows:

$$\mathbf{S_o} = \frac{1}{L} \sum_{l=1}^{L} [(\mathbf{x_l} - \mathbf{b_o})(\mathbf{x_l} - \mathbf{b_o})^T]$$

Since the retrieved profile is a combination of an inverse solution and an a priori constraint, the main question is what part of the retrieval is sensitive to the constraint we use, and what part not. Ideally, the effect of the constraint is to fill in the information in parts of the solution that are not resolved by the observing system, hence it is an important part of the solution. We will also see that the error covariance matrix is important in determining how the inverse and constraint profiles combine.

We use various constraints in this study. We divide the constraints into two main kinds. The first (referred to as a diagonal constraint) has a diagonal error covariance matrix, which means all vertical correlations in the retrieval are due only to the observing system. The second kind of constraint has a non-diagonal error covariance matrix (we refer to it as a non-diagonal constraint), hence the vertical correlations in the retrieval are due also to the correlations in the constraint. The latter one is used for the operational retrievals. Appendix A.2 describes how we construct the various constraints.

In the following sections we will show results of retrieving various temperature fields from our QG model, using both Chahine's method (CH retrieval) and the method just described, which is commonly referred to as a Minimum Variance method (MV retrieval). We will look at the effect of using various constraints, and corresponding error covariance matrices on the retrieval. Before describing the various runs and results, we will look at the issue of vertical resolution, using some diagnostics on the weighting functions and on the constraint field, that will help explain some of the results we get.

3.3.4 Vertical resolution

There have been quite a few studies that deal with the issue of the vertical resolution of the observing system (see the review by Rodgers, 1976, for a detailed reference list). In this section we will describe a few diagnostics that will be useful in illuminating the issues of resolution and noise sensitivity in our study.

The first such diagnostic is the set of eigenvectors and eigenvalues of the weighting functions, which give us an idea of the vertical structures that can be resolved by the observing system. This was first introduced by Mateer (1965) for Umkehr ozone soundings. Here we follow Rodgers (1976), who discusses this in context of temperature soundings. One way to solve equation 3.12 is to expand $\mathbf{b}$ onto a set of M or less basis functions. An optimal choice of basis functions is a linear combination of the weighting functions[4]. One such combination, which is also orthogonal, is the set

[4]Using a linear combination of the weighting functions as a choice of basis functions minimizes

of eigenvectors of $\mathbf{KK^T}$. Assuming that the radiance measurements are independent of each other and are measured with an error variance of σ, it can be shown that the corresponding eigenvalues (λ_i) are inversely proportional to the measurement error variance, as follows:

$$\lambda_i = \frac{\sigma^2}{\hat{\sigma}_i^2} \tag{3.19}$$

where $\hat{\sigma}_i$ is the measurement error variance corresponding to the i'th eigenvector. We see that eigenvectors with a large eigenvalue will have a small measurement error variance, which means they are resolvable by the observing system. Eigenvectors with corresponding small eigenvalues are hard to measure with accuracy.

Figure 3.2 shows the eigenvectors that were calculated form the weighting functions in figure 3.1. The corresponding eigenvalues are shown by each of them. It is clear that only 3-4 eigenvalues are practically resolvable, because there is a drop of a few orders of magnitude between λ_1 and λ_5. These eigenvalues are the ones with the largest vertical structures. Eigenvalues with smaller structures are not resolvable, which makes sense. It is interesting to note that we have fewer independent pieces of information than satellite channels. This is due to the large overlap between the weighting functions. This fact is one of the main reasons why we can drop HIRS channel 1 in the Chahine retrieval, without significantly decreasing the resolution of the retrieved profile. Also, this explains why we can drop MSU channel 4 (section 3.2.2). We repeated this analysis on the set of 8 weighting functions, including the MSU channel. The resulting eigenvectors were almost the same, and the eigenvalues of the largest ones increased by a few percent. This means the MSU channels does not add much new independent information. This makes sense since it peaks in the lower stratosphere where there is a relatively large overlap between the weighting functions.

Another diagnostic that will prove useful in understanding the minimum variance retrieval is the *Averaging Kernel Matrix*. There are a few references that have used averaging kernels, most notable Backus and Gilbert (1970), Conrath(1972). Here we will describe the diagnostic used by Rodgers (1990).

A retrieved temperature profile can be expressed as a function of the true temperature profile, by substituting the forward calculation for the radiances in the inverse calculation. In the minimum variance case, we simply substitute equation 3.12 for the radiances, into equation 3.18 to get:

---

the sensitivity of the retrieved profile to errors in the radiance measurements. For more details see Rodgers (1976).

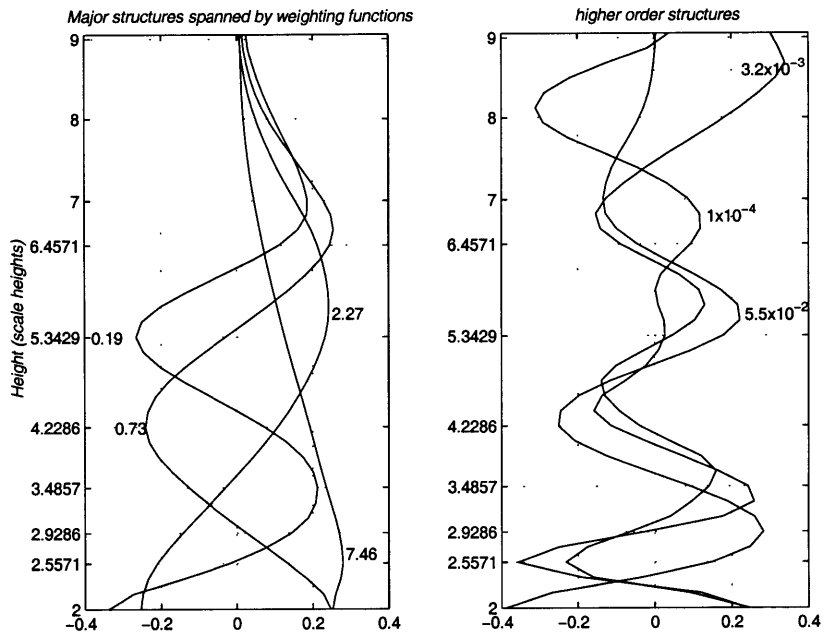

Figure 3.2: The eigenvectors of $\mathbf{KK^T}$, along with the eigenvalues. The structures of the largest four eigenvalues are shown in the left figure and the fifth, six and seventh are shown on the right. See text for details.

$$\hat{\mathbf{b}} = \mathbf{b_o} + \mathbf{S_o}\mathbf{K}^T(\mathbf{K}\mathbf{S_o}\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}\mathbf{K}(\mathbf{b} - \mathbf{b_o}) \equiv \mathbf{b_o} + \mathbf{A_{km}}(\mathbf{b} - \mathbf{b_o}) \tag{3.20}$$

$$\hat{\mathbf{b}} - \mathbf{b_o} = \mathbf{A_{km}}(\mathbf{b} - \mathbf{b_o}) \tag{3.21}$$

$\mathbf{A_{km}}$ is called the *Averaging Kernel Matrix* because the retrieval at a given height is an average of the whole profile weighted by this row. The columns of this matrix (referred to as the *response functions*) are the response of the observing system to a spike of temperature perturbation introduced at a given height[5].

The effect of the constraint on the retrieval solution can be understood more intuitively by rearranging equation 3.21

$$\hat{\mathbf{b}} = \mathbf{A_{km}}\mathbf{b} + (\mathbf{I} - \mathbf{A_{km}})\mathbf{b_o} \tag{3.22}$$

[5]Backus and Gilbert (1970) introduced the Averaging Kernel Matrix and used the averaging kernels to define a resolution of the observing system, in the context of solid earth remote sensing. They also developed a retrieval algorithm based on this matrix that optimizes the resolution we can obtain with respect to the sensitivity of the system to noise. Conrath (1972) applied their method to the temperature sounding problem.

where $\mathbf{I}$ is the identity matrix. If we also decompose $\mathbf{A_{km}}$ into its eigenvectors, we have

$$\mathbf{A_{km}U} = \mathbf{U\Lambda} \tag{3.23}$$

and

$$\mathbf{U^{-1}A_{km}} = \mathbf{\Lambda U^{-1}} \tag{3.24}$$

where $\mathbf{U}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues.

Multiplying both sides of equation 3.22 by $\mathbf{U^{-1}}$ gives:

$$\mathbf{U^{-1}\hat{b}} = \mathbf{\Lambda U^{-1}b} + (\mathbf{I} - \mathbf{\Lambda})\mathbf{U^{-1}b_o} \tag{3.25}$$

If we now use the eigenvectors as a set of basis functions to expand $\mathbf{b}$, we have $\mathbf{b} = \mathbf{Uu}$, $\mathbf{u} = \mathbf{U^{-1}b}$, where $\mathbf{u}$ is a vector of the coefficients of the eigenvectors in the expansion of $\mathbf{b}$. We similarly expand $\mathbf{b_o}$, and get:

$$\hat{u}_i = \lambda_i u_i + (1 - \lambda_i)u_{oi} \tag{3.26}$$

The retrieval solution can be described as a linear combination of the eigenvectors of $\mathbf{A_{km}}$, where the coefficient of each of them is a weighted average of the corresponding coefficients from the inverse solution (u_i) and the constraint (u_{oi}), with a weight of λ_i, the corresponding eigenvalue.

3.4 Results

We apply the retrieval algorithms described above to many temperature fields. The fields we choose vary in the shape of the waves, as well as in the basic state. Some have smaller vertical scales than others. We also use many constraints and error covariance matrices in the MV retrieval. Due to lack of space, we will only present results that are needed to illustrate our conclusions. Unless specifically noted, the results presented are general characteristics of the retrievals in this study.

3.4.1 A single profile

In this section we will discuss the results for a single profile, in order to gain some understanding of the general properties of the retrievals. Figure 3.3 shows a temperature profile (line with solid circles), and various retrievals of it. The continuous lines are different MV retrievals, all using the same constraint profile (thin dotted line)

with a diagonal error covariance matrix but with different values of variance (using a constant value of variance for all levels). The circles show the Chahine retrieval. Also shown are the true minus retrieved profiles for each of the retrievals, to highlight the deviations.

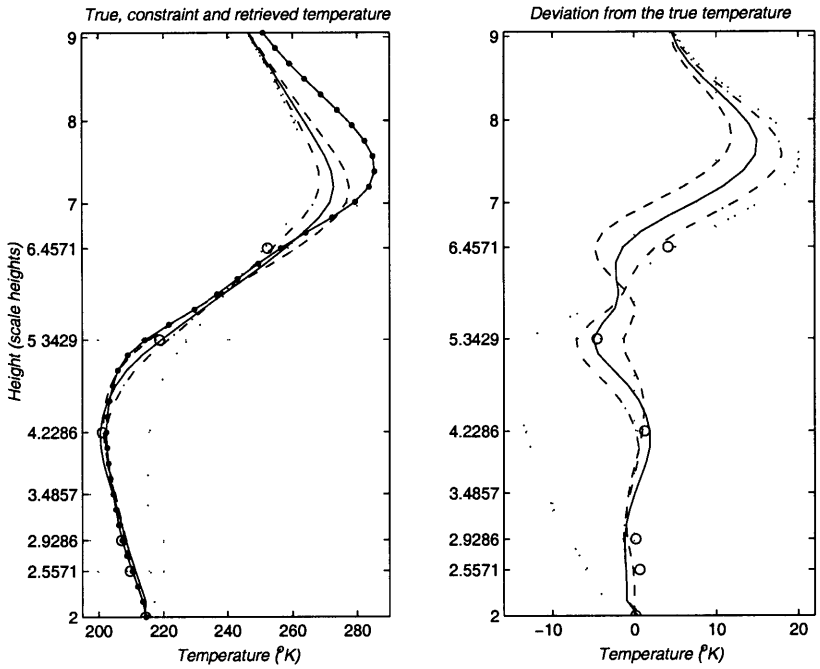

Figure 3.3: Left: A 'true' temperature profile (line with solid circles), and various retrievals of it: the Chahine retrieval (circles), and diagonal MV retrievals using the constraint profile shown (thin dotted line) and a constant variance of 5 °K (dash-dot), 10 °K (solid) and 20 °K (dashed). Right: The true minus retrieved and true minus constraint profiles shown on the left (using same line types).

The retrieval solutions below 4.5 scale heights (30 km) are within 1-2 °Kelvin of the true temperature. The errors grow above that, with largest errors at the top of our model, above the peak of the top weighting functions. Errors at the height of the top Chahine level are typically between 4-12°K (for all methods), depending on the specific profile resolved. It is illuminating to compare the diagonal MV retrievals for various values of variance. Since the constraint errors are uncorrelated in the vertical, the general behavior of the retrieval is to follow the inverse solution (i.e. the true profile) at the lower levels, where the variance in the measurement is smaller than the variance in the constraint, and to follow the constraint at higher levels, where there are no observations. In between, the solution is a combination of the two, where we find that a larger constraint variance allows the retrieval to follow the inverse solution more closely and over a larger region.

The problem with specifying too large a variance is highlighted by comparing the

results of using noisy and exact retrievals. It is important for a retrieval scheme to have low sensitivity to noise in the measurements because in reality we do not know the radiances exactly. Figure 3.4 shows the RMS retrieval errors due to introducing some errors in the radiances. This was calculated by subtracting the exact from noisy retrievals of many different profiles, and calculating the statistics of these deviations for each level. We use white noise errors with a standard deviation about zero of the value of the instrumental uncertainty (see table 3.1). We see that larger values of constraint variance are associated with a larger sensitivity to noise. The noise variance is generally between 1 and 2°K for a constraint variance of 10°K. At 20°K variance, the standard deviation oscillates about a value of 2 °K in most of the domain, and reaches a peak value above 5°K near the bottom (this may be a result of not having the troposphere in our retrieval). For smaller variance, the errors above the weighting functions' span region are very small because the retrieval follows the constraint. The Chahine method has a relatively large sensitivity, equivalent to the sensitivity of the MV retrieval with 20°K variance. Also shown in figure 3.4 are the maximum error values found in the ensemble used for the MV retrievals. The limitations of using a very large variance are obvious, because the errors can reach values larger than 5°K throughout most of the domain. The maximum errors of the Chahine retrievals (not shown) are very large (15°at most levels and 60°K at level 2). When we look at 3 dimensional wave patterns, these large errors usually occur in isolated grid points, while the general noise level corresponds more to the RMS one. These isolated very large errors are most likely due to a bad convergence of the iteration routine. A more likely estimate of an upper bound on the errors is 10°K.

This brings out one of the main issues of inverse solutions, namely, the more you constrain the inverse solution, the less sensitive it is to noise. A solution that has no constraint (for example, what you would get from inverting the matrix K in equation 3.12), will have a sensitivity to noise that will render the solution impractical (the roundoff error will be sufficient to give you a very wrong answer). We therefore have to introduce some constraint. In the MV retrieval, the constraint is very obvious. In the Chahine retrieval it is hidden in the assumptions we make (some form of interpolation) when we integrate the discrete temperature profile in the vertical to calculate radiances in the iteration. The ideal retrieval will have the correct balance between constraint and sensitivity to noise. In the diagonal MV case this seems to be a variance of around 7-10°K.

We can gain more insight as to what the diagonal constraint MV retrieval is doing by looking at the response functions, which show the response of the measuring system to a spike of temperature. Figure 3.5 shows the response to putting a spike

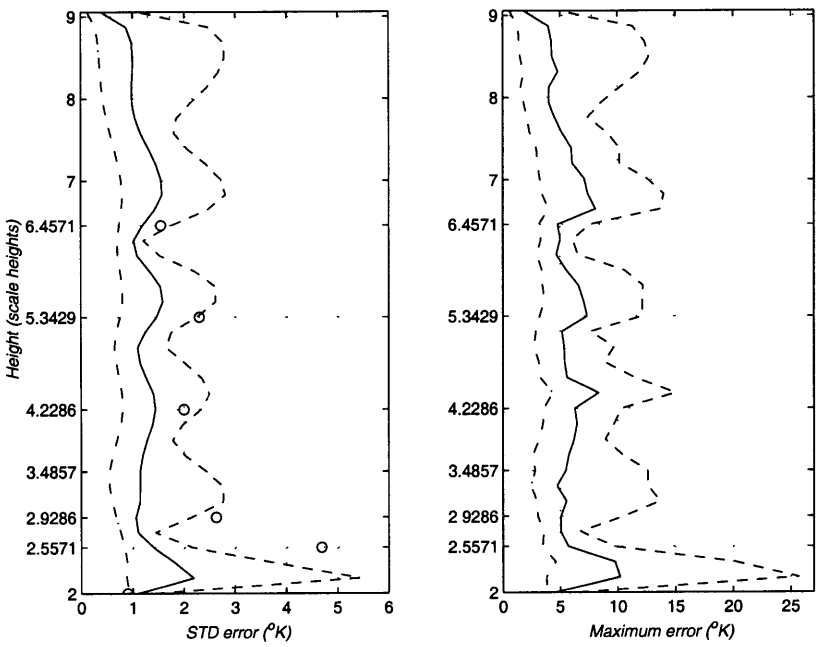

Figure 3.4: The STD (left) and maximum (right) errors in the temperature retrievals (°K), resulting from putting an error in the radiances as described in table 3.1, for various retrievals: Diagonal MV retrievals with a constant variance of 5°K (dot-dash), 10°K (solid) and 20°K (dashed) lines, and (on the left) the Chahine retrieval (circles

.

at a few different heights, for the different values of variance. The response generally peaks at or near the height of the perturbation it is responding to. An exception is the response to perturbations at levels above the peak of the top weighting functions. Those result in a maximum response that is just above the height of the top weighting function, and the response decreases the higher the perturbation. The response has some vertical spread, which can be taken as a measure of the resolution of the system. Generally, the amplitude of the response is confined to a region around the spike, and there is hardly any "remote" response. This will not be true when we introduce vertical correlations into the constraint. Comparing the response functions for the different values of variance, we see that a larger variance results in responses with a narrower and larger magnitude. However, when the variance is increased enough, we start seeing larger overshoot, which will result in a larger sensitivity to noise, as expected.

Figure 3.6 shows the 10 largest eigenvalues of $\mathbf{A_{km}}$ for various values of variance. The first thing to note is that no more than six or seven eigenvalues are significantly larger than zero, which makes sense because we only have seven observations. The largest eigenvalues approach 1.0. Figure 3.7 shows the first six eigenvectors (ordered

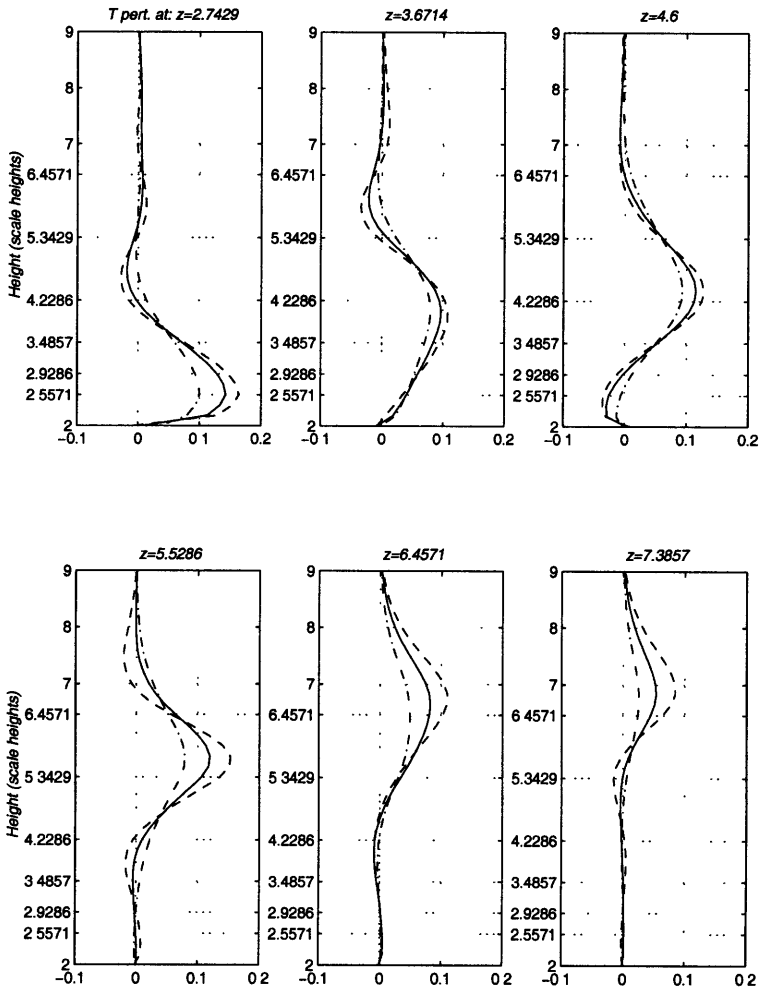

Figure 3.5: The response functions to a spike perturbation of temperature at various heights, for a diagonal MV retrieval using a constant variance of 5°K (dash-dot), 10°K (solid) and 20°K (dashed). Starting from the top, left to right, the response to perturbations at 19.2, 25.7, 32.2, 38.7, 45.2, 51.7 km, (he height in scale heights is given at the top of each sub-plot)

by ascending eigenvalue) for the various values of constraint variance. We see that the shapes of the vectors do not change too much with variance, and that the dominant vertical scale decreases with eigenvalue. According to equation 3.26, the first few eigenvectors (which have the largest vertical structures) with eigenvalues close to 1.0, will be determined mostly by the inverse solution. The structures in between (eigenvalues of 0.33-0.67) will be a mixture of the inverse solution and the constraint. Structures with very small eigenvalues (and small vertical scales) will be affected by the constraint only. As the variance of the constraint is increased, the eigenvalues become larger, and more eigenvectors are affected by the inverse solution. For example, the retrieval using a variance of 2.5°K will draw only the first one or two eigenvectors

from the inverse solution. The resultant retrieval will therefore be very different from the true temperature profile, because it can follow it only with very coarse resolution. We also note that the first two eigenvectors are heavily weighted towards the lower levels, hence the errors are smallest at the lowest levels and increase with height. In contrast, solutions with very large variance (e.g. 40°K) will have 4 or even 5 eigenvalues that are affected mostly by the retrieval solution. That is not very good as well, since the observations do not resolve that many, as demonstrated earlier by looking at the eigenvalues of the weighting functions (section 3.3.4). The resulting retrieval will be sensitive to noise because the projection of the inverse solution on to the higher structures is sensitive. Accordingly, we expect to see these structures emerge in the difference between the true and retrieved fields. Note that the fifth eigenvector is the most stable structure in the sense that it is the least affected by the variance of the solution. This suggests it is some measure of the resolution of the observing system, hence its structure is determined by it, and not by the constraint.

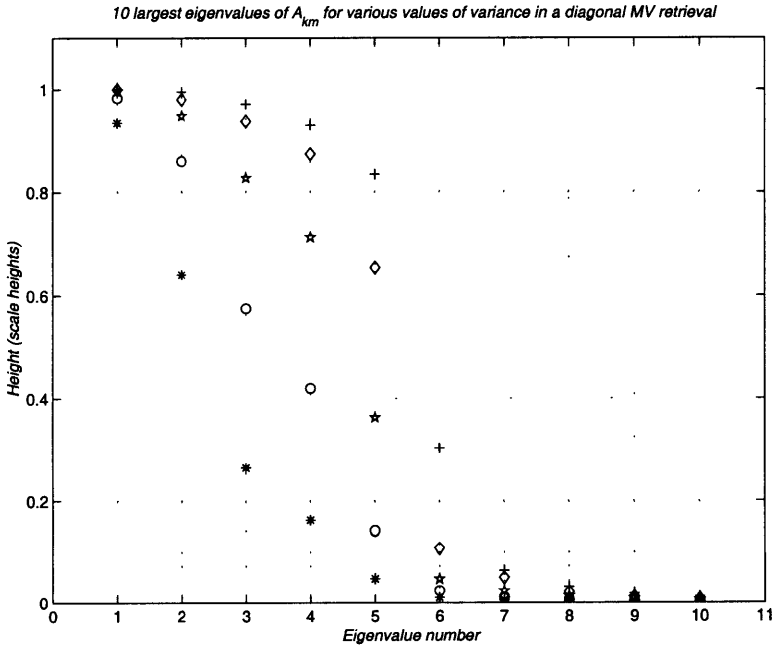

Figure 3.6: The 10 largest eigenvalues of the Averaging Kernel Matrix for a diagonal MV retrieval, with a constant variance of 2.5°K (astrisks), 5°K (circles), 10°K (stars),20°K (diamonds) and 40°K (plusses).

Ideally we want the retrieval to have large eigenvalues for the first 3 structures, because they are well resolved by observations, and to have small eigenvalues for all structures above and including the 5th, because those are not resolvable by the observing system. This leaves the fourth eigenvalue to be a combination of the inverse and constraint profiles. Looking at figure 3.6, this corresponds to a variance of about

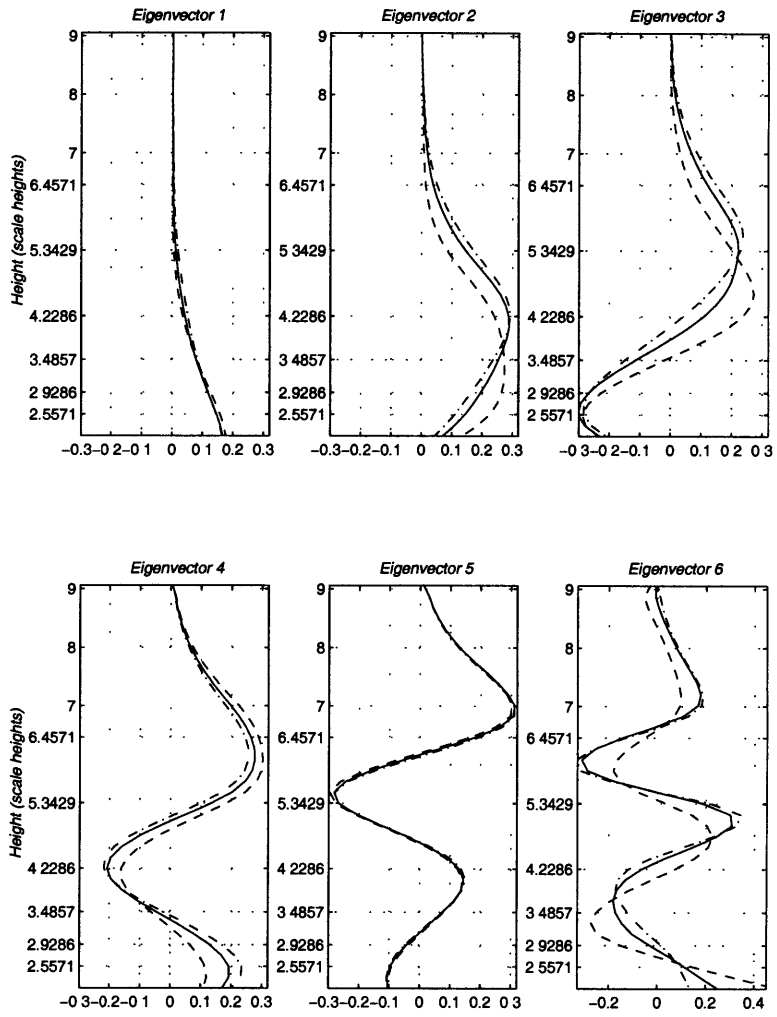

Figure 3.7: The first six eigenvectors of the Averaging Kernel Matrix for a diagonal MV retrieval using a constant variance of 5°K (dash-dot), 10°K (solid) and 20°K (dashed).

$7°K$.

To summarize, a retrieval solution is capable of retrieving a temperature profile to within 2-3°K, below 5 mb. Above that height, the quality of the response decreases with height, and it depends to a large extent on the retrieval method. There is an inherent resolution of the system, which shows up as the characteristic scale of errors when the solution is weakly constrained. This error pattern vertical scale is around 1.5 scale heights (10km).

After getting a good sense of the behavior of the retrieval in one dimension, it is essential to look at the three dimensional fields because it is unclear what horizontal structures the errors in the retrievals will assume, and how they will affect the vertical-horizontal cross sections of the waves. This is the subject of the next few sections.

3.4.2 Three dimensional fields- Chahine's method

Figure 3.8 shows the basic state temperature along with wave number one amplitude and phase and a longitude-height plot of the wave at latitude $y = 2.83$ radii of deformation, (see appendix A.1 for details on the calculation of the true field). Figure 3.9 shows the Chahine retrieval of this wave, along with the true minus retrieved fields.

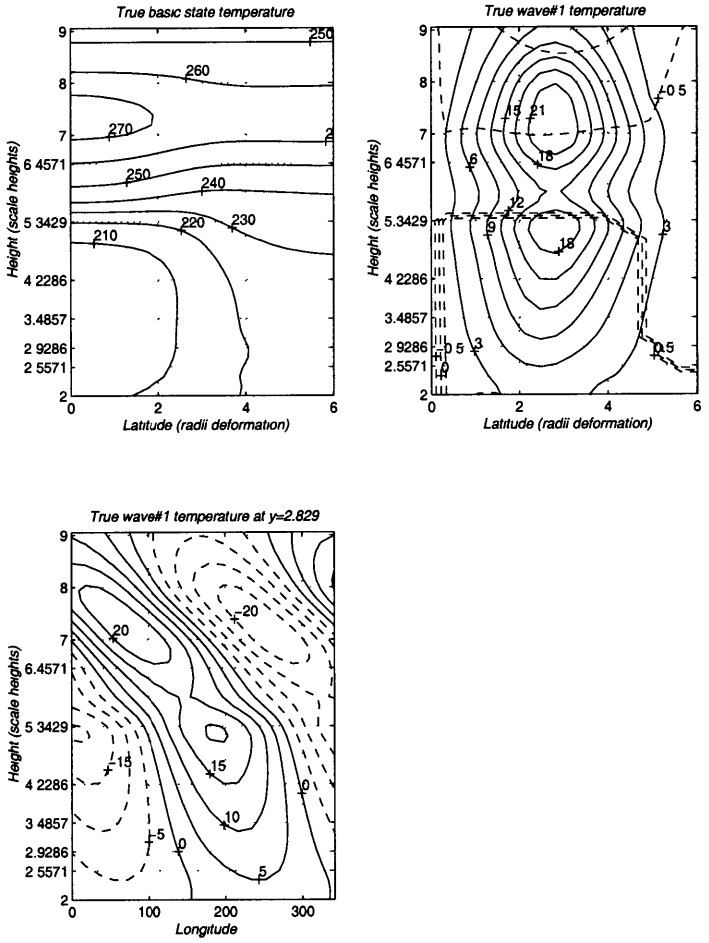

Figure 3.8: The 'true' temperature field, generated by the model and used for the retrievals shown later. Top left: The basic state temperature. Top right: Wave number 1 temperature amplitude (solid) and phase (dashed). Bottom: A longitude-height section of the temperature wave number 1, at a latitude of y=2.83. In all relevant figures, contour intervals for the phase is 0.5π, latitude is in units of radii of deformation, height in scale heights, and negative values are dashed.

In general, the properties of the one dimensional retrieval, described in the previous section, hold here. The interesting result is that the horizontal pattern of the wave is captured well by the retrieval, errors being mostly in the amplitude of the wave, and not in the phase structures. The lower maximum of wave amplitude (with a magnitude of 18°K) which is at roughly 5 scale heights is captured very well, with

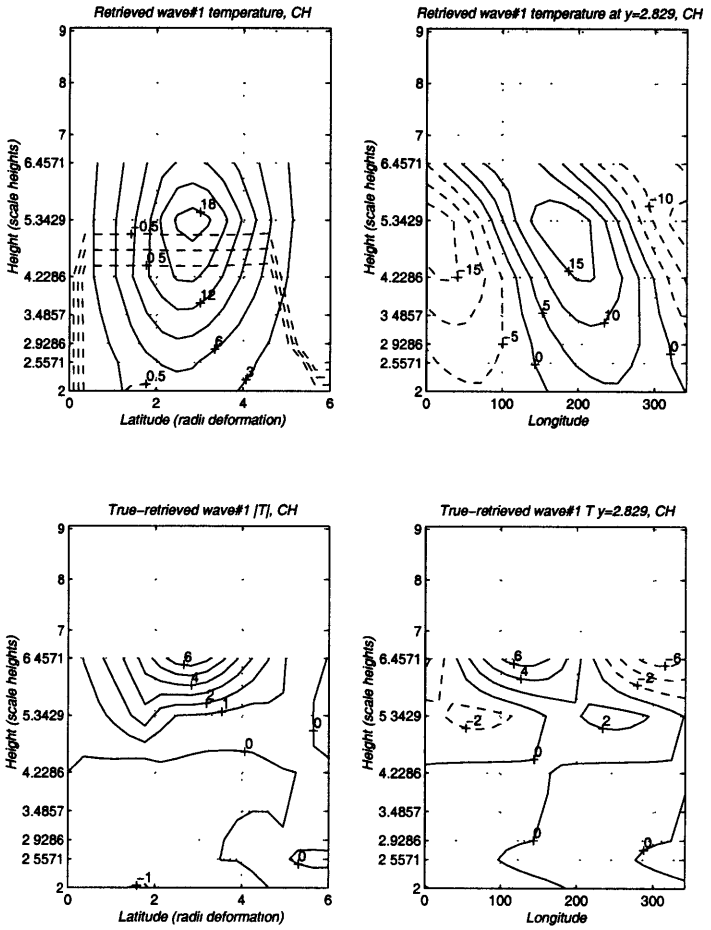

Figure 3.9: The Chahine retrieval of the wave temperature fields shown in figure 3.8 (top) and the true minus retrieved fields (bottom). Left: Wave number 1 temperature amplitude (solid) and phase (dashed, not shown in the true minus retrieved field). Right: A longitude-height section of the temperature wave number 1, at a latitude of y=2.83.

a 1°K error. However, above that, the errors grow, reaching 6°K at the top Chahine level. The excellent retrieval of the phase structure makes sense because the radiance fields reflect the horizontal pattern of the wave and there is nothing in the retrieval algorithm that will change the horizontal pattern of the radiances.

The effect of adding random noise to the radiances is to add a white noise field to the temperature of roughly $2-3°K$. The resultant retrieved wave field looks a bit noisy, but the overall large scale pattern is still evident. When a Fourier decomposition is made and only wave one retained, there is a very small difference between the exact and noisy fields, meaning that the noise does not project onto wave 1. A similar noisy appearance is in fact a feature of observed fields.

3.4.3 Three dimensional fields- Minimum variance method

Diagonal constraint

The minimum variance method is used to retrieve temperatures, with various constraints and error covariance matrices. First we show results from runs using a diagonal error covariance matrix. We showed earlier that having a diagonal constraint means that it has no vertical correlations, hence the vertical correlations in the retrieval solution are due to the observing system. We also showed that for a single profile, there is a tradeoff between resolution and sensitivity to noise.

What we look at in this section is how this shows up in a three dimensional field. Figure 3.10 shows a diagonal MV retrieval, with a constant variance of 10°, of the wave shown in figure 3.8. The general features shown in the single profile case hold here as well. The errors are largest at the top of the domain, and are less than $2°K$ in the lower stratosphere. As in the Chahine retrieval, the main errors below the peak of the highest weighting function are in the amplitude of the wave, while the phase structure is captured well. Above this level (where we have no Chahine retrieval) there are errors in both the amplitude and the phase of the wave that increase with height. To get an idea of magnitudes, the phase error at 8 scale heights is about 60°, and it never exceeds 90° in our runs. The amplitude errors can be as large as the waves are (when the retrieval follows the constraint, the retrieved wave amplitude is zero).

The phase shift needs some explaining, because the constraint, which contains no waves, cannot be the source of it. The constraint can affect the amplitude of the waves, but not the phase. Hence, the explanation lies in the inverse solution. It is interesting to note that the phase shift is always such as to decrease the vertical tilt of the wave. This suggests the inverse solution above the peak of the top weighting function (6.45 scale heights) is some vertical average of the true profile, between the peak of the top weighting function and the point of retrieval, because the observing system sees less and less of the top levels as we go up. This is consistent with the behavior of the response functions, which show that the response to a temperature perturbation at any height above 6.45 scale heights peaks slightly above 6.45 scale heights, and the higher the perturbation, the weaker the response (figure 3.5).

Non-diagonal constraint

In this section we check the effect of having vertical correlations in the constraint. It is important to check this because the operational retrievals calculate the constraint and the error covariance matrix from a climatology which may contain waves and other

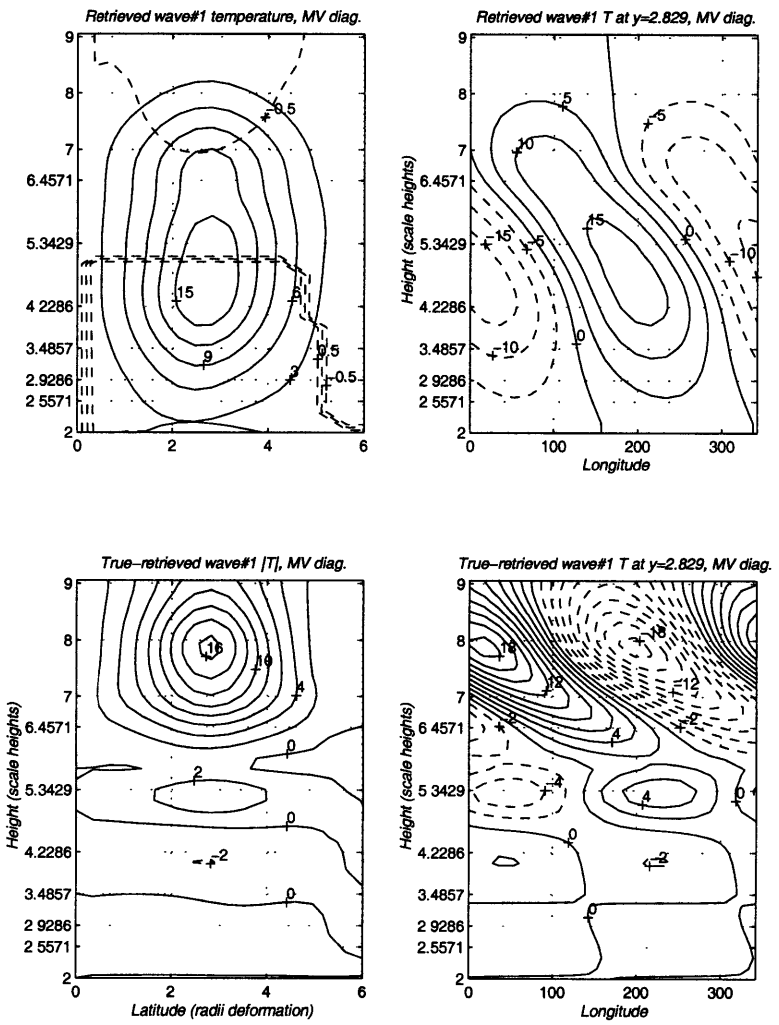

Figure 3.10: As in figure 3.9, only for the diagonal MV retrieval with a constant variance of 10°K.

physical processes that have correlations in the vertical (see appendix A.3). Having a vertical correlation in the constraint is equivalent to introducing new information to our retrieval, which may not necessarily be a good thing, because the new information may be wrong. The constraint we use in most runs is specified from a climatology we create for our model, by running it with a time dependent forcing. We describe the constraints we use and how we calculate them in appendix A.2. Most of our retrievals contain a constraint calculated from a run where we superpose a transient and a stationary wave number 1. The structure of waves in this 'climatology run' is different from the structure of stationary waves or transient waves.

Figure 3.11 shows some of the response functions (see section 3.3.4) for the non-diagonal constraint case, and for a case where only the diagonal elements of the error

covariance matrix were retained, while the rest are set to zero[6]. We see that the main effect of the vertical correlations is to spread the response to non-adjacent layers. The response functions have two or three peaks, instead of a single major one, meaning that a spike of temperature perturbation will introduce responses at remote layers. In this case this is mostly the result of having waves in the climatology from which the constraint was calculated. The largest effect is on the retrieval at high levels, in some cases, leading to a larger response at high levels than at the level at which the perturbation is introduced. This is a result of the increase of wave amplitude with height in the model run used for the constraint calculation. Note however, that the remote responses to perturbations at various heights may cancel each other.

Figure 3.12 shows the eigenvectors of the 6 largest eigenvalues (listed in the figure) of the averaging kernel matrix for the diagonal and non-diagonal constraints that are shown above. The first five eigenvalues of the diagonal case are larger (or equal) to the non-diagonal ones, meaning the constraint has more influence on the retrieval solution in the non-diagonal case (equation 3.26). The eigenvectors of the two cases have similar general features (similar vertical scales and number of peaks) but the non-diagonal ones peak higher up and the highest peaks are larger, relative to the bottom ones.

Figure 3.13 shows a minimum variance retrieval of the fields shown in figure 3.8, using the non-diagonal constraint used in the above analysis. Overall the retrieval does a good job, similar to the diagonal minimum variance and the Chahine retrievals. We see that in this case the retrieval does better at higher altitudes than the diagonal retrieval does (figure 3.10). The improvement is most striking in the phase structure of the wave, which is now almost the same as the true one. The error in amplitude is also about half of that in the diagonal case.

The vertical correlations in the constraint also have an effect of reducing the sensitivity of the solution to noise. Figure 3.14 shows the RMS error due to putting noise in the radiances (the same noise used in figure 3.4) for a non-diagonal MV retrieval, using a few constraints. Shown are the non-diagonal constraint used above (referred to as *standard*), its corresponding diagonal constraint, and two additional constraints that are calculated from the same model run as the standard one, once with doubling and once with halving the amplitude of the waves. This has an effect of increasing or decreasing the variance, respectively. We see, as in the diagonal case,

[6]The difference between the current diagonal covariance matrix and the ones tested in the previous section is that the variance varies with height, whereas the constraint of the previous section assumed a constant variance. The current variance has larger values in the middle and upper stratosphere (around 13°K) and smaller values at the lowest and highest levels (around 3°K).

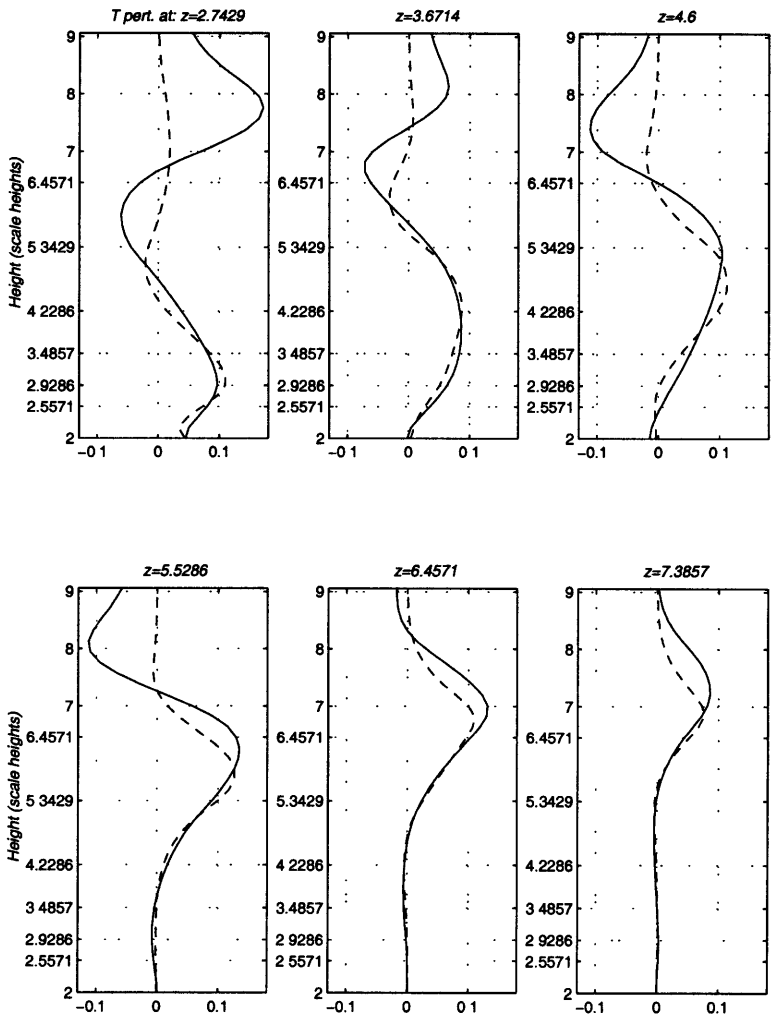

Figure 3.11: The response functions to a spike perturbation of temperature at various heights, as shown in figure 3.5 for a non-diagonal (solid) MV retrieval and the corresponding diagonal (dashed) retrieval (see text for details).

that the more constrained retrievals (those with smaller waves in the constraint) have a smaller sensitivity to noise. We also see, by comparing the standard constraint to its diagonal version, and to the diagonal retrievals of figure 3.4, that the non-diagonal terms in the error covariance matrix act to reduce the noise sensitivity below the peak for the top weighting function, and act to increase it above. In the lower part, that observations see, the RMS error is around 0.5-1.5°K (depending on the constraint used), and the maximum error is less than 4°K. At the upper levels, the RMS error is 2-4°K and the maximum error is 10-20°, depending on the constraint, while the diagonal STD is up to 2°K less and the maximum error up to 10°K less than the corresponding non-diagonal case. This large sensitivity at higher levels is due to the projection of the noise at lower levels on higher levels, through the vertical

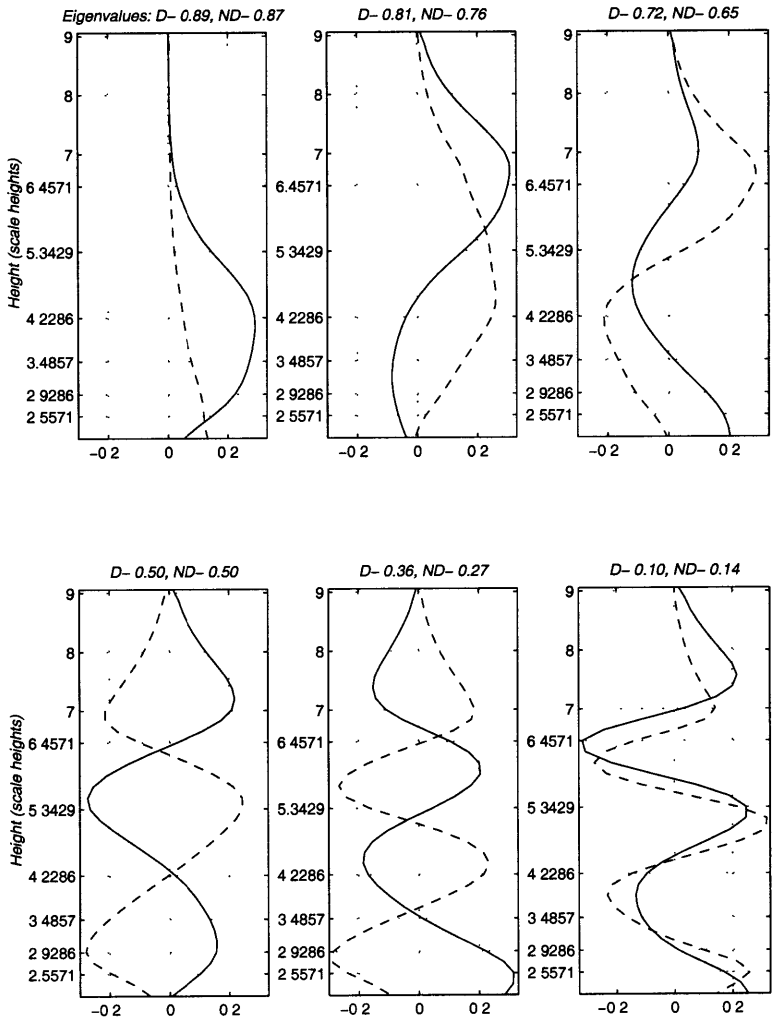

Figure 3.12: The first six eigenvectors of the Averaging Kernel Matrix for a non-diagonal (solid) MV retrieval and the corresponding diagonal (dashed) retrieval. The corresponding eigenvalues for the two cases are given in the title of each subplot. The diagonal case is denoted by D and the non-diagonal by ND.

correlations.

The reasoning behind having a non-diagonal constraint is clear. First, the noise sensitivity is greatly reduced in most of the domain as a result of having vertical correlations in the constraint. Second, if we have a general idea of the structure of waves in the stratosphere, obtained by other means of observation (e.g. radiosondes and rocketsondes), we can use this as an extrapolation tool to supply the observations. If the observing system detects a wave in the middle of the stratosphere, the constraint will supplement its structure at the top of the stratosphere. Indeed our results show that when the true wave field does include waves, the non-diagonal retrieval is capable of doing an excellent job, and if the climatology used to calculate it includes suffi-

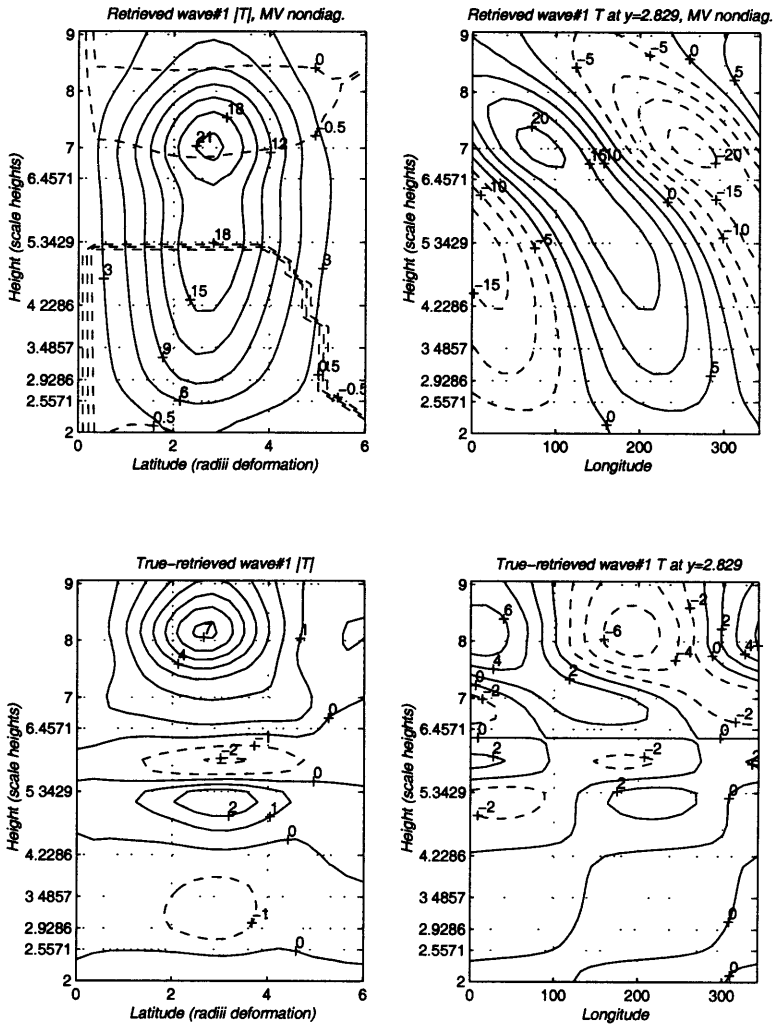

Figure 3.13: As in figure 3.9, only for the non-diagonal MV retrieval.

ciently large waves, it usually is an improvement over a diagonal minimum variance or a Chahine retrieval. However, there is no way to insure this improvement, because the retrieval system does not have a way of knowing when the vertical correlations actually exist in reality. Also, the noise sensitivity tests suggest that small errors in the retrievals can lead to large variations in the retrieved wave structures at the top of the domain. There is no clear way to estimate what part of the retrieved wave is not due to the observations but is artificially provided through the constraint, unless we know what the true wave field is.

We will illustrate this point with an example. We run our steady state model with a basic state that is characteristic of early summer, where the winds become easterly at around 30 km and with a wave number 1 forcing at the bottom. The resultant wave geopotential height decays rapidly above the zero wind line, and the corresponding

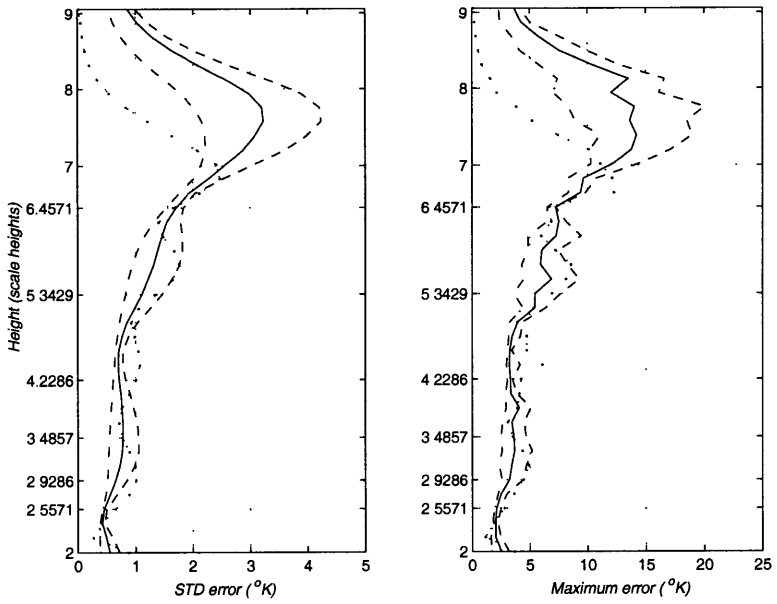

Figure 3.14: As in figure 3.4, only for a non-diagonal MV retrieval, using different values of wave amplitude in the constraint. Shown are the *standard* constraint, which was used in figure 3.11 (solid) and the diagonal version of it (dotted). Also shown are constraints that are calculated as the standard one, only using half (dash-dot) and twice (dashed) the amplitude of waves.

temperature wave field is cut off sharply. The temperature field also has some very small scale structure near the critical layer. Figure 3.15 shows the wave number 1 temperature amplitude and a longitude height cross section, at latitude y=2.83, along with the non-diagonal MV retrieval of these fields. We see, first of all, that the vertical correlations of the retrieval introduce waves into the upper part of the domain. The retrieved wave has some characteristic vertical structure which corresponds quite well to the fourth eigenvalue of the averaging kernel matrix (see figure 3.12, solid line), which has an eigenvalue of 0.5, meaning the constraint contributes as much as the inverse solution. We also see that not only is a wave introduced to upper levels, but the retrieval has a hard time with the true wave at lower levels because the vertical structure is too small for the observing system to see. As a result, the retrieved waves looks nothing like the true wave.

The operational sounding product uses a minimum variance retrieval to obtain temperature on 32 TOVS levels (see appendix A.3). This relatively high resolution profile is averaged over specific layers, and these averages are then used to calculate temperature on 18 levels (9 of which are in the stratosphere), assuming linear interpolation. The reasoning is that the satellite sees layer averages rather than a continuous temperature profile, and the overall effect is to smooth small scale features. We ap-

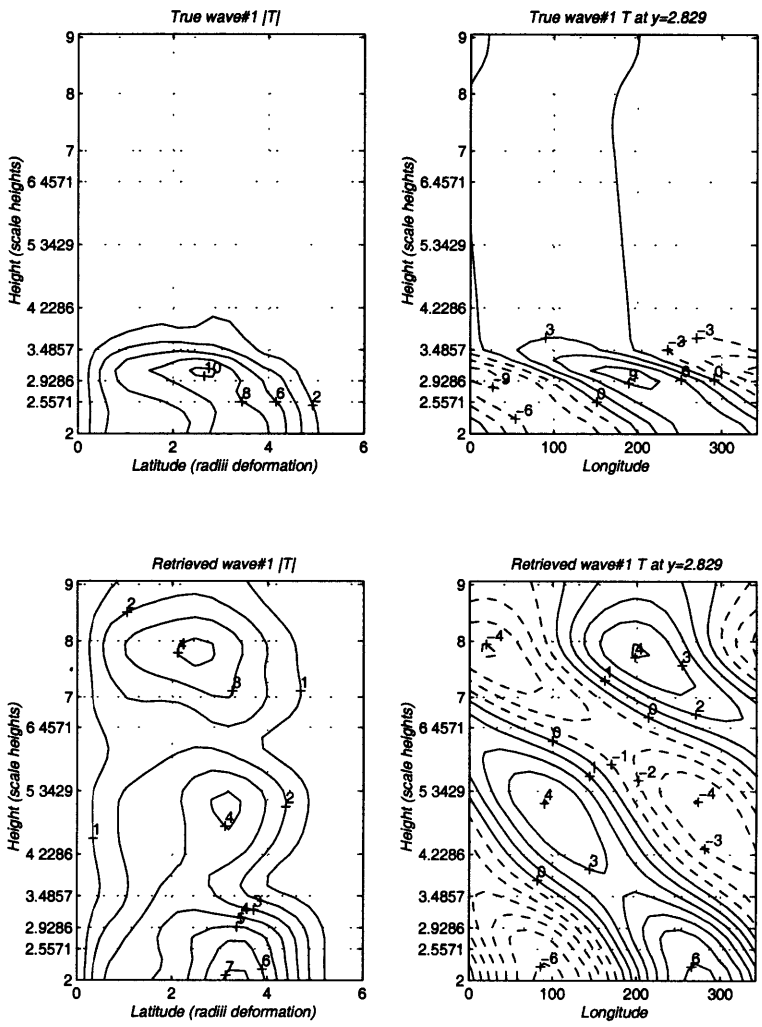

Figure 3.15: The 'true' temperature field generated by the model (top) and its non-diagonal MV retrieval (bottom). Left: Wave number 1 temperature amplitude. Right: A longitude-height section of the temperature wave number 1, at a latitude of y=2.83. The non-diagonal constraint is exactly the same one used in figure 3.13.

plied such an averaging to the retrievals, to see if it gets rid of the relatively small scale artificial waves introduced by the vertical correlations of the constraint. The resulting fields are shown in figure 3.16, where the grid is now on the official levels at which observations are reported. We see that the artificial wave is not averaged out, it just looks a bit smoother, with the top part appearing more connected to the bottom part of it.

In order to test how well the sharp temperature structure of figure 3.15 can be resolved, without having vertical correlations, we repeat the retrieval using a diagonal MV method with a constant variance of 10°K (figure 3.16), and a Chahine method (not shown). Both methods give roughly the same results. We basically get a very

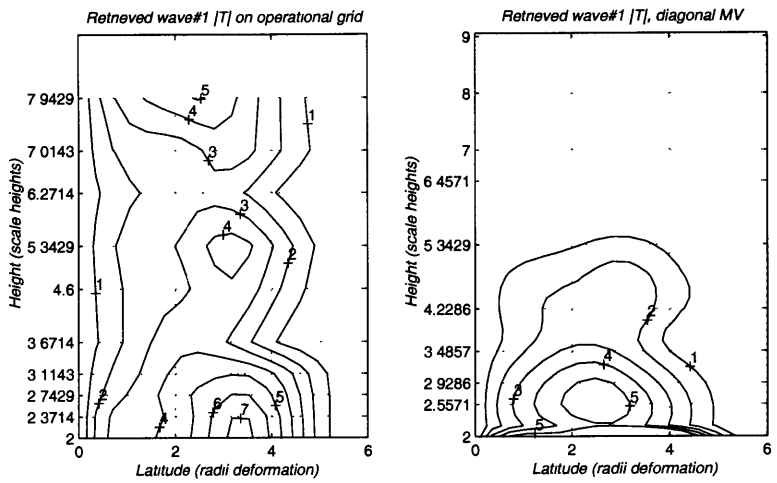

Figure 3.16: The retrieval of the wave 1 temperature amplitude shown in the top left plot of figure 3.15. Left: after applying a vertical averaging over specified layers to get geopotential heights, then applying a linear interpolation to get temperatures. Right: A diagonal MV retrieval with a constant variance of 10°K. Note that the vertical grid of the left hand figure is different from all other figures. The grid corresponds to levels that are reported operationally.

smoothed version of the true wave, without the projection of vertical scales onto the top part of the domain. In this case therefore, the CH and diagonal MV methods do much better than the non-diagonal MV method, not just at the top levels above the weighting function peaks but everywhere.

The main question is do we expect such high structured perturbations that are confined to the lower stratosphere to occur in nature? If yes, then we may have a problem distinguishing between them and a vertically propagating wave. We will address this question in section 3.5. Before we do that however, we will look at the effect of having a spatially varying constraint profile.

Horizontally varying constraint

Operational retrievals use a horizontally varying constraint field (see appendix A.2 for more details). This means that waves are introduced into the constraint field, and not only by the observations. We saw, in a single profile case that in the absence of vertical correlations, the retrieved fields tend towards the constraint at upper levels (section 3.4.1, figure 3.3). This would suggest that having a wave in the constraint profile can introduce a wave into the solution, even if the true field does not contain one, and even in the absence of vertical correlations in the constraint. Figure 3.17 shows the retrieval of the same summer wave field shown in figure 3.15, using a diagonal MV retrieval with a constant variance of 10°K, and a non-zonal constraint

field, which is a wave number 1. Also shown is the constraint wave field. We see that the retrieval at lower levels looks like the diagonal MV retrieval with one constraint profile for all grid points, and at higher levels, above the peak of the top weighting functions, the retrieval looks like the constraint. It is important to note that this kind of retrieval will introduce a spurious wave only at regions where the observations are poor, which is not the case with vertical correlations. In the operational retrievals, a combination of both the vertical correlations and a wave in the constraint profile will contribute to the retrieved wave fields at upper levels.

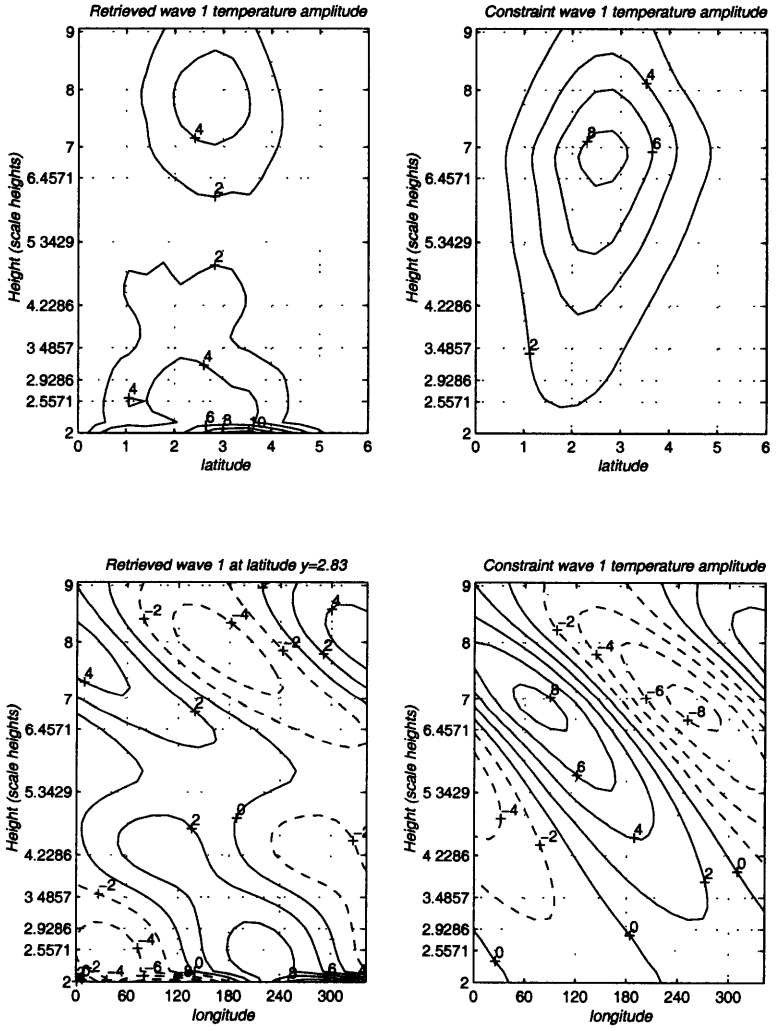

Figure 3.17: A diagonal MV retrieval of the longitude-height section of wave 1 temperature shown in the top right plot of figure 3.15 (left) using a spatially varying constraint that has a wave 1 structure (shown on the right). The diagonal MV retrieval is done with a constant variance of 10°K.

3.5 Summary and implications to observations

We have seen in past sections that the retrievals are capable of doing a very good job of retrieving waves, with a few limitations. There is a limit to the resolution which the retrievals can get. Any features with higher resolution are spurious and may be due to noise or to having small scale structure in the constraint. Above the peak of the top weighting function (1.5 mb), there are not enough observations to resolve the waves or even the zonal mean temperature field. Errors start being large even from the peak of the second highest weighting function, at 5 mb. The operational retrievals use additional information to fill in those areas. When we come to look at the observations, we know that any small scale features are questionable, and any information above 1.5 mb is derived purely from the climatology which forms the basis of the constraint. We can tell if the the observed fields above 1.5 mb are a reasonable continuation of the rest of the wave field, but we can not asses how real they are. Indeed, data on the top analysis level of 0.4 mb often looks obviously wrong. However, some situations may be more complicated. We have seen that in certain cases, the retrievals will produce waves where there are none. Moreover, it is not obvious, looking at such waves, that they are spurious. One way of telling is by looking at the radiances. In the minimum variance method, the radiances of the retrieved profile do not necessarily equal the true radiances, especially at the highest level, because above the peak of the top weighting function the retrieval follows the constraint and not the true profile. In all the cases of a spurious wave retrieval that we checked, the radiances of the highest channel (SSU 3) calculated from the true and retrieved profiles both have very small amplitude waves in them (the retrieved spurious wave seems to arrange itself so as to have very small radiances in the top channel, for example by having a node right at that height) but the wave patterns are 180°out of phase with each other. This can also be seen when comparing the pattern of the retrieved temperature at the peak of the top weighting function with the pattern in the radiances. The radiances of the true and retrieved fields at lower channels are not so distinguishable because they are in phase and have a similar horizontal pattern, only the retrieved radiances have a slightly larger wave amplitude than the true ones. In cases where the true waves span the entire depth of the stratosphere, and the minimum variance method supplies information mostly above the top weighting function, the radiances calculated from the retrieved waves are similar to the true ones for all channels. Checking the radiances of the highest weighting function is therefore one way to tell when the retrieval is likely to be off.

Another approach is to decide, on physical grounds, which is more likely- for the

observed wave to be close to the truth or for it to be spurious. The question then becomes one of listing the scenarios where we have a wave only in lower levels, that is cut off abruptly above some height. There are two main possibilities we expect from theory. The first is waves in summer, where the basic state is easterly and will not allow wave propagation. The second is smaller scale waves, which can't propagate vertically, both in winter and in summer. In both of the above cases we expect to see large signals at and below the tropopause and maybe in the lower stratosphere. Since the geopotential height decays quite abruptly above some height, the temperature field may have small scale structures in the vertical. As shown in section 3.4.3, our model runs show this to be true especially at critical surfaces, which are always present in summer. Our model runs also suggest that waves in winter span the entire depth of the stratosphere, and the existence of small scales depends strongly on the basic state winds having small scale structures.

In this context, it is important to know what kind of constraint is used operationally. The details of this are discussed in the appendix, but it suffices to say here that in the stratosphere, especially in the southern hemisphere, the constraint is taken from a rocketsonde data set. A constraint profile is specified by extrapolating radiosonde profiles upwards, using a covariance matrix calculated from the rocketsonde data set. The data is divided into high, middle and low latitudes, and into seasons. The error covariance matrix of the constraint is calculated directly from a data set of upwards extrapolated radiosonde profiles, meaning the vertical correlations of the rocketsonde data set are dominant at upper levels. Closer to winter, we expect to have more vertical correlations, because there is more wave activity. In summer we expect to have fewer correlations. Since in summer we do not expect to see waves very high up, we checked one summer of southern hemisphere observations (November-February 1996) to see if there are any.

On January 28, 1996, we see perturbations in the upper stratosphere that can best be explained as being a spurious retrieval. The zonal mean winds are easterly above 30 mb. Figure 3.18 shows the temperature perturbation at several levels. We see a wave packet in the troposphere, extending from South America to Australia, that is also evident in the stratosphere. At 5 mb we do not see this pattern but we do see it clearly at 30mb and partly at 10 mb (not shown). At 0.4 we see the full pattern again (it is also evident at 1 and 2 mb, not shown). This is most likely spurious because the stratospheric pattern clearly follows the tropospheric one, and we do not expect a perturbation of such a small wave number to propagate vertically through easterlies. Also, the oscillation of amplitude in the vertical is also a feature of the spurious waves of figure 3.15. The magnitude of the wave at 0.4 mb is 2°K. This is

the clearest case of projection of tropospheric features found in the summer of 1996. This gives us an estimate of the magnitude of this effect. We do however expect it to be larger in winter, because the constraint is calculated from a set of collocated radiances and radiosonde profiles measured in the two weeks prior to the retrieval. Also, the rocketsonde data set used for the extrapolation to the upper stratosphere has more waves built into the vertical correlations in winter than in summer.

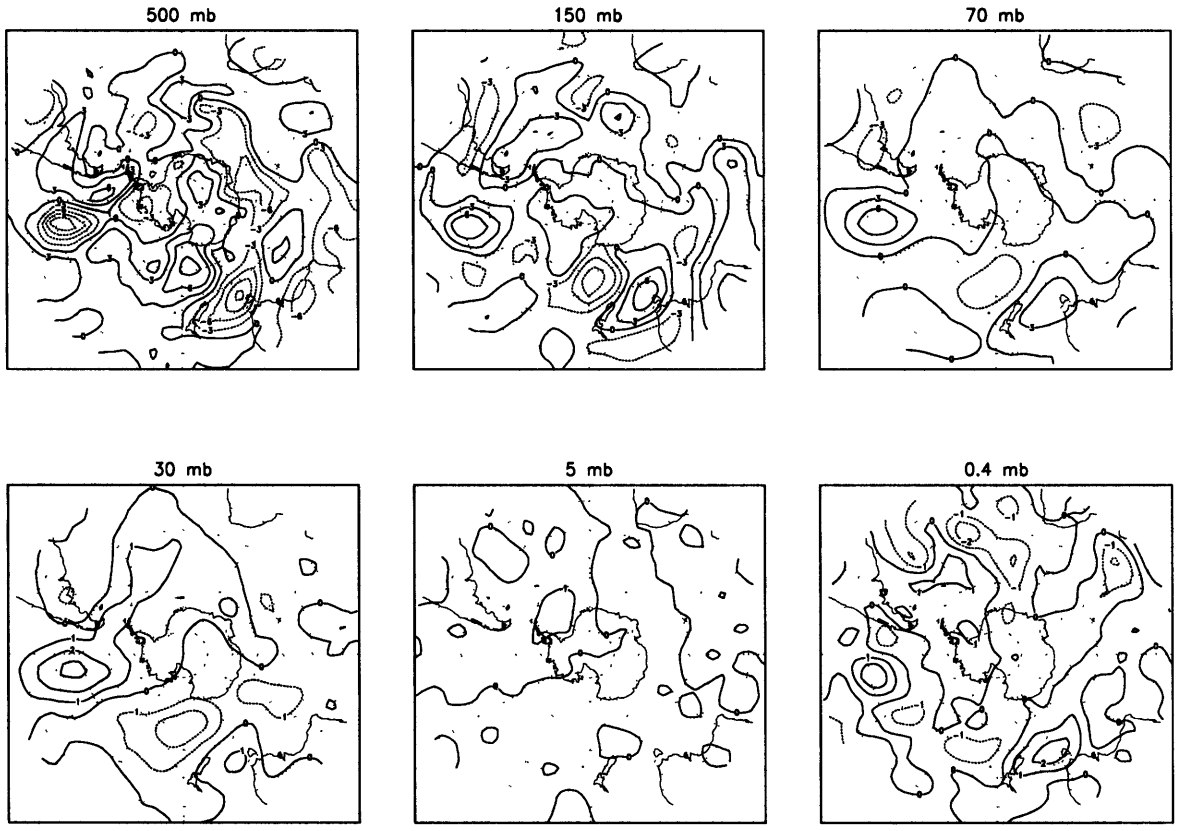

Figure 3.18: Observed temperature perturbation (temperature minus zonal mean) on January 28th, 1996, in the southern hemisphere, at different levels. From top to bottom, left to right: 500mb, 150mb, 70mb, 30mb, 5 mb, 0.4 mb. Contour intervals are 3°in the top figures and 1°in the bottom ones.

A more ambiguous case is the diminishing of wave activity at the end of the winter that is associated with the breakup of the winter polar vortex. In the southern hemisphere summer of 1996 the zonal mean winds gradually shift from a state of easterlies above 2-5 mb poleward of 40°S in the beginning of November, to having a westerly jet in mid-December that is centered around 50°S, and reaches a maximum height of 30 mb. The easterly jet also grows stronger in mid-winter, reaching a maximum speed of -60m/sec in January. The easterly jet starts diminishing in early February. The wave activity follows this cycle. In the beginning of November we

see considerable wave activity throughout the stratosphere. The temperature waves have a double peaked amplitude, with the lower, slightly larger peak at around 70 mb and 70°S and the higher peak at 2-5 mb, 65°S. This activity reduces gradually until in the last third of December, there are no waves in the stratosphere (apart for some exceptions like the case of January 28th). What is interesting is the way the wave activity decreases. The higher peak of the waves decreases and moves down to 5-10 mb, and disappears only around December 20. This is a bit surprising since by December 10th the zero wind line is already at 30 mb. In fact, On December 9-20 the zero wind line often seems to lie just between the two peaks of wave amplitude. The higher peak extends much more in the vertical beyond the critical level and is much broader than what our model runs seem to suggest would be the case. The upper peak magnitude is around 3°K. An example is seen in figure 3.19. The largest perturbations in the stratosphere lie above perturbations of opposite sign, that are just below the critical layer. The fact that we have a critical layer and a wave on an easterly basic state suggests that we may be seeing a case of spurious waves. Another possibility is a case where the true fields are very sharp and we are seeing a smoothed out version of them. However, things are not as clear as the January 28th case, because the phenomena we are seeing has a more coherent time evolution, and the theory is more complicated. The breakup of the polar vortex and the onset of easterlies is a highly transient phenomena. It is not clear how waves would react to the basic state changing so rapidly. Also, in a transient situation, some component of the wave field may be excited westerly waves, which may be able to propagate upwards in the easterlies. Further model studies and observational studies need to be conducted in order to understand this more. In any case, it is important to keep in mind that the breakup of the polar vortex is a time where the wave fields may have very sharp structures and may diminish very rapidly above some height. These conditions are when the retrievals are likely to have the largest errors. Moreover, since the operational radiosonde data set used to calculate the constraint consists of measurement from a period of about two weeks prior to the observation day, it is likely that in this period, the constraint contains significant vertical correlations.

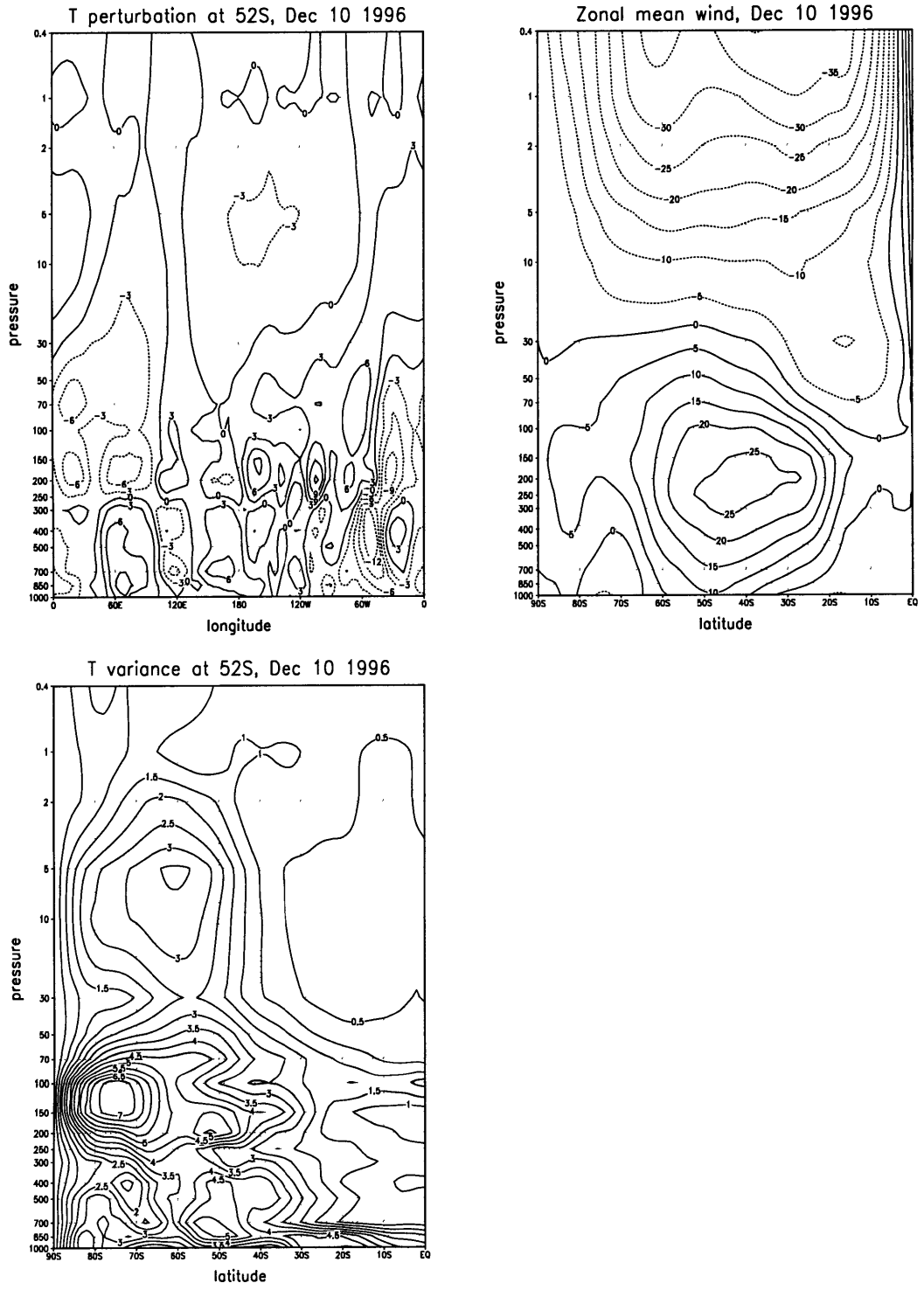

Figure 3.19: Observations of the southern hemisphere on December 10th, 1996. Top left: Height-longitude map of temperature perturbation (temperature minus zonal mean) at latitude 52S (°K). Top right: The zonal mean wind (m/sec). Bottom: Temperature wave number 1 amplitude (°K).

Chapter 4

The dependence of normal mode structure on the wave geometry of vertically varying basic states

4.1 Introduction

In this chapter, we study the effect of wave geometry on the vertical structure of stratospheric waves, with the goal of testing the validity of these relations for observations. We start with the much simpler case where the basic state varies only with height. In chapter 5 we will generalize the results to include meridional variations.

We choose to study vertical wave structures in the context of finding the normal modes of the system, and we discuss a few basic state wave geometry configurations which support different kinds of normal modes, that differ in the location of the interaction with the mean flow. Observationally, the stratospheric basic state varies considerably with time, latitude and hemisphere. The meridional PV gradient structure is very sensitive to the wind and temperature profiles (Sun and Lindzen, 1994). The result is that a very large variety of stratospheric PV gradient structures can be constructed that have fairly realistic looking winds and temperatures, which can support a large variety of wave structures. Observations are too coarse in the vertical to rule out many possibilities. The modes we will discuss most, which are relevant for the rest of our study, are tropospheric baroclinically unstable modes that have a continuation into the stratosphere. As will become clear, the vertical structure-wave geometry relations of these modes applies to the forced wave problem as well (stratospheric waves forced at the tropopause). We solve the eigenvalue problem for modes rather than the forced problem, because it highlights the distinction between the var-

ious kinds of wave geometry configurations, and because of the possible relevance to the eastward propagating waves observed in the southern hemisphere (1.2.3).

We need to say a few words about the differences between the current study and related past studies of the normal modes of the troposphere-stratosphere system (e.g. Geisler and Garcia, 1977, Kuo, 1979, and Fullmer, 1982, for a β plane; Hartmann, 1979, and Strauss, 1981, for a sphere). Most of the discussion in past studies is in terms of Green (1960) modes and Charney (1947) modes, which are the different normal mode solutions of the Charney problem. For a given wavenumber, a few normal mode solutions exist, but only one is unstable while others are decaying. The different modes grow in different regions of wavenumber space. The Charney modes grow at shorter wave lengths, and are the most unstable. The Green modes are unstable at longer wavenumbers, and are mathematically separated from the Charney modes by a neutral wavenumber (zero exponential growth rate). While Charney modes have a shallow vertical structure with amplitudes that peak at the surface, Green modes are deeper and typically have large amplitudes in the stratosphere. This type of discussion does not lend itself well to generalizing the results to other basic states. It is important to note that most of these studies were done before the emphasis on PV dynamics became widespread. The notion of wave geometry (regions of wave propagation and evanescence and the location of critical layers), for which PV is central, will provide a more physically illuminating approach, which will allow us to generalize to the latitude and time dependent problems, and apply our results to specific observed waves on a range of time scales. We will also highlight the fact that the growth mechanism of the Green and Charney modes is physically similar, and the distinction between them is a mathematical artifact of specific features of the wave geometry.

4.2 The model

A one dimensional (height) quasi-geostrophic (QG), β-plane model is used to calculate the normal modes of a given basic state. The formulation follows Lindzen (1994a,b). The nondimensionalized equations of conservation of pseudo potential vorticity, linearized around a zonal mean basic state are:

$$(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x})q' + v'\frac{\partial \overline{q}}{\partial y} = \frac{1}{\rho}\frac{\partial}{\partial z}\left(\frac{\rho \mathcal{H}'}{N^2}\right) + (\nabla \times \mathcal{F}') \cdot \hat{\mathbf{k}} \tag{4.1}$$

q', $\overline{q}$ are the perturbation and zonal mean PV:

$$q' = \zeta' + \frac{1}{\rho}\frac{\partial}{\partial z}(\frac{\rho T'}{N^2}) = \frac{\partial^2 \phi'}{\partial x^2} + \frac{\partial^2 \phi'}{\partial y^2} + e^z \frac{\partial}{\partial z}(\frac{e^{-z}}{N^2}\frac{\partial \phi'}{\partial z}) \tag{4.2}$$

$$\frac{\partial \overline{q}}{\partial y} = -\frac{\partial^2 U}{\partial y^2} + \beta_e - e^z \frac{\partial}{\partial z}(\frac{e^{-z}}{N^2}\frac{\partial U}{\partial z}) \tag{4.3}$$

x, y, z, t are the zonal, meridional, height and time coordinates. ϕ', v', u', T' and ζ' are the perturbation geopotential stream function (see B.1), meridional and zonal winds, temperature and the vertical component of vorticity respectively. U and N^2 are the basic state zonal mean wind and Brunt-Vaisala frequency respectively, taken to be independent of y. β_e is the nondimensional β parameter (see appendix B). The nondimensional density is assumed to be $\rho = e^{-z}$. We express all variables in terms of a geopotential stream function, as follows (nondimensionalized):

$$\begin{aligned} \zeta' &= \tfrac{\partial v'}{\partial x} - \tfrac{\partial u'}{\partial y} = \nabla^2 \phi' \\ v' &= \tfrac{\partial \phi'}{\partial x} \\ u' &= -\tfrac{\partial \phi'}{\partial y} \\ T' &= \tfrac{\partial \phi'}{\partial z} \end{aligned} \tag{4.4}$$

See appendix B for the nondimensionalization constants and other details. $\mathcal{H}'$ and $\mathcal{F}'$ are the heating and momentum damping terms, assumed to act on the perturbation fields only. $\hat{\mathbf{k}}$ is a unit vector in the vertical direction. In most runs, we have no heating and damping, since we are interested first in understanding the unforced, undamped, normal mode structure. We use Newtonian heating and Rayleigh damping when we do retain these terms.

The boundary conditions are a rigid lid at the surface (applied by setting the vertical velocity to zero in the thermodynamic equation), or an Ekman boundary condition, and a radiation condition at the top (see appendix B). For brevity, we will drop the primes from all the perturbation quantities except for the temperature perturbation T', to distinguish it from the zonal mean temperature T.

We assume perturbations on the mean flow which have a sinusoidal horizontal structure with a zonal complex phase speed $c \equiv c_r + ic_i$:

$$\phi(z) = \varphi(z) \cdot e^{i \cdot k(x-ct)+i \cdot ly} \tag{4.5}$$

This results in the following equation:

$$\frac{N^2}{\rho}\frac{\partial}{\partial z}\left(\frac{\rho}{N^2}\frac{\partial \varphi}{\partial z}\right)+\left(\frac{N^2\overline{q}_y}{U-c}-K^2N^2\right)\varphi=\frac{iN^2}{\rho k(U-c)}\frac{\partial}{\partial z}\left(\frac{\rho\alpha}{N^2}\frac{\partial \varphi}{\partial z}\right) \tag{4.6}$$

where $K^2 \equiv k^2 + l^2$ is the total wavenumber and k and l, the zonal and meridional wavenumbers respectively. We have assumed no friction ($\mathcal{F} = 0$)[1] and Newtonian damping $\mathcal{H} = -\alpha\frac{\partial \varphi}{\partial z}$. For given zonal and meridional wavenumbers, the complex phase speed c and the corresponding eigenfunction $\varphi(z)$ are found. The numerical method is described in Harnik and Lindzen (1998).

Since the zonal mean flow has no meridional variation, we are free to choose a meridional wavenumber for our perturbations. Since the wavenumber in the undamped unforced eigenvalue equations appears only as part of the total wavenumber, we do not have to specify it while solving the equations.

The basic states are calculated separately in the troposphere and stratosphere. Unless otherwise noted, the tropospheric PV gradient is specified to be between zero and β,[2] and wind and temperature profiles calculated from it[3]. In the stratosphere, we specify two of the three variables wind, temperature and PV gradient, and calculate the third from them. When we decide to calculate PV gradients from wind and temperature, we specify wind shear and temperature lapse rates, and match the wind and temperature to the tropospheric values at the tropopause (following Lindzen, 1994a). A spline smoother is then applied at a narrow region around the tropopause to keep the first two derivatives of wind and temperature continuous. The results are found not to be sensitive to the width of this smoothing region. In order to facilitate the application of a radiation condition, the wind and temperature are held constant at the top levels. All profiles thus specified have $O(\beta)$ PV gradients in the stratosphere, and a sharp tropopause, characterized by a narrow region of very large PV gradients (10-30 β) between the troposphere and stratosphere.

Specifying PV gradients and one other variable is slightly more involved than calculating PV from winds and temperature. It turns out that in a one dimensional

[1]We use Rayleigh damping in our model studies mostly in order to parameterize a radiation condition. In the one dimensional model this is unnecessary, since it is very easy to implement a radiation condition. In two dimensions it is much harder and we use a sponge layer instead.

[2]Harnik and Lindzen (1998) found that varying the tropospheric values of $\overline{q}_y$ from 0 to β did not significantly affect the long waves.

[3]The PV gradients yield values for isentropic slopes, with arbitrary values at the ground. The partition of the slopes into wind and temperature also involves an arbitrary constant. All of these choices do not affect the results relevant to this study, which makes sense because we are interested in the stratospheric structures. See Harnik and Lindzen (1998) for more details.

model there is an integral constraint on the mass weighted vertical integral of the PV gradient. From equation 4.3 we get:

$$\int_0^{top} e^{-z}\overline{q}_y dz = \frac{U_z(0)}{N^2(0)} - \frac{U_z(top)e^{top}}{N^2(top)} + \beta_e\left(1 - e^{-z}\right) \tag{4.7}$$

Applying the condition that shear at the top is zero (necessary for the implementation of a radiation condition) yields a strong constraint on the mass weighted integral of $\overline{q}_y$. Note that even if we do not require zero isentropic slopes at the top, the value of $\frac{U_z}{N^2}$ is very sensitive to the integral of $\overline{q}_y$ because of the density factor, hence an arbitrary set of $\overline{q}_y$ and N^2 will, in most cases, yield wind profiles that blow up at $z \to \infty$[4]. We therefore need to make sure that the $\overline{q}_y$ we specify satisfies equation 4.7. The tropopause is taken at $z = 1$ (8.9km).

4.3 A wave geometry classification of basic states

In this section we will introduce the wave geometry concept as it applies to our problem. We also identify the different basic state configurations and the possible modes they support. We follow Lindzen et al. (1980) who identify overreflection to be the mechanism by which shear instabilities occur, and classify the wave geometry configuration necessary for instability. The reader is referred to this paper and the references therein for a more thorough treatment of the results we present.

We start by transforming equation 4.6 into canonical form:

$$\psi_{zz} + \left(\frac{\overline{q}_y}{U - c} - K^2 + F(N^2)\right) N^2\psi = 0 \tag{4.8}$$

where we use the following transformation of variables:

$$\varphi = e^{\frac{z}{2}}\sqrt{N^2}\psi \tag{4.9}$$

and

$$F(N^2) \equiv \frac{e^{\frac{z}{2}}}{N}\frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\frac{\partial}{\partial z}(e^{\frac{-z}{2}}N)\right) \tag{4.10}$$

Equation 4.8 is a wave equation, and the index of refraction for vertical propagation

[4]This may be a peculiar property of the one dimensional model, because in two dimensions we have an additional degree of freedom in the meridional wind curvature term.

is:

$$n_{ref}^2 = \frac{N^2 \overline{q}_y}{U - c_r} - k^2 N^2 + F(N^2) N^2 \tag{4.11}$$

where c_r is the real part of the phase speed (see footnote 6). Under WKB conditions[5], the solution is of the form $e^{\pm i \int n_{ref} dz}$, hence in regions where the real part of n_{ref} is non-zero we have wave propagation, and in regions where its imaginary part is non-zero we have exponential behavior[6].

Wave geometry is the configuration of the basic state in terms of vertical wave propagation and evanescence regions. These are separated either by a turning point ($n_{ref}^2 = 0$) or a singular point ($n_{ref}^2 \to \pm\infty$) which in our case happens at the critical level where $U = c_r$. When a wave approaches a turning point, it reflects back because it can't propagate beyond it, unless it can tunnel through to another wave propagation region. In this case we get partial reflection. Waves approaching a critical surface from a wave propagation region are absorbed in it (in the linear limit[7]). If, however, the critical level is separated from the wave propagation region by an evanescent region (i.e. a turning point exists in between), the wave may tunnel to the critical level, in which case it will overreflect from it (Lindzen and Tung, 1978). Tunneling to the critical level will occur only if there is a propagation region or a sink of wave activity beyond the critical layer[8] (otherwise waves will reflect from the turning point before they reach it). Figure 4.1 shows a schematic of these wave geometry configurations.

[5]WKB is valid if the wave length of the solution is much smaller than the variations of the medium, allowing us to make the separation between a wave and a slowly varying basic state. See sections 5.2.1 and 5.3.5 for a quantitative discussion.

[6]Note that we use only the real part of the phase speed in our calculation of the index of refraction. This gives an exact picture of the wave geometry for neutral waves, a good approximation for slowly growing modes and just a crude picture for fast growing modes. A growing wave will decay away from its source (the troposphere in this case) just as waves that are forced by a wave maker that is increasing its forcing amplitude in time will decay in amplitude away from it. Calculating the index of refraction using the real phase speed only will distinguish between decay that is due to the growth of the waves and decay that is due to the medium not supporting wave propagation. Also, using the real phase speed is relevant to waves that have saturated nonlinearly, hence are not growing in time. We expect the vertical propagation of such waves to still be affected by the wave geometry.

[7]We should note here that in the nonlinear limit waves propagating directly to a critical level will oscillate between overreflection and partial reflection, eventually reaching a fully reflecting steady state (e.g. Warn and Warn, 1978).

[8]This is usually the case because $\frac{\overline{q}_y}{U-c}$ (which is the dominant term in n_{ref}^2 near the critical level, equation 4.11) changes sign at the critical level. This is not true in some very special cases, i.e. when $U = c$ at a minimum or maximum of U where $\overline{q}_y$ is monotonous, or when $\overline{q}_y$ changes sign at the critical level, or when $\overline{q}_y = 0$ in the vicinity of the critical level.

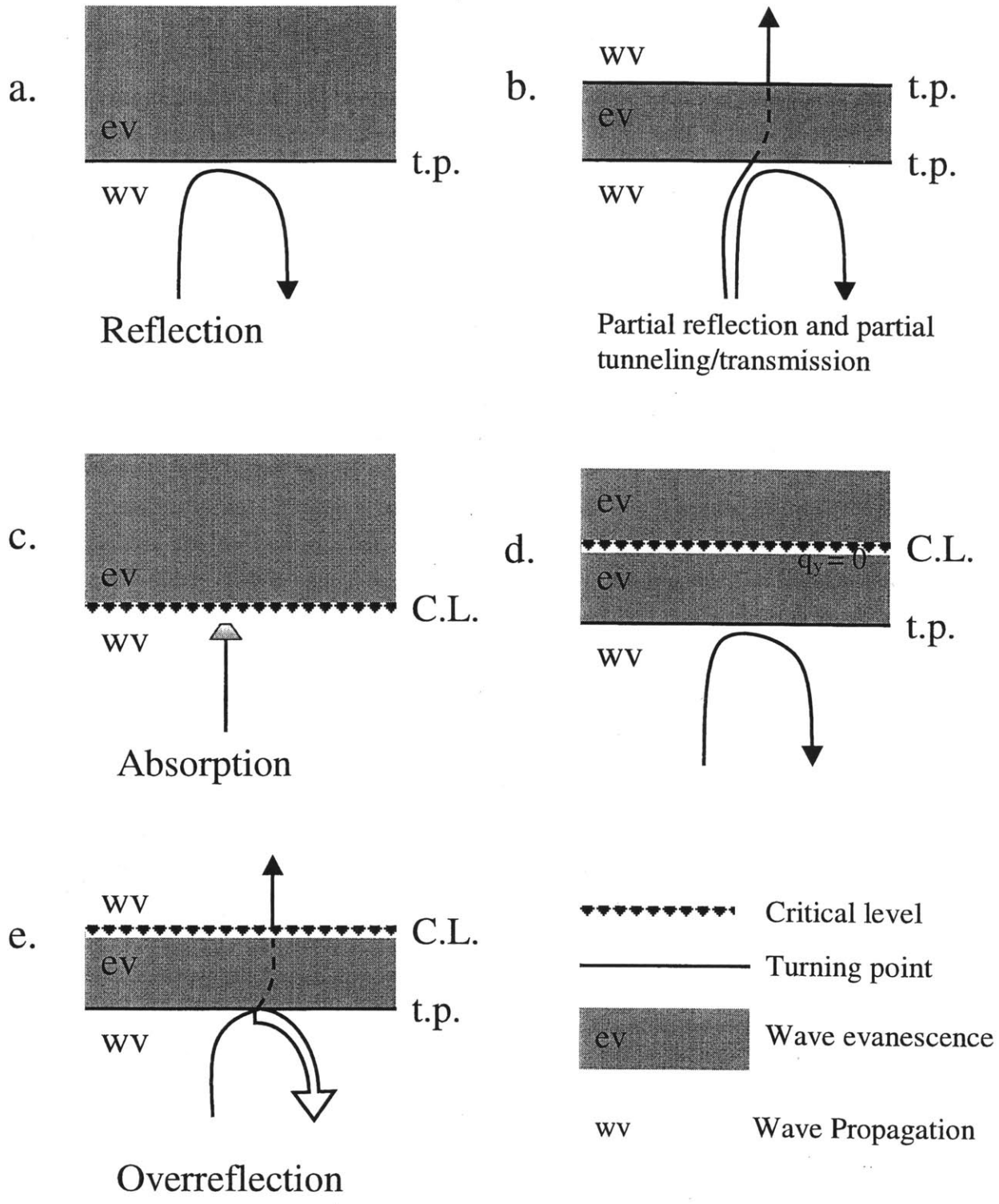

Figure 4.1: The wave geometry of a fully (a) and partially (b) reflecting turning point, and an absorbing (c), reflecting (d) and overreflecting (e) critical level.

The physical mechanism behind overreflection is suggested by Lindzen and Barker (1985) to be the Orr mechanism (Orr, 1907), where a perturbation that is tilted against the shear and is moving with the flow at some level, will be tilted to a more vertical configuration, causing growth of the perturbation[9]. If the wave that overreflects off the critical level reflects back from a turning point (or the surface) in a way that interferes constructively with itself, we can have sustained modal growth through a continuous overreflection-reflection process. For a given wavelength and basic state, the condition for constructive interference will be satisfied only for a certain phase speed (Lindzen and Rosenthal, 1976). Thus, a quantization condition is at the heart of the dispersion relation.

4.3.1 The tropospheric wave geometry.

The main characteristics of the tropospheric basic states we use are order β PV gradients in the troposphere and a large spike of PV gradients at the tropopause. We have shear at the surface, which is equivalent to a δ function of negative PV gradients (Charney and Stern, 1962). Since $U - c$ is also negative at the surface, the index of refraction (equation 4.11) is large and positive there, resulting in an infinitesimal wave propagation region (Lindzen and Tung, 1978). Above the surface, $\overline{q}_y > 0$ and $U - c < 0$, hence $n^2_{ref}$is negative all the way to the critical level. Above the critical level, $\frac{1}{U-c}$ is very large and positive, and we have a wave propagation region. Surface waves can tunnel to the critical level, overreflect off it, and reflect back up from the surface to form a growing mode. This was shown by Lindzen et al. (1980) to be the mechanism of instability in the Charney model. This configuration applies to all wave geometries we consider in this study[10]. Also, the tropopause will always be a wave propagation region because $\overline{q}_y$ is very large and positive, resulting in a positive n^2_{ref}. This tropospheric wave geometry configuration is relevant both to the long waves (Green modes) and the most unstable medium scale waves (Charney modes), and only the stratospheric wave geometry is different. Since the instability results from an interaction of the wave with the mean flow in the troposphere, the

[9]Correspondingly, if the perturbation is tilted in the direction of the shear, it will decay.

[10]Snyder and Lindzen (1988) considered profiles for which the negative PV gradient region extends over some depth, rather than be confined to a δ function at the surface. In particular, they examined states where the $\overline{q}_y < 0$ propagation region extends higher than the critical level, which results in wave propagation below the critical level and wave evanescence above. Such basic states allow overreflection from the tropopause wave propagation region above. Snyder and Lindzen referred to such cases as 'upper level baroclinic instability'. In addition, they considered states with no shear at the ground, but $\overline{q}_y < 0$ in a region above the ground. Since our interest is in the stratosphere, we do not discuss such configurations here.

physical growth mechanism of the long and medium scale waves is similar and the differences between them stem from the differences in the stratospheric basic state. We will elaborate on this point later on.

4.3.2 The stratospheric wave geometry.

Using the conditions for wave absorption, reflection and overreflection, we can now distinguish between three qualitatively different stratospheric basic states, which result in qualitatively different normal modes:

I. PV gradients are positive everywhere, and no critical layers in the stratosphere.

The corresponding normal modes are like the classical baroclinic instability modes which grow by overreflection in the troposphere only (e.g. the Charney model, see section 4.3.1). From equation 4.11 we see that shorter waves will be more likely to have a negative n^2_{ref}. Small waves will therefore have a turning point above the tropopause, and will be evanescent in the stratosphere. Very long waves will in most cases propagate throughout the stratosphere and radiate out through the top of our model. The medium waves will be somewhere in between, with the possibility of having one or more wave ducts in the stratosphere (see figure 4.5, which shows $n^2_{ref}(K)$).

Whether a mode is propagating or evanescent has a large effect on the variation of amplitude with height; WKB theory tells us that to first order, the solution at a given height, away from turning points and the critical level has the form:

$$\phi = \left(\frac{AN(z)}{\sqrt{n_{ref}(z)}} e^{i\int n_{ref}(z)dz} + \frac{BN(z)}{\sqrt{n_{ref}(z)}} e^{-i\int n_{ref}(z)dz} \right) e^{\frac{z}{2} - c_i t} e^{ik(x-c_r t)+ily} \qquad (4.12)$$

where A and B are integration constants. In regions where n^2_{ref} is positive, the zero'th order behavior of the amplitude is of the form $e^{\frac{z}{2}}$. If n^2_{ref} is negative, then $e^{\frac{z}{2} - \int \sqrt{|n^2_{ref}(z)|}dz}$ is the zero'th order behavior[11]. First order effects of n_{ref} come in through the ${n_{ref}}^{-1/2}$ factor. This factor is necessary in order to satisfy wave activity conservation. The wave geometry in the stratosphere is important first of all for the vertical structure of the waves. We will see later that in some cases it can also affect the phase speed and growth rate of the modes (sections 4.4.2, 4.4).

[11] We choose the minus sign before the integral because generally, the waves decay away from their source (the troposphere in this case) if they are not able to propagate away from it.

II. PV gradients are positive everywhere, with one or more critical layers in the stratosphere.

Since $U - c$ and $\bar{q}_y$ are positive in the upper troposphere and lower stratosphere, a critical layer in the stratosphere will cause $n^2_{ref} \to \infty$ below it, and $n^2_{ref} \to -\infty$ above it (this configuration is similar to figure 4.1.c). The corresponding normal modes grow by tropospheric overreflection, but at the same time they are absorbed at the stratospheric critical level. Such basic states occur in summer, when there are easterlies in the stratosphere, and also possibly in spring, when the polar vortex breaks down. The effect of the stratospheric critical level is first of all on the amplitude of the waves, which drops sharply to zero above the critical level. Apart from that, there may be some effect on the growth rates. Since the critical level acts as a sink of wave activity, its effect is similar to putting damping in the stratosphere. We will not show examples of modes with a stratospheric critical level, rather, the reader is referred to section 4.4.2 (also figure 4.8) where we discuss the effects of Newtonian damping.

III. There are one or more regions of negative PV gradients *and* one or more critical levels in the stratosphere.

The existence of one or more regions of negative PV gradients may allow an interaction with the stratospheric critical level (an overreflection from it), resulting in *stratospheric* modes. There are few different configurations which can allow for qualitatively different modes. Not considering the special cases when n^2_{ref} does not change sign at the critical level (footnote 8), we distinguish between critical levels that have a wave propagation region below, and an evanescent region (underneath another wave region) above, and the opposite, when the evanescent region is below the critical level. The difference is in the direction from which overreflection can occur. What determines which configuration we have is the sign of shear and PV gradient at the critical level. Figure 4.2 shows the different wave geometry configurations that support overreflection. We divide them into a critical level in a region of positive shear, with negative (a) or positive (b) PV gradients, and a critical level in a region of negative shear with negative (c) or positive (d) PV gradients.

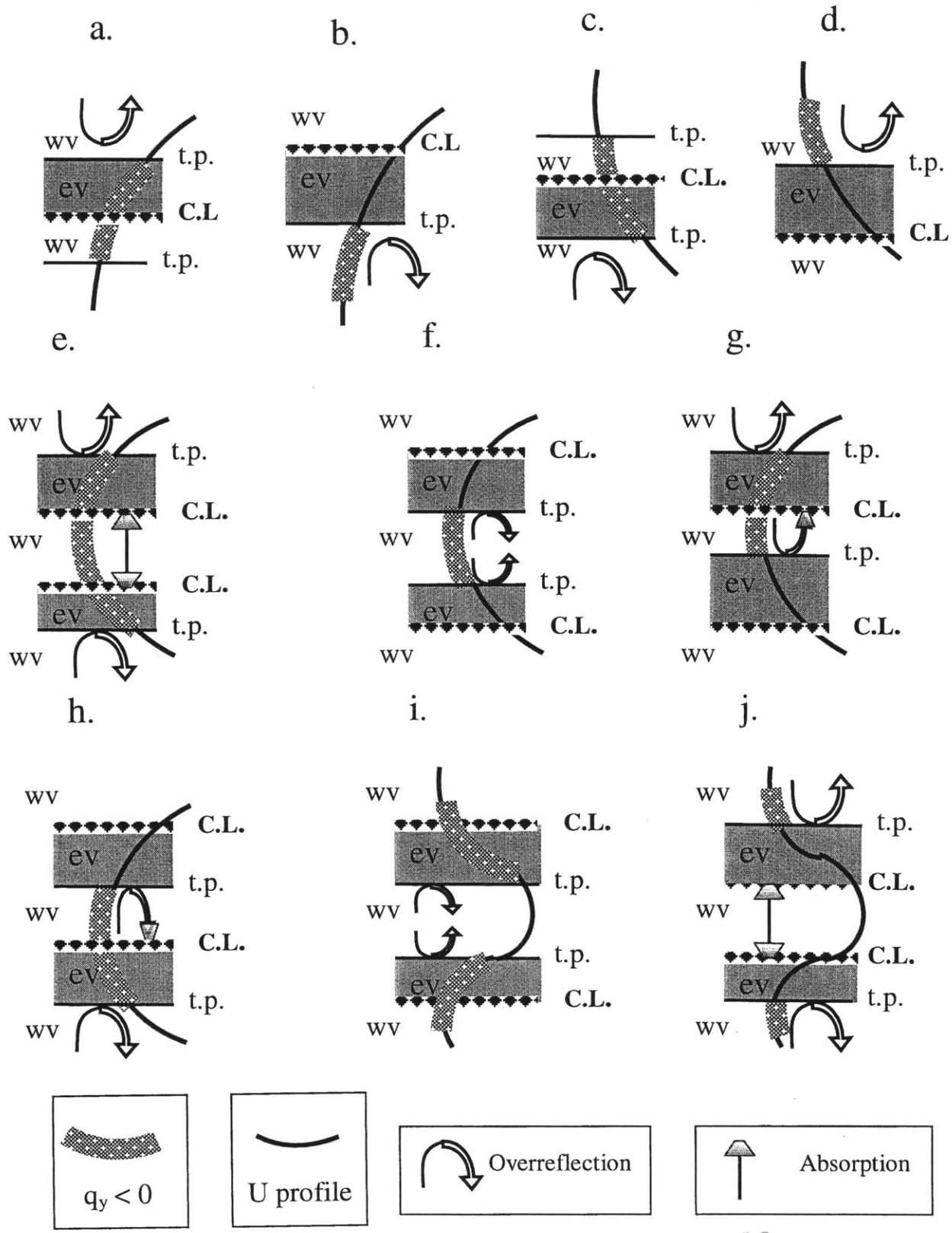

Figure 4.2: The various wave geometry configurations that support different kinds of stratospheric modes. See text for details.

We see that when the shear and the PV gradient have the same sign at the critical level, overreflection occurs below the critical level, and when they have opposite signs, overreflection occurs from above. When the PV gradients are positive at a critical level in a region of negative(positive) shear, an unstable mode can exist only if there is a region of negative PV gradients above(below) the critical level, which serves as the wave propagation region from which waves tunnel to the critical level and overreflect[12]. An important point to note is that for PV gradients to be negative, we need a large enough positive wind curvature, or alternatively a large enough negative shear (equation 4.3). Negative PV gradients are therefore more likely to occur near a minimum in westerly winds (or a maximum in easterly winds).

Also shown in figure 4.2 are possible combinations of two critical levels. We expect to see two critical levels, not just one, in cases where the region of negative PV gradients is near a local wind minimum or maximum. The different possibilities are:

1. A wave propagation region that has on each of its sides a turning point with a further critical level, allowing overreflection from both directions. The overreflection off of each side enhances the other, as long as the quantization condition is satisfied (plots f, i).

2. A wave propagation region bounded on both sides by critical levels. There are also two wave propagation regions beyond the critical levels and their evanescent regions. Waves can overreflect off of both these critical levels from the outer wave regions. The wave activity that tunnels into the middle wave propagation region from overreflection off one critical level gets absorbed in the other, i.e. each critical level acts as a wave energy sink for the other (plots e, j).

3. A wave propagation region bounded by an absorbing critical level on one side and a turning point and further critical level on the other. The waves overreflect off one side of the wave propagation region and get absorbed at the other. The only way such a basic state can sustain an exponentially growing mode is if waves are continuously excited in the propagation region by overreflection from the other side of the critical level (plots g, h).

Since the phase speed of the modes (the location of the critical level) is part of the solution, it is quite tricky to construct some of the configurations above. Having two

[12]As was pointed out by Lindzen and Tung (1978), this is the Charney-Stern (1962) necessary condition for instability

critical levels rather than one will affect the vertical structure of the modes, and will probably also affect their phase speed and growth rate. It could be, however, that the interaction with one of the critical levels will be dominant, while the interaction with the other will be forced. This is certainly so when one of the critical levels is the tropospheric one (which is dominant). By comparing modes of basic states that differ only in the stratosphere, we find that the tropospheric instability is dominant in our results in the sense that it sets the phase speed (the interaction of the mode with any critical levels in the stratosphere may have a small but secondary effect on the magnitude of the phase speed). It is important to note that we did not necessarily find all possible normal mode solutions because our search routine only finds the most unstable modes. We did not find any internal stratospheric instabilities with phase speeds that are higher than tropospheric phase speeds (i.e. any modes that do not interact with the mean flow at a tropospheric critical level). We did not, however, look for purely stratospheric modes on basic states that do not support tropospheric baroclinic instability[13]. It is very possible that stratospheric modes with higher phase speeds do exist in our model. We leave these for future study.

The fact that the tropospheric basic states set the phase speeds to tropospheric values means the stratospheric critical levels exist only on zonal wind profiles that have a decrease in wind above the tropopause. In section 4.5 we will show an example of a mode which is essentially like figure 4.2.f, in addition to having a critical level in the troposphere.

We decided to focus on the deep tropospheric instabilities, and leave the internal stratospheric modes for future study for two reasons. First, most observed waves are quasi stationary, and the prominent eastward stratospheric propagating waves occur in winter in the southern hemisphere, when there are no critical levels in the stratosphere (the basic state wind in the stratosphere is much larger than the phase speed of the modes). The deep tropospheric modes are therefore the most likely candidates to explain observed waves. Such waves are not observed in southern hemisphere spring (March, April) and in the northern hemisphere spring and fall (March, April, September, October, also maybe in February), when the basic states that are most likely to support internal stratospheric instability occur (basic states with a minimum in westerly wind at latitudes 50-70°, above the tropospheric jet). The second reason is that even if internal stratospheric modes do exist, observing them is hard because they have very sharp features (e.g. figure 4.10). Our results from chapter 3 show that

[13]Internal baroclinic instability of the *mesospheric* summer easterly jet has been suggested in the past as a source for the mesospheric 2-day waves (Plumb, 1983). As far as we know, there have been no studies of internal baroclinic instability of stratospheric basic states.

the retrievals of such waves may be very problematic (sections 3.4.3, 3.5). In addition, the observations of zonal mean wind may not be good enough to distinguish between a basic state with negative PV gradients and one with only small but positive $\overline{q}_y$. It should be noted that the stratospheric part of the deep tropospheric instability modes can be regarded as forced, with a forcing that has a zonal phase speed and is increasing with time. Therefore, the relations between wave structure and the wave geometry apply to quasi-stationary waves, and to eastward propagating waves even if they are forced by nonlinear interactions at the tropopause and not by instability (see discussion in section 1.2.3).

4.4 The normal modes on basic states with no critical levels in the stratosphere

In the following section, we will show how the wave geometry view applies to our model results, for basic states that have no critical levels in the stratosphere. In section 4.4.1 we show the results of a specific run to illustrate the general features of the solutions. In section 4.4.2 we discuss the effect of wave geometry and Newtonian damping on the growth rates, in section 4.4.3 we discuss the effect of surface damping and in section 4.4.4 we discuss the sensitivity of the results to various parameters.

4.4.1 The relation between the index of refraction, the dispersion relation, and the vertical structure of the modes

Figure 4.3 shows the basic state wind, PV gradient, N^2, and temperature of this run. The PV gradient in the troposphere is set to 0.5β. In the stratosphere, the PV gradients are of order β. The wind profile is characteristic in shape and magnitude of the winds of a southern hemisphere June-August basic state.

Figure 4.4 shows the real and imaginary phase speeds as a function of nondimensional total wavenumber (K) for unstable modes on this basic state. On top is the dispersion relation for a large range of wavenumbers. On the bottom is a blowup of the long wave spectrum. There are three main regions, based on the dispersion relation (top figure); The *medium waves* which are the fastest growing, the *short waves* which grow very slowly (in the limit of zero PV gradients in the troposphere these waves are neutral, as in the Eady model), and the *long waves* which grow slower than the medium scale waves.

The long and medium waves are the traditional Green and Charney modes re-

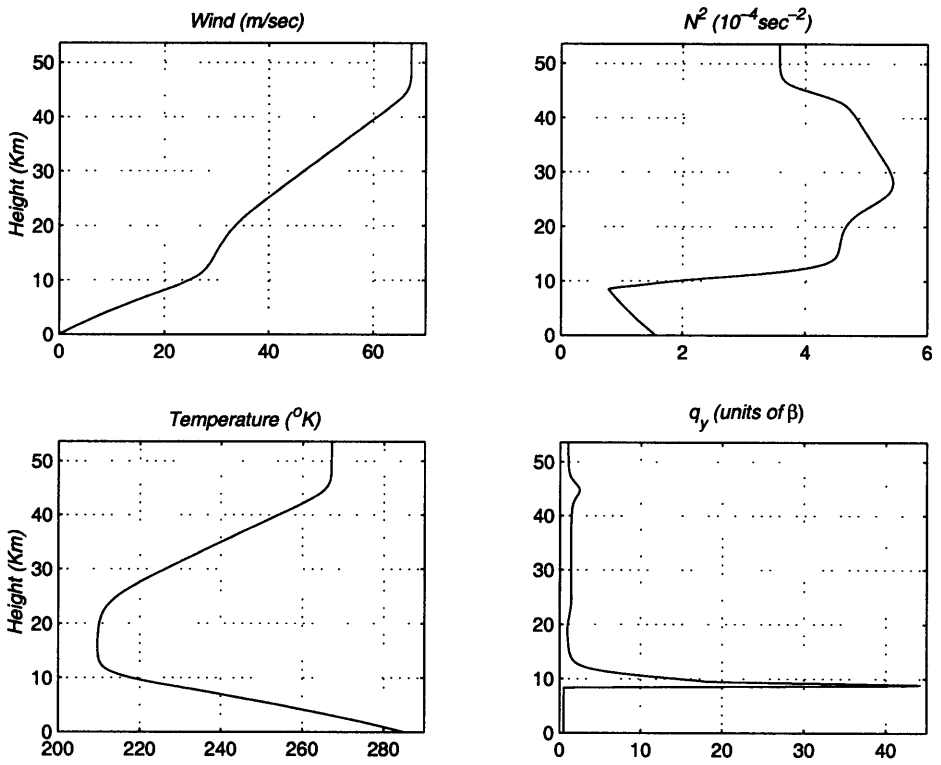

Figure 4.3: The basic state wind (top left), Brunt Vaisala frequency (top right), Temperature (bottom left) and PV gradients (bottom right) used in the standard model runs. Height is in kilometers, wind in m/sec, N^2 in $10^{-4}sec^{-2}$, temperature in $^\circ K$ and PV gradients in units of β.

spectively, and the neutral point that, by definition, separates them exists in this run at $K = 0.5$. There is another neutral point at longer waves ($K = 0.22$). A closer look at the long wave region (bottom figure) reveals a lot of structure in both the real and imaginary phase speeds. The neutral points discussed above appear to be neutral regions ($K = 0.4 - 0.5$, $K = 0.2 - 0.22$). Between them we find regions of relatively large growth rates. The longest waves also have non-zero imaginary phase speeds.

The index of refraction also has a few distinct regions in wavenumber space, that correspond to the different regions in phase speed. Figure 4.5 shows height-wavenumber contour plots of the index of refraction squared calculated using equation 4.11, for the long and medium wavenumbers shown in the bottom of figure 4.4. On the right(left) is n^2_{ref} calculated with(without) the N^2 derivative terms in $F(N^2)$ (equation 4.10). The wavenumbers shown are the same as in the bottom of figure 4.4. The term $F(N^2)$ has a contribution from the density factor ($-\frac{N^2}{4}$) and a contribution from N^2, which depends only on its first and second derivatives with height. If N^2

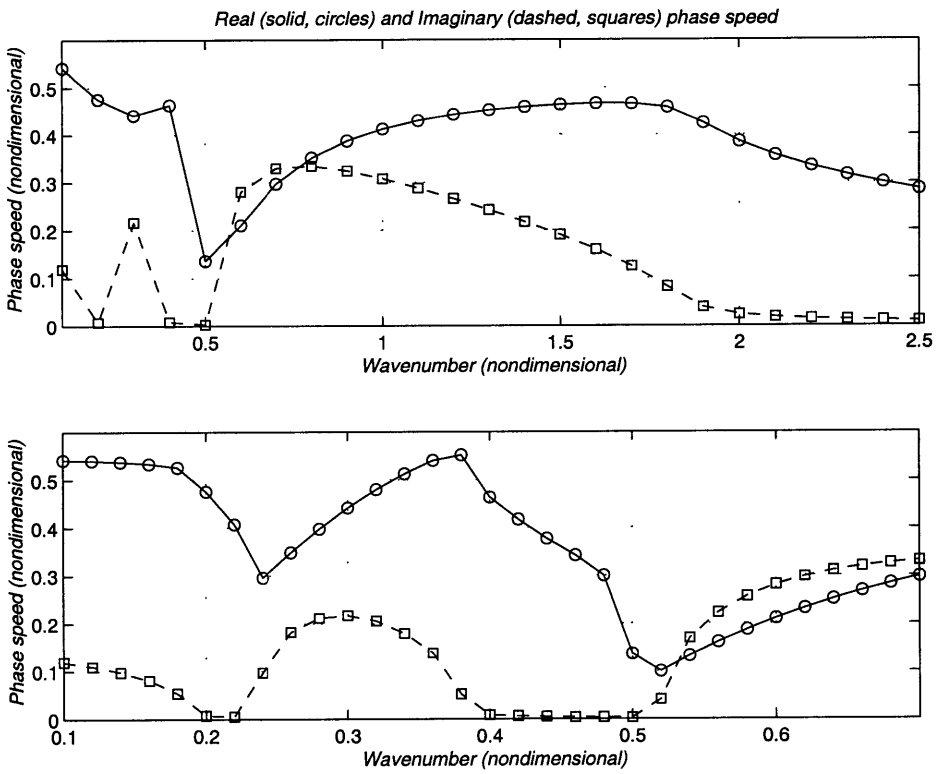

Figure 4.4: The dispersion relation for the basic state of figure 4.3. Real (solid, circles) and imaginary (dashed, squares) phase speed. Multiply by 30 to get in units of m/sec. Circles and squares mark the wavenumbers that are actually calculated. Wavenumbers are nondimensional. Top: Full range of long, medium and short wavenumbers (k=0.1:0.1:2.5). Bottom: The long wave region only (k=0.1:0.02:0.7).

is constant, $F(N^2) = -\frac{N^2}{4}$. As mentioned before, the index of refraction without the N^2 derivative terms is the one relevant for φ (equation 4.6). The full n_{ref} is relevant for ψ. We see that the differences are generally small except at regions where N^2 varies rapidly (near the tropopause and stratopause). In these regions, n^2_{ref} with the derivative terms is negative, suggesting wave evanescence. In the simpler form of n^2_{ref}, this evanescence still exists, but it is hidden in the rapid variations of φ in regions where N^2 changes rapidly. In these regions the wave nature of φ is violated (it varies faster than a wavelength)[14]. These regions have little effect since they are very narrow and waves tunnel through them easily. Since the wave geometry looks simpler without the N^2 derivative terms, we prefer looking at it. Looking at the simpler form, without the N^2 terms, we see that the longest waves, beyond the long

[14]A quantitative example is shown in figure 5.8, where the magnitude of $\frac{d}{dz}\left(\frac{1}{m}\right)$ is plotted. The regions where this term is much larger than 0.5 are shaded and they coincide with regions where N^2 (figure 5.1) varies rapidly.

wave neutral point ($K < 0.2$) can propagate all the way through the stratosphere. Waves between $K = 0.2 - 0.5$ have a turning point in the upper stratosphere, and wavenumbers larger than $K = 0.5$ (the Charney-Green neutral point) are evanescent in the stratosphere. Note that all the wavenumbers have similar tropospheric wave geometries (propagation in the tropopause region, where the PV gradients are large enough to allow it, and in a narrow region above the critical level). The main differences are in the stratospheric wave geometry.

As we already pointed out, the division of stratospheric n^2_{ref} into regions of wavenumber space is similar to the division in the dispersion relation. This is expected to some extent, since the index of refraction is a function of the phase speed. The coincidence of the neutral point with the boundary between stratospheric propagation and evanescence is not general. On some basic states (mostly those that have a minimum in winds above the tropospheric jet peak), the longest medium scale waves may have a propagation region in the stratosphere, and correspondingly, quite large amplitudes there.

The main point we want to show in this section, however, is that the index of refraction (which is part of the solution to the extent it depends on phase speed) has a direct effect on the vertical structure of the waves. Vertical structures are calculated for all wavenumbers for which c is calculated. As expected, we find that the modes in each of the regions of n^2_{ref} and the dispersion relation curves have distinct vertical structures. Figure 4.6 shows vertical structures of waves from each of these regions and of the neutral wave ($K < 0.2$, $0.2 < K < 0.5$, $K = 0.5$, $K > 0.5$). Shown are the amplitude and phase of the temperature perturbation and of the geopotential stream function perturbation (see appendix B for exact definition), with and without the density contribution. For easy comparison, we also show the real part of the index of refraction for vertical wave propagation for these wavenumbers. Since we are looking at φ and not at ψ (equation 4.9), we use n_{ref} without the vertical derivatives of N^2.

We see that the two long waves have relatively similar tropospheric structures, with a minimum of $|\phi|$ at the critical level (where the phase increase with height is most rapid) and relatively constant temperature amplitude and phase. Long wave amplitudes are much larger in the stratosphere than in the troposphere[15]. This may be important because one of the main problems in associating the eastward propagating observed waves with normal mode instability is their apparent lack of a tropospheric continuation. This may be due to a much smaller signal-to-noise ratio in the tropo-

[15]In the current example, the geopotential height can be more than 10 times as large at 40 km than in the troposphere. This however will depend strongly on damping, which is not included in this run.

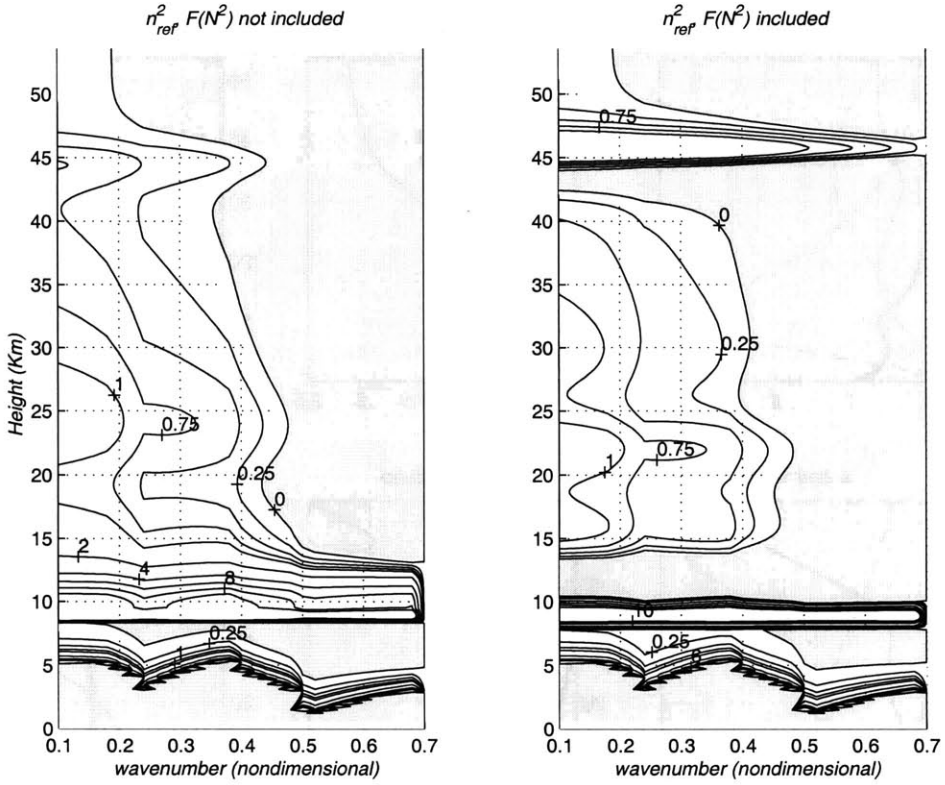

Figure 4.5: Height-wavenumber plots of the index of refraction squared (n^2_{ref}), not using the derivative terms in F(N^2) (left) and including these terms (right, see text for explanation), for the long waves shown in the bottom of figure 4.4. $n^2_{ref}$is in nondimensional units. Contour values are 0:0.25:1,2:2:10. Negative (wave evanescence) regions are shaded. Height is in kilometers.

sphere. We also see that the phase of the long waves increases with height, indicating a westward tilt of phase lines with height and vertical propagation. This fits nicely with the picture of an unstable wave growing in the troposphere and propagating upwards into the stratosphere.

There are, however, large differences in the stratospheric temperature structure of the long waves of the different regions. The temperature amplitude and phase of the longest waves ($K = 0.1$, solid line) increase throughout the stratosphere, while the slightly shorter long waves ($K = 0.3$, dashed line) have a more confined structure with a broad maximum in the upper stratosphere and a decrease to tropospheric amplitudes above 45 km, where the phase is constant with height. These differences correspond well with the differences in n^2_{ref}: while the longer waves propagate all the way through the stratosphere, the shorter waves have a turning point at around $45km$ (n^2_{ref} becomes negative). These differences are also manifest in the geopotential height, but not as clearly ($|\phi|e^{-\frac{z}{2}}$ is more constant with height for the vertically

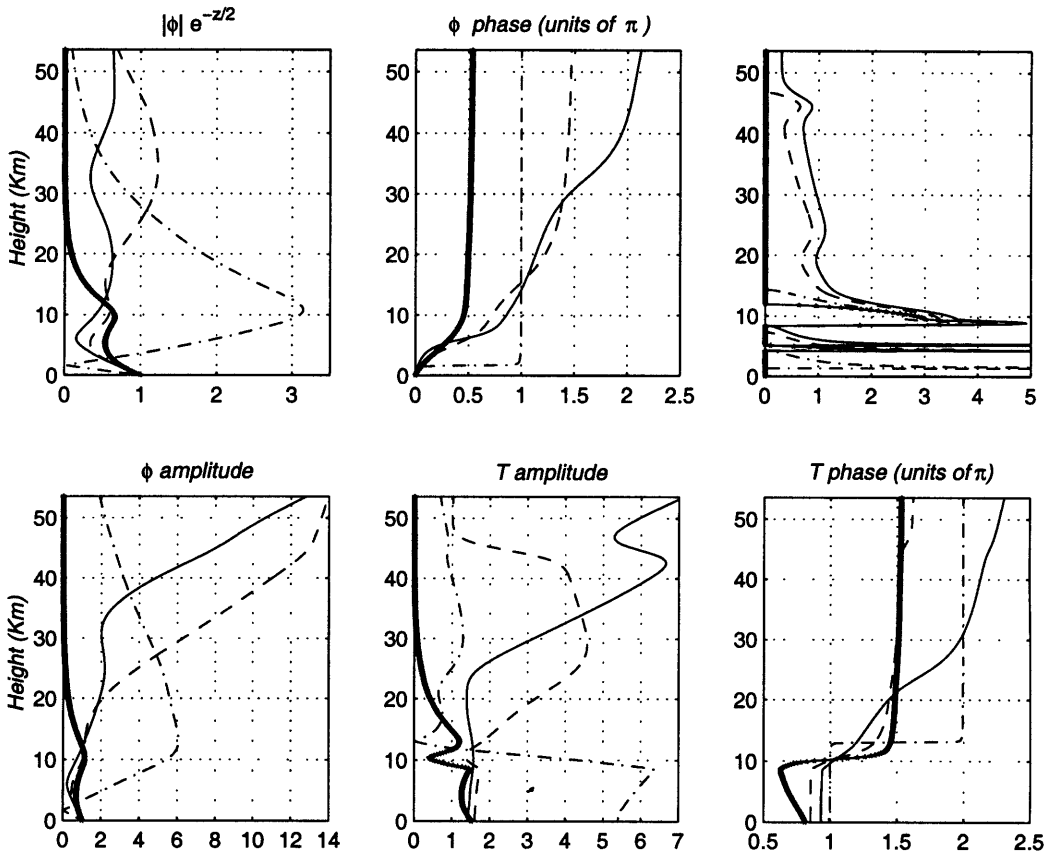

Figure 4.6: Top: Amplitude (left) and phase (middle) of the geopotential stream function, without the density contribution ($\varphi e^{-\frac{z}{2}}$), and the real part of the index of refraction (n_{ref}, not using N^2 terms, see text for details). Bottom: The geopotential stream function amplitude including the density effect (left), and the temperature amplitude (middle) and phase (right), plotted for a few wavenumbers. Wavenumbers shown are $K = 0.1$ (thin solid), $K = 0.3$ (dashed), $K = 0.5$ (dash-dotted), and $K = 1.1$ (thick solid). Height is in kilometers. Phase is in units of π. $n_{ref}$is in nondimensional units.

propagating $K = 0.1$, while it has a peak in mid-stratosphere which is indicative of downward reflection, for $K = 0.3$). Note however, that the $K = 0.1$ wave has a small amount of partial downward reflection (evident from the small undulations in $|\phi|e^{-\frac{z}{2}}$ and in temperature amplitude), and the $K = 0.3$ is not fully reflected downward, and some of it leaks through the top of the model (a non-zero, but small increase of phase with height even above 45km).

The medium waves ($K = 1.1$, thick line) have a very different stratospheric structure- their amplitude decays rapidly above the tropopause, and the phase is constant with height (the tropospheric structure is quite similar to the long waves). This decay, along with no phase shift with height, corresponds well with the fact that

medium waves are evanescent in the stratosphere. It is interesting to compare the medium waves to the neutral wave ($K = 0.5$, dash-dotted), which is also evanescent in the stratosphere. The latter decays with height in the troposphere, but much more slowly. This highlights another factor which contributes to the decay with height of the medium scale waves, and that is their growth in time. Charney and Pedlosky (1963) were the first to point out in this context that growth in time will be manifest in a decrease in amplitude away from a source of wave energy which is growing in time (the source in our case is the troposphere).

Finally, while the unstable waves have a westward phase tilt with height (some larger than others, and the medium scale waves have a phase tilt only in the troposphere), the neutral waves have a vertical phase structure in which all the phase variation is concentrated at nodes (where the phase jumps π radians). This is true both in the troposphere and stratosphere.

Figure 4.7 shows longitude-height structures of a few of the long waves, assuming a zonal wave 1, for the purposes of comparing to observations and 2 dimensional model studies later on (e.g. figures 1.6 and 5.5). Shown are the two long waves of figure 4.6. We see that the $K = 0.3$ wave, which is evanescent at the top of the stratosphere, has a smaller ϕ phase tilt at the top of the domain, compared to wavenumber 0.1 which has no turning point. Also, the temperature structure is more vertical and more confined. It peaks in mid-stratosphere, as opposed to the upper stratosphere. The temperature structure in general shows more variability because it is a vertical derivative of ϕ. This makes the combination of ϕ and temperature fields a very useful diagnostic for wave geometry.

Also shown in figure 4.7 is the neutral long wave ($K = 0.2$). The wave is clearly a standing wave in the vertical. There is no phase tilt with height and there are two nodes in ϕ, one in mid-stratosphere and the other at the critical level in the troposphere. The pattern of this wave looks like a version of the longest wave ($K = 0.1$), that was tilted into the vertical.

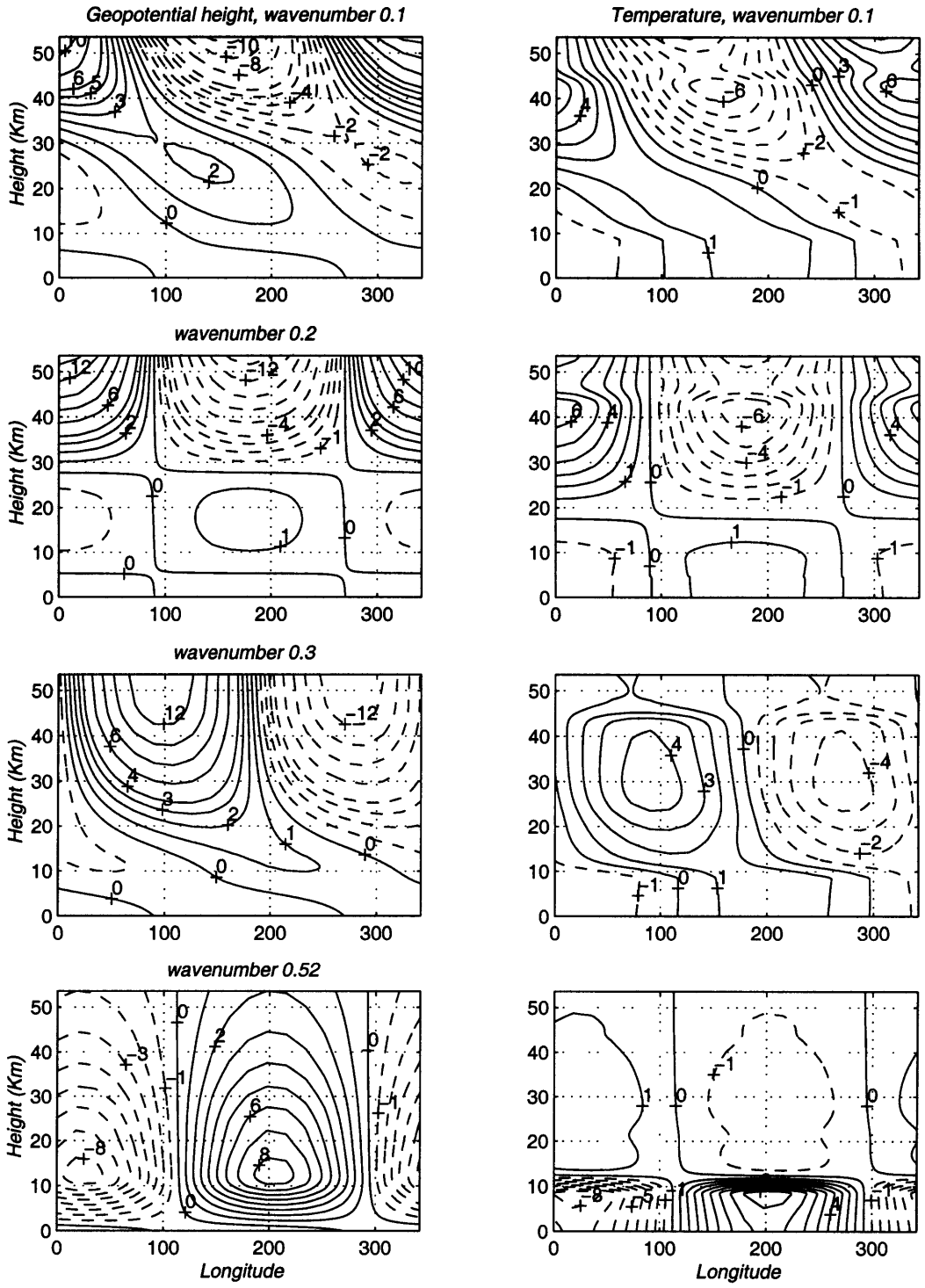

Figure 4.7: Longitude-height structure of geopotential stream function (left) and temperature (right), for various total wavenumbers K, assuming a zonal wave one ($S_x = 1$). From top to bottom, we have the longest wave ($K = 0.1$), the long neutral wave ($K = 0.2$), the fastest growing long wave ($K = 0.3$), and the longest medium wave ($K = 0.52$). Height is in kilometers. Negative values are dashed.

An interesting point to note about the node at the critical level is the following: A strictly neutral wave, assuming normal mode structure, has a real phase speed, and in the absence of damping, at least locally at the critical level, the equations become singular:

$$(U - c)q' + \overline{q}_y \phi' = 0 \tag{4.13}$$

In the Eady model $\overline{q}_y = 0$, hence this is not a problem. In the Charney model $c_r = 0$ when c_i=0, which places the critical level at the ground where again $\overline{q}_y = 0$. In the present case, the neutral waves do have a critical level, and $\overline{q}_y \neq 0$, hence the geopotential stream function has to vanish, by having a node. Bretherton (1966) was the first to note that neutral waves can not occur unless specific conditions are met at the critical level, however, he stressed the possibility of $\overline{q}_y = 0$ at the critical level, and assumed the case where $v' = 0$ ($\phi' = 0$) at the critical level is rare. Since long waves generally have a node (or almost node, when $c_i > 0$) somewhere in the troposphere, it is not as surprising to find modes that have $v' = 0$ at the critical level.

Finally, also shown in figure 4.7 is the structure of the longest medium wave, just beyond the neutral point. We see that the ϕ perturbation is maximum at the tropopause, and decays quite slowly in the stratosphere (its growth rate is quite small). The structure is almost vertical, consistent with a perturbation that is slowly growing, and is evanescent in the stratosphere. The temperature perturbation is much larger in the troposphere than in the stratosphere. Note that both neutral waves are fully reflected downward from a turning point, and the main difference between them is in the location of the turning point.

4.4.2 The dependence of growth rate on the wave geometry and Newtonian damping

As was shown by Ioannou and Lindzen (1986), the meridional wavenumber makes a very big difference for the maximum growth rate of the long waves (because the growth rate is kc_i, and for a given total wavenumber, l determines k). In the real atmosphere, we will only see integral zonal wavenumbers. Therefore, assuming the meridional structure is determined externally by the basic state confinement (see chapter 5), the wavenumbers that will be relevant to the stratosphere will be set by the basic state. For example, for no meridional confinement only waves 1 and 2 have large amplitudes in the stratosphere (wave 3 is a medium scale wave), while for large enough values of l, the very long waves are not relevant- they have imaginary zonal wavenumbers. The structure of the dispersion relation (bottom of figure 4.4) in this case becomes

crucial, because the waves will have large or small growth rates depending on whether their integral zonal wavenumbers lie in the regions of very slow growth or not. It is therefore important to understand what causes some wavenumbers to grow and others to be neutral. This also highlights the importance of understanding the latitudinal behavior of the perturbations (which we discuss in more detail in chapter 5).

In the Charney model, the neutral point forms as a result of the formation of a standing wave pattern above the critical level, which forces a node at the critical level (Lindzen et al., 1980). The wave geometry framework allows one to generalize from simpler models like the Charney model to more complex basic states like our own. We expect vertical reflection off of turning points in the stratosphere to interfere destructively with the perturbation at the critical level, just as in the Charney model.

A simple way to test this is to inhibit the reflection at the turning point, by putting damping there. The damping we use is Newtonian cooling, which is specified to grow gradually to effective magnitudes only in the upper stratosphere. Figure 4.8 shows the results of adding Newtonian damping to the undamped model described in previous sections. Shown is the dispersion relation for the long waves (top), along with the vertical structure (left and middle of two bottom rows), the index of refraction for wavenumber $K = 0.22$ (middle row on right), and the damping coefficient α (middle row on right, thick line). The damping time scale reaches an amplitude of 1 day at around 40 km. It is essentially infinite below 20 km. The damping affects the phase speed of the longest waves, those that are propagating in the damping region. We see that the damping homogenizes the real and imaginary phase speeds of the modes. In particular, the undamped neutral modes that have a turning point in the damping region ($K = 0.2 - 0.22$) are not neutral when the damping is added. The neutral modes that are not affected by damping ($K = 0.4 - 0.5$) have a turning point below the damping region (see n_{ref} in middle row, right panel).

These results support our picture that the neutral points in our model result from interference of the downward reflected wave with the wave field at the tropospheric critical level. In the real atmosphere, we do not expect this to happen so easily for a few reasons. First, we have some damping which will most likely cut at least part of the downward reflection. Second, the reflection is from a surface which is not necessarily simple geometrically, making any sort of destructive interference less likely to occur. Third, we expect the wave to take a relatively long time to reach steady state because it involves the wave propagating up to the turning point and back down to the critical level, at least once.

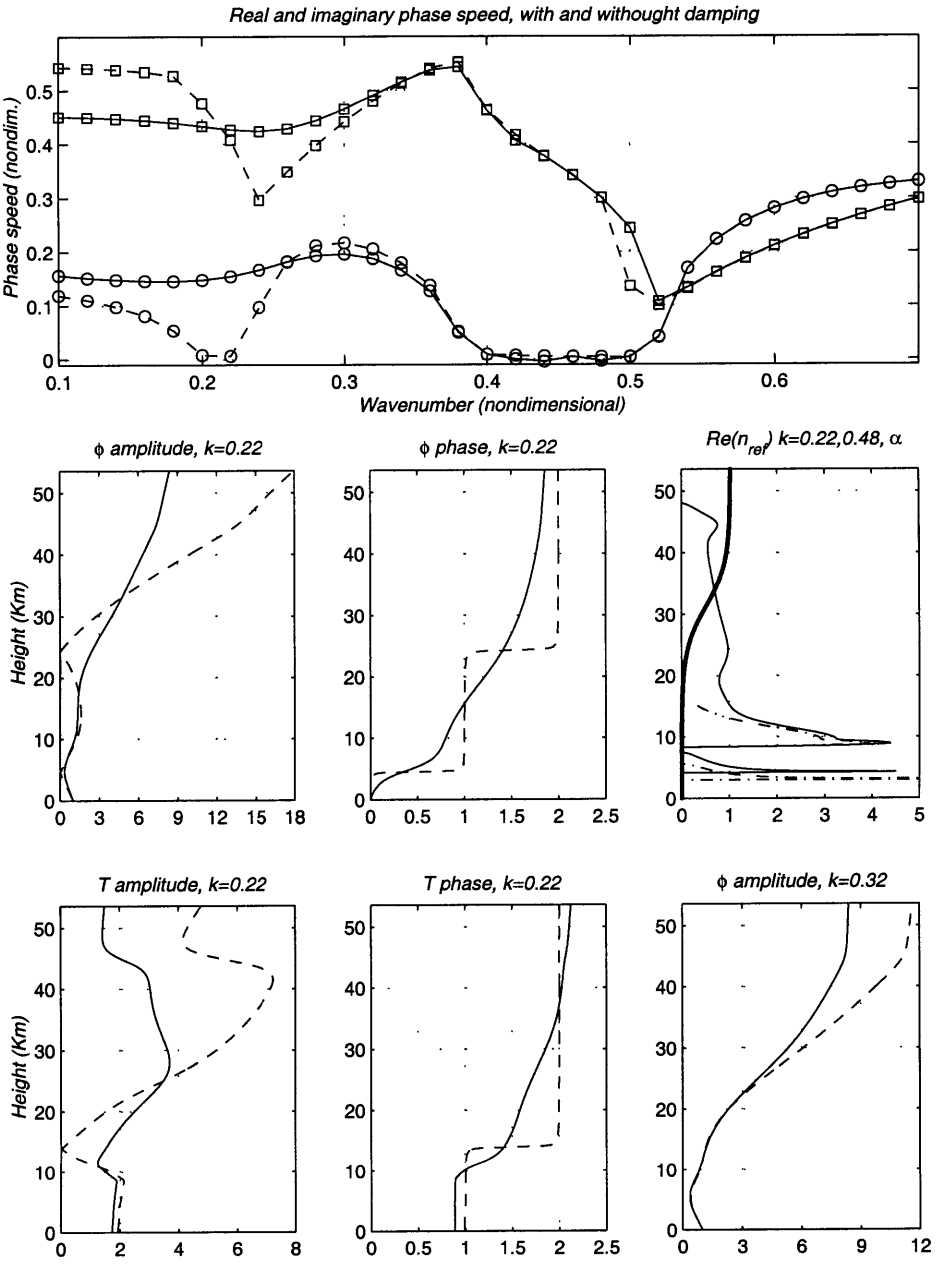

Figure 4.8: Results of the undamped model (dashed) and the model with Newtonian damping (solid). Top: Real (squares) and imaginary (circles) phase speed, nondimensional units. Middle row: ϕ amplitude (left) and phase (middle) for the neutral long wave ($K = 0.22$). n_{ref} (right) for the neutral waves $K = 0.22$ (solid) and $K = 0.48$ (dash-dot), and the damping coefficient α (thick) in units of day^{-1}. Bottom: Temperature amplitude (left) and phase (middle) for the neutral long wave ($K = 0.22$). ϕ amplitude for the fastest growing long wave $K = 0.32$ (right). Phase is in units of π, amplitudes and n_{ref} are nondimensional.

We show in chapters 6 and 7 that it takes Rossby waves a few days to traverse the stratosphere. Note that initially, when the modes develop in the troposphere and are only on their first time up through the stratosphere, we expect all wavenumbers to have similar phase speeds and growth rates, as in the damped case, because they do not 'see' their turning points yet. This highlights the fact that neutral points do not separate between physically different modes. We can generalize to the neutral point dividing the Charney and Green modes and conclude that the growth mechanism of both is similar— an interaction of the mode with the tropospheric critical level, and the differences are due to the propagation characteristics in the stratosphere or to the existence of a turning point.

We also see that damping affects the modes mainly in two ways. The first is to inhibit reflection from the turning point and make the wave vertically propagating. This can be seen by looking at the phase of the perturbation- damping gets rid of the nodes, and causes the phase to increase with height. The other effect is to reduce the wave amplitudes in the damping regions. Newtonian damping acts directly on temperature, but its effect on ϕ is also to reduce its amplitude. It is interesting that the amplitude of the neutral wave $K = 0.4$ is hardly affected by the damping, even though its amplitude reaches a peak of 4.0 at 40km (not shown). This is because the wave is evanescent in the damping region, and the amplitude peaks there only because of the density effect. Since damping is one of the largest unknowns in the real atmosphere, it is useful to understand its effect on the large scale structure of the waves in order to understand its contribution to differences between observed and modeled waves.

As mentioned in section 4.3, in the Charney model, neutral points form when the phase accumulation of the wave over the wave duct, calculated as follows, is π:

$$\Delta phase = \frac{1}{\pi} \int_{p.r.} n_{ref} dz \tag{4.14}$$

where the integration is over the entire wave propagation region (p.r.). To see if this rule holds, we calculated $\Delta phase$ as a function of wavenumber, for different basic states and parameter values[16]. We find a few things. First, when we have a sharp

[16]In our model, we have more than one propagation region. There is a very narrow region of evanescence for all but the longest wavenumbers, just below the tropopause. We would expect multiple reflections to complicate the interference at the critical level and to get rid of the neutral points, however, this does not happen, probably because the evanescence region is so small that it acts like an internal scattering region. In some cases we have no neutral points, but we do see wavenumbers that have a reduced growth rate and these basic states have a more obvious multiplicity of propagation and evanescence regions, for a larger range of wavenumbers.

tropopause, the neutral point does not occur at a phase accumulation of π, but for a smaller number. This is due to partial reflections the wave undergoes as it propagates through regions where the basic state is varying rapidly. It makes sense therefore that the phase accumulation is less than π. In many of the basic states we have used, there are two neutral points. In most of these cases, the total phase accumulation of these neutral points is roughly π apart. For example, some runs have neutral points at $\frac{\pi}{2}$, $\frac{3\pi}{2}$, etc, and some are at 0.4π, 1.4π, etc. The phase accumulation increases with wavelength because the propagation region increases in size with wavelength. The difference in phase accumulation between the medium waves and the long waves is due to the addition of phase in the stratospheric propagation regions. The common feature of these runs is that these stratospheric propagation regions are simple in the sense that the basic state does not vary rapidly with height and internal partial reflections are small. As a result, the phase accumulation in the stratosphere is simple, making additional neutral points spaced π apart. Figure 4.9 shows the imaginary phase speed as a function of $\Delta phase$ for the long waves of our control run. We have strictly neutral points at $\Delta phase \approx 0.5\pi$ and 1.5π. Also, the wavenumbers with $\Delta phase \approx 0.5\pi - \pi$ are almost neutral. This is a feature that is common to quite a few of the runs. Not all runs have neutral wavenumbers. These are runs that have a complicated wave geometry that does not allow destructive interference to occur. For example, basic states with multiple turning points and/or regions where the basic state varies rapidly and we have partial reflection. The only conclusion we can draw from all these runs is that the mode can interfere with the growth at the critical level, and that basic states that vary with height do not have a simple relationship between the phase accumulation and the growth rate as we find in the Charney model.

This raises the issue of why long waves grow as slowly as they do. The long waves are expected to grow slower than the medium scale waves because the growth rate is proportional to the wavenumber, however, in our case, the imaginary phase speed itself is smaller. This is not obvious, based on simpler models. In the Eady model, for example, the long waves have the largest imaginary phase speed. In the Charney model, the longest waves grow much faster than in our case, even though there is a neutral point to separate them from the medium scale waves. In the Green model (a Charney model with a lid at the tropopause), on the other hand, the long waves grow slowly. The above results would suggest that partial reflections from above the critical level inhibit growth through interference. Most of the partial reflections occur in the tropopause region, where the basic state varies most rapidly. This suggests the tropopause is the reason for slow growth. Our model differs from the Charney model essentially in having a tropopause. In order to test this we ran a series of models

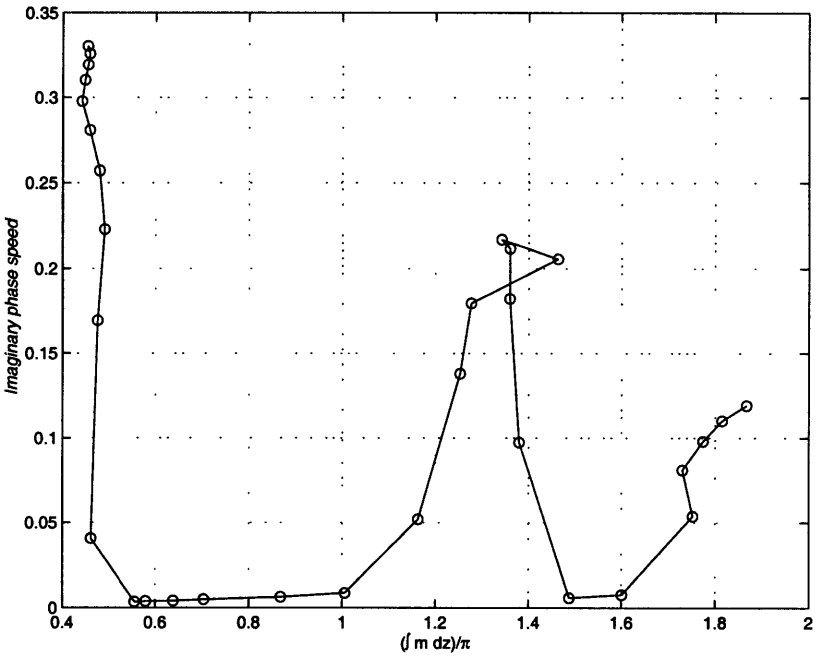

Figure 4.9: The imaginary phase speed (nondimensional units) as a function of the phase accumulation, $\Delta phase$ (in units of π), for the long waves shown in figure 4.4.

that vary gradually from a Charney model where the shear and N^2 are constant with height, to a model where shear and N^2 both vary rapidly at the tropopause. Our results suggest that the tropopause inhibits growth by partially reflecting waves downward, since most models with a sharp tropopause had slower growing long waves than the Charney model. There was, however, one basic state where the long waves actually grew faster when the tropopause was sharper. It is possible that partial reflections sometimes cause the growth rate to increase, instead of decrease.

4.4.3 The effect of surface damping

In section 4.4.2 (and figure 4.8) we saw that the effects of thermal damping that is parameterized as Newtonian damping are mostly to decrease amplitudes of waves in the damping region and to dramatically reduce reflection from turning points, causing the growth rate of long waves to be relatively constant (the neutral waves become unstable and the growth of unstable waves is slightly decreased).

In this section we discuss the sensitivity of our results to the inclusion of surface damping. This is important since the long waves grow slowly, and damping may interfere with this growth to make them irrelevant to the real atmosphere. There have been quite a few studies of the effect of Ekman damping on baroclinically unstable modes (e.g. Card and Barcilon, 1982, Lin and Pierrehumbert, 1988, Snyder and

Lindzen, 1988) however, most studies concentrated on the most unstable medium scale waves, for which they present the results, using quite a large meridional wavenumber.

Surface friction is parameterized as an Ekman boundary condition (see appendix B). Table 4.1 shows the growth rates (in day^{-1}), and the zonal and total wavenumbers of the fastest growing long waves, for various values of Ekman damping coefficient E_k. The nondimensional damping coefficients used are 0.1, 0.2, 0.5, corresponding to eddy viscosity coefficients of 11.1, 44.3, and 277.0 $\frac{m^2}{sec}$, and spin-down time scales of 2.0, 1.0, and 0.4 days (Lin and Pierrehumbert, 1988, considered eddy viscosity coefficients in the range $0 = 100\frac{m^2}{sec}$). For comparison, we also show the growth rates for no damping. Since the inclusion of damping breaks the symmetry between zonal and meridional wavenumbers, a meridional wavenumber has to be specified before solving the equations. The results are shown for three meridional wavenumbers (the equivalence in degrees latitude is also shown). We see that the damping is more efficient for larger meridional wavenumbers. When $l = 0$, a damping of $E_k = 0.1$ reduces the growth rate by 23%. For l=0.23, by 33%, and for l=0.35, by 67%. It is interesting, however, that even with a very large damping $E_k = 0.5$, the long waves are not eliminated. The effect of Ekman damping is not simply to reduce the growth rate by the inverse spin-down time. This is probably because the damping is applied at the surface, and not at the critical level, where the actual growth occurs[17]. For this reason, Ekman damping may not be the appropriate form of damping for our problem. The critical level of the fastest growing long waves lies between 1 and 6 kilometers, which may lie in the boundary layer. The effect of surface damping on baroclinically unstable modes is a still debated question which we will not attempt to answer in this study.

4.4.4 Sensitivity of the results

The results shown so far are for a single basic state. The wave geometry framework allows us to test the sensitivity to the basic state characteristics by identifying and varying the parameters that matter most. What we find is that varying the basic state in the stratosphere affects only waves that have wave propagation regions there (unless

[17]Snyder and Lindzen (1988) distinguished between upper level modes, whose wave geometry is such that there is propagation below the critical layer and evanescence above it, and lower level modes (the classic Charney modes, for example), which have propagation above the critical level and evanescence below. Upper level modes grow due to overreflection of the waves approaching the critical level from above. In these configurations, surface damping has a reduced effect, and may actually increase the growth rate. The reduced effect of Ekman damping (compared to the spin-down time) in our model is not due to the same cause because we have a lower level instability.

Meridional wavenumber l(°lat.)	Ekman damping coefficient	Largest growth rate	Zonal wavenumber k	Total wavenumber K
0.0 (∞)	0.0	0.17	0.32	0.32
0.0 (∞)	0.1	0.13	0.30	0.30
0.0 (∞)	0.2	0.11	0.30	0.30
0.0 (∞)	0.5	0.07	0.28	0.28
0.23 (90°)	0.0	0.12	0.22	0.32
0.23 (90°)	0.1	0.08	0.22	0.32
0.23 (90°)	0.2	0.06	0.22	0.32
0.23 (90°)	0.5	0.03	0.19	0.30
0.35 (60°)	0.0	0.03	0.08	0.36
0.35 (60°)	0.1	0.01	0.19	0.40
0.35 (60°)	0.2	0.01	0.19	0.40
0.35 (60°)	0.5	0.004	0.15	0.38

Table 4.1: Growth rates, zonal, and total wavenumbers of the fastest growing long waves, for various values of Ekman damping coefficients and meridional wavenumbers. Meridional wavenumber is given in nondimensional units as well as in degrees latitude of half a meridional wavelength. Growth rate is in day^{-1}. Zonal and total wavenumbers are in nondimensional units. Ekman damping coefficients are nondimensional, refer to the text for the corresponding dimensional values.

the changes cause more or less waves to propagate in the stratosphere). Changes in the tropopause region affect all wavenumbers because all waves propagate there. The effects depend on the wave geometry. The relation between the vertical structure and wave geometry found in the control run holds for all runs. For example, if we change the winds in the stratosphere in a manner that reduces the index of refraction, the separation between waves that are stratospheric and waves that decay in the stratosphere will shift to longer waves. The effect of wave geometry on the growth rate also holds, hence, the tropospheric basic state will affect the location of the neutral point between the Green and Charney modes, while the stratospheric basic state will affect the existence of neutral waves at longer wavenumbers. The relation between the neutral wavenumbers and the phase integral (equation 4.14), however, is hard to generalize, because of the internal scattering of the waves in the tropopause region. It is important to note that there are two significant wavenumbers, the neutral wavenumber and the wavenumber that separates between the modes that propagate in the stratosphere and the modes that don't. These two wavenumbers coincide in

our control run, but not in general. If we take, for example, the same tropospheric basic state as the control run, and change the winds to have a minimum above the tropopause (as in fall and spring of both hemispheres), the location of the neutral point may not move much. At the same time, some of the waves on the short wave side of the neutral point may have a propagation region in the stratosphere (in the region of minimum winds). Such modes have relatively large amplitudes in the stratosphere (as large as in the troposphere but not much larger) along with relatively large growth rates (compared to the long stratospheric modes). The basic states that allow these modes to exist, however, are not very realistic for winter, and they occur for mid-channel latitudes that are a bit too small for stratospheric planetary waves (e.g. 45°).

Finally, in order to make sure none of the above results depend too strongly on the vertical scale height chosen, we repeated the calculations for different values. We do this because we used a scale height of $H = 8.9km$, while $7km$ is a more characteristic value for the stratosphere. As expected, changing the scale height mostly affects the vertical structure of the waves through the $e^{\frac{z}{2H}}$ factor.

4.5 Internal stratospheric instability

In this section we will show one example of a mode that is drawing energy from a critical level in the stratosphere. We constructed the basic state to have a mid-stratospheric minimum in wind of 10m/sec at 30km, to assure critical levels. Also, we specified the PV gradient to be negative around the minimum in zonal wind (figure 4.10). This specific run has zero PV gradients in the troposphere. The dispersion relation (not shown) is similar to the dispersion relation of runs with no critical levels in the stratosphere, meaning that the tropospheric instability is dominant in setting the growth rates and phase speed of the modes. Also shown in figure 4.10 are the results for wavenumber $K = 0.6$ that has a phase speed of 15m/sec and a growth time of around 7.5 days. We see from the real part of n_{ref} that the wave geometry configuration is like figure 4.2f, where there is a region of negative $\overline{q}_y$ in between, but not touching, two critical levels. As a result, the $\overline{q}_y < 0$ region is a wave propagation region that is bounded by turning points on both sides, with critical levels beyond, such that waves can overreflect from both sides. This is evident from the meridional PV flux (also shown in figure 4.10) which is positive in the region of negative $\overline{q}_y$, and negative elsewhere. There are two peaks of negative PV flux at the stratospheric critical levels, with the lower one being larger. Such peaks in PV flux are a clear indication of an interaction with the mean flow at the critical level. They are not found for long waves that do not have a critical level in the stratosphere. Note that

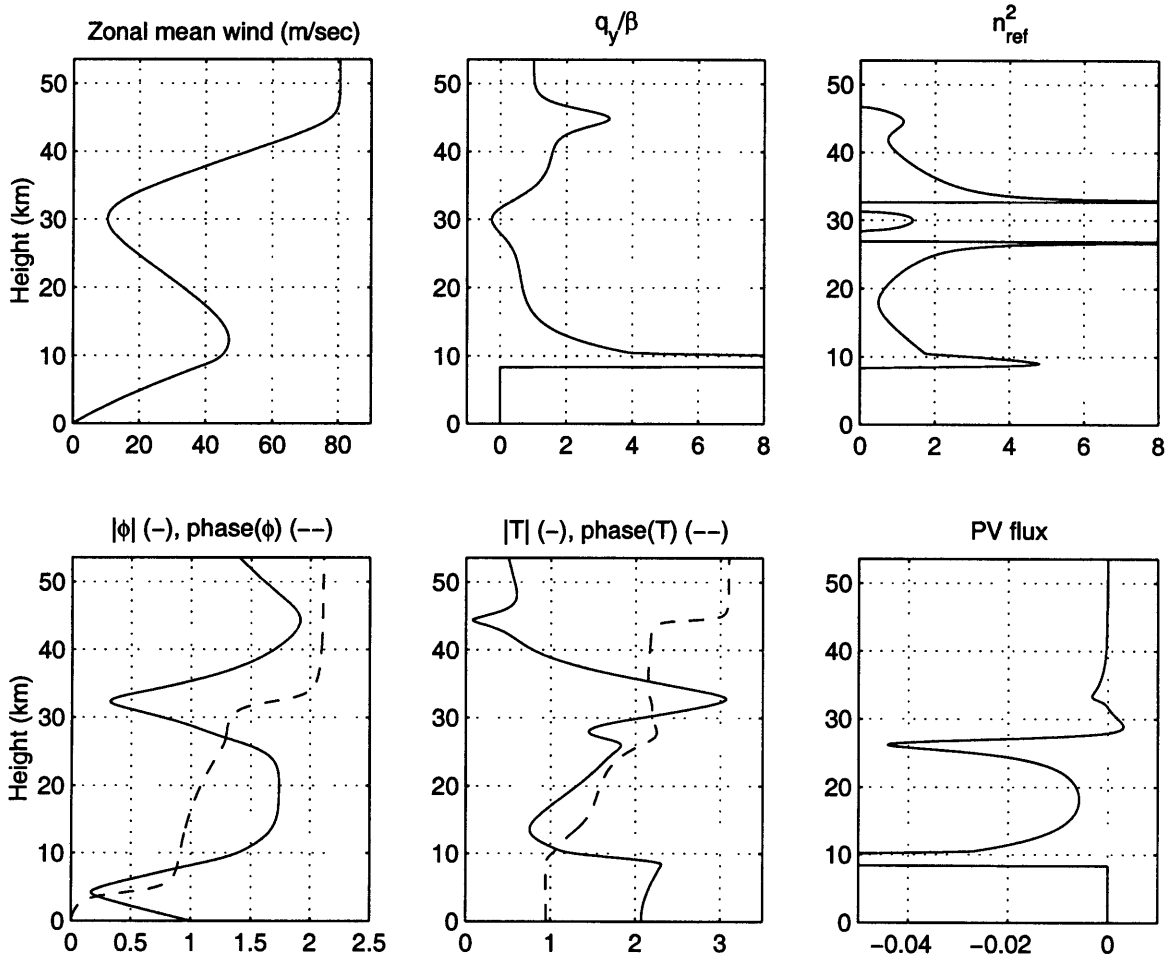

Figure 4.10: Results for a run with stratospheric critical levels and a $\overline{q}_y < 0$ region, for wavenumber $K = 0.6$. Top, from left to right: Zonal mean wind (m/sec), PV gradient (units of β), and the real part of the index of refraction. Bottom: Geopotential height (left) and temperature (middle) amplitude (-) and phase (–), and PV fluxes (right). See text for details.

the PV fluxes in the troposphere are zero, which is due to $\overline{q}_y = 0$ there. The above results hold for runs with non-zero tropospheric $\overline{q}_y$ as well. Also shown in figure 4.10 are the geopotential height and temperature structure of the mode (both amplitude and phase). The vertical structure varies quite a lot with wavenumber. There are however a few robust features. There is a westward phase tilt with height of ϕ at the higher critical level (where the shear is positive) and essentially no phase tilt at the lower critical level and in the negative $\overline{q}_y$ region. The temperature phase tilts slightly eastward (this is similar to the tropospheric instability). The temperature structure (and even the geopotential height field) has relatively small scale features which may pose a problem in retrieving these waves (see sections 3.4.3, 3.5).

Finally, it is interesting to note that Manney et. al (1991a) observed occasional episodes of growth which are confined to the stratosphere and have slightly different characteristics than the more typical growth episodes (e.g. weak equatorward heat fluxes). The type of mode we get here could be a possible explanation for these observations but more study is needed to test this.

4.6 Discussion

In this chapter we studied the normal modes of a troposphere-stratosphere system with basic states that vary with height only, in the framework of wave geometry. We used tropospheric basic states that support baroclinic instability, and identified different kinds of stratospheric wave geometry configurations that support different kinds of modes. The first kind evolve on basic states that have no critical levels or regions of negative PV gradients in the stratosphere. These are tropospherically unstable modes that propagate up to the stratosphere, and are commonly referred to in the literature as Green modes. The second kind of modes evolve on basic states that have both a region of negative PV gradients and a critical level in the stratosphere. This allows the modes to draw energy from the mean flow at the stratospheric critical level. There are various configurations of critical levels and $\overline{q}_y < 0$ regions, that result in different structures of internal stratospheric modes. The existence of regions of negative PV gradients are crucial, since without them the modes just get absorbed at the critical level.

We concentrated on the relation between the normal mode structure and phase speed (real and imaginary) and the basic state wave geometry for the first kind of modes. We studied the second kind of modes only on basic states that support tropospheric instability as well. We found that the tropospheric instability is dominant in the sense that it determines the phase speed, while the stratospheric part of the mode, which now has a critical level in the stratosphere, draws energy from the mean flow there. We leave looking for pure stratospheric instability both in models and observations for future study.

The largest effect the stratosphere has in model runs with simpler basic states (i.e. positive stratospheric $\overline{q}_y$), is obviously on the mode structure there, through determining the vertical propagation characteristics of the waves. The picture is slightly more complicated because the index of refraction depends on the phase speed, which is determined mostly in the troposphere. We saw that to some extent, the phase speed also depends on the stratosphere. When a wave has turning points in the stratosphere, it reflects downward, and interferes with the upward propagating wave.

This affects the interaction of the wave with the critical layer in the troposphere, which affects both the growth rate and the real phase speed.

We can also apply wave geometry reasoning to the time development of the mode. We expect the following picture: the unstable tropospheric mode develops first, with the stratosphere acting as a wave energy sink. In the stratosphere, structure will be controlled by wave propagation, determined by the *local* basic state properties, and by the phase speed of the perturbation. At later stages the mode will reflect down and adjust the phase speed, affecting the stratospheric structure, reflecting back, and so on until a normal mode is fully developed. In the initial stages of mode development, the wave will not see the reflecting surface and we expect the growth rates to be quite constant for the different wavenumbers. Only later (if at all) we expect the stratosphere to affect growth rates[18]. This is a mechanism by which neutral structures can develop, and it is consistent with findings that the neutral modes actually grow linearly in the initial value problem (e.g. Burger, 1966, Farrell, 1982).

We actually do not expect to see a fully developed *linear* neutral wave for several reasons. The first is that it may take the mode a long time to develop, and the basic state may change in the meantime, affecting the wave geometry. Second, the turning point that reflects the wave down is in reality a turning surface, and this surface will most likely not have a simple geometric form. Under these conditions, it is hard to see how we can get the interference at the critical level that is inherent to the neutral modes. Finally, we expect damping to reduce reflections from turning surfaces. We should not, however, dismiss downward reflection. As we will show in the later chapters (mostly chapter 7), downward reflections are observed, and do play a role in the transient evolution of waves. Neutral modes are not entirely irrelevant to the atmosphere, because we may have *nonlinear* equilibration, where nonlinear dissipation of energy balances the linear growth. In such cases we may still expect the wave geometry in the stratosphere to affect the wave structure there, while nonlinearities modify the structure qualitatively like damping. In chapter 7 we show some examples where nonlinearities affect the mode like damping, while the overall behavior is still qualitatively linear.

Finally, even though our results are obtained for the eigenvalue problem, the effect of the basic state wave geometry on the vertical structure of the normal modes is relevant to the forced problem as well. This is important, since we can use our

[18]We will show in chapters 6 and 7 that it takes an order a few days for waves to propagate to stratospheric turning points.

results to test the extent to which linear QG wave theory explains the structure and time evolution of observed planetary waves at a given time or season. The sensitivity of linear QG wave models to details of the basic state and model damping, both of which are not determined from observations in great accuracy, makes it hard to determine why the observations deviate from modeled waves in any given case. By understanding the effects of the wave propagation geometry and damping on the vertical structure of geopotential height and temperature of the waves in a model, we can relate specific large scale (and easily observed) features of vertical structure to specific features of the basic state, and see to what extent they are consistent. Before we do this, however, we need to see how these relations hold when the basic state depends on latitude, and the waves can propagate meridionally. This is the focus of the next two chapters.

Chapter 5

The dependence of stratospheric wave structure on the latitude-height wave geometry of the basic state

5.1 Introduction

In this chapter, we would like to relate the latitude-height wave structure to the propagation characteristics of the basic state (index of refraction). In general, given an index of refraction, there is no unique way to separate wave propagation in the meridional and vertical directions a priori, without obtaining the full solution first. The existence of a waveguide in the stratosphere simplifies matters, because it determines the structure of the perturbations in the across-waveguide direction. This allows us to obtain a Charney-Drazin type criterion for wave propagation in the vertical. This is trivial for a constant-width channel model with a separable basic state and is less obvious for more general conditions. In particular, it is not obvious that this formulation applies at all, because the 'wave propagation - index of refraction' picture is appropriate only when the amplitude of the wave varies on scales larger than the wavelength, which is not obviously the case everywhere in the stratosphere. We discuss this quantitatively in sections 5.2.1 and 5.3.5.

We start by formulating the wave propagation problem in a way that will allow us to diagnose the propagation characteristic of a given wave solution, using an extension of the Charney-Drazin criterion to two dimensions (section 5.2). In section 5.3 we diagnose the propagation characteristics of waves in simple QG β-plane model, and

discuss its effect on vertical wave structure. We then use the diagnostic to explain some of the observed features of waves (section 5.4).

5.2 Formulation of the Charney-Drazin criterion in two dimensions

Throughout our discussion we use the quasi-geostrophic approximation, to study linear wave propagation. We start by using a β-plane and we will extend our analysis to spherical coordinates later on.

The equations we use are the same as 4.1 - 4.3, where z is log-pressure, nondimensionalized by a reference scale height (H_o), and ϕ is the geopotential height (see appendix B).

Assuming a normal mode structure:

$$\phi(y,z) = \varphi(y,z) \cdot e^{i \cdot k(x-ct)} \tag{5.1}$$

we get the following equation for φ:

$$e^z \frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\frac{\partial \varphi}{\partial z}\right) + \frac{\partial^2 \varphi}{\partial y^2} + \left(\frac{\overline{q}_y}{U-c} - k^2\right)\varphi = \frac{i}{k(U-c)}\left[\frac{\partial}{\partial z}\left(\frac{e^{-z}\alpha}{N^2}\frac{\partial \varphi}{\partial z}\right) + r\nabla^2\varphi + \frac{\partial r}{\partial y}\frac{\partial \varphi}{\partial y}\right] \tag{5.2}$$

We have used Newtonian damping of the form $\mathcal{H} = -\alpha \frac{\partial \varphi}{\partial z}$, and Rayleigh friction of the form $(\nabla \times \mathcal{F}) \cdot \hat{\mathbf{k}} = -r\nabla^2\varphi - \frac{\partial r}{\partial y}\frac{\partial \varphi}{\partial y}$

On a spatially varying medium, we can formulate the problem in terms of wave propagation/index of refraction only by assuming the basic state varies on scales larger than the wavelength. We use the WKB approximation to write down a wave solution with an amplitude and wavenumber varying in space on scales that are much larger than the wavelength itself (see 5.2.1). In one dimension, the first order WKB solution is given by equation 4.12. The appropriate equivalent relation in two dimensions is not readily obtained, but we may start by assuming a solution of the form (using the transformation 4.9):

$$\phi = \left[\left(A_1 e^{i\int l dy} + A_2 e^{-i\int l dy}\right) e^{i\int m dz} + \left(B_1 e^{i\int l dy} + B_2 e^{-i\int l dy}\right) e^{-i\int m dz}\right] N e^{\frac{z}{2}} e^{ik(x-ct)} \tag{5.3}$$

where l and m can either be real or imaginary (in which case we have an exponential rather than a wave behavior). As in one dimension, the amplitude functions A_1, A_2,

B_1, B_2, which are functions of y and z, have to satisfy conservation of wave activity, while the meridional and vertical wavenumbers (l and m) satisfy a dispersion relation (see equations 5.10 and 5.11). These conditions, however, do not determine the amplitude coefficients and the wavenumbers uniquely. Our approach is therefore to solve the equations, and to *diagnose* the wavenumbers. We will show later that this approach is useful, since our goal is to determine the wave geometry of the basic state.

We divide equation 5.2 by $\frac{\varphi}{N^2(U-c)}$, equate the real parts, and divide by $U - c_r$ to get:

$$Re\left(\frac{\frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\varphi_z\right)}{\frac{e^{-z}}{N^2}\varphi}\right) + N^2 Re\left(\frac{\varphi_{yy}}{\varphi}\right) + N^2\frac{\overline{q}_y}{U-c_r} - k^2N^2 + \frac{\alpha_z/k}{U-c_r}Im\left(\frac{\varphi_z}{\varphi}\right) +$$
$$\frac{N^2 r_y/k}{U-c_r}Im\left(\frac{\varphi_y}{\varphi}\right) + \frac{c_i+\frac{\alpha}{k}}{U-c_r}Im\left(\frac{\frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\varphi_z\right)}{\frac{e^{-z}}{N^2}\varphi}\right) + N^2\frac{c_i+\frac{r}{k}}{U-c_r}Im\left(\frac{\varphi_{yy}}{\varphi}\right) = 0 \quad (5.4)$$

We now note that under WKB conditions the wavelength and amplitude vary on scales that are much larger than a wavelength, allowing us, for example, to neglect l_y, A_{1y}, A_{2y}, B_{1y}, and B_{2y} relative to l^2. As a result, if we plug a solution of the form 5.3 into equation 5.4, we can relate the first two terms to the vertical (m) and meridional (l) wavenumbers as follows:

$$Re\left(\frac{\psi_{zz}}{\psi}\right) = Re\left(\frac{\frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\varphi_z\right)}{\frac{e^{-z}}{N^2}\varphi}\right) - N^2F(N^2) = -m^2 \quad (5.5)$$

$$Re\left(\frac{\varphi_{yy}}{\varphi}\right) = -l^2 \quad (5.6)$$

where $F(N^2)$ is defined by 4.10, and ψ by equation 4.9. Note that in general l^2 and m^2 are not pure real numbers, and under the conditions assumed for 5.5 and 5.6:

$$Im\left(\frac{\psi_{zz}}{\psi}\right) = Im\left(\frac{\varphi_{zz}}{\varphi}\right) = -Im(m^2) \quad (5.7)$$

$$Im\left(\frac{\varphi_{yy}}{\varphi}\right) = -Im(l^2) \quad (5.8)$$

however, when we have no damping (or very small damping), these terms are small. This can be shown as follows. We transform equation 5.2 to ψ (using 4.9), divide it by $\frac{\psi}{N^2(U-c)}$, equate the *imaginary* parts, and divide by $U - c_r$, to get:

$$Im\left(\frac{\psi_{zz}}{\psi}\right)+N^2Im\left(\frac{\psi_{yy}}{\psi}\right)+\frac{\alpha_z/k}{U-c_r}\left[\frac{1}{2}+\frac{N_z}{N}+Re\left(\frac{\psi_z}{\psi}\right)\right]-N^2F(N^2)\frac{c_i+\frac{\alpha}{k}}{U-c_r}+$$
$$k^2N^2\frac{c_i+\frac{\alpha}{k}}{U-c_r}-\frac{N^2r_y/k}{U-c_r}Re\left(\frac{\psi_y}{\psi}\right)-\frac{c_i+\frac{\alpha}{k}}{U-c_r}Re\left(\frac{\psi_{zz}}{\psi}\right)-N^2\frac{c_i+\frac{r}{k}}{U-c_r}Re\left(\frac{\psi_{yy}}{\psi}\right)=0 \quad (5.9)$$

It is clear from equation 5.9 that if the damping and imaginary phase speed are zero (the two are related in the forced problem since the only source of an imaginary phase speed is damping), the terms $Im\left(\frac{\psi_{zz}}{\psi}\right)$ and $Im\left(\frac{\psi_{yy}}{\psi}\right)$ are zero.

Note that for WKB to hold, we need damping to be small, in which case we may neglect the last four terms on the left hand side of equation 5.4, which leaves us with the following relation:

$$m^2+N^2l^2=N^2\left(\frac{\overline{q}_y}{U-c_r}-k^2+F(N^2)\right)\equiv n^2_{ref} \quad (5.10)$$

n^2_{ref} is the index of refraction of the one dimensional case (4.11).

It is important to point out that if we define m and l as in relations 5.5-5.6, equation 5.10 is still exactly satisfied in the undamped case even if WKB conditions are violated.

When we do have damping, we need to add the damping terms from 5.4 to the dispersion relation 5.10:

$$\frac{m^2}{N^2}+l^2=\frac{\overline{q}_y}{U-c_r}-k^2+F(N^2)+\frac{\alpha_z/k}{N^2(U-c_r)}Im\left(\frac{\varphi_z}{\varphi}\right)+$$
$$\frac{r_y/k}{(U-c_r)}Im\left(\frac{\varphi_y}{\varphi}\right)+\frac{c_i+\frac{\alpha}{k}}{N^2(U-c_r)}Im\left(\frac{\frac{\partial}{\partial z}\left(\frac{e^{-z}}{N^2}\varphi_z\right)}{\frac{e^{-z}}{N^2}\varphi}\right)+\frac{c_i+\frac{r}{k}}{U-c_r}Im\left(\frac{\varphi_{yy}}{\varphi}\right) \quad (5.11)$$

We see that when we have large damping, l and m can change. The terms $Im\left(\frac{\varphi_y}{\varphi}\right)$ and $Im\left(\frac{\varphi_z}{\varphi}\right)$ are proportional to the meridional and vertical components of the Eliassen-Palm flux:

$$F_y=\rho\overline{u'v'}=\rho\frac{k}{2}Im(\varphi_y\varphi)=\rho\frac{k}{2}|\varphi|^2Im\left(\frac{\varphi_y}{\varphi}\right) \quad (5.12)$$

$$F_Z=\frac{\rho}{N^2}\overline{v'T'}=\frac{k\rho}{2N^2}Im(\varphi_z\varphi)=\frac{k\rho}{2N^2}|\varphi|^2Im\left(\frac{\varphi_z}{\varphi}\right) \quad (5.13)$$

Under WKB conditions (section 5.2.1), perturbations of the form $f(y,z)e^{i\int ldy}+$

$g(y,z)e^{-i\int ldy}$ will satisfy:

$$Im\left(\frac{\varphi_y}{\varphi}\right) = Re(l)\frac{|f|^2e^{-2\int Im(l)dy} - |g|^2e^{2\int Im(l)dy}}{|fe^{i\int ldy} + ge^{-i\int ldy}|^2} + Im(l)\frac{fg^*Im(e^{2i\int Re(l)dy})}{2(|fe^{i\int ldy} + ge^{-i\int ldy}|^2)} \tag{5.14}$$

where $f(y,z)$ is the amplitude of the wave propagating in the positive direction (equatorwards in our model) and $g(y,z)$ of the oppositely propagating wave. We are free to assume l is real, by putting the terms $e^{\pm\int Im(l)dy}$ into the coefficients f, g, which leaves us with:

$$Im\left(\frac{\varphi_y}{\varphi}\right) = l\frac{|f|^2 - |g|^2}{|fe^{i\int ldy} + ge^{-i\int ldy}|^2} \tag{5.15}$$

and we see that if the poleward and equatorward perturbations are of equal magnitudes, there will be no EP flux. This makes sense since an EP flux implies net propagation of wave activity and a standing wave has no net propagation. We will have a non-zero EP flux, only if the waves have a sink of wave energy.

We can get a sense of one of the effects of damping on the dispersion relation by looking at the EP flux terms of equation 5.11. When the EP flux is positive, and the Newtonian damping coefficient increases(decreases) with height, the vertical wavenumber increases(decreases). In this case damping increases(decreases) vertical propagation in the sense that it allows larger(smaller) zonal wavenumbers to propagate upwards.

Equation 5.10 (or 5.11) is the dispersion relation in the two dimensional case, and will serve as the basis of our diagnostics in the following sections. Before we apply it, however, we will discuss the conditions under which the wave propagation-wave geometry concept and the WKB assumptions are valid.

5.2.1 Conditions for the WKB approximation to hold

In order for equations 5.5 and 5.6 to hold, with l and m being the wavenumbers of a solution of the form 5.3, we need to assume that the wavelength of the solution is much smaller than the length over which the amplitude of the wave and the wavenumber itself vary, allowing us to neglect l_y, A_{1y}, A_{2y}, B_{1y}, and B_{2y} relative to l^2. These are the assumptions we make for WKB to hold.

In one dimension, we can express the amplitude of the wave in terms of the wavenumber ($A \propto m^{-1/2}$, equation 4.12), hence the above conditions can be stated

in terms of the wavenumber only:

$$\frac{d}{dz}\left(\frac{1}{m}\right) \ll 1 \tag{5.16}$$

The extension to two dimensions is not as simple, since we cannot express the amplitude in terms of the wavenumber explicitly. The best we can do is to obtain a set of necessary conditions by applying 5.16 along with a similar condition on the meridional direction:

$$\frac{d}{dy}\left(\frac{1}{l}\right) \ll 1 \tag{5.17}$$

Since we had to assume WKB in our definition of l and m, we can only test the consistency of our approximations a posteriori[1]. Note however, that even if WKB conditions do not hold, equation 5.10 is still exactly satisfied in the undamped case. Also, the solution may still be of wavelike nature, as long as the $\ll$ in the inequalities 5.16- 5.17 is replaced by $<$. In this case the WKB form of the solution will not be correct, but the qualitative wave features of the solution will still hold. It is interesting to see in how much of the domain the solution violates WKB but is still wavelike. If the left hand sides of 5.16- 5.17 are greater than 1.0, the interpretation of the solution in terms of wave structure and wave propagation is ambiguous. In section 5.3.5 we calculate the size of the left hand terms of the inequalities 5.16- 5.17 for our model run to see where WKB is violated.

5.3 Demonstration on a β-plane model

In the following section we will use a model to obtain a steady state wave solution to a given basic state (solve equation 5.2), and use it to demonstrate the meaning of the vertical and meridional wavenumbers defined in equations 5.5- 5.6.

5.3.1 The model

Our model is quasi-geostrophic, linear and on a β-plane. We specify a basic state wind U to be a function of latitude and height, and a basic state temperature that varies only with height. We have a sponge layer at the top, to approximate a radiation condition, and a sponge layer at low latitudes to include the effect of either absorption at a critical surface or radiation through the tropics. The sponge layers are

[1]Karoly and Hoskins (1982) came up with approximate conditions that are based only on the basic state and do not require solving for the wave in order to test WKB validity, however, since we obtain a wave solution, it is easier to test the validity directly.

a combination of Newtonian cooling and Rayleigh damping with equal coefficients, which we specify such that waves are absorbed before they reach the model boundaries (see appendix B for details). In some runs we raise the sponge layer and add a more realistic Newtonian cooling, that follows the temperature profile, as suggested by Dickinson (1969b). We also add a small constant damping to model runs that have a critical level, to insure numerical convergence. At the poleward boundary we set the perturbation to be zero. Such a boundary condition will reflect the waves equatorwards. However, in the real atmosphere and in our model, there is always a turning point at high latitudes which reflects the waves equatorwards before they reach the pole. Since a channel model can't really get the perturbations at the pole correctly (a polar coordinate model is needed for that) this is the best we can do. We use a mid-channel latitude of 55°, which means the radius of deformation (the horizontal length scale nondimensionalization constant) is: $L_d = \frac{NH_o}{f_o} = 1190Km$ which is 10.7 degrees latitude. See appendix B for more details on the model.

We force the model by specifying the zonal wavenumber and phase speed, and the latitudinal structure of the amplitude and phase of the forcing at the bottom (which is at 2 scale heights, 14km). The forcing is constant with time, apart for a zonal propagation with the prescribed phase speed. The latitudinal variation of the phase of the forcing at the bottom determines the latitudinal direction and magnitude of the EP fluxes there. The model is computationally cheap, and we can run it many times to test the sensitivity to various parameters. The results we will show, unless otherwise stated, are general. The overall picture we get is that the stratospheric jet acts to guide wave activity from the troposphere up, along its maximum. This waveguide however is leaky, with most of the leakage to the equator where we have a sponge layer. This picture of a leaky waveguide has been suggested in the past (e.g. by Dickinson, 1968 and Matsuno, 1970), however, the consequences of such a configuration have not been demonstrated in much detail before. We will show that the consequence of having a waveguide aligned with the jet axis is to set the meridional wavenumber of the perturbation. This will in turn determine the vertical propagation characteristics of the waves, which will determine their vertical structure.

Figure 5.1 shows the basic state for our model run, which we will refer to as our control run in this chapter. The basic state wind is specified analytically. It tilts equatorwards and widens with height, which is characteristic of the early southern hemisphere winter jet. The maximum winds are around 100 m/sec, which is realistic for the early Austral winter. The PV gradient field is also qualitatively like observed. It has a ridge that follows the jet with a maximum of about $5.9\beta = 7.7{\cdot}10^{-11}(sec{\cdot}m)^{-1}$, and negative regions on both sides of the jet. In observations, the maximum we

observe is around $6 \cdot 10^{-11}(sec \cdot m)^{-1}$. We almost always see the poleward negative region and not always the equatorward negative region. Some of these differences are due to the use of a β-plane. It is important to remember that the observations of PV gradients are not very accurate because of the coarse vertical resolution. Also shown in figure 5.1 are the basic state temperature and Brunt-Vaisala frequency that we will use in all runs shown. The temperature is specified analytically to look like a standard midlatitude winter profile. We use both Newtonian damping and Rayleigh friction, which increase from zero at the bottom high latitudes (small y) to a a value of $3day^{-1}$ at the top and equatorial boundaries, with most of the increase above 42 km (z=6) and equatorwards of latitude (y=6.5).

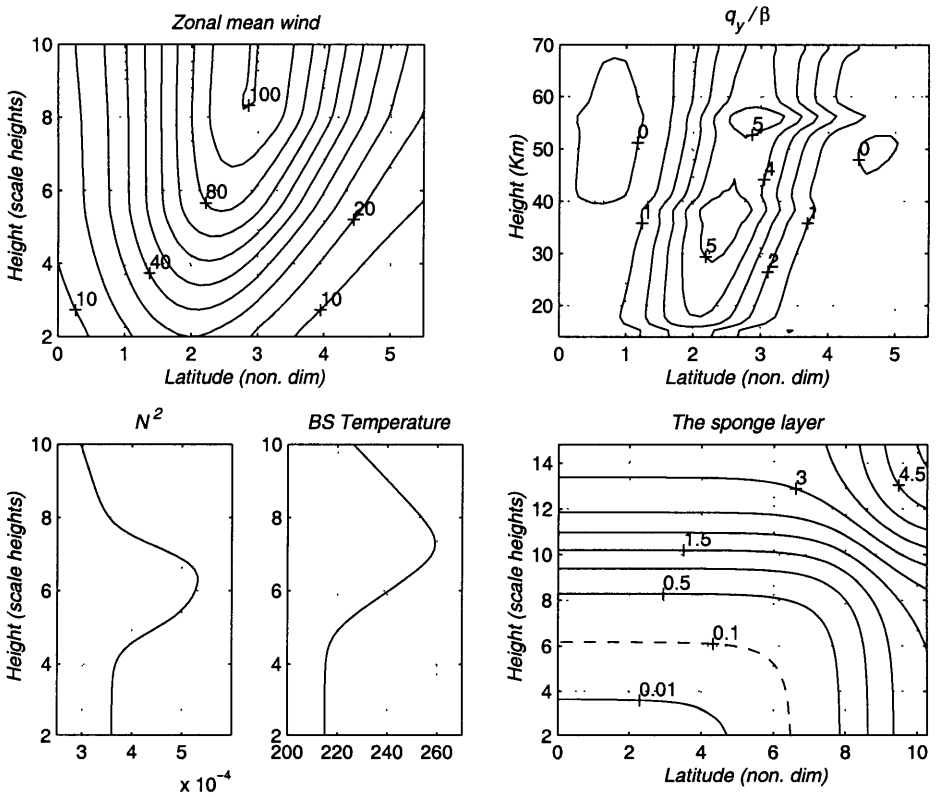

Figure 5.1: The control run- and early southern hemisphere winter idealized run. Top: Left- Basic state zonal mean wind (m/sec). Right- Meridional PV gradients, in units of $\beta = 1.3 \cdot 10^{-11}(sec \cdot m)^{-1}$. Bottom: Left- Basic state N^2 in sec^{-2}. Middle- Basic state temperature (°K). Right- The damping coefficient used (day^{-1}) for both Newtonian cooling and Rayleigh damping. For reference, the vertical coordinate of the PV gradient plot is in kilometers while all the rest are in scale heights. The latitude is in units of radii deformation ($L_d = 1190km$).

5.3.2 Robustness of the meridional wavenumber in a waveguide

Figure 5.2 shows the wave one and two geopotential height perturbations for a stationary forcing that is centered on the waveguide (also shown). The geopotential height perturbation has a peak that is located on the jet axis. The height of the maximum is determined by the damping of the sponge layer (which, in our model stratosphere, is much smaller than available estimates of damping, e.g. Dickinson, 1969b). The wave one geopotential height peak is 21 times as large as the forcing at the bottom. The largest waves we have observed have maximum geopotential height amplitudes of about 2.5km, with an amplitude of about 400m at 150mb (lowest observation level above 2 scale heights). The amplification we get is therefore too large (we will see later that some of the amplification is due to our model being on a β-plane rather than on a sphere). The corresponding index of refraction (equation 5.10) has a ridge aligned vertically along the jet, bounded to the north and south by regions of negative index of refraction. As expected, wave one has a larger index of refraction than wave two. The index of refraction in the middle of the waveguide (i.e. the value at the ridge) generally decreases with height. The PV gradient increases with height up to about 30km, then it is relatively constant, while N^2 is constant up to about 30km, then it increases. The PV gradients and N^2 act to increase the index of refraction (it increases with increasing PV gradient, and is roughly proportional to N^2, apart from the wiggles due to $F(N^2)$). The decrease of the index of refraction with height in this case, is therefore due to the winds increasing. This is as suggested by Charney and Drazin (1961), however the magnitude of the winds needed for trapping waves will be different because they did not include the effects of meridional curvature in their analysis.

Figure 5.3 shows the vertical and meridional wavenumbers (as defined by equations 5.5- 5.6) for waves one and two. Shown are positive values only, denoting the regions of meridional and vertical propagation, respectively. We see that there is meridional propagation in a relatively narrow region that is aligned with the PV gradient ridge, and in the equatorial region, where the index of refraction is very large due to the small winds. In chapter 6 we will show that the waves leak from the midlatitude wave propagation region to the equatorial one. The width of the midlatitude region is much smaller than the jet width. We see that the meridional wavenumber is similar for waves one and two, and the differences in index of refraction are manifest in the vertical wavenumber only. This is a consequence of having a waveguide that is oriented more or less vertically. Note that the meridional wavenumber, which is

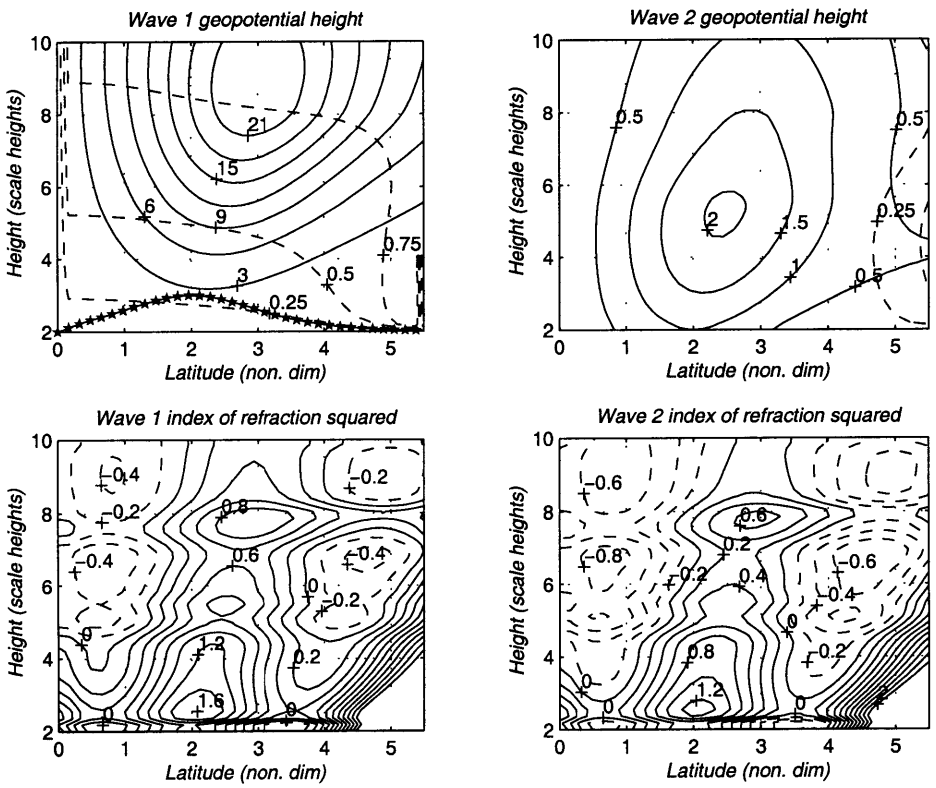

Figure 5.2: Top: Wave one (left) and two (right) stationary geopotential height amplitude (solid, arbitrary units) and phase (dashed, in units of π), for the basic state of figure 5.1. Also shown on the left is the amplitude of forcing at the bottom (thick stars). The magnitude is zero at the sides and one in the middle and there is no phase variation with y. Bottom: The index of refraction (equation 5.10), for wave one (left) and two (right), in nondimensional units. Negative values are dashed.

roughly 1.0 in nondimensional units, is much larger than any wavenumber considered in chapter 4 (e.g. table 4.1). According to those results even wave one should have been evanescent in the vertical. This shows the importance of the meridional curvature, which allows larger total wavenumbers to propagate vertically by increasing the PV gradients. An examination of the vertical wavenumber reveals that wavenumber two is evanescent in the vertical roughly above 5 scale heights, where the index of refraction becomes smaller than l^2N^2. This causes its amplitude to be much smaller than wave one, in accordance with the Charney-Drazin criterion. The $m^2 = 0$ line is referred to as the *turning point*, the *turning surface*, or the *reflecting (reflection) surface*. All mean the same and are chosen randomly. It is interesting that wave one also has turning points. It is evanescent between 6 and 7 scale heights and above 8.5 scale heights. According to our one dimensional runs, this implies downward reflection.

We also test the dependence of the meridional and vertical wavenumbers on other

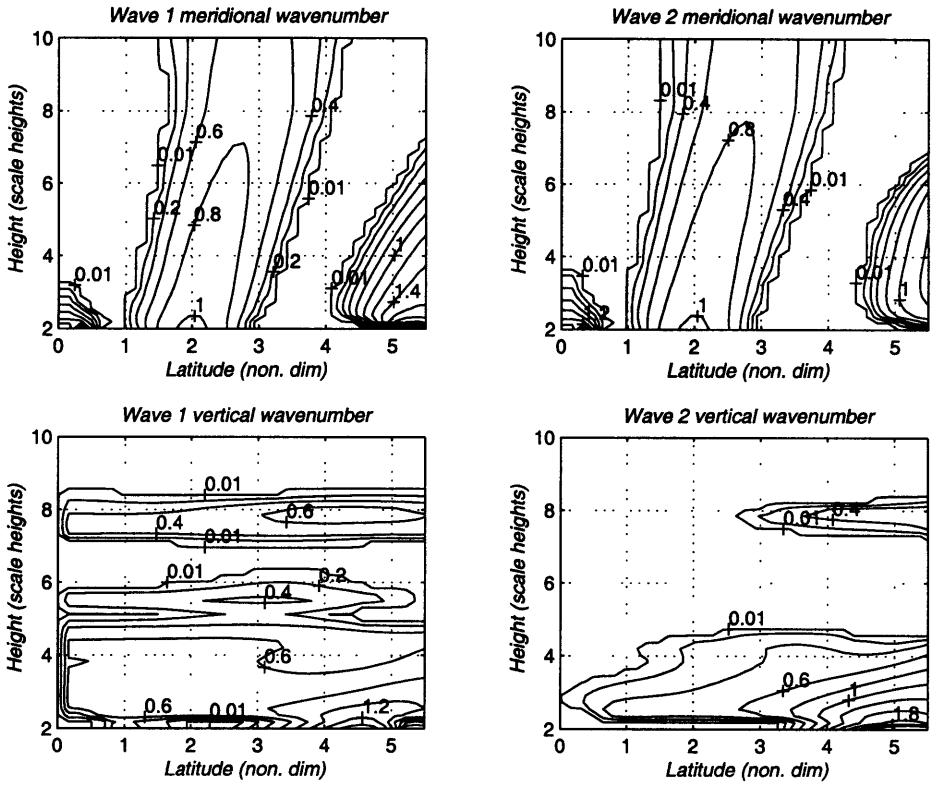

Figure 5.3: Meridional (top) and vertical (bottom) wavenumbers (as defined by equations 5.6 and 5.5) for the stationary wave one (left) and two (right) perturbations of figure 5.2. Only propagation regions are contoured, in nondimensional units.

parameters of the forcing besides zonal wavenumber. Figure 5.4 shows the wave one stationary response to a forcing that is constant in latitude (with an amplitude of 1.0). The geopotential height has a similar amplitude shape as the control run, but the amplitude is much larger, since the total amount of wave activity injected into the waveguide is larger. The phase increase with height at the bottom scale height is stronger than in the control run[2]. Above 3 scale heights, the meridional wavenumbers (and the vertical wavenumbers) are similar. Also shown in figure 5.4 is the response to forcing from a point source, located at the middle of the waveguide. The meridional wavenumber, in this case too, is similar to the control run above 3 scale heights. Since the model is linear, the geopotential height perturbation of a general forcing is a superposition of the response to point sources at the bottom. Apart from being much smaller, the geopotential height perturbation is very similar

[2]A stronger phase increase with height reflects a larger vertical group velocity at the bottom, which can be explained by the increase in the vertical wavenumber at the bottom. This increase has to happen, according to equation 5.10, because the constant forcing imposes a zero meridional wavenumber at the bottom.

to the control run. We also find that the meridional wavenumber of the response to forcing with a different phase structure at the bottom (a non-vertical EP flux at the bottom) is similar to the control run above 3-4 scale heights[3].

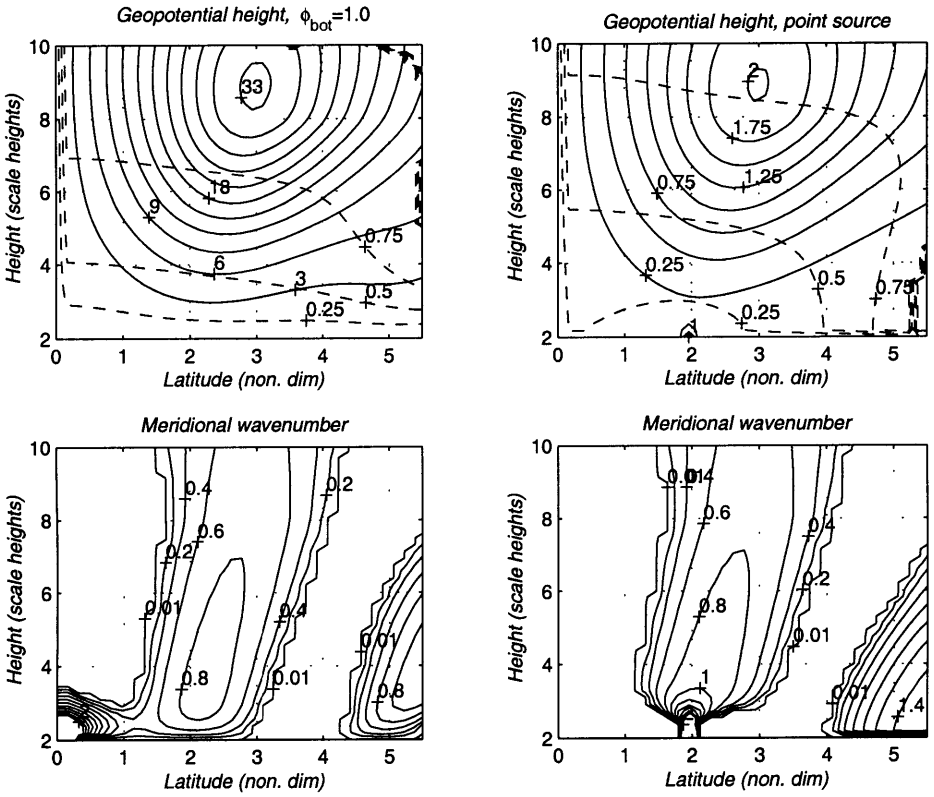

Figure 5.4: Top: Wave 1 stationary geopotential height amplitude (solid), and phase (dashed), for a constant forcing at the bottom of $\varphi = 1.0$ (left), and for a point source at the middle of the waveguide of magnitude 1.0 (right). Amplitude is in arbitrary units, phase is in units of π. The region where many phase lines are bunched together is a jump in phase of 2π, which is no phase shift at all (it is an artifact of the plotting routine). Bottom: The corresponding meridional wavenumbers (constant forcing on left and point source on right), in nondimensional units. Only propagation regions are contoured. The basic state is of figure 5.1.

We also find the meridional wavenumber to be insensitive to zonal phase speed, at least in the middle of the waveguide. Differences between zero and non-zero phase speeds are found only very near the critical surfaces (where the phase speed equals the zonal mean wind). For observed phase speeds, which are usually not more than 15-20

[3]It is interesting to note that Dunkerton et al. (1981) found that the orientation in the vertical-meridional plane of the EP fluxes at the lower boundary did not affect the orientation of the EP fluxes higher up in the stratosphere. Dunkerton et al. tested whether or not the convergence of EP fluxes into the upper stratospheric polar vortex during a sudden warming of the model resulted from a poleward tilt of the EP fluxes at the bottom of the model.

m/sec, the critical surface is located far enough from the jet core to be separated from the waveguide by an evanescent region.

To sum up, the meridional wavenumber is determined by the basic state. The shape of the forcing affects the meridional wavenumber only roughly in the lowest scale height, before the wave reflects off the sides of the waveguide. This means the zonal wavenumber and phase speed of the forcing affect mostly the vertical propagation characteristics (by affecting m through n^2_{ref}), resulting essentially in a Charney-Drazin type criterion for the propagation of waves up the waveguide.

5.3.3 The dependence of wave structure on the wave geometry and damping

In this section, we will show the relation between vertical wave structure and the vertical wavenumber in our model. We will also test the sensitivity of our results to the damping in our model. This is essential for comparison with observations, because damping is a large uncertainty in the atmosphere.

In the control run, the sponge layer (the only damping in our model) roughly coincides in height with the turning point. As we saw in chapter 4 (equation 4.12), the wave amplitude will decrease with height due to evanescence and to damping. We would like to determine which of these effects is dominant in our run. We do this by raising the sponge layer and the lid of our model by 5 scale heights, leaving all other features the same (referred to as the high-sponge run). The location of the turning surface does not change as a result.

Figure 5.5 shows the longitude-height cross section of the temperature and geopotential height fields at latitude $y = 2.45$ (which is roughly in the middle of the waveguide between 4-8 scale heights), for the control run and the high-sponge run. We see that the control run wave tilts westward with height. The tilt in geopotential height is stronger at the bottom scale height, where we almost have a node, as a result of the downward reflection. The temperature amplitude peaks slightly below the highest turning point. The amplitude of the high-sponge wave is larger than the control, and the peak of geopotential height amplitude moves up. This implies that the amplitude maximum in the control run is a function of the damping. Also, the waves in the high-sponge run are almost vertical, indicating an 'almost standing' wave pattern (just a tiny bit of the wave leaks to the sponge layer but most of it reflects downward). This means that the sponge layer in the control run inhibits reflection from the turning points, just as we saw happening in the one dimensional model (section 4.4.2 and figure 4.8). Another way to show this is by looking at ψ

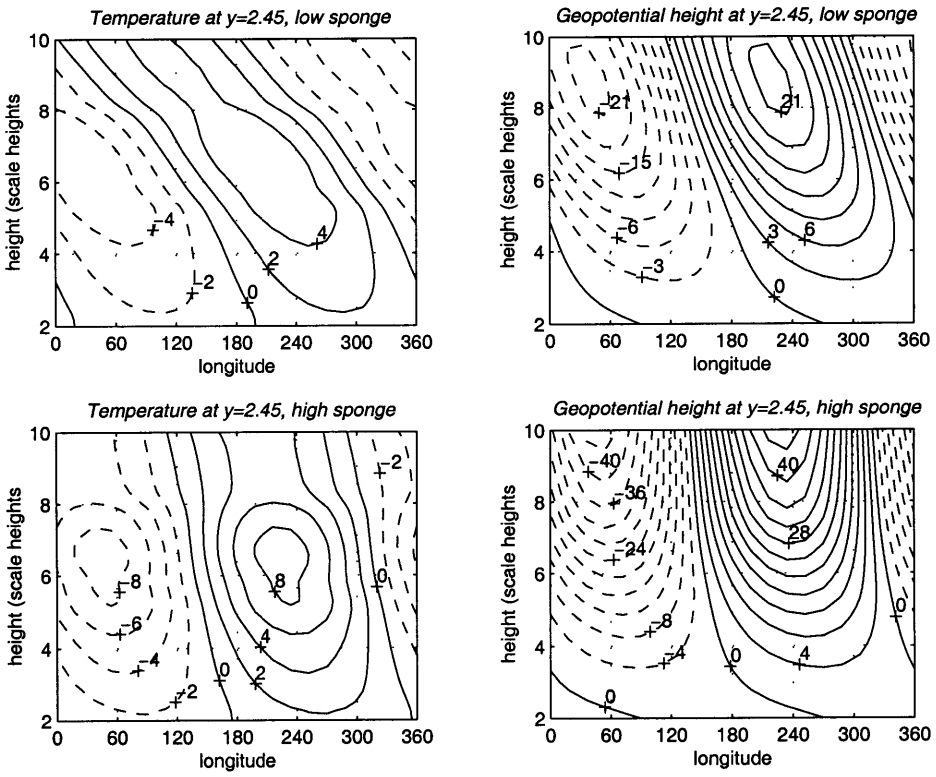

Figure 5.5: Zonal-height cross-sections of temperature (left) and geopotential height (right) for the control run (top) and a similar run with the sponge layer and the top of the model shifted upward by five scale heights (bottom). Units are arbitrary, and the same contour intervals are used for the two runs.

(equation 4.9), which is constant if there is pure propagation with no reflection, and is decaying in regions of wave evanescence or regions with damping. Figure 5.6 shows $|\psi|$ (the solid lines), along with the sponge layer damping coefficients (dashed) in $days^{-1}$, and the meridional wavenumber (dotted, same contour values as in figure 5.3). We see clearly that now that the sponge layer is much higher, ψ decays when it reaches an evanescent region and reflects downward, and not because it reaches the sponge layer. The damping time scale at the highest turning point (z=8) is 100 days, too small to cause the rapid decay in amplitude. The region of minimum amplitude just below z=3 is indicative of downward reflection.

We see that the top sponge layer damping affects the vertical EP flux (the vertical phase tilt with height) without affecting the vertical propagation geometry. This implies that the damping affects the relative magnitude of the coefficients of upward and downward reflection (in the case of meridional propagation, affecting the relative magnitudes of f and g in equation 5.15).

We specified our sponge layer quite arbitrarily, with the only consideration that it

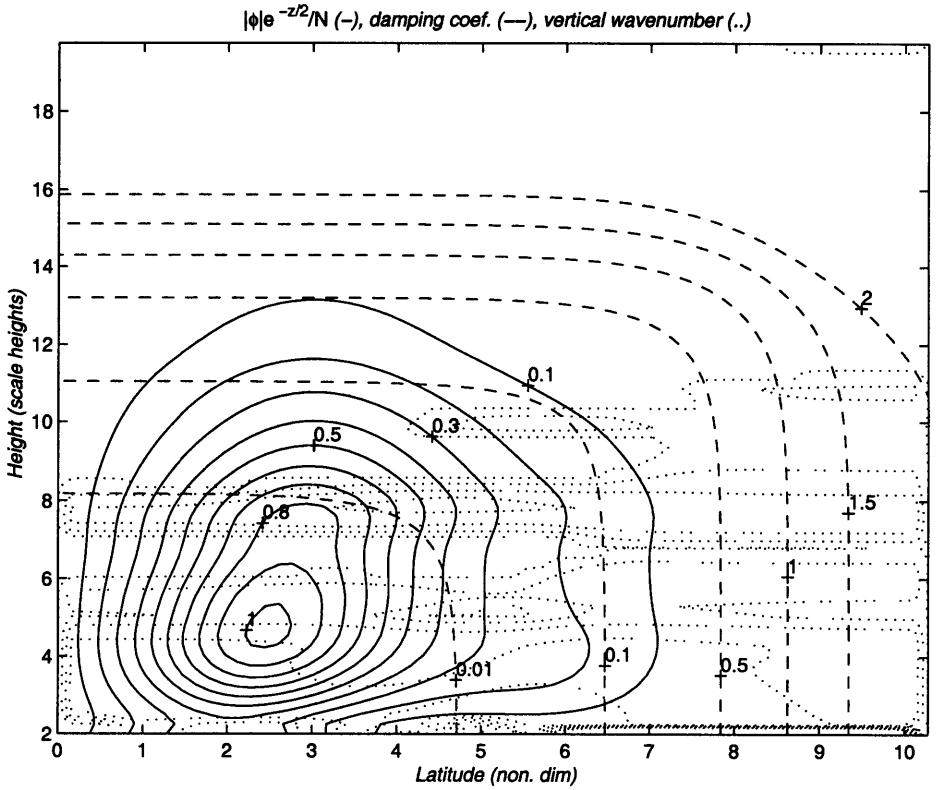

Figure 5.6: $\psi = \phi \frac{e^{-z/2}}{N}$ (equation 4.9, solid lines) in arbitrary units, the sponge layer damping coefficient (both Newtonian cooling and Rayleigh damping) in $days^{-1}$ (dashed) and the vertical wavenumber (dotted line, same contours as in figure 5.3), for a run which is like the control only with the sponge layer and the top of the model shifted upward by 5 scale heights.

absorb waves and not reflect them back, and not based on any realistic damping parameterization. We therefore run the model with a high sponge layer and Newtonian damping that is based on Dickinson (1969b), which follows the shape of the basic state temperature profile[4]. We find that the Newtonian damping does not affect the meridional wavenumber. The main effects are to decrease the amplitude, with a very small effect on the vertical wavenumber of the perturbation. We also see that the radiative damping inhibits downward reflection, although not to the same extent as the sponge layer at 10.5 (the control run).

Finally, we varied the strength and size of the equatorial sponge layer, to test the sensitivity to the damping at the equator. The meridional wavenumber is not affected, as long as the sponge layer does not extend into the waveguide region. We

[4]The damping time scale decreases from 20 days at the bottom of our model to 2 days at 50 km, with 10 days at 30km. Above 50km the damping time scale increases to more than 5 days between 70-80 km above which it decreases as a result of the sponge layer.

expect the magnitude of the damping, however, to affect the meridional EP flux, since damping acts as a sink of wave activity.

5.3.4 The effect of a turning surface on the time evolution of waves

Since the source of planetary waves in the stratosphere is not constant with time (also, the basic state itself varies with time, but we will discuss this later), waves are not in strict steady state. One of our main goals is to understand how the basic state wave geometry affects the time evolution of the waves. The existence of a turning point for vertical propagation will affect the time evolution of the wave most strongly, because the amount of wave that is propagating upward relative to the amount of the wave that is being reflected down will vary with time, causing the phase tilt with height to change. This change will manifest itself at a given height and latitude as a zonal phase translation, that will project onto traveling modes in a Fourier decomposition in time. Since such an analysis is often done to look for traveling modes, it is important to know how much of the signal comes from the transient adjustment of waves to steady state. In this study, we will mostly be concerned with showing that indeed such variations in phase occur in the stratosphere as a result of transient evolution in the presence of turning points (chapter 7), rather than try to estimate quantitatively how much this contributes to the Fourier decomposition statistics.

As a start we use a time dependent version of our quasi-geostrophic, β-plane model. The details of the model are described in appendix B, and it suffices here to say that the setup is like the steady state model, only we specify the wave geopotential height at the bottom as a function of time. Unless otherwise noted, the basic state is kept constant with time. Figure 5.7 shows time-height plots of the wave 1 geopotential height amplitude without the contribution of density, along with the phase, from an integration where the basic state is like the control run, only with a high lid and sponge layer (such that the turning point at 8.5 scale heights is below the sponge damping region). We have an additional constant damping of approximately $0.04 day^{-1}$ to assure numerical stability. We initialize the model with no wave, and force it as in the steady state control run, and turn the forcing off on day 30 (over 3 days). This is equivalent to switching on a source instantaneously (and switching it off later). We see that initially, the westward phase tilt (phase increase with height) increases with time, reaching a maximum at model day 5. This is when the wave front reaches the turning point. After this time, part of the wave reflects down, resulting in a decrease of the phase tilt with height. The downward reflection is followed by further

adjustments to the steady state, which are leakage to the equator, equilibration with the damping and some weak reflection back up from the surface and the sides of the waveguide. When the forcing is shut off, the upward propagating part of the wave decreases first, causing the wave to tilt vertically (day 37), and finally eastward with height. A Karoly and Hoskins ray tracing (see section 6.5) estimate of the vertical group travel time gives roughly 3-4 days to reach the highest turning point at 8.5 scale heights. A different way of estimating this time, by tracking wave packets (which is more appropriate for our calculation, see chapter 6), gives an estimate of 4-5 days. The time scales in our model run seem slightly longer. It takes the wave front 5 days to reach the turning point after the forcing is turned on, and about 6-8 days both to reflect back down (the time between maximum and minimum phase tilts with height), and to reach the maximum eastward phase tilt when the forcing shuts down. There is some slowing down and ambiguity of the travel times, associated with the wave front being spread out over a region. Apart from the beginning of the run when the forcing is turned on, and the end, when the forcing shuts off, the phase tilt with height hardly changes below 5 scale heights. This may be due to a leaking out to the equator of the downward reflected part of the wave, such that when the wave reaches 5 scale heights, all of the reflected part leaked. Time-latitude phase plots confirm that there is constant leakage to the equator.

It is clear from this model run that the wave sees the turning point, and that the time evolution is affected by it, mostly when the bottom forcing changes rapidly. This stresses the fact that the wavenumbers diagnosed from the steady state solution are a diagnostic of the *basic state* wave propagation characteristics rather than of the wave structure[5]. We can therefore use the steady state solution to an observed instantaneous basic state as a diagnostic of the propagation characteristics of that basic state. Later on we will use this to determine whether observed variations of wave structure are consistent with variations of the basic state, both on seasonal (section 5.4.2) and on daily (chapter 7) time scales.

Another point to make is that the transient evolution of the wave is manifest as a temporary propagation in the zonal direction (i.e., a non-zero zonal phase speed), which would temporarily change the local index of refraction relevant to the perturbation. It is therefore not obvious that the wave will 'see' the index of refraction that is relevant to the stationary wave throughout the integration. However, calculations using the maximum transient phase speed in our model run, show this has a very

[5]Note that wavenumbers that are calculated from an instantaneous wave field are meaningless when the time variations are large.

small effect on n^2_{ref} hence this is not an issue here.

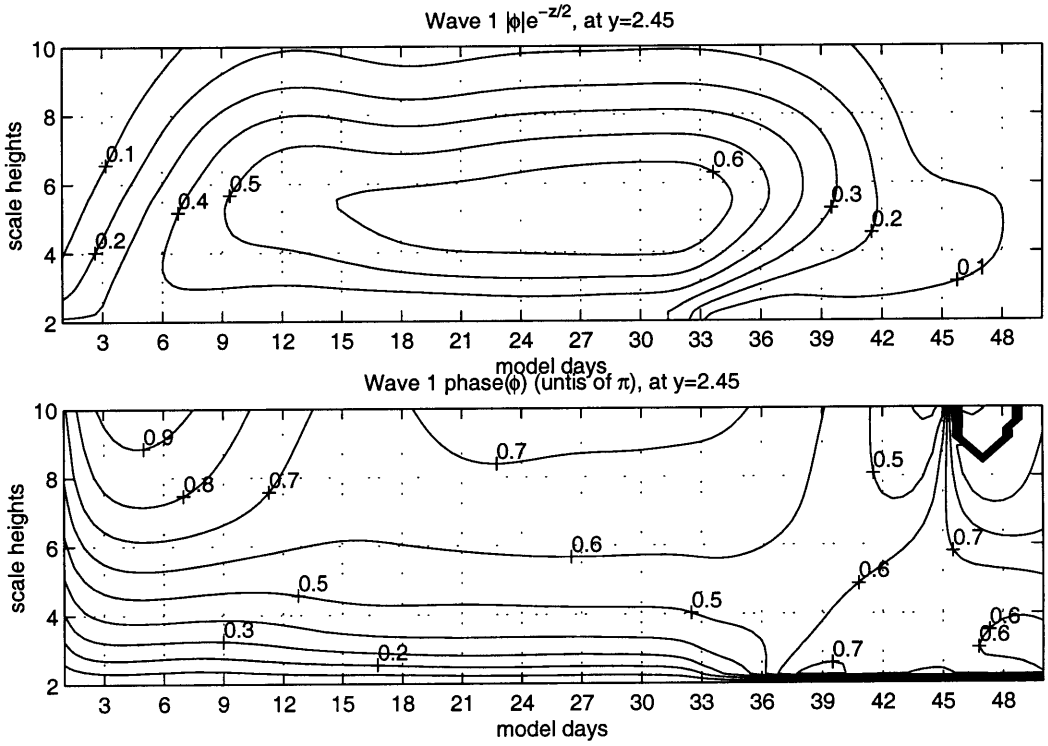

Figure 5.7: Height-time plots of geopotential height wave 1 amplitude multiplied by the square root of density $|\phi|e^{-z/2}$ (top) and phase (bottom), at y=2.45, for a model run like the control, only with a high lid and sponge layer, where the forcing is turned on, then off (see text for details). Geopotential height amplitude is nondimensional and phase is units of π. Time is in model days (roughly 1 day).

5.3.5 Validity of the WKB approximation

In section 5.2.1 we wrote down conditions on the meridional and vertical wavenumbers for the WKB approximation to be valid (5.16- 5.17), and for the solution to be a wave (5.16- 5.17, with $\ll$ replaced by $<$). We now check whether these conditions are satisfied, and where. Figure 5.8 shows the absolute value of the left hand sides of 5.16- 5.17, calculated from the run using the high lid and sponge layer (the control run yields similar results).

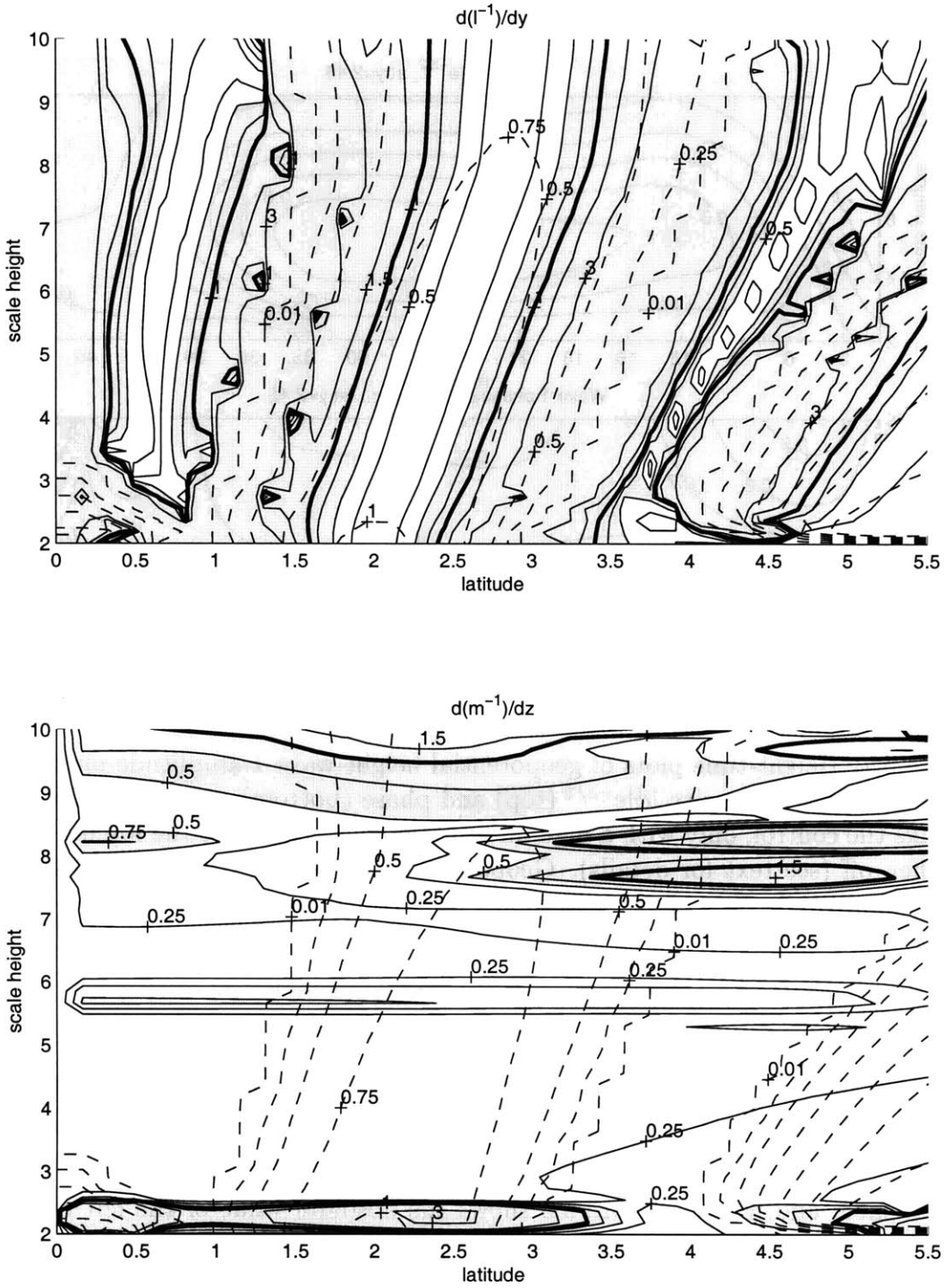

Figure 5.8: The validity of the WKB approximation: Absolute values of $\frac{d}{dy}\left(\frac{1}{l}\right)$ (top) and $\frac{d}{dz}\left(\frac{1}{m}\right)$ (bottom) (5.16- 5.17), which are conditions on the meridional and vertical wavenumbers, respectively. Regions larger than 0.5 are shaded. Contour values are 0.25:0.25:1, 1.5, 3, and the 1.0 contour is thick. Also plotted for reference is the meridional wavenumber (dashed), with contour values of 0.01, 0.25:0.25:1. See text for details.

For WKB, we need these parameters to be much smaller than 1.0. The shaded regions are for values greater than 0.5, which we consider as regions where WKB is violated. We also highlighted the 1.0 contours by making them thick. We see that the vertical wavenumber satisfies the WKB quite well in most of the domain, and violates it near turning surfaces, and near the bottom. The region near the bottom is where we have a node, due to the downward reflection from the turning point. The meridional wavenumber, on the other hand, satisfies WKB only in a narrow region near the center of the waveguide. Moreover, the solution is wavelike in nature only in the center of the waveguide, in a region that is slightly wider than the WKB region. For reference, we have plotted the meridional wavenumber (dashed). In the next section we will see some possible implications of the meridional WKB condition being satisfied only in a narrow region.

5.3.6 An approximate 1D model of the wave in the center of the waveguide.

The waveguide configuration simplifies the structure of the response by determining the meridional wavenumber of the perturbation and by rendering the response almost separable in the across-waveguide and along-waveguide directions[6]. In this section we try to approximate the response at the center of the waveguide using a one dimensional (1D) model, as follows: We take the basic state from the center of our channel. We also take the damping coefficients and the forcing at the bottom (amplitude, phase, zonal wavenumber and phase speed) from the center of our channel. Finally, we take the meridional wavenumber of the two dimensional (2D) response, also at the center of our channel. Note that the waveguide is tilted, hence for each level we find the latitude of maximum l and take the profile values from there. We then assume this is a vertical profile by ignoring the fact that the waveguide is slanted, and solve equation 4.6 to obtain the 1D approximation of the geopotential height and temperature at the center of the 2D model.

What we can learn from such an exercise is whether our assumption of a WKB solution of the form 5.3 is a good one, and whether the index of refraction and wavenumbers defined in the previous section are meaningful. Also, this allows us to compare the response of the 2D model to that of a 1D model directly[7]. In chapter 4

[6]The response can not be purely separable because there is damping at the equator and not at the pole, causing a leakage to the equator.

[7]This is useful, for example, in light of studies like Plumb (1989) which used a one dimensional wave mean-flow model to explain the mid-winter minimum in wave activity observed in the southern

we saw that the response in one dimension can be very sensitive to the wavenumber of the forcing. It is interesting to see if this holds for the 2D model as well.

We run and compare the 1D and corresponding 2D models for a range of zonal wavenumbers. We do this for the control run of section 5.3.1, for a different basic state, for other mid-channel latitudes and with different damping profiles. The 1D model succeeds qualitatively in reproducing some of the main aspects of the response. Most interesting is that the one dimensional model captures the sensitivity of the response to the zonal wavenumber (with a few exceptions which will be shown shortly), and to the mid-channel latitude. It does not succeed in quantitatively reproducing the amplitude and phase of the perturbation. There are few reasons for this failure. First, the leakage from the side of the waveguide to the equator is not accounted for, which will cause the one dimensional model to overestimate the response. Second, we ignore the tilt of the waveguide and the resultant stretching of the coordinate when setting up our one dimensional model. It is hard to say how this will affect the results. Also, we see that the wave nature of the solution, and the WKB condition (section 5.3.5), strictly hold only in the middle of the waveguide. Note that despite the violation of WKB in the meridional direction, the 1D model succeeds in qualitatively representing the vertical wave propagation along the waveguide. This is not so surprising since in the vertical direction WKB holds quite well in most of the domain.

Figure 5.9 shows the geopotential height and temperature[8] amplitudes for the 1D model and at the center of the waveguide of the 2D model, as a function of zonal wavenumber, for a run that is like the control except for a mid-channel latitude of 45° instead of 55°. This configuration showed the largest differences between the two models. The one dimensional model overestimates the response for all wavenumbers. There is a very pronounced resonant wavenumber ($k = 0.6$)[9]. This resonance is not a feature of the 2D model. On close inspection, we see that this is actually the main difference between the two models. Other than that, the shape of the response is quite similar- the maximum amplitude in temperature and geopotential height is at the same altitudes, and the abrupt cutoff of the response occurs at exactly the same wavenumber (0.7). The resonant wavenumber is reminiscent of the neutral wavenumber response in the 1D model of chapter 4 (see for example the neutral

hemisphere, in terms of the basic state seasonal evolution.

[8]Note that since the temperature is a vertical derivative of the geopotential height, the only way the model can get temperatures correctly is by getting both the amplitude and phase structure of the geopotential height field.

[9]Other 1D model runs, including the control run, did not have such a pronounced resonant wavenumber.

wavenumber in figure 4.6)[10], which suggests constructive interference due to vertical reflection. An inspection of the vertical wavenumber in the middle of the channel (not shown) shows that wavenumbers smaller than $k = 0.6$ propagate up to 8.5 scale heights, and wavenumbers 0.6 and larger have a turning point below six scale heights. This is important since the damping is large enough to inhibit reflection at 8.5 scale heights, but is not large enough to inhibit it at six scale heights (see section 5.3.3). This is an interesting result which we speculated upon in chapter 4, namely that in two dimensions we will not see strong constructive interference, because of the more complicated geometry and the leakage of wave activity to the equator.

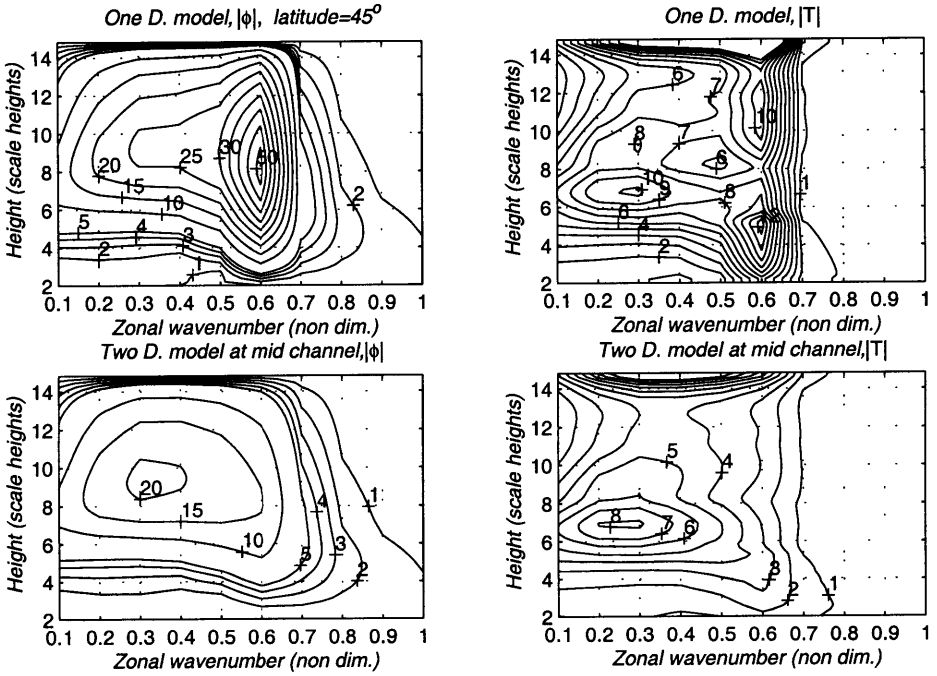

Figure 5.9: Height-wavenumber sections of geopotential height (left) and temperature (right) amplitude, for a one dimensional model (top) and the corresponding response at mid-channel of the two dimensional model run with a mid-channel latitude of 45°, but otherwise like the control run (bottom). To create the figure, we ran the model for nondimensional zonal wavenumbers $k = 0.1, 0.2....1.0$. For reference, waves one and two are $k = 0.325$ and $k = 0.65$ respectively. See text for more details.

[10]We should keep in mind that the 1D model is a stationary wave model, and not a normal mode calculation as is chapter 4, however, the dependence of the waves on wave geometry is the same in both models, which allows them to be compared.

5.4 Applying the diagnostic to observations

One of our main motivations behind studying the relation between the basic state wave geometry and vertical wave structure is to explain vertical structures of observed waves. Our results suggest that the meridional wavenumber is determined by the basic state, regardless of tropospheric forcing characteristics (zonal wavenumber and phase speed and latitudinal shape of the forcing) and of damping. The vertical propagation characteristics are sensitive to these parameters, in a manner which is qualitatively like the one dimensional model. In this section we obtain the propagation characteristics of observed basic states in order to explain a few characteristics of observed waves.

5.4.1 The effect of spherical coordinates and model setup

Before we apply our diagnostic to observations, we need to make sure the results still hold for spherical coordinates. Most important, is the waveguide picture relevant, and does the insensitivity of the meridional wavenumber to forcing parameterizations hold. The equations and formulation of the meridional wavenumber and index of refraction in spherical coordinates are described in appendix D. The model we use spans the southern hemisphere[11] and for simplicity, we use the latitudinal resolution of the operational observation data product (2°). The vertical domain, unless specified otherwise, extends from 2-15 scale heights (14-105km). When we use observed basic states, the observations are interpolated in the vertical to the model grid (which is the same as in the β-plane) below 0.4 mb, and wind and temperature are kept constant above that, in order to apply a sponge layer. For simplicity, we specify a temperature that varies only with height, by taking an average of the observed temperature over 40-70° latitude. A comparison with runs using the full two dimensional temperature field give very similar results (similar enough given the uncertainties in the observations). We also put damping in the equator, as in the β-plane. See appendix B for more details.

The most notable difference between the two coordinate systems is in the relation between the zonal mean wind and the index of refraction. The PV gradient (equation D.8), interestingly enough, is not very sensitive to the coordinate system in the basic states we have used, because most of the contribution is from the meridional and vertical curvature, and the terms that depend on latitude are much smaller. The largest effect is close to the pole, where the spherical terms are large and negative.

[11] Our model is a channel model, in the sense that the polar boundary is a wall. More realistic polar dynamics would require a polar cap model.

There is a strong dependence of the index of refraction on latitude which causes the polar region to be evanescent to wave propagation (the term $\frac{-s^2}{\cos^2(\varphi)}$ in n^2_{ref} becomes large and negative, equation D.14). Also, the index of refraction becomes infinite at the equator, causing waves to refract equatorwards (as was shown by Karoly and Hoskins, 1982).

These differences aside, given a wave geometry, waves behave in a qualitatively similar manner in both coordinate systems. The refraction equatorwards results in smaller amplitudes in the spherical model. Also, the waveguide along the jet axis is much less separated from the equatorial propagation regions. This is a characteristic of the index of refraction, and is evident in the meridional wavenumber. Nevertheless, the insensitivity of the meridional wavenumber to the parameters of the forcing and damping hold for all the basic states we have checked, both ones we specified analytically and from observations.

5.4.2 The differences between mid-winter and later winter wave structure

We have mostly studied the evolution of waves in southern hemisphere winter of 1996. Looking at other years suggests the features we will present in this section are not specific to 1996.

As was shown in figure 1.3, in 1996 there were two major wave 1 events, one in July 18-August 19 (referred to as the mid-winter wave) and the other in September (referred to as late winter). Figure 5.10 shows the latitude-height amplitude and phase structures of the time mean waves in these two periods (both geopotential height and temperature). We see that the mid-winter wave has one temperature amplitude peak in the stratosphere while the late winter wave has two. Correspondingly, the geopotential height amplitude reaches a maximum much lower in late winter. The phase tilt with height in September is also smaller.

To show that the structures shown in figure 5.10 are not dependent on the time averaging, we plot the time-height evolution of the waves (amplitude and phase of the geopotential height and temperature, averaged over 40-70° latitude[12].) in July-August and in September (figures 5.11 and 5.12 respectively). While the geopotential height peaks at or above the top of the observation domain for most of the time in

[12]We have taken a latitude average of these quantities, rather than show their value at a specific latitude because we are interested in structure changes to the whole wave field (as opposed to apparent structure changes at a given latitude that result from a latitudinal shift of wave patterns). In any case, the differences between a latitudinal average and a latitudinal section at 60°S for the observations we will show are minimal.

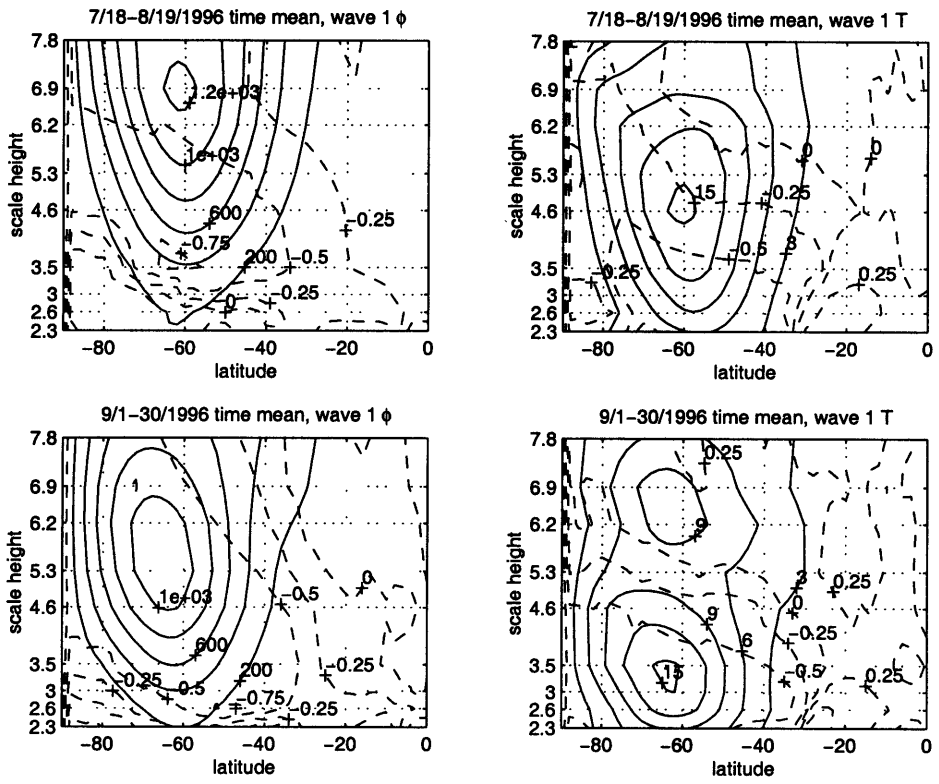

Figure 5.10: Time averaged wave 1 Geopotential height (left) and temperature (right) amplitude (solid) and phase (dashed), for July 18-August 19, 1996 (top) and September 1-30, 1996 (bottom). Geopotential height amplitude is in meters, temperature amplitude in °K and phase in units of π. The vertical grid is the observation grid. Time averaging was done on the amplitude and phase separately.

July-August, it peaks in the middle of the stratosphere during most of September. Correspondingly, the temperature has two amplitude peaks (in height) during most of September and only one in July-August. Note that longitude-height sections at 60°S, on August 8th and September 15th were presented in figure 1.6. The differences in structure are shown clearly there. We also see from figures 5.11 and 5.12 that the waves generally tilt westward with height (phase increases upward) but there are days on which the wave tilts to a vertical position (phase is constant with height, July 31-August 2, August 12, September 14-15, 24). Also, in September, there is one period when we have one peak in temperature amplitude instead of two (September 10-12). We discuss these variations in vertical structure in detail in chapter 7.

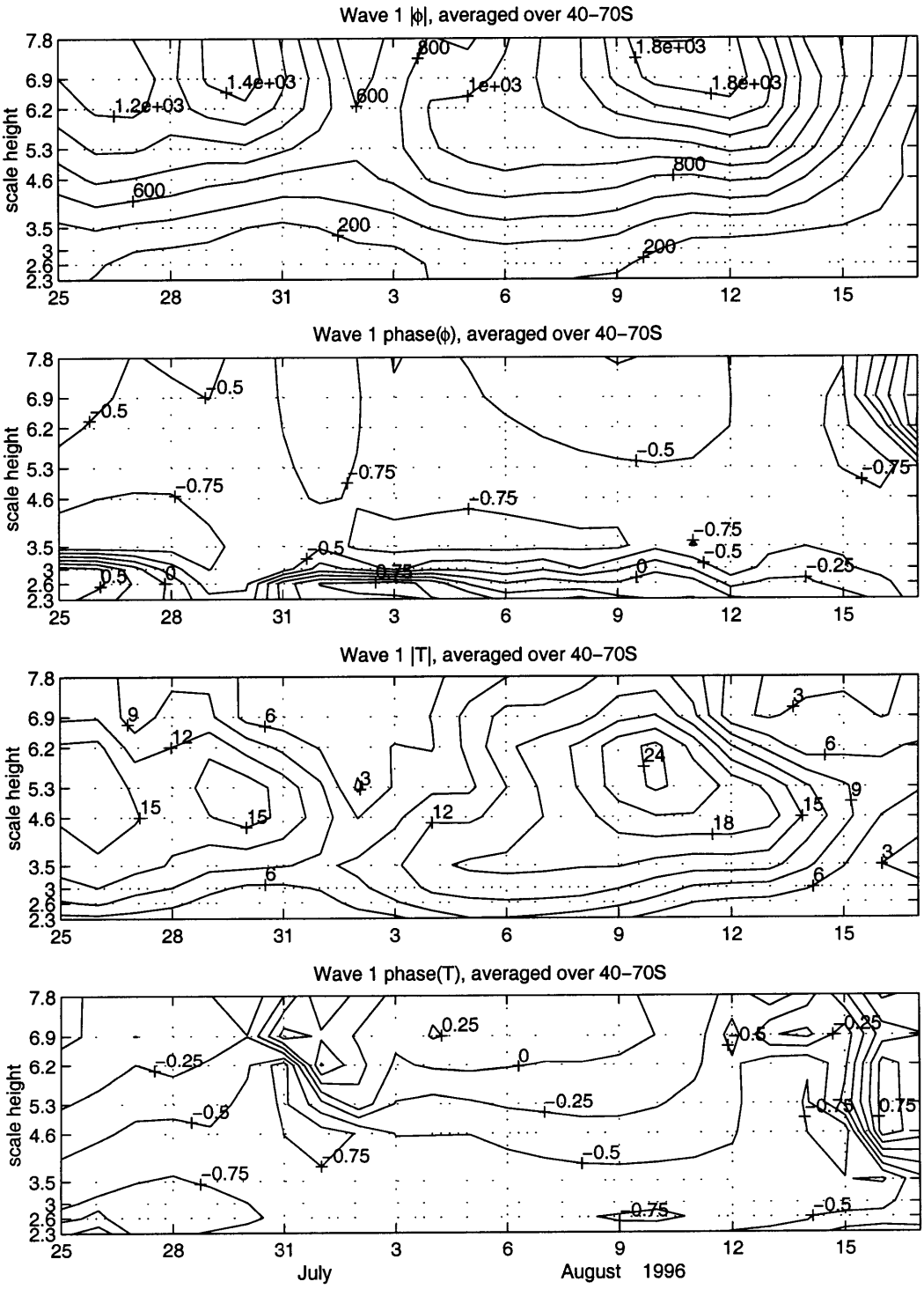

Figure 5.11: Height-time plots of a latitudinal average over 40-70S of wave 1 geopotential height amplitude and phase (respectively in top two plots), and temperature amplitude and phase (bottom two plots), for July 18-August 19, 1996. Geopotential height amplitude is in meters, temperature amplitude in °K and phase in units of π. Vertical grid is the observation grid.

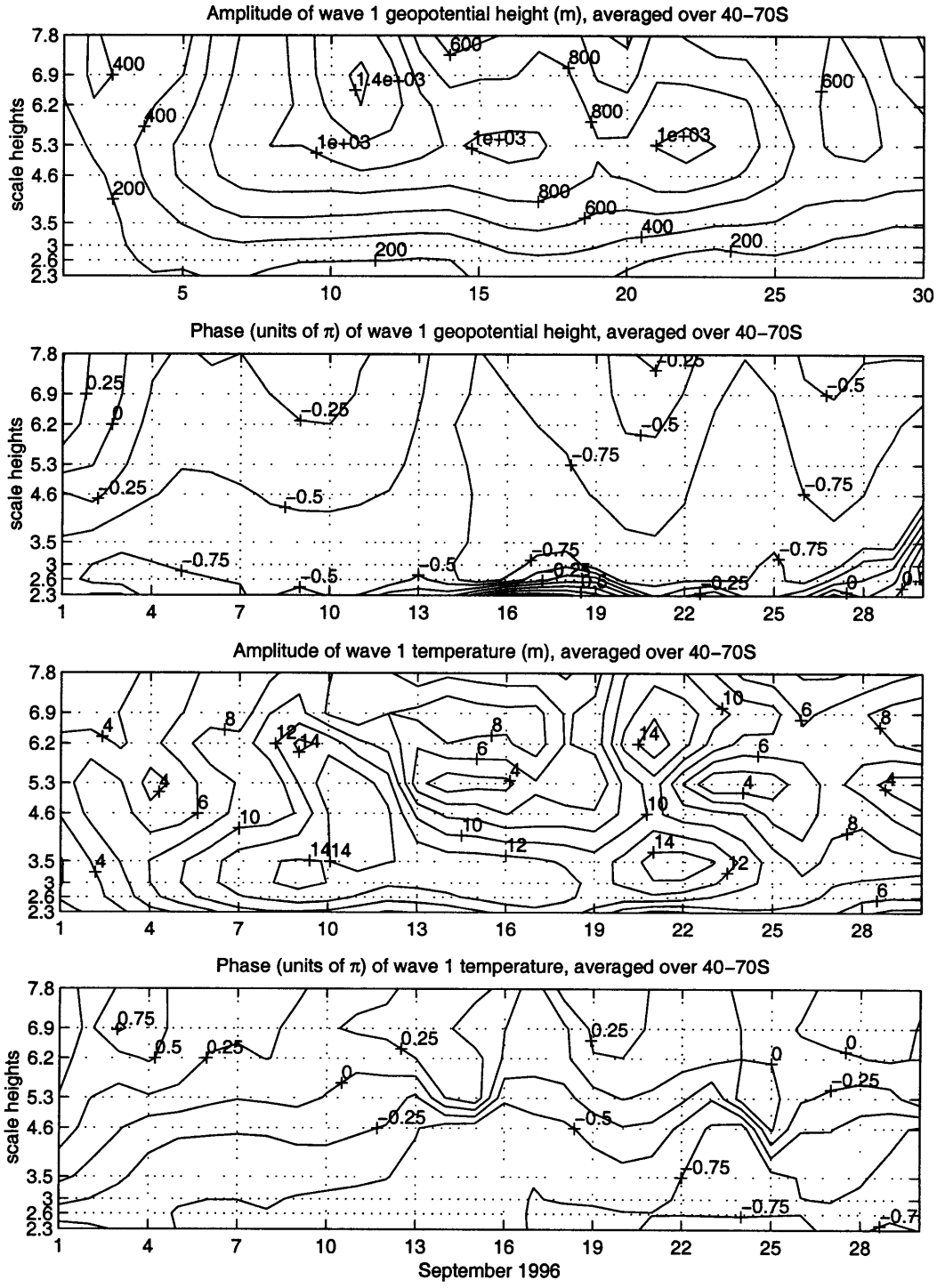

Figure 5.12: As in figure 5.11, only for September 1-30, 1996.

Based on results from this and the previous chapter, the September wave structure is suggestive of downward reflection from a turning point in the upper-middle stratosphere. To check this, we calculate the steady state solution on the observed basic states, using a wave 1 stationary forcing that is constant with latitude (geopotential height of 100 m). We use both the time mean and daily basic states and both lead to the same conclusions. Figure 5.13 shows the observed zonal mean wind averaged over each of the two wave events, along with the meridional and vertical wavenumbers of the corresponding steady state solutions. We see that the difference in basic states leads to a qualitative difference in the vertical propagation characteristics. In August, the zonal jet reaches 70 m/sec at 5.5 scale heights, 50°S, and there is vertical propagation in most of the domain (evanescent regions, where $m^2 < 0$, are shaded). By September, the jet has weakened and moved downward and poleward, reaching 50 m/sec at 5 scale heights, 60°S. As a result, a turning surface develops at 5.5 scale heights in midlatitudes, and the steady state geopotential height peaks at around 5.5 scale heights (not shown). The evanescent region is due to the positive vertical wind curvature on the upward flank of the jet, which causes the PV gradient to be small and even negative (see equation D.8). The resultant vertical decay of geopotential height is much stronger than the decay in an evanescent region that is due to strong winds (the Charney-Drazin criterion) as is usually the case in early and mid-winter[13]. This results in the geopotential height peaking so low (around 5 mb) in September, and the temperature having a node there. Apart from evanescent regions, damping will also cause the geopotential amplitude to decay. We calculated the steady state solution for the September mean basic state, with the lid and sponge layer raised by 5 scale heights, to make sure the decay we see is due to the evanescent region and not to the damping in our model. The results, as expected, were very similar to those obtained with the lower sponge layer.

To sum up, we can explain the seasonal evolution of wave structure in the southern hemisphere winter of 1996 in terms of the change in linear wave propagation characteristics of the basic state. The seasonal evolution of the zonal mean wind in 1996 was such that the jet weakened and moved downward and poleward in the end of August-beginning of September. This caused a turning point to form and the

[13]The reason for this is as follows. Ignoring damping, the decay of geopotential height in an evanescent region is dominated by the term $e^{-\int m dz}$ (equation 5.3). m, in turn, depends on n^2_{ref} and on the meridional wavenumber (equations 5.10 and D.14). Since the meridional wavenumber is not dependent on the type of evanescent region we have (it depends mostly on the meridional wind curvature), the main difference between evanescence in a region of positive vertical wind curvature and in a region of strong winds is that the latter has smaller PV gradients and hence much smaller (even negative) n^2_{ref}.

geopotential height to peak in mid-stratosphere. Correspondingly, the temperature structure assumed a double-peaked structure. Since the time evolution of the zonal mean wind observed in 1996 is not specific to this year (e.g. see Shiotani and Hirota, 1985, and references therein), we expect to see this seasonal change in wave structure in other years. Indeed, we find the double peaked temperature structure in September of other years we have studied (e.g. 1982, figure 7.18), both in wave 1 and wave 2.

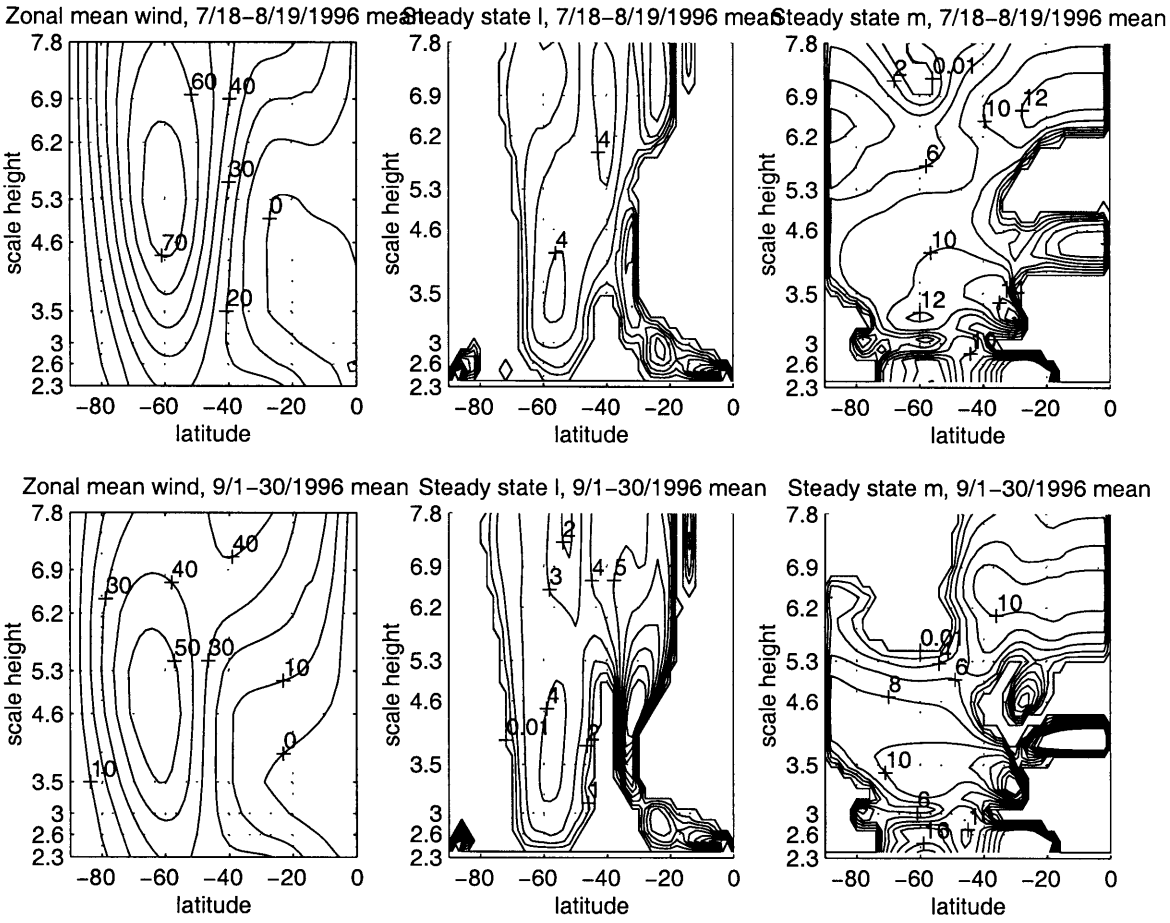

Figure 5.13: Observed time mean zonal mean wind (left) and the meridional (middle) and vertical (right) wavenumbers of the corresponding steady state model solution, for July 18-August 19 (top) and September 1-30 (bottom), 1996. Wind in m/sec, meridional wavenumber in $radians^{-1}$ and vertical wavenumber in $10^{-5}m^{-1}$. See text for details.

5.4.3 Relevance of the steady state solution to instantaneously observed waves

Finally, we will discuss the relevance of the steady state solution to the time evolving observed waves. We will concentrate here on general characteristics of the two. In chapter 7, we will concentrate more on specific cases of variations in wave structure and the basic state. Figures 5.14 and 5.15 show time-height plots of the steady state wave solution (averaged over 40-70°S) for the daily observed time basic state. These should not be taken as a time evolution of waves, since each point is a steady state solution. Also, the forcing is constant, unlike the real atmosphere where it varies from day to day. However, a comparison with figures 5.11 and 5.12 is useful in understanding the relevance of the steady state solution to an instantaneous one.
The first thing we see is that the general structure in terms of the height of amplitude peaks is captured, apart from a few days when the observed waves undergo structure changes, or the wave is very weak. This means that the observed characteristic differences in wave structure between middle and late winter are evident in the steady state solution. Damping, which is a big uncertainty in our model, can affect the height of the geopotential amplitude peak, but our analysis suggests that at least in late winter, the location of reflection surfaces rather than damping determine the height of the peak in geopotential height wave amplitude. Some of the differences we do see, however, are due to damping. Our model thermal damping, is most likely and underestimate in the domain of observations below the sponge layer. It is hard to determine how realistic the momentum damping we have is, since it represents the effect of gravity wave drag, both in the stratosphere and in the mesosphere[14]. We expect thermal damping to reduce both geopotential height and temperature amplitudes (with a larger effect on temperature). Indeed we see that the magnitude of the waves is overestimated by the steady state solution. Also, the phase structure of the waves is not very well captured by the steady state in most of September[15]. We have already shown that one of the main effects of damping, when we have a turning point, is on the phase tilt with height. As we saw in the time dependent model run described in section 5.3.4, some of the differences in phase structure, however, are due to the transience of the wave rather than to damping.

[14]Note that we need to have momentum damping in order to fully absorb the waves in the sponge layers. When we have only thermal damping, the waves reach the model boundaries and reflect.

[15]For example, in September, the observed wave 1 temperature phase tilts westward with height above 3.5 scale heights on most days, while the model steady state solution has an region of eastward or vertical tilt between 5.5 and 7 scale heights.

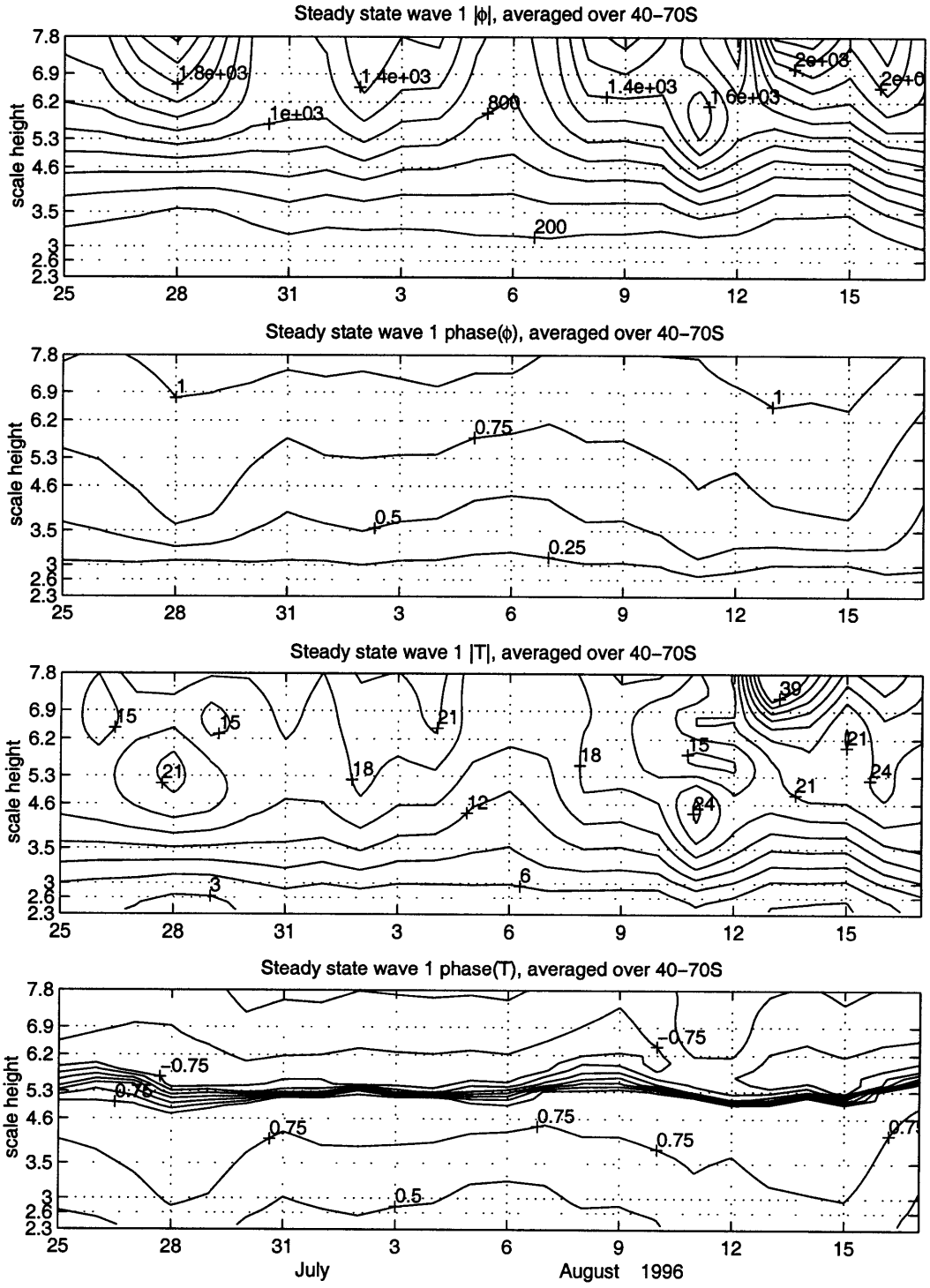

Figure 5.14: As in figure 5.11, only the steady state model solution for the instantaneous observed basic state. Note that the strong localized variation in temperature phase at around 5.5 scale heights is spurious- it is a change of almost 2π radians. It is not exactly 2π because of the latitudinal averaging(the surface is slightly tilted in latitude).

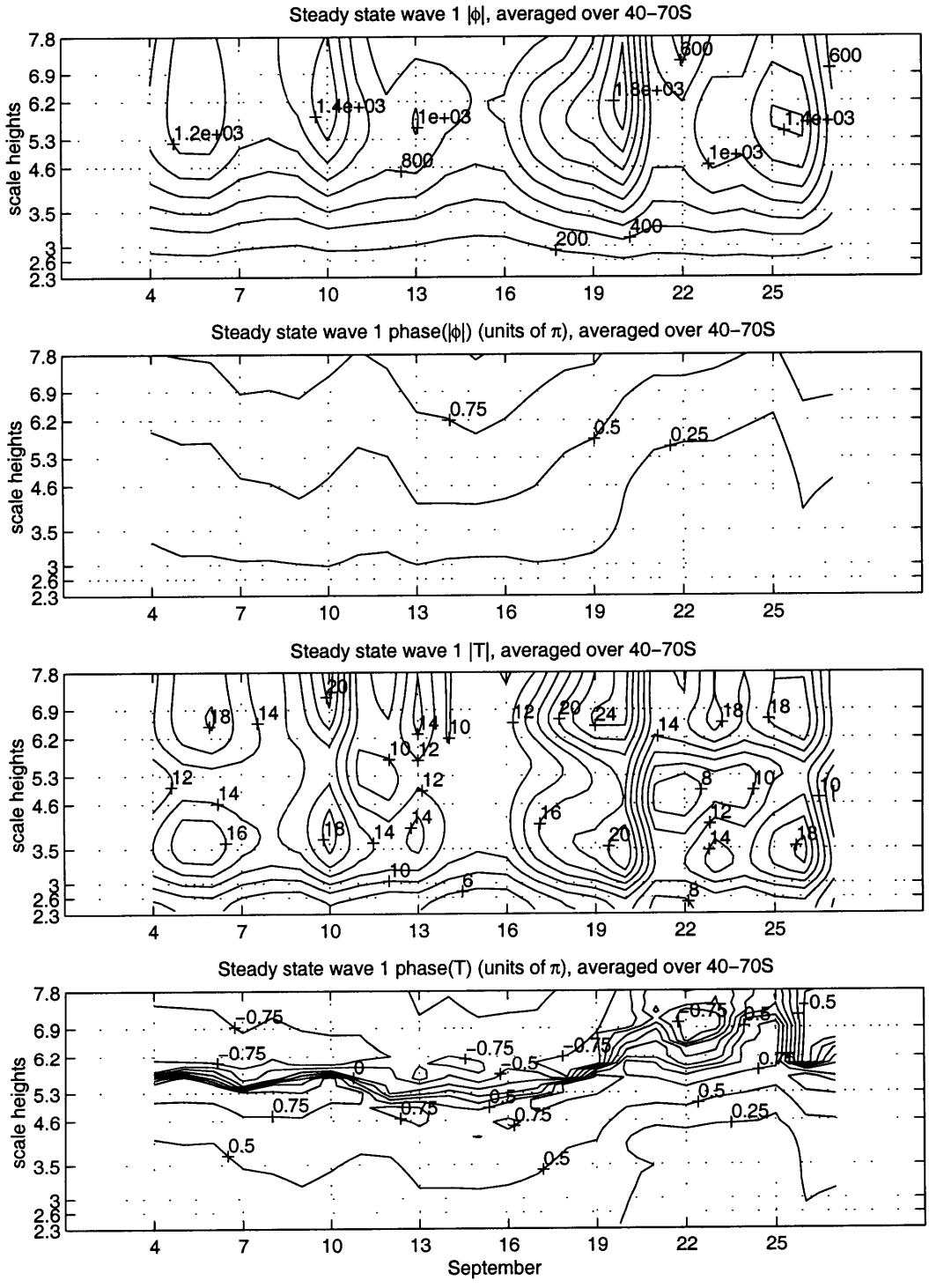

Figure 5.15: As in figure 5.12, only the steady state model solution for the instantaneous observed basic state. Note that the strong localized variation in temperature phase at around 5.5-6.5 scale heights is spurious- it is a change of almost 2π radians. It is not exactly 2π because of the latitudinal averaging(the surface is slightly tilted in latitude).

For example, there are a few days in July-August when the wave tilts vertically for a few days and then resumes a vertically propagating structure. We will discuss these in detail in chapter 7. Note that part of the observed transience is due to the fact that the wave source varies with time, unlike our model forcing.

Our calculations have shown how the steady state solution of a model using the observed zonal mean winds and temperature is useful in revealing the wave geometry of the basic state. While the steady state solution is not a good approximation to actual observed waves, it is useful in explaining general features of their structure. Generally, the shape of the amplitude is captured by the steady state solution (e.g. the height of the peak in geopotential height amplitude or the number of peaks in the temperature structure). The phase structure, on the other hand is not generally captured by the steady state solution, both because of uncertainties in damping and because the wave is not in steady state. Turning points will lead to temporal changes of the phase structure by reflecting the waves as they evolve in time (e.g. section 5.3.4). We will discuss this more in chapter 7.

5.5 Summary

We have shown that the two dimensional wave propagation problem is conceptually simplified by the existence of a waveguide that is oriented along the polar night jet axis, because it determines the meridional wavenumber. As a result, variations in the tropospheric forcing parameters (i.e. zonal wavenumber and phase speed) affect the vertical wavenumber only, which allows us to treat the two dimensional problem essentially as one dimensional.

The main advantages are that we can predict the effect of having damping and turning surfaces (in the vertical direction) on the vertical structure of the waves. For example, damping has a large effect on the phase structure of waves that have turning surfaces because it controls the amount of downward reflection from these surfaces.

We also tested the analogy to a 1D propagation problem quantitatively, by formulating an equivalent one dimensional model, using the meridional wavenumber and basic state from the middle of the waveguide. We find that the evolution in the waveguide is qualitatively like a 1D model, but not quantitatively. Most important is the fact that the sensitivity to zonal wavenumber, in particular the cutoff wavenumber to vertical propagation, are the same for the two models, as well as the sensitivity to parameters like the mid-channel latitude. The discrepancies are explained by the leakage of a large part of the perturbation to the equator, and also, because the WKB condition in the meridional direction is violated outside of a narrow region in

the middle of the waveguide. In the vertical direction, however, WKB holds quite well in most of the domain. A feature of the 1D model that was not found in the analogous 2D model is resonance of specific wavenumbers. Overall, the concept of meridional and vertical wavenumbers and the wave propagation/index of refraction picture work remarkably well.

We also applied the wavenumber diagnostics to observations. Testing linear wave propagation concepts on observations constitutes one of the main goal of this thesis. It is therefore exciting to find a seasonal transition in vertical wave structure which is nicely explained by the evolution of the basic state wave geometry. Towards end of winter, when the jet moves downward and poleward, and weakens, a turning point forms at around 5 mb. This turning point is not a result of the winds becoming too large for propagation (the Charney-Drazin criterion), but rather a result of the PV gradient becoming small and even negative when the vertical wind curvature becomes positive above the jet peak. The downward reflection from the turning point is clearly evident in the vertical structure of the waves. Most notable is the effect on vertical temperature structure, which develops a node or almost-node at the turning point. By almost-node we mean a minimum in amplitude at a region of rapid vertical phase variations. Whether there is an actual node or an almost-node depends on the degree of reflection from the turning point. The most important factor to determine that is damping.

Finally, we discussed briefly the effects of wave geometry on the time evolution of waves, in particular, when we have a time varying source. We regard the steady state solution using the observed basic state as a tool to obtain the propagation characteristics of the basic state. The actual observed instantaneous wave is influenced by many factors which do not come into the steady state solution. For example, the time evolution of the tropospheric forcing, and/or the basic state will affect the instantaneous wave structure. Damping will also play a role. Our knowledge of the wave geometry and how waves evolve in it are useful in understanding the more complicated time evolution of the waves. We will apply this to observations in chapter 7.

Chapter 6

The structure of stratospheric planetary waves from a wave activity point of view

6.1 Motivation

This chapter is a short digression from applying the wavenumber diagnostics to observations, in which we view a wave field as the propagation of many 'wave activity packets' with a velocity analogous to group velocity, and develop a diagnostic technique to study the evolution of these packets within a stratospheric planetary Rossby wave. Our diagnostic is based on defining a coordinate system that follows the propagation of wave packets.

This idea has stemmed from a few general thoughts. First, in thinking about distinguishing in observations between a propagating mode and a quasi stationary wave that undergoes vertical structure changes (see discussion in section 1.2.2), we come up against the question of how to track a wave as it propagates up through the stratosphere. In particular, tracking a wave becomes tricky when we realize that as it propagates up through the stratosphere a few factors affect its amplitude, namely, time variations in its source, refraction due to variations in the index of refraction, and dissipation. Second, the relation between the index of refraction and the group velocity of a pure plane Rossby wave is very simple ($\vec{C}_g$ refracts up the gradient of n^2_{ref}), however, the relation between the index of refraction, the EP flux of the wave, and the vertical-latitudinal structure is not as straightforward (see chapter 5). Finally, the nonseparability of the basic state complicates our thinking of the waves in terms of vertical propagation of a given meridional structure (see discussion in section 1.1).

Our original intention in defining a coordinate system that follows the wave field was to come up with a coordinate in which the wave field is separable. We realized this is impossible since the distribution of damping is asymmetric relative to the basic state (causing for example the waveguide to be leaky only on its equatorial side). Looking at the wave in terms of wave activity, and defining a wave based coordinate does, however, highlight the relation between n^2_{ref}, group velocity, EP fluxes and wave structure, and it allows to distinguish between the various factors that control wave activity and hence amplitude. In addition, there are some uses for analyzing observations.

We will start by presenting the basic wave activity formulation (section 6.2). We then define our coordinate system, and use it to analyze a steady state (section 6.3) and a time dependent (section 6.4) wave field. In section 6.5 we discuss the relation between our diagnostic and the ray tracing technique of Karoly and Hoskins (1982). Finally, we discuss the limitations and uses of our diagnostic, as well as the application to observations (section 6.6).

6.2 Formulation- Tracking wave packets along Eliassen-Palm Flux lines

Wave activity (which is a form of wave action) is an important dynamical quantity of the waves because it obeys a simple conservation relation. We follow Andrews et al. (1987) in the following derivation. The wave activity equation is obtained by multiplying equation 4.1 by q' (where primes denote deviations from a zonal mean), and taking a zonal average (denoted by an overbar):

$$\frac{\partial}{\partial t}\left(\frac{\overline{q'^2}}{2}\right) + \overline{v'q'}\frac{\partial \overline{q}}{\partial y} = \frac{1}{\rho}\overline{q'\frac{\partial}{\partial z}\left(\frac{\rho\alpha T'}{N^2}\right)} + \overline{q'\frac{\partial(ru')}{\partial y}} \tag{6.1}$$

We have assumed Newtonian cooling and Rayleigh damping, as in chapter 5. The PV flux, $\overline{v'q'}$, is related to the EP Flux (equations 5.12, 5.13) in the following way:

$$\rho\overline{v'q'} = \rho\overline{v'\zeta'} + \overline{v'\frac{\partial}{\partial z}\left(\frac{\rho T'}{N^2}\right)} = -\frac{\partial}{\partial y}\left(\rho\overline{u'v'}\right) + \frac{\partial}{\partial z}\left(\frac{\rho}{N^2}\overline{v'T'}\right) = \nabla\cdot\vec{F} \tag{6.2}$$

We have used equations 4.2 and 4.4 and the fact that the zonal mean of zonal derivatives is zero. Multiplying 6.1 by $\frac{\rho}{\overline{q}_y}$, and assuming the PV gradient is constant with

time we get:

$$\frac{\partial A}{\partial t} + \nabla \cdot \vec{F} = D \tag{6.3}$$

where A is a wave activity *density*:

$$A \equiv \frac{\rho \overline{q'^2}}{2\overline{q}_y} \tag{6.4}$$

and D is the damping of wave activity. When WKB conditions hold, and we have a simple Rossby wave of the form: $\phi \propto e^{i(kx-\omega t+\int ldy+\int mdz)}$, with the following dispersion relation

$$\omega = kc = kU - \frac{k\overline{q}_y}{k^2 + l^2 + \frac{m^2}{N^2} - F(N^2)} \tag{6.5}$$

where the variables are as defined in chapter 4, we can define a group velocity as follows:

$$\begin{aligned} C_{gy} &= \frac{\partial \omega}{\partial l} = \frac{2\overline{q}_y kl}{(k^2 + l^2 + \frac{m^2}{N^2} - F(N^2))^2} \\ C_{gz} &= \frac{\partial \omega}{\partial m} = \frac{2\overline{q}_y km}{N^2(k^2 + l^2 + \frac{m^2}{N^2} - F(N^2))^2} \end{aligned} \tag{6.6}$$

The EP flux and the wave activity equal:

$$\begin{aligned} F_y &= \rho \frac{k}{2} |\phi|^2 Im\left(\frac{\phi_y}{\phi}\right) = \rho \frac{kl}{2} |\phi|^2 \\ F_Z &= \rho \frac{k}{2N^2} |\phi|^2 Im\left(\frac{\phi_z}{\phi}\right) = \rho \frac{km}{2N^2} |\phi|^2 \end{aligned} \tag{6.7}$$

$$A = (k^2 + l^2 + \frac{m^2}{N^2} - F(N^2))^2 \frac{\rho |\phi|^2}{2\overline{q}_y} \tag{6.8}$$

We see that the following relation holds in this case:

$$\vec{C}_g = \frac{\vec{F}}{A} \tag{6.9}$$

and wave activity is conserved following the group velocity, except for loss by damping on temperature or momentum. In more real-world scenarios, we do not have a pure Rossby wave, but rather, a superposition of waves, moving in all directions and reflecting at turning points. However, we can still define a "wave activity velocity"

which we will denote by $\vec{V}_a$, as follows (e.g. Palmer, 1982).

$$\vec{V}_a = \frac{\vec{F}}{A} \tag{6.10}$$

$\vec{V}_a$ is the velocity at which wave activity propagates along the wave field. Unlike the group velocity, it can only be calculated as a diagnostic of a given wave field. It is not a local property of the basic state, rather it depends on the wave geometry configuration and the global distribution of damping. Plugging into equation 6.3 gives an equation for the variation of wave activity following $\vec{V}_a$.

$$\frac{\partial A}{\partial t} + \nabla \cdot (\vec{V}_a \cdot A) = \frac{\partial A}{\partial t} + \vec{V}_a \cdot \nabla(A) + A\nabla \cdot \vec{V}_a = D \tag{6.11}$$

Following $\vec{V}_a$, wave activity increases when lines of $\vec{V}_a$ converge, and vice versa.

We can define a wave packet as a part of the wave that moves with the wave activity velocity, $\vec{V}_a$. If we now define a coordinate system that follows the wave packet (see next section), we can show in analogy to fluid flow, that the Jacobian of the transformation (J) is the ratio of the wave packet volume to its initial volume (which we take as a unit) and its fractional change following the wave packet equals the divergence of $\vec{V}_a$ (see appendix C for derivation):

$$\frac{1}{J}\frac{dJ}{dt} = \nabla \cdot \vec{V}_a \tag{6.12}$$

Plugging in equation 6.11, we get the following conservation equation:

$$\frac{d}{dt}(AJ) = DJ \tag{6.13}$$

where the material derivative is defined following the wave activity velocity ($\frac{d}{dt} = \frac{\partial}{\partial t} + \vec{V}_a \cdot \nabla$). AJ is the total amount of wave activity in the packet (A is wave activity density and J is the volume of the packet) and DJ is the volume integrated damping of wave activity. AJ is conserved as it moves along $\vec{V}_a$, unless there is dissipation. The wave activity density (which is proportional to the wave amplitude) does change, however, because the volume of the wave packet changes due to divergence of $\vec{V}_a$. In the next section we will define a wave based coordinate which follows wave packets, and illustrate the conservation of wave activity in the new system, using the control run of chapter 5.

6.3 The wave based coordinate: a steady state wave

Figure 6.1 shows the wave activity density (A), along with an EP flux diagram (EP flux arrows and EP flux divergence) for the control run of section 5.3.1. As a reminder, the pole is on the left hand side of our figures (small y). We see two distinct maxima of wave activity density, one in the middle of the wave guide, and the other equatorwards of it. The EP flux vectors are vertical at lower levels and tilt towards the equatorial sponge layer higher up. $\nabla \cdot \vec{F}$[1] is large at the equatorial sponge layer where there is a maximum of wave activity density, and in the upper stratosphere in the middle of the waveguide.

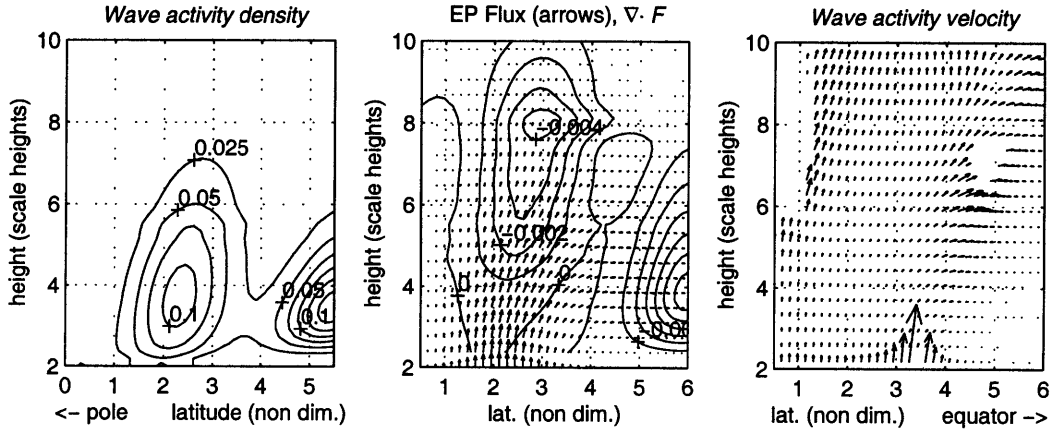

Figure 6.1: Latitude-height plots. Left: Wave activity density (A, arbitrary units). Middle: An EP flux diagram- the EP flux (arrows) and EP flux divergence (contours, see footnote 1 for dimensions). Right: Wave activity velocity, $\vec{V}_a$.

Also shown is the wave activity velocity $\vec{V}_a$. There are isolated regions where $\vec{V}_a$ is very large, on both sides of the waveguide. These regions coincide with regions of small or negative PV gradients where there is wave evanescence. Large $\vec{V}_a$ in evanescent regions is consistent with wave tunneling, which is much faster than wave propagation. There is also a region of large $\vec{V}_a$ and small wave activity density at the surface, which is due to vertical reflection. The location of the 'almost-node' depends on the specific run, and it is not always so close to the surface.

[1]To get the actual deceleration from $\nabla \cdot \vec{F}$ of figure 6.1 we need to put it in dimensional units and multiply by density. For dimensional units we need to multiply by $\frac{(U_o f_o)^2}{|\phi|_o^2}$, where ϕ_o is what we nondimensionalized geopotential height by (433m in this case). The resultant deceleration (taking into account the density factor) reaches a maximum in the middle of the waveguide at 10-11 scale heights. For a geopotential height perturbation of 100m at the bottom, the deceleration is 1(2) $\frac{m/sec}{day}$ at 8(10) scale heights.

We define a coordinate system (referred to as the $s - r$ coordinate) that follows a wave packet. Figure 6.2 shows this coordinate, along with $\vec{V}_a$ arrows. One coordinate, denoted by s, represents the time it takes a parcel to reach its location, and we obtain it by calculating integral lines of $\vec{V}_a$[2] (see appendix C). The other coordinate, denoted by r, represents the latitude at which a wave packet enters the stratosphere at the bottom. From figure 6.2 we see that s lines (which we refer to as either *wave packet paths* or *rays*) are everywhere tangential to $\vec{V}_a$. The value of r is constant along each of these lines, and is equal to the latitude of the ray at the bottom.

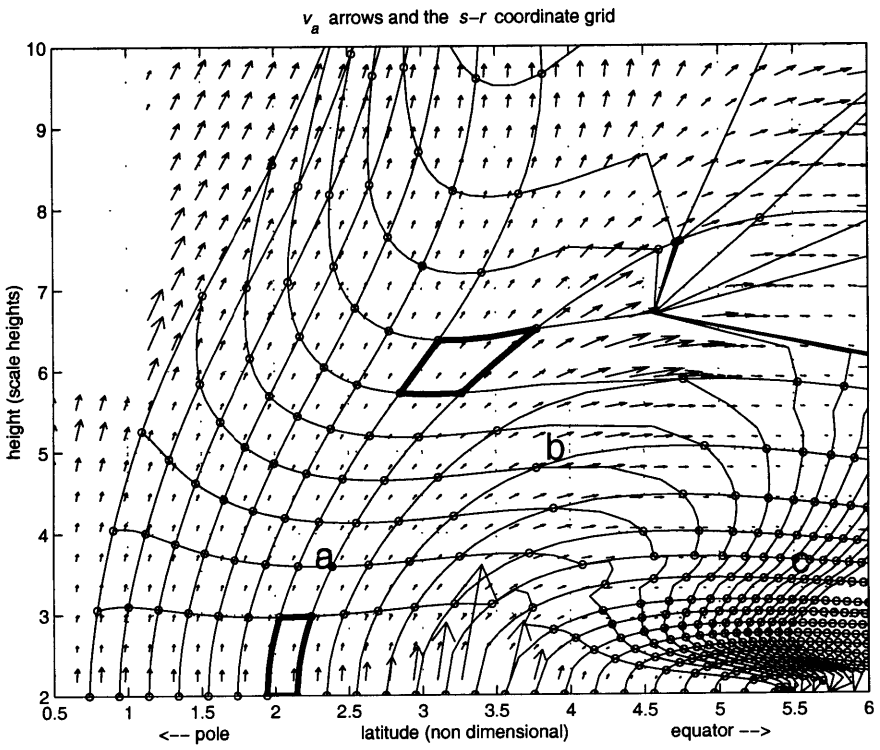

Figure 6.2: The wave based coordinate, plotted in geometric space (latitude-height). The s coordinate is tangent to $\vec{V}_a$ (arrows). The circles are spaced one day apart on s lines. **a**, **b**, and **c** are also marked in figure 6.3. See text for details.

Wave packets are defined as a grid box in $s - r$ space (assuming each grid box is one unit volume in $s-r$ space). The wave packets move along s lines, and change their volume both because the spacing between s lines changes (refraction) and because the magnitude of $\vec{V}_a$ changes along the packet path. The variation of the volume of a

[2]Since the definitions of wave activity and wave activity velocity are ambiguous in regions of zero and negative PV gradients, and since the regions of negative PV gradients are small, we have artificially set $\vec{V}_a$ to zero there. This explains why wave activity paths end in a point in the middle of the domain.

wave packet in $y - z$ space is clearly illustrated by the two thick grid boxes marked on figure 6.2.

According to equation 6.13, the total wave activity in the wave packet is constant (apart from loss to damping), hence an increase in packet volume will be accompanied by a decrease in wave activity density, and vice versa. This is nicely illustrated when we transform to $s - r$ coordinates. The transformation of a scalar field is done simply by interpolating the field to the grid points of the $s - r$ coordinate (the circles in figure 6.2), and plotting on a Cartesian $s - r$ plot (unskewing the $s - r$ coordinate). Figure 6.3 shows the wave activity density in both coordinate systems. To facilitate the comparison, we marked the location of three points with the letters a, b, c, on both plots, as well as in figure 6.2. In the $s - r$ coordinate system, the 'vertical' axis is along the wave packet path, while the 'horizontal' axis denotes the latitude at which the packets entered the stratosphere at the bottom of the model. Quantities that are conserved following the packet are constant along the 'vertical' s axis.

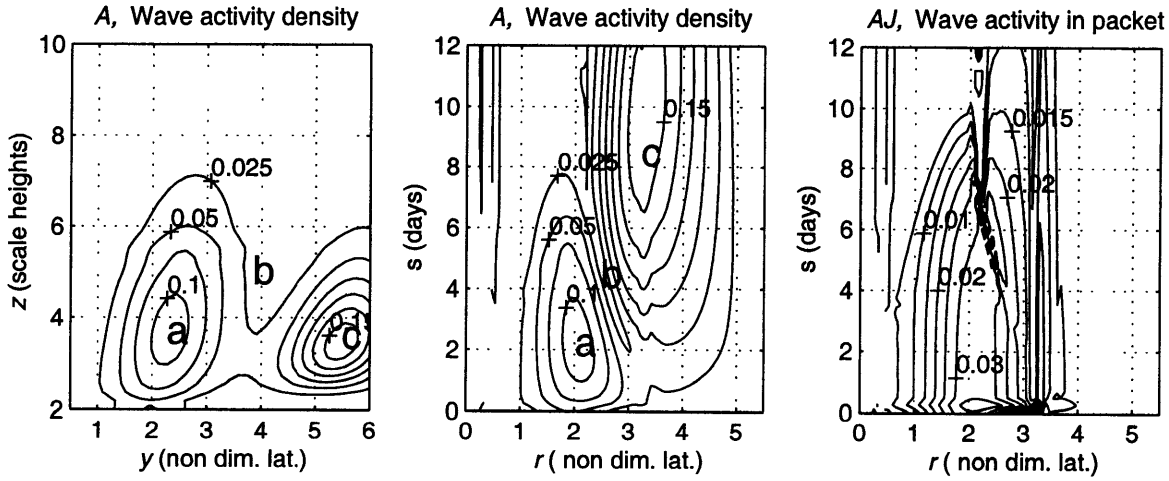

Figure 6.3: Wave activity density (A) plotted on $y - z$ (left) and $s - r$ (middle) coordinates, and the total wave activity in a wave packet (AJ) plotted on $s - r$ coordinates (right). The s axis is in days, the z axis in scale heights (7km), and the r and y axes in nondimensional latitude ($1190km$). Note that the pole is at small y/r.

Looking at the wave activity density on $s - r$ coordinates, we see that it varies a lot along the packet paths. This variation is due both to variations in wave packet volume, and to damping. To isolate the effect of damping, we multiply A by the wave packet volume (the Jacobian of the transformation), and plot on $s - r$ coordinates. The result is also plotted in figure 6.3. We see that the contours in the bottom 1/3 of the figure are 'vertical', meaning the total wave activity in the wave packets is conserved along their path for the first four days. After four days AJ decreases along the packets' paths. Using figure 6.2, we see that after four days (the r coordinate

lines are spaced 1 day apart) the packets reach 5-7 scale heights, depending on their latitude. The damping we specify has a time scale of more than 20 days at 5 scale heights and it becomes significant (less than 10 days) only above 6 scale heights.

It is very simple to calculate material derivatives in the new coordinate system, because it is simple a derivative along the s axis ($\frac{d}{DT} \equiv \frac{\partial}{\partial s}$). We can calculate the contributions of various factors to the wave activity budget, as follows:

$$\frac{DA}{DT} \equiv \frac{\partial A}{\partial s} = \frac{1}{J}\frac{\partial AJ}{\partial s} - \frac{A}{J}\frac{\partial J}{\partial s} \tag{6.14}$$

Figure 6.4 shows the contributions to variations in wave activity due to damping and to changes in wave packet volume plotted on latitude-height coordinates (the first and second terms of equation 6.14, transformed back from $s-r$ to $y-z$ coordinates). For easy comparison, we marked the three locations **a**, **b**, and **c** which are also marked on figures 6.3 and 6.2.

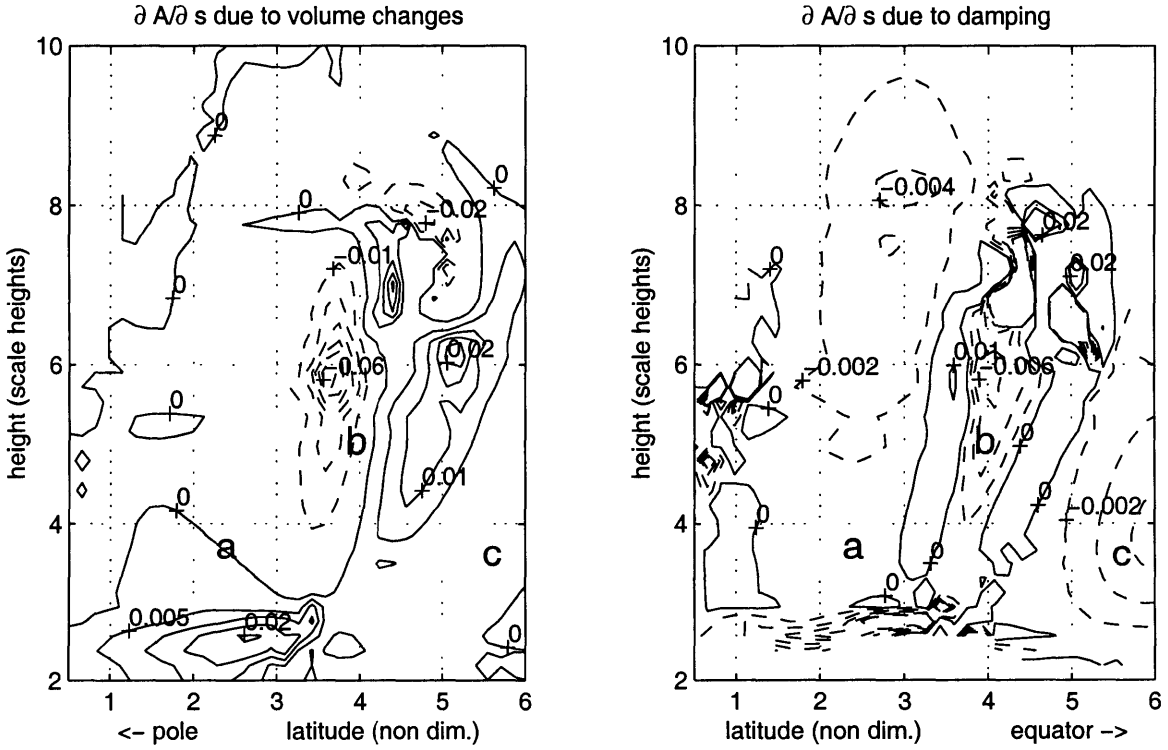

Figure 6.4: The variations of wave activity along wave packet paths due to changes in packet volume (left) and damping (right). **a**, **b**, and **c** are the same as in figure 6.3. See text for details.

The packet volume effect increases wave activity in the waveguide (roughly y=1-3) in the lowest third of the domain and decreases it above. Note that the wave activity peak marked as **a** falls exactly on the zero line. There is also a decrease in

wave activity between the midlatitude and equatorial wave regions (point **b**, y=3-4, where $\bar{q}_y$ is small or negative and the wave has to tunnel). This is due to $\vec{V}_a$ and the wave packet volume being very large there. Beyond the evanescent region ($y > 4$), $\vec{V}_a$ and packet volume are small, hence A is large (point **c**). It is important to note that in most cases we can not assume causality, only consistency. For example, the midlatitude peak in wave activity (point **a**) is a result of downward reflection (wave activity is constant in the absence of reflection), as well as the minimum near the surface (in the case of full reflection we would have a node, here we have and 'almost-node' due to partial reflection). $\vec{V}_a$ has to be very large at a 'node' of A, since the EP flux is relatively constant (see for example figure 6.1). This results in a divergence/convergence pattern of $\vec{V}_a$, that is consistent with an increase/decrease of packet volume, and a decrease/increase of wave activity density at the 'node'. We can point out to some causality in the equatorial peak of wave activity, since wave packets reach the equatorial region mostly by leaking out of the midlatitude waveguide, and not through direct upward propagation in the equatorial region. Since $\vec{V}_a$ is very large in the evanescence region (wave tunneling), wave packets increase before (poleward of) and decrease after the tunneling region. As a result, wave activity density decreases in the evanescent region and increases beyond it. This effect is larger at upper levels.

The damping effect is to decrease wave activity density in the upper and equatorial sponge regions (figure 6.4, right). We expect the contribution of damping to be mostly negative (the global integral should be negative). There are regions with large positive contributions, near the regions of tunneling. These regions, as well as the negative region at point **b** are spurious, and highlight the limitations of this diagnostic. Calculating the damping effect involves transforming A from $y - z$ coordinates to $s - r$ coordinates and multiplying by J. This calculation is messy in regions of large $\vec{V}_a$ (and J), because we are under-sampling (the $s - r$ grid is very large). This is clear in figure 6.3, both in the region of tunneling (point b and the diagonal messy line in AJ), and near the node of A (the kinks in AJ lines near the surface). The consequent derivative of AJ along s is also very noisy, and the transformation back to $y-z$ coordinates spreads the region of noise back to a large region in geometric space. These problems are not as large in the calculations of the volume effect because we do not have a transformation from $y - z$ to $s - r$ coordinates. Note that in steady state, the damping equals the EP flux divergence, and it is easier and less noisy to calculate $\nabla \cdot \vec{F}$ directly. In the next section we will discuss the case of a time varying wave, where the damping felt by a wave packet is not equal $\nabla \cdot \vec{F}$.

6.4 The wave packet propagation in a time dependent case

Observed stratospheric waves occur in episodes and are constantly varying in time (see chapter 1). When the wave field varies with time, the wave activity, the EP flux, and hence $\vec{V}_a$ and our wave-based coordinate also vary with time. In analogy to fluid flow, we have the distinction between *streamlines* (integral lines of the $\vec{V}_a$ field on a given day) and *trajectories* (integral lines of $\vec{V}_a$ following a wave packet, taking into account the time variations of $\vec{V}_a$). For illustration purposes, we use the time dependent model described in section 5.3.4 (also appendix B). We force it with a stationary wave 1 that has the same latitudinal shape as the control run forcing, and turn the forcing on over a period of 12 days and decrease it to 0.6 of its maximum by day 19 (figure 6.5 shows the amplitude of the wave at the latitude of maximum forcing, which is at the middle of the waveguide). The basic state, which is constant in time, is the same one used in the previous section (and section 5.3.1).

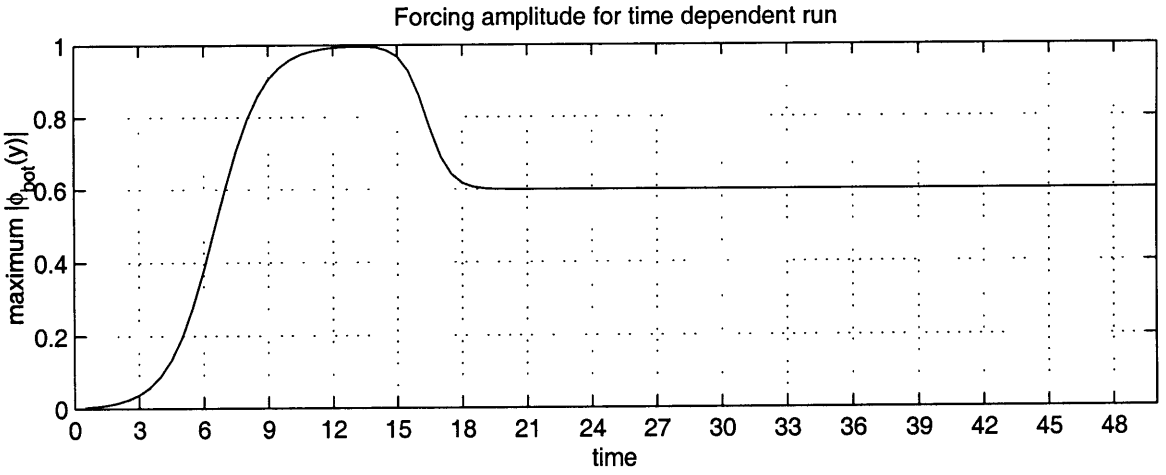

Figure 6.5: The forcing of the time dependent model used in the next four figures. Shown is the maximum amplitude (the latitudinal shape of the forcing is as the control run, figure 5.2).

Figure 6.6 shows the geopotential height amplitude, the wave activity, and the wave activity *flow lines* for a few days of the model run. *Flow lines* are the integral lines of the daily snapshots of $\vec{V}_a$ (analogous to streamlines). Looking at the wave activity density, we see the perturbation propagating up the waveguide and then 'spreading sideways'[3]. This is also evident in the wave activity flow lines.

[3]For reference, the wave eventually reaches a steady state (small oscillations in amplitude persist till about day 50), which is similar to that of figure 5.2, scaled by 0.6 due to the smaller forcing.

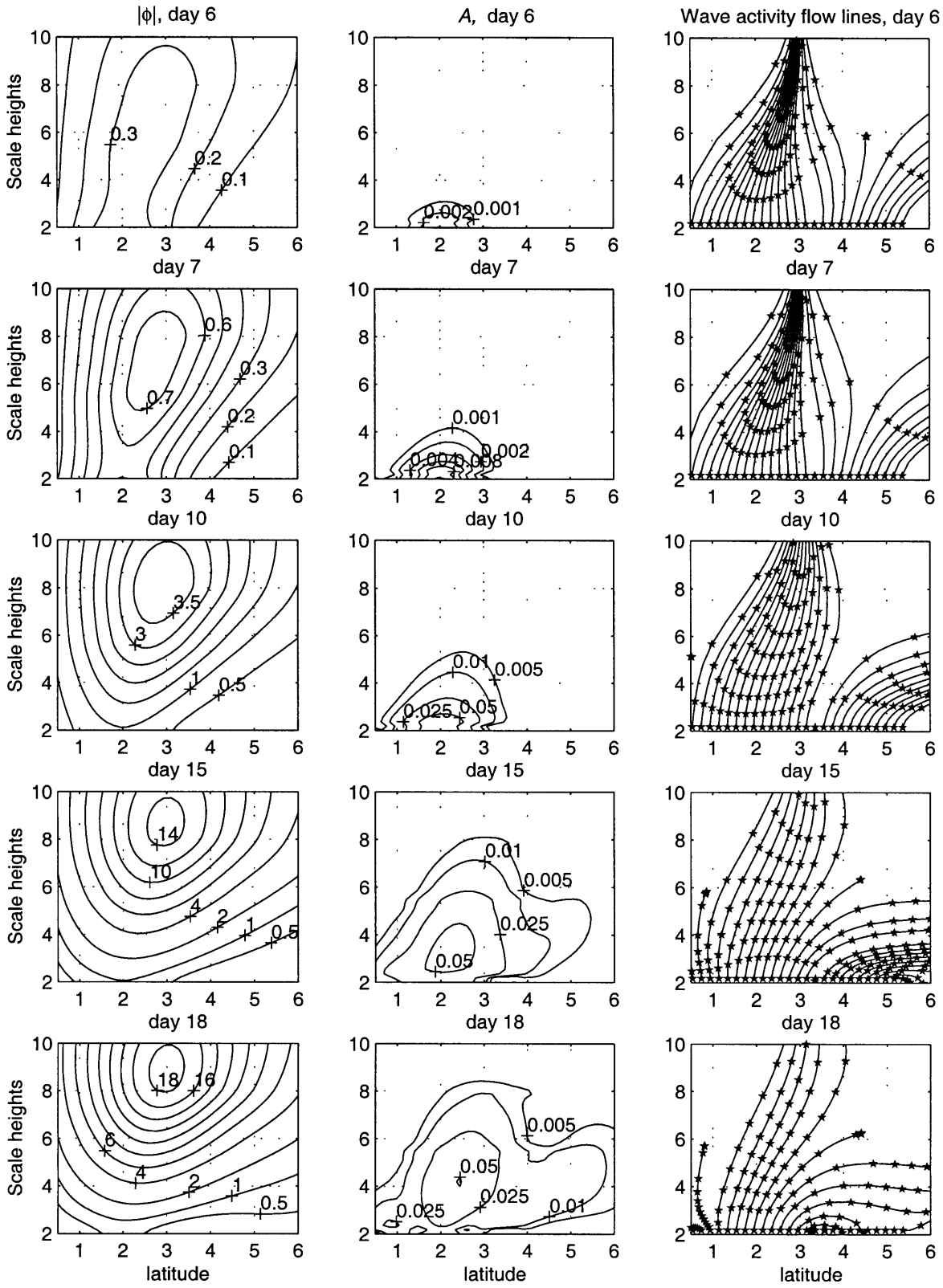

Figure 6.6: Five days (6, 7, 10, 15, 18) of geopotential height amplitude (left), wave activity density (middle) and wave activity flow lines (left, stars mark day intervals), for the model run of figure 6.5 (see text for details). Note that contour intervals are not the same for all days.

Initially, the wave field is 'not aware' of the large index of refraction at the equator, and travels up the local n^2_{ref} gradient into the middle of the waveguide. Afterwards the wave tunnels out of the waveguide to the equator and gets absorbed in the equatorial sponge layer, resulting in the wave activity flow lines tilting equatorwards. This is a leaky waveguide signature. It is interesting that Randel et al. (1987) in a study of life cycles of stratospheric planetary waves find an initial baroclinic stage (EP fluxes point upwards), followed by a barotropic stage (EP fluxes point equatorwards), suggestive of the wave evolution shown here. In section 6.6 we will show an example from observations. Note that this behavior is much less obvious in geopotential height.

The wave activity density peak increases until day 15 and decreases afterwards, due to the decrease in forcing at the bottom. The time dependence introduces an effect on the vertical-latitudinal structure of the wave. For example, on day 18, A increases with height below 4 scale heights. Apart for variations in wave packet volume and damping, part of this increase with height may be due to the vertical advection of the decrease in forcing over days 14-18. This effect will be evident if we repeat calculations of the previous section using the instantaneous wave fields. In particular, AJ contours plotted on $s-r$ coordinates calculated from a daily $\vec{V}_a$ field will not be 'vertical' (AJ will not be constant along s lines), even in regions of no damping, because of the advection of time variations in the source of A. We can get rid of this effect by 'hopping onto a wave packet'. Essentially this means repeating the exercise of integrating $\vec{V}_a$ lines to obtain $s-r$, but keeping track of the time variations of $\vec{V}_a$ as the wave packet moves along. Figure 6.7 shows *wave packet paths*, which are the paths that a set of wave packets that leave the bottom of the model on a certain day follow. We mark intervals of 1 model day by circles, and highlight the locations of packets on day 12. For example, packets that leave the bottom on day 6 (9) reach the locations marked by $\triangle$ ($\square$) on day 12. We can also combine the information from a few of these to create a plot of wave packet locations (day 12 is shown here), where each packet is tagged according to the day it left the bottom ($\triangle$ for day 6, $\square$ for day 9). Apart for being an illustrative tool for looking at the evolution of a given wave field in time (as we do in sections 7.1.1 and 7.1.4), we can use this 'dissection' of the wave field into wave packets to calculate the effects of packet volume and damping on the wave activity density of the packet as it moves along. Note that unlike the steady state case where $\nabla \cdot \vec{F}$ (which is much easier to calculate) equals the damping of A, in the time dependent case, $\frac{\partial A}{\partial t}$ has to be taken into account. Also, daily plots of $\frac{\partial A}{\partial t} + \nabla \cdot \vec{F}$ do not follow a given wave packet.

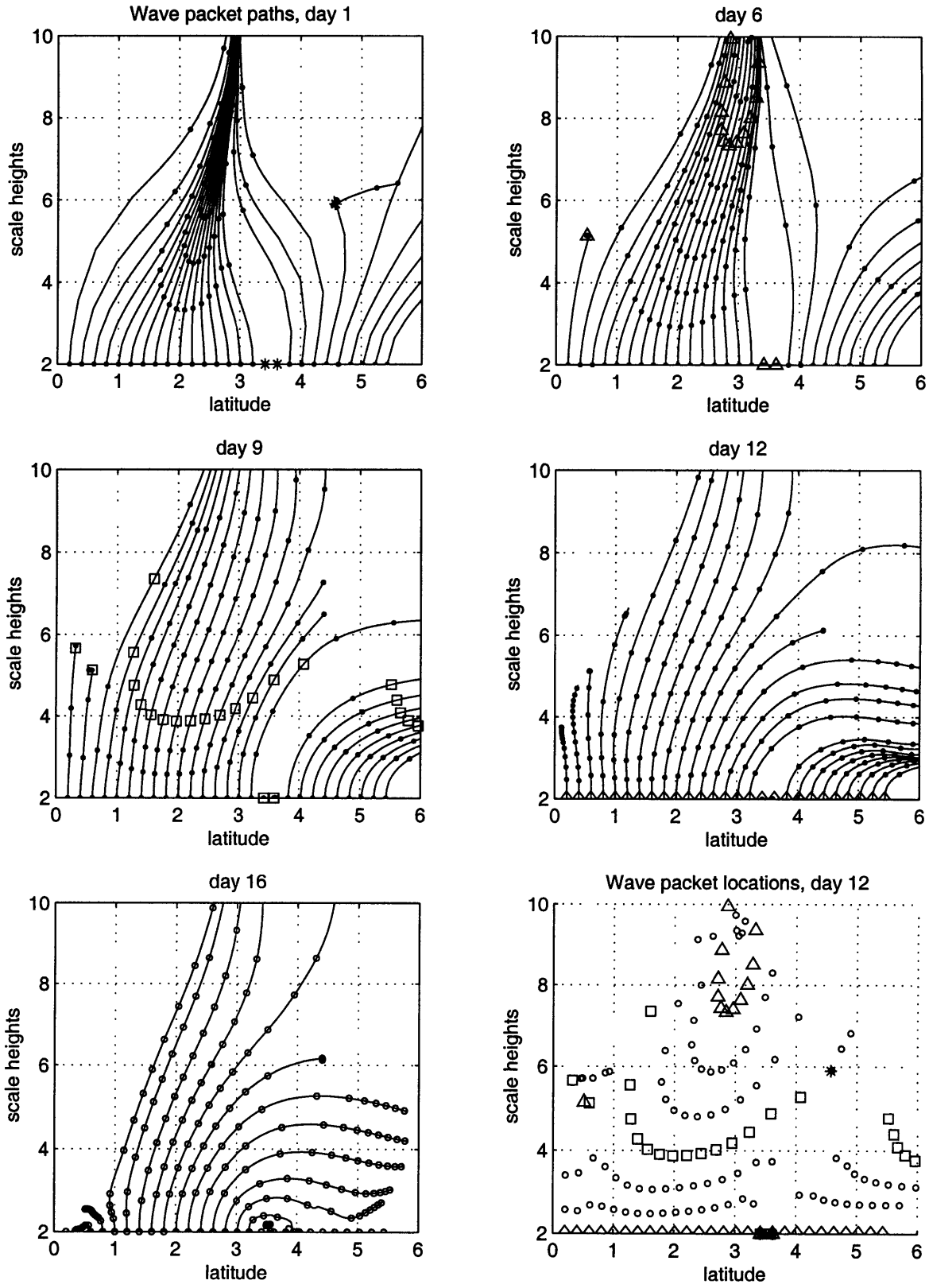

Figure 6.7: Wave packet paths (with day intervals marked), for wave packets that left the bottom on days 1,6,9,12, and 16, and the location of wave packets on day 12, for the model run of figure 6.6. See text for details.

To calculate the evolution of wave activity in a wave packet, we define the wave packet paths to be our $s - r$ coordinate system. The Jacobian of the transformation denotes the volume of the wave packets as they move along. Since s denotes the time a certain wave packet reaches a given location, we can determine A of the packet from the corresponding day and location. This is easy to do if we transform the wave activity fields of each day (actually our model output is in increments of half days but for brevity we will use 'days') onto the $s-r$ coordinate, choose from each day the appropriate value, and stack them (i.e. the composited $A(s,r)$ for wave packets that leave the bottom on day 16 is constructed by choosing the $s = 0$ row of A from day 16, the $s = 1$ row from day 17, the $s = 2$ row from day 18, etc., and stacking them in order). To illustrate, figure 6.8 shows A of wave packets that left the bottom on day 16, plotted on $s - r$ coordinates (middle). The values of A that are plotted on the $s = i$ row are taken from the $s = i$ row of the A field of day $16 + i$, where $i = 0...12$. We see that A varies as the wave packets move along. This is due both to variations in packet volume and to damping. Since we are tracking given wave packets, we do not have a contribution from the time variation of the forcing. To view in geometric space, we transform this composited A field back to $y - z$ space (left). Also plotted are the corresponding wave activity paths. The packets leave the bottom on day 16 and travel along these paths, taking one day to travel the distance between the circles, while wave activity density varies according to the contours[4]. As in the steady state case, we get rid of the volume effect by multiplying A by the Jacobian to get the total wave activity in the wave packets as they move along (shown on the right). As expected, AJ is quite constant for the first 3 days, roughly the time it takes the packets to traverse 4 scale heights. Note that AJ decreases more rapidly than in the steady state case (figure 6.3). This is because in the time dependent model runs, in addition to the sponge layers, we have a small constant damping (time scale of 25 days) to assure numerical convergence. The strong decrease in AJ near the surface is probably not real, for a few reasons. The Jacobian, which involves derivatives along the rays, is not very accurate at the bottom. Also, the resolution of A in $s - r$ space is only $\Delta s = 0.5$days, which is the time resolution of our model output, while the resolution of s is $1/8$ of a day[5]. We choose a relatively low time resolution for our model output since the observations do not have a high resolution. Also, as with the

[4]The resolution in $s - r$ coordinates is half a day. Note that the transformation from $s - r$ back to $y - z$ coordinates, which is a transformation from an irregular to a regular grid, is not well defined over roughly the lowest scale height (the distance traveled in 1 day, which is equivalent to two s grid points).

[5]As explained in appendix C, we interpolate $\vec{V}_a$ in space and time when we integrate it to get the wave packet rays.

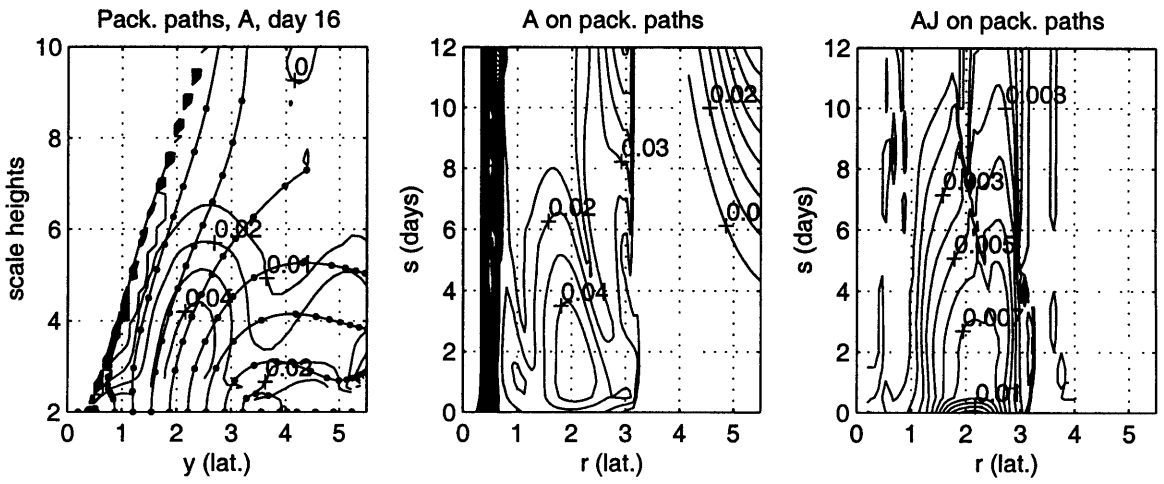

Figure 6.8: Middle: The wave activity density following wave packets that leave the bottom on day 16, plotted on $s-r$ coordinates. Left: A transformed back to $y-z$ coordinates, along with the $s-r$ coordinate (day intervals marked). Right: the total wave activity in a the wave packets (on $s-r$ coordinates). Note that the pole is at small y/r.

steady state case, the calculation is not very accurate near the surface where wave activity almost has a node.

Finally, we calculate the contribution of the volume and damping terms to the wave activity budget, as was done for the steady state case (equation 6.14, figure 6.4), and transform back to $y-z$ coordinates. Figure 6.9 shows these fields, along with the corresponding wave packet paths. Some of the features of the steady state solution are found here. The most striking is the increase in volume (hence a decrease in wave activity density) poleward (small y) of the tunneling region (y=3-4, z=3-7), and a corresponding decrease in volume (increase in A) beyond it. The magnitude of the volume effect is similar to the steady state wave. This suggests that most of the volume effect is due to variations in $\vec{V}_a$ as a result of the tunneling of the wave to the equator. The damping effect is similar to the steady state in that it contributes only negatively, but it is roughly twice as large between 4-6 scale heights in the time dependent case. This is because we added a constant damping of $0.04 days^{-1}$ to insure numerical stability of our control run. This is four times as much as the steady state damping at 4 scale heights, and half the damping at 6 scale heights (see figure 5.1).

One of the goals in developing this diagnostic was to distinguish between the different factors that affect wave activity and wave amplitude in observations, in particular to get an estimate of the amount of damping felt by the waves. Application to observations, however can be quite problematic. Before we discuss the limitations of this technique, however, we will digress a bit and show how the wave based coordinate

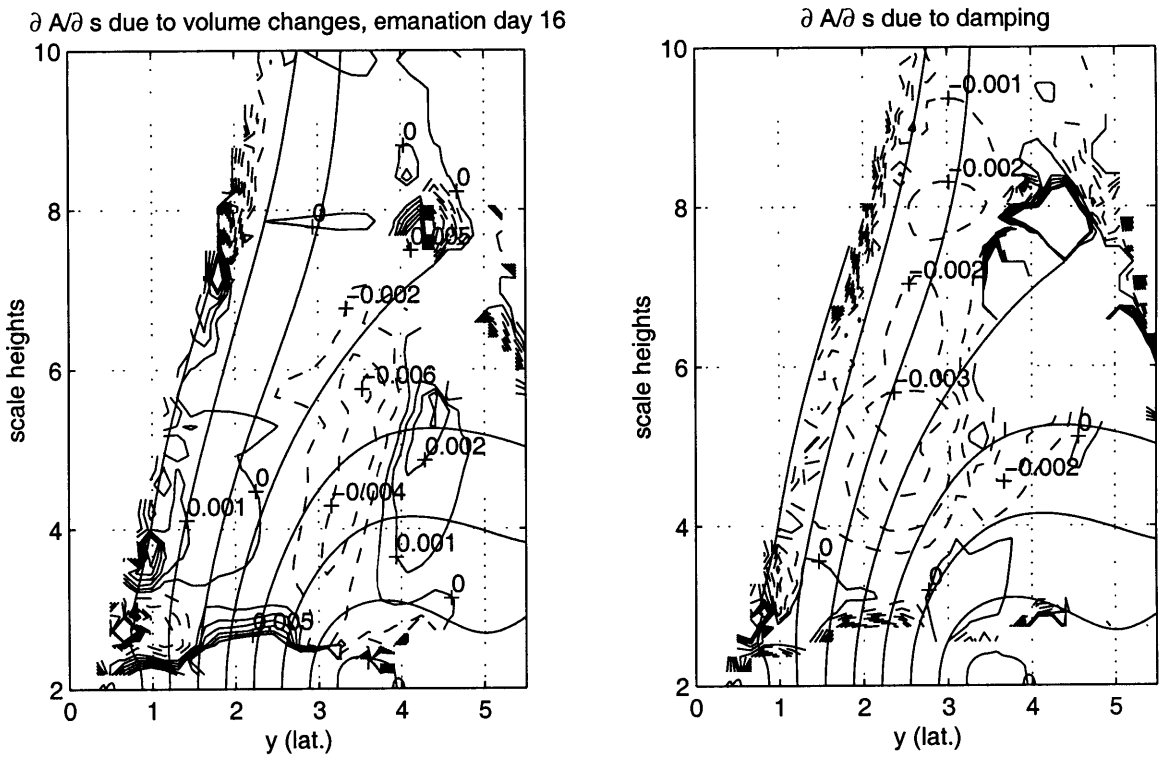

Figure 6.9: The variations of wave activity density along wave packet paths due to changes in packet volume (left) and damping (right), for packets that leave the bottom on day 16. The fields were calculated on $s - r$ coordinates using equation 6.14, and transformed back to $y - z$ coordinates. Negative values dashed. Also plotted are the corresponding $s - r$ coordinate lines (the wave packet paths, intervals of one day are marked by circles). The pole is at small y.

is related to ray tracing (as in Karoly and Hoskins, 1982).

6.5 The relation to Karoly and Hoskins' ray tracing

The integral lines of wave activity velocity are reminiscent of ray tracing calculations by Karoly and Hoskins (1982)[6], referred to from now on as KH. Since ray tracing involves integrating the group velocity, which is equal to $\vec{V}_a$ in the case of a pure plane

[6]The first to introduce ray tracing (that we know of) are Landau and Lifshitz (1959). However, Lighthill and Whitham (1955), and Whitham (1960) redeveloped the formulation unaware of Landau and Lifshitz's work. Since the original work was not widely known, Lighthill and Whitham's work was published as an expository article. While Lighthill and Whitham developed the kinematic approach to group velocity (i.e. ray tracing) for one dimensional wave propagation, Whitham (1960) extended the derivation to two and three dimensions.

Rossby wave, the two are related. The main difference, however, is that ray tracing is an analytic calculation of certain wave propagation properties on a given basic state, while our calculations are diagnostic in nature, meaning we need to have a wave field, as well as the basic state on which it travels. We will start by a short description of the ray tracing formulation of KH (see also Andrews et al., 1987, Appendix 4A):

Given a basic state, we assume a pure plane Rossby wave of the form $\phi(x,y,z,t) = \phi_o e^{i(kx+ly+mz-\omega t)}$. (k,l,m) are the wavenumbers in the (x,y,z) direction and ω is the frequency, which is related to the wavenumbers through a dispersion relation:

$$\omega = \Omega(x,y,z,t,k,l,m) \tag{6.15}$$

The dispersion relation for Rossby waves is given by equation 6.5. Wave rays are integral lines of group velocity, which is defined as:

$$(C_{gx}, C_{gy}, C_{gz}) = (\frac{\partial\Omega}{\partial k}, \frac{\partial\Omega}{\partial l}, \frac{\partial\Omega}{\partial m}) \tag{6.16}$$

The group velocities in our case are given by equations 6.6[7]. We need to know the vertical and meridional wavenumbers in order to integrate the group velocities. We can again use the dispersion relation to write down equations for the variation of l and m along a wave ray, which leaves us with the following set of equations:

$$\begin{aligned} \frac{dy}{DT} &= C_{gy} = \frac{\partial\Omega}{\partial l} \\ \frac{dz}{DT} &= C_{gz} = \frac{\partial\Omega}{\partial m} \\ \frac{dl}{DT} &= -\frac{\partial\Omega}{\partial y} \\ \frac{dm}{DT} &= -\frac{\partial\Omega}{\partial z} \end{aligned} \tag{6.17}$$

Given initial conditions y_o, z_o, l_o, m_o, at $t = 0$, we integrate this set of equations to obtain the wave rays. KH rays essentially show where wave activity will propagate if a *point source* (y_o, z_o) is put into the medium, for a given initial angle of propagation (l_o, m_o). Karoly and Hoskins (1982) calculate rays for a range of initial propagation angles, on specified stratospheric basic states and various locations of point source. They use the rays as an indication for where waves will propagate, in order to gain insight into their structure. They also show that wave rays are refracted up the

[7]Since the basic state, hence also the zonal wavenumber, phase speed, and group velocity are constant in the zonal direction (x), we will omit it from the rest of our discussion.

local gradient of n^2_{ref} (and in spherical coordinates rays refract equatorwards). The relation, however, to a given wave field, and in particular to the EP fluxes, is not straightforward, and is not really discussed in KH. We can gain some intuition by comparing KH rays with our wave packet paths, for a given point source. Figure 6.10 shows the KH rays along with $\vec{V}_a$ lines for a point source that is turned on at the bottom of the wave guide at $t = 0$, both for wave packets that leave the bottom during the initial stages of wave development (0.1 days) and for the steady state solution. The basic state is the same as used in previous sections. The circles/squares on both mark 1/2 day intervals. We also plot the KH rays superposed on n^2_{ref}, to show that the rays are reflected back and forth in the latitudinal direction along the n^2_{ref} waveguide. We see that initially, the wave field is related to the KH response quite strongly. In the lowest 1.5 scale heights, below the level of meridional reflection of the KH rays, $\vec{V}_a$ lines and KH rays are quite similar. Above that, the KH rays oscillate around the wave packet paths, which are concentrated in the middle of the waveguide. The steady state response, on the other hand, is not similar, because of leakage to the equator. There are few points to note. KH rays reflect *local* wave propagation properties, hence they do not 'tunnel' through evanescent regions. They are also not affected by damping. $\vec{V}_a$ lines, on the other hand, show the flow of wave activity in the *total* wave field, which is determined non-locally (i.e. we have tunneling through evanescent regions, and damping may affect the wave non-locally). Even when non-local effects are not present, as is the case for wave packets that leave the bottom on day 0.1 (the wave has not had time to 'feel' beyond its local surroundings), $\vec{V}_a$ at any given location is a superposition of the northward and southward components of the wave field. As a result, $\vec{V}_a$ lines concentrate into the middle of the wave guide[8]. When a point source is turned on, a wave front spreads out from the source, and propagates according to the ray equations. One caveat is that KH ray tracing is for a pure plane wave with a given phase speed, and very close to the time when the source is turned on there are many frequencies.

[8]Similar concentration into the middle of the wave guide is found even in steady state, for the $\vec{V}_a$ lines that do not refract equatorwards, in regions of vertical wave evanescence ($m^2 < 0$). In figure 6.2 (as well as 6.10), for example, $\vec{V}_a$ lines that reach the top concentrate in to the middle of the wave guide above 10 scale heights. The reason for this concentration is not completely understood.

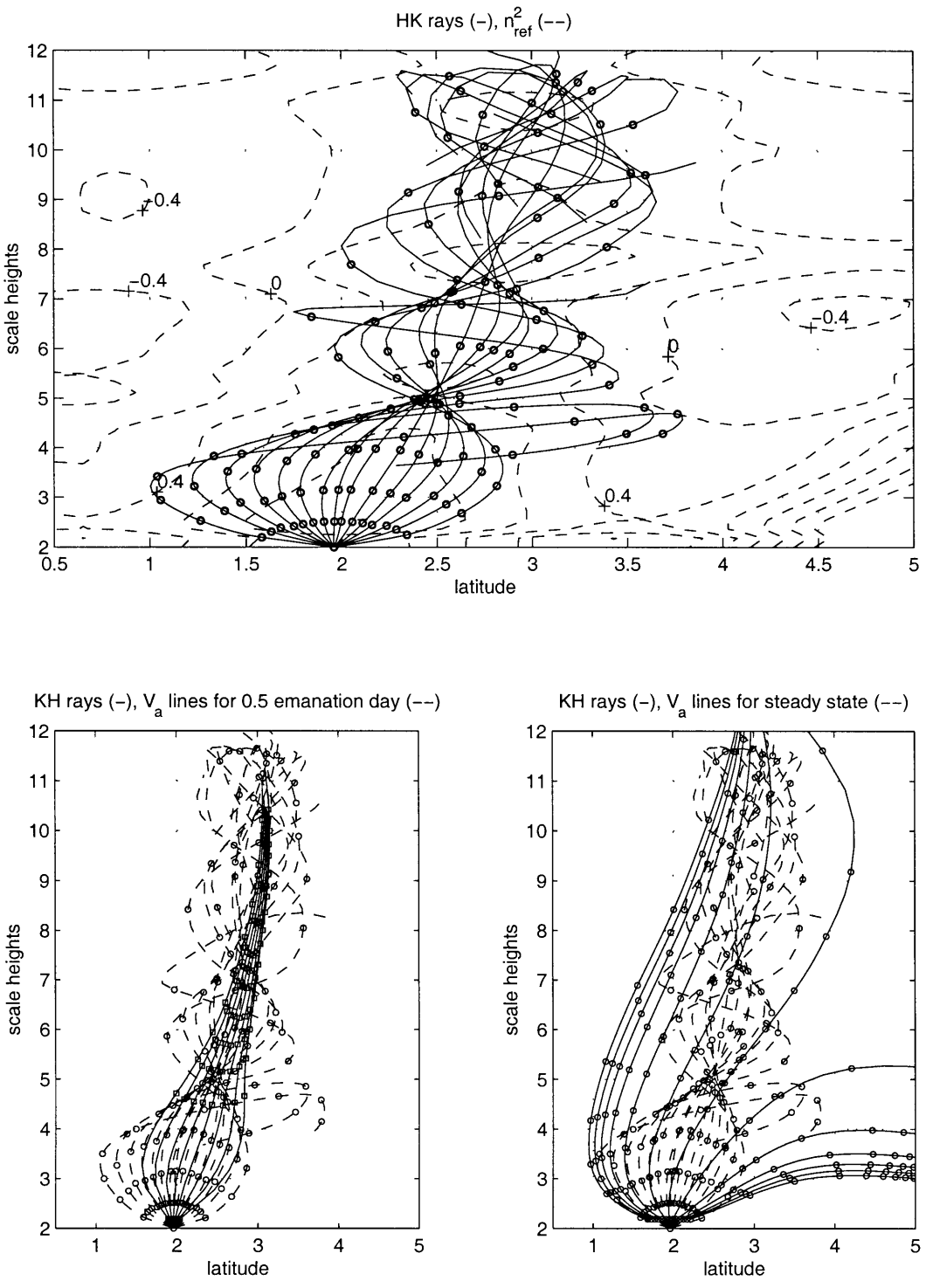

Figure 6.10: Top: Karoly and Hoskins rays (solid), superposed on the index of refraction squared (dashed). Bottom: A comparison between Karoly and Hoskins rays (dashed) and wave packet paths (solid), for wave packets that leave the bottom at day 0.1 (left) and for steady state (right), for a model run with point forcing turned on at day 0, over four days. Half-day intervals are marked on all rays and packet paths. See text for details.

This may explain why the time scales of KH rays do not match the 0.1 day $\vec{V}_a$ as well as they do later times (packets that leave the source at day 0.1 move 4.5 scale heights in the first half day, while KH rays move 0.5 scale heights). When the wave front reaches the sides of the waveguide, it reflects onto itself and the similarity between KH rays and $\vec{V}_a$ lines is lost. The relation between $\vec{V}_a$ and KH rays is even more complicated if the wave field is a response to a continuous forcing (i.e. a superposition of the response to many point sources). Finally, while ray tracing holds for highly idealized waves, for which WKB applies strictly (a small wavelength limit), the wave activity formulation is not so restricted. In fact, we do not need to assume WKB to derive the wave activity conservation equation (6.3), or to define $\vec{V}_a$.

6.6 Summary: uses and application to observations

In this section we will discuss the potential uses and limitations of our wave based coordinate and the wave activity diagnostics that stem from it. So far in this chapter we have used it as an illustrative tool, or a different approach to looking at wave structure and evolution. Another obvious application, is to observations, with two main kinds of calculations. The first has to do with tracking wave packets, and observing the evolution of the wave field in this way. In particular, we can estimate propagation time scales. The second has to do with the wave activity budget, and calculating the various terms that contribute to it.

In order to apply the coordinate diagnostics to observations, we interpolate observed geopotential height, zonal mean wind and temperature onto a high resolution grid, and then calculate the coordinate system, the wave activity and other diagnostics. In order to check the effect of low resolution on these calculations, we simulate this process using our model. We sample the geopotential height and the basic state at 18 equally spaced levels (9 of which are below 8 scale heights, the top observations level), which corresponds roughly to the resolution of the operational product. Unlike observations, the levels are evenly spaced, and we are assuming they are perfect. We then interpolate the low resolution fields back to the high resolution of the control run using a standard spline interpolation routine (which is what we use with real data), calculate A, the coordinate system and its Jacobian, and compare $A(y,z)$, and $AJ(s-r)$ to the original high resolution version (figure 6.11, compare to figure 6.3). We use the steady state model run of section 6.3. We see that the sampled-interpolated version does remarkably well in capturing the fact that AJ is

constant, except near the node in A, which is problematic in the high resolution calculations also. The errors in wave activity are more evident, because the effect of the sampling is to move the node up to the new lowest grid point, at around 3 scale heights, and to spread its effect in the vertical. We should note, that since the node of A in our control run is very close to the surface, it does not affect most of the domain, but in cases when the node is in mid-stratosphere, the errors due to low resolution sampling affect much more of the analysis. Since the sampling is even spaced, and our 'observations' are perfect, these errors should be taken as a lower bound for real observations.

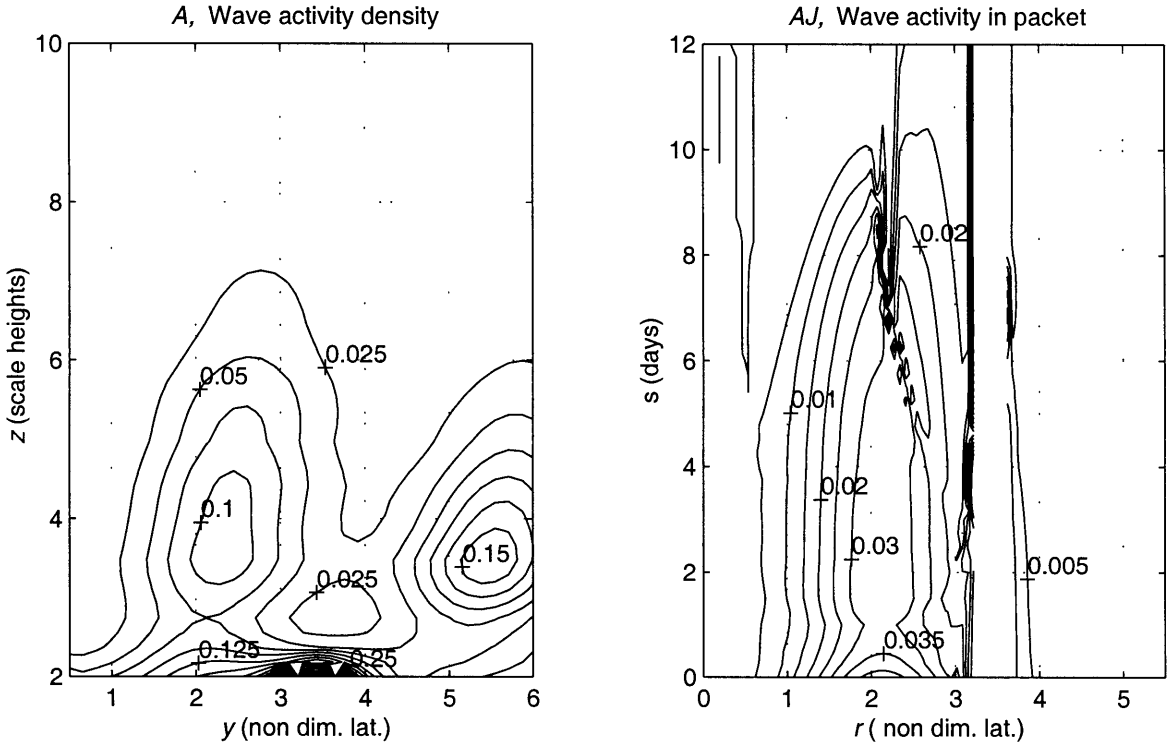

Figure 6.11: Wave activity density on $y - z$ coordinates (left), and the total wave activity in a wave packet on $s - r$ coordinates (right) for the steady state run, only using low vertical resolution sampled fields as the basis for the calculations. Compare to figure 6.3. See text for details.

We have done both kinds of calculations, tracking wave packets, their paths and travel times, and calculating the various contributions to the wave activity budget, using observations from the southern hemisphere winter of 1996. For the first kind of calculations, our diagnostic is quite useful. One feature of our model which is qualitatively found in observations is the initial concentration and eventual refraction toward the equator of the wave activity flow lines of a perturbation in its growing stages (e.g. figure 6.6). This is a signature of a leaky wave guide, hence it is interesting

to find it in observations. Randel et al. (1987) found a similar signature using time-lag correlations of wave amplitude with EP fluxes. Figure 6.12 shows an example of the wave activity flow lines for August 3-7, 1996, in the southern hemisphere. At this time the wave is growing (see figure 5.10). We see that on August 3rd and 4th the wave activity lines concentrate into a narrow region, while on the 5th and 6th, the lines are spread out and tilt equatorwards. It is important to note, however, that out of all the periods of wave growth we analyzed in winter of 1996, this was the cleanest example of such a leaky waveguide signature, and more wave events have to be analyzed in order to establish this as a characteristic behavior. In section 7.1.4 we use wave packet locations and wave activity paths to study the time evolution of the wave, and also discuss the accuracy of these diagnostics. We also use $\vec{V}_a$ lines to estimate travel times. There are existing methods for calculating time scales for propagation through the stratosphere, namely KH ray tracing discussed in the previous section and space-time lag correlation diagrams (Randel et al., 1987, Randel, 1987b). Ray tracing is a theoretical calculation, for the propagation of a wave front from a point source, which may or may not relate to actual wave propagation time scales (see sections 5.3.4, 7.1.4, and 7.1.3, where we compare the two). Time-lag correlations are statistical in nature, hence they give climatological time scales. There are, however, occasions when we want to estimate the propagation time scales of a given wave event, for example, as a consistency check on the applicability of linear theory to observations (section 7.1.4). It is important to note one limitation of our diagnostic, namely, that it reflects the wave activity flow in the total wave field, which is a superposition of upward and downward propagating components. This will result in an overestimation of vertical propagation time scales. The only way to get time scales for vertical propagation that reflect only a purely upward propagating component is to diagnose the time scales from the wave at the initial development stage, before it reflects downwards.

Calculations of the wave activity budget are more complicated, because the accuracy of wave activity observations is very low. Since the uncertainties in the calculations are too large to really make sense of them, we will not show any results here. The most widely used wave activity based diagnostic is the EP flux divergence, as a measure of damping (Edmon et al., 1980). Such estimates of damping are not very reliable, and the various operational data products may differ by a factor of two in the southern hemisphere, with the main sources of discrepancy are errors in the base level analysis and errors in the retrieval process (Miles and O'Neill, 1989, and references therein). Other sources of error are the low vertical resolution of observations, and the fact that winds are calculated from geopotential height via some balance assumption (Robinson, 1986). As much as $\nabla \cdot \vec{F}$ calculations are inaccurate,

they give much smoother fields than our coordinate system. Advantages of our diagnostic are that we can follow a wave packet and keep track of its wave activity, and distinguish between the various factors that contribute to it. One of the main problems we find is that the coordinate system itself is quite messy, more so during later stages of wave development, when there is downward reflection. The transformation itself is ill defined in times and places where the coordinate lines cross each other. This renders many days of observations useless for our calculations. During the growth stages of a wave episode, $\vec{V}_a$ lines are quite well behaved in a large range of latitudes. Corresponding AJ calculations, however, are very hard to interpolate, because in general AJ varies along packet paths (at the very best it appears constant for a day or two, which is meaningless given our time resolution is of one day). It is then hard to say if variations in AJ are due to real damping or to errors in the data. One limitation appears to be that we only have daily time resolution. To test the effects of daily sampling, we calculated our coordinates and the corresponding wave activity budget using only daily output from our model run. The results we get are quite similar to the half-day sampling of our control run. This suggests undersampling in time is not in itself a problem, however, we should note that wave activity density in our model is quite smooth and does not change much on daily time scales. In observations, wave activity density varies quite a lot on daily time scales, but we are not sure if this is due to a real variation or to large errors in the wave activity fields. Since wave activity density is a messy field, small shifts in the wave position can cause large local variations with time. A possible way to gain more insight into the need for more temporal resolution (which we leave for a future study) is to apply the diagnostic on an assimilation product that has 6 hour time resolution, but this involves model uncertainties. Disregarding these practical problems (we can always apply this diagnostic to a GCM), we still have not found a practical use for knowing the daily wave activity budget of a given wave event (this whole diagnostic started as a thought experiment and an illustrative tool to view waves differently), although we suspect there are dynamic situations where it is useful to be able to distinguish between effects of wave packet volume changes, time variations in the source for wave activity, and damping. A potential use (only speculative at this point) is in explaining the budgets of other quantities that may be affected by the waves on short time scales, for example chemical concentrations.

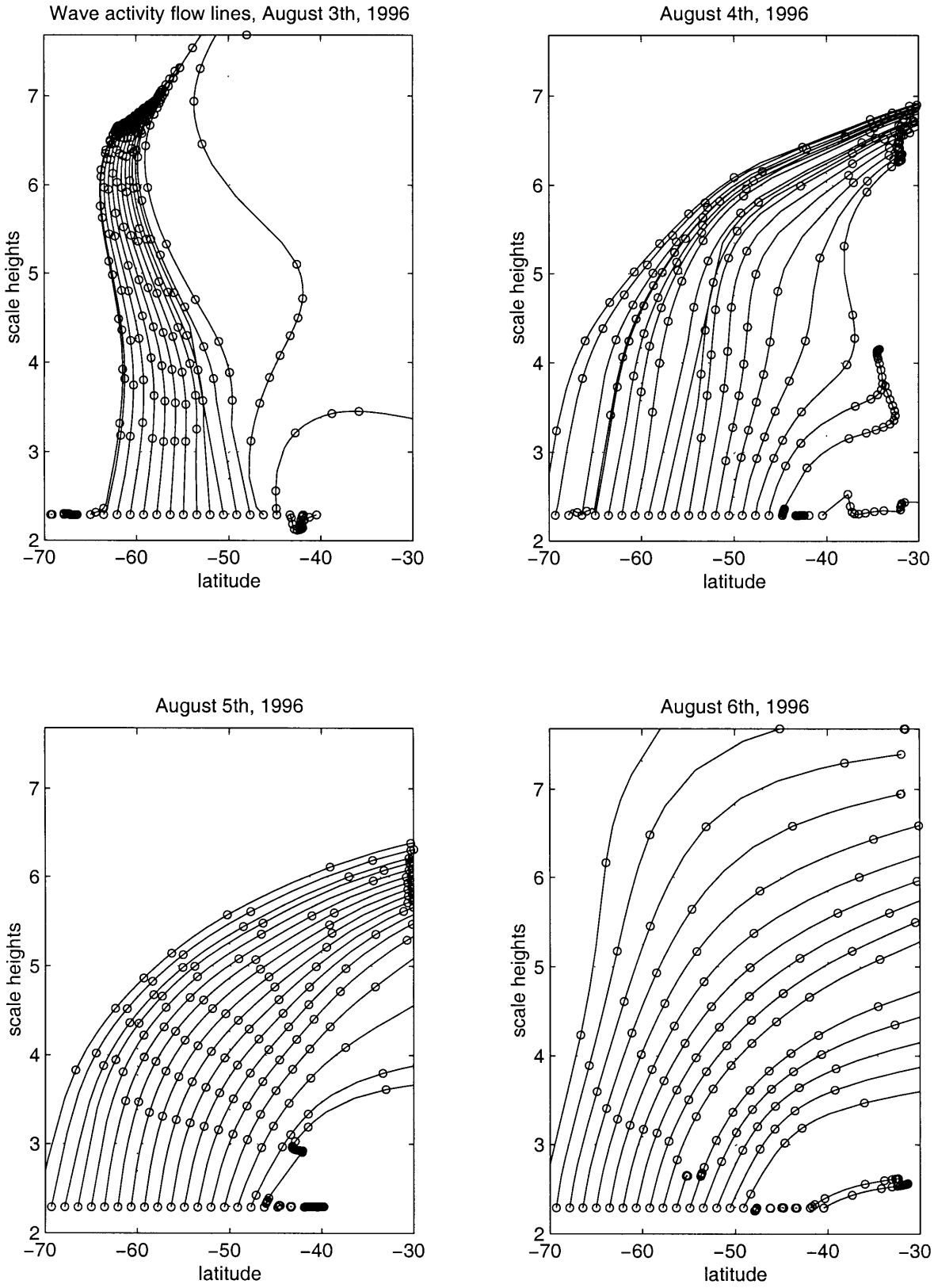

Figure 6.12: Wave activity flow lines for wave 1 in the southern hemisphere, on August 3-7, 1996 (during the growth stages of the wave). See text for details.

Chapter 7

Applying the diagnostics: explaining observed variations of wave structure on daily time scales

In the following sections we will use the various diagnostics developed so far to study in detail the relations between observed variations in vertical wave structure and variations in the basic state and tropospheric forcing. The motivation is both to gain some understanding of the linear transient evolution of the stratospheric waves and to demonstrate the use of the diagnostics themselves. We will show how the index of refraction, which is calculated assuming a steady state wave is relevant for the transient evolution of the waves, and in particular, how it is relevant to the evolution of specific observed waves. This increases our confidence in both the relevance of the theory to the atmosphere and in the observations themselves.

7.1 Wave 1 event of July-August 1996

Figure 7.1 shows the time series of wave 1 geopotential height amplitude, zonal mean wind and its acceleration, and the EP flux divergence term in the zonal momentum equation (equation D.11), for the period of July 18th-August 19th 1996, averaged over the latitudes 40-80°S[1]. The amplitude of wave 1 was shown in figure 5.11, and is shown here again to facilitate the comparison with the other quantities plotted. There

[1]We have taken a latitude average of these quantities, rather than show their value at a specific latitude, to account for variations due to latitudinal shifts of the jet. This is more important for the acceleration term because the averaging will distinguish between true net acceleration and large deceleration/acceleration dipole patterns that are associated with the jet shifting position without changing its strength.

are strong decelerations of the zonal mean wind at the end of July and in mid-August. A comparison of the EP flux divergence and the observed acceleration shows a strong relation between the two, with the former being much larger and preceding roughly by a day or two. The fact that the $\nabla \cdot \vec{F}$ term is much larger makes sense because part of it goes to driving a mean meridional circulation. During both deceleration periods the amplitude of the wave starts decreasing roughly when minimum winds are reached. Figure 5.11 shows that along with the decrease in amplitude there is a change in the vertical structure of the wave. Figures 7.2 and 7.3 show the evolution of the vertical wave structure of geopotential height and temperature at 60S during these two periods[2]. We show these two figures, in spite of the fact that they do not add any information to what is shown figure 5.11, because it is easier to visualize the structure changes from them. We see that the wave has a structure of an upward propagating wave (westward phase tilt with height) during most of the wave episode (before July 7/29, August 5-11). At the time of maximum deceleration and a few days afterwards (7/31-8/2, 8/12-16), the phase of the wave tilts into the vertical (characteristic of a standing wave in the vertical), and eventually tilts eastward with height (characteristic of a downward propagating wave).

[2]As is evident from figure 5.11, the structure variations occur over a wide range of latitudes, and not just at 60°S.

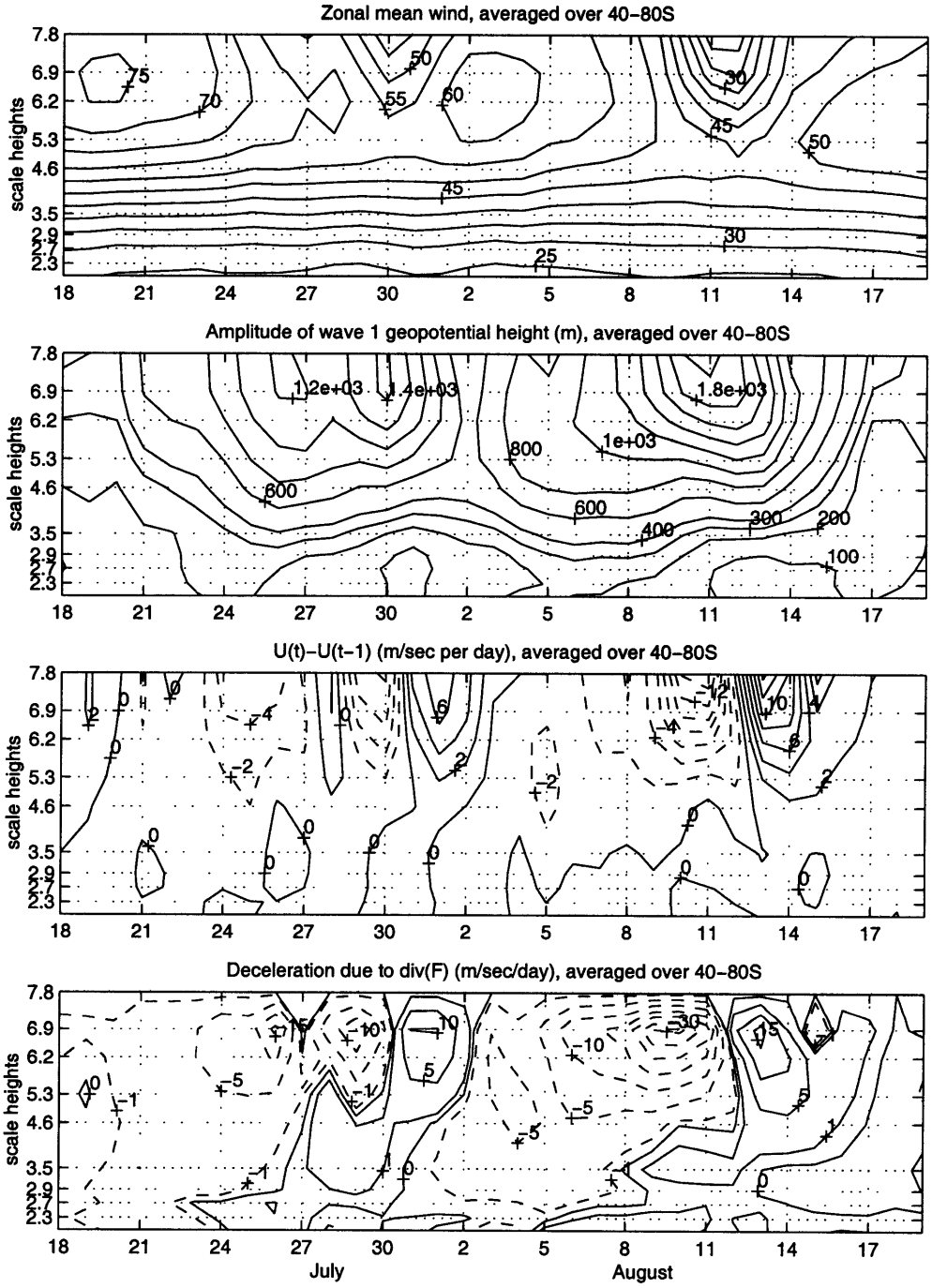

Figure 7.1: Longitude-time sections (July 18-August 19, 1996) of the 40-80°S average of (top to bottom): A. Zonal mean wind (contour interval of 5 m/sec). B. Wave 1 geopotential height amplitude (contours at 0,100,200:200:2000 m). C. The change in zonal mean wind over 1 day (U(t)-U(t-1)). Contour interval is 2m/sec, negative values dashed. D. The acceleration due to wave driving: $\frac{\nabla \cdot \vec{F}}{a_e \rho \cos\varphi}$. Contours at $\pm 0, 1, 5 : 5 : 30 m/sec/day$, negative values dashed. All quantities, except the wave geopotential height amplitude are volume averaged over latitude (weighted by $\cos\varphi$).

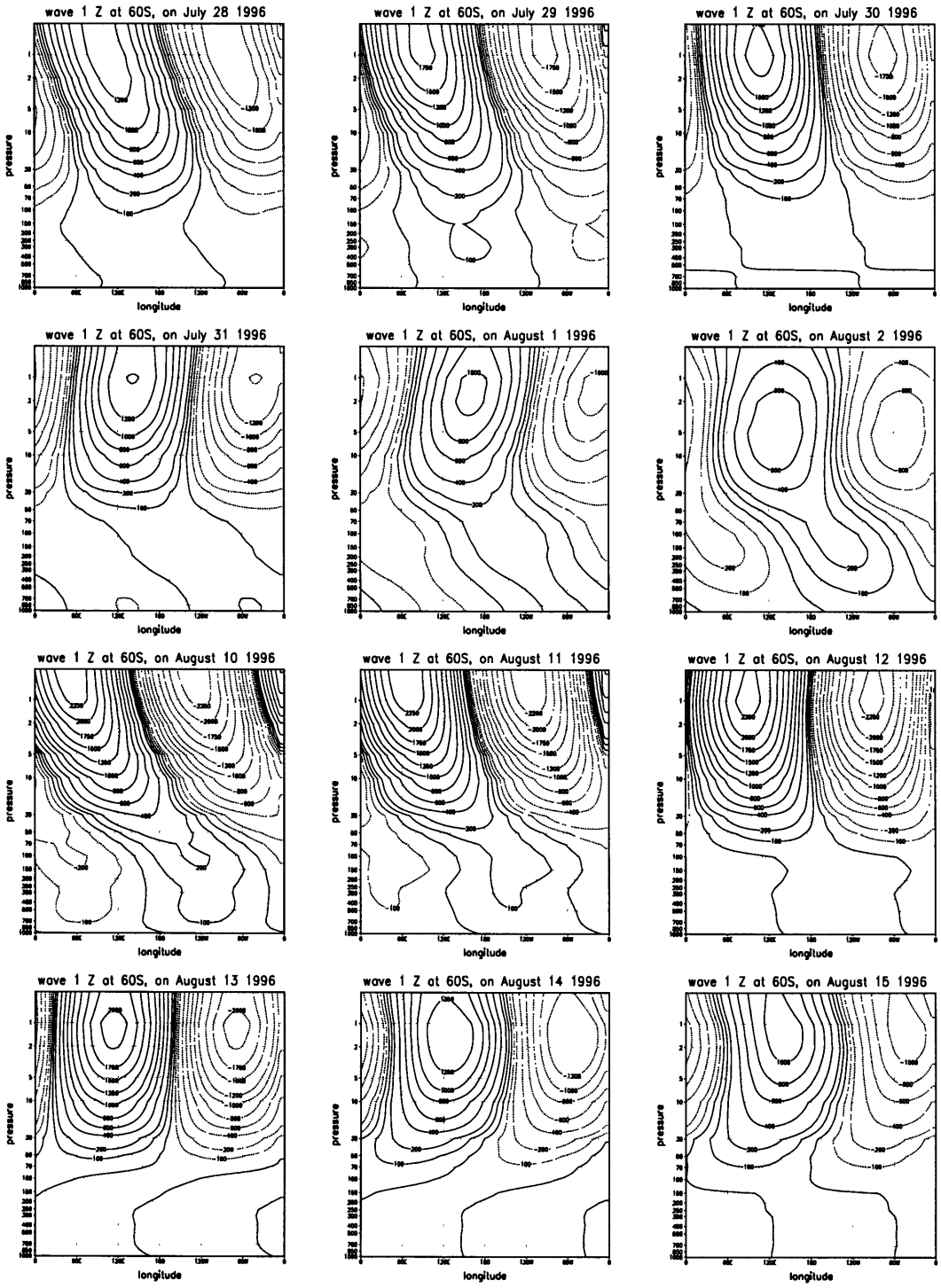

Figure 7.2: Daily longitude-height cross-sections at 60°S of wave 1 geopotential height for July 28-August 2 and August 10-15, 1996. Contour intervals are at 0, ±100, ±200, ±400, ±600, ±800 , ±1000:250:2500. Negative values are dashed.

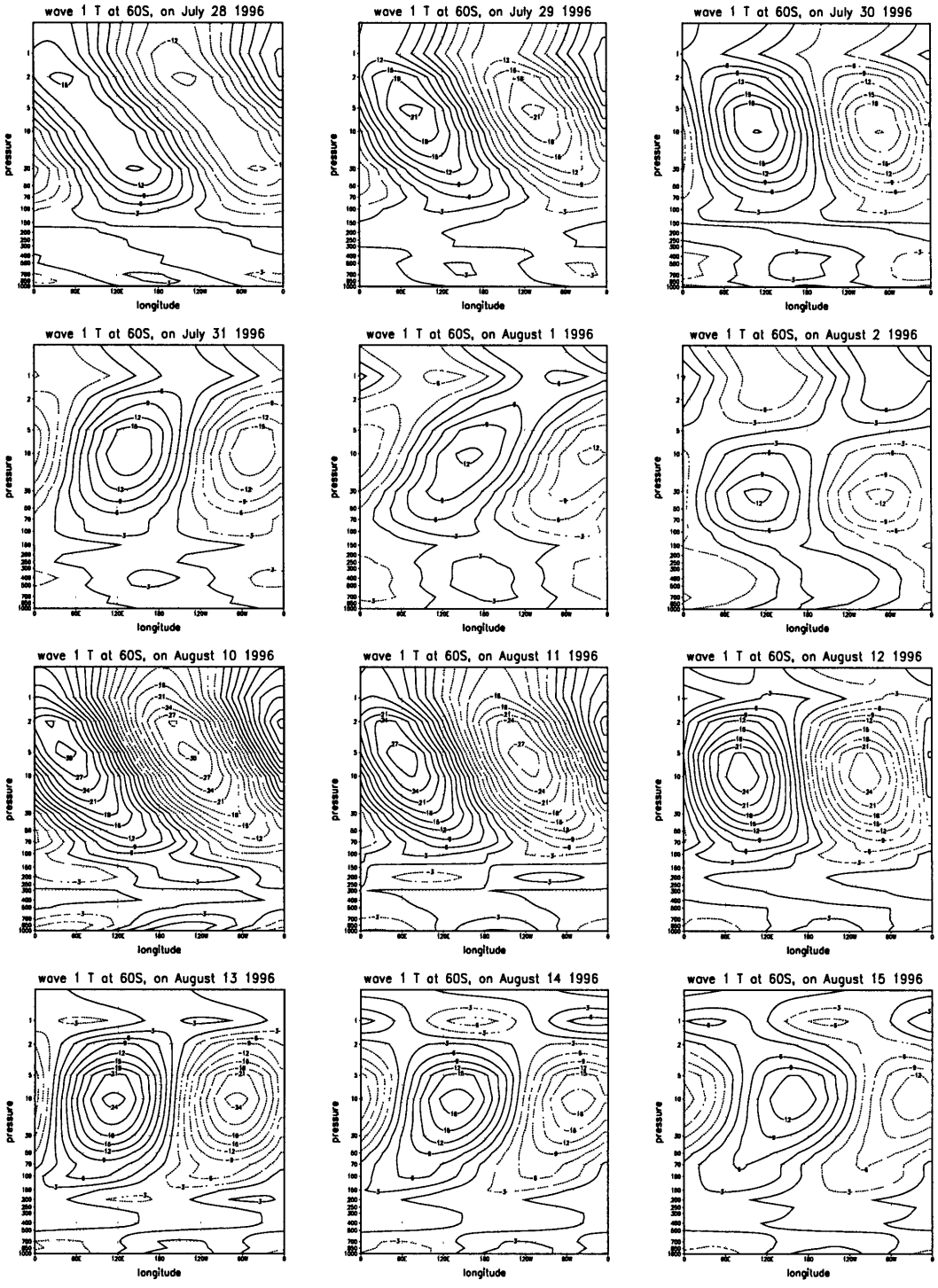

Figure 7.3: As in figure 7.2, only the temperature perturbation, with a contour interval of 3°K.

7.1.1 The formation of a turning point and its effect on wave structure

We will now proceed to show that the observed time evolution is a wave-mean flow interaction, where the wave responds qualitatively linearly to the basic state changes (which appear to be wave-induced), and that the time evolution of both seem to be dynamically consistent. Figure 7.4 shows the zonal mean wind, meridional PV gradient ($\overline{q}_y$, equation D.8) and index of refraction squared (n^2_{ref} equation D.14) for stationary wave 1, on August 8th and 11th. The strong deceleration on August 8-12 results in the formation of a region of negative $\overline{q}_y$ and n^2_{ref} in the upper stratosphere, between 55-65°S, on August 11-12[3]. As was shown in chapter 3, the top observation level, (0.4 mb, 7.8 scale heights) cannot be trusted because it is above the highest weighting function. We believe, however, that the deceleration and the formation of negative PV gradients are real features, because they are observed at and above 5mb (the second highest weighting function, which is relatively reliable). Also, the coincidence with consistent variations in wave structure encourages us to believe at least the qualitative nature of the observations.

In a one dimensional model, the formation of a region of negative n^2_{ref} would lead to downward reflection. The present case is more complicated, since the PV gradients become negative only in a midlatitude region, and essentially the waveguide in the upper stratosphere splits into poleward and equatorward branches (the latter is more pronounced). It is unclear if this would cause the wave to reflect downward, or to bypass the negative $\overline{q}_y$ region and propagate up one or both of the branches of the split waveguide, and how this would affect the vertical structure.

One way to test the propagation characteristics of a given basic state was introduced in chapter 5. We showed that the meridional and vertical wavenumbers calculated from the steady state response to forcing of a given basic state are an indication of the propagation characteristics of waves on this basic state, relevant both to their transient evolution[4] and to their steady state structure.

Figure 7.5 shows the meridional and vertical wavenumbers calculated from the steady state response (also shown) to a steady forcing that is constant with latitude, for the basic states of figure 7.4. Regions of evanescence are shaded. The model we use

[3]The regions of negative $\overline{q}_y$ and n^2_{ref} are later on reduced by the strong acceleration in the upper stratosphere on August 13-17. The early August basic state, however, is not regained, rather we see a slow climatological transition to the September-October basic state (see 5.4.2).

[4]Note that by *transient* we mean the time dependent response to time dependent forcing of a given zonal wavenumber and phase speed. This has to be distinguished from traveling waves which have a non-zero zonal phase speed (and may have a constant structure with time).

is the spherical steady state model described in 5.4.1 and appendix B. The vertical wavenumber indicates clearly the formation of a reflecting surface ($m^2 = 0$ surface) on August 11th, which is not present on August 8th. The meridional wavenumber on August 11th has a very clear split waveguide structure, while on the 8th it looks like a waveguide about to split in to two in the upper stratosphere. These changes in the basic state seem to explain the observed structure changes of the wave. This is not so clear when we note that the turning surface on August 11th, which is as low as 6 scale heights, is only 1 scale height lower than the turning surface on some days when the observed wave tilts westward with height (e.g. July 29th, figure 7.6). This westward tilt (implying only partial downward reflection), in spite of the existence of the turning point at 7 scale heights (assuming the observations are sufficiently good), indicates the presence of damping at or above the reflection surface, and/or a higher region of propagation beyond the domain of observations. In our *steady state* model runs we never get a pure vertical standing wave. The vertical wavenumber is always large enough to feel at least the top sponge layer, resulting in some vertical propagation and westward tilt with height (see figure 5.5). A slow formation of the turning point will not cause a tilting into the vertical, if the transition is slower than the relevant damping time scale. The abruptness of the basic state changes is therefore important.

The importance of transience is highlighted by looking at the earlier deceleration event (which is weaker by almost 50%). Figure 7.6 shows the PV gradients and the steady state meridional and vertical wavenumbers of the basic states of July 29th and August 1st. The PV gradients and the meridional wavenumber show that on July 29th, there is a well defined, vertically oriented waveguide. The deceleration causes this waveguide to split into two branches (August 1st). The vertical wavenumber, on the other hand, does not change as dramatically as in mid-August, and the changes are opposite to what is expected. On July 29th, there is a turning surface at 6.9 scale heights, that spans the latitudes of the waveguide. On August 1st, on the other hand, the region of negative m^2 is confined and is roughly between the two branches of the waveguide, allowing vertical propagation at least in the equatorward waveguide. While in mid-August we expect to get strong downward reflection as a transient response to the basic state changes, it is not so clear in the end of July. The question is will the transient response to the observed splitting of the waveguide lead to a temporary downward reflection of the wave?

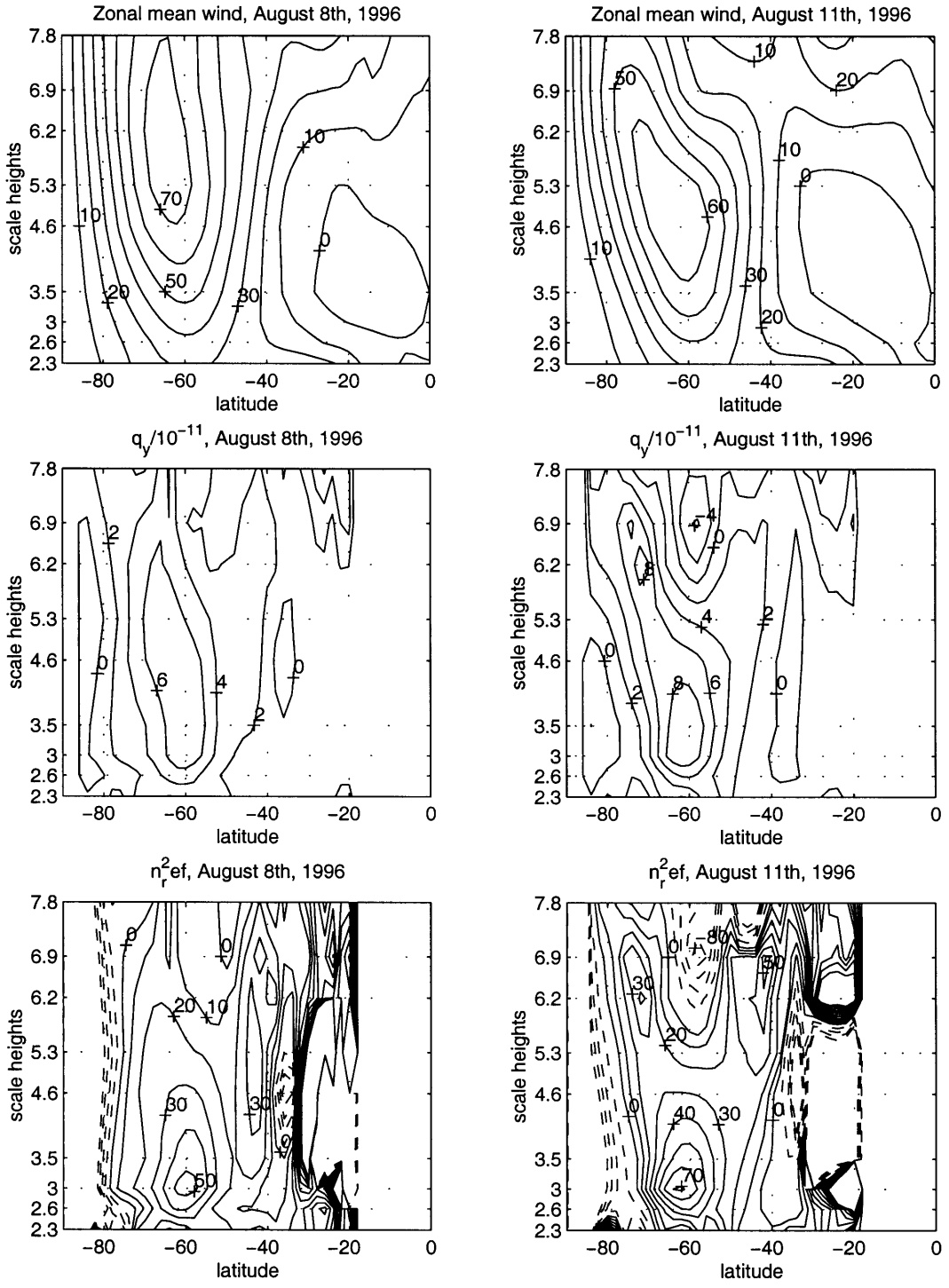

Figure 7.4: Top: Zonal mean wind (contour interval of 10m/sec). Middle: Meridional PV gradient (units of $10^{-11} sec^{-1} m^{-1}$, contours at -1,0,2:2:8). Bottom: Index of refraction squared (nondimensional, see equation 5.10, contour interval is 10, negative values are dashed), on August 8th (left) and 11th (right), 1996.

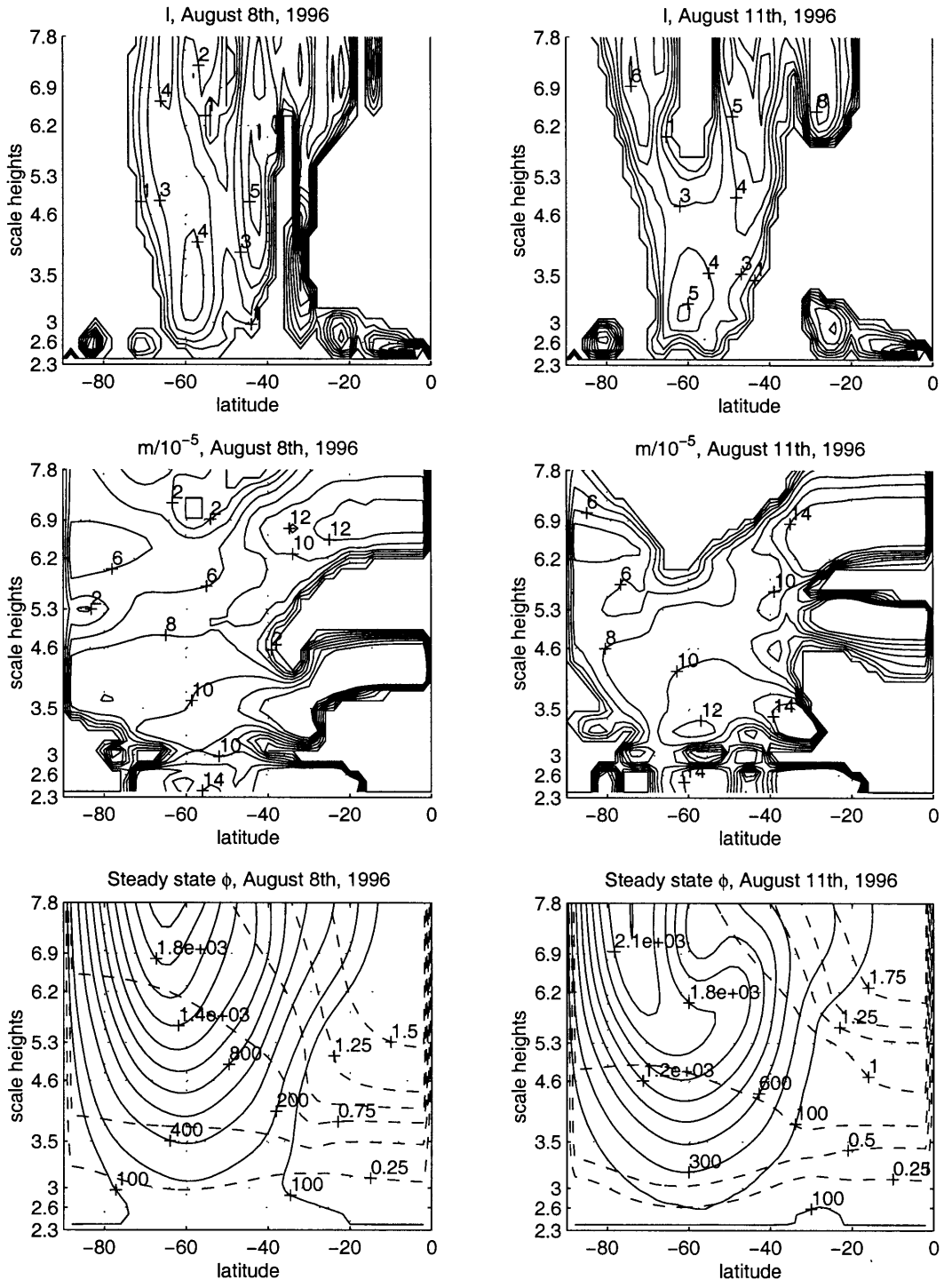

Figure 7.5: Latitude height sections of the steady state solution to the observed basic state on August 8th (left) and 11th (right), 1996. Top: Meridional wavenumber (contours at 0.01,1:5 $a_e^{-1}m^{-1}$)). Middle: Vertical wavenumber (contours at 0.01,2:2:20 $10^{-5}m^{-1}$). Bottom: Wave 1 geopotential height amplitude (solid, in meters) and phase (dashed, in units of π). Regions of evanescence (negative l^2, m^2) are shaded.

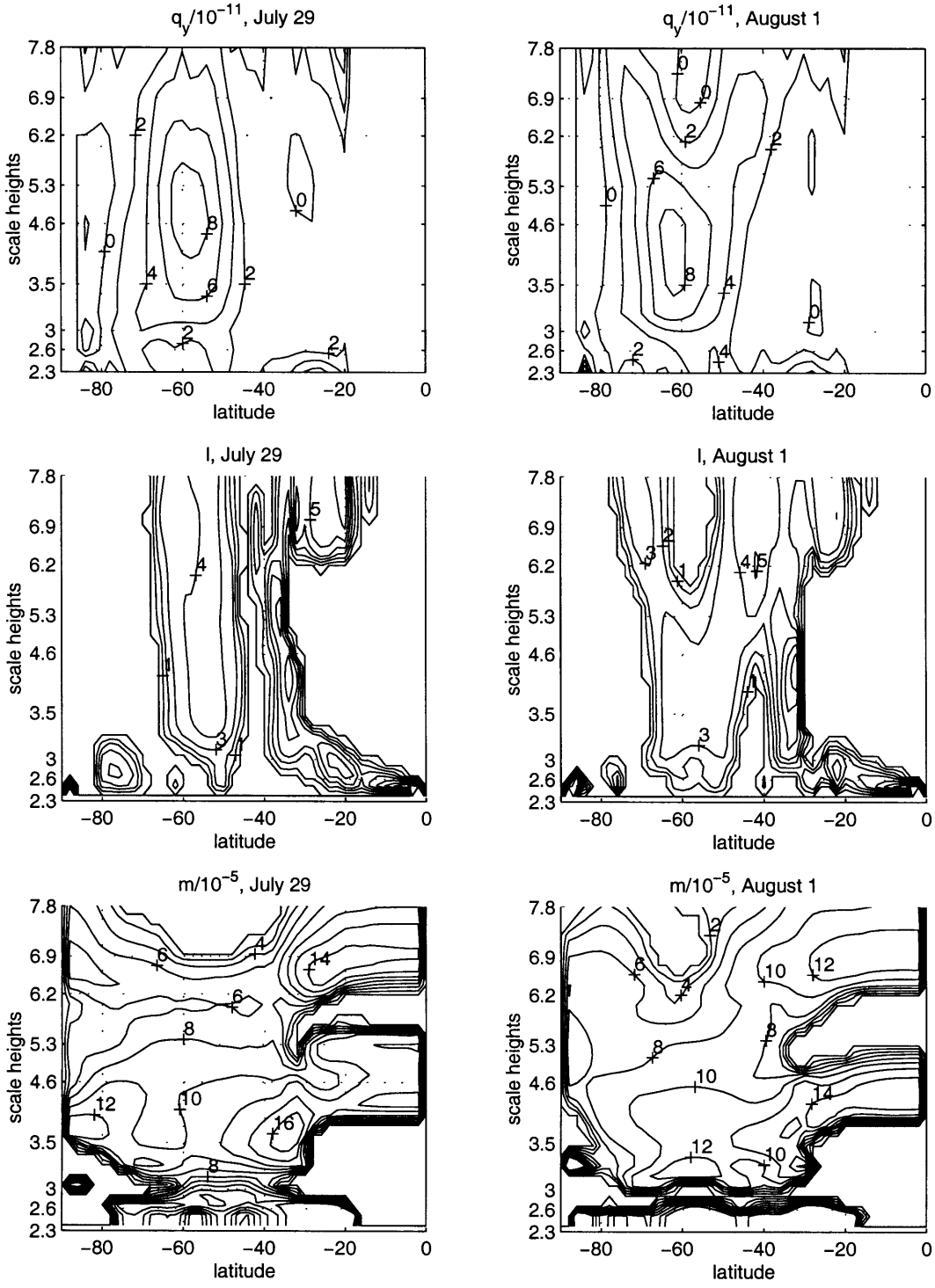

Figure 7.6: Latitude height sections of the steady state solution to the observed basic state on July 29th (left) and August 1st (right), 1996. Top: Top: Meridional PV gradient (units of $10^{-11} sec^{-1} m^{-1}$, contours at -1,0,2:2:8). Middle: Meridional wavenumber (contours are 0.01,1:5 $a_e^{-1} m^{-1}$). Bottom: Vertical wavenumber. Contours are 0.01,2:2:20 $10^{-5} m^{-1}$. Regions of evanescence (negative l^2, m^2) are shaded.

To check this we run our time dependent model with a zonal mean wind that changes in a way similar to the observed. The initial state is specified analytically to have a well defined vertically oriented waveguide, and the final state is calculated by adding the observed deceleration between August 11-13, above 4.5 scale heights[5]. We vary the wind linearly between the initial and final states, and initialize the model with the steady state response to the initial wind, while keeping the bottom forcing constant. The wind starts changing after model day 12.5 and reaches the final state on model day 17.5. Negative PV gradients appear on day 16. Figure 7.7 shows the initial and final zonal mean winds, along with n^2_{ref} and the meridional wavenumber calculated from the steady state response. We see that as the zonal mean wind changes, the initially vertically oriented waveguide shifts poleward in the upper stratosphere, effectively forming a 'turning surface' at a range of latitudes. There are no striking variations in the vertical wavenumber between the two states (not shown), making this scenario more like the early winter (July 29-August 1st) observed event. Note that our model is on a β-plane, hence there are some differences in the relation between a given wind field and the index of refraction. As a result, even though we use a deceleration similar to observed in mid-August, the changes in the model basic state are more like the earlier deceleration event. Apart from this difference in the dependence of the index of refraction on the basic state, we expect the waves to behave qualitatively the same in the two coordinate systems (see section 5.4.1 for a more detailed discussion).

Figure 7.8 shows zonal-height sections of the wave at 60S at a succession of times, as it responds to the changes in the basic state. The phase shift with height decreases with time, such that on days 17-18, the wave has a barotropic structure (phase lines are vertical) above about 3.5 scale heights, after which the phase tilt increases and readjusts to the final westward tilting steady-state solution. The temperature field in this run changes its structure mostly above 4 scale heights, where a minimum forms at six scale heights on day 17, and eventually becomes a second peak. The main feature is the cutting-off of the amplitude at the height of the turning point. The details of the temperature structure above this are dependent on the damping we use in the model. The corresponding observed evolution of temperature (figure 7.3) shows a similar cutting-off of the amplitude roughly at the height of the turning point. We also see the formation of a second peak above (e.g. August 14, at 1 mb), however, we believe that at best, this small scale feature is a distorted representation of a real

[5]We do not use observed basic states because they are noisy and our model is a β-plane, which means the propagation characteristics for a given wind field are different. Rather, we specify winds such that the propagation characteristics are qualitatively like the observed.

feature which is relatively small in its vertical extent.

It is interesting to look at the 'wave activity flow lines' (lines tangent to $\vec{V}_a$, 6.10) and the 'wave packet paths' (paths that wave packets follow, calculated for a given initial day and height) which were introduced in section 6.4. Figure 7.9 shows the flow lines (left) for three days. Before the deceleration (model day 13), the wave activity flow lines concentrate into the initial waveguide. As the winds change, they split around the region of negative n^2_{ref}, with most of the flux going up the poleward waveguide (model days 15, 50). Also shown are the wave packet paths (right) for three emanation days. The stars are separated one model day apart. We see that before the basic state changes the packets move upwards, but after the changes, wave packets slow down completely, for about one day, before moving polewards or equatorwards up one of the newly formed waveguides. It is interesting how the packets that left the bottom on day 10 reach 6 scale heights before they split sideways, on day 16, while the packets that left the bottom on day 13 reach only 5 scale heights, and split sideways on day 17. This behavior indicates that the response of the wave field moves downwards in time, as expected in a downward reflecting wave.

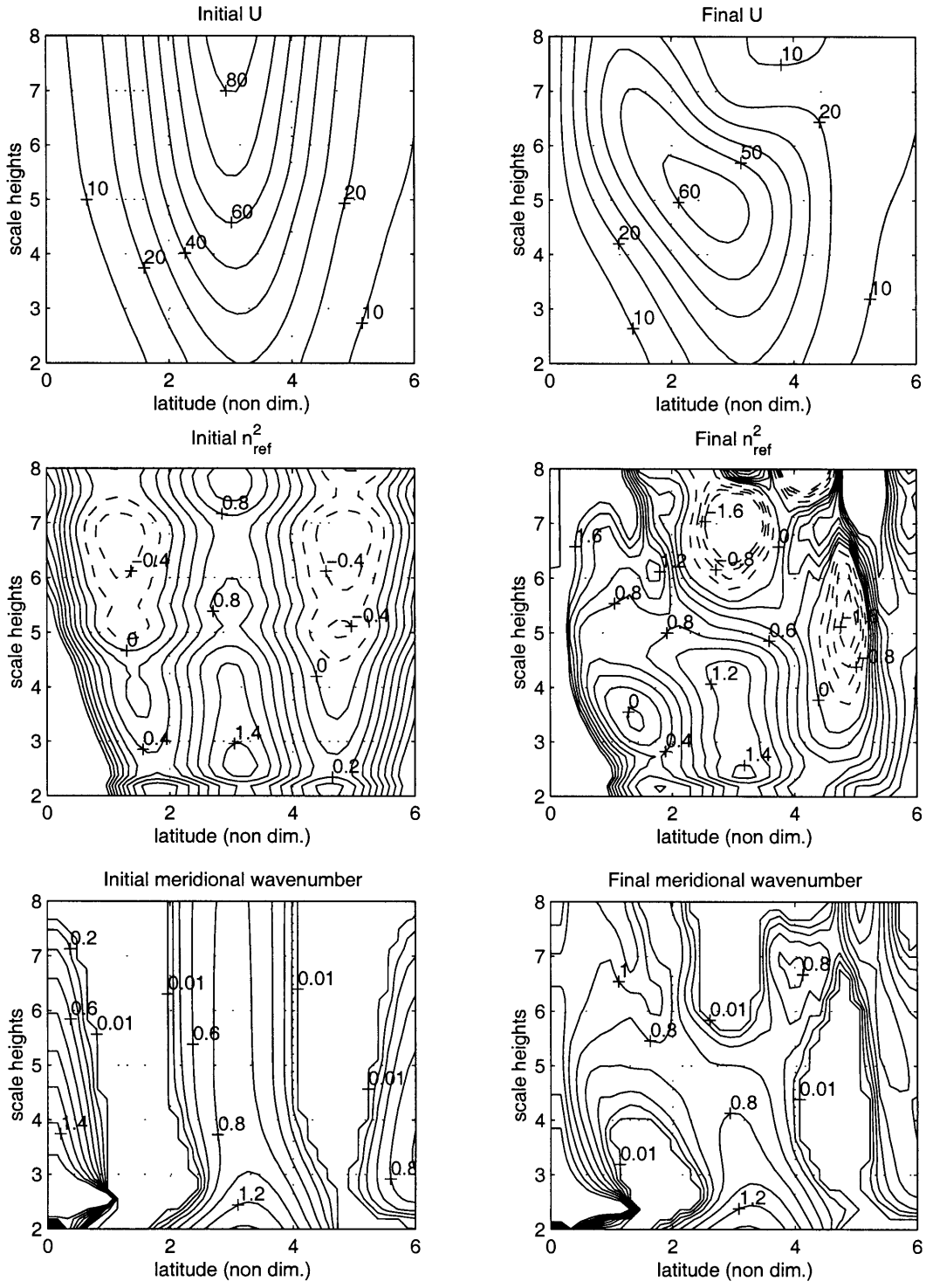

Figure 7.7: Characteristics of the initial (left) and final (right) basic states of the model run described in the text. Top: Zonal mean wind. Middle: Index of refraction squared for stationary wave 1 (negative values are dashed). Bottom: The meridional wavenumber of the steady state response. Only regions of propagation ($l^2 > 0$) are contoured (within the 0.01 contour line). Latitude increases equatorwards (0 is the pole).

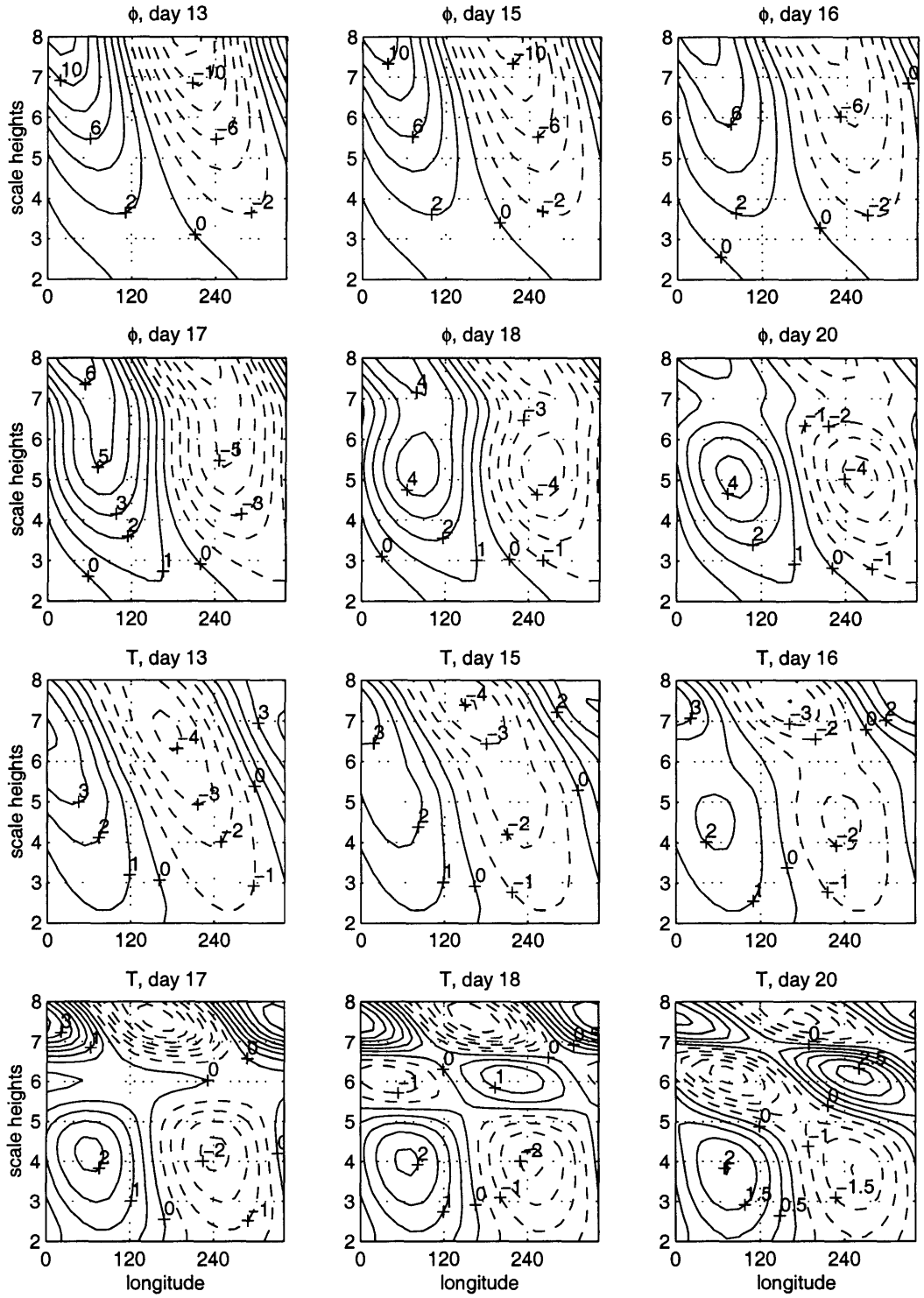

Figure 7.8: Zonal-height sections of the geopotential height (top two rows) and temperature (bottom two rows) perturbations for six days in the model run. Day 13 is the wave field of the initial state while day 20 is pretty much the final steady state wave. Sections are taken at a latitude of 3.0.

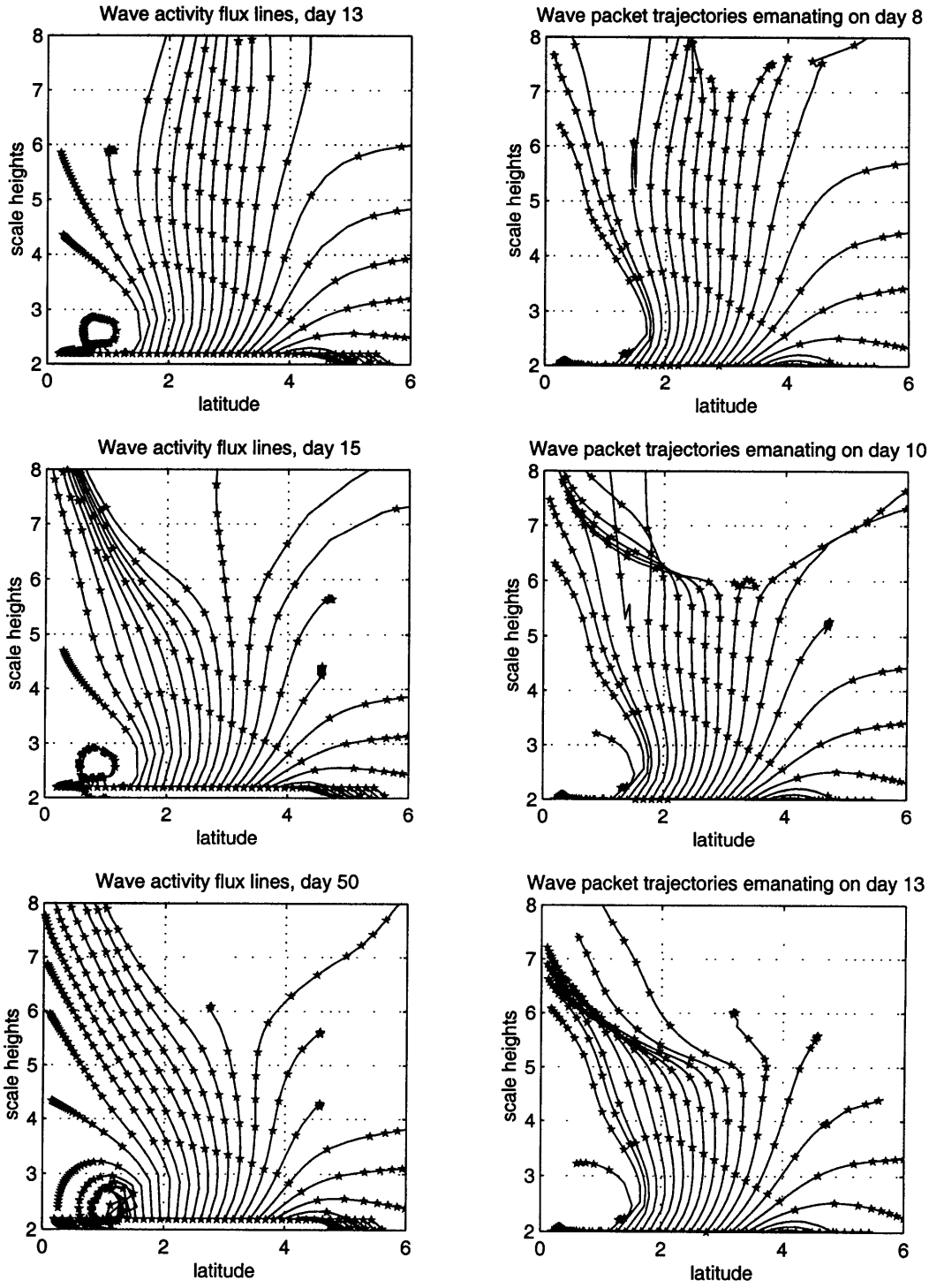

Figure 7.9: Left: Wave activity flow lines on days 13, 15, and 50 of the model run. Right: The paths followed by wave packets that emanated at the bottom of the model (2 scale heights) on days 8, 10, and 13. Stars are separated one day apart. See text for details.

7.1.2 The role of time variations in forcing

Note that there is no eastward phase tilt with height in our model run. Such a tilt requires a larger downward than upward propagating component, which can only happen if the wave forcing decreases while the wave reflects downwards. A reduction in wave forcing is indeed observed during the periods of structure changes. We run our model with a bottom forcing that decreases at the same time that the turning point first forms and the wave starts tilting to a vertical position. The result is that the wave goes on tilting to an eastward position before decaying. This raises the question of whether the observed structure changes are not just the result of the forcing decreasing. As we have shown in section 5.3.4, decreasing the forcing in a basic state that has a turning point will cause the wave pattern to tilt eastward with height because we are shutting off the upward propagating wave first, and only the downward reflecting part of it remains. From figure 7.1 we can see that a decrease in the amplitude of the wave at 150mb slightly precedes the decrease of the wave (and the variations in structure which accompany it) in the upper stratosphere in the first deceleration period, and coincides with the changes in structure in the second deceleration period. The basic state, however, does not seem to have a turning point before the deceleration changes the basic state (as is shown in figure 7.5). It appears that two simultaneous things happen- a turning point forms and the bottom forcing decreases. This naturally raises the question of whether these two are connected, and in what way? Do the variations of the basic state lead to a weakening of the wave or does the weakening of the wave lead to enhanced deceleration of the mean state? At present, we do not know the answer. The former possibility suggests an effect of the stratosphere on tropospheric wave structure. Some evidence for such influence is consistent with the one dimensional model results of chapter 4, where we saw that the existence of a turning point in the stratosphere has an effect on the phase speed and growth rate of normal modes, which are due to tropospheric instability. Since our study focuses on the stratosphere only (and the tools we have developed are not easily applicable to the troposphere), we leave addressing these issues for future research.

7.1.3 The consistency and estimation of time scales

The patterns of our model run and observations are similar. A main issue we still need to check is the consistency of time scales of variations in structure with the vertical group speed in the atmosphere. In the idealized case when a turning point is inserted into the path of an upward propagating wave, a wave front forms at the turning point, that propagates downwards with the group speed. The time it will take the wave in

the domain to reach a vertical position is roughly the time it takes a wave packet to travel from the turning point down to the bottom[6]. In our model run we do not have an instant formation of a turning point and we have gradual and partial reflection, hence the time it takes the wave to tilt should be larger than this group propagation speed (see section 5.3.4). From figure 7.8 we see that it takes the wave roughly two days to tilt into the vertical (days 15-17, this is an under-estimate because the wave also decreases its tilt slightly on days 14 and 18). The wave activity paths (e.g. top right of figure 7.8) give us an estimated propagation time of four days between 2 and 5.5 scale heights (the height of the turning point). The wave packet paths, however, in most cases over-estimate travel times because they take into account both upward and downward propagation. If there is partial downward reflection, the net vertical propagation speed will be reduced (it is zero for full reflection). We can also use ray tracing (Karoly and Hoskins, 1982, see section 6.5). Figure 7.10 shows the ray tracing calculations for the initial and final states of our runs, superposed on the PV gradient fields. Circles are spaced one model day apart. The estimated propagation time from these calculations is a bit more than 2 days to travel from 2 to 5.5 scale heights, which is consistent with a tilting into the vertical over approximately two days. To obtain corresponding estimates of travel times from observations, we need to apply our wave packet diagnostics to them.

7.1.4 Evolution of the wave using the wave packet formulation

It is illuminating to look at the time evolution of the wave field using the wave packet framework developed in chapter 6 (see also appendix D for the spherical coordinate version of the diagnostics). It is especially interesting to superpose the wave packet diagnostics on the index of refraction or on the meridional and vertical wavenumbers of the steady state solution, to see how they relate, since they are obtained using different and somewhat independent calculations. The wavenumbers are calculated using observed zonal mean quantities and a model, n^2_{ref} is calculated using observed

[6]This analysis also holds for the case where we have a standing wave and the forcing at the bottom is turned off, causing the wave to tilt eastward. In this case we have a wave packet tail forming, and propagating upward with the group speed. The vertical group propagation time is the time it will take the wave pattern to reach its maximum eastward tilt. If both a turning point forms and the forcing at the bottom is turned off in a vertically propagating wave field, both effects will occur at once. After the time of group propagation from the bottom to the turning point the wave will be have an eastward phase tilt with height, while after half this time it will have a barotropic structure.

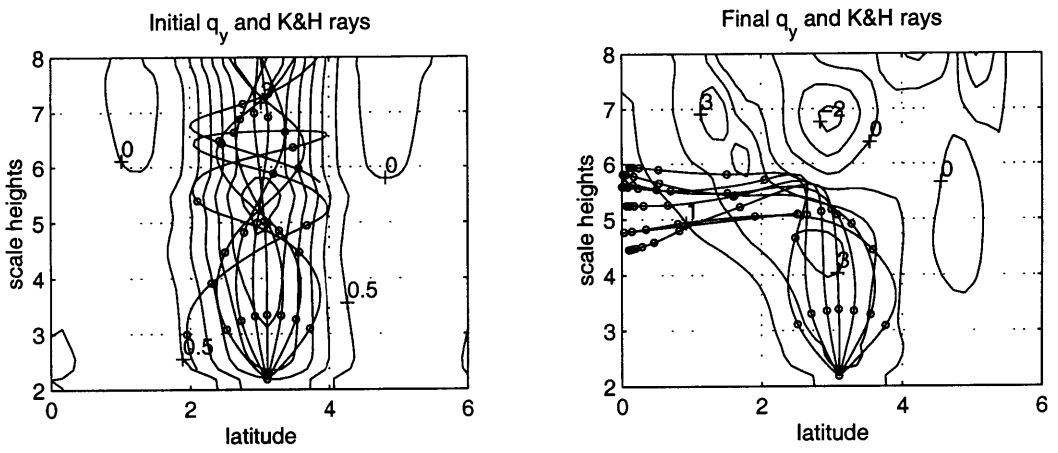

Figure 7.10: Latitude-height plots of wave rays, calculated using Karoly and Hoskins ray tracing, for the initial (left) and final (right) basic states shown in figure 7.7, for a source at 2.2 scale heights and latitude y=3. Different lines are for different initial propagation angles. Circles are spaced one day apart. Also plotted are the corresponding PV gradients.

zonal means, and the wave packet diagnostic uses observed wave and zonal mean quantities. Figure 7.11 shows the paths that wave packets follow, starting from a height of 2.3 scale heights on July 23rd and August 7th, with shaded regions denoting no meridional propagation ($l^2 < 0$) on July 31st and no vertical propagation ($m^2 < 0$) on August 12th, respectively. Note that l (m) are relevant only to the motion of wave packets on July 31 (August 12). The time is marked on each packet path by the circles, which are spaced one day apart. Looking at the July 23rd packets, we see one packet path that changes its direction sharply on July 30-31. This packet path clearly shows the earlier period of downward and equatorward reflection due to the splitting of the waveguide on July 31st. The mid-August downward reflection shows up much clearer, as is evident from the August 7th packet paths. We see that downward reflection occurs at the location where a turning surface forms on August 11-12th. An advantage of this diagnostic is that we can get a sense for the time evolution of the two dimensional field in one plot. It is interesting that the different nature of structure changes between the two deceleration events shows up so clearly in these wave packet paths. In mid-August a vertical turning point develops at all latitudes, and the reflection is very strong, causing all wave packets in midlatitudes to move down. In the July-August event the waveguide only splits in two with no turning point for vertical propagation developing and the reflection is weaker, more concentrated in latitude, and for some wave packets is more equatorward than downward.

It is important to note that the location at which wave packets reflect downward does not necessarily coincide with the turning surfaces, rather, it depends on the

location of the wave packets on the days that downward reflection developed.

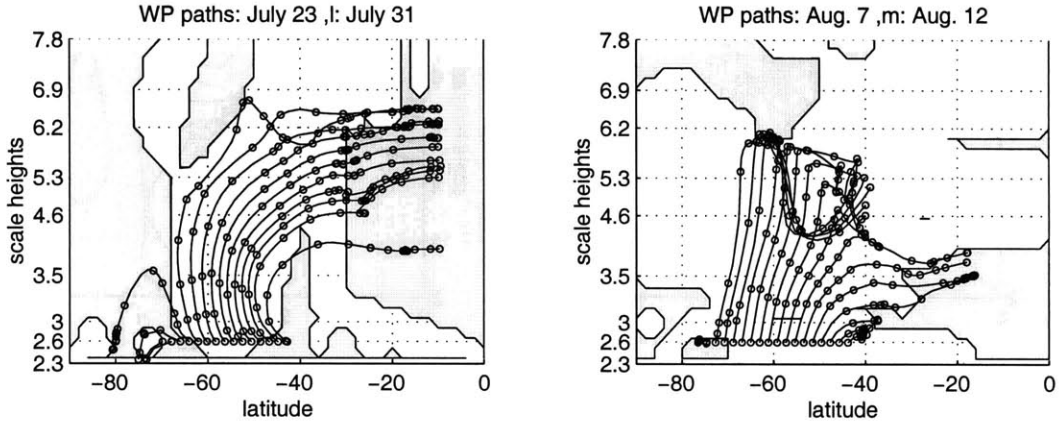

Figure 7.11: Latitude-height plots of wave packet paths, for packets emanating at the bottom (2.3 scale heights) on July 23rd (left) and August 7th (right), 1996. Circles are spaced one day apart. Regions where $l^2 < 0$ on July 31st (left) and $m^2 < 0$ on August 12th (right) are shaded. See text for details.

Figure 7.12 shows wave packet locations (section 6.4) for different stages of the July-August wave event, superposed on the corresponding wave geometry (regions of negative n^2_{ref} are shaded). Each symbol represents a wave packet. Packets are marked by the day they emanate at the bottom (2.3 scale heights in this case), and the color and symbol are kept the same in all plots for a given emanation day. For example, on July 21st, we see the locations of packets that left the bottom on July 18 (magenta o's), July 19 (yellow o's), July 20th (black squares) and July 21st (the purple line at the bottom). On these days, the wave is in its growing stage. We can see that it took the packets of July 18 three days to ascend from 2 to 5-6 scale heights (meaning 3-4 days to reach 6 scale heights, where the turning point develops). Packet locations on earlier days (not shown) suggest an even faster vertical propagation, which makes sense because wave packets slow down as the wave reaches the stratopause and some partial downward reflection develops. This time scale is consistent with the time it takes the wave to tilt into a vertical, and eastward position (see section 7.1.3, and footnote 6 for a discussion). There are two periods of relatively fast vertical propagation during the wave event, once at the beginning (July 21) and one after the first deceleration-wave tilting event when the wave grows again (see large spacing between red, light-blue and magenta lines on August 9th, representing packets that left the bottom on August 6-8 respectively). During these rapid propagation and growth periods, wave activity first propagates vertically up the waveguide, and only later it leaks out sideways to the equator (packets are more concentrated in the waveguide on July 26 and August 9th and are more spread out

on July 31 and August 14th). Also clearly shown is the splitting of the wave field into the two waveguides on July 31st-August 2nd[7] along with a downward motion of the wave packets as a result of the downward reflection. Downward propagation is also clear on August 14-15 (15th not shown), when the wave essentially breaks down.

A note is needed as to the quality of these diagnostics. First, in our calculations, we set the wave activity velocity to zero in regions of negative PV gradients, hence, the bunching up of packets at equatorward vertical lines. To some extent, the good correspondence between packet locations and the wave geometry is due to the dependence of the magnitude (not the direction) of the wave activity velocity on PV gradient, which is also a large factor in n^2_{ref} ($\vec{V}_a = \frac{\vec{F}}{A} \propto \frac{\vec{F}}{1/\overline{q}_y} \propto \overline{q}_y$). The direction of the propagation, however, depends on the EP fluxes of the wave. In particular, the boundaries of the downward reflection region depend on where the EP fluxes change from upward to downward, which depends on the phase structure of the wave. It follows, that the location of the reflection front is a more reliable feature than the actual wave packet locations. Plots of successive wave packet locations are useful in showing us patterns, but it is hard to determine whether we can trust the exact locations of a given wave packet on a given day. An idea of the uncertainty in packet locations can be drawn from the large differences between calculations of $\nabla \cdot \vec{F}$ of the various analyses products, especially for the southern hemisphere.

One specific problem with the observations is that we do not observe a time progression of the reflection region, as we see in the model run. Instead, the downward reflection appears and lasts throughout the upper half of the stratosphere (above 4 scale heights) simultaneously (July 30-August 2). The downward progression of the reflection front, however, is not something we expect observations to be able to detect easily, given the coarse vertical and time resolution. It does, however, suggest we should look at other mechanisms that may cause a tilting of the wave.

[7]The poleward upper branch of the waveguide disappears by August 2nd. n^2_{ref} of August 1st is shown to highlight that packets move up the poleward branch of the waveguide.

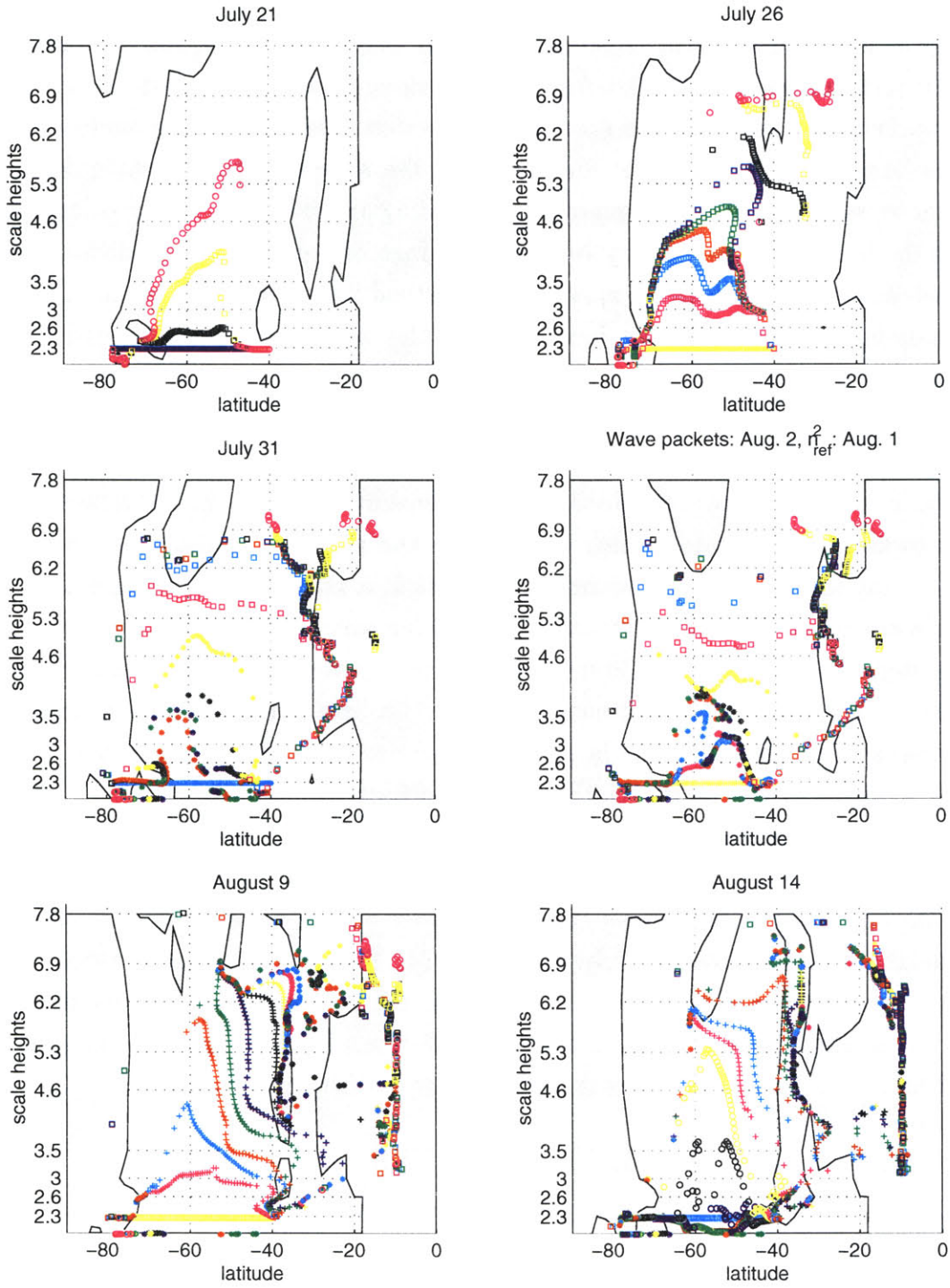

Figure 7.12: Latitude-height plots of wave packet locations, for July 21, 26, 31, August 2, 9, and 14, 1996, superposed on the shading of regions of negative n^2_{ref} of the same day, except for packets of August 2nd, for which the shading is of August 1st. Each symbol denotes a wave packet. Packets are plotted for each day, and the colors and symbols mark the day of emanation at the bottom. These are consistent within the different plots, for example, magenta circles denote packets that emanated on July 18th. See text for details.

7.1.5 Alternative possibilities

Superposition of stationary and transient waves

A possible source of variations in vertical structure is the superposition between a transient and a stationary wave, each of which has a constant vertical structure with time (e.g. Salby and Garcia, 1987, Lindzen et. al., 1982). One of the characteristics of such a superposition is a periodicity in the changes, that should have the period of the transient wave. Figures 7.2 and 7.3 show that the variation of structure occurs twice in July-August of 1996, suggesting a period of 13-14 days. The main problem with testing this possibility from observations is how to separate between the stationary and transient waves, especially since their amplitude changes with time as the waves grow and decay. One possibility is to assume the amplitude of the waves is constant with time at least over one period, hence a time mean of that period is the stationary wave. Two choices that we tried are the period between the two days on which the wave is vertical (July 30th-August 12th), and the entire July 18-August 19th period. In both cases, the transient wave, which is the total wave minus the stationary wave, do not look like a traveling mode. At a given level, the phase speed changes with time. Also, the wave does not propagate at the same phase speed at all levels. In the period of August 4-13, the transient wave at 30mb propagates westward, while at 10 and 2mb it propagates eastward, for both cases. This means the vertical structure changes in time during this period. It is possible that we are seeing a superposition between two different transient waves, however, the additional degree of freedom in choosing the two wave phase speeds makes it even harder to separate the two modes. A main problem of the superposition theory is having to account for a source of traveling waves in the stratosphere, especially modes that last for a whole month. Given this, and the fact that our simple attempts do not support the superposition theory, it is more likely that what we are seeing is a transient response of the wave to the changes in basic state and forcing.

Non-modal decay by shearing

The coincidence of decay with a tilting of the waves from a westward to an eastward phase shift with height is reminiscent of the non-modal decay of the Orr mechanism (1907). Such a decrease in amplitude has to do with the changes in spacing between PV lines, caused by shearing. The tilt of the perturbation in this case is an advection by the mean flow and in not associated with the propagation of waves in the vertical

direction, hence is not a result of downward propagation or reflection[8]. In this section we test the possibility that the observed southern hemisphere wave 1 in July-August 1996 undergoes such a decay-shearing stage, once the source of the perturbation is shut off. This would account for the tilting appearing over a deep level simultaneously.

We use our time dependent β-plane QG model to test this with a few different runs. The basic state we use in all runs is the control run of chapter 5, with a sponge layer at 10.5 scale heights. In chapter 5 we saw that in such a basic state, stationary waves propagate vertically to the sponge layer where they are damped. Stationary waves were shown to have a turning point in the sponge layer but its effect is not felt strongly by the wave because the damping is sufficient to cut the downward reflection.

Since the initial tilting of the wave against the shear is associated with its vertical propagation from the troposphere, we look at what happens to the perturbation after we shut off its source (reduce the bottom forcing to zero). The resultant behavior depends on whether there are turning points, or actually, on whether the initial steady state has a downward propagating component in it or not. When it does not, the perturbation propagates vertically and eventually dissipates in the sponge layer. We see a wave packet 'tail' forming and moving upwards. Correspondingly, the tilt remains westward with height the whole time. To further understand the mechanisms associated with shearing the PV by the flow, we initialize our model with a PV blob, and let it evolve without forcing it any further. The blob has a zonal wavenumber 1, and a barotropic structure. A PV blob that is put between y=2-3, and 2.5-3.5 scale heights does shear with the flow, but it induces a vertically propagating component above it, which tilts westward with height, even as it decays in time. It could be that the time it takes the perturbation to shear is larger than the time it takes it to propagate vertically and decay in the sponge layer, hence we put the PV blob in a larger range of heights (3-6 scale heights), to give the perturbation the most favorable structure for shearing. Figure 7.13 shows the evolution of the geopotential height field in this run. Even in this case, we see an initial *westward* tilting of the perturbation, however, starting on day 5, we see an eastward tilt developing (it lasts three days only but we believe reflections from the surface affect the results at this point). This run suggests that the shearing mechanism may work in the stratosphere, but given the unrealistic initial state (a barotropic PV blob), we do not believe it is relevant to the observed case at hand. We do, however, need to understand what causes the

[8]Note that such decay may occur on any wave geometry, even when we have an evanescent wave region (for example when $\overline{q}_y = 0$), and even without the existence of a turning point. This makes the shearing decay mechanism different from the case of tilting and decay that are caused by turning off the tropospheric forcing.

perturbations to tilt against the shear, in order to determine in what cases, if at all, we expect perturbations in the stratosphere to be sheared by the basic state.

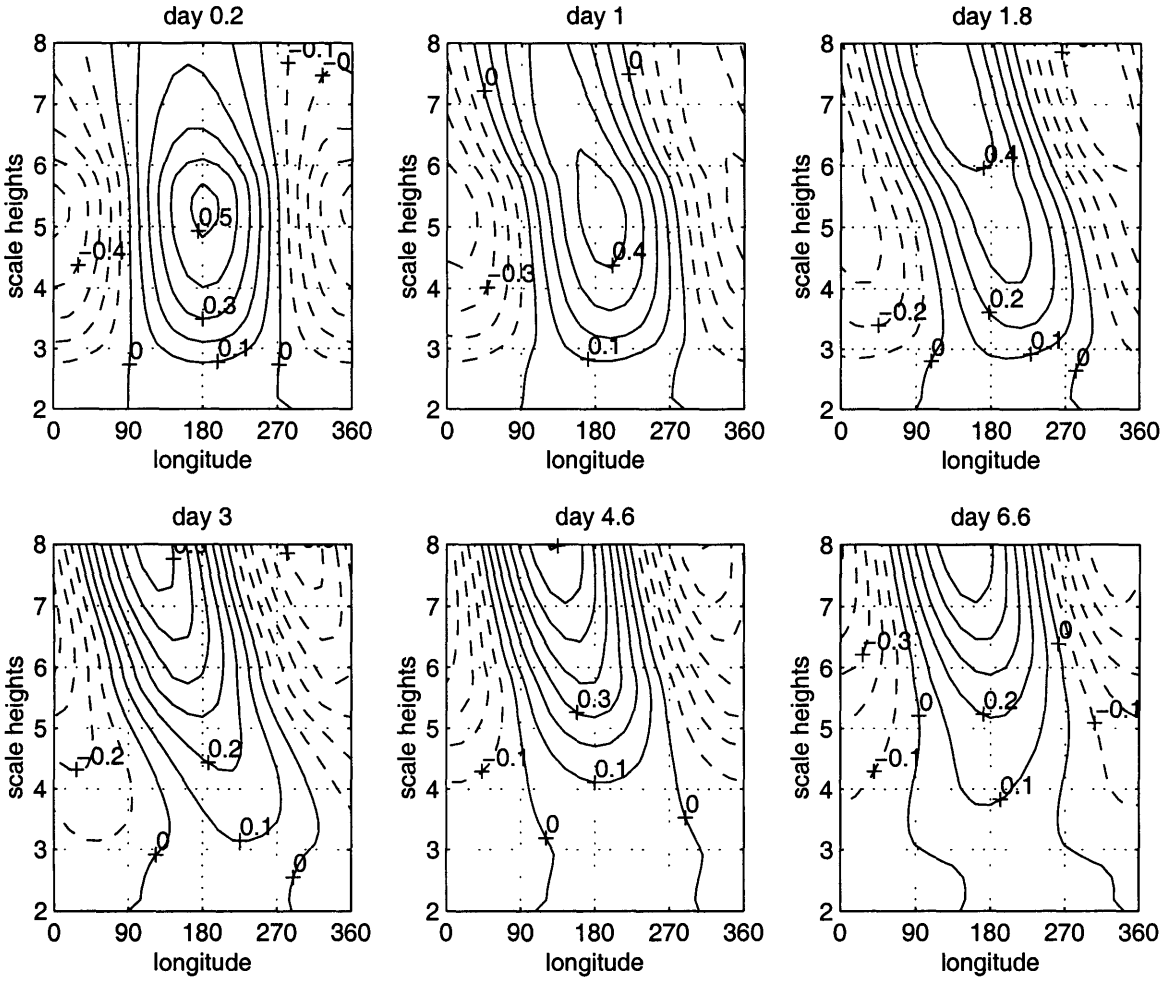

Figure 7.13: Longitude-height plots of geopotential height perturbation, at different times, for the run that is initialized by a barotropic wave 1 PV blob at y=2-3, z=3-6. See text for details.

A possible explanation we can think of is the following. There are two opposing mechanisms affecting the tilt of the PV perturbation. One is the advection by the mean flow, which will tend to tilt the perturbation with the shear. The other is the vertical propagation, which is manifest by the advection of basic state PV by the meridional flow that the PV perturbation induces. It can be shown that above the PV perturbation, this mechanism will tend to tilt the wave against the shear (Heifetz, 1999, personal communication), which is expected since the wave is upward propagating. What causes the latter effect to win seems to be the density effect. Some support for this comes from runs we have done where we initialized the model with a random PV perturbation (wave 1, random initial amplitude and phase at each

grid point) to see what patterns emerge. Large scale patterns occasionally emerge, that are very weak, but they have a westward tilt with height. Since a δ-function PV perturbation will propagate in all directions, the only reason why the vertical propagating structure emerges is the amplification related to the density effect of amplifying vertically propagating waves. We plan to test this by running a Boussinesq flow model, among other things. As for the issue at hand, it seems that in the present case at least, the observed tilting of the waves in July-August 1996 is not due to shearing during the decay stage of the wave.

7.2 The September version of reflection from a turning point

In this section we will discuss variations of wave structure that occurred in wave 1 of September 1996, which are in essence similar to the July-August case shown above, but with a few differences. We concentrate here on presenting the differences.

Figure 7.14 shows time-height plots of the zonal mean wind, the wave 1 vertical wavenumber calculated from the steady state solution using the daily basic state, the zonal mean wind acceleration, and the contribution to the acceleration from wave 1 $\nabla \cdot \vec{F}$ (RHS of equation D.11) for September 1-30, 1996. All quantities are averaged over 40-80°S, except the vertical wavenumber (m) which is averaged over 56-76°S (these latitudes were chosen to represent the values of m in middle-high latitudes, see for example figure 5.13). The evolution of wave 1 geopotential height and temperature is shown in figure 5.12. There are two strong deceleration events during this period (9/9-12, 20-22), followed by acceleration (larger on September 13-15). Wave 1 $\nabla \cdot \vec{F}$ is strong enough to account for the deceleration, but not for the acceleration which follows. On the other hand, a strong positive wave 2 $\nabla \cdot \vec{F}$ on September 12-14 (not shown) can account for it. We will comment on the appearance of wave 2 later (before September 12 its $\nabla \cdot \vec{F}$ is negligible). Following both decelerations, wave 1 tilts to a vertical position (geopotential height phase becomes constant with height, and a π jump in temperature phase forms around 5 scale heights). In between, the wave returns to a westward phase tilt with height (September 16-21).

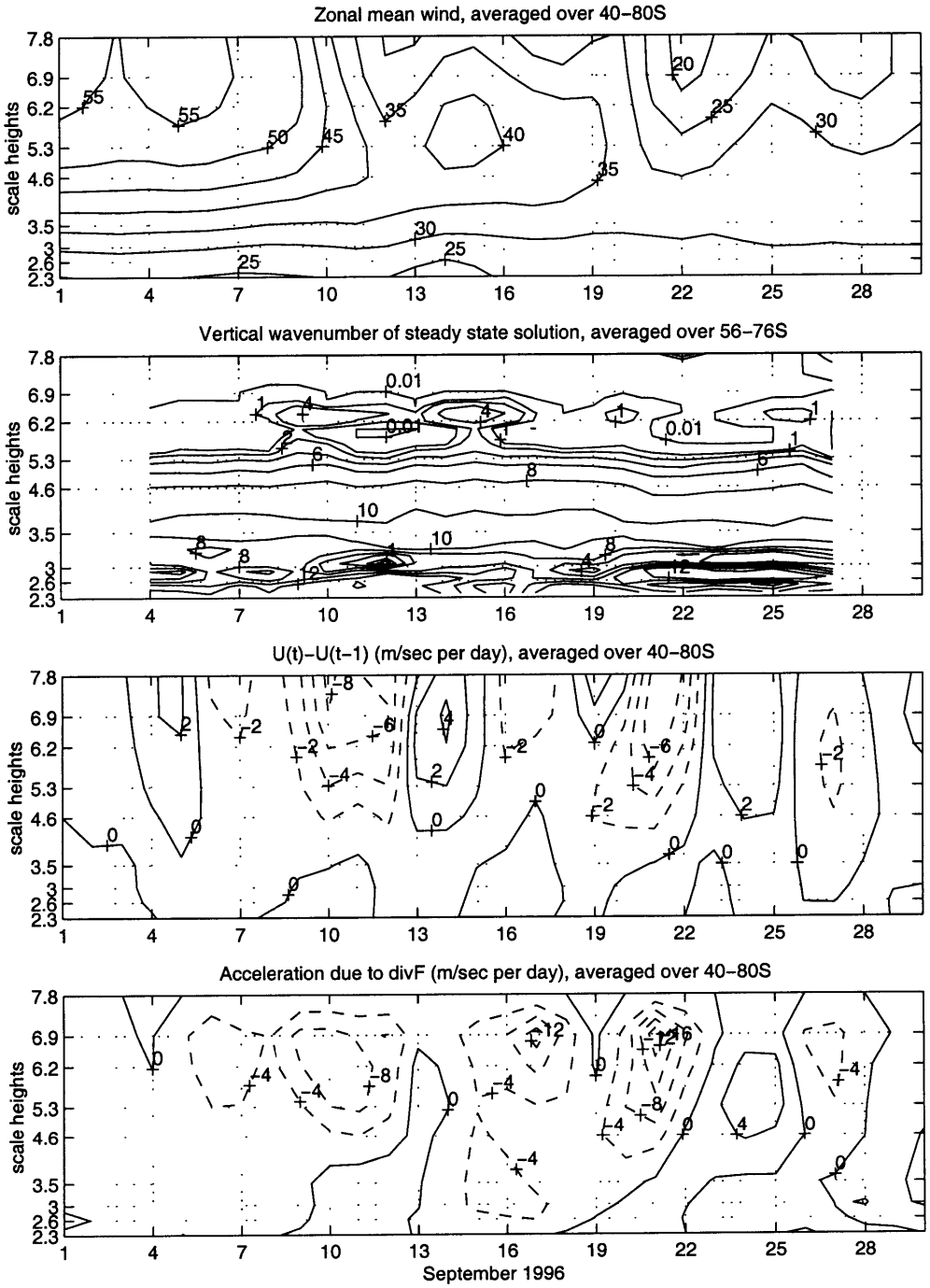

Figure 7.14: Height-time sections (September 1-30, 1996) of a latitudinal average of (top to bottom): A. Zonal mean wind (m/sec). B. Vertical wavenumber (m), calculated from the wave 1 steady state solution to the daily observed basic state ($10^{-5}m^{-1}$). C. The change in zonal mean wind over 1 day: (U(t)-U(t-1), m/sec, negative values dashed). D. The acceleration due to wave 1 driving: $\frac{\nabla \cdot \vec{F}}{a_e \rho \cos\varphi}$ (m/sec per day, negative values dashed). All quantities, except m are volume averaged in latitude over 40-80°S (weighted by $\cos\varphi$). m is averaged over 56-76°S.

In mid-August we saw reflection that resulted from the formation of a turning point. In September, this mechanism doesn't obviously work because there is a turning point to begin with (as we saw in 5.4.2, the climatological September basic state has a turning point at around 5.5 scale heights). Looking at the time series of m in figure 7.14, we see that a turning point exists within the observation domain throughout September, but that on 9/11-13 and 9/21-24, the turning point dips down[9]. By comparing this time evolution of m with the evolution of the wave phase tilt, it seems that the wave phase tilt roughly follows m, such that a downward motion of the turning point causes the wave to reflect downward and tilt vertically, and an upward motion of the turning point causes the wave to tilt more westward with height. As was already pointed out in 5.4.3, the observed temperature phase is not similar to that of the steady state solution, which is never vertical (figures 5.12 and 5.15). This emphasizes the importance of transience when the adjustment of the wave involves reflections.

Another change in structure which is not seen in mid-winter is the transition from a double to a single peak in temperature amplitude (September 10-12). This is most clearly seen when looking at longitude-height sections of temperature (shown in figure 7.15 for September 8-13, at 60°S)[10]. Unlike the phase of the wave, the observed temperature amplitude seems to roughly follow (with a time lag of approximately 3 days) the amplitude of the steady state solution to the time evolving basic state (compare figures 5.12 and 5.15). For example, a single peak in temperature, like observed on the 10-12th, appears in the steady state solution on the 7-9th. Also, the steady state solution has a single peak on the 14-16th, while the observations almost have a single peak (very shallow double peaks) on the 18th.

Since the variations in the height of the turning point are not very large, and in some cases the tilting into the vertical appears quite suddenly in observations (see figure 7.15, September 12-13)[11], we test the adjustment of waves to changes in the height of the turning point in a model. We want to see whether a small change in the height of the turning point can cause the wave to tilt as much as observed, and whether observed time scales for these changes are reasonable. We run our time

[9] $m = 0.01$ denotes the turning point.

[10] This change is also evident in the latitudinal average (figure 5.12), hence it is not due to latitudinal shifts of a complicated wave pattern, but to a robust change over all latitudes.

[11] Our intuition is that the transition from a westward tilting wave (9/12) to a vertical wave with a double temperature peak (9/13) is robust, because both structures are observed on the few days before and after this transition, but the observations of the transition stage are biased by the asynoptic sampling (which increases eastward phase progression), and in addition, there is probably some error in the highest levels, where most of the transition occurs.

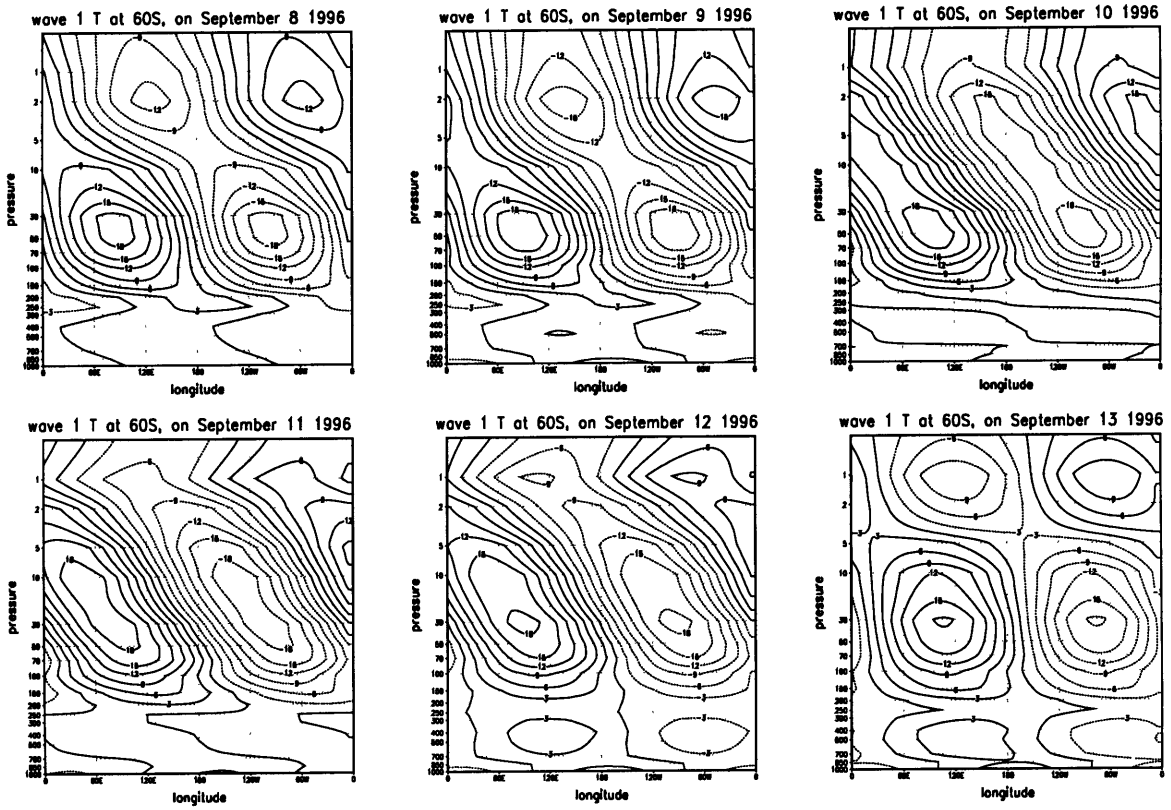

Figure 7.15: As in figure 7.3, only for September 8-13, 1996.

dependent model (of sections 7.1.1, 5.3.4) with a constant forcing at the bottom, and over a period of three days change the zonal wind (in both directions) between the observed September 6th and 15th winds (turning points at 5.5 and 7 scale heights, respectively)[12]. When the turning point is lowered, we see a tilting of the wave to a vertical position within 3-4 days, but only in the run where the lid and the sponge layer are high, so that the damping at the turning points is minute[13].

We also ran the model using the basic state of other days. One point that comes out clearly is the role of the wind structure above the domain of observations. Since the turning point in late winter is due to a positive vertical curvature of the wind above its peak, the propagation characteristics can be sensitive to the winds above the domain of observations. This is true if the turning point is close enough to the highest observation level (1.5 mb), because a region of small enough vertical curvature may result in another turning point, above which we will have vertical propagation. If the two turning points are close enough, the waves will easily tunnel. The phase

[12]These days were chosen to represent a low and high turning point in the β-plane model.

[13]The lid and sponge layer are raised by 5 scale heights compared to the control run, and a constant damping of $0.04day^{-1}$ is applied.

structure of the waves is sensitive to this wave geometry, as well as to the damping. We can, however, get vertical phase lines as a transient structure, without it being a characteristic of the steady state solution before or after the basic state changes. This sensitivity to high level winds may be more important for the amplitude structure. It is unclear if we can get the temperature amplitude to 'stretch' to one peak as a strictly transient response, or whether it happens only if the steady state solution to the basic state at some point has only one peak. In our model run, we do manage to get the two temperature peaks to become one, within three days of moving the turning point up, but only in the run with a high lid and sponge layer. The steady state solution to the basic state with the high turning point does not, however, have a large upper temperature peak to begin with.

The sensitivity of the response of our model waves to the basic state in the upper stratosphere raises the issue of the reliability of the observations. The changes in zonal mean wind occur mostly above 5 scale heights, where observations are less reliable, but they extend down to 4.6 scale heights most of the time. The time variations in the vertical wavenumber, however, occur above 5 scale heights. Since m is diagnosed from a model that uses vertically interpolated winds it is hard to judge how accurate it is. The variations in temperature amplitude are concentrated at high levels as well, and do not involve very large amplitudes. For example, the difference between September 9 and 10 in figure 7.15 is of 3-6°K at 5-10 mb. The analysis of chapter 3 suggests the retrievals are capable of resolving such a feature, but errors can be as large as most of this difference. The fact that they occur simultaneously at all latitudes raises our confidence in these variations, because they are not due to measurement noise. As was shown in chapter 3, large scale errors may be a result of sharp features projecting onto the vertical correlations of the error covariance matrix of the retrieval. It is hard to say if sharp vertical features exist in September. As we will show in a moment, there is wave breaking going on in some of the days, however, the high peak in temperature amplitude disappears during one of the wave breaking events, contrary to what we expect from spurious vertical correlations. Overall, however, since the variations in basic state and in the waves agree quite well, we believe they are real and at the most the errors are quantitative, not qualitative.

Another mechanism which might play a role in affecting vertical wave structure is wave breaking. The increase in phase tilt with height, and the disappearance of the higher temperature peak, is consistent with a 'pulling up of the wave' (or alternatively, an elimination of the downward reflected component of the wave), which could be achieved by increased damping in the vicinity of the turning point. This, along with the sudden appearance of wave 2 $\nabla \cdot \vec{F}$ mentioned above, suggests looking for wave

breaking. By Rossby wave breaking, we mean an irreversible pulling out of material PV from the polar vortex (McIntyre and Palmer, 1983). In terms of a linear wave formulation, the nonlinearities effectively become large enough to project onto higher wavenumbers, and act as effective damping on the wave. Plots of Ertel PV on θ surfaces show a clear event of wave breaking[14] (the formation of a comma shaped vortex and an eventual detachment of one or more blobs of PV from the vortex) over the Indian ocean, on September 10-12 (figure 7.16). These are precisely the days when the temperature structure becomes one peaked. While we do not find large scale wave breaking in 1996 before September 10th, we see a weaker case of breaking on September 17-19 over the Pacific ocean, accompanied by a westward tilting of the wave, and a weakening of the upper temperature peak. Wave breaking however does not always occur simultaneously with a stretching westward of the wave, as in September 20-21, when we see quite large breaking over the Indian ocean, but no change in wave structure. At this stage, without running a nonlinear model that can exhibit wave breaking, it is hard to say more about the effects of nonlinearities on the vertical structure of the waves. The only thing we can do is show another example when wave breaking occurs simultaneously with vertical structure changes. Analyzing another case is useful because we can see if the vertical changes in wave structure are also accompanied by an upward shift in the location of the turning point. If not, it is more likely that the wave breaking acts like damping, and reduces the downward reflection from the turning point.

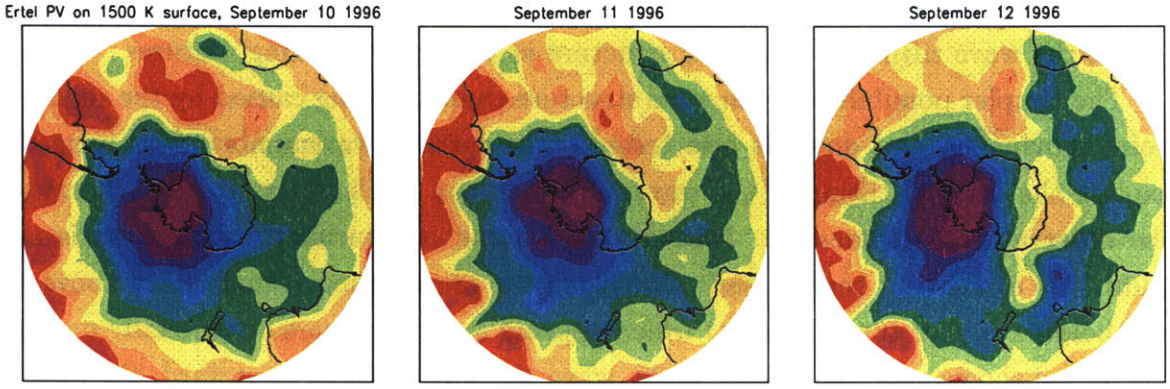

Figure 7.16: Ertel PV on the 1500°K θ surface, on September 10, 11, 12, 1996. Contour intervals are $2 \times 10^{-4\circ} Km^2/Kg/sec.$

[14]Wave breaking is observed above the 1000°K potential temperature surface (roughly 5mb) on these days, and extends at least up to 1500°K (the highest level we checked, roughly 2mb).

7.2.1 Another example from September 1982

Mechoso et al. (1988), in their discussion of the final warming of 1982 in the southern hemisphere, pointed out the appearance of an upward wave 1 EP flux in the stratosphere (without a corresponding tropospheric signal) alongside wave breaking, in September 1982. They discussed this evolution in terms of *internal stratospheric nonlinear evolution.* This signature is similar to what we observe on September 10th, 1996, hence we decided to analyze the 1982 case and compare.

Figure 7.17 shows time-height plots of the zonal mean wind, wave 1 geopotential height and temperature amplitudes and geopotential height phase, averaged over 40-70°S, for the period September 20th-October 9th, 1982. We see, as in 1996, a strong deceleration (9/25-29), followed by the wave tilting into the vertical (9/28-10/3). Just when the deceleration starts, we see an increase in the geopotential height phase tilt with height (9/23-25) and the transition from a double to a single peaked temperature amplitude (9/25-27). It easier to visualize these changes, and to compare them to the 1996 case, by looking at longitude-height sections of wave 1 temperature, shown in figure 7.18. Figure 7.19 shows the Ertel PV on the 1500°K potential temperature surface (roughly 2 mb), for September 25-27[15]. As in 1996, wave breaking occurs on the days that the temperature structure changes from a double to a single peak. We also see that a few days later, the wave tilts to a vertical position. Nonlinear reflection from the critical level is one possibility we would like to test in a more comprehensive study of the effects of wave breaking on wave structure. Both in 1996 and in 1982, we see the wave tilting to a vertical structure as the wave breaking matures.

The simpler possibility of the wave structure changes being caused by the turning point moving up and down is also tested. The vertical wavenumber does seem to explain the vertical tilting of the wave, although not as clearly as in 1996, because the observations are messier. Rather than having one turning point, there is an additional evanescent region (such that a narrow propagation region forms above it and below a higher turning point) which appears and disappears occasionally during the period shown. On September 28-29, when the wave reflects downward, the upper propagation region disappears, and there is a turning point at 5 scale heights. There is no increase in the height of the turning point on the days preceding the westward tilting of the phase lines on September 25-27, which supports the hypothesis that wave breaking causes the top peak to disappear. As we pointed out before, it is hard to judge how robust small scale features in m are (e.g. the multiple turning points),

[15]Similar to the 1996 case, wave breaking is observed above about 1000°K on these days, and extends at least up to 1500°K.

since we use a high resolution model with interpolated winds and temperature. It is important to note that after 1982 (end of 1988) the operational retrievals changed from a statistical to a minimum variance method, meaning, there is less dependence on statistics in the retrievals after 1988. In the southern hemisphere in particular the statistics are not very robust, resulting (at least in theory) in much improved retrievals.

Finally, it is interesting to read the discussion of the observed wave life cycles in Mechoso et al. (1988), in light of the current study. They discuss the evolution of waves 1 and 2 (both of which appear in 1982) in terms of life cycles of growth, eastward progression and decay. They do not explain the source of eastward propagation. While wave 2 is clearly an eastward propagating mode, wave 1 has periods of eastward propagation, separated by westward propagation. This cycle of growth-eastward phase progression-decay is what you get when a wave grows, reflects downward and decays.

7.3 Summary

In this chapter we have used the diagnostics developed in previous chapters, to ask to what extent the time evolution of observed waves can be explained by linear wave propagation theory, given the observed basic state. We find that the observed time evolution is a wave-mean flow interaction, where the wave responds qualitatively linearly to the basic state changes (which appear to be wave-induced), and that the time evolution of both seem to be dynamically consistent.

The purpose of these calculations, apart from testing the applicability of linear theory, is to test the use of the diagnostics and the reliability of the observations. Much of the time variations we are interested in occur at or above 5 mb, where the observations start losing reliability. The coincidence and consistency of variations in the basic state wave geometry and the vertical structure of the waves increases our faith in these observations. The wave geometry diagnostics we have used (l, m) are diagnostics of the *basic state* propagation characteristics, even though they were derived from the steady state *wave* solution. Being basic state diagnostics, they are relevant for the transient evolution of the waves, even on daily time scales. The diagnostics based on the steady state solution to an instantaneous basic state are useful for understanding the wave structure even when the basic state varies with time and the waves never have the opportunity to reach their steady state.

The phase structure of the waves, which is a diagnostic of the direction of propagation, is much of the time in transience. In particular, vertical or eastward phase

tilts with height are not a steady state response, hence they always result from abrupt changes in the basic state or forcing, or from nonlinearities (given the observations are real). During late winter, and occasionally during mid-winter, we have a turning surface which reflects waves downwards. Variations of the mean flow (which may be wave-induced) cause such turning points to either form abruptly, or if a turning point existed to begin with (as in late winter), to shift downwards. The abruptness of these changes causes the waves to reflect downwards for a few days. We also get reflection in the presence of a steady turning point with a transient wave source. Downward reflection and the associated structure changes will appear as an eastward phase propagation at some levels. Much of the discussion of planetary wave variability has been in terms of a Fourier decomposition in space and time. In the one winter we analyzed (1996, southern hemisphere), the life cycle of growth, followed by reflection (and sometimes decay) was quite abundant. While we did not study other years in great detail, we skimmed a few years of data for signatures of a deceleration of the upper stratospheric zonal mean winds, followed by a poleward heat flux (downward reflection). Such signatures were found in the southern hemisphere for wave 1 in September 1986, and 1983, and for wave 2 in September 1983, and in the northern hemisphere, for wave 1 in winters 1990-91 and 1995-6. While these signatures are not a proof of formation of a turning point and subsequent wave reflection, they are necessary conditions. While we have concentrated on the southern hemisphere, we expect to find such wave behavior in the northern hemisphere, except for times when the waves are extremely large and nonlinear (e.g. sudden warmings).

Finally, our analysis suggests that in September, when the polar vortex starts its breakdown, wave nonlinearities are noticeable in terms of their effect on the vertical wave structure, and at least qualitatively, nonlinearities can act as damping on the waves.

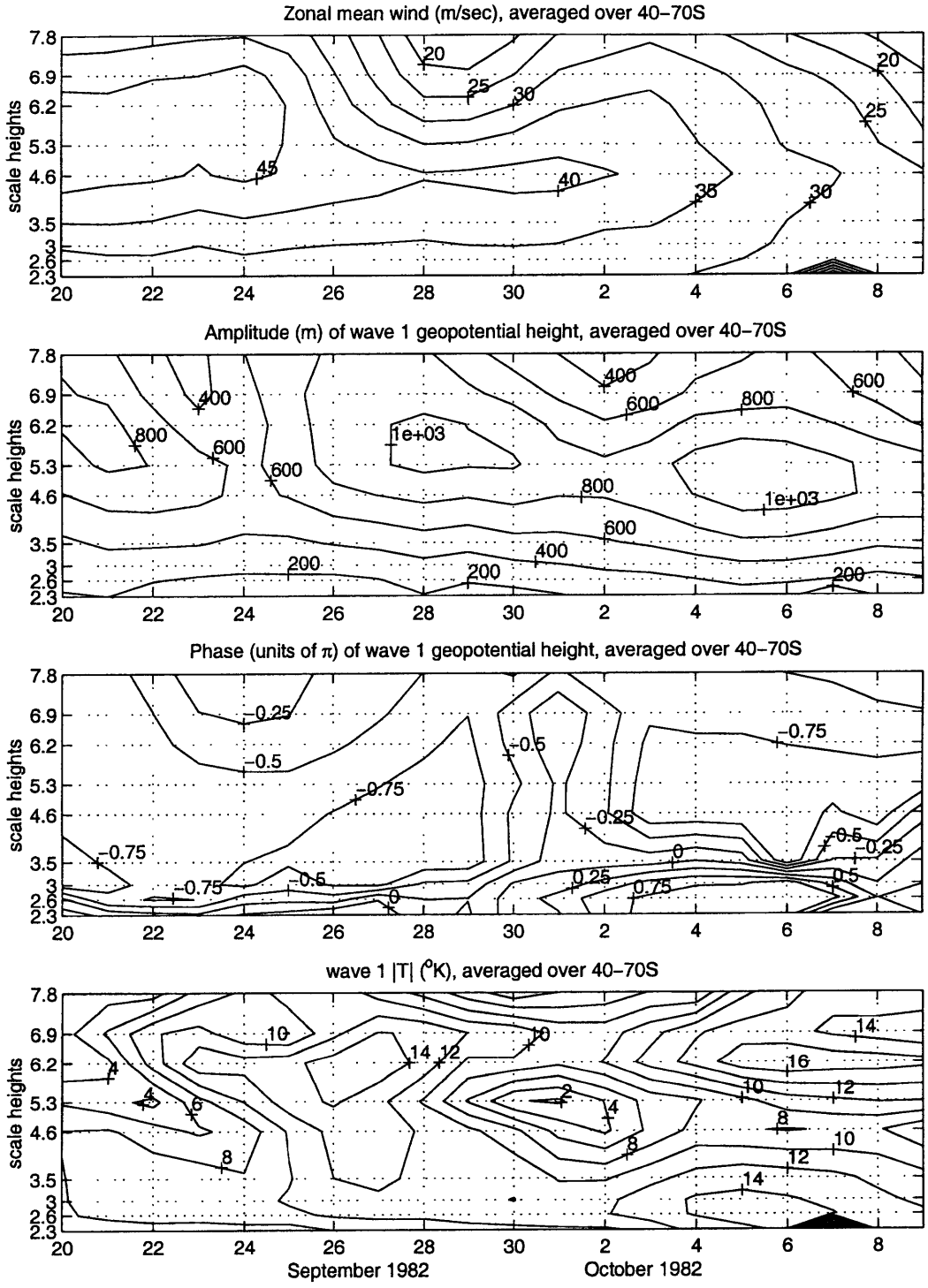

Figure 7.17: Height-time sections of the 40-70°S average of (top to bottom): A. Zonal mean wind (weighted by $\cos\varphi$, contour interval is 5 m/sec). B. and C. Wave 1 geopotential height amplitude and phase, respectively (amplitude in meters, phase in units of π). D. Wave 1 temperature amplitude (°K) for September 20 - October 9, 1982. The vertical grid is the observations grid.

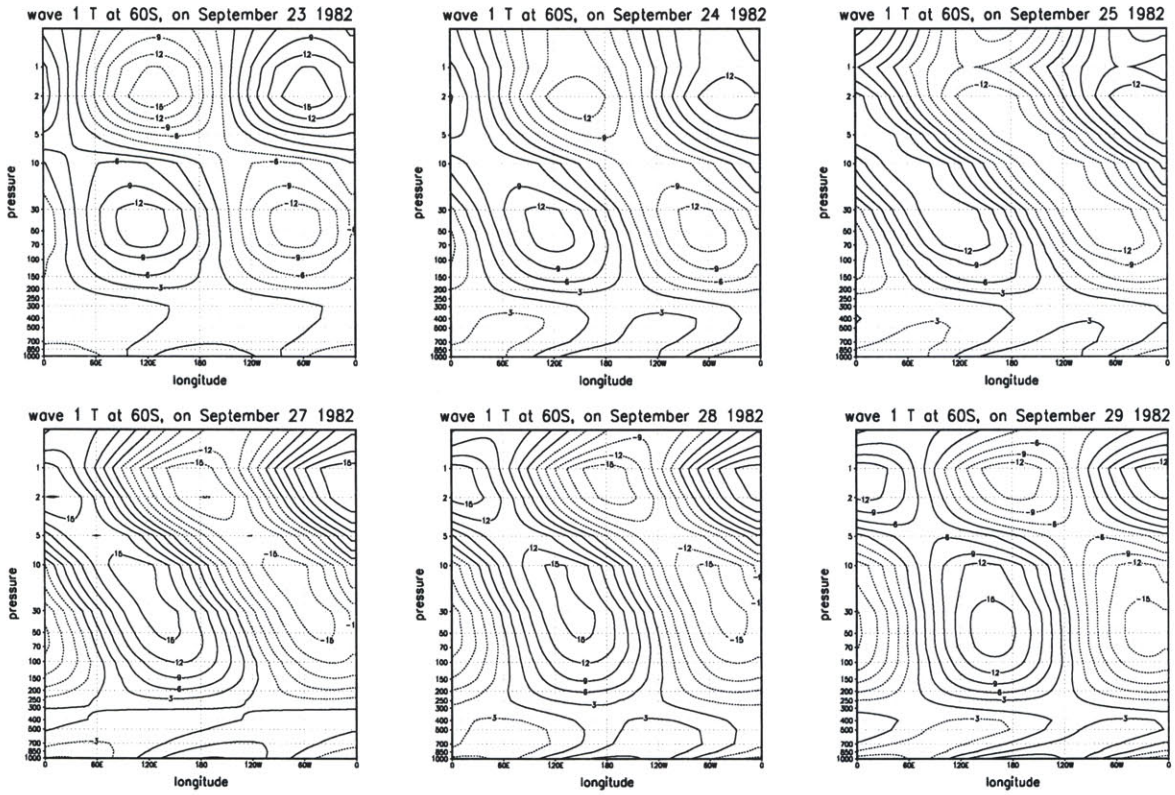

Figure 7.18: Longitude height sections of wave 1 temperature at 60°S, for September 23, 24, 25, 27, 28, and 29, 1982.

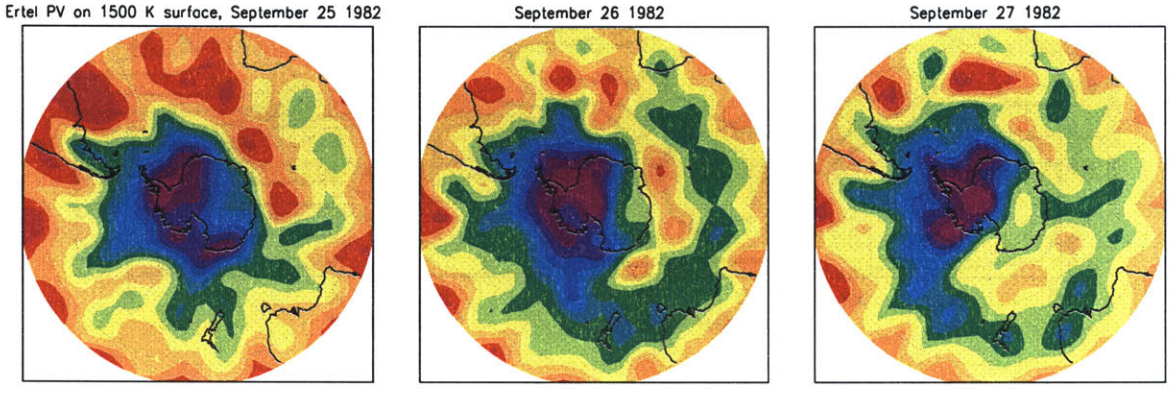

Figure 7.19: Ertel PV on the 1500°K θ surface, on September 25, 26 and 27, 1982. Contour intervals are $2 \times 10^{-4} {}^{\circ} Km^2/Kg/sec$.

Chapter 8

Summary and conclusions

In this thesis we study the vertical structure of stratospheric planetary waves and its variability, both in models and in observations. We concentrate on observations of vertical wave structure in the southern hemisphere, but a skim of some northern hemisphere data indicates that much of the wave features are observed there as well. Stratospheric waves usually appear in episodes, with a characteristic zonal wavenumber. In the northern hemisphere we observe mostly quasi-stationary waves 1 and 2, while in the southern hemisphere only wave 1 is quasi stationary and wave 2 is eastward propagating. We find a large variety of vertical wave structures, which varies from one wave episode to another, with a seasonal cycle. Also, within a given episode waves occasionally undergo changes that last a few days. These variations, are most notable when looking at longitude-height sections of the waves. It is important to point out the inherent difference between the occasional phase propagation associated with structure changes and the the modal phase propagation that is observed, for example for wave 2 in the southern hemisphere, which is coherent at all latitudes and heights[1].

A simple model study of normal modes on vertically varying basic states also reveals a large variety of vertical structures, that is similar to observations. In models, the factor that introduces the largest variability in the longitude-height structure of the waves is a reflection of the wave off of turning surfaces in the vertical direction. The existence and the location of turning points depends on the basic state. Using a diagnostic of the basic state wave propagation characteristics, we find in observations that a turning point exists in the upper stratosphere in late winter, and that

[1]Note that eastward propagating wave 2 also exhibits occasional vertical structure changes, where the associated phase propagation (which varies with height) is superposed on the modal eastward propagation.

occasionally it forms for a few days during early and mid-winter. Analysis of a few observed wave events shows the time evolution of vertical structure to be consistent with the formation of turning points and with time variations in the forcing. We view this as a qualitative assessment of the relevance of quasi-linear wave theory to the real atmosphere as well as an assessment of the observations (much of the relevant variability in the basic state occurs above 5mb, where observations are less reliable). A clear advantage of using vertical wave structures of geopotential height and temperature is that these fields are the most directly observed wave quantities. Higher order diagnostics have errors that relate to the base level analysis, to vertical and horizontal differentiation and to the balance assumption used to calculate winds. We do need to worry, however, about the ability of satellite retrievals to resolve the vertical structure of the waves.

We will proceed to summarize and discuss the main results of this work, which consists of theoretical model studies of the relation between vertical wave structure and the wave geometry of the basic state, an assessment of the ability of the observing system to resolve the wave phenomena we are interested in, and a diagnostic study of the observations to determine whether the theoretical relations we find hold.

8.1 Assessing observations

We start the thesis with a thorough discussion of the operational observations product we use, and the various stages involved in obtaining it (chapters 2, 3). The observations we use are based almost entirely on satellite measurements of radiance, from which temperature profiles are retrieved[2]. There have been many studies of the errors and uncertainties in the various stages of the observational product (see sections 2.1, 2.5 for references). The largest uncertainty we need to be concerned with is the ability of the satellite retrievals to resolve the vertical structure planetary waves, given the coarse vertical resolution of the observations. Past studies that have tested the ability of a retrieval algorithm to reproduce a temperature profile have mostly emphasized the ability to simulate the vertical structure of the total temperature field, and not wave structures, which are deviations from the zonal mean. It is true in general, that studies that asses the quality of observations are done by the scientists who design the observations, and not by the scientists who use them. These studies, therefore, do not usually test specifically the ability to resolve the phenomena of interest.

[2]The quantity observed is neither temperature nor geopotential height, rather it is layer mean temperature, or thickness. Combined with a surface height, this gives us the geopotential height. Using some interpolation method, we can get the temperature at a given level.

We have therefore decided to use our model to check whether the retrievals are able to resolve the vertical amplitude and phase structures of the waves by calculating the radiances that a virtual satellite sitting at the top of our model would see, inverting these radiances to obtain retrieved temperature fields, and comparing to the model 'true' fields. We find that retrievals are capable of capturing the important features of the waves, in particular their vertical phase and amplitude structure, to within a few °K, with a few limitations and exceptions. Above the peak of the top weighting function, at 1.5 mb, the retrievals contain little or no information from the radiance measurements, rather, the source of data in the operational products (which are given on levels as high as 2, 1, and 0.4 mb) is the *additional information* that is put into the retrieval to make it stable. This additional information is usually statistical. Errors in the retrievals start being large above 5mb, the peak of the second highest weighting function. Also, small scale features can not be resolved, but since most waves have quite large vertical wavelengths, the retrievals are able to resolve their general features quite well. When the wave fields terminate sharply at some level, the retrievals may spuriously create waves above that level, depending on the vertical correlations in the *additional information* used. This may happen, for example, when waves encounter critical levels, as is the case in summer or when the polar vortex breaks down in late winter. We should note that the above limitations reflect the information content in the radiance measurements, hence they pertain to any retrieval algorithm, including the direct assimilation of radiances. The differences between retrieval methods is in how data voids are filled, and how the solution combines the observations with the additional information.

Another concern we have is the ability of the satellites to resolve rapid variations in vertical structure, given the asynoptic nature of satellite sampling (section 2.3). A simple test of this effect (sampling a specified time varying wave as the satellite would and plotting the fields as if they were taken synoptically, as is done in the operational product) shows that the resultant distortion is usually a decrease in amplitude which is noticeable only for very rapid variations in vertical wave structure, and is quite small for observed ones. The distortion increases with increasing phase speed and is slightly larger for eastward moving patterns. Sampling errors may be partially responsible for the observed decreases in amplitude in the upper stratosphere that accompany structure changes.

Our results are reassuring, because they indicate the observations are capable of resolving the features we are interested in. We should be cautious, however, of observations above 5 mb, and of small scale features. Also, we should be cautious of the observations of the breakup of the polar vortex, when a critical level descends,

and time variations are very rapid. One way to verify the retrievals in such cases is to look at the radiances directly. Also, we can repeat our 'virtual satellite' exercise on a model of the polar vortex breakup. Also, an assimilation product where the radiances are assimilated directly, will assure that the additional information is dynamically consistent with an absorption of waves at a critical level, because the additional information used in the retrieval is based on the model dynamics (hence it will not introduce spurious waves in the easterlies above a critical level, as a non-diagonal minimum variance constraint would).

8.2 Theoretical model studies

The theoretical part of this work consists of a few studies:

- A study of the normal modes on tropospheric-stratospheric basic states that vary only with height, in the framework of wave-geometry and overreflection theory (chapter 4).
- A study of linear forced stratospheric waves on basic states that vary with latitude and height, both steady state and time dependent, where we come up with a diagnostic of the wave propagation geometry (chapter 5).
- A more diagnostic study, in which we view the waves as consisting of many wave activity packets that propagate from the troposphere through the stratosphere. We use a wave-based coordinate to study the evolution and budget of wave activity within the wave packets as they propagate through the stratosphere, both for steady state and time dependent waves (chapter 6).

Our goal in the first two studies is to categorize the waves in terms of their vertical structure, and to determine how the basic state and other factors like damping affect it. To this end we categorize the basic state in terms of the geometry of wave propagation and wave evanescence regions, separated either by a critical level or a turning point (*wave geometry*). This approach is useful because it lends itself well to generalizing our results to many basic states. While Charney and Drazin (1961) showed that wave geometry is relevant to the forced stratospheric wave problem, Lindzen et al. (1980) showed that it is relevant to baroclinically unstable modes, which can be explained in terms of wave propagation/overreflection. In chapter 4 we extend the normal mode analysis using overreflection theory to the normal modes of the tropospheric-stratospheric system. The modes we study draw energy from an

interaction with a critical level in the troposphere, and, depending on the wavenumber, propagate vertically in the stratosphere. We find a large variety of vertical wave structures, which are similar to the observed. The variability is both for different wave numbers on a given basic state, and for a given wavenumber on different basic states. It stems from having different stratospheric wave geometry configurations for the different modes. Viewed in this way, it becomes obvious that there is not much difference between the stratospheric part of forced and unstable waves, hence results from the normal mode analysis are relevant to the observed quasi stationary waves, as well as to the eastward propagating wave 2 observed in the southern hemisphere. The feature which affects wave structure most is the existence of a turning point, which reflects the waves downward. The zonal-vertical structure is directly affected because the orientation of phase lines in the longitude-height plane is indicative of the direction of wave propagation. When damping is added in the vicinity of a turning point, it cuts down the reflection, causing the phase tilt with height to be more westward. Also, it decreases the amplitude of waves that are propagating in the region of damping. Temperature amplitude is also very sensitive, because it is a vertical derivative of geopotential height. We commonly find a node in temperature at or above the turning point, along with a peak in geopotential height (in the case of partial reflection we have an 'almost node' in temperature).

When the basic state varies with latitude as well as with height, we need to worry about meridional propagation (chapter 5). In general, given an index of refraction, there is no unique way to separate wave propagation in the meridional and vertical directions a priori, without obtaining the full solution first. Given a steady state solution, we can however calculate meridional and vertical wavenumbers. In the stratosphere, the meridional wavenumber is determined by the shape of the polar night jet, which acts as a waveguide. As a result, the meridional wavenumber is insensitive to the zonal wavenumber and phase speed of the wave, as well as to the damping[3] and shape of the forcing at the bottom[4]. This leaves only the vertical wavenumber free to vary with the zonal wavenumber and phase speed of the wave. Most important, the meridional and vertical wavenumbers that are calculated from a steady state solution to a given basic state, using arbitrary forcing and damping, are a diagnostic of the *basic state* wave propagation characteristics. This diagnostic allows

[3]The insensitivity to damping holds as long as the equatorial damping does not get into the waveguide and as long as it does not vary on spatial scales of the order of or smaller than the wavenumber.

[4]The shape of the forcing at the bottom affects only the lowest scale height. Above that, the meridional wavenumber is determined by the shape of the waveguide.

us to 'see' turning points (for propagation in the vertical). We also find that the effects of turning points and damping on the wave phase and amplitude structure are qualitatively like in the one dimensional model. The analogy to the one dimensional model is further tested by comparing the solution in the middle of the waveguide to that of an approximate one dimensional model, with the basic state, damping, and meridional wavenumbers taken from the middle of the waveguide (section 5.3.6). The sensitivity of the response to zonal wavenumber is very similar in the two models, with the exception that resonant wavenumbers are not found in the two dimensional model. This is due to leakage of the perturbation to the equator. Quantitatively, the 1D approximation is not very good, with some of the difference also being due to leakage to the equator. Since the zonal and vertical wavenumbers are diagnostics of the basic state propagation characteristics, they are relevant for time evolving waves as well as for steady state. Note that using an instantaneous wave structure to calculate wavenumbers is meaningless if we have large time variations.

There are a few additional points to make. In solving the 1D model for normal modes, we find that when stratospheric turning points exist, the downward reflection of the wave interferes with the interaction of the mode with the critical level in the troposphere, causing one or more wavenumbers to be exponentially neutral (section 4.4.2). In the real world, however, we do not expect strong interference to develop because the reflection is from a surface that is not necessarily simple geometrically, the damping will most likely reduce the reflection, and by the time the wave propagates up to the turning point and back, the basic state may change. Also, since it takes time for the wave to propagate up to the turning point and back down to the critical level, we expect the modes initially to grow like the faster growing adjacent wavenumbers. This highlights the fact that the growth mechanism of the neutral wave and the waves adjacent to it are similar, namely, an interaction with the critical level in the troposphere. Generalizing to the neutral wavenumber of the in the Charney model, which defines the separation between the long Green modes and the medium scale Charney modes, we note that the physical growth mechanism of Green and Charney modes is the same, and the distinction between them is an artifact of having a turning point.

Solving the normal mode problem also allows us to identify other possible basic state configurations that support different kinds of modes, differing in the location of interaction with the mean flow. Basic states that have one or more regions of negative PV gradients in the stratosphere along with one or more critical levels may support modes that draw energy from the mean flow at the stratospheric critical level. However, we can exclude these modes as being a dominant source of variability in

the stratosphere because the phase speeds of the dominant observed waves are generally smaller than stratospheric winds (which excludes the possibility of stratospheric critical levels). Basic states that are more likely to support internal stratospheric instability, given the observed phase speeds, are found in fall and spring of both hemispheres, when the zonal wind has a minimum in the lower stratosphere. Waves during these periods, however, are typically small.

Finally, we side step from looking at vertical structures in terms of wave geometry, and take a different approach to stratospheric waves. In chapter 6 we develop a diagnostic technique to study the evolution of a 'wave activity packet' within a stratospheric planetary Rossby wave. We view the wave field as the propagation of many such wave packets with a velocity analogous to group velocity. We define a coordinate system that follows this propagation, and use it to track the packets as they propagate. This also allows us to keep track of the evolution of the wave activity in the packet, and to distinguish between the contribution to the wave activity budget of wave refraction, damping, and time variations in the source. It also allows us to obtain a time scale for propagation of the waves through the stratosphere. We also relate our coordinate system to Karoly and Hoskins (1982) ray tracing, which highlights the diagnostic vs. analytic nature of these two calculations, respectively. Theoretically, this wave activity approach is mostly an interesting alternative way to view waves and their evolution. In some cases it is particularly illuminating. For example, a perturbation that is forced at the tropopause initially concentrates into the center of the wave guide as it propagates along it. After some time it spreads out and leaks through the equatorial boundary and tunnels to the critical level at the equator. This is a consequence of having a leaky wave guide, where initially the wave only feels the local index of refraction, but eventually, when the perturbation fills the waveguide, it 'sees' the equator where the index of refraction and the damping are large.

8.3 Applying to observations

In the observational part of our study we analyze a few cases where the vertical structure of the waves is observed to vary on time scales of a few days (chapter 7). We analyze one event from mid-winter in the southern hemisphere (July-August 1996) and one from late-winter/spring (September 1996). During each episode the waves have a characteristic vertical structure which occasionally changes for a few days, mostly suggestive of downward reflection. We use our wavenumber diagnostic to diagnose the propagation characteristics of the basic state (by finding the steady state

wave solution). As in our model, we find turning points from which waves can reflect downwards. There are two kinds of turning points, one due to large zonal winds, as in the Charney-Drazin criterion, and the other due to small or negative PV gradients forming in the region of positive vertical wind curvature above the jet maximum. The decay of the waves above the turning point is much sharper for the latter. The zonal mean jet has a well observed seasonal cycle- its peak moves downward and poleward towards the end of winter, such that it peaks in mid-stratosphere in September (e.g. Shiotani and Hirota, 1985). As a result, in late winter a region of small or negative PV gradients forms in the upper stratosphere, resulting in a turning point of the second kind. While in mid-winter we see a turning point in the stratosphere only occasionally, in September we usually observe one at around 5-7 scale heights. The observed seasonal cycle in wave structure is consistent with the basic state (section 5.4.2). In mid-winter the waves have a westward phase tilt with height and their amplitude increases in the stratosphere, corresponding to an upward propagating wave. In late winter, when the jet shifts downward and a turning point forms, the geopotential height amplitude peaks in mid-stratosphere, slightly above the turning point, while the temperature has a node. Also, the waves have a smaller westward phase tilt with height.

We also study the variations in wave structure on daily time scales, and look for consistent variations in the wave geometry of the basic state. In the cases we studied we find that the observed time evolution is a wave-mean flow interaction, where the wave responds qualitatively linearly to the basic state changes, which appear to be wave-induced (based on EP flux divergence calculations). A quantitative consistency check is a calculation of vertical propagation time scales, where the time over which vertical structure changes occur is related to the time it takes the wave to propagate to the turning point. Our wave based coordinate diagnostic allows us to calculate the time it takes a wave packet to propagate through the stratosphere[5]. Calculations, based both on our diagnostic and on Karoly and Hoskins (1982) ray tracing, show a consistency of wave response times. Apart from confirming the relevance of linear wave propagation theory to observations on short time scales, this suggests the observations are at least qualitatively correct. This is important since much of the important variations in the basic state occur above 5 mb, the level above which satellite retrievals become less reliable. Our analysis also suggests that in September, when the polar vortex starts its breakdown, wave nonlinearities are noticeable in terms of

[5]We need to calculate this time in the initial stages of wave growth because once partial downward reflections develop, our diagnostic shows longer time scales because the wave is a superposition of the upward and downward propagating components.

their effect on the vertical wave structure, and at least qualitatively, nonlinearities can act as damping on the waves.

Finally, we apply our wave based coordinate diagnostics to observations, for two kinds of calculations. The first has to do with tracking wave packets, and observing the evolution of the wave field in this way (one use already mentioned is the estimation of propagation time scales). This type of calculation works quite well, especially in highlighting the time evolution of the wave in a leaky waveguide, with or without a turning point (section 7.1.4). The second kind of calculation has to do with the wave activity budget, and the contributions to it from various terms. Advantages of this diagnostic are that we can follow a wave packet, keep track of its wave activity, and distinguish between the various factors that contribute to it. The accuracy of wave activity calculations from observations is very low, however, causing the uncertainties in the calculations to be too large to really make sense of.

8.4 Discussion

In the introduction, we described a few of the outstanding issues regarding stratospheric planetary waves that have been the general motivation of our work. In this section we will comment briefly about how our results relate to these issues.

The extent to which linear wave theory explains the structure and time evolution of observed planetary waves at a given time or season in the stratosphere is still debated (e.g. O'Neill and Pope, 1988 and references therein). An obvious way to test this is to compare observed waves with modeled ones, however, we need to account for the differences between them. Differences are expected because models are sensitive to details of the basic state and damping, both of which are not determined from observations in great accuracy. This sensitivity, on the other hand, makes it hard to determine why the observations deviate from modeled waves in any given case. It is also unclear whether discrepancies are due to a model deficiency or to observational uncertainty, especially if we use higher order diagnostics. It is important to be able to generalize features of the waves to different basic states and damping. To this end, looking at time variations of vertical wave structure can be very useful, because the response to time variations in the basic state is reflected in the large scale wave structure, which is easily and relatively accurately observed. Also, the wave geometry framework allows us to estimate the effects of unknown parameters like damping, hence to account at least for some of the discrepancies between modeled and observed waves.

Most of the variations in vertical structure, both in models and in observations,

have to do with reflection of the waves at a turning surface. Since there is no source for the waves in the mesosphere, waves with an eastward phase tilt with height (which are downward propagating) have to be transient. Sources of transience may either be variations in the forcing at the bottom, in the presence of a turning surface, or variations in the basic state that either create or shift an existing turning surface. It is interesting that in most cases we analyzed both seem to happen. It is unclear whether this is only coincidence, and if it is not, which is the cause and which the effect. This leads us to the following issue: what causes stratospheric waves to grow at some times and not at others. In particular, waves appear in episodes of a few weeks. To what extent does the existence of waves depend on the tropospheric forcing, and to what extent does it depend on the stratospheric basic state? Is it the lower stratospheric basic state or is it the wave geometry in the middle and upper stratosphere as well? A better understanding of the sources of waves in the troposphere, as well as a comprehensive observational study of both the troposphere and stratosphere are needed. Our wavenumber diagnostic may be a useful diagnostic for studies of these issues.

Another aspect of the question of what causes stratospheric waves to grow when they do is the seasonal cycle in wave amplitude. As we have shown in the introduction (figures 1.4, 1.5), some years in the southern hemisphere have a mid-winter minimum in wave variance, but other years show a succession of wave events throughout the winter. One explanation for the mid-winter minimum (Plumb, 1989) is that the waves respond linearly to the seasonal cycle of zonal mean winds, and that in mid-winter the zonal mean wind becomes large enough for a turning point to form, which causes the wave response to decrease. Although we have analyzed a year where we do not find a characteristic mid-winter minimum cycle, our results suggest that a turning point does not necessarily inhibit wave growth. On the contrary, we are more likely to find a turning point in late winter, when the jet peaks in mid-stratosphere. This turning point is of the second kind, which is due to the strong positive curvature above the jet, and not due to the winds being too strong as in the Charney-Drazin criterion. Results of our study do suggest however, that at least in terms of their vertical structure, observed waves respond quasi-linearly to the seasonal evolution of the basic state in the southern hemisphere middle to late winter. Note that Wirth (1991) did not manage to reproduce the seasonal cycle using a Matsuno type steady state model using monthly mean basic states (he did not manage to reproduce the early winter peak). Calculations of vertical and meridional wavenumbers, using the steady state solution to observed monthly mean winds of various years, may shed some light on this issue. Also, the question remains as to what causes some years

to have the characteristic mid-winter minimum cycle, some years to only have a late winter peak, and other years to have a succession of wave events? In particular, are the differences due to the tropospheric forcing or to the stratospheric basic state structure or to both?

We have chosen to study the transience of the waves as a daily time scale variability in the vertical structure. This makes sense given that we find a corresponding variability of the basic state. A different approach, which is common in the literature, is to assume the total wave field is made up of many different modes and to study the variability using a time-space Fourier decomposition of the waves (e.g. Mechoso and Hartmann, 1982). We argue that for much of the variability, these are two different approaches to looking at the same thing, since downward reflection and the associated structure changes appear as an eastward phase propagation at some levels, and they are quite abundant in the years we have looked at. A more comprehensive study of more years is needed to estimate how much of the transience can be accounted for by vertical structure changes.

Finally, we should note that the observational examples we have shown are from the southern hemisphere, however, we believe that our results are relevant to the relatively quiescent periods of the northern hemisphere, when waves behave more linearly.

Appendix A

The 'Virtual Satellite' problem

A.1 The basic state temperature

The temperature field we use for our retrieval exercise consists of a zonal mean and a perturbation. We calculate the zonal mean from the thermal wind relation as follows:

$$T(y,z) = -\int_{y_o}^{y} \frac{f_o T_o}{g} U_z dy + T(y_o, z) \tag{A.1}$$

where z is log pressure, $H_o = \frac{RT_o}{g}$ is a reference density scale height, T_o a reference temperature and R the gas constant. $T(y_o, z)$ is the temperature profile that corresponds to the basic state N^2. y_o is chosen at some mid-channel latitude (we use $y_o = 3900km$). Other model details and parameters are given in appendix B. An example of a basic state temperature field is shown in figure 3.8.

The temperature perturbation field is calculated from the model geopotential height perturbation:

$$\frac{\partial \varphi}{\partial z} = g\frac{T}{T_o} \tag{A.2}$$

since our model is linear, the wave amplitude is arbitrary, and we choose it specifically for each case, depending on the wave field itself. Wave amplitudes we use vary between a few degrees and 20 degrees Kelvin, which is the normal range of wave amplitudes observed in the stratosphere.

A.2 The minimum variance constraint

The diagonal constraint:
We specify the variance of temperature as the diagonal terms in the error covariance matrix, and use the zonal mean basic state at the middle latitude as the constraint.

We also use other profiles as the constraint in some runs.

The non-diagonal constraint:

The constraint is calculated using 50 days of a time dependent run where we have a superposition of a transient wave with a phase speed of $15\frac{m}{sec}$ and a stationary wave, both wave 1. We turn the forcing on, equally for both waves, over a period of 8 days and let them evolve. The wave field changes periodically (with a period of 20.6 days), due to the different phase superpositions of the standing and traveling waves. The vertical structures of the combined wave fields are quite different from the stationary wave alone. Our control experiments use all the grid points between y=1 and y=4, and all days, as a climatology from which to calculate the constraint. The waves in this run reach a maximum amplitude of 10°K. In our 'control' constraint, we multiply the wave by an amplitude factor of 4.0 before adding it to the basic state. The constraint profile is just the time and spatial average of temperature, and the error covariance matrix is calculated as shown in section 3.3.3, in the footnote. The choice of an amplitude factor of 4 for the control run needs to be explained because the control run waves reach an unrealistic amplitude of 40°K on some of the days. However, the total variance (square root of the diagonal elements of the covariance matrix, see table A.1) reaches much more reasonable values, which is why we chose this value as the control. It is important to remember that in reality, there are variations of the basic state as well as variations in the forcing at the bottom, which would be additional sources of variance.

Variations on the control run include choosing a different latitude range over which to average and choosing a different linear wave amplitude (ranging from 0.0 to 8.0). We are free to calculate an error covariance matrix using one climatology and the constraint itself using another. None of our results depend on the exact parameter values we use.

A spatially dependent constraint:

We use a specific wave field (a single day out of the runs described above or a wave field from a steady state model run) to specify a different constraint profile at each grid point (referred to as the *constraint field*). The error covariance matrix is calculated in the fashion described above. We assume it is the same for all grid points, out of computation time considerations.

A.3 The operational constraint

A dynamic data set of a few thousands co-located radiance measurements and radiosonde profiles is constantly compiled out of the previous few weeks of observations.

Height (km)	Variance (°K)
14.0	3.2
20.5	10.2
27.0	11.2
33.5	11.7
40.0	12.5
46.5	13.4
53.0	11.0
59.5	5.1

Table A.1: The variance of the standard constraint (see text for details).

For each grid point, a constraint profile is chosen by searching this data set for the 10 closest sets of radiances. The mean of the corresponding temperature profiles is taken to be the constraint profile. In the upper stratosphere, where there are no radiosondes, the profiles are extrapolated upward using the covariance of a rocketsonde data set. There are different rocketsonde data sets for high, middle and low latitudes and the different seasons. Note that most rocketsonde stations are in the northern hemisphere. The constraint field is non-zonal, since its shape at high altitudes is determined by the observed field at lower altitudes (the radiosonde measurements) and its upward extrapolation using the rocketsonde data. Since the constraint has a set of radiances that is close to the measured ones, it is to some extent close to the true profile, however, all the limitations of the observing system apply here.

The error covariance matrix is calculated once every few weeks from the radiosonde/rocketsonde data set, and it is the same for all grid points in a specific region (high, middle or low latitudes). The constraint is therefore a non-diagonal one.

Appendix B

The models used

B.1 The 1 dimensional model

B.1.1 Parameters and nondimensionalization constants

The linear QG pseudo-PV equation in *height* coordinates is derived as in Pedlosky (1987). The geopotential stream function and temperature variables are defined as follows:

$$\phi^* = \frac{p_{tot} - p_s}{\rho_s} \tag{B.1}$$

$$T^* = \frac{T_{tot} - T_s}{T_s} \tag{B.2}$$

$$\theta^* = \frac{\theta_{tot} - T_s}{\theta_s} \tag{B.3}$$

where the subscript *tot* denotes the total dimensional field, while the subscript s denotes a representative *horizontal* mean of the total dimensional field. Note that since $()_s$ is a horizontal average, $()_{tot} - ()_s$ has a zonal mean component that varies with latitude.

Equations 4.1-4.3 are derived by nondimensionalizing the linear pseudo-PV equation and the definition of PV in terms of ϕ as follows (' denotes a deviation from the zonal mean, T^*, θ^* and ϕ^* are as defined above, otherwise, $*$ denotes the dimensional

variables):

$$\begin{aligned} z &= \frac{z^*}{H} \\ (k,l) &= (k^*,l^*)L \\ u' &= \frac{u'^*}{U_T} \\ v' &= \frac{v'^*}{U_T} \\ w' &= \frac{w'^*}{U_T}\frac{L}{H} \\ U &= \frac{U^*-U(0)}{U_T} \\ c &= \frac{c^*-U(0)}{U_T} \\ \phi' &= \frac{\phi'^*}{U_T f_0 L} \\ \theta' &= \frac{gH}{U_T f_0 L}\theta'^* \\ \overline{\theta} &= \frac{gH}{U_T f_0 L}\overline{\theta^*} \\ N^2 &= \frac{N^{*2}}{N_0^2} \\ q_y &= \overline{q}_y{}^* \frac{H^2 N_0^2}{U_T f_0^2} \end{aligned} \tag{B.4}$$

N^* is given by $N^{*2} = \frac{g}{\theta_s}\frac{d\theta_s}{dz^*} = \frac{g}{T_s}(\frac{dT_s}{dz^*} + \frac{g}{C_p})$, where C_p is the heat capacity of air at constant pressure and g the gravitational acceleration.

There are two nondimensional parameters, $\beta_e \equiv \frac{\beta H^2 N_0^2}{U_T f_0^2}$ (equations 4.3, 4.7) which is the nondimensional β, and a factor multiplying the horizontal wavenumbers $\mu \equiv \frac{HN_0}{Lf_0}$, which we set equal to 1 by choosing the horizontal scale to be the radius of deformation: $L = L_d = \frac{N_0 H}{f_0}$.

Standard values are used for the different parameters: $U_T = U^*(z^* = H) = 22msec^{-1}$, $T_0 = 285K$, $N^2{}_0 = 1.1\times10^{-4}sec^{-2}$, $H = 8.9km$, $\beta = 1.14\times10^{-11}sec^{-1}m^{-1}$, $\beta_e = 0.29$. Unless specified differently, the tropopause is taken at $z = 1$ i.e. at 8.9km. The top of the 'troposphere-stratosphere' model is chosen at $z = 6$ (53.5 km). Also, $f_o = -1.26\times10^{-4}$ (corresponding to a latitude of -60°), and the radius of deformation is $L_d = \frac{N_o H_o}{f_o} = 745km$.

The numerical method, which is described in Harnik and Lindzen (1998), is essentially the algorithm used by Kuo (1979). We have 400 grid points in the vertical (a grid spacing of 134m for the top at $z^* = 53km$).

B.1.2 The boundary conditions

Our lower boundary condition is either a rigid surface or an Ekman pumping condition. We use the temperature equation on the lower boundary, and specify the vertical velocity according to the boundary condition.

For a rigid surface, the ground is our lower boundary and we set the vertical

velocity to zero there:

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)\theta'+v'\frac{\partial\overline{\theta}}{\partial y}=0 \tag{B.5}$$

To implement an Ekman parameterization, our lower boundary represents the top of the Ekman boundary layer, and we use the Ekman pumping parameterization for the vertical velocity there:

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)\theta'+v'\frac{\partial\overline{\theta}}{\partial y}=-w'N^2=-N^2E_k\zeta' \tag{B.6}$$

where $E_k \equiv \frac{N_o}{U_T}\sqrt{\frac{\nu}{2f_o}}$ is the Ekman damping coefficient, ν is the vertical eddy viscosity coefficient, and ζ' is the vertical component of the vorticity. To solve, we write θ' and ζ' in terms of the geopotential stream function (4.4), and solve along with equation 4.6.

The top boundary condition is a radiation condition. To implement it, wind and temperature are held constant at the top scale height of the model. We use the transformation 4.9 to put the equation in canonical form (equation 4.8). Taking into account that N^2 is constant at the top, the solution becomes analytically tractable, and has the form of a superposition of an upward and a downward propagating waves. Only the upward propagating part is chosen. The solution is then transformed back into our original variable φ:

$$\varphi_z+\left(-\frac{1}{2}+in_{ref}(c)\right)\varphi=0 \quad at\ z=Top \tag{B.7}$$

where

$$n_{ref}=\sqrt{-(k^2+l^2)^2N^2-\frac{1}{4}+\frac{N^2\overline{q}_y}{U-c}} \tag{B.8}$$

and we choose the square root that yields $Im(n_{ref}) \geq 0$.

B.2 The 2D steady state β-plane channel model.

We use the 2D model to solve the *forced* problem (equation 5.2). The nondimensionalized linear QG β-plane pseudo-PV equation in *log-pressure* coordinates, as well as the definitions of PV, winds and temperature in terms of the geopotential stream function (equations 4.1-4.3 and 4.4) are similar to their height coordinate versions, only the geopotential stream function is the geopotential height, and the temperature variables are the total temperature fields:

$$z^* \equiv -H_o\ln\frac{p}{p_o}=H_o\int_0^{z_g}\frac{dz_g}{H}=T_o\int_0^z\frac{dz}{T^*} \tag{B.9}$$

$$d\phi^* = g dz_g \qquad \phi^* = g z_g(z) \tag{B.10}$$

$$T^* = T_{tot} \tag{B.11}$$

$$\theta^* = \theta_{tot} \tag{B.12}$$

The same notation as in section B.1.1 is used, with the exception that z^* and z are the dimensional and nondimensional log pressure vertical coordinates, while z_g is the geometric height. $H \equiv \frac{RT^*}{g}$ in the density scale height, and $H_o \equiv \frac{RT_o}{g}$ and T_o are reference scale height and temperature. All variables (including the newly defined ϕ^*, T^*, and θ^*) are nondimensionalized as in the 1D model, and z^* is nondimensionalized by H_o. The parameters we use in our model control run are: $H_0 = 7km$ (corresponding to $T_o = 239°K$) $U_T = 30\,m\,sec^{-1}$, ${N^2}_0 = 4.1 \times 10^{-4} sec^{-2}$, $\beta = 1.31 \times 10^{-11}\,sec^{-1}\,m^{-1}$, $f_o = -1.19 \times 10^{-4}$ (corresponding to a mid-channel latitude of -55°), and the radius of deformation is $L_d = \frac{N_o H_o}{f_o} = 1190km$.

The bottom of our model is at 2 scale heights (14 km). We force the model by specifying a zonal wavenumber and a wave amplitude and phase at the bottom. At the top and side boundaries we set the perturbation to zero. We have a sponge layer at the top and equatorial boundaries, resulting in an absorption of the wave there. We make sure reflections from the boundaries and the sponge layer are minimal by running the model with a larger sponge layer (by increasing the model domain) and making sure the results are similar. We find that it is necessary to include Rayleigh damping of momentum in order to absorb the waves in the sponge layers (using only Newtonian damping results in much less absorption). The polar boundary, which has no sponge layer is fully reflecting. Since in spherical coordinates the index of refraction becomes negative close to the pole (equation D.14), we expect Rossby waves to be reflected equatorwards anyway. Also, in our β-plane model, waves reflect off the poleward side of the waveguide (section 5.3.1).

The sponge layer damping is specified as follows:

$$\begin{aligned} r(y,z) = \alpha(y,z) \;=\; & \tfrac{1}{2}A_1\left(\tanh(\tfrac{z-A_2}{A_3}) - \tanh(\tfrac{z_{bot}-A_2}{A_3})\right) + \\ & \tfrac{1}{2}B_1\left(\tanh(\tfrac{y-B_2}{B_3}) - \tanh(\tfrac{y_{pole}-B_2}{B_3})\right) \end{aligned} \tag{B.13}$$

where the Rayleigh (r) and Newtonian (α) damping coefficients are equal. In the control run $A_1 = 1.5$, $A_2 = 10.5$, $A_3 = 2.5$, $B_1 = 1.0$, $B_2 = 8.75$, $B_3 = 1.5$. To test the effect of thermal damping, we raise the sponge layer by 5 scale heights (i.e. $A_2 = 15.5$), and add the following α (following Dickinson, 1969b):

$$\alpha = 0.45(days^{-1}) e^{-\left(\frac{z^*-50km}{13km}\right)^2} + 0.05(days^{-1}) \tag{B.14}$$

We solve equation 5.2 for the steady state solution by finite differencing the equations and using a direct solver, based on Lindzen and Kuo (1969). We have 71 grid points in the vertical direction and 64 in latitude. The top of our model (unless specified otherwise) is at 15 scale heights (105km), and the latitudinal boundary is at y=10.27.

The basic state wind and temperature are specified analytically. The latter we specify to be similar to a US standard winter midlatitude stratosphere profile. N^2 is then calculated from temperature, and $\overline{q}_y$ from U and N^2, using equation 4.3. The control run basic state is shown in figure 5.1.

B.3 The 2D time dependent β-plane channel model.

Our time dependent model consists of solving equation 4.1, assuming a normal mode solution in the zonal direction:

$$\phi' = \varphi'(y,z,t)e^{ikx} \tag{B.15}$$

plugging in equation 4.1, we get:

$$\frac{\partial q'}{\partial t} \equiv F(q') = -ik(Uq' + \varphi'\overline{q}_y) + \mathcal{D}' \tag{B.16}$$

where $\mathcal{D}$ is a damping term. We solve this as follows:

1. Start with a PV perturbation distribution, $q'(t)$.
2. Invert $q'(t)$ to get $\varphi'(t)$, using equation 4.2.
3. Calculate $\frac{\partial q'}{\partial t}(t)$ using $q'(t)$ and $\varphi'(t)$ (equation B.16).
4. Integrate equation B.16 in time to get $q'(t+\Delta t)$.
5. Return to step 2 and repeat for $t = t + \Delta t$.

To get geopotential height we invert equation 4.2 using the same numerical algorithm, and the same boundary conditions ($\phi = 0$ at the top and sides, ϕ specified at the bottom) as in the steady state model.

The time integration of equation B.16 is done using a third order Adams-Bashford method (we follow Durran, 1991):

$$q'(t+\Delta t) = q'(t) + \frac{\Delta t}{12}\left[23F(t) - 16F(t-\Delta t) + 5F(t-2\Delta t)\right] \tag{B.17}$$

We integrate the first two time steps using a second order Runge-Kutta method.

The basic state and the forcing at the bottom are specified every time step, and are either constant or time dependent. The damping is held constant in all runs, and is equal to the β-plane model damping, except that we add a constant damping in the form of an imaginary phase speed (constant and equal Newtonian and Rayleigh damping coefficients) with a damping time scale of 25 days to assure numerical stability. The spatial resolution is the same as in the β-plane model. Unless otherwise specified, we output our results every 50 time steps ($\Delta t = 0.02$) which is every 1 nondimensional time unit. In dimensional terms, a model time unit is $\frac{U_T}{L_d} = 0.46 days$. For simplicity we call this half a *model day.*

B.4 The 2D steady state spherical hemispheric model.

Since we use the spherical coordinate model mostly in order to calculate the steady state solution to observed basic states, we keep the variables dimensional and use log-pressure coordinates. The dimensional, linear, QG PV equations, written in terms of a geopotential stream function are described in appendix D. We solve for the steady state solution to a prescribed forcing using the same numerical algorithm, and essentially the same forcing and boundary conditions as in the 2D β-plane model, with the following differences. The latitude and width of the sponge layer are $B_2 = -20°$ and $B_3 = 10°$ respectively (equation B.13). The model domain is the southern hemisphere (-90°to 0°). The vertical resolution is the same as in the β-plane model but latitudinal resolution is the same as the operational observations product (2°latitude)[1].

The basic state is either taken from observations, or specified analytically. In the latter case, the wind and temperature profiles are as in the β-plane model, but the corresponding $\overline{q}_y$ is different (compare equations 4.3 and D.8). When we use observations, we first extend the fields in the vertical to the model domain (15 or 20 scale heights, depending on the run) by keeping wind and temperature constant above 0.4mb, and then interpolate in the vertical to the high resolution model grid using a cubic spline interpolation. We calculate the PV gradients from the high resolution interpolated fields using equation D.8.

[1]This resolution is of the operational product, which is interpolated from the satellite retrievals. The actual resolution of the observations depends on the scanning patters of the satellite instruments (see the Kidwell, 1986, for actual numbers).

Appendix C

Tracking wave packets: a wave activity based coordinate

C.1 The relation between the Jacobian and $\nabla \cdot \vec{V}_a$.

We show the relation between a velocity field and the Jacobian of the transformation to coordinates that follow the velocity (equation 6.12). The mathematics was developed for studying the kinematics of fluids, but it applies to our wave activity flow field as well. The following derivation is taken from Aris (1962).

We define a coordinate system that follows our velocity field, and denote a unit volume in this new coordinate as a packet. At time t, the packet which is at the Cartesian position (x_1,x_2,x_3) is denoted by its initial ($t = 0$) Cartesian position ($\xi_1(\mathbf{x},t)$,$\xi_2(\mathbf{x},t)$,$\xi_3(\mathbf{x},t)$). Thus (ξ_1,ξ_2,ξ_3) follow a single packet (material coordinates).

A volume element in the Cartesian coordinates, relates to a volume element in the material coordinates as follows:

$$dV = dx_1 dx_2 dx_3 = \frac{\partial(x_1, x_2, x_3)}{\partial(\xi_1, \xi_2, \xi_3)} d\xi_1 d\xi_2 d\xi_3 = J \cdot dV_0 \tag{C.1}$$

where J is the Jacobian of the transformation:

$$J = \frac{\partial(x_1, x_2, x_3)}{\partial(\xi_1, \xi_2, \xi_3)} = \begin{vmatrix} \frac{\partial x_1}{\partial \xi_1} & \frac{\partial x_1}{\partial \xi_2} & \frac{\partial x_1}{\partial \xi_3} \\ \frac{\partial x_2}{\partial \xi_1} & \frac{\partial x_2}{\partial \xi_2} & \frac{\partial x_2}{\partial \xi_3} \\ \frac{\partial x_3}{\partial \xi_1} & \frac{\partial x_3}{\partial \xi_2} & \frac{\partial x_3}{\partial \xi_3} \end{vmatrix} \tag{C.2}$$

To calculate the material derivative of J we use the rule that the derivative of a determinant equals the sum of three determinants with one row differentiated at a time. Using $\frac{d}{dt}\left(\frac{\partial x_i}{\partial \xi_k}\right) = \frac{\partial v_i}{\partial \xi_k}$, the result of differentiating only the first row is:

$$\begin{vmatrix} \frac{\partial v_1}{\partial x_k}\frac{\partial x_k}{\partial \xi_1} & \frac{\partial v_2}{\partial x_k}\frac{\partial x_k}{\partial \xi_2} & \frac{\partial v_3}{\partial x_k}\frac{\partial x_k}{\partial \xi_3} \\ \frac{\partial x_2}{\partial \xi_1} & \frac{\partial x_2}{\partial \xi_2} & \frac{\partial x_2}{\partial \xi_3} \\ \frac{\partial x_3}{\partial \xi_1} & \frac{\partial x_3}{\partial \xi_2} & \frac{\partial x_3}{\partial \xi_3} \end{vmatrix}$$

where we use the summation rule. For $k = 2$ we get $\frac{\partial v_1}{\partial x_2}$ multiplying a determinant with the first and second rows equal, which is zero. Similarly, for $k = 3$ we get zero. Thus we are left only with the $k = 1$ term: $J \cdot \frac{\partial v_1}{\partial x_1}$. Differentiating the second and third rows, and summing, we get:

$$\frac{dJ}{dt} = J\left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right) = J\nabla \cdot \vec{V}_a \tag{C.3}$$

resulting in equation 6.12.

C.2 Calculating the $s - r$ coordinate from $\vec{V}_a$

We calculate our coordinate system by integrating along $\vec{V}_a$. The variables defined in the previous section (C.1) relate to the variables we use in chapter 6 as follows:

$$\begin{aligned} x_1 &= x_s \\ x_2 &= y_s \\ x_3 &= z_s \\ \xi_1 &= x_s \\ \xi_2 &= r \\ \xi_3 &= s \end{aligned} \tag{C.4}$$

where the subscript s denotes a Cartesian location of the $s - r$ coordinate grid point (or wave packet). Since in our case everything is constant along the zonal direction, $x_1 = \xi_1 = x_s$, we drop the first coordinate and are left with a 2 dimensional system. The Jacobian is therefore:

$$J = \frac{\partial x_2}{\partial \xi_2}\frac{\partial x_3}{\partial \xi_3} - \frac{\partial x_2}{\partial \xi_3}\frac{\partial x_3}{\partial \xi_2} = \frac{\partial y_s}{\partial r}\frac{\partial z_s}{\partial s} - \frac{\partial y_s}{\partial s}\frac{\partial z_s}{\partial r} \tag{C.5}$$

To obtain $(y_s(y,z),\ z_s(y,z))$, we integrate the following set of equations for a set of equally spaced emanation latitudes (y_0):

$$\frac{dy_s}{dt} = V_{ay} \tag{C.6}$$

$$\frac{dz_s}{dt} = V_{az} \tag{C.7}$$

$$\frac{ds}{dt} = 1 \tag{C.8}$$

$$\frac{dr}{dt} = 0 \tag{C.9}$$

subject to the following initial conditions at $t = 0$:

$$\begin{aligned} z_s(t=0) &= 2 \\ y_s(t=0) &= y_0 \\ s(t=0) &= 0 \\ r(t=0) &= y_0 \end{aligned} \tag{C.10}$$

r equals the latitude at which the $\vec{V}_a$ integral line emanates at the bottom ($z = 2$ scale heights), and s is the time it takes to reach the given location ($y_s(y,z)$,$z_s(y,z)$) from the initial location (y_0,2). $\vec{V}_a$ is non-divergent in the new coordinate system, and constant s lines are spaced proportionally to $|\vec{V}_a|$ in geometric space.

Numerically, the calculation is done as follows: Starting from the bottom of our model, we integrate equations C.6-C.9 using a Runge-Kutta integration method, to obtain the location of the wave packets after one time step (y_s,z_s). From our model generated wave we calculate the velocity field $\vec{V}_a$ on the regular model grid (y,z), and interpolate it to (y_s, z_s), using a two dimensional fifth order polynomial interpolation[1]. We then use this interpolated value to integrate equations C.6-C.9 by an additional time step, and so on. Our time step is $\frac{1}{16}day$, and we output the results only every 4 integration steps, resulting in a resolution of 0.25 day for s[2]. We stop our integration after 400 time steps (100 s points). When the packet location reaches the boundaries of our model domain, we set $\vec{V}_a$ to zero. Typically, we have 80 r grid points between $y = 0.2 - 5.5$.

In the time dependent model run (see appendix B for details of the model), we calculate the $s - r$ coordinate in the same way, taking into account that $\vec{V}_a$ is now a function of time. In general, our model integration time step is larger than the coordinate integration time step (the model Δt is $\frac{1}{2}$ day, while the coordinate in-

[1]The interpolation of $\vec{V}_a$ to a given point (y_s, z_s) is done from a 6×6 sub-grid of the regular model grid, centered (as much has possible) as around the point y_s, z_s. If y_s, z_s is closer than three grid points to one of the boundaries of our domain, we use a 6×6 grid starting from that boundary.

[2]We need a small integration step because there are regions of very large $\vec{V}_a$, especially near regions of very small, zero or negative PV gradients, where the wave activity is small. On the other hand, we want an integration long enough for the packets to traverse the stratosphere, and we do not want more than 100 grid points of s.

tegration Δt is $\frac{1}{16}$ day). We therefore interpolate $\vec{V}_a$ at each grid point to all the integration time steps using a fourth order polynomial interpolation in time. We use a linear interpolation for times shorter than the model Δt. We then proceed as in the steady state case, by interpolating $\vec{V}_a$ in space to $y_s(y,z,t), z_s(y,z,t)$, integrating equations C.6-C.9 in time, interpolating in space, and so forth.

We follow the same scheme when we use observations, only the observations have an even lower time resolution of 1 day. Also, we use the spherical coordinate version of the coordinate calculation, which is described in appendix D.

C.3 Transforming scalar fields between the geometric and $s-r$ grids

The transformation of scalar fields onto the new $s-r$ grid is simply an interpolation of these fields from the regular y,z model grid onto the irregular y_s, z_s grid. Since some of our calculations involve differentiating the scalar field in $s-r$ space, we use a bicubic spline interpolation, which insures continuity of first and second order derivatives. The interpolation back to geometric space is slightly more involved since it is an interpolation from an irregular to a regular grid. Since we do this part of our calculations using MATLAB, we use the 'griddata' routine with the 'cubic' option, which fits a surface to the irregularly spaced field values (e.g. $A(y_s, z_s)$), and interpolates this surface to the regularly spaced grid, using a triangle-based cubic interpolation.

Appendix D

Spherical coordinates

D.1 The PV equations.

Formulating the quasi geostrophic equations on a sphere is more complicated than on a β-plane, for two reasons. First, the QG scaling requires the characteristic length scales of the flow to be much smaller than the radius of the earth, which allows us to ignore meridional derivatives of the geometric sphericity factors relative to meridional derivatives of the flow. The characteristic length scales of our waves (planetary) are larger than an earth's radius. In Spite of this, we assume QG and test its validity retroactively, because it is necessary for the formulation of the equations in terms of wave propagation-wave geometry. The results of our observational analysis may be viewed as an assessment of the applicability of QG linear theory to stratospheric planetary waves.

The second complication is that the geostrophic velocity, when defined in the traditional way, is divergent on a sphere. As a result, the EP flux divergence is not proportional to the PV flux as it is on a β-plane (Palmer, 1982). To get round this problem, past studies (e.g. Matsuno, 1970, Palmer, 1982) have essentially redefined the vertical component of the vorticity perturbation as follows: $\zeta' = f\nabla \times \left(\frac{\vec{v}}{f}f\right)$. For our calculations we take the approach of Plumb (1999, personal communication[1]), where instead of redefining the vorticity, we redefine the geostrophic winds by using the geopotential height scaled by the Coriolis parameter, as follows:

$$\Psi' = \frac{\phi'}{f} \tag{D.1}$$

[1]The derivation can be found in R. A. Plumb's Middle Atmosphere class notes. A less detailed derivation is also found in Wirth, 1990.

$$v' = \frac{1}{a\cos\varphi}\frac{\partial\Psi'}{\partial\lambda} \tag{D.2}$$

$$u' = \frac{1}{a}\frac{\partial\Psi'}{\partial\varphi} \tag{D.3}$$

$$T' = \frac{T_o}{g}\frac{\partial\phi'}{\partial z} = \frac{fT_o}{g}\frac{\partial\Psi'}{\partial z} \tag{D.4}$$

where $f = 2\Omega\sin(\varphi)$ is the latitude dependent Coriolis parameter, φ and λ are the latitude and longitude angles, z is log pressure, a the earth's radius, Ω the earth's rotation rate, and all other variables are as in the β-plane model, except that we use *dimensional* variables. Since we use spherical coordinates as a diagnostic of observations it is simpler to keep the variables dimensional.

Using these relations, the vertical component of vorticity is:

$$\zeta' = \nabla\times\vec{v} = \frac{1}{a\cos\varphi}\frac{\partial v'}{\partial\lambda} - \frac{1}{a\cos\varphi}\frac{\partial(u'\cos\varphi)}{\partial\varphi} \tag{D.5}$$

The corresponding potential vorticity equations are derived following Plumb (1999, personal communication, see footnote 1), from the momentum, temperature, and continuity equations. The derivation is essentially the β-plane one, where we neglect meridional derivatives of sphericity factors. The derivation is also similar to Matsuno (1970), except for the differences in the definitions above.

The potential vorticity equations are as follows:

$$\left(\frac{\partial}{\partial t} + \frac{U}{a\cos\varphi}\frac{\partial}{\partial\lambda}\right)q' + \frac{v'}{a}\frac{\partial\overline{q}}{\partial\varphi} = \frac{f}{p}\frac{\partial}{\partial z}\left(\frac{p\mathcal{H}}{N^2}\right) + (\nabla\times\mathcal{F}')\cdot\hat{\mathbf{k}} \tag{D.6}$$

$$q' = \zeta' + \frac{fg}{pT_o}\frac{\partial}{\partial z}\left(\frac{pT'}{N^2}\right) = \frac{1}{a^2\cos^2\varphi}\frac{\partial^2\Psi'}{\partial\lambda^2} - \frac{1}{a\cos\varphi}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial\Psi}{\varphi}\right) + \frac{f^2}{p}\frac{\partial}{\partial z}\left(\frac{p}{N^2}\frac{\partial\Psi'}{\partial z}\right) \tag{D.7}$$

$$\frac{\partial\overline{q}}{\partial y} = \frac{1}{a}\frac{\partial\overline{q}}{\partial\varphi} = \beta - \frac{1}{a^2}\frac{\partial}{\partial\varphi}\left(\frac{1}{\cos\varphi}\frac{\partial(U\cos\varphi)}{\partial\varphi}\right) - \frac{f^2}{pa^2}\frac{\partial}{\partial z}\left(\frac{p}{N^2}\frac{\partial U}{\partial z}\right) \tag{D.8}$$

where $p = p_o e^{-z/H_o}$ is pressure, and $\beta = \frac{2\Omega\cos\varphi}{a}$. All other variables are as defined in the β-plane model.

D.2 The transformed Eulerian mean zonal momentum equation.

The residual mean meridional circulation in spherical coordinates is defined as follows[2]:

$$\overline{v}^* = \overline{v} - \frac{1}{p}\frac{\partial}{\partial z}\left(\frac{pg}{T_o N^2}\overline{v'T'}\right) \tag{D.9}$$

$$\overline{w}^* = \overline{w} - \frac{1}{a\cos\varphi}\frac{\partial}{\partial \varphi}\left(\frac{g\cos\varphi}{T_o N^2}\overline{v'T'}\right) \tag{D.10}$$

Plugging into the zonal momentum equation, and taking a zonal mean results in the following second order equation for the zonal mean wind acceleration:

$$\frac{\partial U}{\partial t} + \overline{v}^*\left(\frac{1}{a\cos\varphi}\frac{\partial(U\cos\varphi)}{\partial\varphi} - f\right) + \overline{w}^*\frac{\partial U}{\partial z} - \mathcal{F}_\lambda = \frac{1}{\rho a\cos\varphi}\nabla\cdot\vec{F} \tag{D.11}$$

The term $\frac{1}{\rho a\cos\varphi}\nabla\cdot\vec{F}$, is plotted in figures 7.1 and 7.14 and is expected to be larger than the actual acceleration since part of it goes into the mean meridional circulation.

D.3 The linear, QG, spherical wave equations: Index of refraction and wavenumbers.

To get the linear wave equation we assume a normal mode solution in longitude and time, and use the transformation 4.9:

$$\Psi = \tilde{\Psi}e^{is(\lambda - \frac{c}{a\cos\varphi}t)} = \psi e^{\frac{z}{2H_o}}Ne^{is(\lambda - \frac{c}{a\cos\varphi}t)} \tag{D.12}$$

Using relations D.1-D.4, and equations D.7, and D.12 in equation D.6, and rearranging:

$$\frac{1}{\cos\varphi}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial\psi}{\partial\varphi}\right) + \frac{f^2a^2}{N^2}\frac{\partial^2\psi}{\partial z^2} + \left(\frac{a\overline{q}_\varphi}{U-c} - \frac{s^2}{\cos^2\varphi} + a^2f^2F(N^2)\right)\psi = damping \tag{D.13}$$

where $F(N^2)$ is defined in 4.10. As in the β-plane, we define an index of refraction and meridional and vertical wavenumbers (equations 4.11, 5.6, and 5.5):

[2]The derivation in this section was first suggested by Boyd (1976) and Andrews and McIntyre (1976).

$$n_{ref}^2 = N^2 \left(\frac{a\overline{q}_\varphi}{U-c} - \frac{s^2}{\cos^2\varphi} + a^2 f^2 F(N^2) \right) \tag{D.14}$$

$$Re\left(\frac{\frac{1}{\cos\varphi}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial\psi}{\partial\varphi}\right)}{\psi} \right) = Re\left(\frac{\psi_{\varphi\varphi} - \psi_\varphi \tan\varphi}{\psi} \right) \equiv -l^2 \tag{D.15}$$

$$Re\left(\frac{\psi_{zz}}{\psi} \right) \equiv -m^2 \tag{D.16}$$

Note that Plumb's derivation of the equations results in a cleaner definition of the meridional wavenumber because the meridional derivative term is the meridional component of ∇^2 in spherical coordinates. Matsuno's (1970) derivation, on the other hand, results in a different, more complicated meridional derivative term.

D.4 Wave activity conservation and the wave based coordinate.

The wave activity equation in spherical coordinates is derived the same way as in the β-plane (see section 6.2), and is exactly like equation 6.3, only the definitions of A, $\vec{F}$, $\nabla \cdot \vec{F}$, and D are different:

$$A = \frac{a^2 \rho \cos\varphi \overline{q'^2}}{2\overline{q}_\varphi} \tag{D.17}$$

$$F_\varphi = -a\rho\cos\varphi \overline{u'v'} \tag{D.18}$$

$$F_z = -a\rho\cos\varphi \frac{gf}{T_o N^2} \overline{v'T'} \tag{D.19}$$

$$\nabla \cdot \vec{F} = \frac{\partial F_z}{\partial z} + \frac{1}{a\cos\varphi}\frac{\partial (F_\varphi \cos\varphi)}{\partial\varphi} \tag{D.20}$$

To calculate the wave activity coordinate of chapter 6, we define the wave activity velocity as in the β-plane (equation 6.10), and integrate D.21 along with equations C.7-C.9:

$$\frac{dy_s}{dt} = \frac{V_{ay}}{a} \tag{D.21}$$

subject to the initial conditions C.10, where $y_o = a\varphi_o$.

The transformation to wave activity coordinates and the calculations of wave activity budget that follow are similar to the β-plane model (see chapter 6 and ap-

pendix C), with the exception that we add a term to the Jacobian, to account for the effects of sphericity:

$$J = \frac{\partial x_2}{\partial \xi_2}\frac{\partial x_3}{\partial \xi_3} - \frac{\partial x_2}{\partial \xi_3}\frac{\partial x_3}{\partial \xi_2} = \left(\frac{\partial y_s}{\partial r}\frac{\partial z_s}{\partial s} - \frac{\partial y_s}{\partial s}\frac{\partial z_s}{\partial r}\right) a^2 \cos\varphi \tag{D.22}$$

References

Andrews, D. G., J. R. Holton and C. B. Leovy, 1987. *Middle atmosphere dynamics.* Academic Press, 489pp.

Andrews, D. G., and M. E. McIntyre, 1976. Planetary waves in horizontal and vertical shear: the generalized Eliassen-Palm relation and the mean zonal acceleration. *J. Atmos. Sci.* **33**: 2031-2048.

Aris, R., 1962. *Vectors, tensors, and the basic equations of fluid mechanics.* Dover, New York. 286pp.

Backus, G. E., and J. F. Gilbert, 1970. Uniqueness in the inversion of inaccurate gross earth data. *Philos. Trans. R. Soc. London, Ser. A,* **266**: 123-192.

Barnett, J. J., and M. Corney, 1984. Temperature comparisons between the Nimbus 7 SAMS, rocket/radiosondes and the NOAA 6 SSU. *J. Geophys. Res.* **89**: 5294-5302.

Boville, B. W., 1960. The Aleutian stratospheric anticyclone. *J. Meteor.* **17**, 329-336.

Boyd, J. P., 1976. The noninteraction of waves with the zonally averaged flow on a spherical earth and the interrelationships of eddy fluxes of energy, heat and momentum. *J. Atmos. Sci.* 33: 2285-2291.

Bretherton, F. P., 1966. Critical layer instability in baroclinic flows. *Quart. J. Roy. Meteor. Soc.* **92**: 325-334.

Burger, A. P., 1966. Instability associated with the continuous spectrum in a baroclinic flow. *J. Atmos. Sci.* **23**: 272-277.

Card, P. A., and A. Barcilon, 1982. The Charney problem with a lower Ekman layer. *J. Atmos. Sci.* **39**: 2128-2137.

Chahine, M. T., 1968. Determining of the temperature profile in an atmosphere from its outgoing radiance. *J. Opt. Soc. Amer.* **58**: 1634-1637.

Chahine, M. T., 1970. Inverse problems in radiative transfer: A determination of atmospheric parameters. *J. Atmos. Sci.* **27**: 960-967.

Charney, J. G., 1947: The dynamics of long waves in a baroclinic westerly current, *J. Meteor.* , **4**, 135-162.

Charney, J. G., and P. G. Drazin, 1961. Propagation of planetary scale disturbances from the lower into the upper atmosphere. *J. Geophys. Res.* **66**: 83-110.

Charney, J. G., and J. Pedlosky, 1963. On the trapping of unstable planetary waves in the atmosphere. *J. Geophys. Res.* **68**: 6441-6442.

Charney, J. G., and M. E. Stern, 1962: On the instability of internal baroclinic jets in a rotating atmosphere. *J. Atmos. Sci.* , **19**, 159-172.

Claud, C., J. Ovarlez, and N. A. Scott, 1998. Evaluation of TOVS-derived stratospheric temperatures up to 10 hPa for a case of vortex displacement over western Europe. *J. Geophys. Res.* **103**: 13743-13761.

Conrath, B. J., 1972. Vertical resolution of temperature profiles obtained from remote radiation measurements. *J. Atmos. Sci.* **29**: 1262-1271.

Cressman, G. P., 1959. An operational objective analysis system. *Mon. Wea. Rev.* **87**: 367-374.

da Silva, A. M. and R. S. Lindzen, 1987. A mechanism for excitation of ultra-long Rossby waves. *J. Atmos. Sci.* **44**: 3625-3639.

da Silva, A. M. and R. S. Lindzen, 1993. On the establishment of stationary waves in the northern hemisphere winter. *J. Atmos. Sci.* **50**: 43-61.

Deland, R. J., 1973. Analysis of Nimbus 3 SIRS radiance data: Traveling planetary-scale waves in the stratosphere temperature field. *Mon. Wea. Rev.* **101**: 132-146.

Dickinson, R. E., 1968a. On the exact and approximate linear theory of vertically propagating planetary Rossby waves forced at a spherical lower boundary. *Mon. Wea. Rev.* **96**: 405-415.

Dickinson R. E., 1968b. Planetary waves propagating vertically through weak westerly wind wave guides. *J. Atmos. Sci.* **25**: 984-1002.

Dickinson, R. E., 1969a. Theory of planetary wave-zonal flow interaction. *J. Atmos. Sci.* **26**: 73-81.

Dickinson R. E., 1969b. Vertical propagation of planetary Rossby waves through and atmosphere with Newtonian cooling. *J. Geophys. Res.* **74**: 929-938.

Dunkerton T., C.-P. F. Hsu, and M. E. McIntyre, 1981. Some Eulerian and Lagrangian diagnostics for a model stratospheric warming. *J. Atmos. Sci.* **38**: 819-843.

Durran, D. R., 1991. The third-order Adams-Bashford method: An attractive alternative to Leapfrog time differencing. *Mon. Wea. Rev.* **119**: 702-720.

Edmon, H. J., Jr., Hoskins, B. J., and McIntyre, M. E., 1980. Eliassen-Palm cross-sections for the troposphere. *J. Atmos. Sci.* **37**: 2600-2616; corrigendum:**38**: 1115 (1981).

Farrell, B. F., 1982. The initial growth of disturbances in a baroclinic flow. *J. Atmos. Sci.* **39**: 1663-1686.

Finger F. G., H. M. Woolf, and C. E. Anderson, 1965. A method for objective analysis of stratospheric constant-pressure charts. *Mon. Wea. Rev.* **93**: 619-638.

Fullmer, J. W. A., 1982. The baroclinic instability of highly structured one dimensional basic states. *J. Atmos. Sci.* **39**: 2371-2378.

Garcia, R. R., and J. E. Geisler, 1981. Stochastic forcing of small amplitude oscillations in the stratosphere. *J. Atmos. Sci.* **38**: 2187-2197.

Geisler, J. E., and Dickinson, R. E., 1975. External Rossby modes on a β-plane with realistic vertical wind shear. *J. Atmos. Sci.* **32**: 2082-2093.

Geisler, J. E., and R. R. Garcia, 1977. Baroclinic instability at long wavelengths on a β-plane. *J. Atmos. Sci.* **34**: 311-321.

Gellman, M. E., and R. M. Nagatani, 1977. Objective analyses of height and temperatures at the 5, 2, and 0.4 mb levels using meteorological rocketsonde and satellite radiation data. *Space Res.* XVII: 117-122.

Gellman, M. E., A. J. Miller, K. W. Johnson, and R. M. Nagatani, 1986. detection of long term trends in global stratospheric temperatures from NMC analyses derived from NOAA satellite data. *Adv. Space Res.* **6**: 17-26.

Geller, M. A., M. F. Wu, and M. E. Gellman, 1983. Troposphere-stratosphere (surface-55 km) monthly winter general circulation statistics for the northern hemisphere-four year averages. *J. Atmos. Sci.* **40**: 1334-1352.

Geller, M. A., M. F. Wu, and M. E. Gellman, 1984. Troposphere-stratosphere (surface-55 km) monthly winter general circulation statistics for the northern hemisphere-interannual variations. *J. Atmos. Sci.* **41**: 1726-1744.

Gent, P. R., and J. C. McWilliams, 1983. Regimes of validity for balanced models. *Dyn. Atmos. Oceans* **7**:167-183.

Graves, D. S., 1986. Evaluation of satellite sampling of the middle atmosphere using the GFDL SKYHI general circulation model. *Ph.D dissertation.* Princeton university. 314 pp.

Green, J. S. A., 1960: A problem in baroclinic stability. *Quart. J. Roy. Meteor. Soc.* , **86**, 237-251.

Grose, W. L., and A. O'Neill, 1989. Comparison of data and derived quantities for the middle atmosphere of the southern hemisphere. *PAGEOEPH* **130**: 195-212.

Harnik, N. and R. S. Lindzen, 1998. The Effect of Basic-State Potential Vorticity Gradients on the Growth of Baroclinic Waves and the Height of the Tropopause, *J. Atmos. Sci.* **55**: 344-360.

Hartmann, D. L., 1976. The structure of the stratosphere in the southern hemisphere during late winter 1973 as observed by satellite. *J. Atmos. Sci.* **33**: 1141-1154.

Hartmann, D. L., 1979. Baroclinic instability of realistic zonal-mean states to planetary waves. *J. Atmos. Sci.* **36**: 2336-2349.

Hartmann, D. L., 1983. barotropic instability of the polar night jet stream. *J. Atmos. Sci.* **40**: 817-835.

Hartmann, D. L., 1985. Some aspects of stratospheric dynamics. *Adv. Geophys.* **28A**: 219-247.

Hartmann, D. L., C. R. Mechoso, and K. Yamazaki. 1984. Observations of wave-mean flow interactions in the southern hemisphere. *J. Atmos. Sci.* **41**: 351-362.

Harwood, R. S., 1975. The temperature structure of the southern hemisphere stratosphere August-October 1971. *Quart. J. Roy. Meteor. Soc.* **101**: 75-91.

Hirota, I. T., 1971. Excitation of planetary Rossby waves in the winter stratosphere by periodic forcing. *J. Meteorol. Soc. Jpn.* **49**: 439-449.

Hirota, I., T. Hirooka, and M. Shiotani, 1983. Upper stratospheric circulation in the two hemispheres, observed by satellites. *Quart. J. Roy. Meteor. Soc.* **109**: 443-454.

Hirota, I., K. Kuroi, and M. Shiotani, 1990. Mid-winter warmings in the southern hemisphere stratosphere in 1988. *Quart. J. Roy. Meteor. Soc.* **116**: 929-941.

Hirota I., and Y. Sato, 1969. Periodic variation of the winter circulation and intermittent vertical propagation of planetary waves. *J. Meteorol. Soc. Jpn.* **47**: 390-402.

Ioannou, P., and R. S. Lindzen, 1986. Baroclinic instability in the presence of barotropic jets. *J. Atmos. Sci.* **43**: 2999-3014.

Jacqmin, D., and R. S. Lindzen, 1985. The causation and sensitivity of the northern winter planetary waves. *J. Atmos. Sci.* **42**: 724-745.

Kaplan, L. D., 1959. Inference of atmospheric structure from remote radiation measurement. *J. Opt. Soc. Amer.* **49**: 1004-1007.

Karoly, J. D., 1989. The impact of base-level analyses on stratospheric circulation statistics for the southern hemisphere. *PAGEOEPH* **130**: 181-194.

Karoly, J. D., and Graves, D. S., 1990. On data sources and quality for the southern hemisphere stratosphere. *Dynamics, transport and Photochemistry in the middle atmosphere of the southern hemisphere* (A. O'Neill ed.): 19-32. Kluwer Academic Publishers.

Karoly, D. J., and B. J. Hoskins, 1982. Three dimensional propagation of planetary waves. *J. Meteorol. Soc. Jpn.* **60**: 109-123.

Kidder, S. Q. and T. H. Vonder Haar, 1995. *Satellite meteorology*, Academic press. 466 pp.

Kidwell, K. B., 1986. *The NOAA Polar Orbiter Data user's guide.* NOAA/NESDIS satellite data services division, Washington DC;
An updated version: http://www2.ncdc.noaa.gov:80/docs/podug/.

Kuo, H. L., 1979. Baroclinic instabilities of linear and jet profiles in the atmosphere. *J. Atmos. Sci.* **36**: 2360-2378.

Lait, L. R., and J. L. Stanford, 1988a. Applications of asynoptic space-time Fourier transform methods to scanning satellite measurements. *J. Atmos. Sci.* **45**: 3784-3799.

Lait, L. R., and J. L. Stanford, 1988a. Fast, Long-lived features in the polar stratosphere. *J. Atmos. Sci.* **45**: 3800-3809.

Landau, L. D., and E. M. Lifshitz, 1959. *Fluid Mechanics.* London: Pergamon press. 536pp

Leovy, C. B., and P. J. Webster, 1976. Stratospheric long waves: Comparison of thermal structure in the northern and southern hemispheres. *J. Atmos. Sci.* **33**: 1624-1638.

Lighthill, M. J. and Whitham, G. B, 1955. On kinematic waves. I. Flood movement in long rivers. *Proc. Roy. Soc. A* **229**: 281-316.

Lin, B.-D., 1982. The behavior of winter stationary planetary waves forced by topography and diabatic heating. *J. Atmos. Sci.* **39**: 1206-1226.

Lin, S.-J., and R. T. Pierrhumbert, 1988. Does Ekman friction suppress baroclinic instability? *J. Atmos. Sci.* **45**: 2920-2933.

Lindzen, R. S., 1990. *Dynamics in atmospheric physics.* Cambridge University Press, 310pp.

Lindzen, R. S., 1994a. The effect of concentrated PV gradients on stationary waves. *J. Atmos. Sci.* **51**:3455-3466.

Lindzen, R. S., 1994b. The Eady problem for a basic state with zero PV gradients but $\beta \neq 0$. *J. Atmos. Sci.* **51**: 3221-3226.

Lindzen, R. S., and Barker, J. W., 1985. Instability and wave over-reflection in stably stratified shear flow. *J. Fluid Mech.* **151**: 189-217.

Lindzen, R. S., B. Farrell, and K. K. Tung, 1980. The concept of wave overreflection and its application to baroclinic instability. *J. Atmos. Sci.* **37**: 44-63.

Lindzen, R. S., B. Farrell, and D. Jacqmin, 1982. Vacillations due to wave interference: Applications to the atmosphere and to analog experiments. *J. Atmos. Sci.* **39**: 14-23.

Lindzen, R. S., and H. L. Kuo, 1969. A reliable method for the numerical integration of a large class of ordinary and partial differential equations. *Mon. Wea. Rev.* **96**: 732-734.

Lindzen, R. S., and Rosenthal, A. J., 1976. On the instability of Helmholtz velocity profiles in stably stratified fluids when a lower boundary is present. *J. Geophys. Res.* **81**: 1561-1571.

Lindzen, R. S., and K. K. Tung, 1978. Wave overreflection and shear instability. *J. Atmos. Sci.* **35**: 1626-1632.

Madden, R. A., 1975. Oscillations in the winter stratosphere: Part 2. Theory of horizontal eddy heat transports and interaction of transient and stationary planetary scale waves. *Mon. Wea. Rev.* **103**: 717-729.

Madden, R. A., 1983. The effect of interference of traveling and stationary waves on time variations of the large-scale circulation. *J. Atmos. Sci.* **40**: 1110-1125.

Manney, G. L., J. D. Farrara, and C. R. Mechoso, 1991a. The behavior of wave 2 in the southern hemisphere stratosphere during late winter and early spring. *J. Atmos. Sci.* **48**: 976-998.

Manney, G. L., C. R. Mechoso, L. S. Elson, and J. D. Farrara, 1991b. Planetary-scale waves in the southern hemisphere winter and early spring stratosphere: stability analysis. *J. Atmos. Sci.* **48**: 2509-2523.

Mateer, C. L., 1965. On the information content of Umkehr observations. *J. Atmos. Sci.* **22**: 370-381.

Matsuno, T., 1970. Vertical propagation of stationary planetary waves in the winter northern hemisphere. *J. Atmos. Sci.* **27**: 871-883.

McIntyre, M. E., and T. N. Palmer, 1983. Breaking planetary waves in the stratosphere. *Nature* **305**, 593-600.

McMillin, L. M., L.J. Crone, M. D. Goldberg, and T. J. Kleespies, 1995. Atmospheric transmittance of an absorbing gas. 4. OPTRAN: a computationally fast and accurate transmittance model for absorbing gases with fixed and with variable mixing ratios at variable viewing angles. *Appl. Opt.* **34**: 6269-6274.

McPherson, R. D., K. H. Bergman, R. E. Kistler, G. E. Rasch, and D. S. Gordon, 1979. The NMC operational global data assimilation system. *Mon. Wea. Rev.* **107**: 1445-1461.

Mechoso, C. R., and D. L. Hartmann, 1982. An observational study of traveling planetary waves in the southern hemisphere. *J. Atmos. Sci.* **39**: 1921-1935.

Mechoso, C. R., and D. L. Hartmann, and J. D. Farrara, 1985. Climatology and interannual variability of wave mean-flow interaction in the southern hemisphere. *J. Atmos. Sci.* **42**: 2189-2206.

Mechoso, C. R., A. O'Neill, V. D. Pope, and J. D. Farrara, 1988. A study of the stratospheric final warming of 1982 in the southern hemisphere. *Quart. J. Roy. Meteor. Soc.* **114**: 1365-1384.

Miles, T. and A. O'Neill, 1989. *Comparison of satellite derived dynamical quantities in the stratosphere of the southern hemisphere.* Proceedings of MASH workshop, Williamsburg, VA.

Morgan, M. C., 1995. An observationally and dynamically determined basic state for the study of synoptic scale waves. Ph.D. Thesis.

Muench, H. S. 1965. On the dynamics of the wintertime stratospheric circulation. *J. Atmos. Sci.* **22**: 349-360.

Nigam, S., and R. S Lindzen, 1989. The sensitivity of stationary waves to variations in the basic state zonal flow. *J. Atmos. Sci.* **46**: 1746-1768.

O'Neill, A. and V. D. Pope, 1988. Simulations of linear and nonlinear disturbances in the stratosphere. *Quart. J. Roy. Meteor. Soc.* **114**: 1063-1110.

O'Neill, A. and C. E. Youngblut, 1982. Stratospheric warmings diagnosed using the transformed Eulerian-mean equations and the effect of the mean state on wave propagation. *J. Atmos. Sci.* **39**: 1370-1386.

Orr, W. McF., 1907. Stability or instability of the steady motions of a perfect liquid. *Proc. R. Irish Acad.* **27**: 9-69.

Palmer, T. N., 1981. Diagnostic study of a wavenumber-2 stratospheric sudden warming in a transformed Eulerian-mean view point. *J. Atmos. Sci.* **38**: 844-855.

Palmer, T. N., 1982. Properties of the Eliassen-Palm flux for planetary scale motions. *J. Atmos. Sci.* **39**: 992-997.

Pedlosky, J., 1987. *Geophysical Fluid Dynamics,* Springer Verlag, 709pp.

Phillpot, H. R., 1969. Antarctic stratospheric warming reviewed in the light of 1967 observations. *Quart. J. Roy. Meteor. Soc.* **95**: 329-348.

Pierce, J. C., 1991. Timing of NASA afternoon passes. *Int. J. Remote Sensing* **12**: 193-198.

Plumb, R. A., 1981. Instability of the disturbed polar night vortex: a theory of stratospheric warming. *J. Atmos. Sci.* **38**, 2514-2531.

Plumb, R. A., 1983. Baroclinic instability of the summer mesosphere: A mechanism for the quasi-two day wave? *J. Atmos. Sci.* **40**, 262-270.

Plumb, R. A., 1989. On the seasonal cycle of stratospheric planetary waves. *Pure Appl. Geophys.* **130**: 233-242.

Polvani, L. M., and R. A. Plumb, 1992. Rossby wave breaking, microbreaking, filamentation and secondary vortex formation: the dynamics of a perturbed vortex. *J. Atmos. Sci.* **49**: 462-476.

Polvani, L. M., D. W. Waugh, and R. A. Plumb , 1995. On the subtropical edge of the stratospheric surf zone. *J. Atmos. Sci.* **52**: 1288-1309.

Randel, W. J., 1987a. The evaluation of winds from geopotential height data in the stratosphere. *J. Atmos. Sci.* **44**: 3097-3120.

Randel, W. J., 1987b. A study of planetary waves in the southern winter troposphere and stratosphere. Part I: Wave structure and vertical propagation. *J. Atmos. Sci.* **44**: 917-935.

Randel, W. J., 1988. The seasonal evolution of planetary waves in the southern hemisphere stratosphere and troposphere. *Quart. J. Roy. Meteor. Soc.* **114**: 1385-1409.

Randel, W. J., 1990. A comparison of the dynamic life cycles of the tropospheric medium-scale waves and stratospheric planetary waves. *Dynamics, transport and Photochemistry in the middle atmosphere of the southern hemisphere* (A. O'Neill ed.): 19-32. Kluwer Academic Publishers.

Randel, W. J., 1992. Global atmospheric circulation statistics, 1000-1 mb. *NCAR Technical Note*, NCAR/TN-336+STR, 156 pp.

Randel, W. J., and J. L. Stanford, 1985. The observed life cycle of a baroclinic instability. *J. Atmos. Sci.* **42**: 1364-1373.

Randel, W. J., D. E. Stevens, and J. L. Stanford, 1987. A study of planetary waves in the southern winter troposphere and stratosphere. Part II: Life cycles. *J. Atmos. Sci.* **44**: 936-949.

Robinson, W. A., 1986. The application of the quasi-geostrophic Eliassen-Palm flux to the analysis of stratospheric data. *J. Atmos. Sci.* **44**, 1017-1023.

Rodgers, C. D., 1976. Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys. and Space Phys.* **14**: 609-624.

Rodgers, C. D., 1984. Workshops on comparison of data and derived dynamical quantities during Northern hemisphere winters, *Adv. Space Res.* **4**: 117-125.

Rodgers, C. D., 1990. Characterization and error analysis of profiles retrieved from remote sounding measurements. *J. Geophys. Res.* **95**: 5587-5595.

Salby, M. L., 1982a. Sampling theory for asynoptic satellite observations. Part I: Space-time spectra, resolution and aliasing. *J. Atmos. Sci.* **39**: 2577-2600.

Salby, M. L., 1982b. Sampling theory for asynoptic satellite observations. Part II: Fast Fourier synoptic mapping. *J. Atmos. Sci.* **39**: 2601-2614.

Salby, M. L., 1984. Survey of planetary-scale traveling waves: The state of theory and observations. *Rev. Geophys. and Space Phys.* **22**: 209-236.

Salby, M. L., and R. R. Garcia, 1987. Vacillations induced by interference of stationary and traveling planetary waves. *J. Atmos. Sci.* **44**: 2679-2711.

Schmidlin, F. J., 1984. Inter-comparisons of temperature, density, and wind measurements from in situ and satellite techniques. *Adv. Space Res.* **4**: 101-110.

Schoeberl, M. R., and M. A. Geller, 1977. A calculation of the structure of stationary planetary waves in winter. *J. Atmos. Sci.* **34**: 1235-1255.

Scinocca, J. F., and P. H. Haynes, 1998. Dynamical forcing of stratospheric planetary waves by tropospheric baroclinic eddies. *J. Atmos. Sci.* **55**: 2361-2392.

Shiotani, M., and I. Hirota, 1985. Planetary wave-mean flow interaction in the stratosphere: a comparison between northern and southern hemispheres. *Quart. J. Roy. Meteor. Soc.* **111**: 309-334.

Shiotani, M., K. Kuroi, and I. Hirota, 1990. Eastward traveling waves in the southern hemisphere stratosphere during the spring of 1983. *Quart. J. Roy. Meteor. Soc.* **116**: 913-927.

Simmons, A. J., 1974. Planetary-scale disturbances in the polar winter stratosphere. *Quart. J. Roy. Meteor. Soc.* **100**: 76-108.

Smith, W. L., 1970. Iterative solution of the radiative transfer equation for the temperature and absorbing gas profile of an atmosphere. *Appl. Opt.* **9**: 1993-1999.

Smith, W. L., and H. M. Woolf, 1976. The use of statistical covariance matrices for interpreting satellite sounding radiometer observations. *J. Atmos. Sci.* **33**: 1127-1140.

Smith, W. L., H. M. Woolf, and C. M. Hayden, 1979. The TIROS-N operational vertical sounder. *Bull. Am. Meteorol. Soc.* **58**: 1177-1187.

Snyder, C. M., and R. S. Lindzen, 1988. Upper level baroclinic instability. *J. Atmos. Sci.* **45**: 2445-2459.

Straus, D. M, 1981. Long wave baroclinic instability in the troposphere and stratosphere with spherical geometry. *J. Atmos. Sci.* **38**: 409-426.

Sun, D. Z., and R. S. Lindzen, 1994. A PV view of the zonal mean distribution of temperature and wind in the extratropical troposphere. *J. Atmos. Sci.* **51**: 757-772.

Trenberth, K. E., and J. G. Olson, 1988. An evaluation and inter-comparison of global analyses from the National Meteorological Center and the European Center for Medium Range Weather Forecasts. *Bull. Am. Meteorol. Soc.* **69**: 1047-1057.

Taylor, F. W., J. T. Houghton, G. D. Peskett, C. D. Rodgers, and E. J. Williamson, 1972. Radiometer for remote sounding of the upper atmosphere. *Appl. Opt.* **11**: 135-141.

Warn, T. and H. Warn, 1978. The evolution of a nonlinear critical level. *J. Atmos. Sci.* **33**: 2021-2024.

Whitham, G. B., 1960. A note on group velocity. *J. Fluid Mech.* **9**: 347-352.

Wirth, V., 1990. The seasonal cycle of stationary planetary waves in the southern stratosphere: A numerical study. M.S. thesis, Massachusetts Institute of Technology.

Wirth, V., 1991. What causes the seasonal cycle of stationary waves in the southern stratosphere? *J. Atmos. Sci.* **42**: 11-28.

Young, R. E., and H. Houben, 1989. Dynamics of planetary scale waves during southern hemisphere winter. *J. Atmos. Sci.* **46**: 1365-1383.

Zhang, K.-S., and T. Sasamori, 1985. A linear stability analysis of the stratospheric and mesospheric zonal mean state in winter and summer. *J. Atmos. Sci.* **42**: 2728-2750.

5733-47