4

Theoretical investigations

4.1 Introduction

During the 1970s, research into the theoretical understanding of high latitude mesocyclones was focused on the basic mechanisms of development of the more intense systems, known as polar lows. The aim was to explain the striking differences between polar lows and other extratropical cyclones, namely the small size and rapid growth rates of polar lows, and their favoured formation within cold air masses over the oceans in winter. It will become apparent by the end of this chapter that these fundamental questions have not been completely answered. However, considerable progress has been made, and new areas of research have been opened up regarding the life cycle of polar lows, and their interaction with the broad-scale atmospheric flow.

The construction of mathematical and theoretical models of mesocyclones is not simple, because there are many types of vortices occurring in the high latitude areas. They vary widely in horizontal and vertical extent, in intensity and in structure. A mesocyclone may be a powerful system, extending through the depth of the troposphere, with intense deep convection and hurricane force winds, or a weak swirl in the boundary-layer cloud, clearly visible on satellite imagery but with little significant weather at the Earth's surface. The environment in which the vortex forms may differ widely being, for example, a low-level frontal zone, or a flaccid low-pressure region at the centre of a decaying synoptic cyclone. The lows sometimes resemble more familiar weather systems, such as baroclinic waves or tropical cyclones, but between the more identifiable systems lies a variety of transitional types (see Table 3.1).

The resemblance of mesocyclones to other, more studied, weather systems provides a useful starting point for an analysis of their structure and mechanisms of formation/development, and indeed many of the advances in the

theoretical understanding of polar lows have occurred in the context of work done on other maritime cyclones. These include rapidly intensifying synoptic cyclones, where the importance of latent heat release has been demonstrated, frontal wave cyclones, synoptic systems that are initiated on low-level frontal zones, and tropical cyclones.

This chapter will draw on work from all of the above fields, in addition to studies directed explicitly towards polar lows. The first two sections that follow will discuss individually two of the principle fluid dynamical instabilities that are believed to contribute to mesocyclone formation, namely baroclinic and barotropic instability. We then examine the insight that can be gained into polar low formation and structure using the powerful tool of potential vorticity. We then consider the important role that thermal instability plays in some of the more vigorous systems. Finally, we attempt to draw together the preceding material into a current picture of our understanding of polar low development and structure and try to identify the most pressing questions for future research.

4.2 Baroclinic instability

Baroclinic instability is a type of dynamical instability associated with a baroclinic region of the atmosphere, i.e. an area where the density depends on both the temperature and pressure, or, expressed slightly differently, temperature varies along the pressure surfaces. Baroclinic instability is associated with the vertical shear of the mean flow, which is related to the horizontal temperature gradient by the thermal wind equation. Instabilities in a baroclinic region grow by converting potential energy associated with the mean horizontal temperature gradient into kinetic energy through ascending warm air and descending cold air. In the following, only a short overview will be given presenting features pertinent for the arguments elsewhere in the book. For a detailed discussion of baroclinic instability theory the reader is referred to textbooks, such as Holton (1992) or Bluestein (1993).

The mechanism through which a baroclinic wave amplifies within a region of strong north–south temperature gradient is illustrated by Figure 4.1. Assuming that a weak wave-like perturbation is initiated by some process in an otherwise uniform zonal flow, the meridional motions associated with the perturbation will distort the originally straight east–west oriented isotherms causing a wave in the temperature field to form. This wave will be displaced one quarter of a wavelength to the west of the wave in the pressure (geopotential) field. In the absence of other influences, horizontal temperature advection associated with the geostrophic wind field will further distort the isotherms

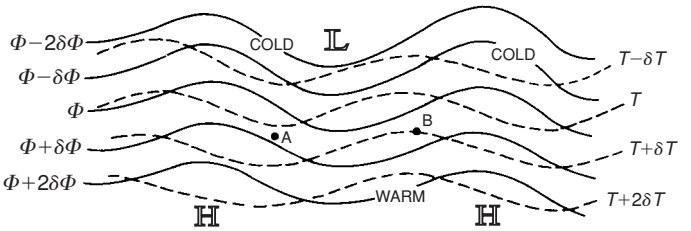


Figure 4.1. The distribution of geopotential height (solid lines) and temperature (broken lines) on a constant pressure surface in a developing baroclinic wave in the Northern Hemisphere. The pressure surface is located near the level where the speed of the wave is the same as the speed of the mean zonal flow (from Wallace and Hobbs, 1977).

from their original east–west orientation, causing the temperature wave to further amplify. In order for the wave to grow, the kinetic energy must increase. The mechanism by which this happens is a thermally direct circulation in which cold air at A (see Figure 4.1) sinks and warm air at B rises thus lowering the centre of gravity of the fluid, i.e. converting potential energy into kinetic energy.

By the 1950s, it was widely accepted that baroclinic instability acting within the belt of the mid-latitude westerlies was responsible for the development of the majority of extratropical, synoptic-scale cyclones. From theoretical as well as observational studies it was possible to infer the various stages of development of a baroclinic cyclone and the schematic flow in a typical system is shown in Figure. 4.2.

In this sequence of diagrams the developing low (Figure 4.2a) is seen as a perturbation on the baroclinic zone. During the rapid development phase there is a cooperative interaction between the flows at upper levels and near the surface, with strong, low-level cold air advection west of the low and weaker warm advection to the east. This particular pattern of thermal advection derives from the fact that the 500 hPa trough is to the west of the surface low with the mean geostrophic wind in the 1000–500 hPa layer being across the 1000–500 hPa thickness lines towards larger thickness values west of the surface low and towards smaller thickness east of the surface low. As the system continues to develop (Figure 4.2b) the distortion of the upper-level flow leads to a sharper 500 hPa trough and growing upper-level divergence with pressure falls at the surface leading to further development of the low (self development).

Figure 4.2 suggests that the low-level cyclone and the upper-level system (trough) develop simultaneously as a continuous process as part of the baroclinic development. In many cases, however, cyclone development starts in such a way, that a baroclinic wave that is well defined at upper or middle levels

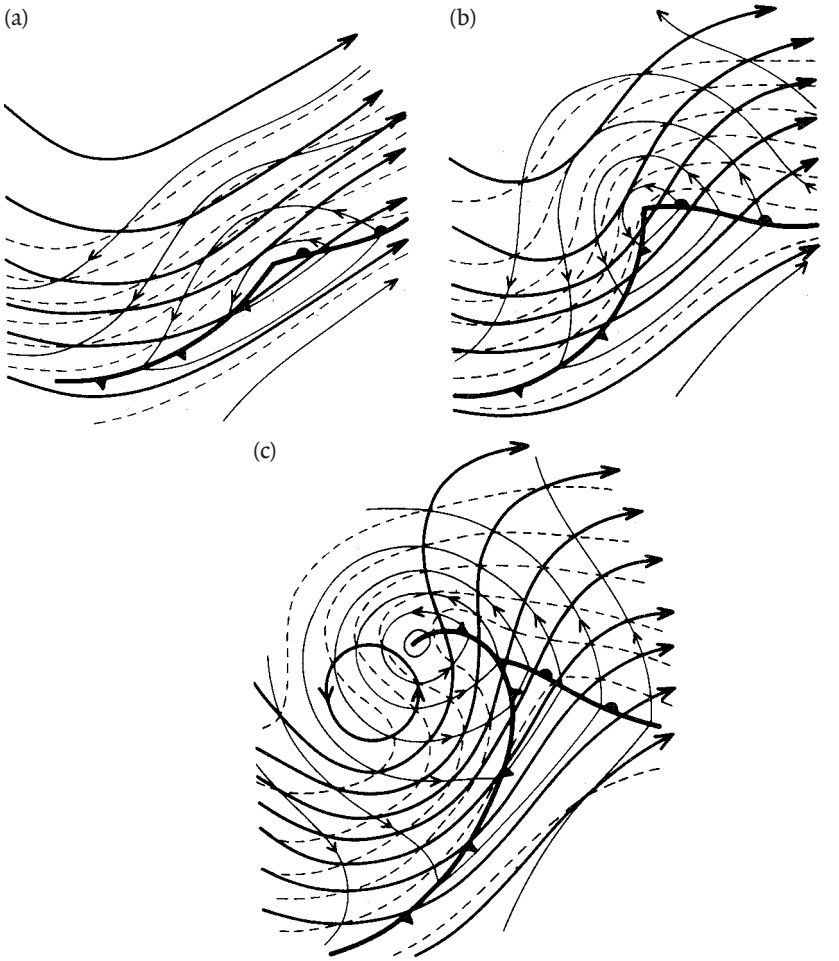


Figure 4.2. Schematic 500 hPa contours (heavy solid lines), 1000 hPa contours (thin lines), and 1000–500 hPa thickness (dashed lines), illustrating the 'self-development' process during the growth of an extratropical cyclone (from Palmén and Newton, 1969).

(an upper-level, cold short-wave trough, Figure 4.3a), but weak in the lower atmosphere, moves over a pre-existing low-level frontal zone. When the region of upper-level vorticity advection associated with the short-wave trough approaches the low-level front with its associated strong horizontal temperature gradients, low-level thermal advection will become increasingly important leading to hastening of low-level cyclogenesis in a manner similar to that illustrated on Figure 4.2 (Figure 4.3b and c).

The role of upper-level troughs for the development of baroclinic waves was elaborated by Pettersen and Smebye (1971). They distinguished between

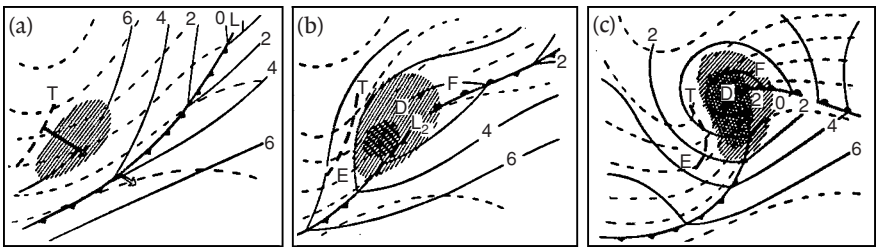


Figure 4.3. Stages in a Type B extratropical cyclone development. As the upper-level cold trough advances and the area of vorticity advection aloft (hatched area) spreads over the low-level frontal zone, the imbalance created results in convergence at low levels. When the thermal field has become distorted through cyclonic circulation the system can develop further through the self-development process illustrated in Figure 4.2 (from Palmén and Newton, 1969).

two types of development leading to the formation of extratropical cyclones. One (Type A) was characterized by an initial development under a more or less straight upper-level current and an initially large, low-level amount of baroclinic instability that decreases as the wave occludes. The other type of development (Type B) commences, according to Pettersen and Smebye, 'when a pre-existing upper trough, with strong vorticity advection on its forward side, spreads over a low-level area of warm advection in which fronts may or may not be present'. For the 'pure form' of Type B development the amount of baroclinicity initially is relatively small but increases as the storm intensifies. In the present text we use the term 'Type B development' simply to designate a baroclinic development initiated by the arrival of an upper-level disturbance, generally an upper-level trough (upper-level potential vorticity anomaly), over a low-level baroclinic zone or front.

A schematic picture of cyclogenesis associated with the arrival of an upper-level positive PV (potential vorticity) perturbation over a low-level baroclinic region illustrating the same process as Figure 4.3 but from a PV perspective is illustrated on Figure 4.4 (the PV approach is discussed in more detail in Section 4.4).

Although the role of baroclinic instability on the development of synoptic-scale depressions was well understood by the 1970s, it was unclear as to which process or processes were responsible for the development of polar lows and other mesoscale lows found in the polar regions and the degree to which baroclinic instability played a part. The role of baroclinic instability in the development of polar lows has been examined via different types of theoretical/numerical investigations. First, simple linear models, in both dry and moist forms, have been applied based on data from case studies. These have used normal mode techniques to examine the growth rates of

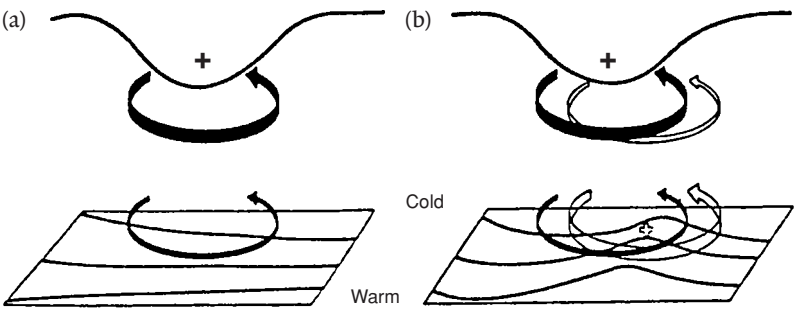


Figure 4.4. A schematic picture of cyclogenesis associated with the arrival of an upper-level positive PV anomaly over a lower-level baroclinic region. (a) The circulation induced by the upper-level vorticity anomaly is shown by solid arrows. The thin lines show potential temperature contours at the lower boundary. The advection of potential temperature by the induced low-level circulation leads to the formation of a warm anomaly slightly east of the upper-level PV anomaly. This in turn will induce a cyclonic circulation as shown by the open arrows in (b). The induced upper-level circulation will reinforce the original upper-level anomaly and can lead to amplification of the disturbance (after Hoskins *et al.*, 1985).

disturbances of different wavelengths. The 'problem' that polar lows have a much shorter wavelength than that which follows from the standard form of baroclinic instability theory has been explained by the fact that that the cold air masses involved are quite shallow, being confined to the lowest 1000–2000 m (e.g. Mansfield, 1974; Wiin-Nielsen, 1989). Another type of approach is the initial value method in which the general perturbations that trigger the storms have a more complex structure (type B developments). Also, full primitive equation mesoscale models have been used to simulate selected cases with experiments being undertaken with different parameterization schemes. Such work has shed light on the role of baroclinic instability and other instability mechanisms, such as Conditional Instability of the Second Kind (CISK).

Considering baroclinic instability in polar regions, it is possible to distinguish between low-level instability mainly confined to shallow boundary layers, typically, but not always, along the ice edges, and 'deep instability' involving deeper baroclinic layers. Furthermore, it is useful to distinguish between 'ordinary' baroclinic instability, where the direction of the surface wind, the thermal wind and the direction of wave propagation are all the same, and the so-called 'reverse shear instability' where the thermal wind is opposite to the direction of the surface wind and to the direction of the wave propagation. In this section we will consider these different types of investigation carried out over the last 30 years and relate the results to the observational studies presented in Chapter 3.

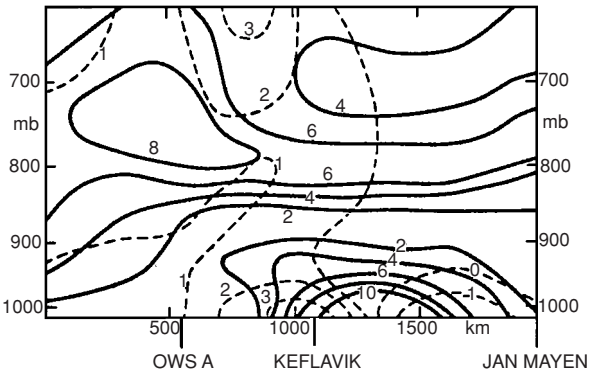


Figure 4.5. A cross-section perpendicular to the direction of travel of a polar low on 0000 GMT 7 December 1967. Isopleths of the vertical gradient of potential temperature ($\partial\theta/\partial z$) in °C km^{-1} (solid lines) and horizontal gradient of θ in °C (100 km)$^{-1}$ (broken lines): positive θ increases to the left (from Mansfield, 1974).

The first attempt to apply baroclinic theory to polar lows using a linear model was by Mansfield (1974). This study was motivated by the observational work of Harrold and Browning (1969), who documented the structure and evolution of a polar low that crossed the British Isles using Doppler radar and conventional radar and synoptic observations. From their analysis, they concluded that the system was basically a baroclinic cyclone. As would be expected for a baroclinic cyclone, the polar low was, during its passage over the British Isles, predominantly associated with slantwise precipitation along a narrow tongue, rather than convective precipitation (see Section 3.1.2).

Mansfield (1974) constructed a typical cross-section of the region of formation of the Harrold and Browning polar lows, based on station and ocean weather ship radiosonde ascents. This cross-section, showing the horizontal and vertical gradients of potential temperature, is reproduced in Figure 4.5.

Mansfield noted the existence of a shallow layer of reduced static stability below a strong inversion at 850 hPa, and suggested that an instability would likely be confined to this layer. Both the low vertical stability and the shallow depth could significantly modify a baroclinic instability. To quantify these effects, Mansfield applied the Eady model (Eady, 1949; see also Gill, 1982) for the growth of perturbations in a plane-parallel flow. The basic state was assumed to have a constant static stability, constant vertical wind shear, and to be bounded above by a rigid lid. The parameter values that were chosen on the basis of the cross-section in Figure 4.5. were Brunt–Väisälä frequency $N^2 = 1.6 \times 10^{-4}$ s^{-2}, the horizontal gradient of potential temperature, $\partial\theta/\partial y = 1.4$ K (100 km)$^{-1}$ and layer depth $H = 1.6$ km. The solution of the Eady model is an exponentially growing wave. Using the above

values and a Coriolis parameter of $f = 1.22 \times 10^{-4}$ s^{-1}, Mansfield calculated various parameters for the growing wave. He found that the fastest growing normal mode would have a wavelength of $l = 645 \pm 10$ km, a phase speed of $c = 6 \pm 1$ m s^{-1}, and a growth rate with e-folding time of $t = 28 \pm 1$ h. These values agree remarkably well with the estimates from observations of $l = 650 \pm 100$ km, $c = 6 \pm 4$ m s^{-1}, and $t = 24 \pm 6$ h. It is particularly notable that the calculation predicted the small wavelength and rapid growth rate of the system. The wavelength of the fastest growing Eady mode is proportional to the Rossby radius of deformation NH/f, which shows that the small scale predicted for the polar low was due mainly to its shallow depth. The growth rate, $\sigma = t^{-1}$, is proportional to the isentropic slope $N(g/\theta_0)(\partial\theta/\partial y)$ and thus has a relatively large value owing to the large horizontal temperature gradient. Low values of static stability could also have contributed to rapid growth at a short wavelength, but the value of N was unexceptional.

In retrospect, the agreement between model and observations may perhaps have been fortuitous. As Mansfield shows (Figure 4.6), factors not included in this analysis, such as friction and surface heat fluxes would tend to damp the growing wave. On the other hand, it will be seen later that latent heat release will tend to increase the growth rate. It should be noted that Rasmussen (1979) reproduced basically the same growth rates and horizontal scale of the disturbance studied by Harrold and Browning, and Mansfield using a CISK-type model.

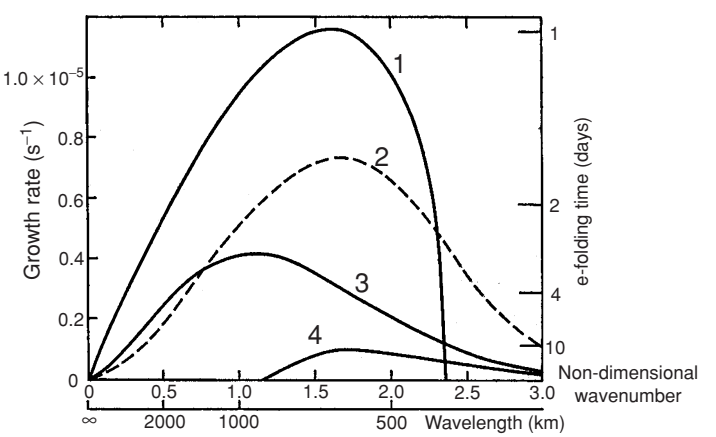


Figure 4.6. The effect of the inclusion of surface friction and heating in Eady's model. Growth rate versus wavelength for: 1, surface wind speed (U_0) zero (no heating or friction); 2, heating only $U_0 = 15$ m s^{-1}; 3, friction only $U_0 = 15$ m s^{-1}; 4, friction and heating $U_0 = 10$ m s^{-1} (from Mansfield, 1974).

In his analysis, Mansfield assumed that a stable layer acting as a lid confined the disturbance to the layer below 1.6 km. The data presented by Harrold and Browning, on the other hand, show that the low during its passage over the British Isles was a fairly deep system with cloud reaching above 5 km.

Baroclinic developments confined to very shallow layers may be found along the ice edges in the Southern as well as the Northern Hemisphere. During winter shallow baroclinic zones tend to form along the ice edges bordering the ice- or snow-covered polar regions. Over the years, developments along these low-level, shallow baroclinic zones have been considered essential for the formation of polar lows, cf. for example the classification scheme of Businger and Reed (1989b) in which one of the three main types of 'elementary polar low developments' is the so-called 'Arctic-front type' associated with ice boundaries.

A type A mesoscale cyclone formation over the northern part of the Fram Strait along a shallow frontal zone situated over the marginal ice zone along the northeast coast of Greenland, was studied by Rasmussen *et al.* (1997). The mesoscale cyclone in this case formed in a region covered by sea ice, but within an area of a very strong horizontal temperature gradient (Figure 4.7a). During the formation of the mesoscale cyclone a sharp low-level cold front formed. As the front passed the two observation camps in the region (Camp A and Camp O indicated on Figure 4.7a) the temperature dropped from around $-5\,^{\circ}$C to $-30\,^{\circ}$C and the wind increased from a few metres per second to nearly 10 m s^{-1}.

The air over the ice was extremely stable and convection could entirely be ruled out as contributing to the development.

The vertical structure of the atmosphere within the region where the disturbance developed is illustrated by the radiosonde ascent shown on Figure 4.7b. The ascent shows a very shallow cold air mass which was only around 200 m deep and neutrally stratified, capped by a somewhat deeper frontal zone extending up to around 500 m. While the development of the weak low was accompanied by the formation of a sharp low-level cold front, the advance of warm air preceding the passage of the cold front had the character of a 'warm surge' and only a weak warm front could be identified.

The shallow character of the air masses involved in such systems indicates that only modest development in terms of strength of the resulting circulation can be expected. After its formation south of Camp A early on 11 April the low moved north and disappeared over the ice during the evening (Figure 4.7c).

During the time of development, the upper-air flow was mainly anticyclonic and no discernible upper-level system triggered or influenced the development of the near-surface system, which in this case must be classified as a baroclinic

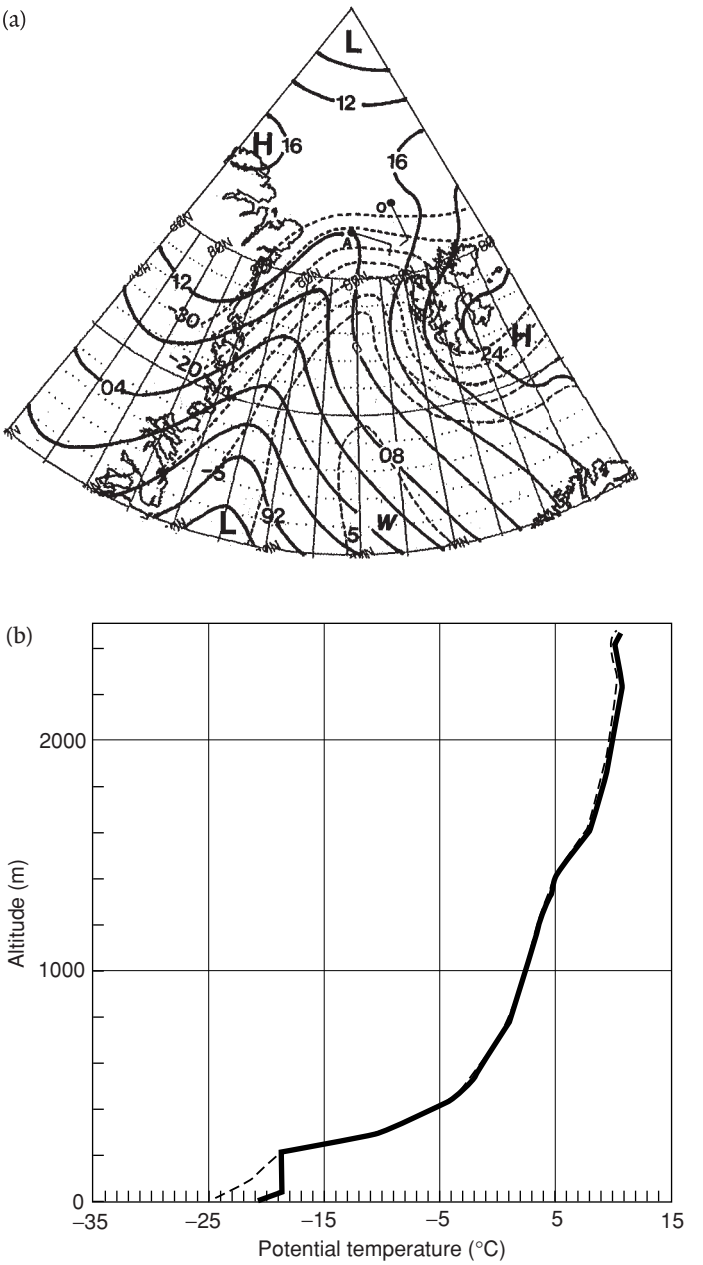


Figure 4.7. (a) Surface pressure and temperatures at 0000 GMT 11 April 1989. Pressure in 4 hPa intervals and isoterms in 5 °C intervals. Dots marked with the letters A and O show positions of observation camps. (b) Radiosonde ascent from Camp O showing potential temperature (solid) and dew point potential temperature (dashed) as a function of height at 2254 GMT 11 April. (c) Surface pressure and temperatures at 1800 GMT 11 April. Pressure and isotherms as in (a) (from Rasmussen *et al.*, 1997).

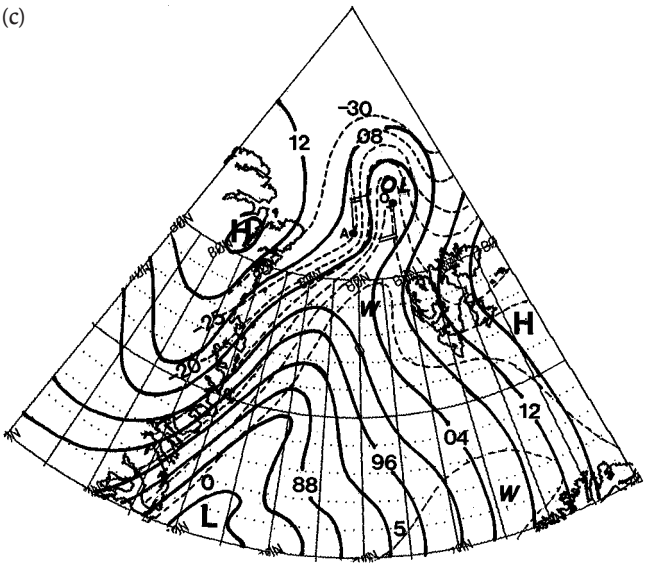


Figure 4.7 (*cont.*).

type A system being confined to the lowest few hundred metres above the surface.

No satellite images showing this disturbance were available. Figure 4.8, however, shows a wave train of two mesoscale cyclones within the same region, albeit over the sea. The satellite image shows that the meso-vortices forming the wave train in this case were characterized by low-level cloud. Another example of a baroclinic wave formed in the pack ice region along the northeast Greenland coast is shown on Figure 1.22.

Other minor mesoscale cyclones with spiral cloud structures are regularly observed on satellite imagery over the sea between northeast Greenland and Svalbard. Some of these vortices form as the result of barotropic instability along minor shear lines (see Section 4.3). They are rather insignificant weather features accompanied only by low wind speed.

Tsuboki and Wakahama (1992), in a study of mesoscale cyclones with a diameter of 200–700 km developing off the west coast of Hokkaido, Japan, found that baroclinicity in the lower troposphere was important for the formation of these cyclones. Based on an analysis of satellite images, the mesoscale cyclones could be classified into two types according to their horizontal scale: Type I (200–300 km in diameter) and Type II (500–700 km). From a linear instability analysis with a basic flow based on observed wind profiles, two unstable modes were found: Mode I, with a wavelength of 200–300 km and another, Mode II, of 500–700 km. Comparisons between theoretical and observational results

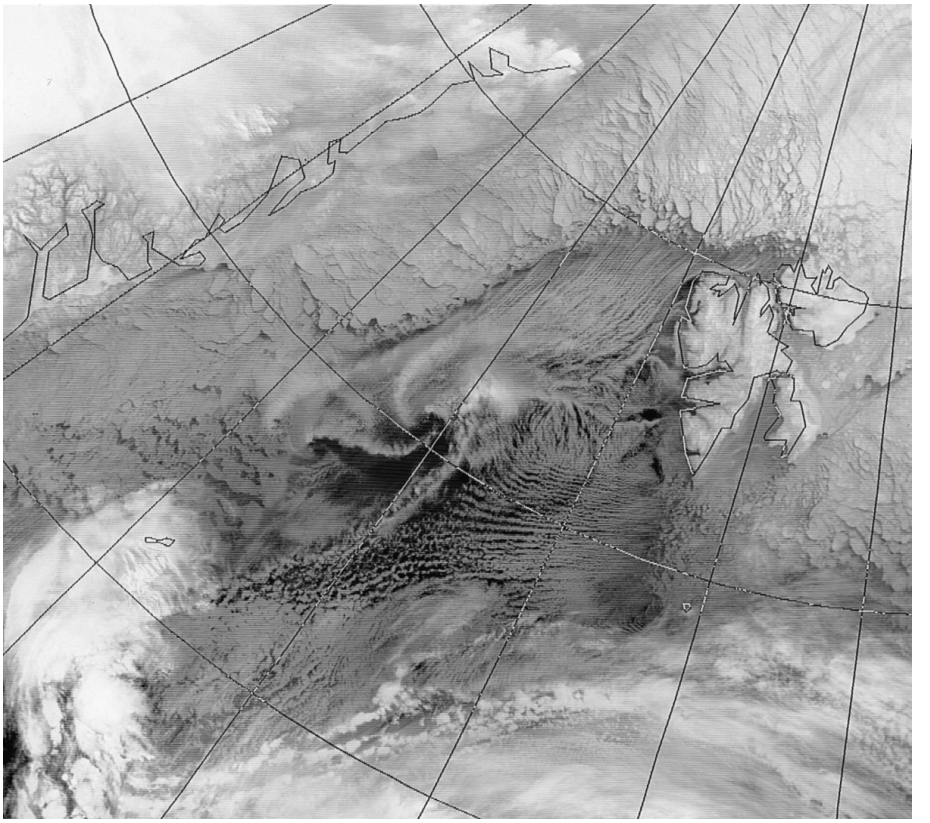


Figure 4.8. An infra-red satellite image for 1137 GMT 19 December 1989 showing two mesoscale cyclones that formed within the shallow baroclinic zone at the edge of the pack ice along the northeast coast of Greenland. (Image courtesy of the NERC Satellite Receiving Station, University of Dundee.)

indicated that the modes I and II could account for the characteristic properties of Types I and II systems respectively. The energetics showed that both modes I and II were maintained by the increase of eddy available potential energy and its conversion into eddy kinetic energy. The Type I systems were shallow disturbances confined below 850 hPa, while Type II systems were rather deep, extending to *c.* 500 hPa. The lifetimes of Type I and Type II systems were 0.4–1.6 and 1.0–3.0 days, respectively.

On Figure 4.9 is shown the growth rate diagram of the unstable waves, the abscissa representing the magnitude of the non-dimensional wavenumber vector and the ordinate its direction. The figure has two maxima: one corresponding to a growth rate of *c.* 2.4 day^{-1} located at wavenumber 8, and another corresponding to *c.* 2 day^{-1}, at wavenumber 2.7. The corresponding dimensional wavelength of the former wave is *c.* 240 km (Mode I), and that of the latter

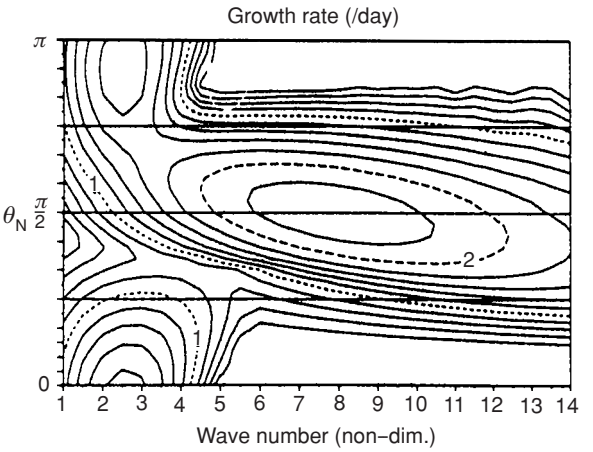


Figure 4.9. Growth rate diagram of unstable waves. The abscissa represents the magnitude of non-dimensional wavenumber vector and the ordinate represents its direction for representative parameters of the basic flow (from Tsuboki and Wakahama, 1992).

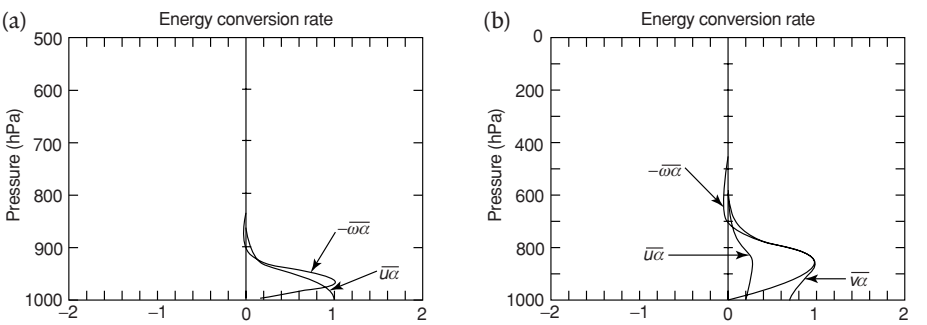


Figure 4.10. Vertical profiles of energy conversion rate of (a) Mode I, corresponding to shallow, short wavelength systems, and (b) Mode II, corresponding to a deeper system of longer wavelength (in arbitrary units). α, specific volume; ω, vertical velocity in isobaric coordinates; $\overline{u\alpha}$, zonal heat transport; $\overline{v\alpha}$, meridional heat transport (from Tsuboki and Wakahama, 1992).

c. 700 km (Mode II). Compared with the growth rates found by Mansfield (1974) and Duncan (1977) in their respective linear analyses of polar low growth, the growth rates are relatively large.

Figures 4.10a and b show the vertical profiles of energy conversion rate for Mode I and Mode II systems. The profiles show clearly that Mode I lows, corresponding to the short wavelength systems, are shallow, confined to a thin layer of a depth around 100 hPa, while the Mode II systems are about twice as deep.

Based on the data analysis of observations and the linear instability analysis, Tsuboki and Wakahama concluded that the mesoscale cyclones off the

west coast of Hokkaido and Sakhalin were baroclinic disturbances. The systems studied by them developed over the Japan Sea within a northwesterly, light winter monsoon wind, and, judging from the information in the paper by Tsuboki and Wakahama, along a boundary layer front (BLF) rather similar to the BLFs frequently observed along the west coast of Svalbard. As argued in Section 4.3, the numerous minor vortices along such BLFs most probably owe their existence to barotropic instability. The results from the analysis discussed above do not necessarily contradict this statement since Tsuboki and Wakahama excluded small vortices of diameter less than 100 km from their study, i.e. those vortices most likely to be caused by barotropic instability.

The possible role of low static stability near the Earth's surface in the growth of polar lows was explored by Duncan (1977). He used a quasi-geostrophic model to look at normal mode solutions for unstable disturbances including both vertical and horizontal wind shears, with low static stability near the Earth's surface. Three cases, all of them developments at a fairly southerly latitude, were analysed. In each situation the developments took place within a baroclinic zone below an upper-level wind speed maximum. Good predictions were made for the wavelength of two of the observed disturbances. In the third case, Duncan suggested that strong horizontal variations in the static stability, which cannot be represented in a quasi-geostrophic model, were responsible for the poor prediction of wavelength. In all three cases the perturbation amplitudes were very small at upper levels and largest close to the surface, implying that the growth of the perturbations by baroclinic processes should be expected primarily at low levels.

Duncan concluded that polar air depressions could be considered as shallow baroclinic waves and that the conversion of available potential energy to eddy kinetic energy occurs in the lowest 200–300 hPa of the atmosphere when the low-level static stability is small. No significant growth could be ascribed to barotropic processes. The rate at which the energy conversion occurred, and therefore the growth rate, depended on the thermal wind and the static stability. In general, the largest vertical wind shear at low levels on the synoptic scale exists below the jet stream so that this is a likely region for polar low developments.

Duncan, in his 1977 paper discussed above, noted that a necessary condition for polar air depressions to develop 'is that a vertical wind shear exists such that the thermal wind and the mean flow at low levels are almost parallel'. The concept of the development of polar lows in a *reverse shear flow* (where the thermal wind and the mean flow are still parallel but of opposite direction) was first put forward by Duncan (1978). While most large, baroclinic waves develop when the surface wind, the thermal wind and the progression of the systems

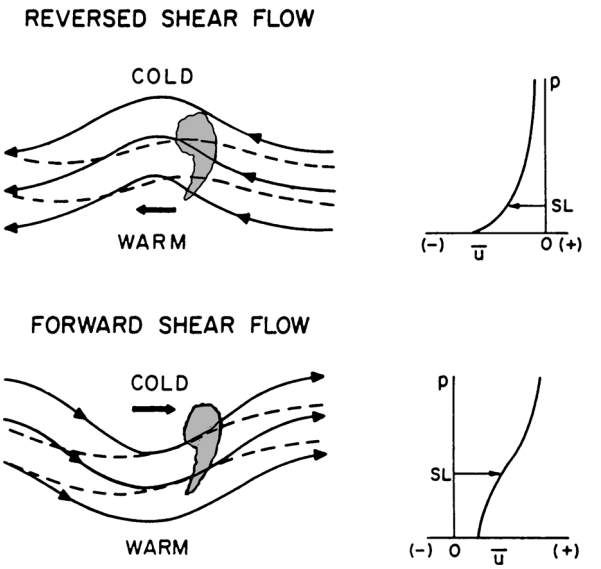


Figure 4.11. Comparison of the structure of disturbances in reverse-shear flow (above) and (normal) foreward-shear flow (below). Solid lines show streamlines, broken lines isotherms at the steering level; heavy arrows show phase propagation vector and steering level (SL) wind; stippling indicates the extent and position of an associated comma cloud (from Businger and Reed, 1989b, in Twitchell *et al.*, 1989. © A. Deepak Publishing).

are in the same direction, Duncan considered disturbances where the surface wind and thermal wind were in opposite directions, and the magnitude of the horizontal wind decrease with height. The configurations of reverse shear and forward-shear systems are illustrated in Figure 4.11.

In this situation, relative to the motion of the system, there is warm air to the left of the path and colder air to the right. The effects of horizontal advection will be to move warm air behind the trough so that kinetic energy will be gained at the expense of available potential energy if ascending motion predominates behind the trough with descending motion in the cold air ahead. Duncan used the model described above to investigate the growth of disturbances in a reverse shear flow using atmospheric conditions from a polar low development from 10 December 1976. It was found that the most unstable wave in the model had a wavelength of 900 km and a phase speed of 10.8 m s^{-1} resembling the observed polar low. Based on this, Duncan suggested '…that the observed disturbance was baroclinic in nature and that its structure was probably similar to that of the 900 km wave in the numerical model.'

Haugen (1986) used a three-dimensional primitive equation model to simulate a reverse shear polar low development within a channel over the Norwegian

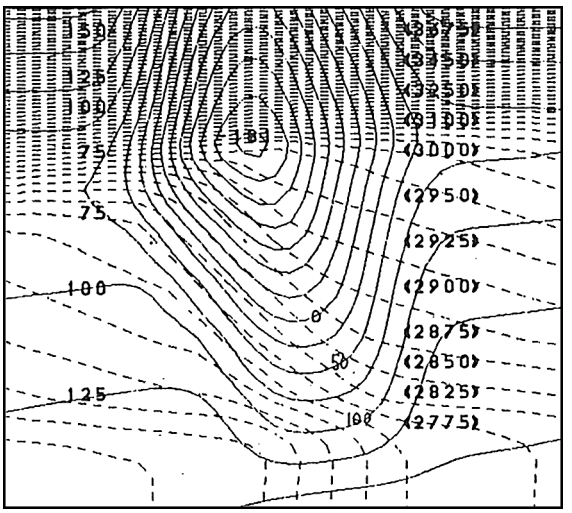


Figure 4.12. East–west cross-section across the Norwegian Sea, with Greenland to the left and the Norwegian coast to the right. Solid lines show wind components perpendicular to the cross-section and broken lines the potential temperature at the start of simulation. Labels on contours show the basic current in units of 0.1 m s^{-1} (positive wind velocities towards south) and the potential temperature in units of 0.1 K (from Haugen, 1986). (Figure used courtesy of the Polar Lows Project, Norwegian Meteorological Institute.)

Sea from 15 to 17 December 1982. A large-scale low with its centre over the northern part of Scandinavia advected polar air southwards over the Norwegian Sea, the wind direction in the lower and middle troposphere being opposite to the thermal wind direction. An east–west cross-section across the basic flow showing the wind component perpendicular to the cross-section and the potential temperature, is shown on Figure 4.12.

Greenland is situated to the left in the figure and the Norwegian coast to the right. The capped boundary layer with its constant potential temperature in the vertical can be maintained and increase in thickness when the polar air mass moves southwards. The east–west baroclinicity, concentrated in the middle of the channel with the coldest air to the west (close to Greenland) caused the low-level northerly basic flow to decrease with height, creating a weak southerly jet stream at tropopause height. The troposphere was initially stable on the cold side (except within the shallow, cold boundary layer) and in the middle of the region, but close to being conditionally unstable on the warm side.

The development of a fairly large polar low (diameter *c.* 1000 km) from the initial basic conditions shown on Figure 4.12 is illustrated in Figure 4.13, showing geopotential height and potential temperature after 24 and 36 h of simulation.

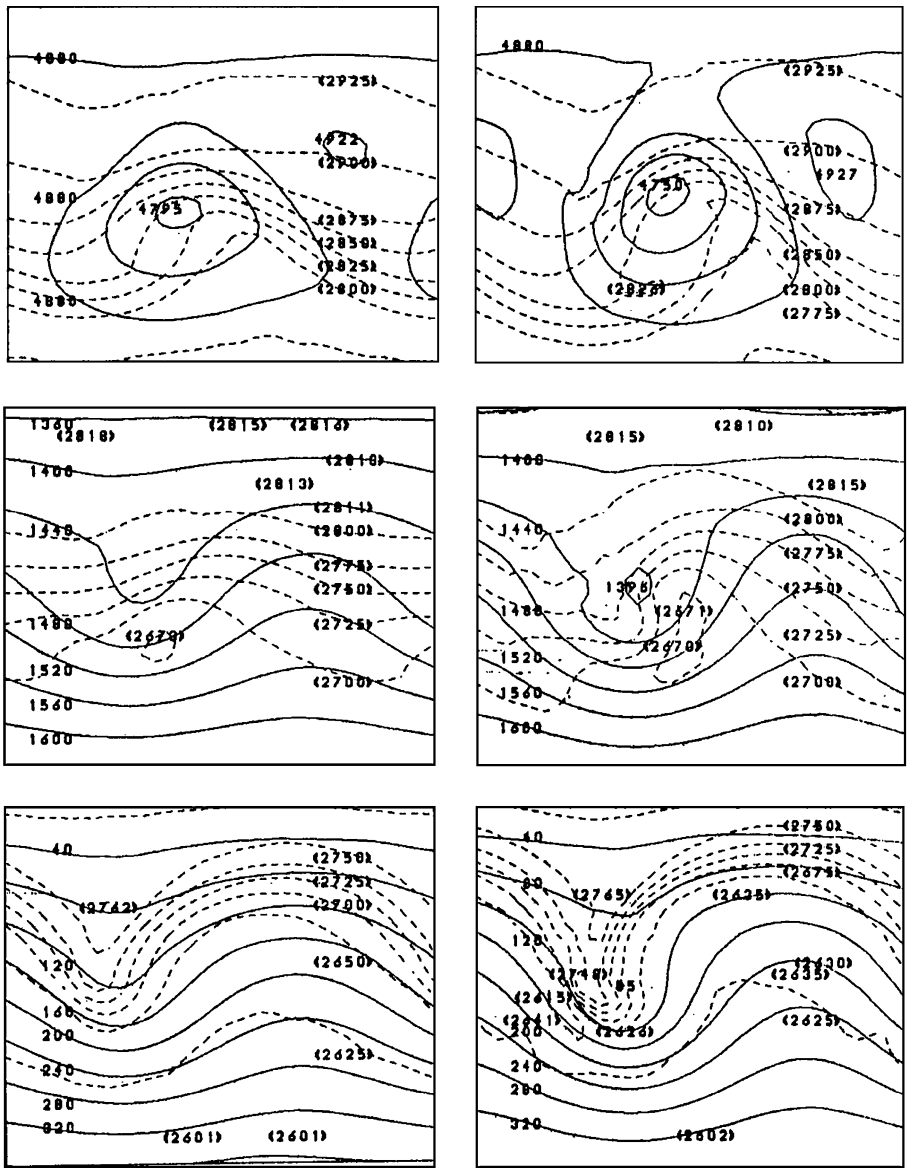


Figure 4.13. Horizontal contours of geopotential height (solid lines) and potential temperature (broken lines) after, respectively, 24 h of simulation (left column), and 36 h (right column) illustrating a reverse shear development corresponding to the basic flow shown on Figure 4.12. Pressure levels, from top, 500, 800 and 900 hPa. Labels on contours show heights (in m) and potential temperatures (in units of 0.1 K) with intervals 40 m and 2.5 K, respectively. The Greenland side is at the bottom of the figures (from Haugen, 1986). (Figure used courtesy of the Polar Lows Project, Norwegian Meteorological Institute.)

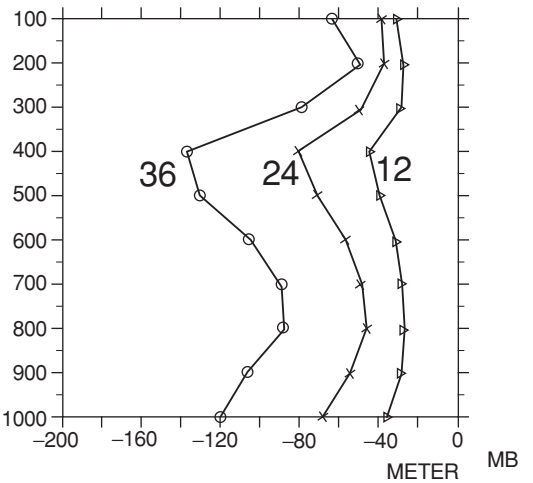


Figure 4.14. The vertical variation of the maximum amplitude of the low seen on Figure 4.13 as a function of time (12, 24, 36 h). The values describe the maximum height difference of a pressure surface relative to that of the initial (after Haugen, 1986). (Figure used courtesy of the Polar Lows Project, Norwegian Meteorological Institute.)

Greenland is at the bottom of the figures and the low moved southwards in the direction of the basic low-level flow. The development *seems* small at the surface, because of the strong pressure gradient of the basic flow. The basic flow across the channel was smallest at the 500 hPa level where a closed circulation developed. As a measure for the growth rate of the low, Haugen used the difference in height of a pressure surface from the initial stage. The vertical variation as a function of time is shown on Figure 4.14, with greatest amplification at the surface and at the tropopause level, in accordance with Eady (1949) in his analysis of the development of a baroclinic wave.

During the simulation of the reverse shear development the static stability changed because of differential horizontal advection with heating near the surface, as well as due to vertical stretching. As a result of this, the low levels, in particular, became conditionally unstable (Figure 4.15). Haugen interpreted this as a possibility of enhanced growth of such systems through release of latent heat from convection. This result agrees well with the observation of reverse shear systems (see Section 3.1.4) in the way that widespread deep convection often accompanies these developments.

A further study of reverse shear instability, by Reed and Duncan (1987), examined the development of four polar lows that had formed in January 1993 over the Greenland Sea in a shallow baroclinic zone at a time when the flow was from the northeast at the surface and the thermal wind was from the opposite

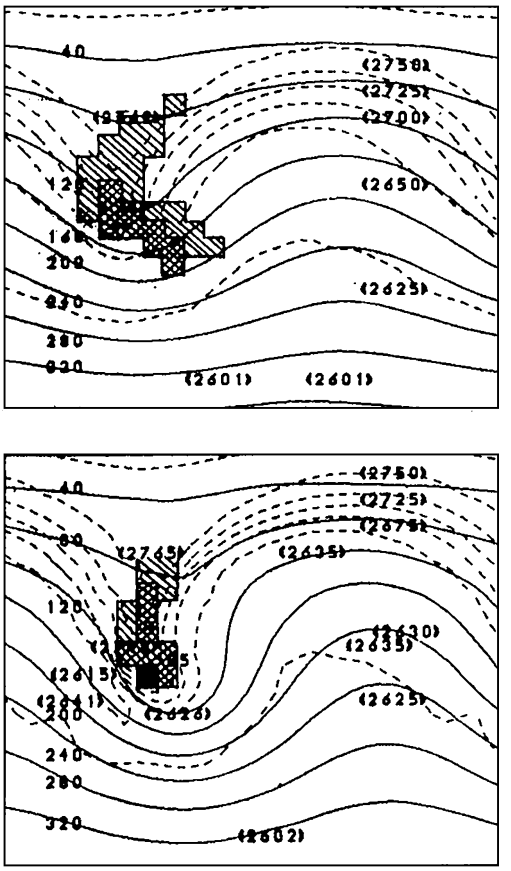


Figure 4.15. Horizontal contours of geopotential height (solid lines) and potential temperature (broken lines) at the 950 hPa surface after 24 h (upper) and 36 h (lower) of simulation (units as Figure 4.13). Areas with conditional instability are hatched. Single-hatched indicates instability over 900–850 hPa; cross-hatched, 900–800 hPa; solid, 900–700 hPa (from Haugen, 1986). (Figure used courtesy of the Polar Lows Project, Norwegian Meteorological Institute.)

direction. At the 500 hPa level, only a weak flow, generally from the north, was observed. Reed and Duncan applied their quasi-geostrophic model (Duncan, 1977) and obtained solutions for the fastest growing mode for wavelengths in the range 300–1000 km. The growth rates of these modes are shown in Figure 4.16 and indicate that wavelengths around 500 km are the most unstable. These model disturbances were confined almost entirely to the fastest moving lower layers adjacent to the surface as illustrated on Figure 4.17. Reed and Duncan suggested that the propagation of the disturbances in the opposite direction to the shear was explained by the fact that short baroclinic waves moved approximately with the mean wind in the layer in which they are embedded.

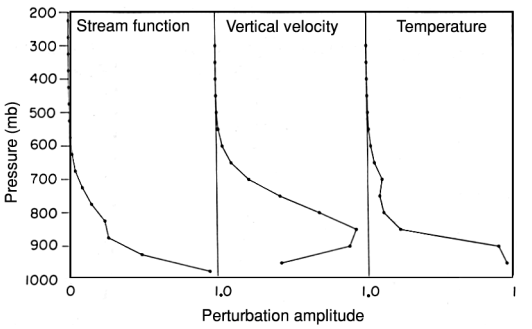


Figure 4.16. Growth rate of pertubations within a reverse shear flow as a function of wavelength for an experiment with no horizontal shear and vertical wind profile observed at the storm track (from Reed and Duncan, 1987).

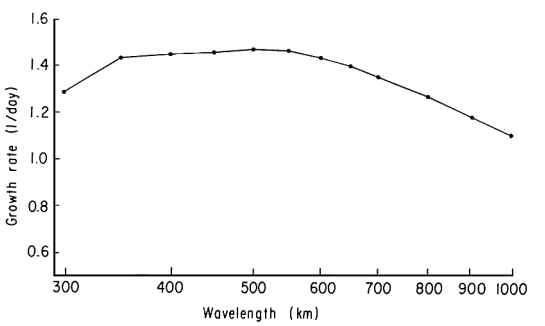


Figure 4.17. Perturbation amplitudes for the stream function, vertical velocity and temperature, for the 500 km wave of Figure 4.16. The magnitudes are in normalized arbitrary units (from Reed and Duncan, 1987).

Moore and Peltier (1989) extended the study of baroclinic wavetrains by Reed and Duncan arguing that these wavetrains should be considered as a new cyclone-scale mode of baroclinic instability discovered by them (Moore and Peltier, 1987). This new mode, however, is filtered out by both the quasi-geostrophic and geostrophic momentum approximations to the primitive equations. Specifically, Moore and Peltier in their 1989 study argued that the quasi-geostrophic approximation used by Reed and Duncan was invalid because of two factors: first, the static stability in the environment in which the polar lows grew varied strongly in the horizontal, and second, the background Richardson number field was of the order unity.

Using the primitive equations, Moore and Peltier considered the stability of a two-dimensional baroclinic zone to three-dimensional, small amplitude perturbations and demonstrated that the environment in which the polar low wavetrain was observed to develop was unstable to the 'cyclone scale branch of baroclinic instability' found by them. The predicted doubling time and

wavelength of the most unstable wave were in good agreement with the observations made by Reed and Duncan (1987).

The short wave-length of the disturbances found in theoretical studies were in good agreement with the observed systems, but as noted by Reed and Duncan the predicted propagation speed was too fast (that is, the steering level in the reverse shear flow was too low). They suggested that cumulus convection might have served to deepen the system in the vertical, thus slowing its motion. They also suggested that baroclinic instability alone was not capable of explaining the rapid development and that another mechanism, possibly latent heat release from deep convection, also played an important role.

To date there has been no work on reverse shear flow in relation to mesoscale lows in the high latitude areas of the Southern Hemisphere. However, there is no reason to assume that such vortices are confined only to the Arctic.

Further work on baroclinic instability, using the Eady model but extended to two layers, has been carried out by Blumen (1979). In this study the static stability was horizontally uniform but different in each layer and the wind shear was uniform throughout both layers. An analysis of the unstable growth rates showed that the instability was associated with the delta function distribution of potential vorticity at one boundary and at the interface between the two layers. This work showed that the short- and long-wave baroclinic instabilities depend on the relative layer depths, along with the jump in static stability between the two layers. He found that the model gave a maximum of baroclinic instability of much shorter wavelengths when the lower layer had a stratification close to an adiabatic lapse rate. This work was important in confirming the earlier results of Mansfield and Duncan in showing that small-scale instabilities can grow as a result of abrupt changes in static stability and/or wind shear.

The Blumen model was successfully used by Rasmussen and Aakjær (1992) to explain a baroclinic polar low development over the North Sea and Denmark on 28–29 March 1985 (see Section 1.6.6). A radiosonde ascent close to the region of development showed that the conditions as set up by Blumen for this type of development were clearly fulfilled. The growth rate curve computed from the data observed is shown on Figure 4.18. The left branch of the curve with a 'most preferred wave length' around 4000 km corresponds to the 'normal' one-layer Eady solution, whereas the right branch, with a maximum at a much shorter wavelength, is a feature caused by the extra degree of freedom in the vertical. Waves within this interval may, provided they form within a cold air mass, as in the case studied by Rasmussen and Aakjær, be interpreted as short-wavelength, baroclinic polar lows.

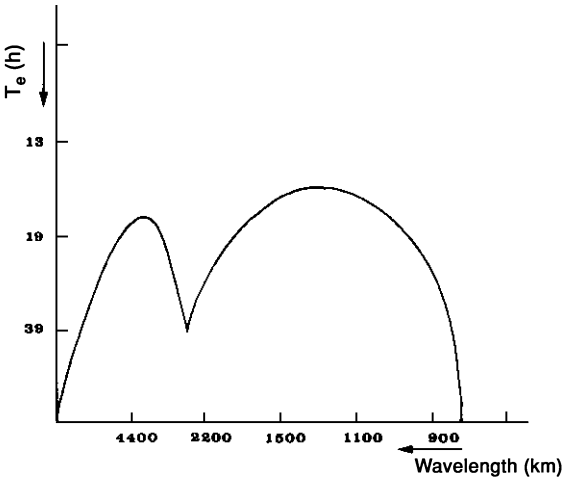


Figure 4.18. *e*-folding time (decreasing upwards) as a function of wavelength for unstable waves obtained from Blumen's model (Blumen, 1979) for a baroclinic polar low development in the North Sea (from Rasmussen and Aakjær, 1992).

In the studies discussed above a normal mode approach has been used to find the growth rate and the wavelength of the most rapidly growing polar lows within a baroclinic unstable environment. An alternative method of instability analysis is the initial value approach motivated by the recognition that in general the perturbations from which storms develop cannot be described as single normal mode disturbances, but may have a more complex structure. As discussed above strong dependence of cyclogenesis on initial conditions occurs when an short-wave, upper-level trough or a corresponding upper-level potential vorticity anomaly is advected into a region with a pre-existing meridional temperature gradient at the surface.

From the very start of the intensive research on polar lows in the 1970s and 1980s it was realized that virtually all high latitude polar low developments were triggered by upper-level disturbances in the form of short-wave cold troughs (e.g. Rasmussen, 1985b; Businger, 1985). Following Pettersen's ideas, it was assumed that the short-wave, upper-level troughs interacted with the shallow baroclinic zones along the ice edges resulting in the formation of a low-level circulation. Such a low-level circulation, generated through baroclinic instability, subsequently might act as a trigger for a 'convective' polar low (two-stage development).

While this type of baroclinic development triggered by short-wave, upper-level troughs undoubtedly occurs leading to the formation of polar lows (an example is presented below) then on the other hand its importance probably has been exaggerated. Forcing from an upper-level trough will generate

a response whose vertical scale is proportional to the horizontal scale of the upper-level system. As shown by Holton (1992) for quasi-geostrophic systems, forcing at a given altitude will generate a response whose vertical scale is given by λ^{-1}, where λ is defined by $\lambda = (k^2 + l^2)\sigma f_0^2$, k and l being the wavenumbers in the x and y direction respectively, f_0 the Coriolis parameter and σ the vertical stability. Upper-level vorticity advection associated with disturbances of large horizontal scale and/or a small vertical stability in the environment will thus produce geopotential tendencies that extend down to the surface with little loss of amplitude, while for disturbances of small horizontal scale and large static stability the response is confined close to the levels of forcing. From a PV perspective, this is expressed in the way that the Rossby penetration depth, which describes to what extent an upper-level PV anomaly may penetrate towards lower levels, is given by $H = fL/N$, L being the horizontal scale of the system and N the Brunt–Väisälä frequency. It is doubtful, therefore, whether the small-scale, upper-level troughs can trigger low-level baroclinic developments, unless the Arctic air mass has become de-stabilized due to adiabatic heat transfer from the surface (see Sections 4.4.2 and 4.4.4 for more details).

Another factor that may limit the influence of the above discussed 'self development process' is the fact that the ice edge-generated baroclinic zones are very shallow. Some scientists have questioned the importance of baroclinic instability for polar low developments involving shallow air masses and their corresponding low-level baroclinic zones generated along the edge of the polar ice. Økland (1989), considering the baroclinicity over the ocean associated with cold air outbreaks from snow/ice-covered regions, pointed out that the baroclinicity 'is caused by the downstream increase in temperature and depth of the convective layer, and it is highly questionable if this baroclinicity can support baroclinic wave development'.

Albright *et al.* (1995), in a study of a polar low development over the Hudson Bay (see Section 5.1 for a discussion of this case), likewise questioned the role of these shallow, ice edge-generated baroclinic zones as a source of low-level baroclinic instability. The low studied by them formed over a relatively small ice-free region in the eastern part of the otherwise ice-covered bay. Albright *et al.* concluded that this case provided an example in which baroclinicity appeared to have played a minor role in the polar low development. Instead they pointed to latent heating released in deep convection as the overwhelming cause of the intensification.

In some cases, however, when the horizontal scale of the upper-level disturbance is sufficiently large and the low-level baroclinic zone is deep enough, baroclinic developments through mutual interaction between an upper-level trough and a low-level baroclinic zone may take place as illustrated

schematically on Figures 4.3 and 4.4. Nordeng (1990) in a diagnostic study of two polar low developments over the Norwegian Sea found that both of the relatively weak developments were baroclinic and noted that 'the main triggering mechanism for both cases was approaching upper-level troughs which interacted with low-level baroclinic zones created by 'fixed surface forcing', i.e. sea surface temperature anomalies and ice-edges'. In both cases the main baroclinic zone, i.e. the polar front, or a branch of this zone was situated close to the place of development near Svalbard. As such, the two developments studied by Nordeng were representative of only a rather small group of polar lows, i.e. lows that develop close to the main baroclonic zone (for more details see Section 5.1.2, 'Nordeng's cases').

Occasionally, strong vortices qualifying as polar lows form along the semipermanent baroclinic zone over and off the pack ice adjacent to the northeast Greenland coast. An example of this was presented by Rasmussen and Cederskov (1994). In this case a strong polar low, which dominated the weather in the region for more than two days, formed near the ice edge. The development was dominated by baroclinic forcing involving a marked low-level baroclinic zone as well as strong upper-level forcing, i.e. a type-B development. Convection, on the other hand, played a very minor role throughout the lifetime of this low, which achieved wind speeds around 25 m s^{-1}.

The development was triggered on 2 March 1989 as an upper-level, synoptic-scale cold trough approached the region around Scoresbysund (70.4° N, 21.4° W) from the west after crossing the ice cap. As the region with differential vorticity advection ahead of the upper-level trough spread out over the low-level baroclinic zone along the northeast Greenland coast a strong polar low developed within 12 h. The structure of the low-level baroclinic zone along the northeast Greenland coast just prior to the polar low development is illustrated through the 2 m temperature field shown in Figure 4.19a. The small wave (indicated by an arrow) and the associated increased gradient in the isotherm field near 75° N, 12° W northeast of Scoresbysund marks the centre of a cyclonic disturbance being formed within the zone of pronounced temperature gradient. The precise location of the development in this case was probably determined by the distribution of the sea ice in the region, the cyclonic disturbance initially forming in a region with only low ice concentration. In the following hours the cyclonic disturbance developed rapidly into a (baroclinic) polar low with well-defined low-level frontal zones (Figure 4.19b). A nearby radiosonde ascent (Figure 4.19c) from Danmarkshavn (76.8° N, 18.7° W, position indicated on Figure 4.19a) at 1200 GMT 2 March 1989, only about 200 km away from the location of the incipient low, illustrates the very stable conditions within which this low developed.

(a)

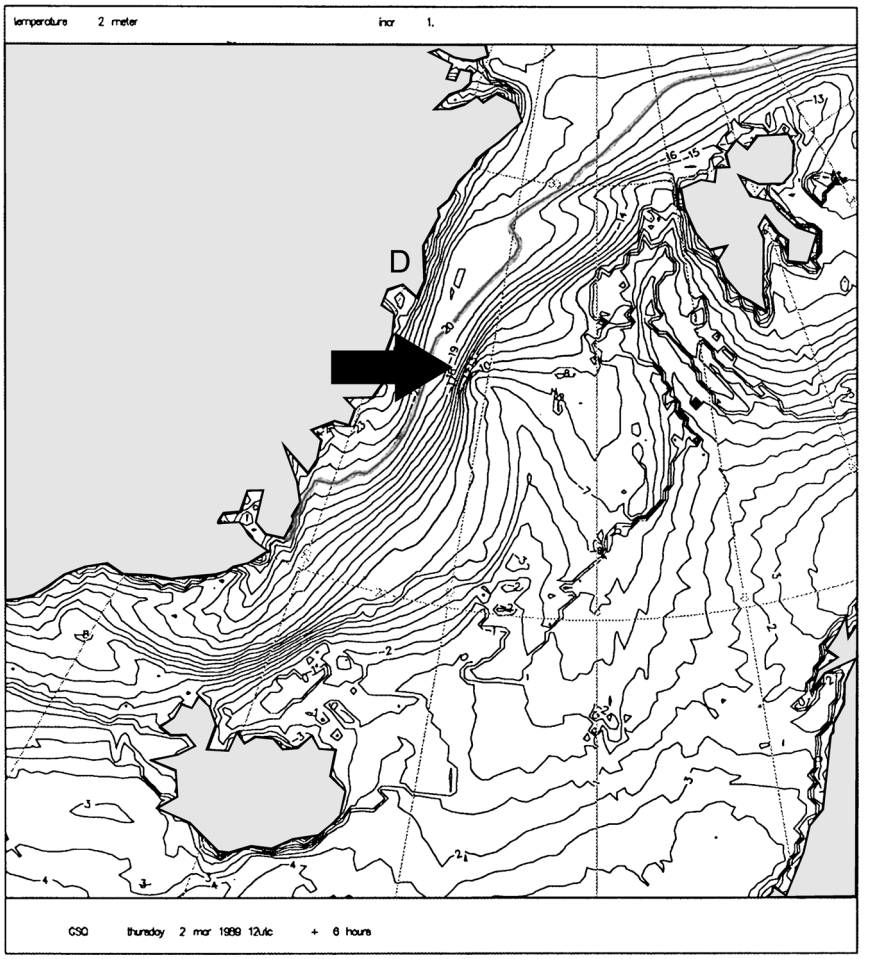
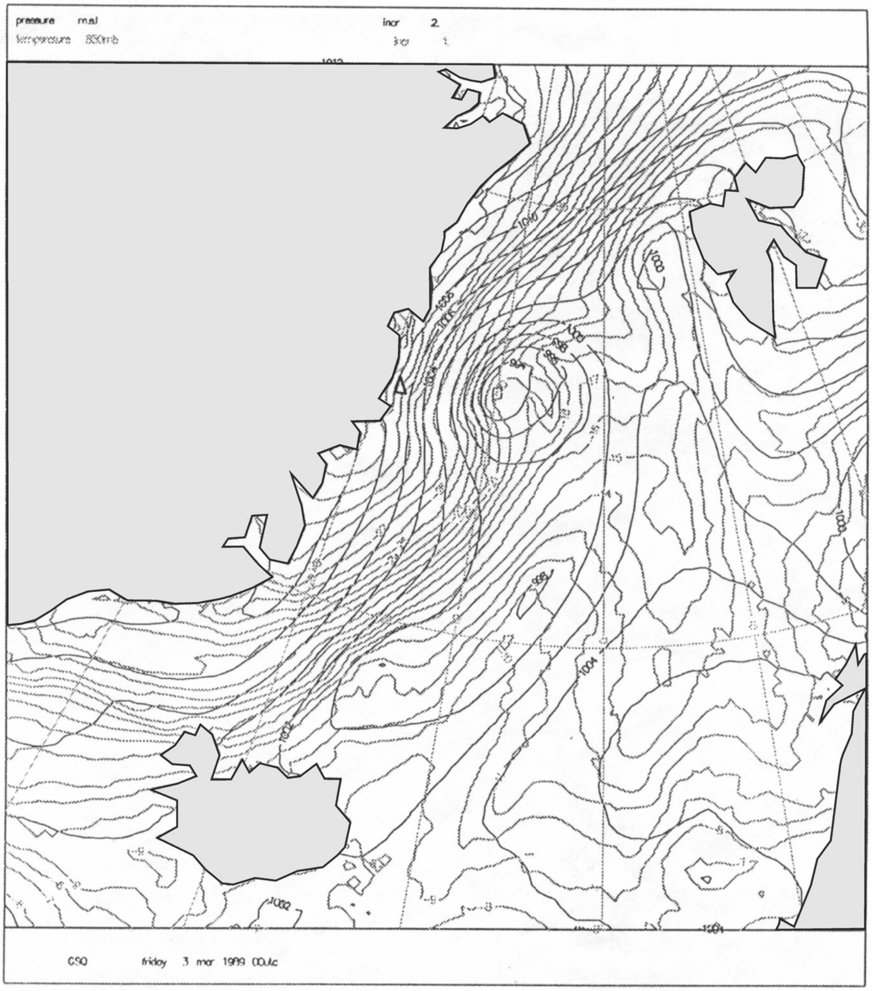


Figure 4.19. Meteorological fields and satellite imagery for the 2–4 March 1989 polar low case. (a) 2 m temperature (intervals 1 °C) at 1800 GMT 2 March. A zone of pronounced temperature gradient along the pack ice adjacent to the northeast Greenland coast shows the position of a strong, low-level baroclinic zone. A region of increased temperature gradient near 75° N, 12° W indicates the position of the centre of a cyclonic disturbance which rapidly developed into a polar low. D indicates the location of Danmarkshavn. (b) Surface pressure (thin solid lines, 2 hPa interval) and 850 hPa temperatures (dotted lines, 1 °C interval) at 0000 GMT 3 March showing a polar low northeast of Scoresbysund on the low-level baroclinic zone along the northeast Greenland coast. (c) Radiosonde ascent from Danmarkshavn (76.8° N, 18.7° W) (position indicated on (a)) at 1200 GMT 2 March. (d) 2 m temperature (intervals 1 °C) at 1800 GMT 3 March. The southern part of the zone of pronounced temperature gradient situated along the coast on 2 March (a) had now, following the development of the polar low, been displaced east. (e) Surface (10 m) wind field at 1800 GMT 3 March. The winds are plotted in the conventional way, a long barb signifying 10 kt, a short barb 5 kt. Thin solid lines show isotacks at 1 m s^{-1} intervals. (f) Surface pressure (thin solid lines, 2 hPa intervals) and 850 hPa temperature (dotted lines, 1 °C intervals) at 0000 GMT

(b)



Caption for Figure 4.19 (*cont.*). 4 March, showing the polar low in its mature stage between northeast Greenland and Svalbard. (g) An infra-red satellite image for 1316 GMT 3 March showing the mature polar low between northeast Greenland and Svalbard. (Image courtesy of the NERC Satellite Receiving Station, University of Dundee. Charts and the radiosonde ascent courtesy of the Danish Meteorological Institute.)

Following the formation of the low, a strong outbreak of Arctic air, lead by a sharp Arctic front, affected the region north of Iceland, while advection of warm air took place further east and north. On the afternoon of 3 March the surface temperature field (Figure 4.19d) showed the characteristic 'T-bone structure' of some large, synoptic-scale cyclones as described by Shapiro and Keyser (1990). Very pronounced horizontal wind shear was present over the leading edge of the frontal zones (Figure 4.19e). The model-derived wind field at this time showed wind velocities exceeding 28 m s^{-1}. Late on the evening

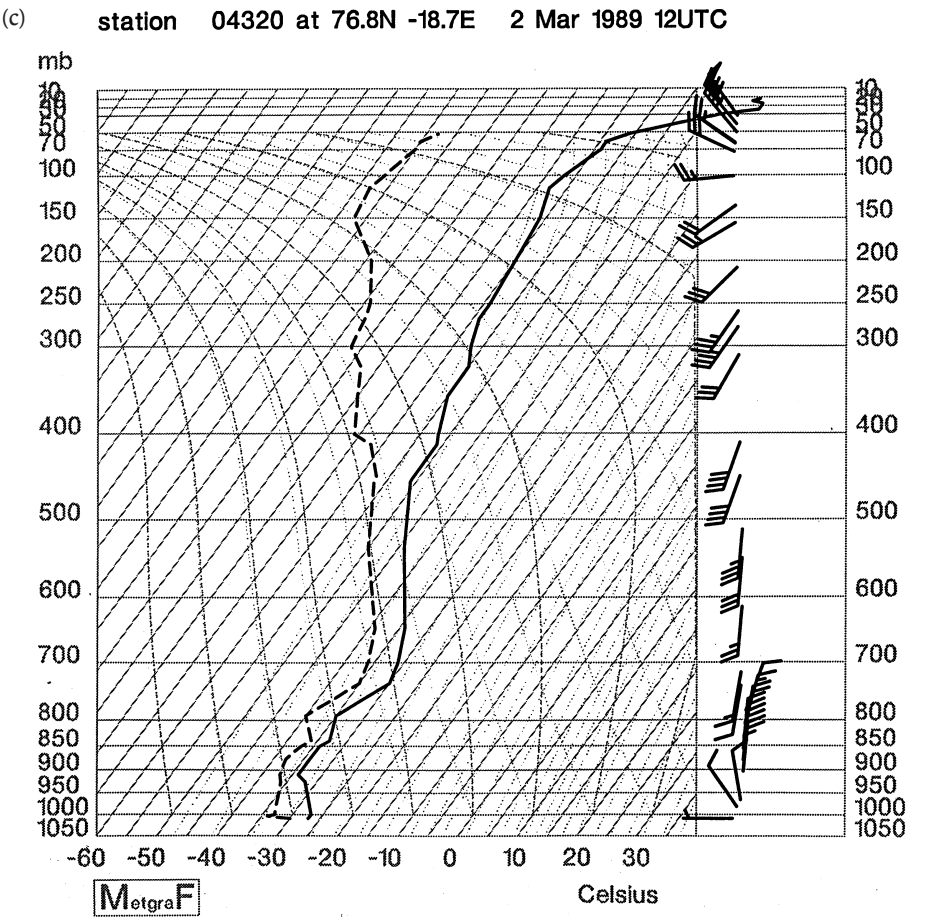


Figure 4.19 (cont.).

of 3 March the surface pressure within the centre of the low reached a minimum around 980 hPa. The surface pressure and 850 hPa temperature field from 0000 GMT 4 March, illustrating the mature stage of the low, is shown in Figure 4.19f.

This baroclinic development was further documented by a number of satellite images that confirmed the results from the numerical model. The polar low, as seen from a satellite during its mature stage around noon on 3 March, is shown in Figure 4.19g. The centre of the low, seen as a cloud spiral composed of a large number of low-level cloud streets converging into a common centre, was situated close to the ice edge. No deep convection can be seen within the region of the low in contrast to most polar low developments further east.

The relative importance of baroclinic instability versus latent heat release associated with deep convection was investigated by Sardie and Warner

(d)

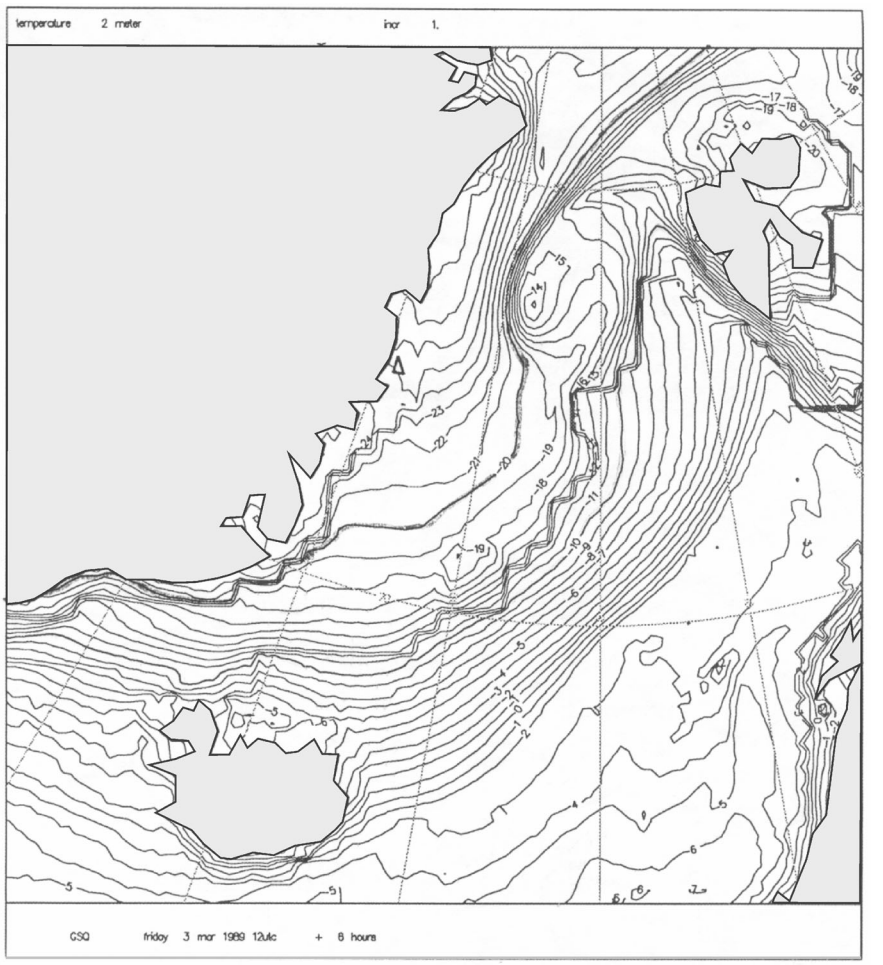


Figure 4.19 (*cont.*).

(1983) who used a three-layer, two-dimensional, quasi-geostrophic model that incorporated both these processes. The model included parameterizations of stable precipitation associated with moist baroclinic processes and convective precipitation associated with CISK. Seven case studies were used in the investigation, the model being run six times for each case. The model runs incorporated pure dry baroclinicity, pure CISK, moist baroclinicity and a combination of these processes. The study showed that the moisture in the boundary layer, the vertical distribution of convective latent heating and the mode of heat release (moist baroclinicity or CISK) were important parameters in determining the form of polar lows in their early stages. For the moist baroclinic modes, greater release of latent heat increased the maximum growth rate and decreased the wavelength of the system.

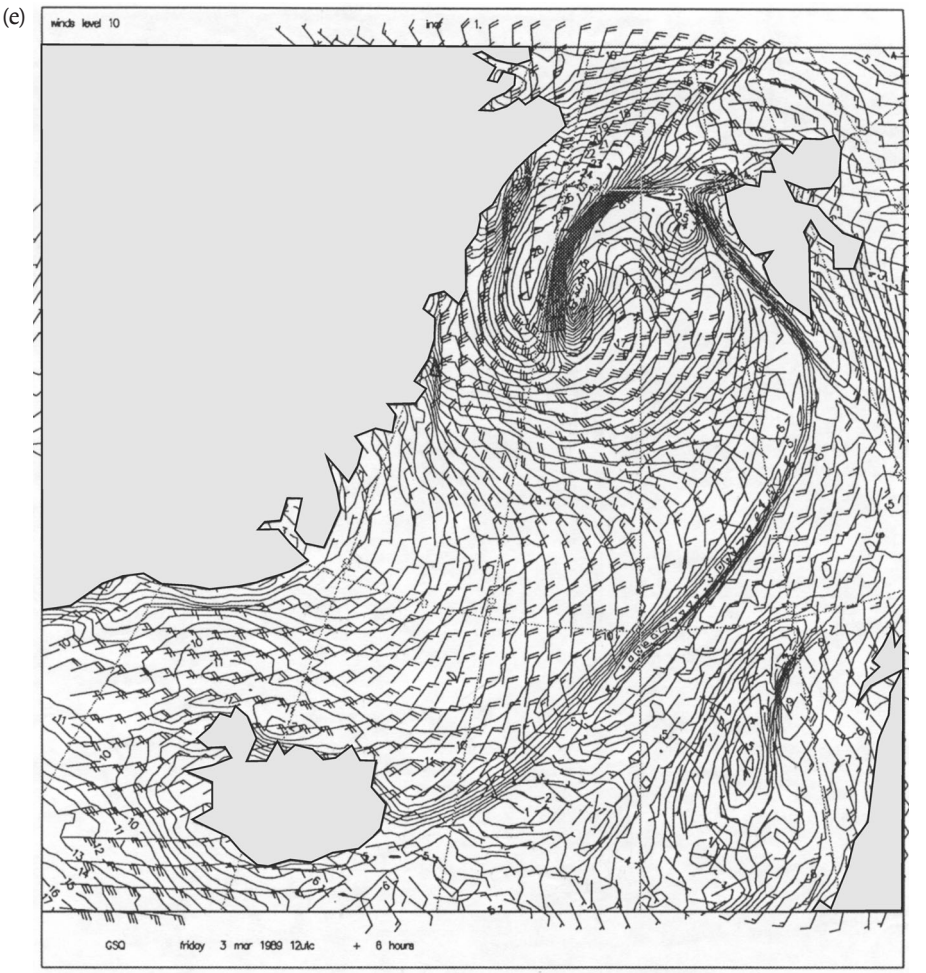


Figure 4.19 (*cont.*).

Overall the model results of Sardie and Warner (1983) showed that neither CISK nor dry baroclinicity on their own provided the necessary forcing to allow instabilities to grow to the observed wavelengths at the observed rates. They found that baroclinicity was important in the formation of polar lows in both the Pacific and the Atlantic, noting that the average latitude of the maximum of baroclinicity was 30–40° N in the central and western Pacific, 30–50° N in the western Atlantic and 20–40° N over Europe. Therefore polar lows in the Pacific forming around 40° N benefit from the presence of strong, deep baroclinicity, while polar lows over the Atlantic, developing around 60° N near Iceland are too far north to take advantage of any strong baroclinicity. Concerning the latter developments they suggested that CISK must operate in conjunction

(f)

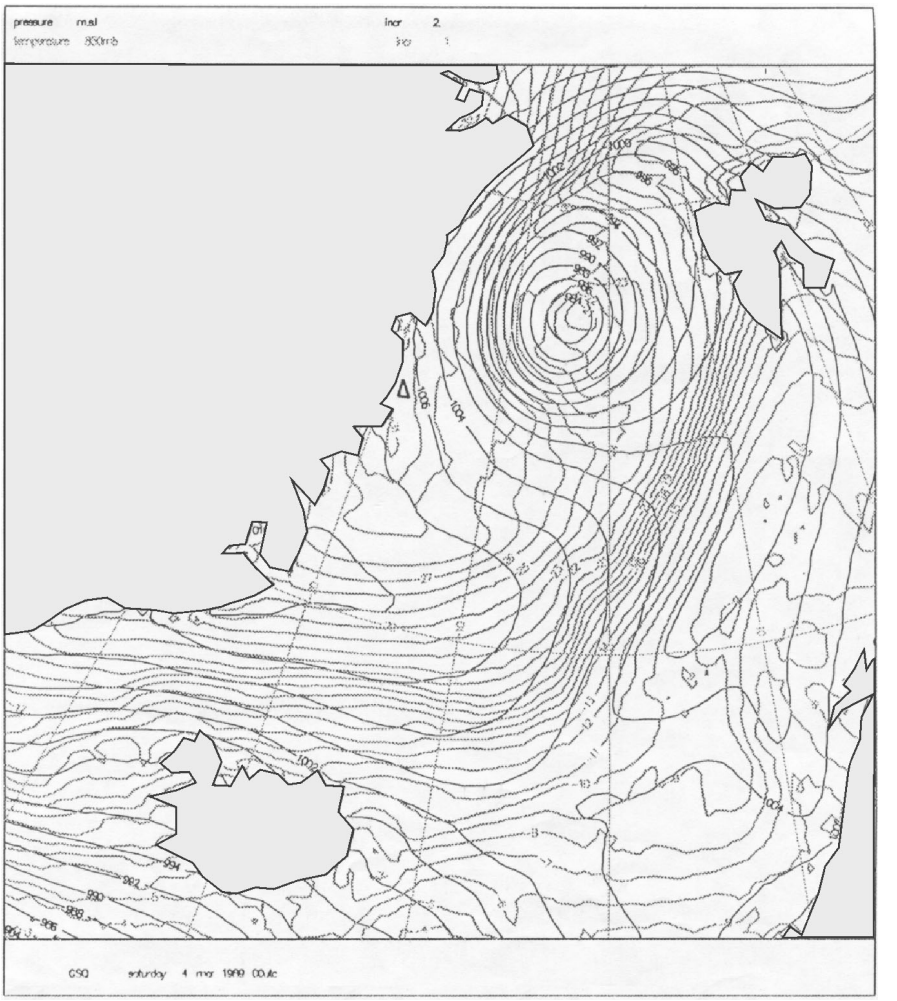


Figure 4.19 (*cont.*).

with shallow baroclinicity from residual circulations or occluded depressions. In this way moist baroclinic processes, not restricted to low levels, on their own may explain the genesis of polar lows observed in the Pacific, such as the systems with well-developed comma-shaped clouds observed within the strong baroclinic zone (Reed, 1979). These conditions provide a source of available potential energy (APE) throughout the whole troposphere to allow the perturbation to grow. On the other hand, moist baroclinicity as well as CISK were felt to be necessary in the development of Atlantic systems. The low-level wind shear, providing a low-level source of APE, complements an upper-level source due to CISK.

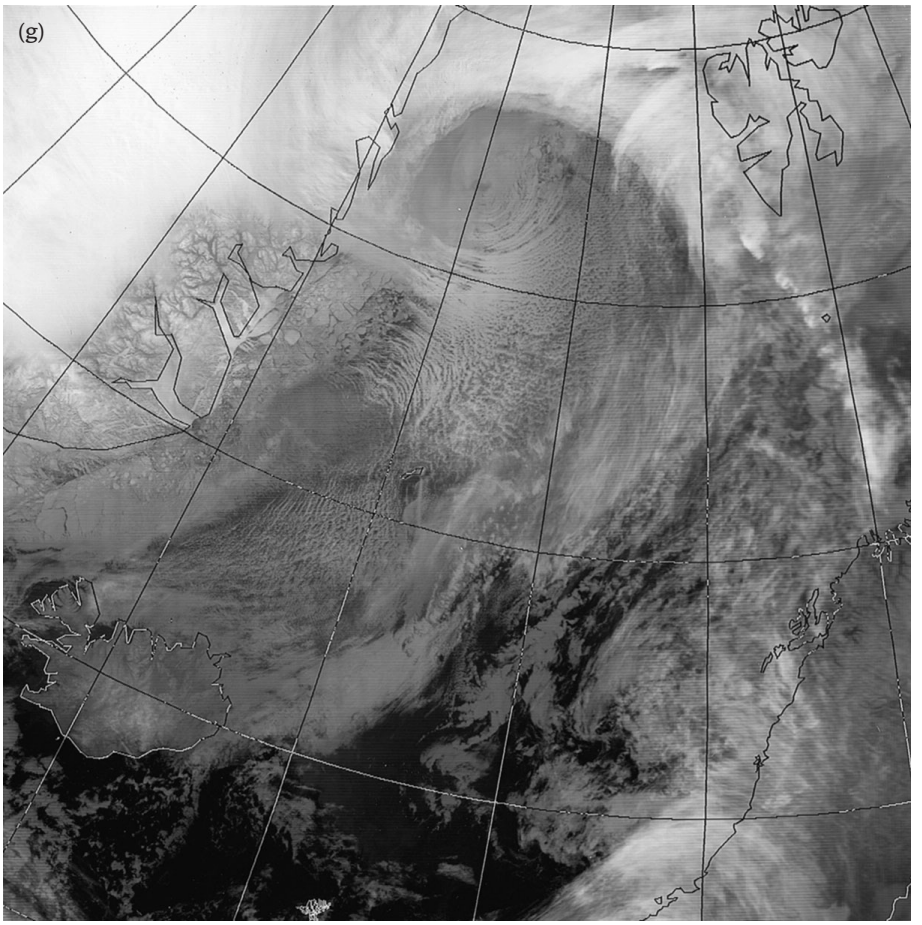


Figure 4.19 (*cont.*).

Another investigation into the nature of moist quasi-geostrophic instability was carried out by Mak (1982). In this study condensational heating was parameterized within a simple analytical model allowing the instability properties to be examined. It was found that as the heating intensity parameter was increased, the most unstable wave's growth rate increased significantly, its wavelength decreased significantly and its phase speed increased. The study also suggested that there is an upper limit for the growth rate, of about five times the dry model value, and a lower bound for the wavelength of the most unstable wave of about one-third the dry model value. This study also supported the concept that the baroclinic forcing in a disturbance could organize the condensational heating on a scale comparable to the wave itself.

High resolution, mesoscale models have now been applied to a number of polar low cases and used to examined the role of various instability mechanisms and synoptic situations. Grøndås *et al.* (1987a) used such a model to investigate developments within a reverse shear flow as discussed above. They found that the model was successful in reproducing the development of polar lows in such a flow with the output of the model being very similar to the observational data.

Building on their earlier study with a simple linear model, Sardie and Warner (1985) employed a full mesoscale model to examine two polar low developments in the Pacific and Atlantic Oceans. The Atlantic polar low examined developed on an intense but shallow baroclinic zone in the Denmark Strait region. The Pacific low also developed in a very strong baroclinic zone, but in this case the zone was deep. They found that the dry baroclinic effects accounted for much of the development in the Pacific case but that CISK was needed to correctly model the full development. It was also found that the development in the model was very sensitive to the shape of the vertical heating profile. Their simulation of the Atlantic case suggested that the baroclinicity was sufficient to allow a realistic initial development to take place while the polar low was close to the baroclinic zone. In this case baroclinicity was the dominant mechanism for development during the first 24 h of the simulation. However, it was found that both convective and non-convective latent heating and surface fluxes of sensible and latent heat were needed for simulation of the observed development after the polar low moved away from the baroclinic zone, which occurred during the 24–48 h period of the simulation. So moist baroclinicity and CISK were both found to be important to the observed development of both polar lows.

4.3 Barotropic instability

'Barotropic instability is a wave instability associated with the horizontal shear in a jet-like current. Barotropic instabilities grow by extracting kinetic energy from the mean flow field' (Holton, 1992). A number of researchers have, over the years, considered the possible role of barotropic instability in the formation of polar lows. Duncan (1977), in his study of three polar lows, found no significant growth due to barotropic processes. Reed (1979), Mullen (1979) and Sardie and Warner (1985) all concluded that, although barotropic instability may be present in a number of polar low developments, it nevertheless represents a minor contribution or no contribution at all to these developments. These results, however, were all based on considerations of the structure of the upper-level (polar) jet stream, and, according to Reed (1979), 'it seems doubtful

that a jet stream can ever be sharp enough to account for the very small-scale systems that develop over the oceans in winter…'.

Rasmussen (1983), on the other hand, in a discussion of polar low developments along BLFs and their associated shear zones west of Svalbard pointed out that polar lows may form as 'shear vortices', i.e. through 'low-level barotropic instability', along these lines. As discussed in Section 3.1, such shear lines are frequently observed to 'roll up' forming numerous vortices on different scales from the very small, with a horizontal scale of a few kilometres, up to much larger vortices, occasionally on the scale of a polar low (see Figure 3.10).

Bond and Shapiro (1991) considered barotropic instability as a possible mechanism for the formation of polar lows over the Gulf of Alaska, but could not conclude decisively whether this effect was important.

Nagata (1993) in a study of meso-β-scale vortices along the Japan Sea Polar Air mass Convergence Zone (JPCZ) noted that barotropic shear instability may be expected to work dominantly for the development of relatively small (meso-β-scale) vortices when a large amount of vorticity is concentrated into a narrow shear zone within a few tens of kilometres. Using a high resolution (6 km horizontal grid) model, Nagata simulated the formation of meso-β-scale vortices along the convergence zone. The simulated vortices appeared as waves on belts of concentrated positive vorticity of a width of a few tens of kilometres along the JPCZ (Figure 4.20a). They became increasingly sharp, and eventually the troughs of the vorticity belt formed mesoscale lows with pressure deficits of 2–4 hPa. The disturbances were characterized by spiral cloud bands around a 'dry eye' with a warm core structure. The vortices had a core of large positive vorticity of around 80 km diameter.

To confirm that the barotropic process dominated in the energetics of the developing disturbances, Nagata calculated the barotropic and baroclinic energy conversion rates within a 'strip region' around vortex V1 (see Figure 4.20). The basic zonal flow within the strip, as well as the eddy momentum fluxes, are shown on Figure 4.21a.

In a mean flow U with zonal and meridional perturbations u' and v' the variation of eddy kinetic energy (K_E) (Eqn. (4.1)):

$$K_E = \overline{1/2(u'^2 + v'^2)} \tag{4.1}$$

across the strip region on four pressure levels is shown on Figure 4.21b, while the barotropic and baroclinic energy conversions calculated as:

$$C(K_Z \rightarrow K_E) = -u'v'\overline{\mathrm{d}U/\mathrm{d}y} \tag{4.2a}$$

and $$C(A_E \rightarrow K_E) = \overline{-\omega'\alpha'} \tag{4.2b}$$

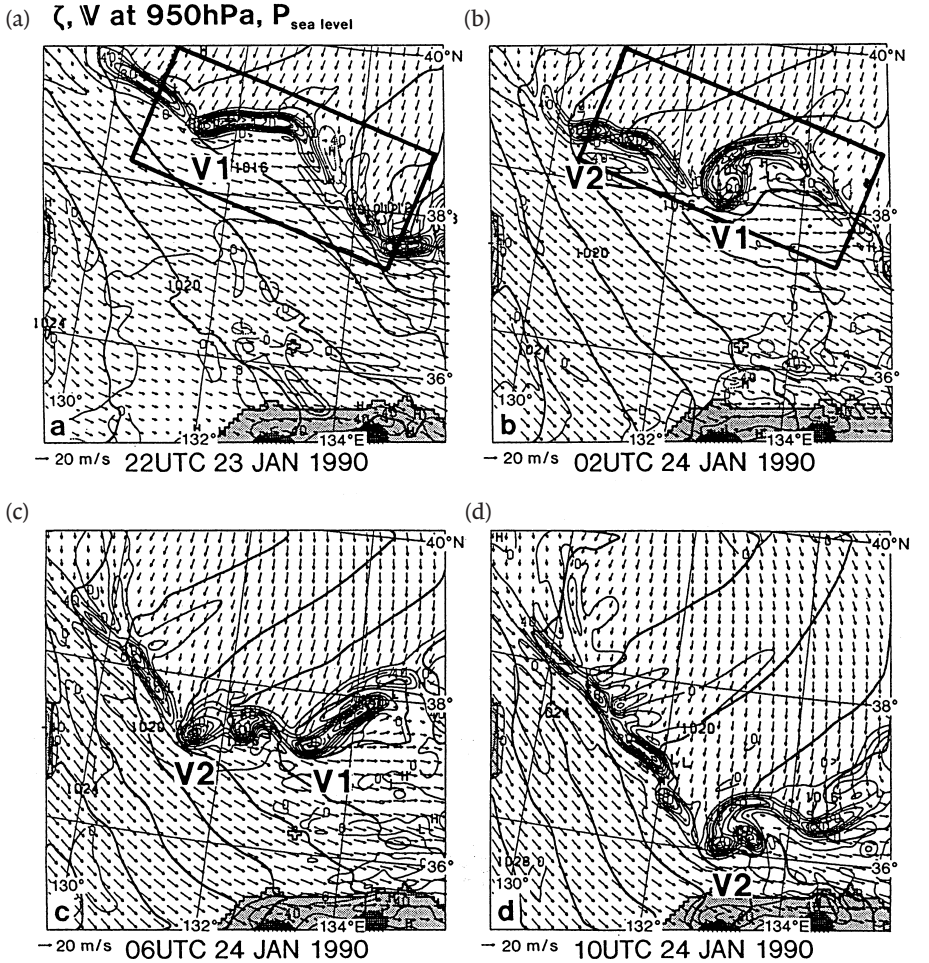


Figure 4.20. Evolution of simulated mesoscale vortices V1 and V2 along a shear line, seen in the 950 hPa relative vorticity field (thin lines; units 10^{-5} s^{-1}, contour intervals 20×10^{-5} s^{-1}; broken lines denote negative values). Wind vectors at every three grid points on the same level and sea level pressure (hPa) with 2 hPa contour intervals (thick lines) are also shown. Shading shows low land areas: (a) $t = 22$ h, (b) $t = 26$ h, (c) $t = 30$ h, (d) $t = 34$ h (after Nagata, 1993).

are shown on Figure 4.21c. The figures show clearly that the barotropic energy conversion dominates in most of the shear zone where the eddy kinetic energy is concentrated, while the baroclinic energy conversion gives a minor contribution. From the results of the energy conversion analysis, together with an agreement in spatial scale and growth rate between theory and the numerical simulation, Nagata concluded that the meso-β-scale vortices developed mainly due to barotropic shear instability.

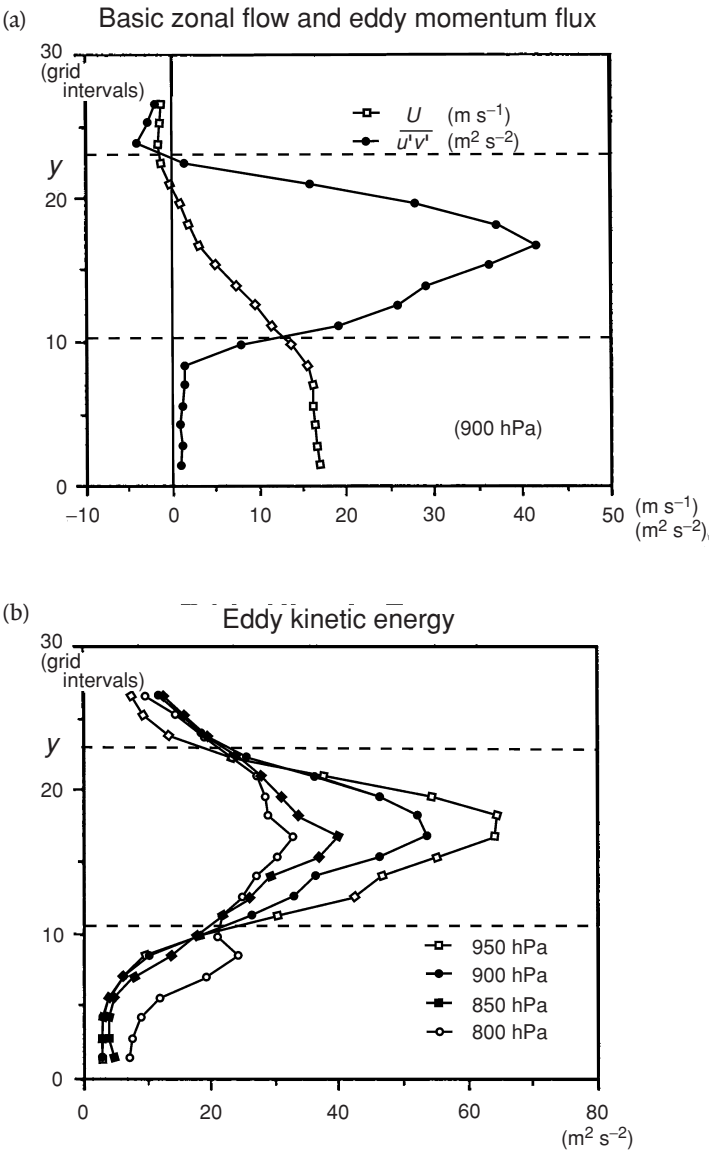


Figure 4.21. Basic zonal flow, eddy momentum flux, eddy kinetic energy and energy conversion rates for the strip region shown on Figure 4.20a. (a) Basic zonal flow ($\mathrm{m\ s^{-1}}$) and eddy momentum flux ($\mathrm{m^2\ s^{-2}}$) at 900 hPa. (b) Eddy kinetic energy ($\mathrm{m^2\ s^{-2}}$) distribution along y on four pressure levels. (c) Barotropic ($K_Z \rightarrow K_E$) and baroclinic ($A_E \rightarrow A_E$) energy conversion rates ($\times 10^{-3}\mathrm{m^2\ s^{-3}}$) along y at 900 hPa (from Nagata, 1993).

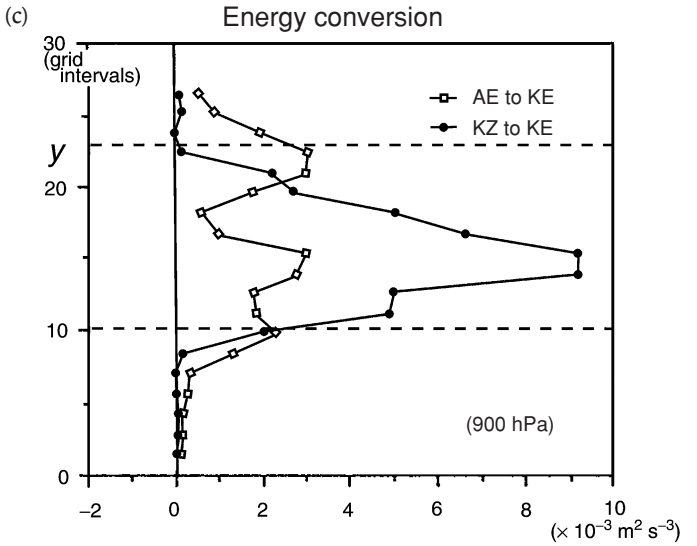


Figure 4.21 (*cont.*).

Detailed studies, such as the one by Nagata, have not yet been carried out for the Nordic Seas region. Satellite images, however, on numerous occasions have documented the presence of minor vortices along narrow shear lines within polar air outbreaks prior to polar low developments (see Section 3.1.4, 'Example: 25–27 January 1982'). Theory as well as observations therefore indicate that barotropic vortices forming along a shear line within a convectively unstable environment may 'focus' deep convection within a limited region affected by the vortex and in this way trigger a convectively driven polar low. On the other hand, there is no evidence that barotropic instability alone can lead to the formation of a polar low.

4.4 Potential vorticity thinking

4.4.1 Introduction

The PV (potential vorticity) concept has over the last decade received increased attention in theoretical work as well as in forecasting. PV is conserved in adiabatic, frictionless flow, and makes this dynamical parameter a very useful tool in identifying upper-air forcing that may trigger cyclogenesis. The effects from latent heating are probably more clearly understood by the PV approach, than by more traditional quasi-geostrophic methods.

For consistency, some important parts of the theory are repeated here.

Quasi-geostrophic PV (q_p) is given by the expression (see e.g. Holton 1992, pp. 164–5):

$$q_p = \nabla^2\psi + f + f_0^2 \frac{\partial}{\partial p}\left(\frac{1}{\sigma}\frac{\partial \psi}{\partial p}\right) \tag{4.3}$$

Where ψ is the streamfunction, the constant f_0 is the planetary vorticity or Coriolis parameter at a particular standard latitude, f the Coriolis parameter at the latitude we consider, σ is a static stability parameter, which is given as $\sigma \equiv -\frac{\alpha}{\theta}\frac{\partial \theta}{\partial p}$, α is specific volume, θ potential temperature and p pressure. q_p is thus expressed by, from left to right, the relative vorticity, planetary vorticity and stretching vorticity. If diabatic effects and friction are neglected, quasi-geostrophic PV is conserved following the geostrophic flow:

$$\frac{\mathrm{d}q_p}{\mathrm{d}t} \equiv \left(\frac{\partial}{\partial t} + \vec{v}_g \cdot \nabla_p\right) q_p = 0 \tag{4.4}$$

The unit of q_p is the same as for vorticity, s^{-1}.

The last term on the right hand side (r.h.s.) of Eqn. (4.3), the stretching vorticity, can be written as:

$$f_0^2 \frac{\partial}{\partial p}\left(\frac{1}{\sigma}\frac{\partial \psi}{\partial p}\right) = \frac{f_0}{S_p}\frac{T}{\theta}\left(-\frac{\partial \theta}{\partial p}\right) \tag{4.5}$$

In this relation $S_p \equiv -\frac{T}{\theta}\frac{\partial \theta}{\partial p}$ is another expression for the static stability parameter. S_p varies slowly with height in the troposphere.

An air column that moves adiabatically, confined between two selected isentropic surfaces, is stretched vertically as it moves into a region where the isentropic surfaces have wider separation. With downward motion in the lower portion of the column and upward motion in the upper part, the upper part must cool and the lower part warm adiabatically (see e.g. Holton, 1992, fig. 4.7). In such a region the expression (4.5) above becomes smaller (increasingly negative), and relative vorticity, given by the first term on the right hand side of Eqn. (4.3), has to increase in order to conserve q_p assuming that the planetary vorticity changes are small. In the case of shrinking, a decrease of relative vorticity takes place.

Quasi-geostrophic PV is used for describing large-scale (synoptic-scale) flow and it is necessary to obtain an expression for PV which describes more general flow systems, including those with large curvature and large Rossby numbers (Ro $\approx$ 1).

For this purpose the Ertel potential vorticity (EPV $\equiv q$; in the following referred to as PV) is used and is given by

$$q = \frac{1}{\rho}(\vec{\zeta}_a \cdot \nabla\theta) \tag{4.6}$$

The changes of q due to diabatic heating and friction are expressed by the following important relation (Hoskins *et al.*, 1985):

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{1}{\rho}(\vec{\zeta}_a \cdot \nabla\dot{\theta}) + \frac{1}{\rho}(\nabla\times \vec{F} \cdot \nabla\theta) \tag{4.7}$$

where ρ is air density, $\vec{\zeta_a}$ is the total vorticity, $\dot{\theta}$ denotes diabatic heating, and $\vec{F}$ is the friction force. Relation (4.7) shows that if the gradient of diabatic heating has a component along the total vorticity vector, the first term on the right hand side of (4.7) contributes to increased PV below the level of the diabatic heating maximum and a decrease of PV above. The spatial location of the diabatically-induced PV anomalies tends to be oriented along the direction of the absolute vorticity vector.

Thus an upper, negative PV anomaly tends to develop due to diabatic heating. As will be discussed in the next paragraph, an anticyclonic circulation is associated with a negative PV anomaly and this circulation may counteract and delay an advancing upper positive PV anomaly, contributing to a prolonged deepening phase of the cyclone (Stoelinga, 1996).

The second term on the right hand side of (4.7) is the friction term. It generally contributes to a decrease of PV, but in regions where the low-level wind has a component directed opposite to the thermal wind (warm fronts) it will contribute to an increase of low-level PV (Stoelinga, 1996). It is seen that for adiabatic, frictionless flow, q is conserved, as for the quasi-geostrophic case.

The expression for EPV above can be written in a somewhat simpler (and more familiar) way. It may be obtained by introducing isentropic coordinates in relation (4.6) for q. We thus have the relation for PV, referred to as *isentropic potential vorticity*:

$$q = (\zeta_\theta + f)\left(-g\frac{\partial\theta}{\partial p}\right) \tag{4.8a}$$

ζ_θ is the relative vorticity on an isentropic surface. q is usually expressed in PV units defined by

$$\{q\} = 10^{-6}\,\mathrm{m^2\,s^{-1}\,K\,kg^{-1}} \equiv 1\ \mathrm{PVU} \tag{4.8b}$$

(see Hoskins *et al.*, 1985).

The conservation of q has important implications. As shown in e.g. Holton (1992), conservation of PV, $\frac{\mathrm{d}}{\mathrm{d}t}q = 0$, for adiabatic, frictionless conditions describes how flow over a large-scale mountain barrier produces a stationary Rossby wave, partly explaining important features of the observed mean tropospheric wave pattern.

Another important consequence of the conservation of q is seen in the exchange of air between the lower stratosphere and the troposphere. At the upper jet core, air from the lower stratosphere may easily be advected into the troposphere. Assuming adiabatic conditions, a requirement frequently met at the tropopause, this advection takes place along isentropic surfaces. If an air column confined between two isentropic surfaces is advected from the lower stratosphere, into the troposphere, the lower static stability (decrease of $-g\frac{\partial\theta}{\partial p}$ in relation (4.8a) above) in the troposphere gives a compensating increase of relative vorticity (ζ_θ). Thus air descending from the stratosphere into the troposphere tends to acquire a cyclonic rotation. q is nearly conserved in the stratosphere, and frequently during the initial descent into the troposphere (radiative cooling is important, but works rather slowly). However, after onset of cyclogenesis, diabatic and frictional effects become essential, as described in relation (4.7).

4.4.2 The invertibility principle

The invertibility principle provides the streamfunction or geopotential height field from which wind and temperature fields are obtained once a PV anomaly and suitable balance and boundary conditions are provided. This may in a simplified way be illustrated by considering the relation for quasi-geostrophic PV (relation (4.3)) which is rewritten

$$\nabla^2\psi + f_0^2\frac{\partial}{\partial p}\left(\frac{1}{\sigma}\frac{\partial\psi}{\partial p}\right) = q_p - f \tag{4.9}$$

f may be considered as a reference PV field and $q_p - f$ a PV anomaly. Relation (4.9) provides the streamfunction associated with the PV anomaly, i.e. the departure from the average streamfunction.

If we assume, for simplicity, that σ is constant, and we further assume that the ψ field is expressed by a single component $\psi = \Psi(t)\sin kx \sin ly \sin mp$ (real fields may be expressed as a sequence of such terms; a Fourier sequence), we obtain by inserting in (4.9):

$$-K^2\psi \approx q_p - f \tag{4.10}$$

where $K^2 \equiv k^2 + l^2 + \frac{f_0^2}{\sigma}m^2$. Static stability ($\sigma$) is assumed positive, a necessary condition when using the quasi-geostrophic condition.

The assumption of constant σ is less reasonable since generally σ is changing in the troposphere. However, the result of this discussion would qualitatively be the same if σ were allowed to vary.

If q_p in an area is larger than the reference PV, f, the right hand side of (4.10) expresses a positive PV anomaly that yields a negative ψ or negative geopotential ϕ. Since these values are the departures from average, there is a lower geopotential associated with a positive PV anomaly than in the surroundings. Likewise we find a higher geopotential associated with a negative PV anomaly. Generally, inversion is carried out by solving Eqn (4.9) by standard numerical methods (relaxation). Usually there are several PV anomalies and inversion is done for each one separately. Since (4.3) and (4.9) are linear, adding the associated geopotential fields yields the total flow-field. This is an advantage of quasi-geostrophic PV. When the more general Eqn. (4.6) is used, the appropriate balance condition is Charney's nonlinear balance condition (Charney, 1955):

$$\nabla^2\varphi = \nabla\cdot(f\nabla\psi) + \left(\frac{\partial^2\psi\,\partial^2\psi}{\partial x^2\partial y^2} - \left(\frac{\partial^2\psi}{\partial x\partial y}\right)^2\right) \tag{4.11}$$

By omitting the non-linear term and letting f be constant, we get geostrophy as balance condition, as for the case above (relation 4.9). For a circular vortex relation (4.11) reduces to the gradient wind relation. Relations (4.6) and (4.11) and suitable boundary conditions provide the equations for PV inversion in the more general case. Unfortunately, the nonlinearity of (4.6) introduces an ambiguity in the solution. The flow fields obtained from inverting a number of PV anomalies do not generally add to give the total flow. However, mathematical methods have been developed that cope with this problem, see Davis and Emanuel (1991), and Davis (1992).

To illustrate the magnitude of PV and its distribution and variability in the troposphere and the stratosphere we consider a typical north–south cross-section across the northern part of continental Europe and Scandinavia, aligned along the 10° E meridian (Figure 4.22). The cross-section represents a 24 h simulation by the Norwegian HIRLAM. The tropopause is represented by the PV = 2 contour.

If we follow the 330 K isentrope, we identify a negative PV anomaly in the region 57° N to 52° N and consequently a high tropopause (at 200–250 hPa) is present. There is a high pressure area south of 57° N and this is associated with the negative upper PV anomaly. The positive PV anomaly between 57° N and 64° N is identified by the lowered tropopause, which makes a 'dip' in this region. As an example, the 310 K isentrope is intersecting the PV = 2 contour (tropopause) at 58° N and at 64° N, defining a positive PV anomaly on this

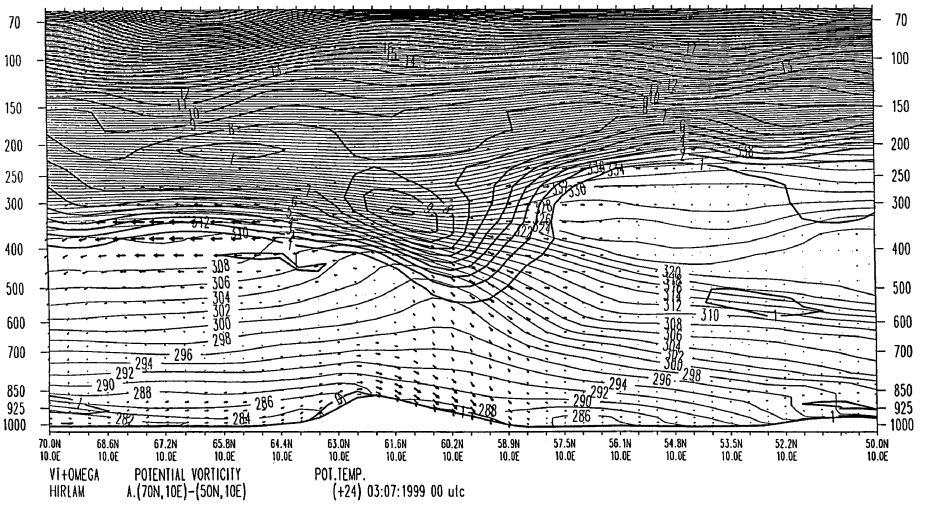


Figure 4.22. North–south cross-section along 10° E showing typical PV distribution in the northern part of continental Europe and Scandinavia (in PV units, thick lines) and potential temperature (in degrees K. thin lines). A negative (anticyclonic) PV anomaly is seen between 52° N and 57° N and a positive (cyclonic) PV anomaly between 58° N and 64° N. Arrows represent tangential winds and omega (vertical wind component in p-coordinates).

isentropic surface. The positive PV anomaly is associated with an upper tropospheric trough. The isentropes are seen to bow upwards below the positive PV anomaly and they are more widely separated as well, meaning that there is colder and less stable air in the region. Thus an upper trough is colder and contains less stable air than in the surroundings. Below the negative PV anomaly the isentropes are bowing downwards and they are generally squeezed more together (though less so above 400–500 hPa). The key point is that a deep tropospheric anticyclone is warmer and the air is more stable in the mid- and lower troposphere than in the surroundings.

These patterns are nearly identical with the idealized cross-sections presented by Thorpe (1985) and also discussed by Hoskins *et al.* (1985, their fig. 15). The results are obtained from a PV inversion considering a circular vortex and gradient wind used as balance condition. Fig. 15 in Hoskins *et al.* (1985) also presents the associated wind field, which is strongly cyclonic in the positive PV anomaly case, and strongly anticyclonic for a negative PV anomaly.

The degree of vertical penetration of the wind field associated with a PV anomaly is expressed by the following characteristic dimensions of a system:

$$H = \frac{fL}{N} \tag{4.12}$$

(see Section 4.5.7) where H is the Rossby penetration depth, which indicates to what extent the associated wind field of an upper PV anomaly is able to penetrate towards lower levels of the troposphere. The Rossby penetration depth depends on the horizontal dimension of the system, L, the planetary vorticity (Coriolis parameter) f, and the static stability expressed by the static stability parameter N, which is referred to as the Brunt–Väisälä frequency given by $N^2 \equiv \frac{g}{\theta}\frac{\mathrm{d}\theta}{\mathrm{d}z}$, where θ is potential temperature.

A low static stability N contributes to a large Rossby penetration depth- and so does a horizontal large scale L of an upper disturbance (e.g. upper positive PV anomaly).

Due to air–sea interaction, deep convective boundary layers may develop during Arctic outbreaks. In such conditions, N may become very small and H may be fairly large, even for a small-scale upper disturbance L. Thus cyclogenesis tends to take place on a small scale.

Figure 4.23, from Thorpe (1985) and Hoskins *et al* (1985), gives the wind and temperature fields associated with a warm (a) and cold (b) surface temperature anomaly. The insets in the right lower corner of the figures illustrate the distribution of isentropes along the Earth's surface at the boundary of the temperature anomaly.

Warm and cold surface anomalies can be regarded as surface cyclonic and anticyclonic PV anomalies respectively. The wind field associated with the warm (cold) anomaly is seen to be strongly cyclonic (anticyclonic), weakening with vertical distance. The static stability above the surface warm (cold) anomaly is smaller (higher) than in the surroundings.

4.4.3 The omega equation

We have seen that a cyclonic vortex is associated with and remains stationary relative to an upper positive PV anomaly. The intensity is unchanged in the absence of diabatic effects and friction. Referring to the thought-experiment described by Hoskins *et al.* (1985), we may consider a low-level wind field, such as a reverse shear flow, which is superposed below a positive PV anomaly (like the one presented in Figure 4.22) and its associated cyclonic vortex. The advection term in the vorticity equation must give a large contribution to the vorticity budget. However, the cyclonic vortex must stay in place relative to the upper positive PV anomaly, according to the invertibility principle. From the vorticity Eqn. (37) presented in Hoskins *et al.* (1985), we see that the advection term must be exactly cancelled by the terms on the right hand side. These terms contain the vertical velocity ω and the vortex stretching term is the most important one. Thus ascent upstream of the

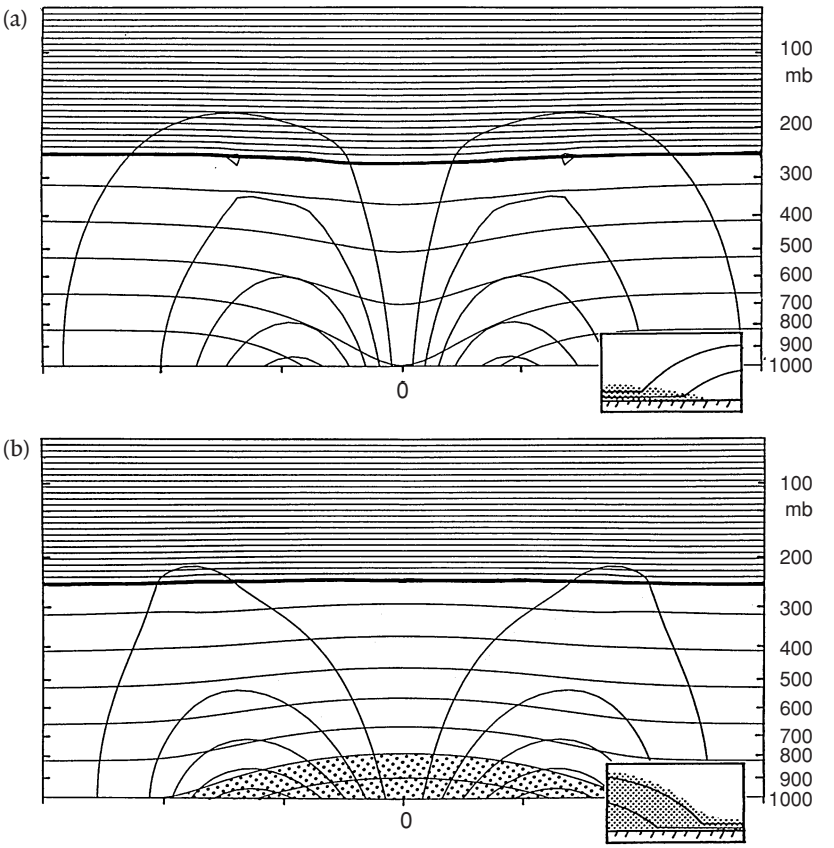


Figure 4.23. Circular symmetric flows induced by simple boundary temperature anomalies. Thick lines represent the tropopause and the two sets of thin lines, respectively, the isentropes for every 5 K and the transverse velocity for every 5 m s^{-1}. (a) A warm surface anomaly, interpreted as a cyclonic PV anomaly; (b) a cold surface anomaly, interpreted as an anticyclonic PV anomaly. The insets illustrate the distribution of isentropes along the Earth's surface at the boundary of the temperature anomaly (from Hoskins *et al.*, 1985).

upper PV anomaly and descent downstream must take place. By a change of coordinate system this thought-experiment may alternatively be described as a propagating upper positive PV anomaly. Ascent will take place ahead of the PV anomaly and descent at the rear. This is of course consistent with traditional quasi-geostrophic theory which gives ascent in response to upper positive vorticity advection. This process is described by the omega equation (Hoskins *et al.*, 1985):

$$\sigma \nabla^2 \omega + f^2 \frac{\partial^2 \omega}{\partial p^2} = f \frac{\partial}{\partial p}(\vec{v}_g \cdot \nabla q_p) \tag{4.13a}$$

where σ is the static stability parameter, as given in relation (4.3), and ω the vertical velocity $\omega \equiv \dfrac{\mathrm{d}p}{\mathrm{d}t} \cong -g\rho\dfrac{\mathrm{d}z}{\mathrm{d}t} = -g\rho w$ and q_p is quasi-geostrophic PV, expressed by (4.3).
Or, in the more conventional form[1]:

$$\left(\nabla^2 + \frac{f_0^2}{\sigma}\frac{\partial^2}{\partial p^2}\right)\omega = \frac{f_0}{\sigma}\frac{\partial}{\partial p}\left[\mathrm{V}_g \cdot \nabla\left(\frac{1}{f_0}\nabla^2\phi + f\right)\right] + \frac{1}{\sigma}\nabla^2\left[\mathrm{V}_g \cdot \nabla\left(-\frac{\partial\phi}{\partial p}\right)\right] \tag{4.13b}$$

An advantage of (4.13a), compared with the traditional formulation (4.13b), is that there is a single term on the right hand side of the equation making qualitative interpretation easier since the cancellation problem is avoided. In the quasi-geostrophic omega equation written in the traditional form (4.13b), one sees that a cancellation between the forcing terms on the right hand side is possible, e.g. when cold air advection and upper level vorticity advection take place in the same region. If we assume that ω is a smoothly periodically varying function, we can write $\omega = W \sin kx \sin ly \sin mp$.

Inserting this expression in (4.13a) yields:

$$-K^2\omega \approx f\frac{\partial}{\partial p}(\vec{v}_g \cdot \nabla q_p) \tag{4.14}$$

where $K^2 \equiv \sigma(k^2 + l^2) + f^2m^2$.

We now see from relation (4.14) that in, for example, regions where there is positive advection of potential vorticity increasing with height, there is ascent, while descent occurs in regions where negative advection of PV is increasing with height. In addition to the effects of PV advection, there is temperature advection in the boundary layer which also affects the vertical velocity field (Hoskins *et al.*, 1985). Thus possible cancellation effects must be considered when the total vertical velocity field is assessed.

4.4.4 Some applications of PV

In high latitudes conditions can frequently be described as a combination of those seen in Figures 4.22 (an upper-level positive PV anomaly) and 4.23b (a low-level negative PV anomaly). Snow-covered land or vast expanses of ice are strongly cooled and a low level inversion develops. The wind field at low levels is anticyclonic, reflecting the higher pressure observed over snow-covered land or extensive ice fields in winter.

[1] Note that for a qualitative estimation of the vertical velocity the left hand side of Eqn. (4.13b) can be set proportional to $-\omega$. See Holton, 1992, sect. 6.4.

Upper troughs sitting above or moving across these areas may thus be visualized by a combination of important features seen in Figures 4.22 and 4.23b: an upper positive PV anomaly, a low-level surface temperature anomaly (corresponding to a negative PV anomaly) and their associated temperature and wind field.

The cyclonic wind field associated with a positive PV anomaly, such as the PV anomaly seen to the left in Figure 4.22 is counteracted by the anticyclonic wind field, mainly at lower levels. Further, the high static stability in the lower troposphere prevents the wind field associated with the upper PV anomaly from reaching the ground. From Eqn. (4.12) we see directly that the large N (static stability) contributes to a small Rossby penetration depth. Thus an upper trough is present, but at low levels the anticyclonic conditions prevail and no cyclonic activity takes place.

As described earlier, conditions in the Arctic, particularly in the regions stretching from Greenland across Iceland to Svalbard and the Barents Sea, are dominated by large contrasts between cold snow-covered land and sea ice on one side and open sea with comparatively high sea surface temperature (SST) on the other. This is illustrated in Figure 2.13, which gives the locations of ice edge and SST in December 1982. Thus strong low-level thermal contrasts are present.

As an upper positive PV anomaly moves from over ice fields or snow-covered land out across expanses of open, relatively warm water, the tropospheric static stability may become very low, resulting in a large Rossby penetration depth. This allows efficient communications between the upper disturbance (e.g. PV anomaly) and low-level features, such as shallow fronts, forced by strong SST gradients near the ice edge, areas of convergence and pre-existing low-level troughs. With such conditions, cyclogenesis may take place.

If the static stability is very low, the characteristic horizontal dimension, L, of the system may be fairly small and still produce a large Rossby depth, H. Consequently, cyclogenesis tends to take place on a smaller scale. The planetary vorticity is large at high latitudes and this could have a positive effect on rapid cyclogenesis as well, contributing to a large H. Polar lows appear to develop in regions where the synoptic-scale flow is cyclonic, and with average relative vorticity comparable with the planetary vorticity or even larger. This gives a modified Rossby depth (Section 4.5.7):

$$H^2 = \frac{f(f + \zeta_{\mathrm{av}})L^2}{N^2} \tag{4.15}$$

where ζ_{av} denotes the average relative vorticity associated with the synoptic-scale flow. In some numerical studies this quantity has been found to be more

than 3*f*. Assuming a relative vorticity of 2*f*, we obtain a Rossby penetration depth nearly twice as large as the one expressed by relation (4.12).

As an example of using PV in explaining a polar low development, we will now describe a case study from Sunde *et al.* (1994). Figure 4.24a shows the synoptic situation at 1200 GMT 4 February 1992 while Figure 4.24b presents the 500 hPa contour and 1000–500 hPa thickness 12 h prognosis for 0000 GMT 5 February. Frontal systems associated with the polar front are seen over the southwestern and eastern parts of the surface charts. The secondary, but fairly strong, frontal zone in the Norwegian Sea and around Bear Island is mainly

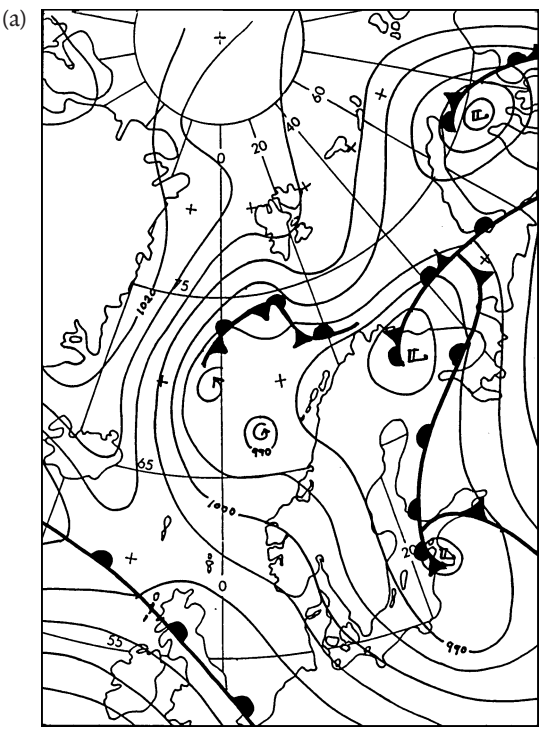


Figure 4.24. (a) Synoptic analysis for 1200 GMT 4 February 1992. (b) 12 h prognosis for 0000 GMT 5 February showing the 500 hPa contours (solid lines) and 1000–500 hPa thickness (broken lines). D indicates the position of the cross-section referred to in the text. (c) 6 h prognosis for 1800 GMT 4 February for cross-section D showing potential vorticity (thick lines, PV units) and potential temperature (thin lines, K). (d) 24 h prognosis for 1200 GMT 5 February for cross-section D showing potential vorticity and potential temperature. (e) 15 h prognosis for 0300 GMT 5 February showing wind and potential vorticity on the 285 K isentropic level. B indicates location of Bear Island. (f) 24 h prognosis for 1200 GMT 5 February for cross-section D with arrows indicating the velocity and thin solid lines the potential temperature. The other cross-section lines A, B, C are described in Sunde *et al.*, 1994, from which these figures are taken.

(b)

(c)

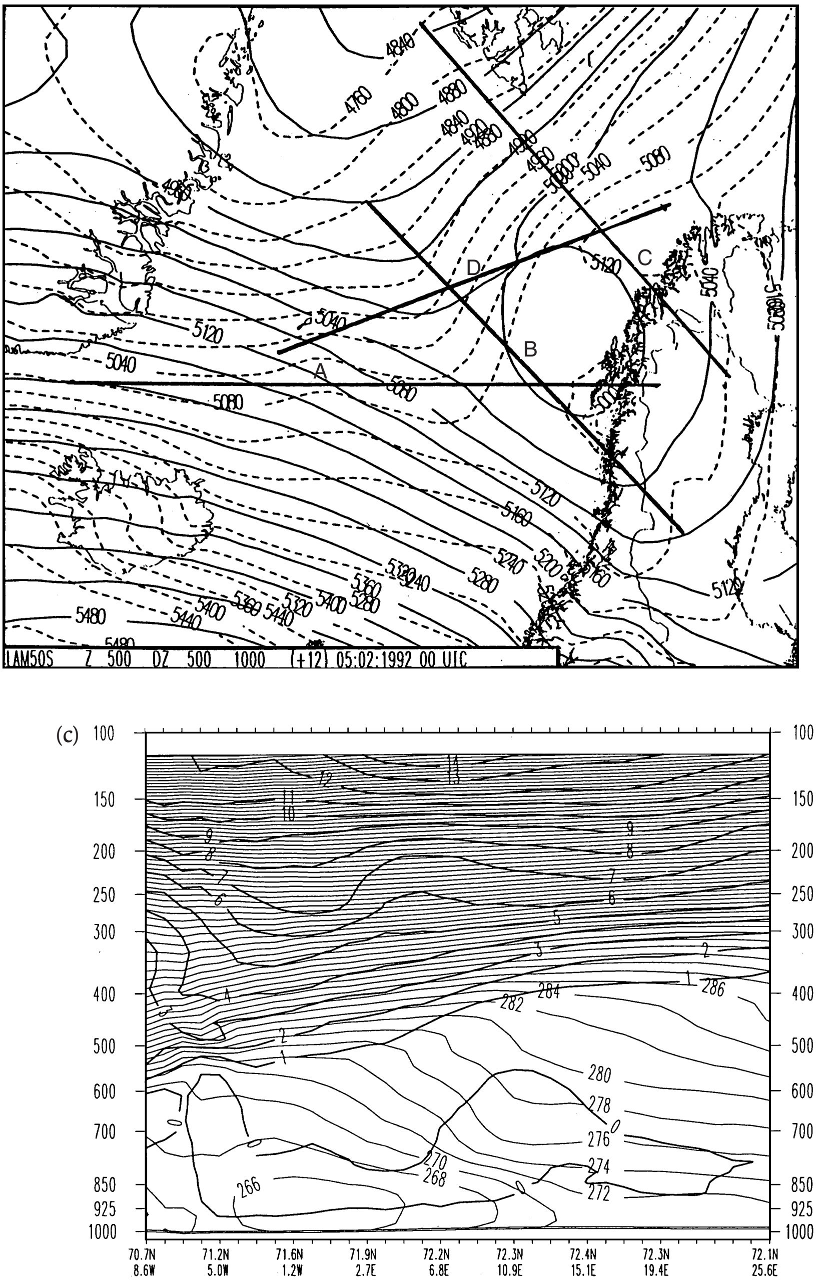


Figure 4.24 (*cont.*).

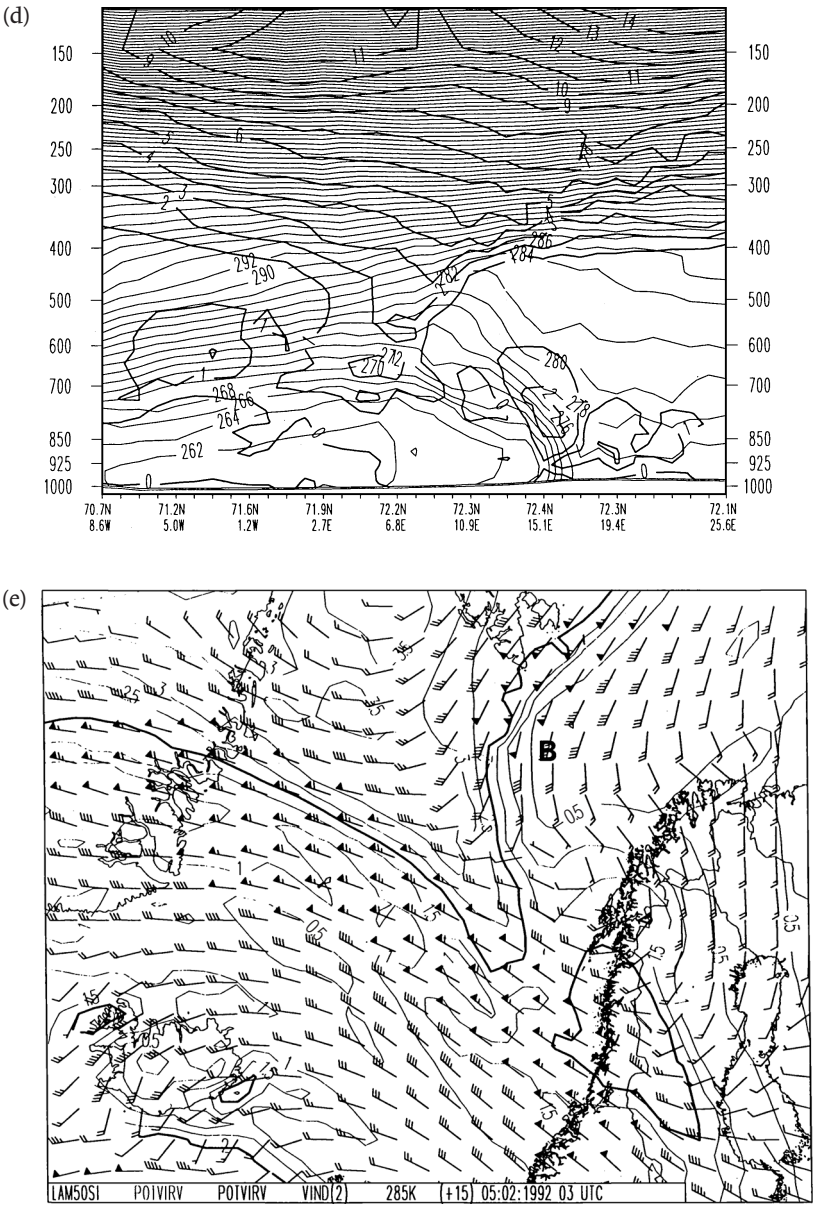


Figure 4.24 (cont.).

confined to the region below 700 hPa, as shown in Figures 4.24c and d. The cross-sections, with location indicated by the bar D in Figure 4.24b, show potential vorticity (in PV units) and potential temperature contours (in degrees K). Figure 4.24e (15 h prognosis) presents wind and potential vorticity on the 285 K isentropic level at 0300 GMT 5 February. The PVU = 2 contour is

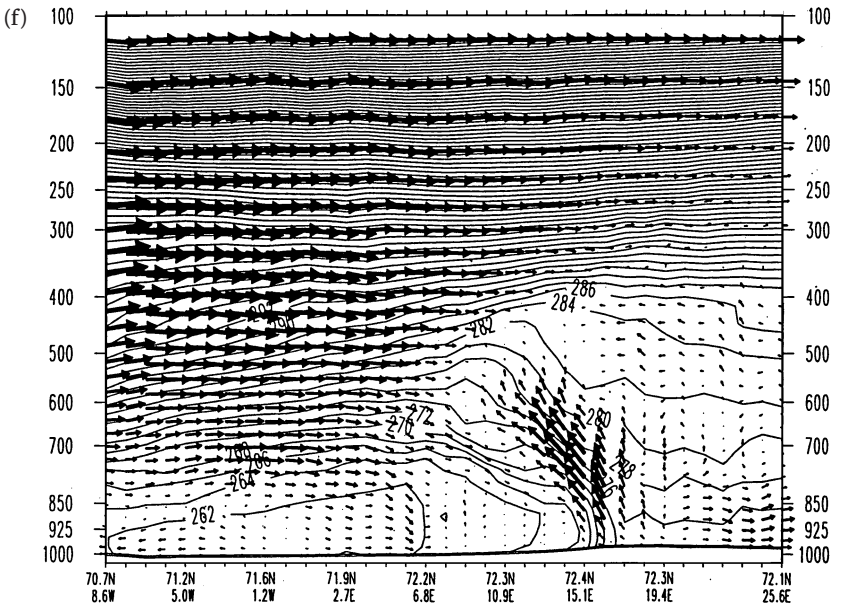


Figure 4.24 (*cont.*).

emphasized. Figure 4.24c shows the positive PV anomaly over the western part of the Norwegian Sea, the rising isentropes indicate colder air and their wider vertical separation means that the static stability was smaller than further east.

There was pronounced upper PV advection approaching the Bear Island region, as seen in Figure 4.24e (Bear Island is indicated by B). Referring to the omega equation (Eqn. 4.13), ascent must take place in the region of upper positive PV advection (PV advection at lower levels is zero or very small), which is readily verified in Figure 4.24f. This is a 24 h prognosis for 1200 GMT 5 February showing a cross-section along line D in Figure 4.24b. Tangential winds, vertical velocity and potential temperature are shown. The cross-section illustrates the classical picture of strong ascent of warm air in the frontal zone. There are also indications of descent in the cold air and upper part of the front, at 700 hPa. In Figure 4.24d, the positive PV anomaly had become more pronounced, and the descending air in the upper part of the front was advecting PV along isentropes, a process referred to as tropopause folding. Thus upper PV was brought further down towards the low-level frontal zone and incipient polar low at Bear Island.

The strong ascent seen in Figure 4.24f, resulting from upper positive PV advection, created release of latent heat and thus increase of PV below the diabatic heating maximum. The vorticity increased in the lower part of the frontal zone,

exceeding 4×10^{-4} s^{-1} (not shown). The low-level PV anomaly, exceeding 2 PV units is seen in Figure 4.24d, at 800–850 hPa levels. This PV anomaly was due to latent heating and was not of stratospheric origin. Since a cyclonic wind field is associated with a positive PV anomaly, the low-level PV anomaly must have been an important contribution to the polar low cyclogenesis. Figures 4.25a and b present satellite images at 1225 GMT 4 February and 0359 GMT 5 February, respectively. They show development of an organized cloud system, with embedded convective clouds at Bear Island as the polar low was developing. The polar low was forced by upper, positive PV advection and, as is usual in the case of many polar lows, with a strong contribution from convection.

4.5 The role of thermal instability in polar low formation and maintenance

4.5.1 Introduction

Since the early investigations of polar lows, it has been recognized that convection plays a significant role in the development of these systems. Examples of this are numerous. Dannevig (1954) linked the formation of 'instability lows' to an organized release of thermal instability and Businger and Reed (1989b) in their definition of a polar low specifically pointed out that the main cloud masses of these systems were 'largely of convective origin'. Also, numerous satellite images of polar lows, of which a large number are shown in Chapter 3, document the occurence of convective clouds in most significant polar low developments.

The precise way in which convection affects the development of a polar low has been disputed over the years. Prior to Dannevig's hyphothesis, some meteorologists believed that a polar low could be considered as one huge cumulonimbus cloud. On the other hand, Harrold and Browning (1969), and a number of other British authors, more or less discarded the role of convection for polar low developments. Rasmussen (1977, 1979) and Økland (1977) revived the idea of the importance of deep convection for polar low developments applying the CISK[2] theory put forward simultaneously by Charney and Eliassen (1964) and Ooyama (1964) to explain the growth of hurricane depressions. According to this theory, a hurricane may intensify through a cooperative interaction between convection and a large-scale, balanced system (for a detailed discussion see the following sections). According to the studies of Rasmussen and Økland, this process might work even over polar/Arctic seas, albeit the

[2] For a discussion of the basic ideas regarding CISK and WISHE, see Section 4.6.1.

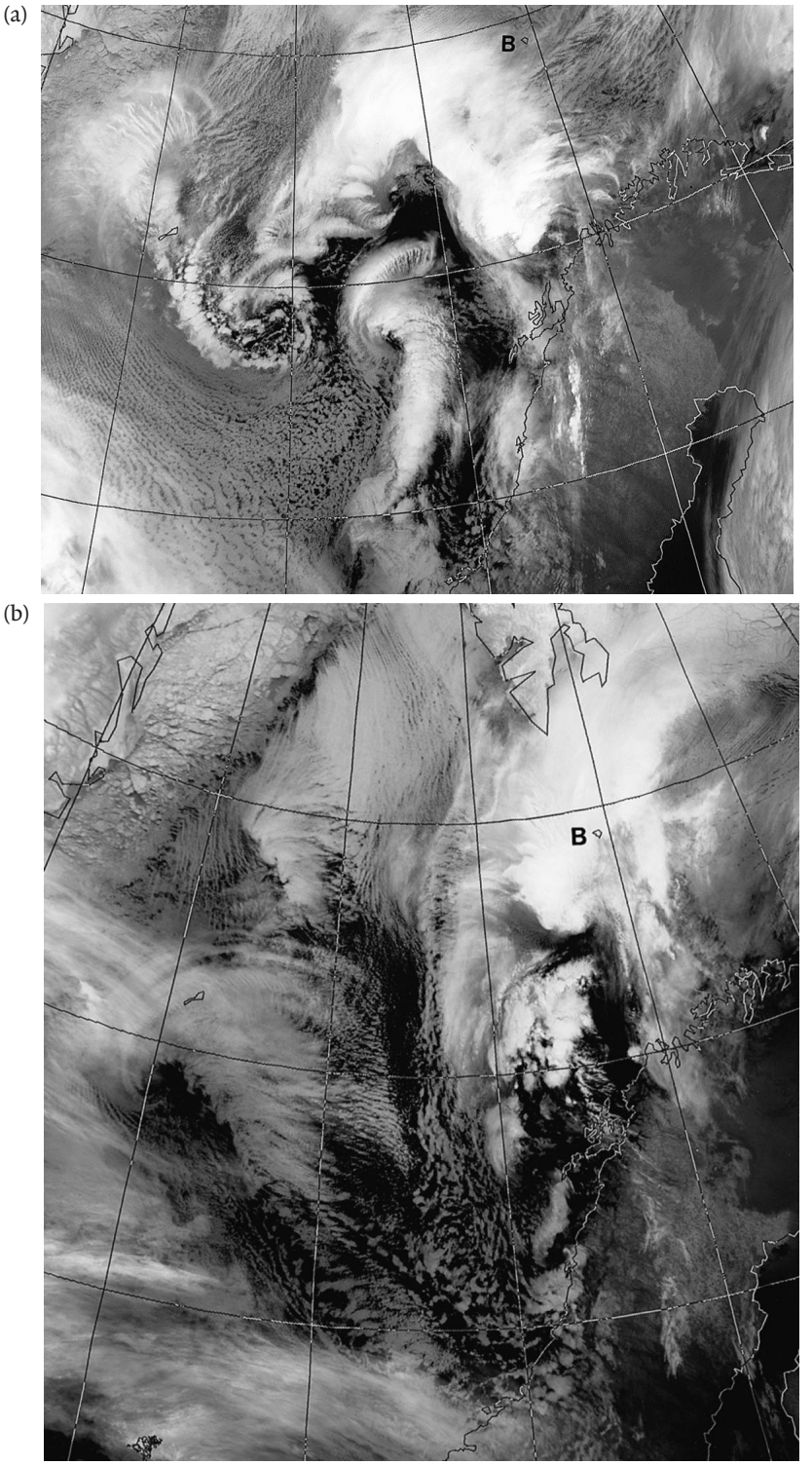


Figure 4.25. Infra-red satellite images for (a) 1224 GMT 4 February 1992 and (b) 0359 GMT 5 February 1992. B indicates location of Bear Island. (Image courtesy of the NERC Satellite Receiving Station, University of Dundee.)

sea surface temperatures in these region are far below the threshold value for the formation of tropical cyclones of around 26 °C.

Following the early studies by Rasmussen and Økland, it was gradually accepted that CISK might be the main dynamic mechanism for a significant group of polar lows, assuming that a reservoir of convective available potential energy (CAPE) was available. However, from the mid-1980s the idea of CISK and its role in the development of tropical cyclones (and polar lows as well) was challenged in a number of papers (Emanuel, 1986a; Emanuel and Rotunno, 1989). These authors argued that the tropical atmosphere was nearly neutral to deep moist convection, and that the reservoir of available potential energy assumed by Charney, Eliasen and Ooyama in their original concept of CISK, apparently did not exist in the tropical atmosphere. According to Emanuel (1986a), tropical cyclones can intensify and be maintained through air–sea interaction instability (ASII, later denoted WISHE (wind induced surface heat exchange); Emanuel, 1986a) *without* ambient conditional instablity (CAPE) providing a starting disturbance of sufficient amplitude exists. In their 1989 paper, Emanuel and Rotunno extended this point of view to polar lows, claiming that 'it seems likely that the convection observed in the environment of polar lows similarly serves to maintain a nearly moist adiabatic lapse rate, with no substantial stored convective available potentiel energy'.

While results from Betts (1982) indicated that the tropical atmosphere was very nearly neutral to deep convection when viewed in a proper thermodynamical framework, the situation may be different in polar/Arctic regions where low-level heating, due to strong air–sea interaction, occurring simultaneously with upper-air cold advection may lead to the formation of significant amounts of CAPE. Rasmussen (1979), in a study of a polar low that formed near Iceland and subsequently moved south over a warmer sea surface, argued that air particles ascending pseudo-adiabatically would achieve a temperature excess relative to their environment of up to 6 °C, corresponding to significant CAPE. In the same paper, another strong polar low development close to the Norwegian coast on 12 October 1971 was studied (the surface chart showing the polar low on 13 October is shown as Figure 3.1). This development took place close to the coast and relatively far south, near 65° N, over an exceptionally warm sea, with sea surface temperatures around 10 °C, as a major upper-level cold trough with 500 hPa temperatures as low as −44 °C approached the coastal region. A radiosonde ascent from the weather ship *Polar Front* (66° N, 2° E) from 0000 GMT 13 October and upstream, but close to the place of formation, is shown on Figure 4.26. Coastal measurements representative of offshore conditions close to the coast where the development took place, showed surface temperatures/dew point temperatures of 8 °C and 0 °C respectively. Air parcels

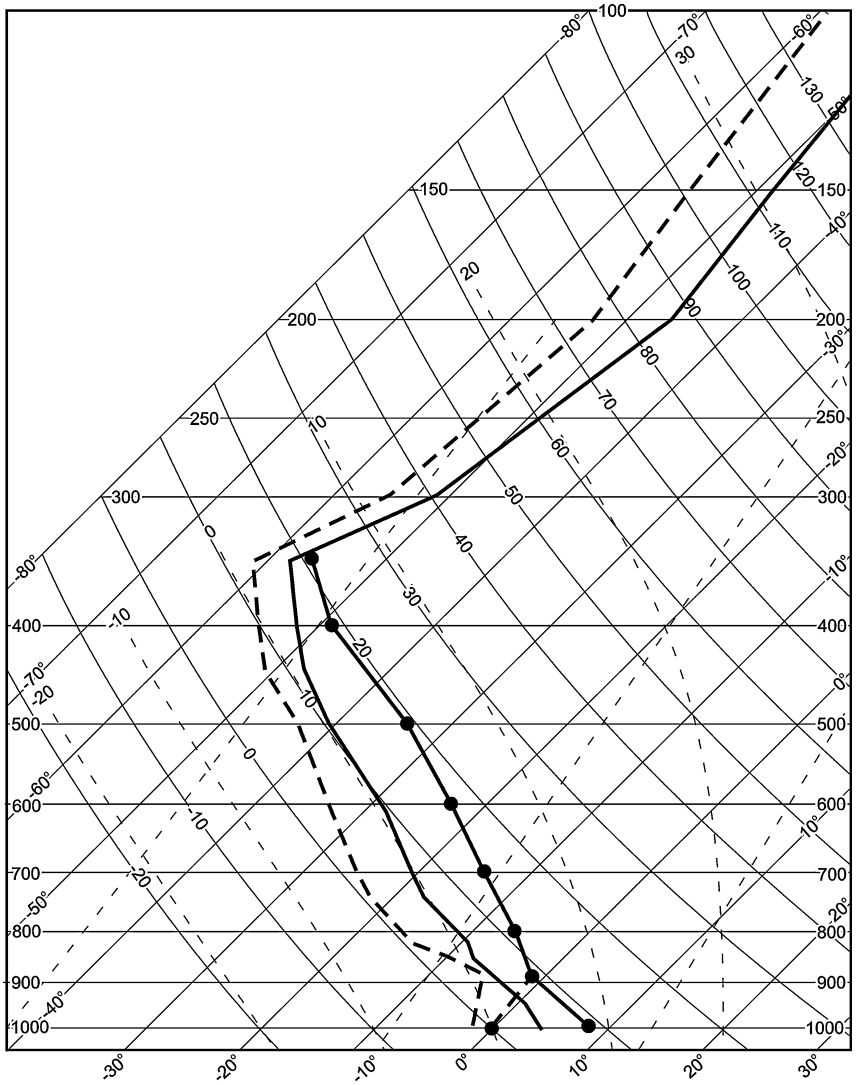


Figure 4.26. Radiosonde ascent from weather ship *Polar Front* (66° N, 2° E) at 0000 GMT 13 October 1971 showing temperatures (solid line) and dewpoint temperatures (broken lines). The solid line with dots shows the temperatures of an air parcel ascending pseudo-adiabatically from the surface with temperature and dewpoint at the start of the ascent as observed in the onshore flow at coastal stations.

ascending from the surface with these values would experience a significant temperature excess relative to the environment, corresponding to a CAPE value of around 1100 J kg^{-1} (a CAPE value of 1100 J kg^{-1} corresponds to a maximum vertical velocity of a pseudo-adiabatically ascending air parcel of around

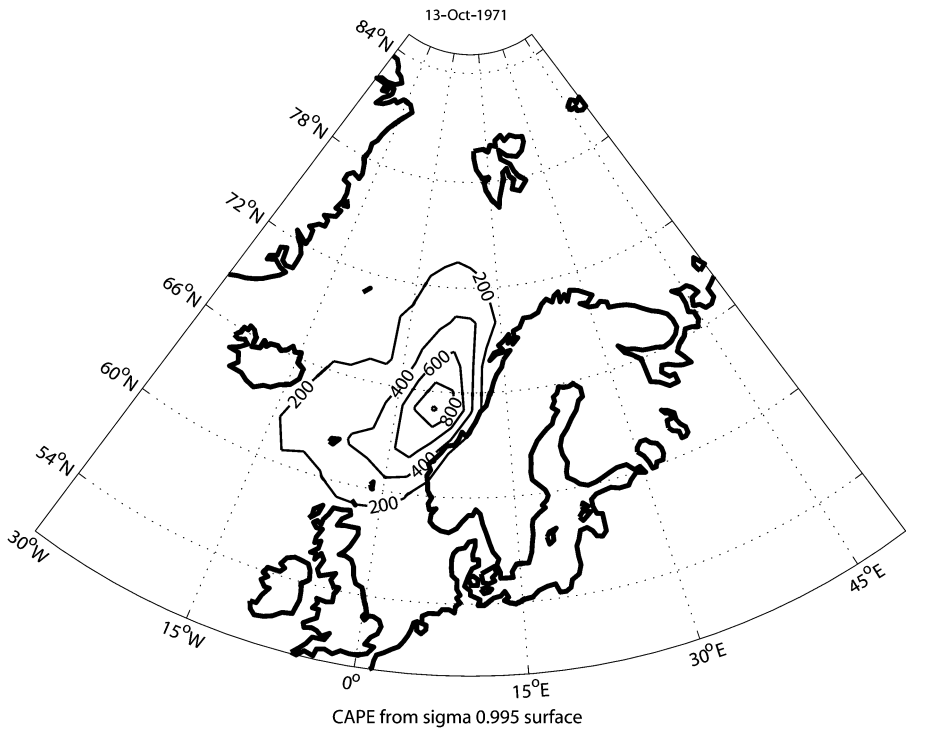


Figure 4.27. CAPE field for 0000 GMT 13 October 1971 (units J kg^{-1}, contour intervals 200 J kg^{-1}, maximum value 1000 J kg^{-1}).

47 m s^{-1}). Assuming that the neutral (dry adiabatic) low-level layer was well mixed (as actually shown by the radiosonde ascent) *all* the parcels within this layer would be approximately equally positively buoyant and the sounding as such 'unambiguously conditional unstable' as defined by Emanuel *et al.* (1994).

The extent of the region associated with high values of CAPE is illustrated by Figure 4.27 showing the CAPE field at 0000 GMT 13 October 1971 (J. Rytter, personal communication).

The CAPE values shown on Figure 4.27 were computed from the NCEP/NCAR re-analysis data set. The CAPE distribution was computed for a number of polar low developments assuming that the particles ascended undiluted from a level near the surface (sigma level 0.995), initially dry adiabatically, and later, after condensation, moist adiabatically carrying their condensation products along.

The CAPE values associated with the 13 October 1971 development are probably the highest documented for a polar low so far and the development as such should not be considered as typical. Wilhelmsen (1986a), on the other

hand, considered 38 cases of gale-producing polar lows and found a conditionally unstable lapse rate between the surface and the 500 hPa level for *all* cases, which indicate that CAPE may be significant for other developments.

In addition to the October 1971 development, the CAPE fields for several other significant polar lows were studied. According to Rytter, the CAPE fields were highly sensitive to the initial conditions, the values being strongly dependent on the particular level from which the particles started their ascent. With this uncertainty in mind, the data indicated that a number of significant polar low developments seem to have been associated with moderate amounts of CAPE, a moderate amount here being defined as *c*.400–600 J kg^{-1} corresponding to a maximum vertical velocity within the convective clouds of around 30–35 m s^{-1}. Occasionally, however, larger CAPE values were found.

The study indicated that the polar lows did not necessarily form within *pre-existing* reservoirs of CAPE but that the reservoirs and the polar lows developed more or less simultaneously as the lows moved over a warm sea surface.

The polar low near Bear Island on 12–14 December 1982 has been discussed in a number of papers (see Section 3.1.5), the development of the system being described as a two-step process; an initial phase followed by a convective stage during which convection was assumed to be important either in connection with CISK or with WISHE.

At the time when the impressive cloud spiral seen on Figure 3.29 had formed in the morning of 13 December only small CAPE values could be detected in the region. Later in the day at 1800 GMT when the cloud spiral had moved south over a warmer sea surface, and just prior to the second phase of development of the polar low, an intensification of the CAPE was indicated by the NCEP/NCAR re-analysis data, resulting in the formation of a rather strong maximum, north to northwest of North Cape (Figure 4.28). As the polar low intensified, some distance east of the centre of the cloud spiral but close to the region of maximum CAPE, the magnitude of the CAPE further increased to around 800 J kg^{-1}. Judging from the data presented by Rytter (EGS Polar Lows Working Group Meeting, Paris, October 2001), the large increase in CAPE was primarily due to the formation of a rather deep, low-level, neutrally (dry adiabatic) stratified layer which formed due to strong sensible heat fluxes from the warm sea surface.

Some preliminary, general conclusions can be drawn from the examples presented above and other cases investigated by Rytter. First, the combination of differential temperature advection and low-level heating may lead to the generation of, and to significant variations in the amount of, CAPE in polar/Arctic regions.

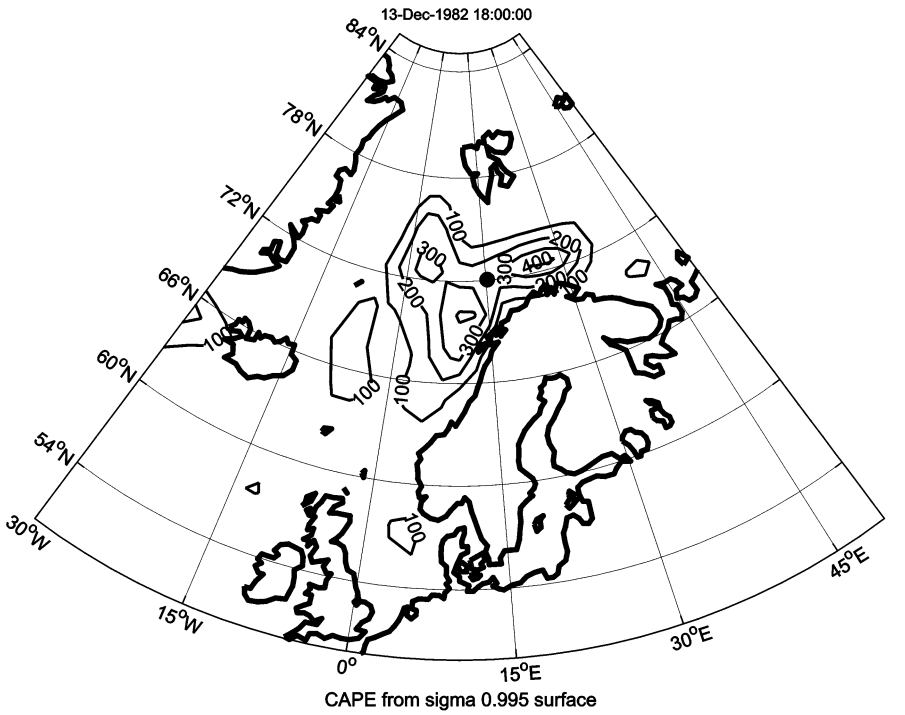


Figure 4.28. CAPE field for 1800 GMT 13 December 1982. The black dot indicates the position of the parent polar low at 1800 GMT. An intense vortex (a new polar low) developed a few hours later east of the position of the parent circulation, close to the centre of maximum CAPE.

Second, the CAPE fields over the Norwegian and Barents Seas seem highly structured, showing well-defined maxima and minima, and CAPE is *not* consumed as quickly as it is generated by large-scale processes. The maxima are well correlated with the positions of the polar lows showing good continuity in space and time.

The amount of CAPE is highly variable during the development of a polar low, being typically rather small during the initial stage, but growing as the lows mature. The reasons for this are not clear, but several factors may be significant, including that most polar lows form initially at high latitudes with relatively low sea surface temperatures. As they move south the sea surface temperature and the surface fluxes will increase significantly modifying mainly the layer adjacent to the surface, leaving the layers above relatively unaffected. A modification of this type will generally lead to an increase of CAPE.

Apart from the question of the mere existence of CAPE, the extent to which it contributes to polar low developments is still an open question as will be apparent from the discussion in the following sections.

4.5.2 Adjustment to heating and cyclone intensification

The supply of water vapour to the atmosphere from the ocean and the heating of the atmosphere when water vapour condenses or freezes in clouds are crucial requirements for the growth and maintenance of polar lows. Sections 4.5.2 to 4.5.13 aim to provide physical insight into the relation between heating and vortex intensification. Employing a variety of model problems, we investigate in detail how heating affects the pressure and potential vorticity distributions in the atmosphere and how this, in turn, affects the air motion. The frequent reference to tropical cyclones, although seemingly inappropriate in a book devoted to polar lows, is motivated by two facts: (some) polar lows and tropical cyclones bear a strong resemblance to each other and most theories discussed in this section were first constructed to explain tropical cyclone growth, but were later adopted by various authors to also explain a significant part of the intensification of polar lows.

Sawyer (1947) summarized the accepted (pre-1947) view of the mechanism for the growth of cyclonic storms in the tropics as follows (Figure 4.29):

> Most meteorologists who have studied the mechanics of the tropical cyclone agree that its formation is connected with the vertical instability in the atmosphere, and that the energy of the cyclone is derived from the latent heat of condensation released during the ascent of warm moist air from the surface. It is explained that the formation of a warm column of air by convection causes the isobaric surfaces at the top of the column to be raised, and an outflow of air at this level results from the rise in pressure. The outflow causes a reduction in total weight of the column

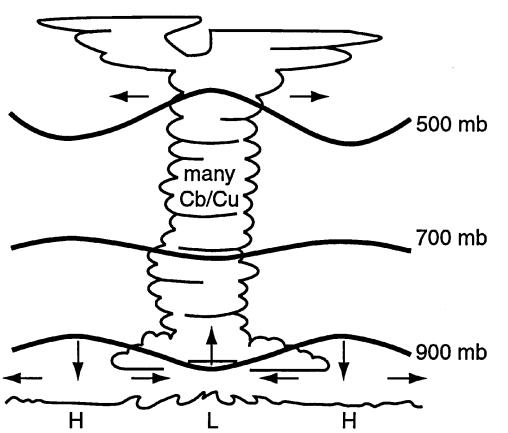


Figure 4.29. Schematic vertical section through a polar low. The arrows indicate the flow direction. The thick solid line indicates the position of a specific pressure level (vertical scale in this case is exaggerated) (based on fig. 8 of Rasmussen, 1979).

> and produces a fall of pressure; the surface winds then converge towards the newly-formed 'low', and produce a cyclonic circulation as a result of the Coriolis effect.

This represents a short summary of the so-called 'thermal theory of cyclones', which had been in existence since the nineteenth century as a theory for the growth of cyclones (Austin, 1951). In more modern terms, Sawyer is in fact referring to the process of hydrostatic adjustment (Bannon, 1995, 1996; van Delden, 2000), and in his final sentence he is referring to the process of geostrophic adjustment (Gill, 1982). Both these processes are more complex than is implied by Sawyer's summary. They involve sound waves, gravity-inertia waves and, usually, convection.

In the 1950s researchers who worked on the thermal theory were frustrated by their inability to make a clear conceptual distinction between buoyant cumulus convection connected to the vertical instability in the atmosphere and the process of adjustment to balance (Bergeron, 1954). The theory of convection resulting from hydrostatic instability, which is based on the Boussinesq approximation, cannot account for the pressure decrease at the Earth's surface.

After preliminary work by Kleinschmidt (1951) and Eliassen (1952), the paper by Charney and Eliassen (1964) marked the beginning of a new approach to the problem. Charney and Eliassen made the following three important simplifying assumptions.

1. The diameter of a tropical cyclone is small compared with its distance from the Equator with the result that the Coriolis parameter can be regarded as independent of the geographical latitude.
2. The ideal tropical cyclone is rotationally symmetrical.
3. Above the friction layer near the ground the hydrostatic and gradient wind balance is achieved to a great degree.

The first assumption is very reasonable and should not lead to much controversy, but the other two assumptions are far from trivial. Although the theory behind the justification of the second assumption is interesting (see e.g. Melander *et al.*, 1987) and is, in fact, still the subject of intensive research, here we will be concerned chiefly with the theoretical justification of the third assumption and its important consequences. Charney, in his conversation with George W. Platzman (see Lindzen *et al.*, 1990, pp. 69–70), states:

> … I became interested in the mechanism of generation of hurricanes and this led to the notion of CISK (Conditional Instability of the Second Kind) … I worked on the problem and one of the things that led me to the formulation of CISK was the idea that the forces in a hurricane must

> be in essential balance, that you were dealing with a balanced flow, not an inertial gravity oscillation … [Earlier authors] had dealt with hurricane motions as a sort of gigantic convection cell. I knew that couldn't be correct, but I was still puzzled by the existence of conditional instability. But it was really Ooyama who pointed out that, despite the fact that the individual cumulus cells were conditionally unstable, that the hurricane as a whole was stable.

Charney and Eliassen (1964), Ogura (1964), Kuo (1965) and Ooyama (1969) constructed models of a tropical cyclone in which the motion was assumed to be in thermal wind balance (i.e. hydrostatic balance and gradient wind balance) at all times. It was hypothesized that the growth of a tropical cyclone is essentially the result of a process of continuous adjustment to thermal wind balance in the presence of processes (such as heating) disturbing this state of balance.

Incorporating (parameterizing) the heating associated with latent heat release in clouds and sensible heat fluxes at the Earth's surface in the balanced cyclone model poses great theoretical problems. Many authors have hypothesized that widespread and prolonged convective precipitation only exists in connection with 'large-scale' lifting of air (the term 'large-scale' is placed within quotation marks because it is not well defined; see the caption of Figure 4.30). One important cause of 'large-scale' lifting in a cyclonic vortex is frictional convergence in the boundary layer, the intensity of which can be shown to be proportional to the vorticity of the balanced flow just above the frictional boundary layer (Holton, 1992). This offers the possibility to theoretically link the heating to the intensity (i.e. the vorticity) of the vortex, yielding a feedback loop between heating and the balanced flow (the vortex). Figure 4.30 visualizes this feedback loop. A closed loop must include the segments shown by the heavy curves. The upper half is referred to as 'control' and the lower half is referred to as 'feedback'. Charney and Eliassen (1964) showed that this loop

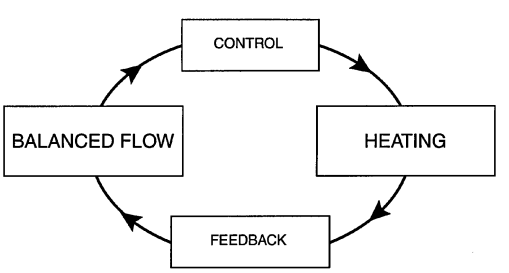


Figure 4.30. Schematic figure showing the interaction between balanced motion and heating (inspired by fig. 1.1 of Arakawa, 2000). Arakawa used the term 'large-scale processes' instead of 'balanced flow' and 'cumulus convection' instead of 'heating'.

can exhibit unlimited growth, and used this fact to explain the intensification of the balanced flow. The instability resulting from this positive feedback was called 'Conditional Instability of the Second Kind'. CISK was also proposed as an explanation of the growth of polar lows (Rasmussen, 1977, 1979; Økland, 1977) and later employed to explain the growth of a certain type of Mediterranean cyclone occurring over the warm waters in the autumn and early winter (Rasmussen and Zick, 1987).

In the formulation of CISK due to Charney and Eliassen (1964), the 'control' (Figure 4.30) is frictional convergence. In more specific words, the intensity of the heating is controlled by (i.e. is proportional to) the vertical motion at the top of the turbulent boundary layer. Ooyama (1969) assumed that the 'constant' of proportionality was a function of the availability of moisture in the boundary layer. Obviously the amount of moisture in the boundary layer depends on the intensity of flux of moisture from the ocean. If this flux is large enough, the reservoir of CAPE can be replenished and continuously released or 'activated' due to frictionally induced upward motion.

Emanuel (1986a) has called into question this particular form of the 'control', which is referred to by Mapes (1997) as a form of 'activation control'. Emanuel argued against the existence of a reservoir of CAPE, pointing out that CAPE is destroyed by convection as quickly as it is created by surface fluxes. This implies that the lapse rate in the convecting part of the atmosphere is constrained to follow the moist adiabatic lapse rate. Simultaneously, the surface fluxes are enhanced by the high wind speeds within a polar low or a tropical cyclone. Therefore, as the vortex intensifies, so does the transfer of heat from the ocean to the atmosphere. The line in a thermodynamic diagram, representing the thermodynamic state of the atmosphere in a tropical cyclone or polar low, would then appear to shift slowly towards the right (increasing temperature) and slightly upward (decreasing pressure). This is visualized in Figure 4.31.

According to Emanuel, this 'convective adjustment' constitutes the 'control' in Figure 4.30. Mapes (1997) referred to this type of control as 'equilibrium control'. Emanuel's theory is now known as 'Wind Induced Sensible Heat Exchange' or 'WISHE' (earlier named 'air–sea interaction instability' (ASII)). Emanuel and Rotunno (1989) investigated how far the theory of WISHE (ASII) applies to polar lows.

Both theories (CISK and WISHE) hypothesize that the vortex evolves through a succession of balanced states. With this hypothesis the 'balanced' dynamics of the cyclone is separated from the much more complicated 'unbalanced' dynamics of cumulus convection. The numerical simulations due to Ooyama (1969) and Sundquist (1970) demonstrated that the structure and growth of a tropical cyclone can be explained by considering explicitly only the

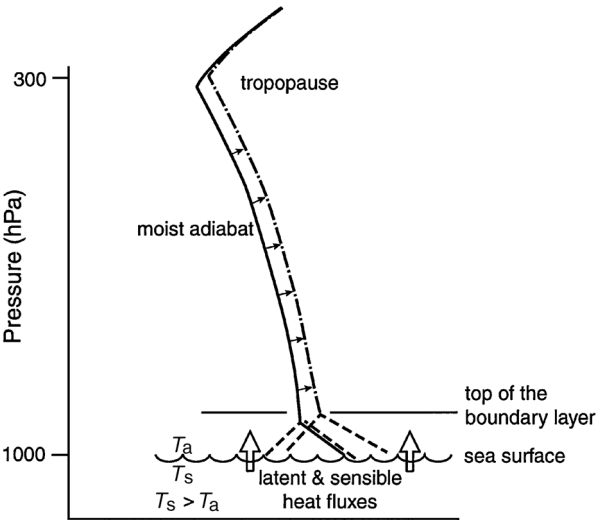


Figure 4.31. Skew-T log p thermodynamic diagram showing schematically two consecutive hypothetical thermodynamic states within a polar low or tropical cyclone, according to the equilibrium control theory. The basic assumption is that the lapse rate above cloud base must always adjust to the saturated adiabat, which is a function of temperature and pressure. Due to heating the temperature profile is shifted slowly towards the right.

balanced part of the motion. The reasons for this are quite subtle and merit a discussion, which will be provided in the following sections.

The question concerning us first is, how does heating affect the balanced flow (the so-called *feedback* in Figure 4.30)? In order to gain a better understanding of this feedback, we must turn to the basic theory of adjustment to balance (hydrostatic balance, geostrophic balance and thermal wind balance).

4.5.3 Hydrostatic adjustment

The first sentences of the introduction to a nearly 50-year-old paper by Scorer (1952) serve to clearly state the problem at hand: 'It is not obvious what will be the consequences of heating the lower layers of the atmosphere over a large region. What will happen to the air over the heated layers? At what levels will pressure gradients be generated and what air motion will ensue? Authors are not unanimous in their answers.'

It is the purpose of this section to shed some light on the answers to these questions. Our starting point is the equation of continuity for an ideal gas, which can be written as follows (Van Delden, 1992, or Durran, 1999):

$$\frac{\mathrm{d}\Pi}{\mathrm{d}t} = -\frac{R\Pi}{c_v}\vec{\nabla}\cdot\vec{v} + \frac{RJ}{c_v\theta} \tag{4.16}$$

where t is time, $\vec{v}$ is the air velocity, J is the heating per unit mass, per unit time, θ is the potential temperature (defined as $\theta = T(p_{\text{ref}}/p)^{\kappa}$, where T is the temperature, p is the pressure, p_{ref} is a constant reference pressure, R is the specific gas constant of air, $\kappa = R/c_p$ with c_p the specific heat at constant pressure), c_v is the specific heat at constant volume, $\mathrm{d}/\mathrm{d}t = (\partial/\partial t + \vec{v} \cdot \vec{\nabla})$ and Π is the Exner function, defined as

$$\Pi \equiv c_p \left(\frac{p}{p_{\text{ref}}} \right)^{\kappa} \tag{4.17}$$

Eqn. (4.16) demonstrates that the pressure inside an air parcel changes due to divergence (first term on the r.h.s.) and due to heating (second term on the r.h.s.).

When an air mass is heated at constant volume, potential energy is introduced in the form of available elastic potential energy (Bannon, 1995). A pressure gradient is set up at the edge of the air mass. The force resulting from this pressure gradient represents the germinal force for the excitation of a wave of expansion, i.e. a thermally forced sound wave. As a result of this force the air mass expands (i.e. $\vec{\nabla} \cdot \vec{v} > 0$) and, in turn, compresses the immediate environment. In this manner the positive pressure perturbation propagates outwards with the phase speed of sound waves (about 300 m s^{-1}). Due to the decrease in its density, the air mass acquires positive buoyancy, which ultimately tends to come in balance (if the environment is stably stratified) with a negative *perturbation* pressure gradient force such that

$$\frac{\partial p'}{\partial z} = -\rho g \tag{4.18}$$

where p' represents the pressure perturbation relative to the hydrostatically balanced state previous to the heating and g is the acceleration due to gravity. These are the essential ingredients of the process of hydrostatic adjustment.

In order to obtain some insight into the process of hydrostatic adjustment to heating we will review some important results of theoretical investigations into the prototype problem of hydrostatic adjustment due to Bannon (1995, 1996) (Figure 4.32). This problem consists of the response of a stably stratified atmosphere to a vertically confined but horizontally uniform instantaneous heating. The problem was originally advanced by Lamb (1908) (see also Lamb, 1932, sections 309–311) and has therefore been termed 'Lamb's problem' by Bannon. However, Lamb did not actually solve the problem of adjustment to hydrostatic balance, but only investigated the properties of the waves excited by a point source of hydrostatic imbalance.

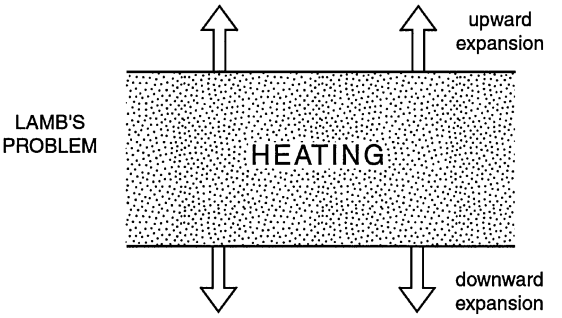


Figure 4.32. Lamb's problem. Horizontally homogeneous heating is applied to a layer of air, as a consequence of which it expands vertically.

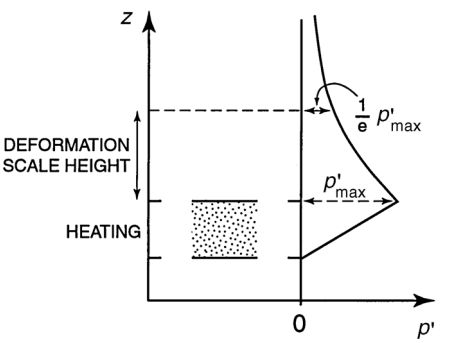


Figure 4.33. Schematic figure showing the pressure perturbation, induced by heating a vertically confined layer of air, as a function of height after adjustment to hydrostatic balance.

A first interesting result, due to Bannon (1996), states that if the heating is horizontally uniform, layers of air *below* the heated layer are not displaced in the final equilibrium state, relative to the initial state. The layers above the heated layer are lifted *uniformly* upward (or downward if there is cooling) with no change in their state variables. At a fixed point above the heated layer, the density increases and the potential temperature decreases adiabatically to produce a continuous pressure field in hydrostatic balance. A heat source in the lower troposphere affects the pressure distribution throughout the entire atmosphere aloft. Of course, the question is how strong this effect is at upper levels.

Inspired by Gill's (1982) linear technique of tackling the geostrophic adjustment problem, Bannon (1995) derived an equation for the pressure as a function of height in the final hydrostatic equilibrium state in an isothermal atmosphere. He found that the pressure increase, relative to the basic state pressure prior to the heating, is greatest at the top of the heated layer and decreases exponentially above the heated layer (see Figure 4.33).

The *e*-folding distance associated with this exponential decrease is proportional to the density scale height in an isothermal atmosphere,

$$H_s = \frac{RT}{g} \tag{4.19}$$

This characteristic height scale, which in the troposphere has a value of about 8 km, is referred to as the *radius of deformation for hydrostatic adjustment*, or perhaps better as the *deformation scale height* in an *isothermal* atmosphere. The effect on pressure of heating in the lower layers of the troposphere will hardly be noticed in the pressure at a height equal to several times the vertical deformation scale height. It appears that the vertical deformation scale in an atmosphere with a realistic thermal structure does not differ very much from the vertical deformation scale in an isothermal atmosphere (Van Delden, 2000). This implies that the effect on pressure of heating in the lower layers of the troposphere will be noticed principally in the troposphere.

The time scale for the purely vertical adjustment (i.e. neglecting horizontal inhomogeneities) is proportional to the time required for the wave of expansion to travel a distance of the order of several times (say 10 times) the deformation scale height. With a typical phase speed of about 300 m s^{-1} and a deformation scale height of about 8 km, the adjustment time is nearly 5 minutes. In other words, within 5 min the pressure perturbation induced by the heating and the vertical readjustment process, has spread out over a characteristic vertical distance equal to 10 times the deformation scale height, and the heated layer has adjusted closely back to hydrostatic balance (Eqn. 4.18).

Lamb's problem gives considerable insight into the process of hydrostatic adjustment to heating, but is a very idealized problem to study. It is worthwhile to imagine what would happen in the more realistic situation of horizontally inhomogeneous heating (Figure 4.34).

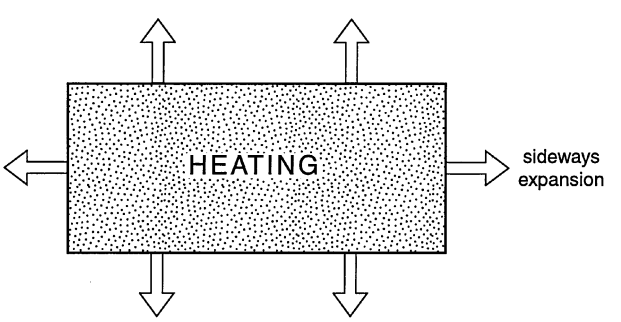


Figure 4.34. Extension of Lamb's problem to the case of horizontally inhomogeneous heating.

If heating is limited in its horizontal extent, for instance, to a circular area with radius, r, it would not only excite a vertically propagating horizontal wave-front at the upper edge of the heated area, but also a spherical wave-front at the outer edge of the heated area propagating horizontally and vertically. Due to its horizontal propagation, this spherical wave-front would affect the one-dimensional (vertical) hydrostatic adjustment process in the centre of the heated area after a time of the order of r/c with c the phase speed of sound waves. If $r \approx 500$ km, this time would be nearly 30 min. If the heating takes place over a sufficiently large region, the adjustment in the centre of a large area experiencing heating, would appear to be a two-step process: first there is a 'vertical' adjustment, involving a horizontal wave-front (i.e. the process described by Lamb's problem), after which there is a 'horizontal' adjustment, involving the spherical wave-front.

This process is illustrated in Figure 4.35a, which shows the pressure perturbation in the vertical plane, calculated with a two-dimensional (x–z) linear numerical model (see Tijm and Van Delden, 1999), 90 s and 270 s after instantaneously heating the lowest 2 km of the atmosphere over a horizontal distance of 100 km.

A basic temperature profile was prescribed as follows:

$$T_0(z) = T_s - \beta z, \tag{4.20}$$

with $T_s = 300$ K, $\beta = (g/c_p) - 0.005$ K m^{-1} for $z < 12$ km, $\beta = 0$ for 12 km $\leq z \leq 50$ km and $\beta = 0.005$ K m^{-1} for $z > 50$ km. The heat source was prescribed such that the lowest 2 km becomes neutrally stratified. The initial potential temperature perturbation, $\theta_i(z)$, is given by:

$$\theta_i(z) = \Theta\left(1 - \frac{z}{H}\right) \text{ for } z \leq H \text{ and } \theta_i(z) = 0 \text{ for } z > H, \tag{4.21}$$

where $\Theta = 10$ K and $H = 2$ km.

Three wave-fronts are formed (see Figure 4.35): a horizontal wave-front moving upwards and two circular wave-fronts (in x–z space) emanating from the edges of the heated layer. These two wave-fronts meet in the centre of the heated layer after about 160 s. The interference of these waves of compression is accompanied by a pressure decrease in the lower half of the heated layer. This is illustrated in Figure 4.35 and also in Figure 4.36, which shows the Exner function perturbation as a function of height in the centre of the domain after 90 s, 180 s and 270 s, respectively.

The profile after 90 s is identical to the profile in Lamb's problem in which there is only a vertically propagating wave. Note that the pressure at the Earth's surface ($z = 0$) does not decrease. However, after 160 s the effect of horizontal

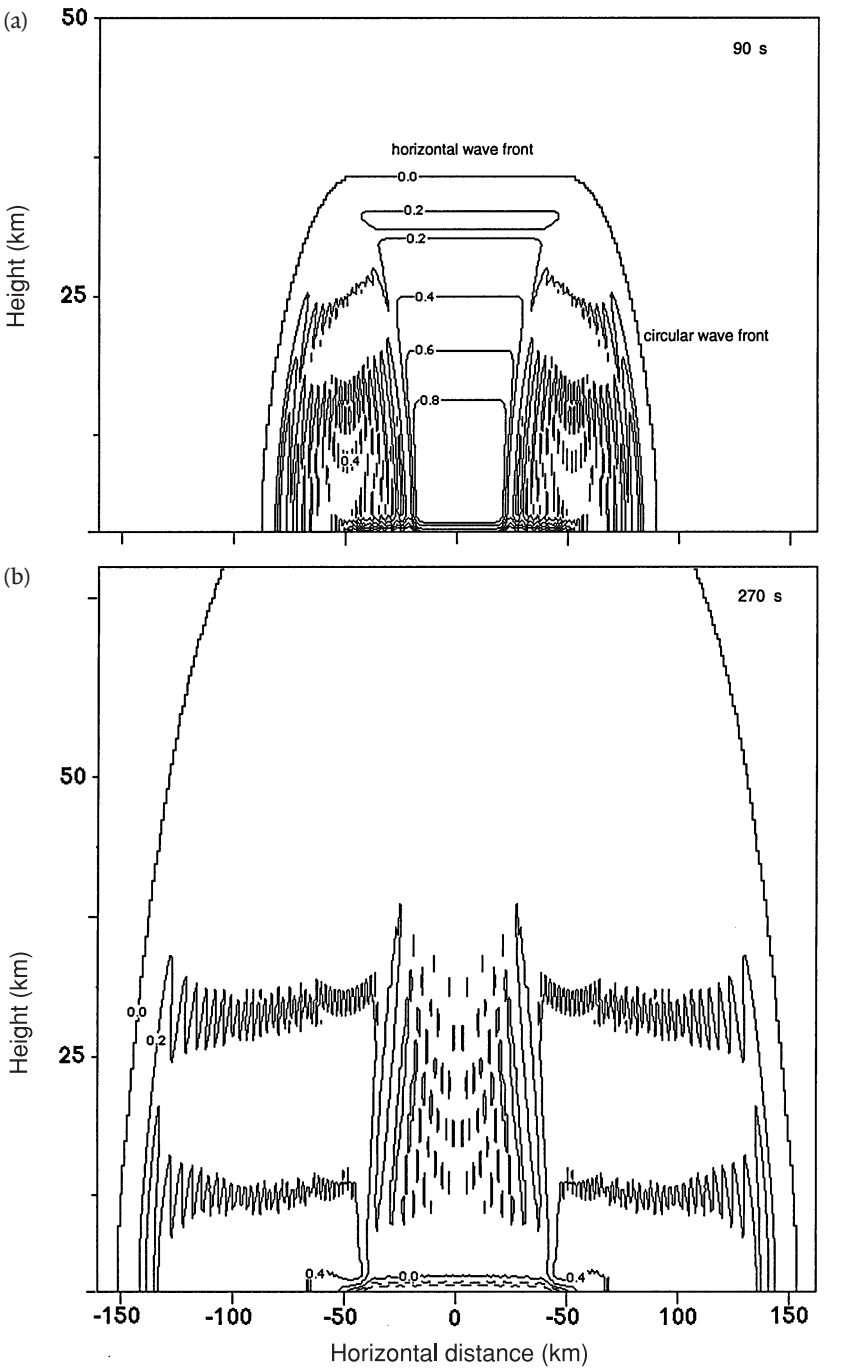


Figure 4.35. Exner function perturbation (J kg^{-1} K^{-1}) as a function of x and z, (a) after 90 s and (b) after 270 s, after the instantaneous heating of the lowest 2 km over a horizontal distance of 100 km in the middle of the domain.

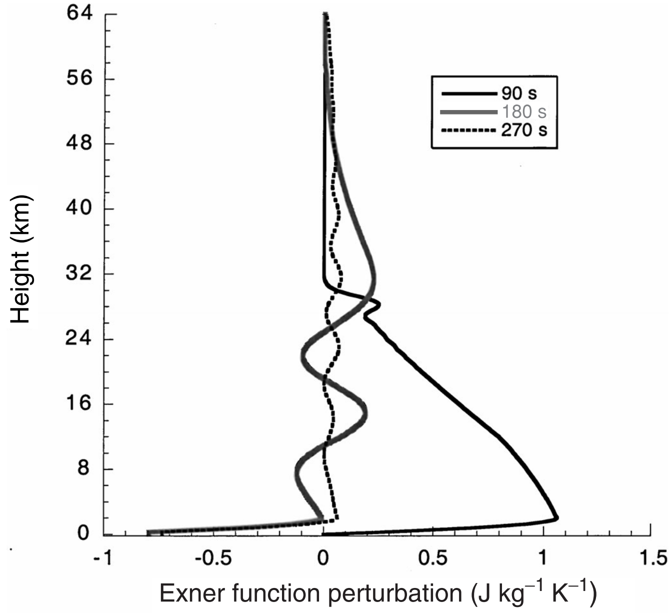


Figure 4.36. Exner function perturbation as a function of height in the middle of the domain ($x = 0$) at three points in time.

expansion of the heated air is felt in the centre of the heated area, leading to a decrease of the total mass above this point and a pressure decrease at the Earth's surface ($z = 0$). The decrease in the Exner function at the Earth's surface after 270 s of 0.8 J kg^{-1} K^{-1} is equivalent to a pressure decrease of about 2.7 hPa (if we assume that the pressure at the Earth's surface is 1000 hPa initially).

In reality, heating does not take place instantaneously. If the heat is added over a finite period, a weaker wave front travels a certain distance from the source before the heating is completed. The wavefront is less well defined than in the case of instantaneous heating. But, qualitatively, the result is the same. A pressure decrease is observed in the lower half of the heated boundary layer while a pressure increase is observed aloft and over the area where no heating has taken place (Tijm and Van Delden, 1999). The vertical pressure gradient force is approximately in (hydrostatic) balance with the buoyancy force, while the horizontal pressure gradient drives a circulation, which, if we apply the model to the core of a cyclone, can be identified with the radial circulation. It is important to note now, that such a circulation arises in an absolutely statically stable atmosphere. That is, CAPE is not required. As soon as the heating stops the non-rotating atmosphere will eventually return to a state of rest. We will see in Section 4.5.4, 'Geostrophic adjustment and the invertibility principle' that this is not the case if the atmosphere is rotating.

In many theoretical models of atmospheric circulations, it is assumed that the atmosphere is in hydrostatic balance at all times. In such a model, hydrostatic adjustment is 'instantaneous'. By integrating the hydrostatic relation and using the ideal gas law, it can be shown that if a layer of fixed thickness Δz in an isothermal atmosphere is heated such that the temperature increases uniformly by ΔT and the pressure remains fixed at the bottom of this layer, the pressure at the top of this layer will increase by a factor $\exp\{\Delta z \Delta T/[H_s(T + \Delta T)]\}$. According to the hydrostatic approximation, the heating of a layer of air gives rise to an instantaneous pressure increase at all levels above the base of this layer of air. The pressure increase above the heated layer is the *same* at all heights, erroneously implying an infinite deformation scale height. This error is a consequence of not taking into account the lifting and consequent adiabatic temperature decrease of the atmosphere above the heated layer.

Assuming the temperature $T = 300\,\mathrm{K}$, $\Delta T = 5\,\mathrm{K}$ and $\Delta z = 2\,\mathrm{km}$, we obtain a factor of 1.0037. If the atmospheric layer in question is located just above the Earth's surface, we obtain a pressure increase of about 3 hPa (assuming that the pressure at 2 km height is 800 hPa prior to the heating). If the consequent divergence of mass were to eliminate this upper-level pressure perturbation, which is not necessarily the case if the effects of rotation (i.e. 'Coriolis' effects) are taken into account, the pressure decrease at the Earth's surface would attain a maximum possible amplitude of about 3 hPa. Of course, convergence of mass near the Earth's surface would tend to partly eliminate this negative pressure perturbation. Eventually, a balance between the inertial forces associated with rotation and the pressure gradient force (i.e. geostrophic balance or gradient wind balance) will be reached. We are, in fact, interested in determining the exact distributions of pressure, temperature and wind associated with this balanced state. The question now is, do we need to know all the details of the sound waves and other slower moving waves emanating from the heated region in order to determine the balanced state? In the next section we will show that (and expain why) this, fortunately, is not the case.

4.5.4 Geostrophic adjustment and the invertibility principle

In this section we illustrate some typical characteristics of adjustment to geostrophic balance using a simple one-layer model of a rotating, density stratified fluid. We intend to demonstrate that the adjustment process is dominated by a materially conserved quantity called potential vorticity. We will show that we need not know the details of the waves and oscillations, excited as a consequence of an imbalance of forces, in order to determine the ultimate state of geostrophic balance to a relatively high degree of accuracy.

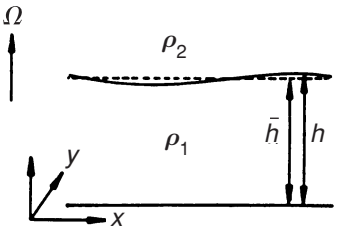


Figure 4.37. Geometry of the shallow slab-symmetric rotating layer of fluid.

We will not adopt the model used in the previous section to make our point, because this would be computationally prohibitive. Instead, we adopt the most simple model of a rotating stratified fluid, which illustrates some essential characteristics of adjustment to geostrophic balance (Figure 4.37). This model consists of a layer of incompressible fluid (depth h and constant density ρ_1) below an infinitely deep motionless layer with density $\rho_2 < \rho_1$. The fluid is rotating with a constant angular velocity equal to $\Omega = f/2$. Assuming hydrostatic balance, the equations of motion and continuity for the lower layer are (Gill, 1982)

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\partial u \tag{4.22a}$$

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -g'\frac{\partial h}{\partial x} + \partial v \tag{4.22b}$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} = -h\frac{\partial u}{\partial x} \tag{4.22c}$$

In these equations $u(x, t)$ and $v(x, t)$ are the x- and y-components of the velocity, respectively, $h(x, t)$ is the height of the lower layer, $\mathrm{d}/\mathrm{d}t = \partial/\partial t + \partial u/\partial x$, and $g' = g(\rho_1 - \rho_2)/\rho_1$ is the reduced gravity. We have neglected derivatives with respect to y.

Suppose we extract a specified volume of mass from the lower layer. To incorporate this effect into the model we set $h = \bar{h} + h'$ at $t = 0$, where $\bar{h}$ is a constant reference height and where

$$h' = h'_{\max}\exp\left\{-\left(\frac{x - x_0}{a}\right)^2\right\} \tag{4.23}$$

This represents a bell-shaped perturbation in the height of the free surface centred at $x = x_0$ with a horizontal scale represented by the parameter a and a maximum amplitude equal to $h'_{\max}$. Due to this perturbation, horizontal pressure gradients are created in the lower fluid layer leading to convergence of mass

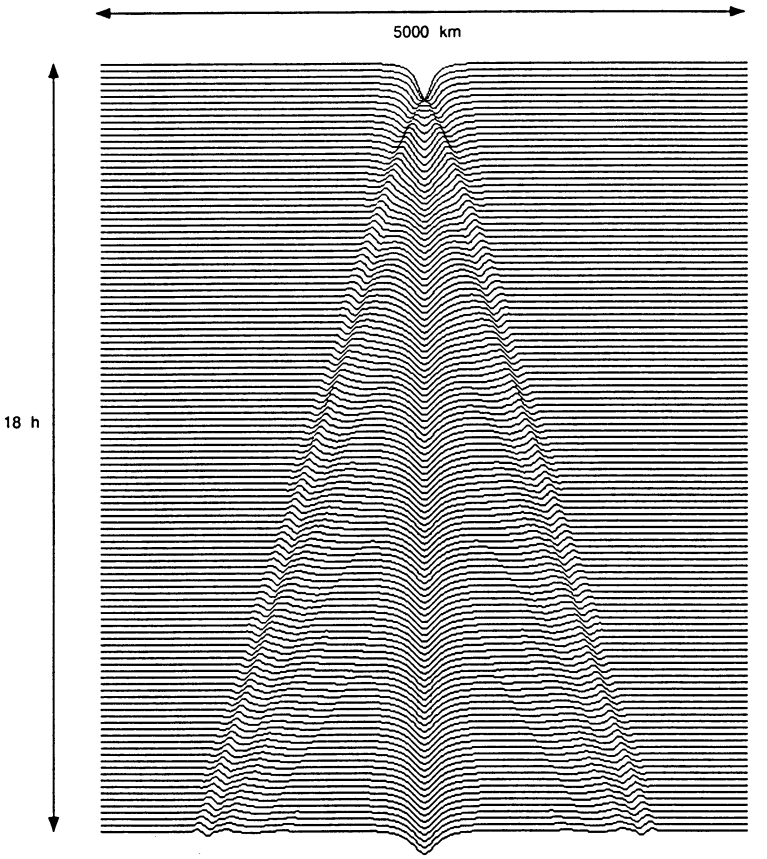


Figure 4.38. Height of the free surface as a function of time and horizontal distance. The initial perturbation in the free surface has a horizontal scale a of 60 km and a maximum amplitude of 50 m at $t = 0$. The height, $\bar{h} = 1000$ m, $f = 0.0005\ \mathrm{s}^{-1}$ and $g' = 1\ \mathrm{m\ s}^{-2}$. The Rossby radius is 63.2 km. The waves observed are gravity-inertia waves.

towards $x = x_0$. A gravity-inertia wave is the result. This wave propagates towards $x = \pm\infty$ (sound waves are not possible here because of the condition of incompressibility). At the same time adjustment to geostrophic balance, given by

$$g' \frac{\partial h}{\partial x} = fv;\ u = 0 \tag{4.24}$$

takes place in the region where the perturbation was inserted. Figure 4.38 visualizes the gravity-inertia waves and the adjustment to geostrophic balance in the centre of the domain.

Although the exact functional relation between h (or v) and x in the geostrophically balanced state can, in principle, be derived from Eqn. (4.22) by

numerical integration, it can also be determined directly from Eqn. (4.24) with the definition of potential vorticity,

$$\zeta_{\text{pot}} \equiv \frac{\zeta + f}{h} \tag{4.25}$$

(the relative vorticity here is $\zeta = \partial v/\partial x$). If we differentiate Eqn. (4.24) with respect to x (assuming g' and f are constant) and substitute the result into Eqn. (4.25), and then again differentiate the resulting equation with respect x and again use Eqn. (4.24), we obtain the following equation for v in the balanced state:

$$\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} - \frac{f\zeta_{\text{pot}}}{g'} v = h\frac{\mathrm{d}\zeta_{\text{pot}}}{\mathrm{d}x} \tag{4.26}$$

Eqn. (4.26), which is of the elliptic type if $\zeta_{\text{pot}} > 0$, is an expression of the so-called *invertibility principle*. This principle states that the velocity distribution in the balanced state can be determined exactly given the potential vorticity distribution and boundary conditions. In the example discussed here (see Figure 4.38), the potential vorticity in the final balanced state is, of course, not known. However, we do know that ζ_{pot} is materially conserved. This can easily be deduced from Eqn. (4.22). If we neglect horizontal advection of ζ_{pot}, the potential vorticity distribution at $t = 0$ is identical to the potential vorticity distribution at any later time. We can then solve Eqn. (4.26) numerically by successive over-relaxation assuming that $h(t \to \infty)$ on the r.h.s. of Eqn. (4.26) is equal to $\bar{h}$, which is reasonable if $h' << \bar{h}$ initially. This yields the solution shown by the thick broken line in Figure 4.39, which, within a certain distance from the place of insertion of the perturbation, apparently is nearly identical to the solution of the time-dependent Eqn. (4.22) after 96 h of integration (the thin solid line in Figure 4.39).

This remarkable fact implies that the potential vorticity, which is inserted initially, indeed practically stays in place. In other words, the potential vorticity perturbation does not propagate away from the source region with the waves. Since the potential vorticity determines the balanced state, this balanced state must therefore be nearly identical in both cases, in spite of the presence of large amplitude waves in one case.

The solution of the homogeneous part of Eqn. (4.26) is of the form

$$v \approx \exp\left(\pm\frac{x}{\lambda}\right) \tag{4.27}$$

with

$$\lambda \equiv \sqrt{\frac{g'}{f\zeta_{\text{pot}}}} \tag{4.28}$$

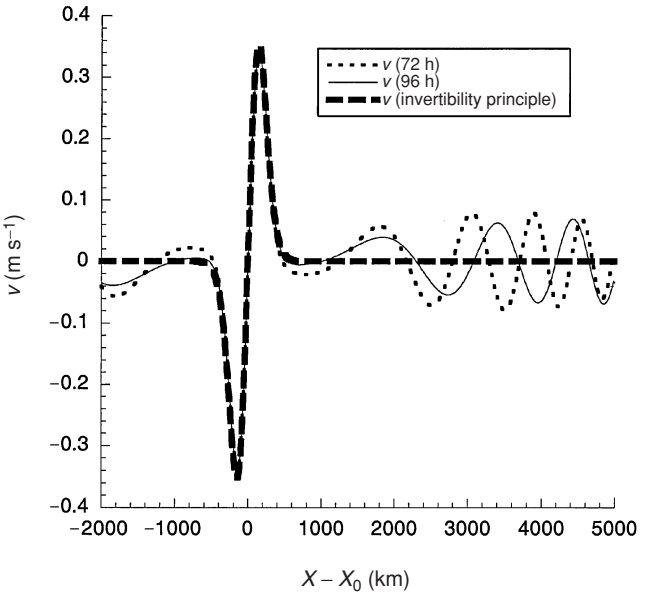


Figure 4.39. Velocity v as a function of horizontal distance according to the invertibility principle (Eqn. 4.26) (thick broken line) compared to the time-dependent solution for $t = 72$ h (dotted line) and $t = 96$ h (thin solid line). The initial perturbation in the free surface has a horizontal scale, a, of 180 km and a maximum amplitude of 50 m at $t = 0$. Height, $\bar{h} = 1000$ m, $f = 0.0005\ \mathrm{s}^{-1}$ and $g' = 1\ \mathrm{m\ s}^{-2}$. The Rossby radius is 63.2 km.

the so-called *Rossby radius of deformation* for geostrophic adjustment. The solution Eqn. (4.27) in effect states that horizontal variations in potential vorticity force or 'induce' a velocity field with a characteristic horizontal scale equal to λ. If $\zeta << f$ the Rossby radius can be expressed in the more familiar form as

$$\lambda \equiv \frac{\sqrt{g'h}}{f} \tag{4.29}$$

i.e. as the ratio of the phase speed of surface gravity waves and the Coriolis frequency.

The Rossby radius of deformation can be viewed as the analogue of the deformation height for hydrostatic adjustment (see the previous section). The Rossby radius of deformation is the e-folding distance characterizing the horizontal scale of the pressure perturbation in the centre of the domain in the final geostrophic equilibrium (Figure 4.38).

The robustness of the invertibility principle is illustrated by Figure 4.40. The solid curve is the result of many 72 h integrations of the shallow water equations (Eqn. 4.22), with varying values of the Coriolis parameter (f) or

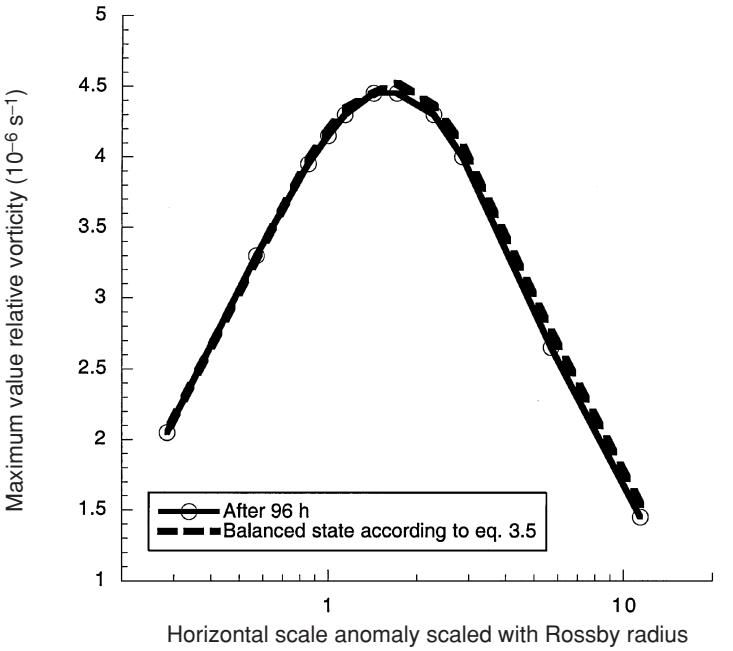


Figure 4.40. The maximum value of ζ (in units of 10^{-6} s^{-1}) as a function of the horizontal scale of the perturbation a, divided by the Rossby radius at $t = 96$ h, according to Eqns. (4.22a, b, c) with Eqn. (4.23) as initial condition and with $h'_{\max} = 50$ m and $a = 180$ km (solid line), and according to the invertibility principle (Eqn. 4.26) using the initial potential vorticity distribution (broken line) $\bar{h} = 1000$ m, and $g' = 1$ m s^{-2}).

equivalently with varying values of the Rossby radius. The maximum value of the relative vorticity after 72 h of integration is shown as a function of the Rossby radius. Also shown is the maximum value of the relative vorticity according to the invertibility principle (Eqn. 4.26), assuming the potential vorticity distribution at $t = 0$ is fixed in time (the dashed curve). The fact that the two curves in Figure 4.40 practically coincide demonstrates that the motion associated with the waves generated by the initial imbalance (i.e. the 'unbalanced' part of the motion) is of negligible importance to the 'balanced' part of the motion, which is represented by the invertibility principle. This is also found in more realistic (i.e. more complex) models of the atmosphere (Davis *et al.*, 1996). We will return to this important conclusion in a later section.

The fluid adjusts to the new distribution of potential vorticity by converting potential vorticity associated with the perturbation in h into potential vorticity associated with relative vorticity, ζ. Apparently (Figure 4.40) this conversion attains an optimum when the scale, a, of the initial perturbation is slightly greater than the Rossby radius. If a is significantly greater than the Rossby radius, the effect of the *rotational stiffness* (or inertial stability) is relatively

strong, so that very little conversion of height-related to vorticity-related potential vorticity is possible. If a is significantly smaller than the Rossby radius, the perturbation spreads out over a distance of the order of the Rossby radius, which is also less effective for the above-mentioned conversion.

The processes just discussed must be considered as an analogue of what happens in the real atmosphere when it is heated locally. The mass extracted is an analogue for the heating, because both effects disturb the pressure distribution as well as the potential vorticity distribution, as we will see in the next section. In the real atmosphere the balanced state need not be in geostrophic balance. For instance, in an axisymmetric cyclone it is gradient wind balance (Section 4.5.6), or in a more complicated flow situation it is described by the nonlinear balance equation (Holton, 1992, p. 387). Nevertheless, in all these cases it is possible to formulate an invertibility principle, which relates potential vorticity to the balanced flow.

4.5.5 How heating changes the distribution of potential vorticity

Since potential vorticity determines the wind field in the balanced state, changes in the potential vorticity distribution will determine changes in the wind field. It is therefore of considerable interest to examine how heating changes the distribution of potential vorticity in a cyclone. Following Kleinschmidt (1951) and many others we assume that the cyclone is axisymmetric, so that we need only examine the structure of the vortex in the radius–height plane. If we further assume hydrostatic balance, it is possible and convenient to use potential temperature as vertical coordinate instead of height. Hydrostatic balance with potential temperature (θ) as vertical coordinate is expressed as (Anthes, 1971)

$$\frac{\partial \Psi}{\partial \theta} = \Pi \tag{4.30}$$

where Π is the Exner function, defined in Eqn. (4.17), and Ψ is the isentropic streamfunction, defined according to $\Psi = c_{\mathrm{p}}T + gz$, with z the height of the isentropic surface.

The time-dependent equations of motion and the continuity equation in cylindrical/isentropic coordinates (r, θ), where r is the radius, are (Anthes, 1971)

$$\frac{\mathrm{d}ru}{\mathrm{d}t} = -rfv \tag{4.31a}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = u\left(f + \frac{u}{r}\right) - \frac{\partial \Psi}{\partial r} \tag{4.31b}$$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = -\sigma\left(\frac{1}{r}\frac{\partial rv}{\partial r} + \frac{\partial}{\partial\theta}\frac{J}{\Pi}\right) \tag{4.31c}$$

Again, we assume that f is constant. Note the mathematical similarity between Eqn. (4.31) and Eqn. (4.22). In Eqn. (4.31), v is the radial velocity, u is the azimuthal velocity (positive if the flow is cyclonic; negative if the flow is anticyclonic), σ is the inverse of the static stability (defined below), and $\mathrm{d}/\mathrm{d}t = \partial/\partial t + v\partial/\partial r + (J/\Pi)\partial/\partial\theta$. The radial derivative is taken at constant θ. The 'vertical velocity' in this coordinate system is equal to (J/Π). The inverse of the static stability is defined as

$$\sigma = -\frac{1}{g}\frac{\partial p}{\partial\theta} \tag{4.32}$$

From Eqn. (4.31a) we find that

$$\frac{\mathrm{d}M_\mathrm{a}}{\mathrm{d}t} = 0 \tag{4.33}$$

where $M_\mathrm{a} \equiv ru + fr^2/2$ is the angular momentum per unit mass. Eqn. (4.33) implies that M_a is materially conserved, even in the presence of heating.

It is easily deduced from Eqn. (4.31a) and Eqn. (4.31c) that

$$\frac{\mathrm{d}Z_\theta}{\mathrm{d}t} = Z_\theta\frac{\partial}{\partial\theta}\left(\frac{J}{\Pi}\right) - \frac{1}{\sigma}\frac{\partial u}{\partial\theta}\frac{\partial}{\partial r}\left(\frac{J}{\Pi}\right) \tag{4.34}$$

where Z_θ is the isentropic potential vorticity, defined as

$$Z_\theta \equiv \frac{\zeta_\theta + f}{\sigma} \tag{4.35}$$

where $\zeta_\theta = \partial u/\partial r + u/r$ is the relative vorticity on an isentropic surface. Eqn. (4.34) tells us *that isentropic potential vorticity is materially conserved if there are no heat sources* (remember: we have neglected friction). The first term on the r.h.s. of Eqn. (4.34) can be interpreted as a 'stretching' or shrinking effect (a vertically-confined source of heat affects the distance between isentropic surfaces). The second term on the r.h.s. of Eqn. (4.34), can be interpreted as a 'tilting' effect. It is non-zero if the vortex is baroclinic and if the heat source is horizontally confined. Isentropic potential vorticity appears partly as absolute vorticity and partly as static stability. The proportions of this partitioning are adjusted through radial and vertical motions.

It is interesting to note that the absolute vorticity can be expressed in terms of M_a as

$$\zeta_{\theta_\mathrm{abs}} \equiv \zeta_\theta + f = \frac{1}{r}\frac{\partial M_a}{\partial r} \tag{4.36}$$

The further interpretation of the terms on the r.h.s. of Eqn. (4.34) is made easier if we first discuss the structure of a balanced cyclone in terms of potential vorticity (Z_θ). But, before we proceed with this, it is perhaps interesting to mention an additional constraint on the potential vorticity budget.

Using the fact that div curl of any vector field is zero and remembering that we have axisymmetry, Eqn. (4.34) can be rewritten as

$$\frac{\mathrm{d}Z_\theta}{\mathrm{d}t} = \frac{1}{\sigma}\vec{\omega}_\mathrm{a} \cdot \vec{\nabla}\left(\frac{J}{\Pi}\right) = \frac{1}{\sigma}\vec{\nabla} \cdot \left(\vec{\omega}_\mathrm{a}\left(\frac{J}{\Pi}\right)\right), \tag{4.37}$$

where $\vec{\omega}_\mathrm{a}$ is the absolute vorticity vector. If we integrate this equation over a material volume V, with a surface S, and use Gauss's divergence theorem, we find

$$\int_V \sigma \frac{\mathrm{d}Z_\theta}{\mathrm{d}t}\mathrm{d}V = \int_S \left(\frac{J}{\Pi}\vec{\omega}_\mathrm{a}\right) . \mathrm{d}\vec{S} \tag{4.38}$$

The r.h.s of this equation is zero if there are no heat sources on the boundary S. In other words, for every positive material tendency of isentropic potential vorticity there is, within a layer bounded by two isentropic surfaces, and with no heating at the lateral edges, a 'compensating' negative material tendency of isentropic potential vorticity (weighted by σ and integrated over a material volume, $2\pi r\,\mathrm{d}r\,\mathrm{d}\theta$). The interpretation of Eqn. (4.38) is discussed at length by Haynes and McIntyre (1987, 1990) and by Danielsen (1990).

4.5.6 The potential vorticity structure of a circularly symmetric cyclone

Let us assume a vortex which is in hydrostatic balance Eqn. (4.30) and in gradient wind balance:

$$u\left(f + \frac{u}{r}\right) = \frac{\partial \Psi}{\partial r} \tag{4.39}$$

The equations for hydrostatic balance (Eqn. 4.30) and for gradient wind balance (Eqn. 4.39) together yield the equation for thermal wind balance:

$$\frac{\partial \Pi}{\partial r} = \frac{c_p}{\theta}\frac{\partial T}{\partial r} = f_\mathrm{loc}\frac{\partial u}{\partial \theta} \tag{4.40}$$

where $f_\mathrm{loc} = f + (2u/r)$. Thus, we see that in a warm core vortex (with $[\partial T/\partial r]_\theta < 0$), u must decrease with increasing potential temperature or height (i.e. the thermal wind is anticyclonic), while in a cold core vortex (with $[\partial T/\partial r]_\theta > 0$), u must increase with increasing potential temperature or height.

Let us assume that the vortex has an azimuthal velocity profile given by

$$u(r,\theta) = \frac{2B(\theta)\hat{u}r}{\hat{r}\left[1+\left(\frac{r}{\hat{r}}\right)^2\right]} \tag{4.41}$$

where $\hat{r}$ is the radius of maximum wind, $\hat{u}$ is the maximum azimuthal wind and $B(\theta)$ is the so-called baroclinicity of the vortex. In a cyclone, the value of $B(\theta)$ varies between 0 and +1. If $B(\theta)$ decreases with height, the balanced vortex has a warm core.

We now specify the following boundary conditions. The pressure and the potential temperature at the Earth's surface ($z = 0$) at $r = 0$ are specified by $p = p_s = 1000$ hPa and $\theta = \theta_s = 275$ K. The potential temperature lapse rate, $\partial\theta/\partial p$, at $r = 0$ is then fixed such that the temperature in the centre of the cyclone decreases from 275 K at the Earth's surface to 225 K at $z = 10\,200$ m ($\theta = 335$ K) and subsequently decreases more slowly to 210 K at $z = 18\,095$ m ($\theta = 445$ K). Therefore, the potential temperature lapse rate in the lower 10000 m at $r = 0$ is *c.* 5 K km^{-1}.

With the balance conditions (Eqns. 4.30 and 4.39) we can now obtain the distribution of M_a and Z_θ as a function of θ and r, belonging to a particular choice of the parameters $\hat{r}$, $\hat{u}$ and $B(\theta)$. The surface pressure, $p_s(r > 0)$, which is required in order to diagnose σ at the lowest isentropic level ($\theta = 280$ K), is obtained from gradient wind balance using (Eqn. 4.41) and the boundary condition that the thermal wind is equal to zero in the layer between the Earth's surface and the lowest isentropic level ($\theta = 280$ K). We therefore neglect the effects of friction and assume that the Earth's surface is an isentropic level. The lowest isentropic level above the ground ($\theta = 280$ K) does not reach the ground at any point within the model domain. We thus avoid the difficulties associated with translating potential temperature anomalies at the Earth's surface to potential vorticity anomalies (see e.g. Hoskins *et al.*, 1985; Bleck, 1990).

A prototype cyclone in the atmosphere has a cold core in the troposphere. The associated thermal wind is cyclonic in the troposphere, while it is anticyclonic in the stratosphere. If we choose $\hat{u} = 20$ m s^{-1} and $\hat{r} = 240$ km and specify $B(\theta)$ as displayed by the thick solid line in Figure 4.41, we obtain a distribution of Z_θ shown in Figure 4.42a. In terms of the wind velocity, the intensity of a cold core cyclone is a maximum at the top of the troposphere. However, as far as the potential vorticity is concerned, this cyclone is most intense in the stratosphere, where it is in fact a 'warm' core cyclone.

A polar low or tropical cyclone typically has a warm core in the lower troposphere implying an anticyclonic thermal wind. Its horizontal scale is

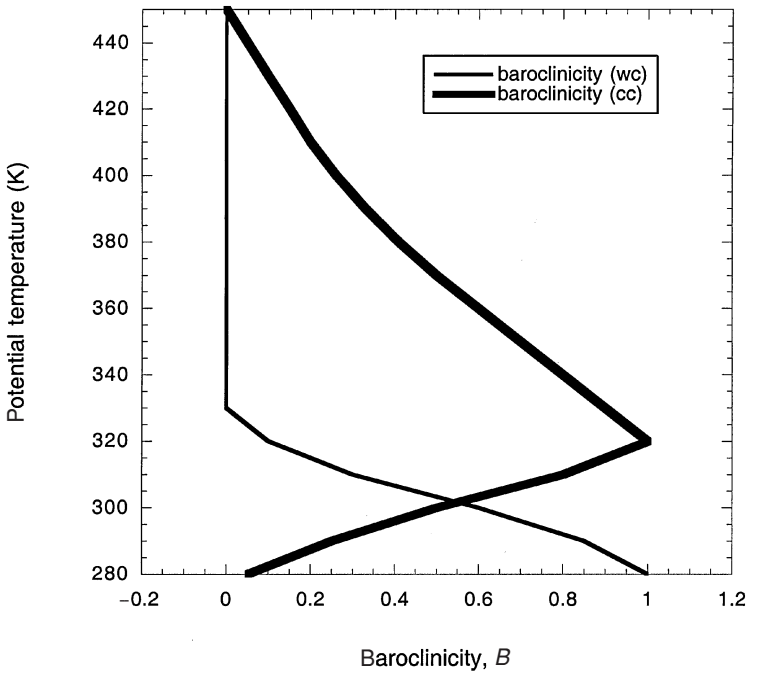


Figure 4.41. Baroclinicity parameter B of the prescribed vortex as a function of potential temperature in the case of a warm core (wc) and in the case of a cold core (cc).

relatively small. A representative value of the radius of maximum azimuthal wind is $\hat{r} = 100$ km. If we choose $\hat{u} = 10$ m s^{-1} and specify $B(\theta)$ according to the thin curve in Figure 4.41, we obtain from the balance conditions (Eqns. 4.30 and 4.39) the distribution of Z_θ as shown in Figure 4.42b. We see that the polar low is a relatively shallow phenomenon. The maximum potential vorticity is found in the lower troposphere. We will see in Section 4.5.8, that this property of warm core cyclones is advantageous for their growth due to heating.

4.5.7 The invertibility principle applied to an axisymmetric cyclone

We have seen that a particular wind distribution is associated with a particular potential vorticity distribution if the atmosphere is in some kind of dynamic balance. Of course, as was explained in the section 'Geostrophic adjustment and the invertibility priciple', the inverse problem is also of relevance: a particular potential vorticity distribution is associated with a specific wind distribution. In this section we explain some dynamical consequences of this so-called invertibility principle. Following Hoskins *et al.*

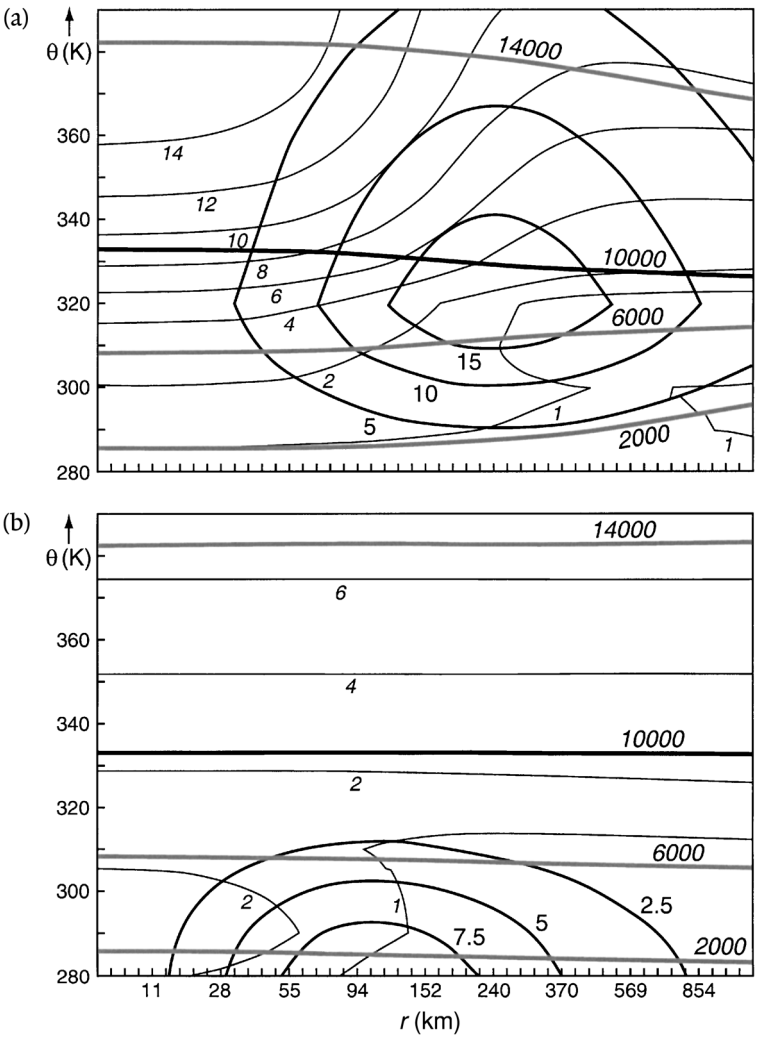


Figure 4.42. Isentropic potential vorticity (labelled in small italic numbers in PV units; 1 PVU is 10^{-6} K m^2 kg^{-1} s^{-1}), height (labelled in large italic numbers in m) and azimuthal velocity (labelled in m s^{-1}) as a function of potential temperature and radius, for (a) a balanced cold core cyclone and (b) a balanced warm core cyclone. The azimuthal velocity is specified by Eqn. 4.41. B(θ) for both cases is plotted in Figure 4.41. Parameters and static stability at $r = 0$ are specified in the text.

(1985), let us differentiate Eqn. (4.35) with respect to r. Using Eqn. (4.32), this yields

$$\frac{\partial}{\partial r}\left\{\frac{1}{r}\frac{\partial(ru)}{\partial r}\right\} + \frac{Z_\theta}{g}\frac{\partial^2 p}{\partial r\,\partial\theta} = \sigma\frac{\partial Z_\theta}{\partial r} \tag{4.42}$$

Now, using the definition of the Exner function (Eqn. 4.17) and the equation of state ($p = \rho RT$), we can express thermal wind balance (Eqn. 4.40) as

$$\frac{\partial p}{\partial r} = f_{\mathrm{loc}} \rho\theta \frac{\partial u}{\partial \theta} \tag{4.43}$$

(f_{loc} was defined below Eqn. (4.40)). Using Eqn. (4.43) to eliminate p from Eqn. (4.42) yields

$$\frac{\partial}{\partial r}\left\{\frac{1}{r}\frac{\partial (ru)}{\partial r}\right\} + \frac{Z_\theta}{g}\frac{\partial}{\partial \theta}\left(\rho\theta f_{\mathrm{loc}}\frac{\partial u}{\partial \theta}\right) = \sigma\frac{\partial Z_\theta}{\partial r} \tag{4.44}$$

This equation, which is another version of the 'invertibility principle' discussed earlier in the context of the one layer model, describes the flow pattern, $u(r, \theta)$, which is associated with a specific pattern of the isentropic potential vorticity in a balanced axisymmetric cyclone. If $f_{\mathrm{loc}} Z_\theta > 0$, Eqn. (4.44) is elliptic. Given the distribution of Z_θ and suitable boundary conditions, the coupled system of Eqns. (4.43) and (4.44) can be solved for the wind $u(r, \theta)$, for instance by relaxation methods (see e.g. Wirth, 2001, p. 29). The term on the r.h.s. of Eqn. (4.44) can be interpreted as 'forcing'. In other words, a radial gradient in Z_θ 'forces' or 'induces' an azimuthal flow.

In order to investigate the implications of the invertibility principle in a little more detail, let us neglect effects of curvature in Eqn. (4.44) by taking the limit $r \to \infty$ and further assume that $\rho\theta f_{\mathrm{loc}} \simeq \rho\theta f$ is a constant. The equation then transforms into the following equation

$$\frac{\partial^2 u}{\partial r^2} + A\frac{\partial^2 u}{\partial \theta^2} = B \tag{4.45}$$

with

$$A \equiv \frac{\rho\theta f(f + \zeta_\theta)}{g\sigma}; \quad B \equiv \sigma\frac{\partial Z_\theta}{\partial r} \tag{4.46}$$

Assuming that A is constant (which makes Eqn. (4.45) linear) and that u depends on r according to

$$u = U(\theta)\sin\frac{2\pi r}{L} \tag{4.47}$$

where L is the typical horizontal scale of the response to the forcing, B, we get

$$A\frac{\partial^2 U}{\partial \theta^2} - \frac{4\pi^2}{L^2}U = B \tag{4.48}$$

The solution of the homogeneous part of the above equation (i.e. with the forcing, $B = 0$) is

$$U = C_1 ex_\rho \left(\frac{2\pi\theta}{L\sqrt{A}} \right) + C_2 ex_\rho \left(\frac{-2\pi\theta}{L\sqrt{A}} \right) \tag{4.49}$$

where C_1 and C_2 are constants determined by the boundary conditions and the nature of the inhomogeneous term (i.e. the forcing, B). If $A > 0$ (which is the case if $Z_\theta > 0$) this solution describes an exponentially decaying function of θ with a maximum or minimum value in the region where B (i.e. the forcing) is non-zero. If B is negative at a certain height, $U > 0$ (cyclonic flow), whereas if B is positive at a certain height, $U < 0$ (anticyclonic flow). The idea is that a potential vorticity anomaly *induces* (or forces) a wind field, $u(r, \theta)$, with a vertical dependence given approximately by Eqn. (4.49) (Figure 4.43). Stated differently, local changes of the potential vorticity (due to for example heating) must be accompanied by local changes in the wind (in order to preserve thermal wind balance). Thorpe (1997) coined the term *attribution* as a slightly weaker form of 'cause-and-effect' to characterize the relation between potential vorticity and the induced wind field. In other words, a particular change in the wind field can be *attributed* to a particular change in the potential vorticity field. The associated vertical scale, $\Delta\theta$, of the response in u is easily distilled from Eqn. (4.49):

$$\Delta\theta \equiv L\sqrt{A} \tag{4.50}$$

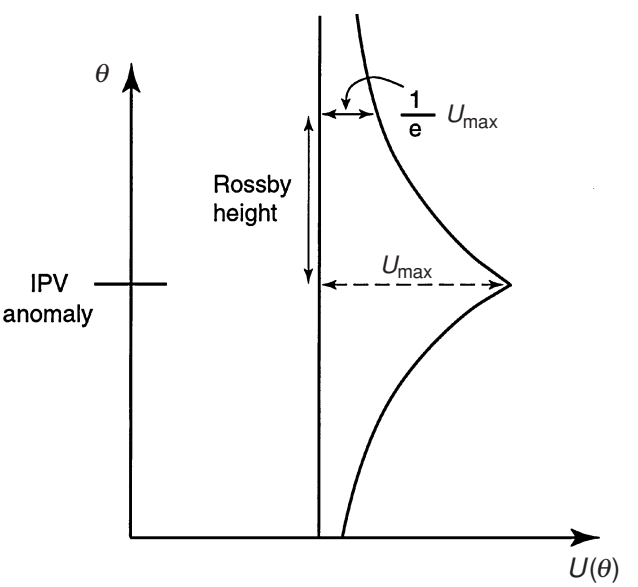


Figure 4.43. Schematic graph of the amplitude of the azimuthal wind as a function of potential temperature, induced by a PV anomaly.

$\Delta\theta$ is referred to by Hoskins *et al.* (1985) as the *Rossby height*. It measures the vertical penetration (in isentropic coordinates) of the flow structure above and below the location of the potential vorticity anomaly, induced by thermal wind adjustment to the anomaly. The potential vorticity anomaly, in turn, may be induced by heating. The concept of Rossby height is complementary to that of the deformation scale height for hydrostatic adjustment (see Eqn. 4.19) or that of the Rossby radius of deformation for geostrophic adjustment (see Eqn. 4.28). Stated shortly, we may say that a potential vorticity perturbation induced by heating 'induces' a perturbation in the wind field with a characteristic vertical scale equal to the Rossby height. The Rossby height for the potential vorticity anomaly shown in Figure 4.42b (with $L \approx 200$ km), is of the order of 10 K in the troposphere and the lower stratosphere.

We can transform the expression for the Rossby height to physical space, using hydrostatic balance, written as $\Delta p = -\rho g \Delta z$, and Eqn. (4.32), written as $\sigma = -\Delta p/(g\Delta\theta)$. This yields

$$\Delta z \equiv \frac{\sqrt{f(f+\zeta_\theta)}L}{N} \tag{4.51}$$

where N is the buoyancy frequency, defined as

$$N \equiv \sqrt{\frac{g}{\theta}\frac{\Delta\theta}{\Delta z}} \tag{4.52}$$

Expression (4.51) for the Rossby height (Δz) demonstrates that the cyclonic wind field 'induced' by potential vorticity anomalies caused by heating has the greatest vertical penetration if the static stability in the environment is low and/or if the absolute vorticity of the environment is high and/or if the horizontal scale of the perturbation is large.

Another interpretation of (4.51) is that, since in general $N >> \sqrt{f(f+\zeta_\theta)}$, balanced circulation systems typically have a large aspect ratio, $L/\Delta z$.

4.5.8 Axisymmetric adjustment to a circular heat source

We now know that there is a specific relation between the isentropic potential vorticity distribution and the wind distribution in a balanced axisymmetric cyclone (Eqn. 4.44). We also know that heating changes the isentropic potential vorticity of an air parcel in a specific way, depending on its potential vorticity (Z_θ) and the local baroclinicity ($\partial u/\partial\theta$) (Eqn. 4.34).

These two facts imply that heating will affect the wind. Let us investigate this effect more in detail by assuming that the atmosphere is in rest initially

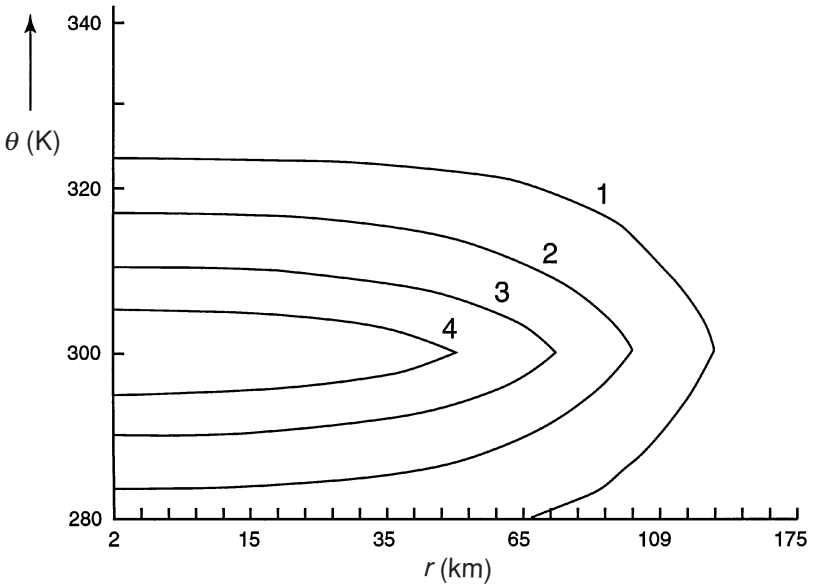


Figure 4.44. Prescribed heating as a function of potential temperature and radius (labelled in units of 10^{-4} K s^{-1}) ($r^* = 100$ km).

and specifying a circularly symmetric heat source, whose intensity depends on r and θ as

$$\frac{d\theta}{dt} \equiv \dot{\theta} = \frac{J}{\Pi} = Q_0(\theta) \exp\left\{-\left(\frac{r}{r^*}\right)^2\right\} \tag{4.53}$$

with $Q_0(\theta)$ specifying the vertical distribution of the heating. Note that the heating according to Eqn. (4.53) has a maximum value at $r = 0$ and that the amplitude of the heating falls off exponentially with increasing r. The associated e-folding distance is $r = r^*$. The net latent heat released to the air at any height depends on the difference between condensation and evaporation and between any freezing and melting at that height. Since evaporation and melting in a precipitating cloud occurs principally below the cloud and condensation and freezing principally higher up within the clouds, the net heating will possess a maximum value at mid-levels in the troposphere. $Q_0(\theta)$ is specified accordingly, yielding a schematic heating distribution as shown in Figure 4.44.

It is important to note that the Earth's surface is assumed to be an isentropic surface, i.e. $Q_0(\theta = \theta_s) = 0$ at the Earth's surface. The maximum value of J/Π is 0.0005 K s^{-1} which corresponds to about 1.8 K h^{-1}.

Note also that the vertical scale of the heating is significantly larger than the

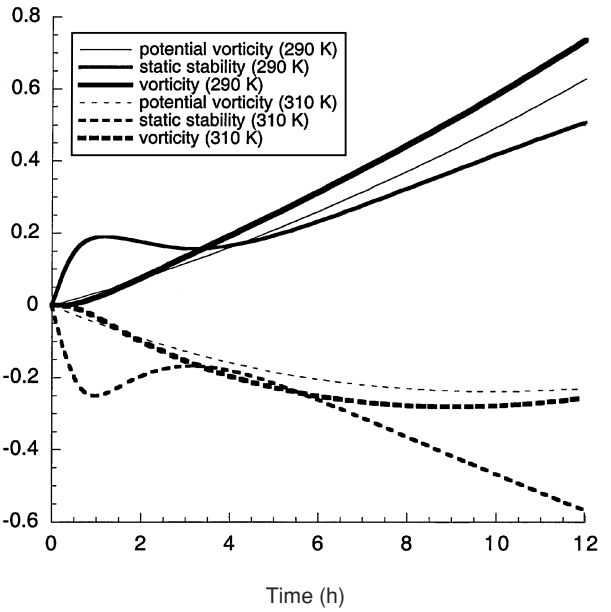


Figure 4.45. Perturbations (deviations from the initial state) as a function of time of potential vorticity in PV units (1 PVU $= 10^{-6}$ K m^2 kg^{-1} s^{-1}), static stability in units of 10^{-4} K Pa^{-1} and vorticity in units of 10^{-4} s^{-1} at two points, i.e. at $r = 19$ km, $\theta = 290$ K (solid lines) and at $r = 19$ km, $\theta = 310$ K (broken lines). The heating is specified as in Figure 4.48.

Rossby height associated with an imposed anomaly with a horizontal scale of 200 km (i.e if we specify $r^* = 100$ km), implying that the balanced response to the heating will have approximately the same vertical scale as the vertical scale of the heating itself.

The closed set of Eqns. (4.30) and (4.31) is approximated by finite differences on a grid within a domain defined by $280 \leq \theta \leq 450$ K and $0 \leq r < 6215$ km. The pressure at the Earth's surface is diagnosed from $\sigma(r, \theta)$, which is a prognostic variable, assuming that $\theta_s(r)$ is constant and that the pressure at the highest isentropic level is constant.

Figure 4.45 shows the evolution in time at $r = 19$ km of the perturbations in, respectively, isentropic potential vorticity, static stability and relative vorticity at, respectively, $(r, \theta) = (19\,\text{km}, 290\,\text{K})$ (below the level of maximum heating) and $(r, \theta) = (19\,\text{km}, 310\,\text{K})$ (above the level of maximum heating). At both points we see an initial oscillation in the static stability associated with the excitation of a Lamb wave as well as internal gravity-inertia waves. Both the potential vorticity and the vorticity increase below the level of maximum heating while they decrease above the level of maximum heating.

These results are as expected, but are not as obvious as they may seem at first sight. A close look at the budget of potential vorticity makes this clear. The budget of isentropic potential vorticity at a fixed point in r-θ space is made up of the following terms (Eqn. 4.34):

$$\text{term 1} = Z_\theta \left(\frac{\partial \dot{\theta}}{\partial \theta}\right); \text{term 2} = -\frac{1}{\sigma}\frac{\partial u}{\partial \theta}\left(\frac{\partial \dot{\theta}}{\partial r}\right);$$

$$\text{term 3} = -\dot{\theta}\left(\frac{\partial Z_\theta}{\partial \theta}\right); \text{term 4} = -v\left(\frac{\partial Z_\theta}{\partial r}\right) \tag{4.54}$$

The sum of terms 1, 2 and 3 is called the diabatic PV forcing (Edouard *et al.*, 1997). Term 4 represents forcing due to advection of potential vorticity along isentropes and is called adiabatic PV forcing. Term 3 represents vertical advection of potential vorticity in a reference frame with potential temperature as a vertical coordinate. In this coordinate system a heat source appears as a vertical velocity which advects potential vorticity and angular momentum. Actually (i.e. in physical space), there is no vertical advection (i.e. vertical transport of air parcels) due to heating: rather, isentropes descend.

Figure 4.46 shows the evolution of the total budget of potential vorticity as well as that of the individual terms listed in Eqn. (4.54) at, respectively, $(r, \theta) = (19\,\text{km}, 290\,\text{K})$ (below the level of maximum heating) and $(r, \theta) = (19\,\text{km}, 310\,\text{K})$ (above the level of maximum heating). As expected, term (1) is dominant at both isentropic levels. In fact, since this term is proportional to Z_θ, and because positive and negative perturbations in Z_θ are formed at, respectively, lower and upper levels, it becomes more effective. The effects of terms 2 and 4 are negligible, while term 3 is negative at both levels. This is due to the basic positive vertical gradient of Z_θ. However, because a low-level positive perturbation in Z_θ, centred below 310 K, is created by the heating, term 3 changes sign at this level after about 4 h of steady heating. In the lower troposphere term 3 remains negative because low values of Z_θ are advected diabatically towards higher values of θ. We will see in the next section that term 3 may dominate the local budget of potential vorticity in a cyclone of realistic intensity.

The negligible amplitude of the only adiabatic PV forcing term (term 4) implies that the potential vorticity perturbation, induced by the heating, hardly propagates away from the source region. The robustness of this remarkable implication can be tested by repeating the numerical experiment with the heating intensity increased 12-fold, but turned off after 1 h. In that case the heating induces an approximately identical potential vorticity

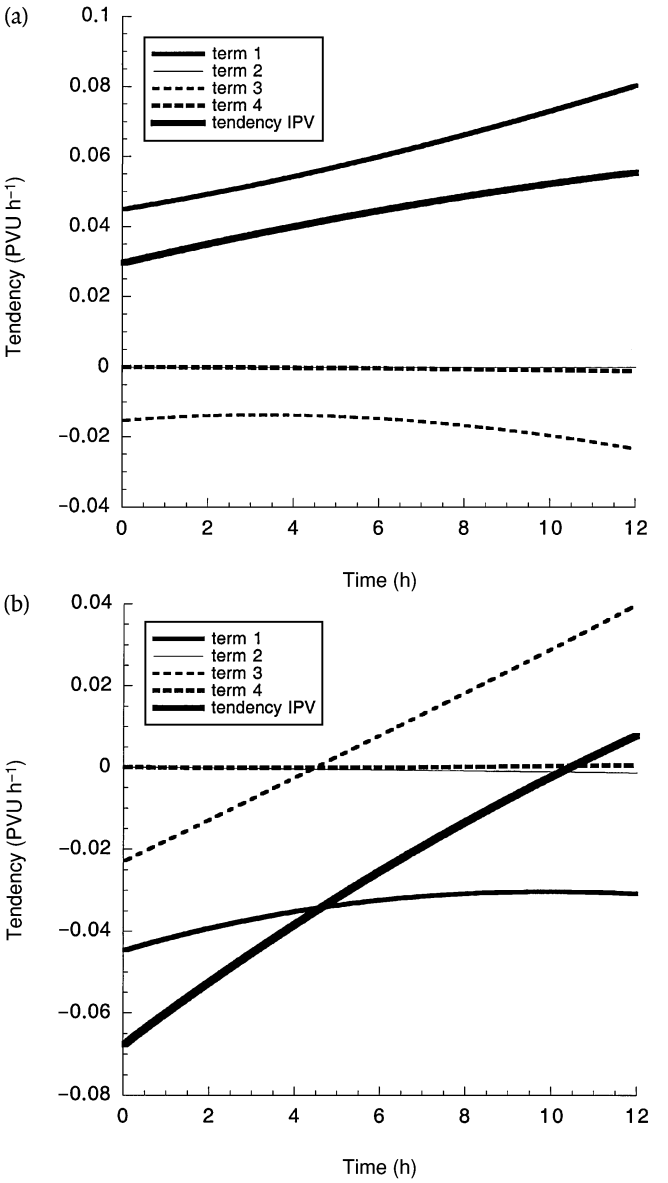


Figure 4.46. Contribution as a function of time of terms listed in Eqn. (4.54) to the total local tendency of the isentropic potential vorticity (also shown) at (a) $r = 19$ km, $\theta = 290$ K, and (b) $r = 19$ km, $\theta = 310$ K.

perturbation in 12 h, but also excites waves of significantly larger amplitude (see Figure 4.47). Despite this, the induced potential vorticity perturbation after 12 h is approximately identical (Figure 4.48). The induced vorticity after 12 hours is also practically identical in both cases, despite

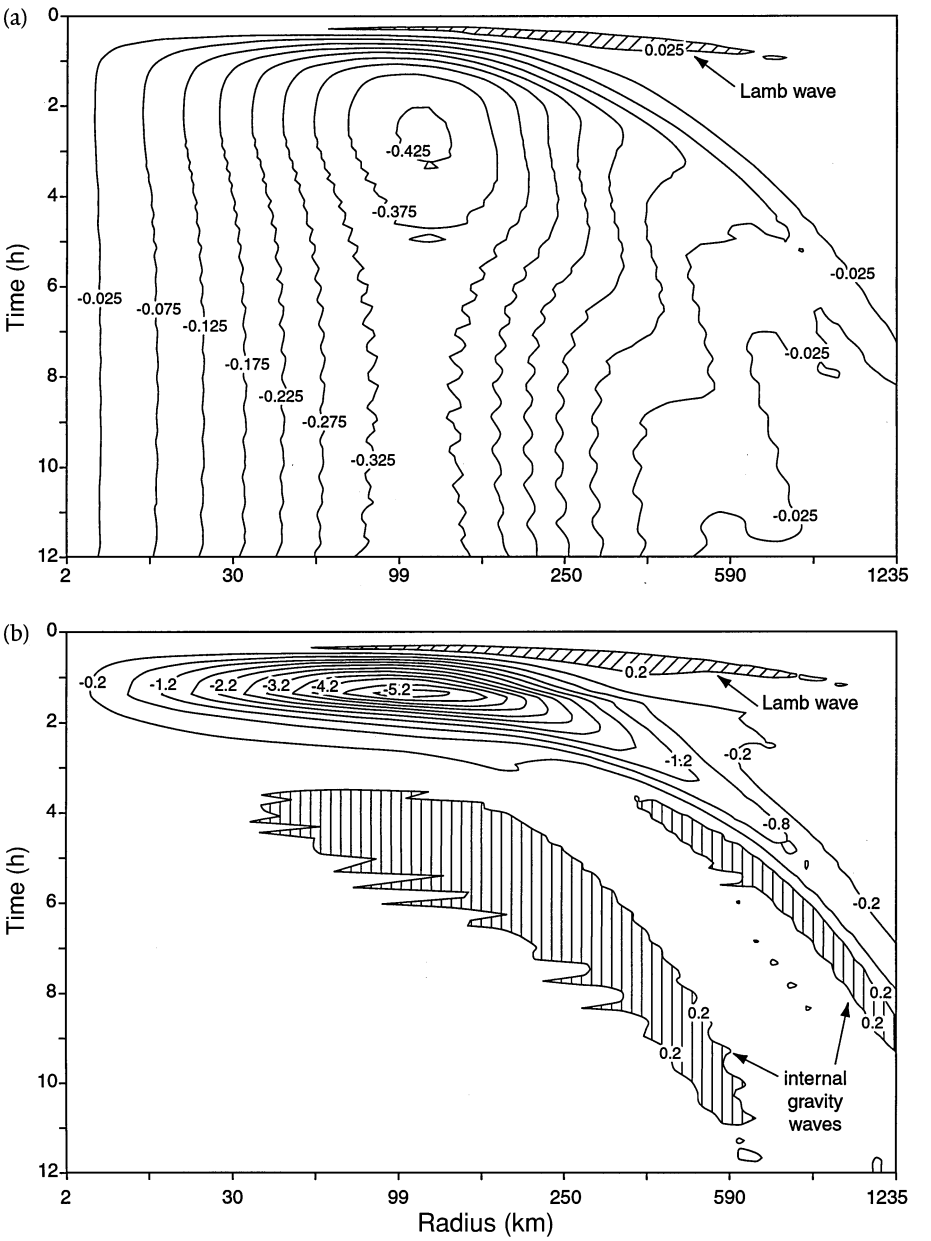


Figure 4.47. Radial velocity at $\theta = 290$ K (c. 2500 m) as a function of time and radial distance in, respectively, case (a) (gentle, slow forcing) and case (b) (fast, intense forcing). In both cases the initial condition (rest) is identical and the final states are practically identical (a weak vortex). Labels indicate values in m s^{-1}. Hatched areas indicate areas where the radial velocity, $v > 0.0025$ m s^{-1} (a) or $v > 0.2$ m s^{-1} (b).

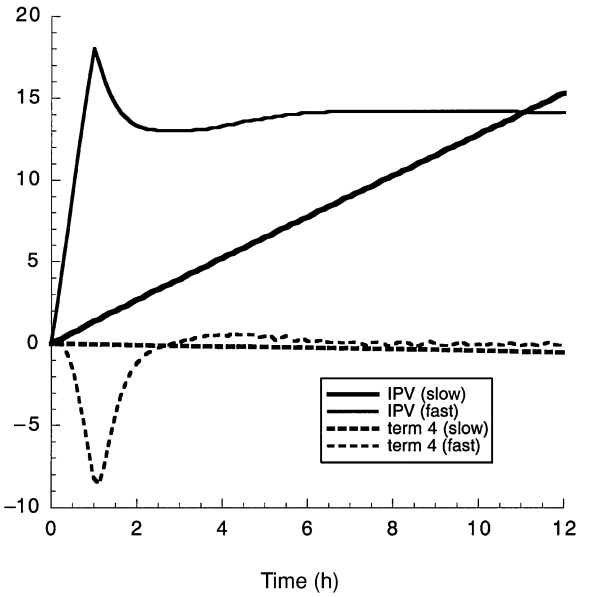


Figure 4.48. Perturbation (deviation from the initial state) as a function of time of isentropic potential vorticity in units of 0.01 PVU (solid lines), and the contribution of term 4 to the total budget of potential vorticity in units of 0.01 PVU h^{-1} (broken lines), at $r = 98$ km, $\theta = 290$ K for, respectively, the experiments with intense forcing (indicated by 'fast') and gentle forcing (indicated by 'slow').

significant departures from gradient wind balance in the case of intense forcing (Figure 4.49).

Figure 4.49 reveals the presence of regular oscillations around gradient wind balance with decreasing frequency, implying that waves with progressively lower frequencies are left in the centre of the cyclone, while the higher frequency waves propagate outwards. In this way the balanced state will eventually be reached.

From these numerical experiments we conclude that the waves and oscillations are of little or negligible importance to the balanced dynamics (represented by the potential vorticity)! This agrees with the results obtained earlier with the shallow water equations (Figure 4.40). Presumably this is the basic theoretical foundation of the success of Ooyama's (1969) balanced model in simulating the life-cycle of a tropical cyclone. In other words, a balanced model, which filters out all motions associated with waves and hydrostatic instability, is sufficient to describe the dynamics of the growth of a cyclone by heating, even quantitatively. Of course, problems and controversy arises (Emanuel *et al.*, 1994; Stevens *et al.*, 1997; Smith, 1997) when one tries to link the heating to the motion, because, for this knowledge of the unbalanced (convective) motion is

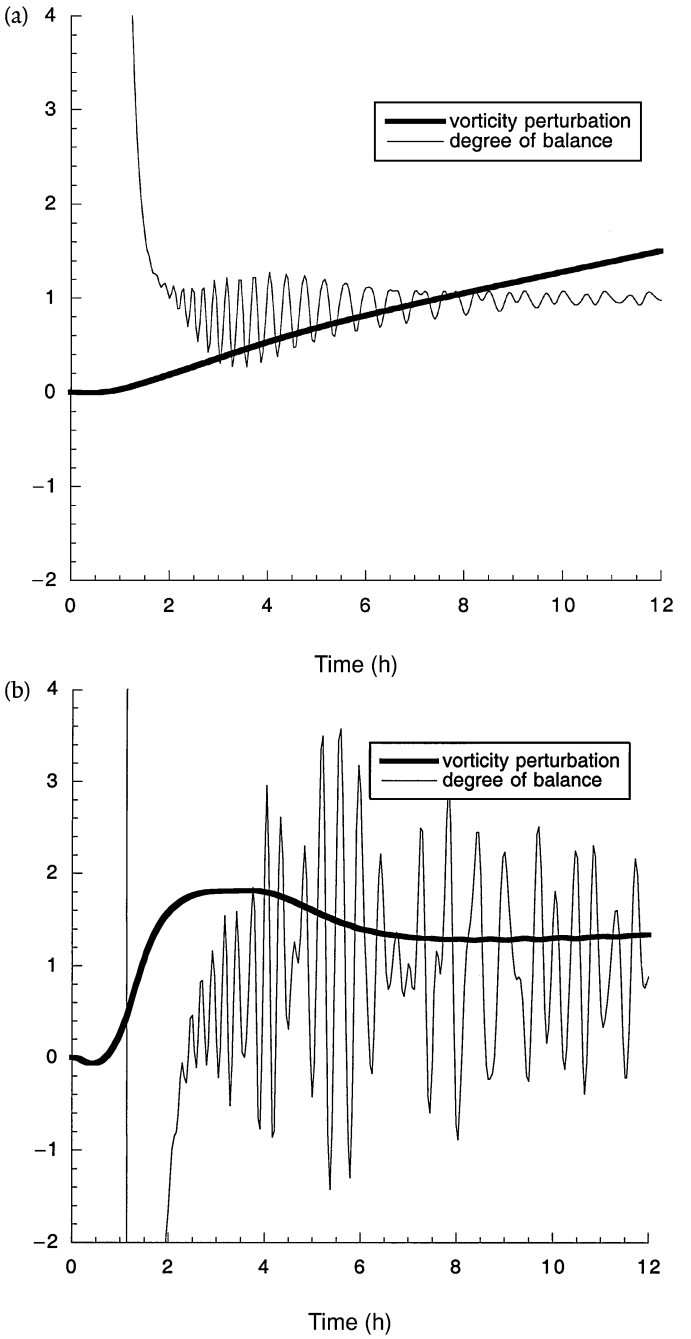


Figure 4.49. Perturbation (deviation from the initial state) as a function of time of the relative vorticity in units of 10^{-5} s^{-1} (thick solid lines) and the degree of balance (thin solid lines), defined as the ratio $\partial\psi/\partial r/[u(f + u/r)]$ (see Eqn. 4.39), at $r = 98$ km and $\theta = 290$ K for, respectively, the experiments with gentle forcing (a) and intense forcing (b). The vortex is in perfect gradient wind balance if the degree of balance is equal to 1. Note that, despite large departures from balance in case (b), the relative vorticity at $t = 12$ h is approximately equal in both cases.

presumably required. A further discussion of this problem will be given in the section 'Processes controlling the heating in a cyclone'.

4.5.9 The intensification of cyclones by heating

On the basis of what we learned from the experiments described in the previous section, we now expect that a warm core cyclone (Figure 4.42b) will react differently to heating than a cold cyclone (Figure 4.42a). In this section we investigate this further by taking the two examples shown in Figure 4.42 and imposing the axisymmetric heating distribution shown in Figure 4.44.

Figure 4.50 shows the isentropic potential vorticity and azimuthal flow as a function of r and θ after 12 h of steady heating for the two cases; one case (case a) with the cold core vortex of Figure 4.42a as initial condition and the other case (case b) with the warm core vortex of Figure 4.42b as initial condition.

It should be stressed that the spatial distribution and intensity of the heating is identical in both experiments. It should also be stressed that these results are meant to illustrate and provide insight into the basic physical mechanism of cyclone growth by heating. The actual heating distribution in a cyclone, such as a polar low, will be discussed further in Section 4.5.11, 'Processes controlling the heating in a cyclone'.

The most remarkable result (Figure 4.50) is that the potential vorticity in the warm core cyclone increases significantly, while the potential vorticity in the cold core cyclone decreases. Associated with this, the maximum azimuthal wind velocity in the cold core cyclone decreases from 20 m s^{-1} to 18.7 m s^{-1}. In the warm core cyclone, on the other hand, the maximum azimuthal wind velocity increases from 10 m s^{-1} to 13.4 m s^{-1}.

The increase in potential vorticity in the centre of the warm core cyclone in this reference frame is again principally due to the effect of terms 1 and 3. This is shown in Figure 4.51.

Term 2 hardly contributes to the local potential vorticity budget. The absolute value of this term initially does not exceed 0.05 PVU per hour. The absolute value of term 4 is zero initially and in both cases never exceeds 0.05 PVU per hour during the 12-hour integration. In this context, it must be noted that the second term in Eqn. (4.54) could become important in a balanced warm core cyclone (with $\partial u/\partial\theta < 0$) if the cyclone possesses an eye with most of the (latent) heating taking place near the radius of maximum wind, as is frequently observed.

In a cold core cyclone, heating is seen to create an intense negative potential vorticity anomaly centred at $r = 0$ and $\theta = 320$ K. Comparatively weak positive potential vorticity tendencies are found outside the core of the vortex, implying that the cold core cyclone will expand due to heating.

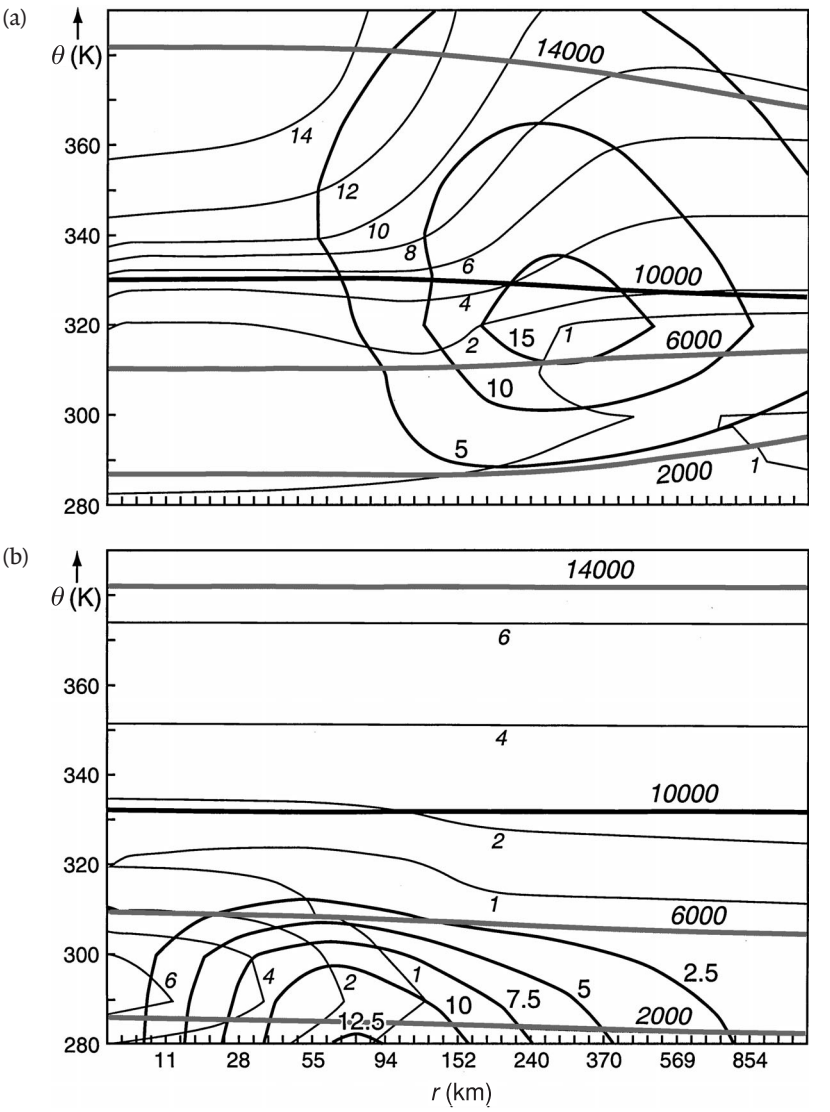


Figure 4.50. Isentropic potential vorticity (labelled in small italic numbers in PV units; 1 PVU is 10^{-6} K m^2 kg^{-1} s^{-1}), height (labelled in large italic numbers in m) and azimuthal velocity (labelled in m s^{-1}) as a function of potential temperature and radius, after 12 h constant heating according to Figure 4.44, for a case (a) with the cold core vortex of Figure 4.42 as initial condition, and for a case (b) with the warm core vortex of Figure 4.42 as initial condition.

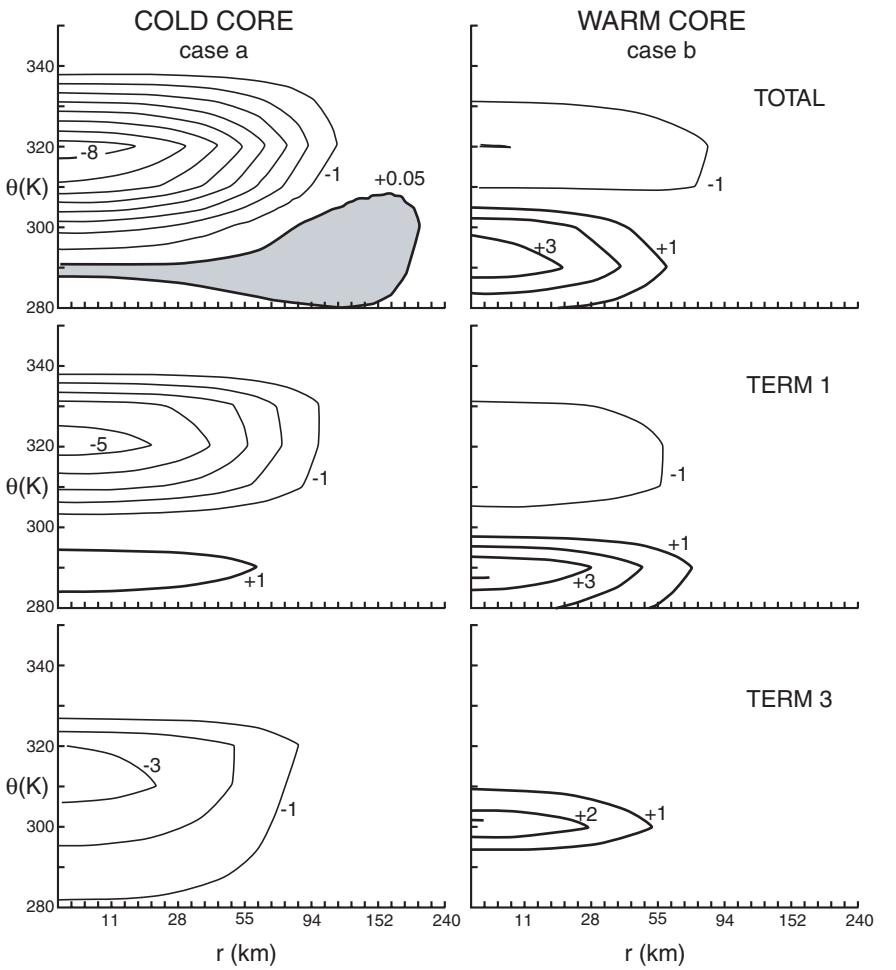


Figure 4.51. Contribution as a function of potential temperature and radius of terms listed in Eqn. (4.54) to the local tendency of the isentropic potential vorticity (Eqn. 4.34) at $t = 0$. The top diagrams show the total tendency for, respectively, the cold core cyclone (case a) and the warm core cyclone (case b). The middle diagrams show the contribution of term 1, while the diagrams below show the contribution of term 3. Labels are in units of 0.05 PVU h^{-1}. Thick contours represent positive values; thin contours represent negative values.

Thus, tropospheric heating in the centre of a warm core cyclone produces a relatively intense positive anomaly of the potential vorticity at low levels and a relatively weak anomaly (positive and negative) at upper levels (a similar result was obtained by Schubert and Alworth, 1987), while tropospheric heating in the centre of a cold core cyclone is likely to produce a negative potential vorticity anomaly at all heights (a similar result was obtained by Wirth, 1995).

Therefore, assuming that the invertibility principle is valid at all times, we may conclude that, in general, the intensity of a warm core cyclone will increase due to heating while the intensity of a cold core cyclone will decrease due to heating. Similar conclusions were reached by Emanuel and Rotunno (1989) and van Delden (1989).

An additional understanding of the mechanism by which the vortex intensifies (i.e. the vorticity increases) can be gained by observing the changes in the angular momentum distribution, since changes in this can easily be translated to changes in the absolute vorticity distribution using Eqn. (4.36), and since angular momentum is materially conserved (even in the presence of heating: Eqn. 4.33). Therefore, these changes are governed only by advection (horizontal and vertical in $r - \theta$ space). Thus, assuming vorticity (or the radial gradient of angular momentum) is a measure of cyclone intensity, following the movement of isopleths of angular momentum and noting where isopleths of angular momentum converge or diverge radially provides insight into the process of cyclone intensity changes.

Figure 4.52 gives an impression of the movement of isopleths of angular momentum in the two examples discussed in this section. The radial diplacements (25 km at the most) of air parcels during a 12 h period of constant heating are small compared to the total radial scale of the cyclone. It is obvious that these small displacements will have very little effect on the potential vorticity anomaly induced by the heating.

In the warm core cyclone the static stability-related potential vorticity perturbation caused by the diabatic PV forcing is partially converted into vorticity related potential vorticity by inward radial advection of high values of the angular momentum and consequent concentration of isopleths of angular momentum.

In the cold core cyclone the situation is very different. The negative perturbation in potential vorticity is accommodated by upward and outward radial advection of angular momentum throughout the whole cyclone, except at low levels near the radius of maximum wind. Note that even above the heat source (well into the stratosphere) there is a displacement of isopleths of angular momentum. Clearly, the cold core cyclone expands as a result of heating in the centre. This is the consequence of the theorem expressed in Eqn. (4.38), which states that a negative (potential) vorticity anomaly induced by internal heating must be 'compensated' in an integral sense, by a positive (potential) vorticity anomaly elsewhere (remember: we did not impose heating at the boundaries). The opposite is of course also true, implying that a warm core cyclone will contract as a consequence of internal heating.

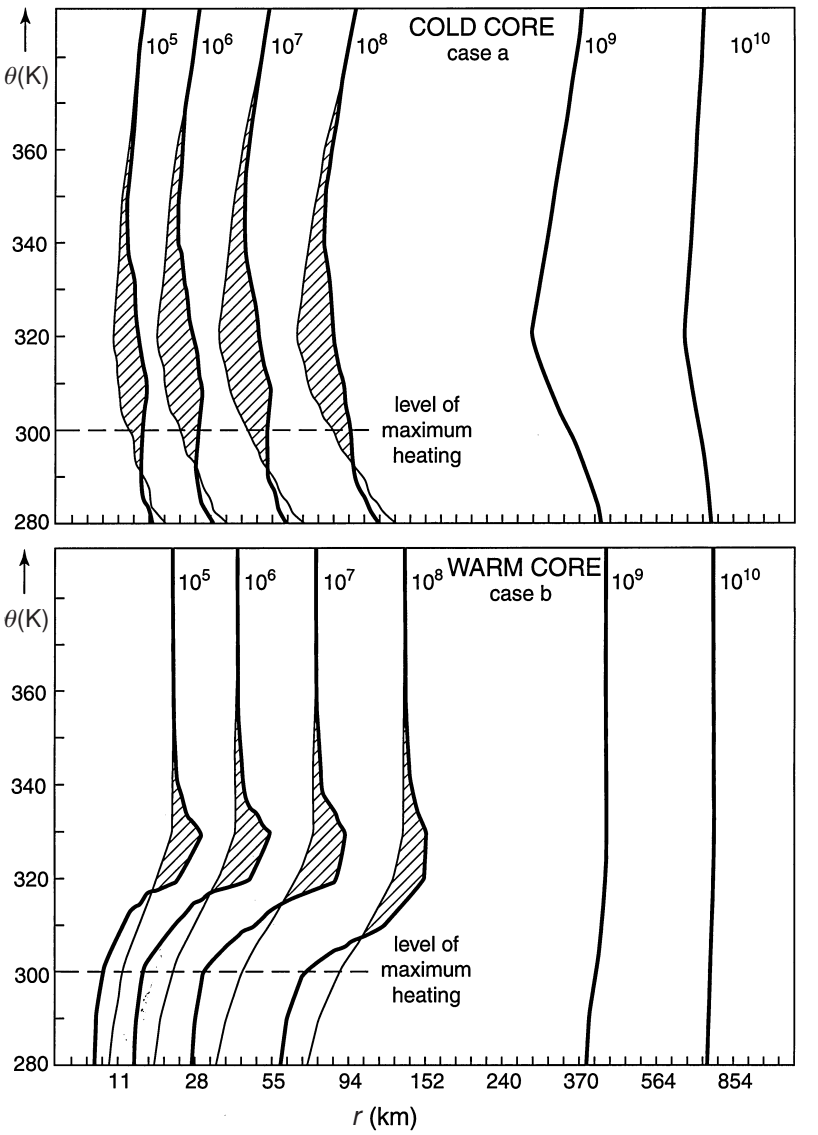


Figure 4.52. Angular momentum per unit mass (labelled in units of $m^2\ s^{-1}$) as a function of potential temperature and radius in case a (above) and in case b (below), initially (thin lines) and after 12 h (thick lines).

In reality, cyclones do not have such a simple potential vorticity structure as illustrated in Figure 4.42. For example, there may be a warm core structure embedded in a larger-scale cold core structure. Indeed, many polar lows develop near or at the centre of a deep, cold core, cut-off cyclone aloft (Businger, 1985; Rasmussen, 1985a). Although cold air aloft is of course favourable for the

development of deep convection and attendant intense latent heating, it is clear now that this will not automatically lead to the growth of a polar low embedded in the cold core cyclone. Of course, if heating continues for a sufficiently long time, a shallow warm core system will eventually develop, embedded within the cold core system, and the chances of producing a positive potential vorticity anomaly due to heating will increase. For example, if we add the azimuthal wind field shown in Figure 4.42a to the azimuthal wind field shown in Figure 4.42b, we are effectively embedding a small scale warm core cyclone in a larger-scale cold core cyclone. The resulting distribution of isentropic potential vorticity is shown in Figure 4.53a.

This potential vorticity distribution does not differ greatly from the distribution shown in Figure 4.42a, except that the $Z_\theta = 2$ PVU isopleth is depressed. However, the effect of a heat source as specified by Figure 4.44 is very different. Term 1 is now much more effective as a source of low-level potential vorticity, while term 3 is less effective as a sink of potential vorticity (Figures 4.53b and 4.54).

Emanuel and Rotunno (1989) performed similar numerical experiments and showed that the growth by heating of a warm core vortex embedded (at low levels) in a tropospheric cold core vortex is possible if the warm core vortex already has a sufficiently large amplitude. Figure 4.55 shows the central surface pressure as a function of time for the three 12 h integrations discussed in this section.

4.5.10 Physical interpretation of Ooyama's balanced model

So how far is the physics outlined in the previous sections represented in the theory of CISK (or cooperative intensification theory) put forward by Ooyama (1969)? We will investigate the answer to this question in this section employing Ooyama's (1969) tropical cyclone model.

Ooyama's tropical cyclone model consists of three axisymmetric superposed incompressible layers (see Figure 4.56). The density of the boundary layer is equal to the density of the layer just above, while the density of the upper layer is a factor ε smaller (i.e. $\varepsilon < 1$). The thickness of the lower layer, i.e. the boundary layer, is assumed constant. Frictional convergence in this layer will give rise to a mass flux into the layer above. The thickness of the upper two layers may vary due to convergence or divergence, or mass fluxes between the layers. Due to the incompressibility, compression waves are not possible. Gravity-inertia waves are not permitted either, because the flow is assumed to be in thermal wind balance permanently. Disturbances to thermal wind balance are accommodated instantaneously and continuously by means of a radial flow. This radial flow is responsible for cyclone intensity changes.

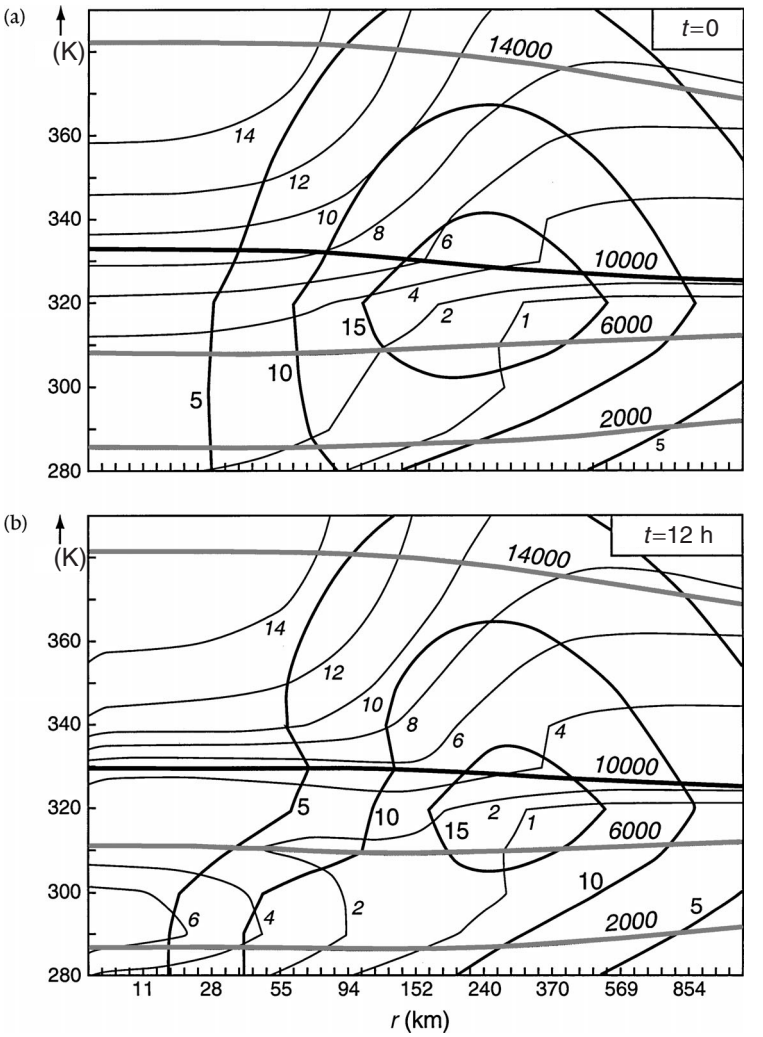


Figure 4.53. Isentropic potential vorticity (labelled in small italic numbers in PV units; 1 PVU $= 10^{-6}$ K m^2 kg^{-1} s^{-1}), height (labelled in large italic numbers in m) and azimuthal velocity (labelled in m s^{-1}) at (a) $t = 0$ and (b) $t = 12$ h as a function of potential temperature and radius, after 12 h of constant heating according to Figure 4.44, for a small-scale warm core vortex embedded in the centre of a large-scale cold core vortex.

The model contains no explicit thermodynamics. How then are the effects of temperature changes due to heating incorporated? The answer to this question nicely illustrates the concept of *parameterization*. Heating in Ooyama's model is represented as a *mass flux* from layer 1 to layer 2, denoted by "Q" in Figure 4.56. Due to this, the thickness of the upper layer grows at the cost of the thickness of

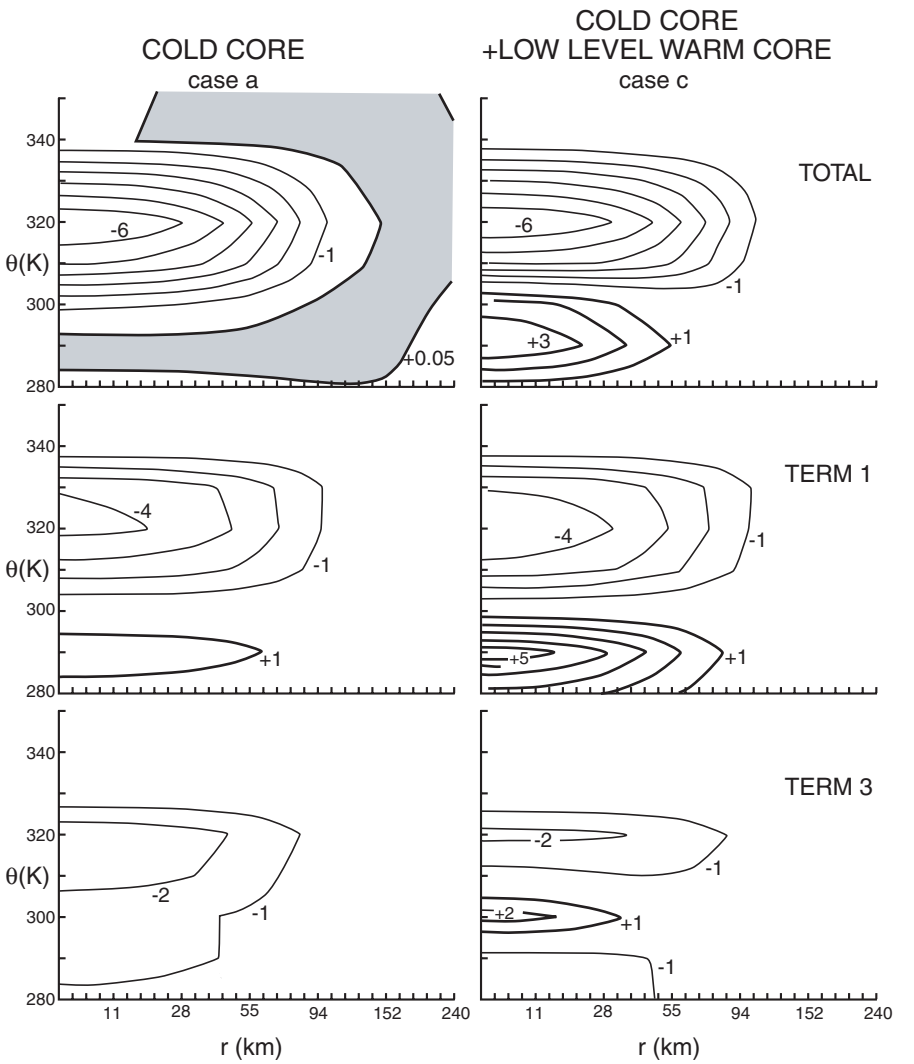


Figure 4.54. Contribution as a function of potential temperature and radius of terms listed in Eqn. (4.54) to the local tendency of the isentropic potential vorticity (Eqn. 4.34) at $t = 3$ h. The top diagrams show the total tendency for, respectively, the cold core cyclone (case a) and the warm core cyclone embedded in the cold core cyclone (case c). The middle diagrams show the contribution of term 1; the lower diagrams show the contribution of term 3. Labels are in units of 0.05 PVU h^{-1}. Thick contours represent positive values; thin contours represent negative values.

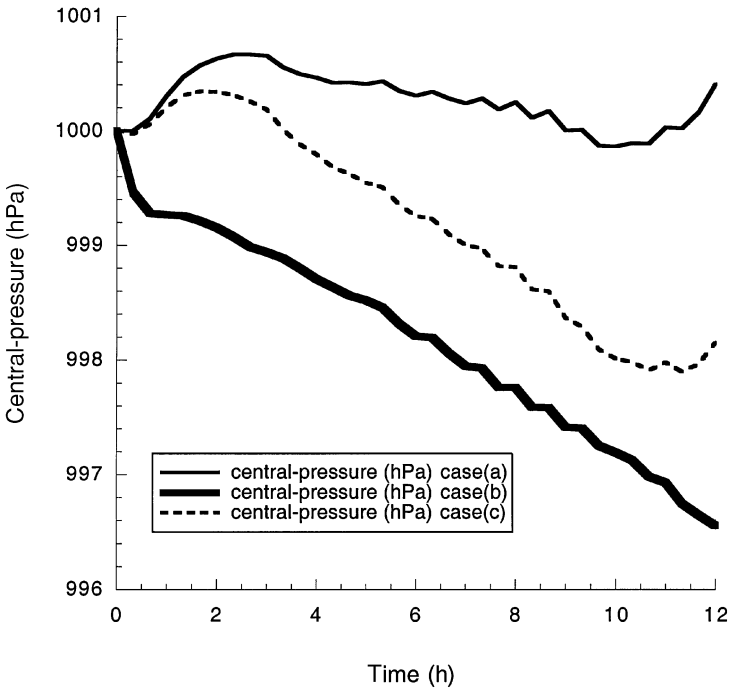


Figure 4.55. Central surface pressure as a function of time for cases a (initial cold core), b (initial warm core) and c (initial warm core embedded in a cold core vortex).

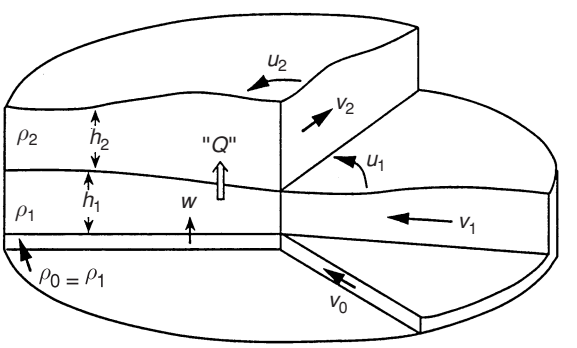


Figure 4.56. Structure of Ooyama's (1969) model (see text for further explanation).

the lower layer. Thus, the pressure in the upper layer increases while the pressure at the Earth's surface remains unchanged. Therefore, the mass flux, "Q", crudely mimics the effect of heating and hydrostatic adjustment in the vertical on the pressure distribution in the atmosphere (Lamb's problem, discussed earlier). This relatively simple parameterization of heating also crudely mimics the effect of diabatic potential vorticity forcing (the effect of term 1 in Eqn. 4.54)

in a warm core cyclone. The potential vorticity for each layer is defined according to Eqn. (4.25). Therefore a mass flux from the lower layer to the upper layer represents a source of potential vorticity in the lower layer and a simultaneous sink of potential vorticity in the upper layer.

Ooyama (1969) assumed that "Q" was proportional to the upward velocity, w, at the top of the boundary layer by

$$\text{“}Q\text{”} = \eta w^{+}, \quad w^{+} = wi \; > 0_{w^{+}} = i \leq 0 \tag{4.55}$$

where η is a function of the latent and sensible heat fluxes from the sea, which, in turn, were assumed to be proportional to the wind speed in the boundary layer. The vertical velocity at the top of the boundary layer was the result of frictional convergence, which increased as the intensity of the azimuthal balanced flow in the free atmosphere increased. Therefore, the balanced vortex controlled the heating or forcing. Hence, we have a feedback loop involving continuous, perfect adjustment to thermal wind balance of a vortex to self-induced disturbances to this state of balance. Ooyama (1969) demonstrated that this feedback loop could become unstable if $\eta > 1$, i.e. if the diabatic mass flux functions as a source of potential vorticity in the lower layer of the (model) free atmosphere.

Ooyama (1969) and many others thereafter interpreted the condition, $\eta > 1$ as a criterion for the existence of conditional instability or positive CAPE. This led to the paradigm: 'CAPE is required for CISK'. Aside from the question whether the condition, $\eta > 1$, can be interpreted in this way in a hydrostatic model consisting of two incompressible fluid layers, it is clear, in the light of the theory discussed in earlier sections, that this interpretation is not needed as an explanation of the intensification. The physical interpretation that does justice to the physical content of Ooyama's model is that the vortex intensifies because the diabatic mass flux induces a positive potential vorticity perturbation at low-levels. This intensification is demanded by the invertibility principle. Relatively few authors have brought potential vorticity into the discussion. One of the early exceptions was Schubert *et al.* (1980), who concluded (p. 1482) that 'the effects of clouds on the potential vorticity field (i.e. the apparent potential vorticity source) is the important ingredient in understanding the feedback of clouds on the large scale fields.'

During the past 10 years more authors have recognized this fact. For example, Montgomery and Enagonio (1998) have taken an approach whereby the vorticity budget of an ensemble of cumulus clouds is parameterized in a quasi-geostrophic model by potential vorticity anomalies having a horizontal scale of approximately 200 km. Both in Ooyama's model and in the approach taken by Montgomery and Enagonio, vertically-confined, mid-tropospheric heating is

seen to create a positive potential vorticity anomaly at low-levels and a negative potential vorticity anomaly aloft. According to the analysis presented in earlier sections, the effect of heating on the potential vorticity is more subtle.

4.5.11 Processes controlling the heating in a cyclone

Can the balanced flow indeed 'take control' of the heating to produce a positive feedback loop between the intensity of the balanced flow and the heating? In other words, how does the upper part of the feedback loop shown in Figure 4.30 work? This question has occupied many researchers since the pioneering papers by Charney and Eliassen (1964) and Ooyama (1969). Basically, two mechanisms have been suggested. As mentioned in Section 4.5.2 and explained in the previous section, Charney and Eliassen (1964) and Ooyama (1969) suggested that frictional convergence in the boundary layer could force moisture (transferred from the sea over a large surrounding area) in and up the core of the vortex leading to condensation of water vapour and, higher up, freezing of cloud water. Both condensation and freezing are accompanied by latent heat release. Thus, they emphasized latent heat release as the principal heat source responsible for the cyclone intensity changes. Ooyama (1969) tried to improve the physical reasoning given by Charney and Eliassen and extended CISK theory to the finite amplitude, non-linear regime. Shapiro and Willoughby (1982), Schubert and Hack (1982) and van Delden (1989) showed that the feedback loop, suggested by Ooyama, was most effective in an intense baroclinic (warm core) vortex. Nevertheless, Emanuel (1986a) called into question this mode of thinking. The essence of Emanuel's criticism is summarized in the following quote from the abstract of Emanuel *et al.* (1994):

> The dominant thinking about the interaction between large-scale atmospheric circulations and moist convection holds that convection acts as a heat source for the large-scale circulations, while the latter supply water vapour to the convection. We show that this idea has led to fundamental misconceptions about this interaction, and offer an alternative paradigm, based on the idea that convection is nearly in statistical equilibrium with its environment. According to the alternative paradigm, the vertical temperature profile itself, rather than the heating, is controlled by the convection, which ties the temperature directly to the sub-cloud-layer entropy. The understanding of large-scale circulations in convecting atmospheres can therefore be regarded as a problem of understanding the distribution in space and time of sub-cloud-layer entropy. We show that the sub-cloud-layer entropy is controlled by the sea surface temperature, the surface wind and the large scale vertical velocity in the convecting layer.

In other words, the 'alternative paradigm' states that convection acts to constrain the temperature profile towards the moist adiabat (see Figure 4.31). This process is called 'convective adjustment'.

Convective adjustment can be incorporated into the calculations described in Sections 4.5.8 and 4.5.9 if we modify Eqn. (4.31c) as follows:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = -\sigma\left(\frac{1}{r}\frac{\partial rv}{\partial r}\right) - \lambda\left(\sigma - \sigma_{\mathrm{ref}}\right) \tag{4.56}$$

where λ is a relaxation constant determining the time scale of adjustment of the static stability $[-\partial\theta/\partial p] = 1/g\sigma$ of a model layer towards a reference value of the static stability, $[-\partial\theta/\partial p]_{\mathrm{ref}} = 1/g\sigma_{\mathrm{ref}}$.

The reference thermodynamic state, represented by the parameter, σ_{ref}, depends on the cloud top and is different for deep (precipitating) and shallow (non-precipitating) convection. If the atmospheric layer is conditionally unstable and saturated this reference state could be represented by the moist adiabat, i.e. $\sigma_{\mathrm{ref}} = \sigma_{\mathrm{s}}$. However, in many model studies a slightly different reference state is adopted in order to correct for effects of cloud water on buoyancy (Betts, 1986). Thus, the problem of convective adjustment consists of determining the value of λ, the value of σ_{ref} as well as finding a criterion to determine whether the adjustment scheme is activated (when it is not activated we may set $\lambda = 0$). This requires an equation for conservation of water vapour in order to determine whether the air at a particular point is supersaturated.

Eqns. (4.56) and (4.31c) imply that

$$\frac{\partial\dot{\theta}}{\partial\theta} = \frac{\lambda}{\sigma}\left(\sigma - \sigma_{\mathrm{ref}}\right) \tag{4.57}$$

This equation can be integrated from the Earth's surface upward with the heating prescribed at the Earth's surface, yielding the heating as a function of r and θ. The heating parameterization problem in this theory, therefore, is linked directly *to the problem of parameterizing the surface heat flux*. In this context, there are two effects that may provide a 'control' as implied in Figure 4.30. First, the intensity of the heat and moisture transfer from the sea surface to the atmosphere is proportional to the wind speed just above the sea surface. Hence, as the wind speed increases, so does the heat transfer, which in turn induces a further intensification of the wind speed and so on. Second, the decrease of the surface pressure in the core of the cyclone is accompanied by a decrease in the air temperature (owing to adiabatic expansion). This leads to a temperature gradient between the sea surface and the air just above. Since the heat transfer is also proportional to the temperature difference between the sea surface and the air just above, this effect will further enhance the heat transfer.

If true, the idea of 'equilibrium control' is very attractive as a mode of thinking about the interaction of cumulus convection with the balanced motion, because it ties the heating to a simple, physically plausible and observationally well documented constraint. Another clear advantage is related to the fact that the convective heating is not related in any way to the divergent part of the flow, which makes it independent of exact knowledge of the 'unbalanced' part of the motion.

It must be remarked, however, that in cases of cold air outbreaks over the oceans in polar regions, the stratification in the lower half of the atmosphere can sometimes be in a state which is very far from the thermodynamic reference equilibrium state. Nevertheless, in principle it is possible to handle this problem by using Eqn. (4.56) and specifying an appropriate value of the relaxation constant, λ.

4.5.12 Travelling upper-level disturbances and cyclogenesis

There is little doubt about the existence in general of strong conditional instability and associated large values of CAPE in regions where, and at times when, polar lows form (Section 4.5.1). Thus, some polar lows may form simply due to heating associated with a sudden release of latent heat associated with CAPE in a limited area of horizontal dimensions comparable to the Rossby radius of deformation. We have seen that if the heat associated with CAPE is released quickly in an area with a low-level positive potential vorticity anomaly, a cyclone may form within a few hours (Figure 4.49). A process that may serve as a 'spatial focusing mechanism' for the release of latent heat is an upper-level potential vorticity anomaly travelling into a region of reduced static stability. Montgomery and Farrell (1992) have investigated this process with application to polar low formation.

By virtue of the invertibility principle, the wind field (or vorticity field) below and above a moving potential vorticity anomaly must undergo changes such that relative vorticity increases in advance of the approaching anomaly, while relative vorticity decreases at the trailing edge of the anomaly. This implies ascending motion in advance of the travelling upper-level anomaly and descending motion at the trailling edge.

This effect can be easily demonstrated with the two-layer model shown in Figure 4.57. This model was adopted by Phillips (1951) and Bretherton (1966) to investigate baroclinic instability. The equations of motion and mass continuity for each layer are similar to Eqns. (4.22a–c), i.e.

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = fv_i - \frac{\partial \phi_i}{\partial x} \tag{4.58a}$$

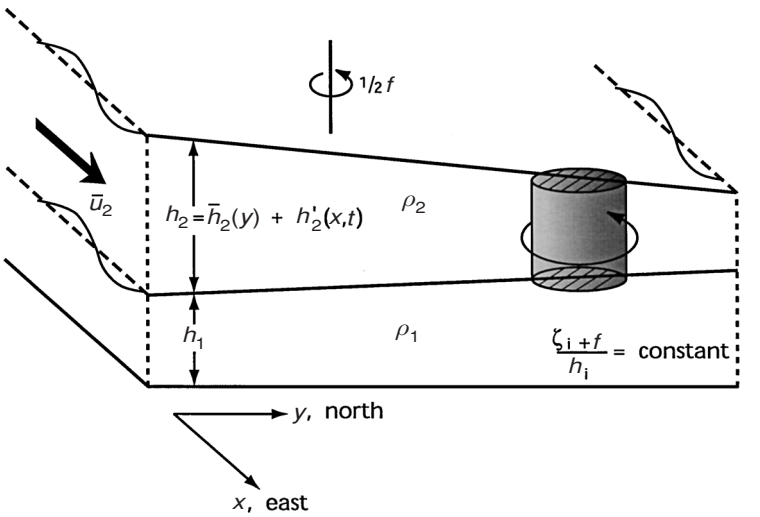


Figure 4.57. Schematic diagram of the two-layer model (see the text for further explanation). The basic potential vorticity gradient (south to north) is positive in the upper layer and negative in the lower layer.

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = f(\bar{u}_i - u_i) \tag{4.58b}$$

$$\frac{\mathrm{d}h_i}{\mathrm{d}t} = -h_i \frac{\partial u_i}{\partial x} \tag{4.58c}$$

with $i = 1$ referring to the lower layer and $i = 2$ referring to the upper layer (Figure 4.57). The geopotential, ϕ_i, defined as

$$\phi_1 = g(h_1 - \bar{h}_1) + \varepsilon g(h_2 - \bar{h}_2) \tag{4.59a}$$

$$\phi_2 = g(h_1 - \bar{h}_1) + g(h_2 - \bar{h}_2) \tag{4.59b}$$

(see Ooyama, 1969). We assume that there is a time-independent meridional potential vorticity gradient aloft, that induces a constant geostrophic velocity, $\bar{u}_i (i = 1, 2)$. In order to have geostrophic balance, this geostrophic velocity is associated with an imposed time-independent gradient in the thickness of the layers. Therefore, the time-mean thickness $\bar{h}_i$ depends only on the meridional coordinate, y. For simplicity, the perturbations are assumed independent of y. Given the potential vorticity distribution, the induced geostrophic meridional flow can be computed by solving the following equation:

$$\frac{\partial^2 v_i}{\partial x^2} - A_i v_i = B_i \tag{4.60}$$

with

$$A_i = \frac{f\zeta_{\text{pot}i}}{g(1-\varepsilon)} \tag{4.61a}$$

$$B_1 = h_1\frac{\partial\zeta_{\text{pot}1}}{\partial x} - \frac{\varepsilon f v_2\zeta_{\text{pot}1}}{g(1-\varepsilon)} \tag{4.61b}$$

$$B_2 = h_2\frac{\partial\zeta_{\text{pot}2}}{\partial x} - \frac{f v_1\zeta_{\text{pot}2}}{g(1-\varepsilon)} \tag{4.61c}$$

and

$$\zeta_{\text{pot}i} = \frac{f+\zeta_i}{h_i} \tag{4.61d}$$

This is the two-layer version of the invertibility principle, analogous to Eqn. (4.26). It is derived from the conditions of geostrophic balance in both layers and from the definition of potential vorticity (4.61d).

Let us assume the existence of a positive potential vorticity anomaly in the upper layer, as shown in Figure 4.58. The physical mechanisms that produce this anomaly are not addressed here. We only assert that polar lows are very frequently observed in the vicinity of such upper-level potential vorticity anomalies. In the lower layer the potential vorticity is constant. The solution of Eqn. 4.60 in terms of the relative vorticity in both layers for this potential vorticity distribution is also shown in Figure 4.58. We see that the potential vorticity anomaly in the upper layer induces a relative vorticity anomaly in the lower layer. The intensity of this relative vorticity anomaly increases with decreasing static stability, represented by the parameter, ε, in this model (Figure 4.59).

If we now set the potential vorticity anomaly in the upper layer into motion (by assuming that $\bar{u}_2 = 30\ \text{m s}^{-1}$; $\bar{u}_1 = 0$), the potential vorticity anomaly will travel eastward (remember that potential vorticity is materially conserved), deforming slightly due to meridional advection of potential vorticity associated with the meridional gradient in the basic state thickness (compare the thick solid line in Figure 4.58 with the solid line in Figure 4.60). The induced vorticity in the lower layer will necessarily also travel eastward. The associated process of adjustment in the lower layer requires convergence in advance of the moving vorticity anomaly (where $\partial\zeta/\partial t > 0$) and divergence behind the moving vorticity anomaly (where $\partial\zeta/\partial t < 0$). This is illustrated in Figure 4.60. Hoskins *et al.* (1985, p. 907) have come up with the following instructive analogy of this process:

> One may think of an eastward-moving upper-air anomaly as acting on the underlying layers of the atmosphere somewhat like a broad very gentle 'vacuum cleaner', sucking air upwards towards its leading portion

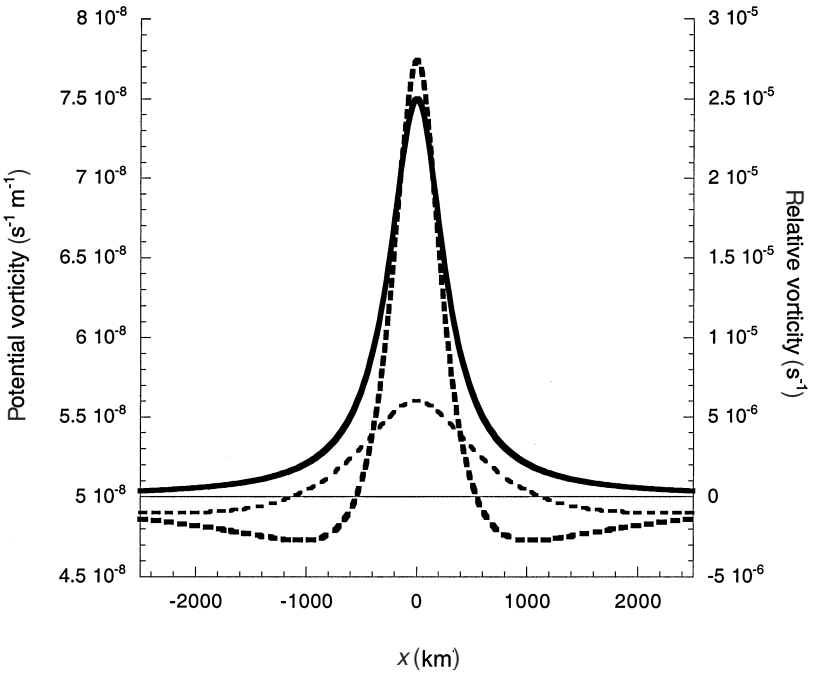


Figure 4.58. Potential vorticity in the upper layer (thick solid line) and in the lower layer (thin solid line) as a function of x, and the induced relative vorticity, according to the invertibility principle (Eqn. 4.60) in both layers (thick broken line: upper layer; thin broken line: lower layer). The potential vorticity anomaly is prescribed using the formula (4.23) (a bell-shaped perturbation) with $a = 300$ km. The static stability parameter, $\varepsilon = 0.8$, $g = 9.81$ m s^{-2}, $f = 0.0001$ s^{-1} and $\bar{h}_i = 2000$ m (for $i = 1, 2$).

and pushing it downwards over the trailing portion. The vertical motion field arises in response to the need to maintain mass conservation and approximate balance. If a potential vorticity anomaly were to arrive overhead without any adjustment taking place underneath it, then the wind, temperature and pressure fields would be out of balance to an improbable extent.

The induced upward motion in advance of the approaching anomaly is very likely to generate precipitating clouds, thereby generating potential vorticity in the lower layers of the atmosphere. On the basis of this idea, Montgomery and Farrell (1992) proposed a conceptual model of polar low development. In the first stage of development, called 'induced self-development', a mobile upper trough initiates a rapid low-level spin-up. This spin-up is especially strong if the trough moves into an area of reduced static stability (see Figure 4.59). A secondary development follows, called 'diabatic destabilization', which is

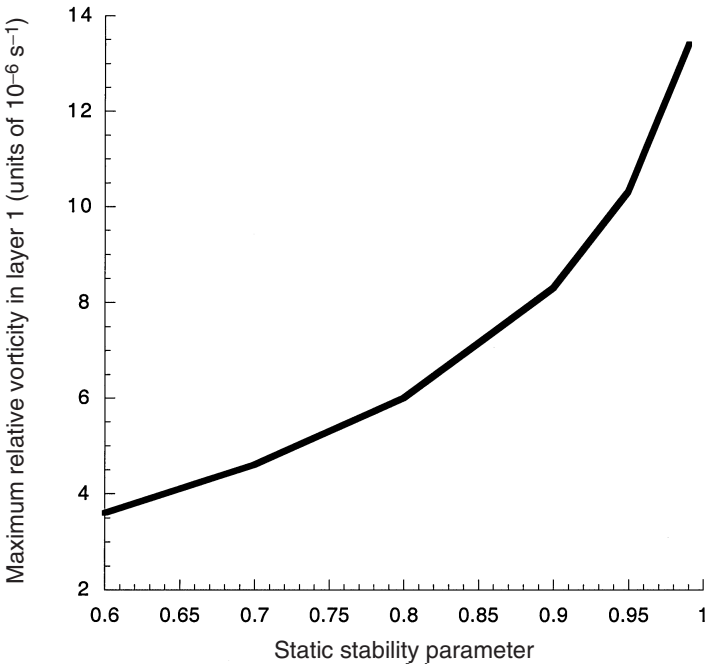


Figure 4.59. Maximum value of the relative vorticity induced in the lower layer by a fixed anomaly in the upper layer (see Figure 4.58) as a function of the static stability parameter, ε. The intensity of the induced vorticity increases with decreasing stability (increasing ε) as was also predicted by Eqn. 4.51.

associated with the production of low-level potential vorticity by heating within clouds forced by the induced upward motion *ahead* of the oncoming upper-level potential vorticity anomaly.

4.5.13 Summary and concluding remarks regarding the role of heating

Although a real polar low is not exactly a balanced flow structure, it is contended that only the 'balanced dynamics' is of importance to understand the intensification and structure of a polar low. Because of this, we can adopt the 'potential vorticity viewpoint' to disentangle from the primitive equations the most important physical principles governing the dynamics of cyclone intensification due to heating. The principle that heating disturbs the state of balance by altering the potential vorticity distribution, the principle of invertibility, which defines the balanced state, and the principle of material conservation of potential vorticity in the absence of friction and heating, offer physical insight into the reasons why a cyclone, such as a polar low, can grow as a result of heating.

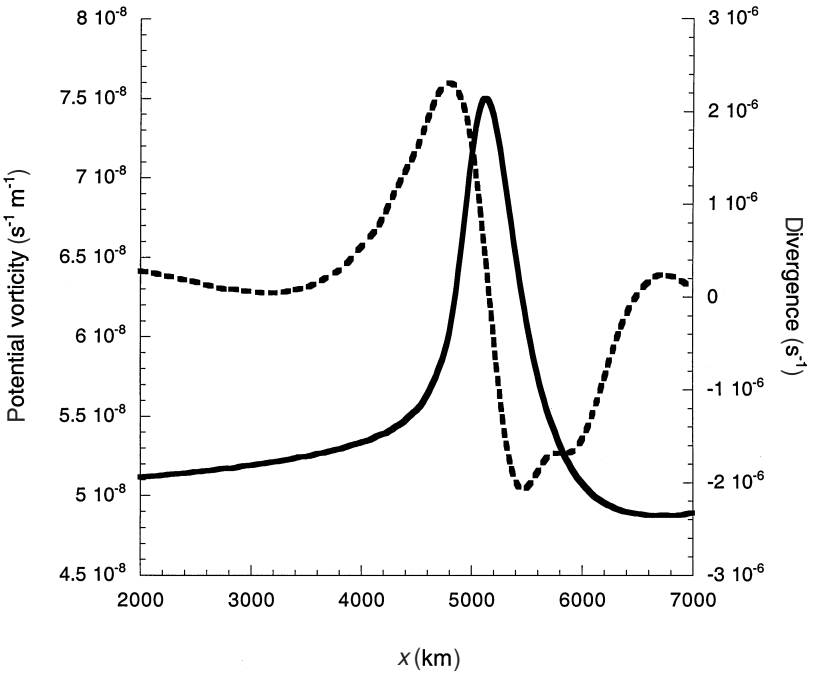


Figure 4.60. The numerical solution of Eqns. 4.58 with $\bar{u}_1 = 0$ and $\bar{u}_2 = 30$ m s^{-1} at $t = 48$ h, in terms of potential vorticity in the upper layer (solid line) and divergence ($\partial u_1/\partial x$) in the lower layer (broken line). The initial condition is a potential vorticity anomaly in the upper layer at $x = 0$ (as shown by the thick solid line in Figure 4.58), conforming to the invertibility principle (Eqn. 4.60). The amplitude of the potential vorticity anomaly remains constant, but the anomaly is deformed slightly at both the leading and trailing edge by meridional advection of basic state potential vorticity.

Because of material conservation of potential vorticity in adiabatic circumstances, the 'balanced dynamics' retains high quantitative accuracy even in the presence of large-amplitude unbalanced motions, such as those associated with gravity waves or convection. This is related to the fact that the 'unbalanced motion' is practically incapable of changing the potential vorticity distribution, i.e. potential vorticity is not carried away or radiated by gravity inertia waves as is the case with energy. We have seen this very clearly in Sections 4.5.4 and 4.5.8.

The intensity and the sign of the perturbation of the potential vorticity depends in particular on the vertical stratification of potential vorticity at the location of the heat source. Heating in the mid-troposphere appears to be most effective as a local source of positive potential vorticity if the cyclone has a warm core in the lower troposphere, because in that case the diabatic advection of

potential vorticity contributes to creating a positive tropospheric potential vorticity perturbation. Therefore, the chances of further intensification of a cyclone due to heating increase as the cyclone acquires a warm core.

Many polar lows are observed in the vicinity of, or directly under, a larger-scale cold core low. The low static stability under such a low stimulates convection and latent heating, but the numerical experiments described in Section 4.5.8 (see also Emanuel and Rotunno, 1989) clearly indicate that the growth of a polar low-like cyclone requires a low-level positive potential vorticity anomaly (i.e. a low-level warm anomaly) that can interact constructively with this heating. A synoptic structure, often observed in association with the early stages of polar low formation, which may play the role of this positive potential vorticity anomaly, is a low-level shear line.

It should be mentioned that deviations from axisymmetry and the consequent excitation of vortex Rossby waves could give rise to changes in the symmetric potential vorticity distribution and, thus, play a role in the evolution of the vortex. This subject is currently under investigation (Möller and Montgomery, 2000; Shapiro, 2000).

The coupling between the heating and the balanced flow has been a controversial topic for many years (Stevens *et al.*, 1997; Smith, 2000). Basically, there are two theoretical views on this problem. The first is based on the assumption that the heating is proportional to the moisture flux convergence, principally in the boundary layer. The second view is based on the assumption that the temperature in the atmosphere, in the presence of moist convection, adjusts to a reference temperature profile with a vertical gradient close to that associated with the moist adiabat. The second view has considerable advantages in a balanced model because knowledge of divergent motion is not required to determine the heating intensity. Agreement on this topic has yet to be reached.

4.6 Further theoretical considerations

In order to summarize and clarify some of the subjects and mechanisms referred to in the preceding sections some additional comments are presented below.

4.6.1 CISK and WISHE

Observations have shown that the development of most polar lows is closely associated with deep convection. Reflecting this fact, two mechanisms, CISK and WISHE, have been suggested to explain the development of a 'large-scale' balanced system, such as a polar low in the presence of deep convection. The persistence of the two modes of thought which have existed in

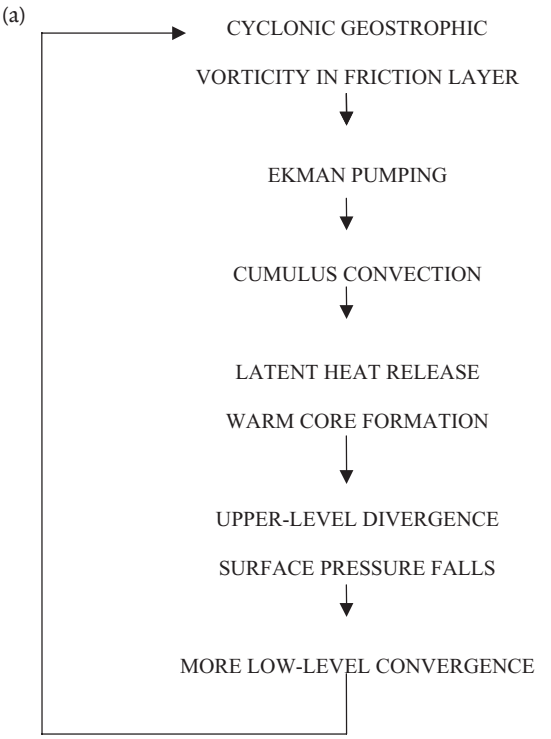


Figure 4.61. Schematic diagrams of (a) CISK (b) WISHE (ASII) (adapted from Bluestein, 1993).

parallel for about 20 years is due at least in part to the difficulty in distinguishing between them with currently available observations (Craig and Gray, 1996).

The two mechanisms are illustrated by the schematic diagrams shown on Figure 4.61 based on Bluestein (1993). Considering Figure 4.61a illustrating the CISK mechanism, it should be kept in mind that a necessary prerequisite for CISK is the presence of CAPE as discussed in Section 4.5.1, so that air parcels, after being lifted to their level of free convection will continue their ascent. In this case the Ekman pumping may trigger and/or enhance cumulus convection. The latent heat release associated with the convection will, through the formation of a warm core, induce upper-level divergence (outflow), falling pressure at the surface and more low-level convergence. Convergence acting on the existing low-level cyclonic vorticity produces more vorticity, which induces more Ekman pumping, and so on.

In CISK the cumulus heating (Q) is controlled by the balanced flow and is often taken to be proportional to the low-level convergence, i.e. $Q = \eta\, w^{+}$ where

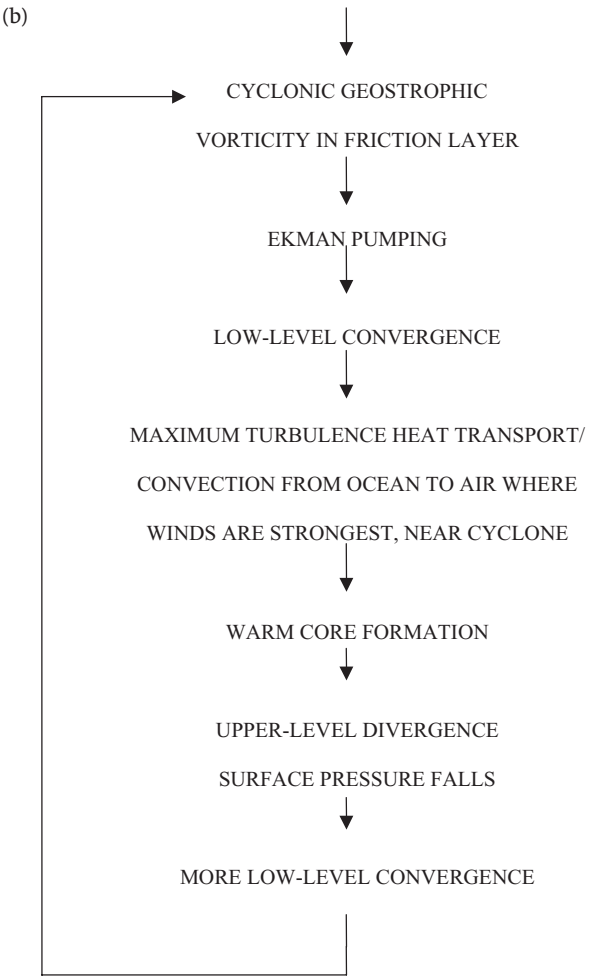


Figure 4.61 (*cont.*).

w^+ is the vertical velocity at the top of the boundary layer. Ooyama (1969) assumed that η was a function of the conditional instability of the atmosphere, while Charney and Eliassen (1964) assumed it to be proportional to the boundary layer specific humidity. In linear CISK theories, η is assumed to be constant, which is equivalent to assuming that surface fluxes act rapidly to replenish any depletion of CAPE or boundary layer moisture by the convection. In nonlinear CISK models, surface fluxes of heat and moisture may contribute to a time evolution of the coefficient η. The possibility of a variable η is especially relevant for polar lows which during their lifetime may move from regions with a relatively cold sea surface to regions characterized by much higher sea surface temperatures and fluxes.

In the ASII/WISHE mechanism (Figure 4.61b) the low-level inflow towards the centre of an incipient tropical cyclone or polar low results in increased surface fluxes from the sea surface in the regions of high wind speed near the centres of the cyclones. Sensible and latent heat is transported upwards by turbulent heat transport and convection. The WISHE theory assumes that the atmosphere is basically convectively neutral, i.e. the lapse rate is being constrained to follow a moist adiabatic lapse rate (see Section 4.5.2 and Figure 4.31) which means that there will be no (or little) CAPE. Also, in this case, a warm core will be formed. It should be noted that although the energy source of the cyclone according to the WISHE theory is through surface fluxes, a small amount of CAPE is not inconsistent with this theory (Emanuel *et al.*, 1994).

Concerning the initial formation of a vortex (polar low) from an 'infinitesimal disturbance', results from simple, linear CISK-driven models (i.e. Rasmussen, 1979; Pedersen and Rasmussen, 1985, Bratseth, 1985) showed growth rate curves with preferred wavelengths and a relatively short e-folding times corresponding to those for low-level baroclinic instability found by Mansfield (see Figure 4.6). These results were interpreted in such a way that small, 'natural' disturbances within a conditionally unstable air mass might grow due to CISK. Disturbances with a wavelength equal to or near the most unstable wavelength would amplify more quickly than disturbances of other wavelengths, and therefore, as long as nonlinear interactions could be neglected, would tend to dominate (Rasmussen 1979).

However, theoretical results have indicated that neither CISK nor WISHE can explain *the initial* development from an infinitesimal disturbance to a polar low. Van Delden (1989) found that a relatively weak cyclone is rather insensitive to diabatic heating, and underlined that '*CISK must be interpreted as a finite-amplitude instability* accounting for the rapid or explosive intensification of a balanced cyclone by diabatic heating'.

Concerning WISHE (ASII), Emanuel and Rotunno (1989) argued, that '*disturbances of substantial amplitude are apparently necessary to initiate intensification by air-sea interaction*'. To the extent that this is a valid finding, it points to the necessity of some presumably non-axisymmetrical dynamical process that operates in the early stages of cyclogenesis. They mention baroclinic instability as an obvious candidate and point out, that the development of polar lows might be thought of as a *two-stage process* for which also other disturbances, such as topographically generated cyclones, might act as starting disturbances.

The theoretical results mentioned above have been supported by earlier observational results. While satellite images have shown that most polar lows are associated with deep convection, then on the other hand many of these

systems initially form in a highly (low-level) baroclinic environment near the ice edges. The coincidence of baroclinicity and convection led Rasmussen (1985a), among others, to propose that convection and baroclinic instability cooperate to produce the polar lows. The presence of an upper-level disturbance in the form of a short-wave trough or a vortex, prior to a number of significant polar low developments, led to the formulation of the widely accepted scenario that high latitude polar lows typically were initiated through a baroclinic process involving an upper-level short-wave trough and a low-level, ice edge-generated, baroclinic zone. As discussed elsewhere in this volume (e.g. Section 4.2), this mechanism may not be so dominant as once believed. The role of the low temperatures associated with the upper-level, cold short-wave troughs may be a crucial factor (see the following discussion of a model proposed by Økland (1987, 1989).

Økland (1987) pointed out that 'although the winter Arctic air mass originally is very stable in its lower layers, the same is not necessarily the case at greater heights'. It should be noted, though, that according to Økland, areas characterized by small vertical stability may be limited to the regions around small-scale, upper-level cold troughs, or emerge during reverse shear baroclinic developments (see Section 4.2 and Figure 4.15). According to Økland, the horizontal scale of circulations (polar lows) will be determined mostly by the size of the region of enhanced convection, i.e. indirectly of the size of the upper-level cold trough, and not through a scale selection for maximum growth rates as envisaged by some linear CISK theories.

Økland (1989) elaborated these ideas arguing that deep convection within Arctic air masses characterized by a low-level inversion is reached only in those cases where a layer of weak stability is present on top of the low-level inversion. When the air moves over the warm ocean, the low-level inversion will gradually be eroded away by a convective boundary layer which increases in thickness and temperature downstream. If this convective boundary layer reaches a height above which the atmosphere is conditionally unstable the increase in depth of the convective layer will be 'explosive', resulting, according to Økland, in a much greater effiency of the CISK process.

Concerning the question of the source of the low mid-tropospheric stability sometimes found in Arctic air masses, Økland, as noted above, pointed towards the role of the small-scale cold, upper-level troughs. Such troughs, as shown by several case studies (e.g. Rasmussen, 1985a; Businger, 1985), are able to trigger polar low developments. Concerning the nature of the triggering process, Økland (1987, 1989), was probably the first to stress the role of the low static stability due to the presence of cold air aloft. Økland (1989) summarized his ideas of the development of a polar low due to latent heat release in deep

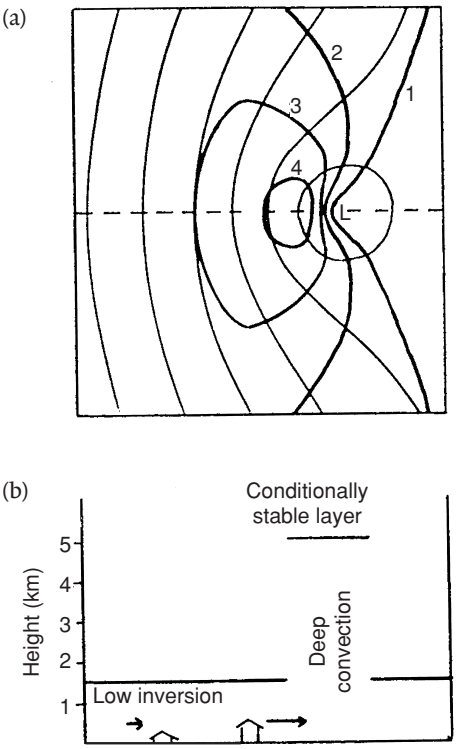


Figure 4.62. Sketch of a polar low caused by convection. (a) Shows isobars (thin lines) and isotachs (thick lines). The numbers on isotachs are relative. (b) A vertical cross-section through the low centre (from Økland, 1989, in Twitchell *et al.*, 1989. © A. Deepak Publishing.)

convection in the schematic diagram shown on Figure 4.62. The polar low is embedded in a northerly flow or located near the shear zone between a strong northerly flow to the west and weak winds towards the east (Figure 4.62a). The asymmetry results in an area of strong wind to the west of the low. In this wind maximum the heat flux from the sea is large. In the centre of the developing vortex the low-level stable layer is missing and if the air mass at greater heights has small static stability the convection will become deep. This will, according to Økland, happen over a fairly restricted area where the mid-level stability has a minimum due to the presence of an upper-level trough (Figure 4.62b). On the figure is shown how, outside the area with deep convection, the low-level inversion usually persists in a large part of the peripheral sections of the low. However, also in this region, the heat and moisture fluxes from the sea surface are comparatively large because of the strong winds associated with the low. The Ekman layer flux towards low pressure brings this air into the central part of the low where the heat and moisture feed the deep convection.

Emanuel and Rotunno (1989) argued that polar lows in general and the 13–14 December 1982 Bear Island case in particular were driven by the WISHE mechanism. However, significant CAPE was indicated by the NCEP/NCAR data (see Section 4.5.1) during the second intense phase of the Bear Island, December 1982 polar low development, pointing towards a contribution from CISK. The very rapidity, only a few hours, of development of this hurricane-like polar low seems to favour a CISK type of mechanism utilizing a source of accumulated CAPE rather than a comparatively slow WISHE mechanism. In the numerical experiments simulating this development using the WISHE mechanism, Emanuel and Rotunno found that the development of the maximum azimuthal velocity around the polar low required a time of the order of two days, whereas the observed time of development was of the order of a few hours.

Craig and Gray (1996) pointed out that the mere presence of CAPE does not provide a conclusive test of CISK and that a small amount of CAPE is not inconsistent with WISHE. To test the relative importance of CISK and WISHE they noted that the sensitivity of the intensification rate of tropical cyclones and polar lows to surface properties, such as surface friction and moisture supply would be different for the two mechanisms. Experiments with an axisymmetric model and explicit convection (to avoid the possibility of prejudicing the results through the choice of a particular parameterization scheme) showed that the intensification rates of a simulated polar low had a strong dependence on the heat and moisture transfer coefficients, while remaining largely insensitive to the frictional drag coefficient. Their results imply that the rate-limiting process for cyclone intensification is the rate of heat and moisture fluxes while frictional convergence is of secondary importance. Based on this they concluded 'that the intensification of the numerically simulated tropical cyclones and polar lows is due to WISHE', and that 'It is anticipated that a similar conclusion would apply in nature'. While it is very likely that WISHE will be the dominant mechanism for some polar low developments it is more doubtful whether this statement is valid in a more general way as advocated by Craig and Gray. They simulated an axisymmetric polar low characterized by a well-developed eye and with only small values of CAPE. In nature the lows will be highly asymmetric, often without an eye-region, and form in regions with relatively large values of CAPE. As seen from satellite images such as Figure 3.31, the intense Bear Island polar low, which formed late on 13 December 1982, had no eye but was characterized by a small cluster of convective clouds within the central region of the low. Økland (1987) argued that polar lows forced by heating due to surface fluxes proportional to surface wind speeds (WISHE) show a tendency to displace the active part of the low from the centre to more

peripheral areas forming an 'eye', whereas heating proportional to the relative vorticity (CISK) tended to form an intense vortex in the central part of the low.

It is also doubtful whether a general conclusion concerning the relative importance of the two mechanisms can be drawn alone based on results on the use of an axisymmetric model. For example, in nature most polar lows are asymmetrical including the presence of shallow low-level fronts. The Arctic air masses in which polar lows typically develop are often quite dry at low levels, which inhibits widespread deep convection. It has been documented from satellite images that deep convection during polar low developments tends to be organized along shallow Arctic fronts along which air parcels are forced to ascend to their level of free convection. The effect of this lifting, which is likely to be important, is limited to polar/Arctic regions and is not included in the traditional formulation of CISK.

4.6.2 The Montgomery–Farrell model

Montgomery and Farrell (1992) in their work utilized observations such as presented by Reed (1979), Rasmussen (1987) and Businger (1987) showing that polar lows are generally initiated by an upper-level disturbance in the form of a mobile, short-wave cold trough. Based on a study utilizing a three-dimensional nonlinear geostrophic momentum model that incorporated moist processes and strong baroclinic dynamics, they proposed, as briefly considered in Section 4.5.12, a conceptual model in which polar low development occurs in two stages. The first stage, the induced self-development, comprises an interaction between an upper-level potential vorticity anomaly, i.e. an upper-level trough, and a low-level potential vorticity anomaly in a nearly moist neutral atmosphere. The enhanced omega response (vertical velocity) in the nearly neutral atmosphere initiates a rapid low-level spin-up and generation of low- and middle-level potential vorticity anomalies that augment the baroclinic interaction between the low-level system and the system aloft. The subsequent secondary development, called diabatic destabilization or diabatic intensification, is associated primarily with the diabatic production of low-level potential vorticity in regions of ascending air and cloud formation (see Section 4.5.12).

Montgomery and Farrell ascribe the fact that polar lows often maintain their intensity, or slowly intensify until they reach land to diabatic destabilization. In exceptional cases instances of polar cyclogenesis with negligible upper-level forcing, diabatic destabilization may also explain the gradual intensification of small-scale vortices in regions of sustained neutrality and surface baroclinicity.

While cooperative intensification processes associated with either CISK or WISHE may be operative in the latter phase of development of polar lows, neither, according to Montgomery and Farrell, appear essential for describing the basic formation process of polar lows. CISK or WISHE may, however, according to them, play the role of a cyclone 'afterburner'.

In conclusion, Montgomery and Farrell hypothesized that the formative mechanisms for polar lows are the same as those for mid-latitude cyclones, with moist processes playing a more central role in the polar low case.

Montgomery and Farrell assumed as a basis for their work that polar lows form in regions of 'strong ambient baroclinicity'. The cases they refer to – comma clouds, polar lows along occluded cyclones and reverse shear systems – may, corresponding to their basic premises, all be characterized as forming within a baroclinic environment. It is doubtful, however, whether the theory put forward by Montgomery and Farrell can be applied to the 'real' or 'true' polar lows which develop at high latitudes away from the main baroclinic zone. In spite of this shortcoming, the ideas put forward by Montgomery and Farrell have been an important contribution to our understanding of polar low dynamics.

4.6.3 The Fantini model

Emanuel (1994) discussed a non-modal mechanism for a polar low WISHE-type development based on work by Fantini (1990). According to this theory, which in many ways is reminiscent of the theory of Montgomery and Farrell discussed above, a surface cyclone is initiated when an upper-level PV anomaly approaches a region of low static stability and high sea–air thermodynamic disequilibrium (Figure 4.63a). The low-level cyclone then amplifies rapidly by the WISHE mechanism forming a highly concentrated, relatively shallow warm core surface cyclone, associated with an intense, small-scale low-level PV maximum. The subsequent overtaking of this low-level PV anomaly by the upper-level anomaly (Figure 4.63b) then results in rapid baroclinic deepening. Numerical experiments with this model (Fantini, 1990) of the development of large-scale maritime storms in a baroclinic, saturated environment resulted in 'explosive' growths with deepening rates near or above the conventional definition of explosive cyclones. The rapid deepening stage was (Emanuel, 1994) 'almost purely baroclinic and should be thought of as a rapid baroclinic growth greatly enhanced by the presence of a low-level PV anomaly created early in the evolution by the WISHE mechanism triggered by an approaching upper-level PV anomaly'.

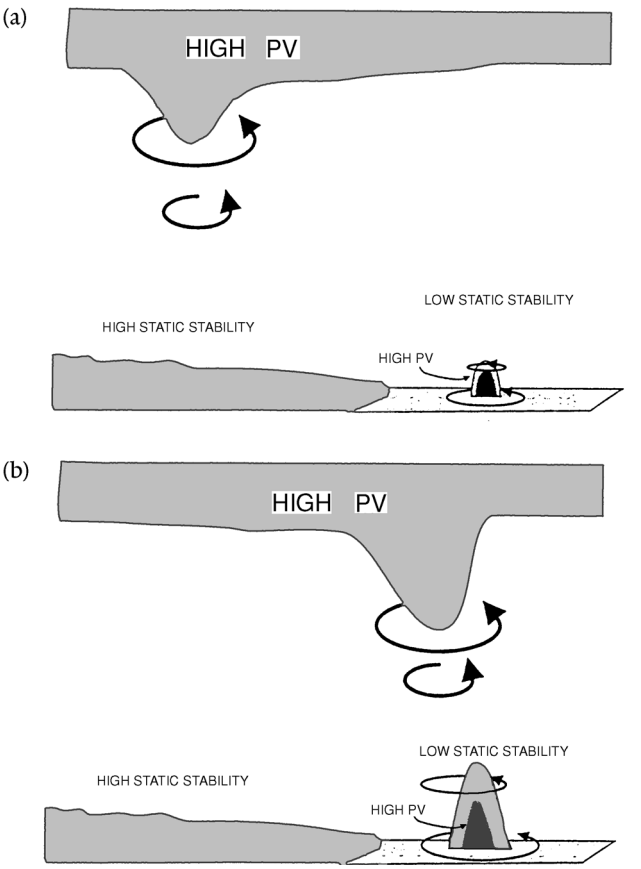


Figure 4.63. An upper-level PV anomaly approaches a region of large sea–air thermodynamic disequilibrium and low static stability and initiates a surface cyclone (a). The WISHE mechanism forms a concentrated low-level PV maximum, similar to that of a hurricane but more shallow and not as intense; this subsequently interacts with the approaching upper-level PV anomaly (b) to produce baroclinic cyclogenesis (from Emanuel, 1994).

4.6.4 The Craig and Cho model

In Section 4.2, two studies of Sardie and Warner (1983, 1985) concerning the relative importance of baroclinic instability and CISK were discussed. Craig and Cho (1989) explored the role of cumulus heating and CISK in a baroclinic environment using a simple linear model that incorporated both baroclinic and convective processes. The model applied was a semi-geostrophic Eady model of baroclinic instability with a continous vertical structure, including a simple parameterization of cumulus heating. Within the context of their linear model they found that the interaction of baroclinic and convective processes could be understood as follows. If the heating rate was not

too great a baroclinic wave would grow in the usual manner, the cumulus heating merely intensifying the existing circulation. The effects of convective heating in this case were essentially the same as a reduction in static stability, which resulted in a faster growing disturbance with a shorter wavelength. CISK occurred in the model when the circulation induced by the heating was sufficient to supply the same or a greater amount of heating. This is only possible when a heating parameter in the model, which is a measure of the efficiency of the heating in resupplying itself, is above a certain threshold. Since the convective and baroclinic processes cooperate at all stages, the onset of CISK is through a smooth transition and there is no precise point where one instability mechanism ceases to dominate and the other starts. Six polar air disturbances considered by Craig and Cho illustrated the full range of behaviour shown by the model. Two polar lows (including the 13–14 December 1982 Bear Island case discussed in several places above) tended to be CISK-dominated, two comma clouds tended to be primarily baroclinic, and the two remaining systems of a transitional nature. Craig and Cho concluded, in accordance with observations, that polar lows may change over time. For example, an initially baroclinic system could become CISK-dominated as surface fluxes of heat and moisture decrease the stability of the atmosphere and enhance cumulus convection.

4.7 Summary and concluding remarks

In the preceding sections of Chapter 4 we have discussed a number of physical mechanisms which, over the years, have been suggested as being responsible for the development of polar lows.

The areas in which polar lows have been observed to form range from highly baroclinic regions near the polar front and along the ice edges, to high latitude, nearly barotropic (or equivalent barotropic) environments. Because of this a variety of forcing mechanisms will be effective, and it is not surprising that polar lows appear in so many forms. The arguments from the late 1970s and early 1980s that polar lows were either baroclinic disturbances, *or* convective systems akin to tropical cyclones, were gradually replaced with the understanding that both mechanisms were important. This advance led to the idea of a 'polar low spectrum' with purely baroclinic systems at one end, and purely convectively driven systems at the other (see Section 3.1.4, The 'Polar Low Spectrum'). In between the two 'pure types', there was room for a variety of hybrid systems for which both mechanisms, i.e. baroclinic instability and a 'convective mechanism' such as CISK or ASII/WISHE, could play a smaller or greater role depending on the precise circumstances under which a particular development took place. Experience has shown that the hybrid types are by far the most common.

Compared to the two principal fluid dynamic instabilities mentioned above, baroclinic instability and thermal (convective) instability, the third main mechanism of barotropic instability, is thought to play a minor role in connection with polar lows. However, increasing evidence indicates, that while this mechanism in isolation may seem only of minor importance, and as such only important for the development of rather insignificant vortices, in combination with other factors it may play an important role as a trigger for subsequently stronger developments as discussed in Section 4.3. In addition, the two main mechanisms of baroclinic and thermal instability may appear in different forms. Concerning baroclinic instability, these forms include low-level baroclinic instability, deep baroclinic instability, reverse-shear baroclinic instability and processes involving a cooperation between upper-level baroclinic systems and low-level baroclinic zones (type B developments).

While it has been generally accepted that convection plays an important role in the development of a number of polar lows, the form through which this influence takes place has been a subject of much discussion. During the 1980s the point of view that CISK was the main dynamic mechanism for a significant group of polar low developments was challenged by a number of authors (i.e. Emanuel and Rotunno, 1989). Arguing that the polar atmosphere, like the tropical one, was nearly neutral to deep, moist convection, they rejected the CISK theory and proposed the WISHE mechanism (Section 4.5.1). In WISHE, sensible and latent heat fluxes from the sea surface maintain the atmosphere in a convectively neutral state with correspondingly negligible amounts of CAPE. There is evidence though that in the polar regions, the combined action of surface fluxes of heat and moisture and upper-air cold advection occasionally may lead to the generation of significant amounts of CAPE. The presence of CAPE plus the fact that some polar lows develop very rapidly indicates that CISK may contribute to these developments. Based on the material presented in this book a preliminary conclusion on the ongoing discussion of the relative importance of CISK versus WISHE for the developments of tropical cyclones and polar lows must be, that for polar lows both mechanisms may, according to the specific circumstances, contribute to the development of the lows. Both theories, CISK and WISHE hypothesize that the large-scale vortex, i.e. the polar low, develops through a succession of balanced states. The question how the heating associated with either the CISK or WISHE affects this balanced flow was dealt with in detail in Sections 4.5.5 to 4.5.10. Clearly future research is required on the role of convection in the development of polar lows.