Under these assumptions, the invertibility principle can be expressed in the following way. If (22) is differentiated with respect to r, using (24), (25) and the assumption of constant f, the result after multiplication by σ can be written as

$$\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial(rv)}{\partial r}\right] - \frac{\zeta_{a\theta}}{\sigma}\frac{\partial \sigma}{\partial r} = \sigma\frac{\partial P}{\partial r}. \tag{26}$$

Differentiating (23) with respect to r and making use of the thermal wind equation (21), we get

$$-g\frac{\partial \sigma}{\partial r} = \frac{\partial^2 p}{\partial r \partial \theta} = \frac{\partial}{\partial \theta}\left[\frac{f_{loc}}{R}\frac{\partial v}{\partial \theta}\right]. \tag{27}$$

(For later reference, the relation between R, σ and the static stability can be shown to be

$$R = g/(\sigma N^2 \theta^2) > 0, \tag{28}$$

N^2 being the static stability expressed as the square of the Brunt–Väisälä or buoyancy frequency.) Using (22) and (27), we can write (26) finally as

$$\boxed{\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial(rv)}{\partial r}\right] + g^{-1}P\frac{\partial}{\partial \theta}\left(\frac{f_{loc}}{R}\frac{\partial v}{\partial \theta}\right) = \sigma\frac{\partial P}{\partial r}} \tag{29}$$

a nonlinear equation which can be solved, for instance by relaxation methods, as described below, for the wind profile $v(r, \theta)$ given the PV distribution $P(r, \theta)$. Note that the *isentropic gradient* of P appears on the right-hand side as a prescribed forcing function.

Together with suitable boundary conditions, and the condition (17a), Eq. (29) expresses the invertibility principle in much the same way as was done in Kleinschmidt's original work. Note that if

$$f_{loc}P > 0, \tag{30}$$

as we shall assume, then Eq. (29) is an elliptic equation, so that the problem is well posed. As is well known, (30), together with (23), also expresses the assumption of static, inertial, and 'symmetric' baroclinic stability previously made in section 1(*d*) (e.g. Hoskins 1974, with f replaced by f_{loc}). Equation (29) is exact; its simple form is due to the assumption of circular symmetry and the use of isentropic coordinates.

Note further that if we were to make the approximations

$$f_{loc} \simeq \zeta_{a\theta} \simeq f, \qquad R \simeq R_{ref}(\theta), \qquad \sigma \simeq \sigma_{ref}(\theta) \tag{31}$$

everywhere except when calculating the forcing function $\partial P/\partial r$ from (22), where $R_{ref}(\theta)$ and $\sigma_{ref}(\theta)$ are the reference-state profiles of R and σ, then (29) would simplify to

$$\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial(rv)}{\partial r}\right) + \frac{f^2}{g\sigma_{ref}}\frac{\partial}{\partial \theta}\left(R_{ref}^{-1}\frac{\partial v}{\partial \theta}\right) = \sigma_{ref}\frac{\partial P}{\partial r}, \tag{32}$$

which is the isentropic coordinate version of the usual quasi-geostrophic approximation to (29). The elliptic operator on the left-hand side of (32) is now linear and if, further, σ_{ref} and R_{ref} were constants, then apart from its slightly different r-dependence the operator would be a three-dimensional Laplacian, after suitably rescaling the vertical coordinate according to the Prandtl–Rossby–Burger relation

$$\Delta\theta \sim fL/(Rg\sigma)^{1/2} \qquad (\text{cf. } H \sim fL/N) \tag{33a}$$

(e.g. Rossby 1938), where L is the horizontal scale of the flow. For a more accurate scale relation corresponding to (29) we may replace f by $(f_{\rm loc}P\sigma)^{1/2}$, giving

$$\Delta\theta \sim (f_{\rm loc}P/Rg)^{1/2}L \qquad (\text{cf. } H \sim (f_{\rm loc}P\sigma)^{1/2}L/N), \tag{33b}$$

which would be relevant near the equator. H and $\Delta\theta$ are respectively the scales in physical, xyz space and in $xy\theta$ space, measuring the vertical penetration of the induced flow structure above or below the location of the IPV anomaly. The relevance of $\Delta\theta$ rather than H in the isentropic coordinate description explains why the square of the Brunt–Väisälä frequency appears in the denominator, rather than the numerator of (28).

We shall call the expressions on the right of (33a) the Rossby heights (in $xy\theta$ and xyz space respectively), and denote them by $\Delta\theta_{\rm Rossby}$ and $H_{\rm Rossby}$. As is well known, the concept is complementary to that of the Rossby radius of deformation, which is the horizontal scale L obtained from (33a) when $\Delta\theta$ or H is given. If $\Delta\theta_{\rm Rossby}$ and $H_{\rm Rossby}$ greatly exceed the corresponding reference-density scale heights $\Delta\theta_{\rm density}$, $H_{\rm density}$ (the scale heights for variation of σ and ρ respectively), then non-Boussinesq effects are important. Order-of-magnitude relations which cover the whole range of $\Delta\theta$ or H are

$$\Delta\theta \sim \min\{\Delta\theta_{\rm Rossby}, \Delta\theta_{\rm density}\} \quad \text{and} \quad H \sim \min\{H_{\rm Rossby}, H_{\rm density}\} \tag{33c}$$

for the vertical scales for downward penetration of the induced wind field, and

$$\Delta\theta \sim \max\left\{\Delta\theta_{\rm Rossby}, \frac{(\Delta\theta_{\rm Rossby})^2}{\Delta\theta_{\rm density}}\right\} \quad \text{and} \quad H \sim \max\left\{H_{\rm Rossby}, \frac{(H_{\rm Rossby})^2}{H_{\rm density}}\right\} \tag{33d}$$

for its upward penetration (Rossby, *op. cit.*). These more general scaling rules are related to well-known results in tidal theory.*

The fact that the inverse Laplacian is a smoothing operator should be kept in mind; for instance it is the essential reason why Figs. 4, 6 and 12 look like smoothed versions of Figs. 3, 5 and 11 (see also (44) below). There is an associated *scale effect*, whereby small-scale features of given strength in the IPV field have a relatively weak effect on the velocity field whereas large-scale features have a relatively strong effect. As already mentioned, this is one of the reasons for supposing that coarse-grain IPV distributions are dynamically meaningful. Note that the smoothing takes place in the vertical as well as in the horizontal.

The exact operator appearing in (29) is nonlinear because of the presence of the unknown functions $f_{\rm loc}$, R and σ (σ appearing on the right-hand side). As Figs. 3, 4, 5, 6, 11 and 12 suggest, this operator very often has the same qualitative character as its approximate counterpart in (32), although it should be remembered that differences near fronts and shear lines can be important (e.g. Hoskins and Bretherton 1972). In many circumstances of interest, (29) may be expected to be soluble iteratively, for given $P(r, \theta)$, having regard to any constraint imposed via (17b). The conceptually simplest albeit not

* The alternatives within braces are the asymptotic forms for large and small $H_{\rm density}/H_{\rm Rossby}$, as the case may be, of the more precise expression

$$H = \left|(2H_{\rm density})^{-1} \pm \sqrt{(H_{\rm Rossby})^{-2} + (2H_{\rm density})^{-2}}\right|^{-1},$$

the + and − signs corresponding to downward and upward penetration respectively, which arises in the theory of very-low-frequency tidal oscillations of negative equivalent depth (Kato 1966; Lindzen 1966). The problem solved by Rossby (1938) is an approximate version of the same problem; both concern the linear response to a forcing effect of a given horizontal scale L at a given level. A convenient reference is Holton (1975), in which it should be noted that the expression (2.85) corresponds apart from sign convention to minus the square of the expression

$$\sqrt{(H_{\rm Rossby})^{-2} + (2H_{\rm density})^{-2}}$$

above, $H_{\rm Rossby}$ being equal to $\sqrt{-gh/N^2}$ in Holton's notation, where h is the equivalent depth.

the most powerful method starts from a solution to (32) as first guess, and then refines the initial approximations (31) by straightforward iteration. Notice carefully how the condition (17a) invoking the reference state has to enter into this process. As soon as a guess for $v(r, \theta)$ has been computed from (29), using the previous guesses for $R(p) = R(r, \theta)$, and $\sigma(r, \theta)$, improved approximations to $R(r, \theta)$ and $\sigma(r, \theta)$ must be derived

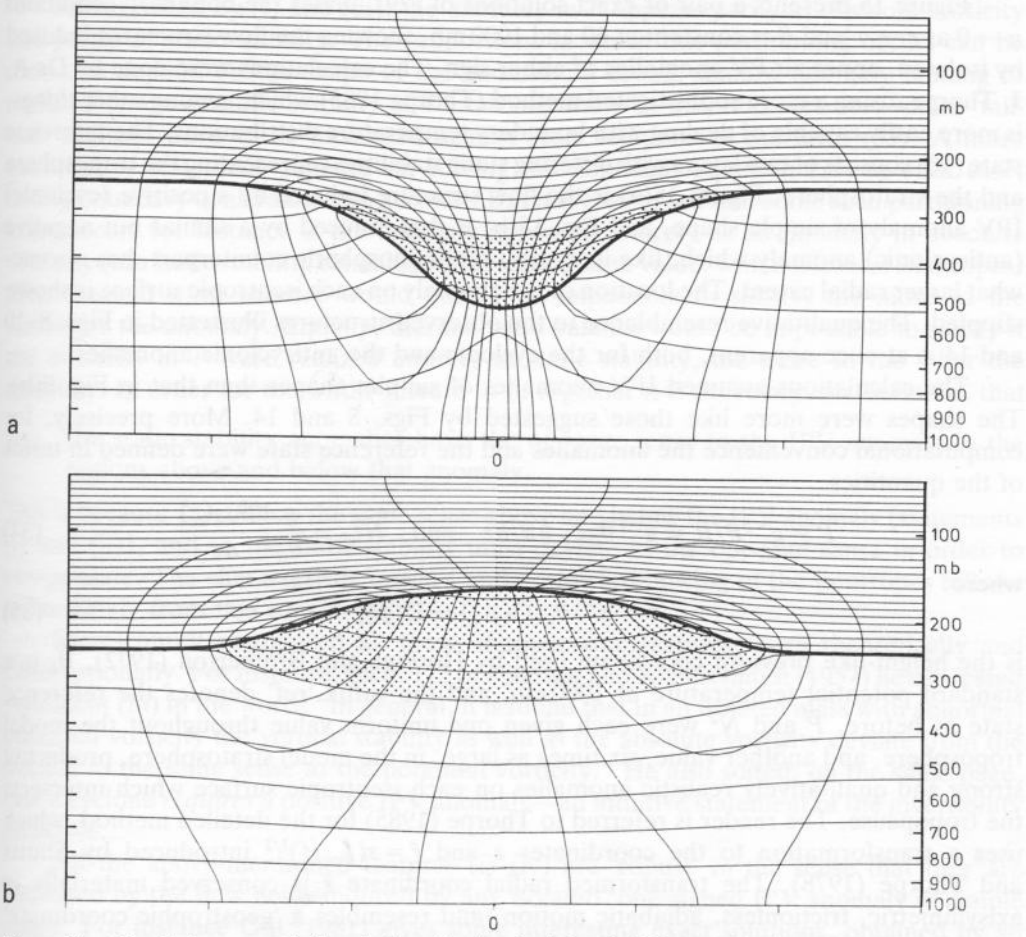

Figure 15. Circularly symmetric flows induced by simple, isolated, IPV anomalies (whose locations are shown stippled) as described in the text. The basic static stability $\tilde{N}$ and therefore $\tilde{P}$ (defined in (34)) was uniform in the tropospheric region and six times larger in the stratospheric region. The vertical coordinate z is nearly the same as physical height but is defined exactly in (35), g/θ_0 being taken to be $(1/30)\,\mathrm{m\,s^{-2}K^{-1}}$. The reference tropospheric 'height' z was 10 km and the total domain 'height' 16·67 km: f was taken to be $10^{-4}\mathrm{s^{-1}}$. The IPV anomaly was defined by taking the tropopause potential temperature to vary in the manner $\frac{1}{2}A\{\cos(\pi\tilde{r}/r_0) + 1\}$ for $\tilde{r} < r_0$, where $\tilde{r} = r(f_{\mathrm{loc}}/f)^{1/2}$. Here the amplitude A was taken to be −24 K in (a) and +24 K in (b) which may be compared with a potential temperature increase of 30 K over the depth of the reference troposphere. The parameter r_0 was taken to be 1667 km. The undisturbed θ distribution was imposed as a boundary condition at $\tilde{r} = 5000$ km, and the solutions obtained had only a weak dependence of $C_b(\theta)$ upon θ as well as a far-field stratification approximating the reference stratification (16). (In terms of our definitions, the IPV anomaly in the stippled regions must therefore strictly speaking be considered to be embedded in a suitable 'surround' of much weaker anomalies, as noted below (17b).) Only the region $r < 2500$ km is shown here, and the tick marks below the axes are drawn every 833 km. The thick line represents the tropopause and the two sets of thin lines the isentropes every 5 K and the transverse velocity every $3\,\mathrm{m\,s^{-1}}$. The zero isotach on the axis of symmetry is omitted. In (a) the sense of the azimuthal wind is cyclonic and in (b) it is anticyclonic, in both cases the maximum contour value being $21\,\mathrm{m\,s^{-1}}$. The surface pressure anomaly is −41 mb in (a) and +13 mb in (b) and the relative vorticity extrema (located at the tropopause) are $1{\cdot}7f$ in (a) and $-0{\cdot}6f$ in (b). The maximum surface winds are $15\,\mathrm{m\,s^{-1}}$ and $6\,\mathrm{m\,s^{-1}}$ respectively. For more details of the method of computation, see Thorpe (1985). Courtesy of A. J. Thorpe.

for use in the next iteration. This requires (18a) to be integrated with respect to r, af which (18b) and (19a) are used to get $p(r, \theta)$ followed by (20) and (23) to get R and When (18a) is integrated with respect to r, an arbitrary function of θ arises as function of integration, and the reference-state condition (17a), along with the bounda conditions, is needed to determine this function of integration.

Figure 15 presents a pair of exact solutions of (29), under the boundary conditio $v \to 0$ as $r \to \infty$ and θ = constant at 60 and 1000 mb, showing the flow structures induc by isolated, upper air PV anomalies of either sign. The calculations were done by Dr J. Thorpe, using a more sophisticated method (Thorpe 1985) which, among other thing is more easily capable of dealing with boundary temperature distributions. The referen state (16) consists of two layers with differing static stabilities representing the troposphe and the stratosphere. Figure 15(a) is the flow structure induced by a positive (cyclon IPV anomaly of simple shape, and Fig. 15(b) is that induced by a similar but negati (anticyclonic) anomaly which, like its typical real-atmospheric counterpart, has a som what larger radial extent. The location of the anomaly on each isentropic surface is shov stippled. The qualitative resemblance to the observed structures illustrated in Figs. 8– and 14 is at once apparent, both for the cyclonic and the anticyclonic anomalies.

The calculations assumed IPV anomalies of simpler shapes than that in Fig. 9(b The shapes were more like those suggested by Figs. 8 and 14. More precisely, f computational convenience the anomalies and the reference state were defined in tern of the quantities

$$\widetilde{P} = g^{-1}P/R = g^{-1}\theta_0 \zeta_\theta \, \partial\theta/\partial z \quad \text{and} \quad \widetilde{N}^2_{\text{ref}} = \frac{g}{\theta_0}\frac{d\theta_{\text{ref}}(z)}{dz} \tag{3}$$

where

$$z = g^{-1}\theta_0\{\Pi(p_0) - \Pi(p)\} \tag{3}$$

is the height-like pressure coordinate used by Hoskins and Bretherton (1972), θ_0 is standard potential temperature at 1000 mb, and the suffix 'ref' denotes the referenc state as before. $\widetilde{P}$ and $\widetilde{N}^2$ were each given one uniform value throughout the mode troposphere, and another value, six times as large, in the model stratosphere, producin strong and qualitatively realistic anomalies on each isentropic surface which intersect the tropopause. The reader is referred to Thorpe (1985) for the detailed method, whic uses a transformation to the coordinates z and $\bar{r} = r(f_{\text{loc}}/f)^{1/2}$ introduced by Shutt and Thorpe (1978). The transformed radial coordinate $\bar{r}$ is conserved materially i axisymmetric, frictionless, adiabatic motion, and resembles a 'geostrophic coordinate (Hoskins 1975, Blumen 1981) except that gradient-wind rather than geostrophic balanc is used. Some quantitative details of the calculations are given in the caption to Fig. 15 note for instance that the values of f_{loc} and the absolute isentropic vorticity $\zeta_{a\theta}$ greatly exceed f near the centre of the cyclonic vortex; for this and other reasons both quasi- geostrophic and semi-geostrophic theory (sections 5(*b*), (*c*) below) would be poor approximations.

Qualitative features to be especially noted about Figs. 15(a), (b) include the facts that:

(i) the circulation in the balanced vortex has the same sense, relative to the earth, as the IPV anomaly which induces it;

(ii) the induced fields penetrate vertically above and below the IPV anomaly, boundary constraints permitting, by amounts consistent with the scale relations (33), especially when f is replaced by $(f_{\text{loc}}P\sigma)^{1/2}$ as in (33b), and

(iii) the static stability N^2, as well as the absolute vorticity, is anomalously high within a high PV anomaly, and low within a low PV anomaly, relative to the static stability of the reference state.

Note that for this purpose the 'anomaly' in the static stability, being the anomaly relative to the reference state, means *isentropic* anomaly, just as it does for P itself; one is comparing the static stability in the centre of the IPV anomaly with the static stability on the same isentropic surface at the periphery of the picture, where the actual state approximates the reference state.

As suggested by (i) and (iii), the anomaly in P appears partly as absolute vorticity and partly as static stability. The proportions in which this partitioning occurs can be shown to depend on the shape of the anomaly, a broad, shallow anomaly tending to realize P more as static stability and a tall one more as absolute vorticity, where 'tall' and 'shallow' have to be measured against scale relations of the form (33). A more quantitative statement depends on solving the nonlinear equation (29) in each case, taking boundary constraints into account. But it is easy to see that some such partitioning must occur; for instance if the anomaly appeared entirely as an anomaly in absolute vorticity (the static stability retaining its reference-state value, with horizontal isentropes) then thermal wind balance would plainly be impossible to satisfy, the more so the shallower the anomaly. Equally, thermal wind balance would be impossible to satisfy if the anomaly in P were realized entirely as static stability, the more so the taller the anomaly. In order for the whole picture to fit together it is necessary, furthermore, that

(iv) the static stability anomalies have the *opposite* sense to the IPV anomaly in the regions above and below that anomaly.

This is because $\zeta_{a\theta}$ still has the same sense above and below the IPV anomaly (statements (i) and (ii)), and so the static stability must deviate in the opposite sense in order to compensate. The characteristic upward and downward bowing of the isentropes follows immediately from this consideration.

Kleinschmidt recognized all these characteristic features, both theoretically and observationally. For instance, on p. 115 of Eliassen and Kleinschmidt (1957) he expressed statement (iii) in the words "In general, it is found that in an isolated mass with abnormal potential vorticity the vertical stability as well as the absolute vorticity deviate from the normal in the same sense as the potential vorticity." He also stated, on the same page, that a cyclone *requires* a positive IPV anomaly—an intuitive statement of the invertibility principle.

All the above-mentioned features (i)–(iv) are 'robust' in the sense that they are exhibited by the flow fields induced by any isolated, one-signed IPV anomaly of simple shape. For instance Gill (1981) gives some interesting exact solutions, obtained by an elegant analytical method, for very strong, anticyclonic, two-dimensional IPV anomalies. In each case the PV in the anomaly is *zero*, and the anomaly is concentrated on a single isentropic surface. All the qualitative features (i)–(iv) are reproduced even in that extreme limiting case. They are likewise reproduced by solutions to the approximate equation (32), which are valid for a sufficiently *weak* anomaly. For instance, let R_{ref} and σ_{ref} be taken to be constants in (32) (a Boussinesq, quasi-geostrophic model with a constant-static-stability reference state) and let P be taken to be constant and to be equal to $\sigma_{\mathrm{ref}}^{-1} f(1 + \varepsilon)$, with $\varepsilon \ll 1$, within an ellipsoidal region $r^2 + \Theta^2 = C^2 = \text{constant}$, and to have its reference value $\sigma_{\mathrm{ref}}^{-1} f$ outside that region, where Θ is the scaled coordinate $f^{-1}(R_{\mathrm{ref}} g \sigma_{\mathrm{ref}})^{1/2} (\theta - \theta_c)$, cf. (33a), θ_c being the value of θ at the centre of the ellipsoidal region. Then it is a straightforward theoretical exercise to construct an analytical solution, again exhibiting the same features (i)–(iv), in which

$$\left.\begin{aligned} v &= \tfrac{1}{3}\varepsilon f r && (r^2 + \Theta^2 < C^2) \\ v &= \tfrac{1}{3}\varepsilon f r \left(\frac{C^2}{r^2 + \Theta^2}\right)^{3/2} && (r^2 + \Theta^2 > C^2) \end{aligned}\right\} \tag{36}$$

terms of order ε^2 being neglected as well as non-Boussinesq effects in the far field. It can also be shown that, of the fractional change comprising the IPV anomaly, one third comes from the static stability and two thirds from the relative vorticity anomaly, in this particular case.*

Kleinschmidt gave various other examples, computed by the approximate methods available to him in the late 1940s and early 1950s, all exhibiting features (i)–(iv). It is worth remarking that one of these examples (e.g. Eliassen and Kleinschmidt, 1957, Fig. 23) is quite like that of Fig. 15(a) in some respects, except that the IPV anomaly is a little lower down, has a somewhat different shape, and is detached from the model stratosphere. The observed case shown in Fig. 9(b) appears to be intermediate between Kleinschmidt's example and cases like those of Figs. 8 and 15(a).

4. On the cancellation of horizontal advection by vertical motion

We are now in a position to appreciate one further, very basic point. It can be brought out most clearly by means of a thought-experiment, to whose wider significance we shall return in sections 6(*e*) and 8. The experiment could be carried out on a computer, with the aid of a multi-layer numerical model, but its outcome can in any case be predicted with certainty since it is an immediate consequence of the invertibility principle together with the robustness of the qualitative features (i)–(iv) just noted.

Consider what would happen if a broad airstream, itself involving comparatively weak IPV and surface θ anomalies, were to flow beneath a stationary, upper air IPV anomaly such as the cyclonic anomaly of Fig. 15(a). To a first approximation (anticipating the theory to be reviewed in the next section), we may superpose the airstream onto the structure shown in Fig. 15(a). The invertibility principle tells us that the cyclonic vortex in the lower troposphere will stay in place beneath the upper air anomaly which is inducing it, despite the 'attempt' by the low-level airstream to advect it downstream.

It is interesting to note what the vorticity budget would look like in this thought-experiment, which in order to highlight the point will be imagined to take place in a completely frictionless, adiabatic atmosphere. The isobaric absolute vorticity $\zeta_{ap} = f + \mathbf{k}.\nabla_p \times \mathbf{v}$ then satisfies

$$\frac{\partial \zeta_{ap}}{\partial t} + \mathbf{v}.\nabla_p \zeta_{ap} = \zeta_{ap} \frac{\partial \omega}{\partial p} + \mathbf{k}.\frac{\partial \mathbf{v}}{\partial p} \times \nabla_p \omega - \omega \frac{\partial \zeta_{ap}}{\partial p}. \tag{37}$$

Beneath the stationary IPV anomaly vorticity advection, the second term on the left, would be an important term, possibly suggesting a tendency for the lower part of the vortex to be carried downwind. However, if accurate vertical velocities ω could be obtained, it would be found that the advection term would be exactly cancelled by the terms on the right-hand side of the equation, of which the dominant term is the vortex-stretching term $\zeta_{ap}\partial\omega/\partial p$ appearing on the first line. There would be ascent upstream

* We note in passing that in the absence of boundary effects this flow field is dynamically stable, by the theorems of Charney and Stern (1962) and Blumen (1968), despite what might be concluded from an incautious application of a 'barotropic instability' criterion. It can easily be verified from the bottom line of (36) that there is a reversal in the sign of the vorticity anomaly on isentropic surfaces near the central, horizontal surface $\Theta = 0$ outside the boundary $r^2 + \Theta^2 = C^2$ of the central IPV anomaly, a phenomenon complementary to that noted in statement (iv) above. An absolute vorticity configuration of this sort might be taken as suggesting, incorrectly, that the vortex could be barotropically unstable to asymmetric disturbances. The suggestion would be incorrect because it is actually the IPV gradients and not the absolute vorticity gradients that are relevant, for reasons to be recalled in section 6(*b*). All the other flow fields described in the present section are likewise dynamically stable.

and descent downstream of the IPV anomaly—just enough to keep the vortex in the same place. Similarly, there would be an exact cancellation in the temperature equation, between temperature advection and adiabatic temperature changes. This complete cancellation of terms in the temperature and vorticity equations occurs if and only if the IPV anomaly, to which the presence of the vortex can be attributed, is stationary overhead. In the case of Fig. 15(a), the situation can be pictured by imagining an air parcel approaching the centre of the vortex on the lowest isentropic surface plotted, not counting the isentrope on the bottom boundary. The trajectory would of course be curved, and its closest approach to the centre of the vortex would depend on the detailed circumstances.

Observational data suggest that the same kind of thing happens in the real atmosphere, and Figs. 8 and 10 appear to provide two examples of it. In the case of Fig. 10 there was a low-level airstream flowing roughly westward relative to the upper air anomaly (Erickson 1971, Figs. 1 and 4). Vertical motion was computed from the omega equation and confirmed by "good agreement between areas of major cloudiness and areas of upward total ω" after an easterly wave in the low-level airstream had supplied enough moisture to the middle troposphere for the cloud to form (*op. cit.*, pp. 74, 76). Vortex stretching was strongly positive to the east of the anomaly, which rapidly became cloudy. It was mostly negative to the west of the anomaly, which remained clearer at first. This is broadly consistent with what one expects from the foregoing reasoning. Of course the effects of neighbouring anomalies, frictional convergence in the planetary boundary layer, and diabatic effects such as latent heat release, would have to be added in order to obtain a completely realistic picture. In the case of Fig. 8, it appeared that air parcels from low levels were approaching the upper air IPV anomaly mostly from the south, the air motion again being viewed relative to the upper air cyclone. There was a corresponding region of precipitation lying mainly to the south (Peltonen 1963; Palmén and Newton 1969, Figs. 10.7a, b), as well as a small region of stronger precipitation just under the centre of the cyclone, which may have been associated with frictional convergence as well as with reduced static stability.

In some cases it may be more natural to think of the upper air IPV anomaly as moving relative to the lower layers; the characteristic pattern of vertical motion may then be described as large-scale ascent and vortex stretching ahead of the anomaly, and descent and vortex shrinking behind it. Such a pattern is conspicuous in, for example, Fig. 11 of Petterssen and Smebye (1971), where vertical motion was estimated for a situation somewhat like that of the present Fig. 5.

One may think of, say, an eastward-moving upper air anomaly like that in Figs. 5(c)–(f) as acting on the underlying layers of the atmosphere somewhat like a broad, very gentle 'vacuum cleaner', sucking air upwards towards its leading portion 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, in the manner assumed in the derivation of theoretical results such as the omega equation, and the Sawyer–Eliassen equation used in discussions of two-dimensional models of frontogenesis. If an IPV 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; cf. the discussion of statements (i)–(iv) of section 3. Note that the strength of the suction effect will increase with the relative speed at which the upper air anomaly is advected.

The sense of the vertical motion, and the associated contribution to surface development, will often be correctly given by the well-known PVA (positive vorticity advection) rule. The foregoing considerations show that, as recently emphasized by Bleck and

Mattocks (1984), the same rule should work when IPV advection is used in place of vorticity advection. Moreover this avoids the dangers (Hoskins *et al.* 1978) inherent in traditional vorticity arguments. The reasons why the two viewpoints agree in many circumstances of interest are, first, the practice of applying the PVA concept only to upper air charts "near the level of maximum wind" (Palmén and Newton 1969, p. 317ff, and refs.), second, the observed fact that the strongest IPV anomalies in the free atmosphere are often upper air anomalies (which is understandable in terms of the proximity of the stratospheric high-PV reservoir), and, third, the theoretical fact that within an IPV anomaly of simple shape the induced vorticity anomaly has the same sign as the IPV anomaly, point (iii) of section 3.

There is a case to be made, perhaps, for permitting the acronym 'PVA' to be read alternatively as (positive) '*potential* vorticity advection'. Be that as it may, the thought-experiment described above is typical of many telling illustrations of the fact that it is often simplest to think directly in terms of IPV anomalies and their advection, whenever IPV information is available. Equivalent descriptions in terms of vorticity advection and stretching, and temperature advection and adiabatic temperature change, will always be valuable, of course, but the conceptual task of linking all these elements in a mutually consistent way is often quite complicated. 'Mutual consistency', in the sense in which it is usually understood (e.g. Palmén and Newton, p. 320) requires explicit consideration of the vertical motion field, simultaneously with the application of some balance condition. The alternative description in terms of IPV advection simplifies the problem, with no loss of accuracy in principle, by allowing for the effects of vertical motion in a way that is implicit but automatically self-consistent. The application of the balance condition, and the extraction of explicit information about the vertical motion field, as well as the other fields, can then be considered as a subsequent, *conceptually separate* problem to be dealt with under the heading of the invertibility principle.

The deduction of the vertical velocity field is discussed further in the appendix, where the direct use of the IPV concept will be compared with other methods. Each method has different advantages, and no single one is best for all practical and conceptual purposes. In connection with the thought-experiment described above, we may also refer to an elegant statement by Kleinschmidt (1957, p. 121) of the essential point about advectively-induced vertical motion. His statement reads "if a cyclone lies in a baroclinic basic flow, ascending motion takes place on that side of the cyclone towards which the thermal wind vector of the basic flow points." This is applicable to a cyclonic IPV anomaly at any altitude.

We next show how, in order to apply the invertibility principle in its full generality, the concept of the 'IPV anomaly' must be enlarged to take account of near-surface temperature anomalies.

5. Anomalies at the lower boundary, and the invertibility principle for general, time-dependent flow

(*a*) *Surface and near-surface anomalies*

Figures 16(a), (b) give two more computations illustrating a contrasting idealized situation in which there is no upper air IPV anomaly but in which θ has an anomaly at 1000 mb. The model atmosphere is still stably stratified all the way down to the surface. The reference-state condition (17a) is interpreted as if the isentropes which intersect the 1000 mb surface were bunched along that surface (see insets to figures), corresponding, as Bretherton (1966a) pointed out in the context of quasi-geostrophic theory, to static stability and PV values of very large magnitude in a region of very small thickness, like

a surface charge in electrostatics. In particular, a warm surface potential temperature anomaly is equivalent to a cyclonic IPV anomaly concentrated at the surface. As statement (i) of section 3 predicts, it induces a cyclonic vortex (Fig. 16(a)). A cold surface anomaly induces an anticyclonic vortex (Fig. 16(b)).

Note that the magnitudes of the induced wind and pressure fields are very significant. For instance the computation on which Fig. 16(a) is based implies that a 10 K warm surface anomaly, on the scale shown, can induce a surface pressure low of −31 mb relative to its surroundings. The figure caption gives more detail.

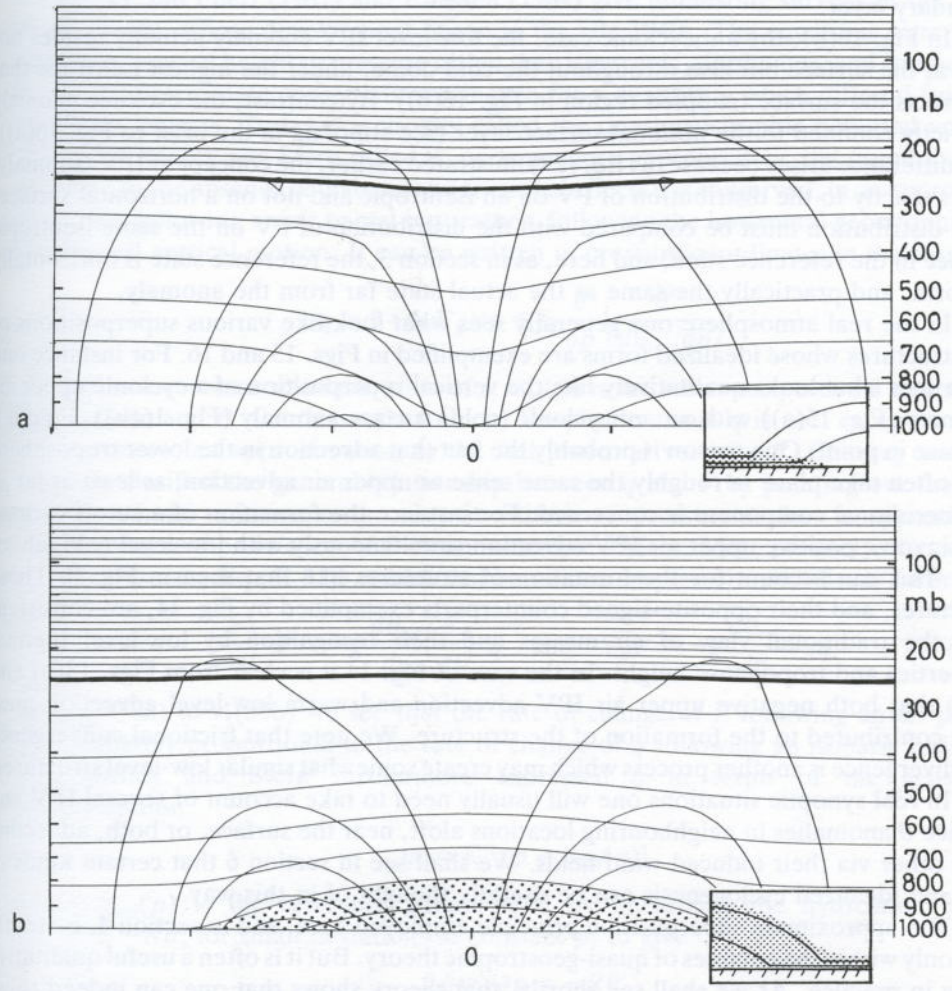

Figure 16. Circularly symmetric flows induced by simple boundary temperature anomalies. The basic situation, and the method of solution, is as in Fig. 15 except that here there is no potential temperature variation along the tropopause, but the boundary potential temperature is taken to vary in the manner $\frac{1}{2}A\{\cos(\pi\tilde{r}/r_0) + 1)$ for $\tilde{r} < r_0 = 1667$ km. The amplitude A was taken to be +10 K in (a) and −10 K in (b). The thick line again represents the tropopause, and the thin lines the isentropes every 5 K and the transverse velocity every 3 m s^{-1}. The zero isotach on the axis of symmetry is omitted. The warm anomaly induces a cyclonic circulation, and the cold anomaly an anticyclonic circulation. The surface pressure anomalies are −31 mb and +18 mb respectively, and the maximum (surface) winds are about 16 m s^{-1} and 17 m s^{-1} respectively. The relative vorticity extrema are at the surface and have values $+0{\cdot}8f$ and $-0{\cdot}5f$ respectively. The insets and stippling suggest the interpretation of the warm surface potential temperature anomaly as equivalent to a cyclonic IPV anomaly, and the cold surface potential temperature anomaly as part of an anticyclonic IPV anomaly (see text). For more details of the method of computation, see Thorpe (1985). Courtesy of A. J. Thorpe.

The qualitative statements (i)–(iv) of section 3 all hold for surface temperature anomalies interpreted in this manner. Note that in order to have positive static stability everywhere, within the notional surface layer as well as above it, the cold (anticyclonic) anomaly must be regarded as being embedded in a layer of high static stability and PV in the reference state, as suggested by the inset to Fig. 16(b). However, this is just a conceptual device and it is probably best simply to think in terms of the surface θ anomalies and the rule that 'warm is cyclonic and cold is anticyclonic'.

In the real atmosphere, θ 'at the ground' is probably best interpreted to mean θ, or temperature T, on some convenient isobaric surface lying just above the planetary boundary layer.

In Fig. 16(b), the anticyclonic case, the low-level IPV anomaly actually resides not only at the surface but also throughout the cold dome, under the highest isentrope that intersects the surface (stippled region in Fig. 16(b)). By contrast, the cyclonic anomaly is strictly confined to the notional surface layer (see stippling in the inset to Fig. 16(a)). The difference arises because, as Eq. (29) illustrated earlier, the concept of IPV 'anomaly' refers strictly to the distribution of PV on an isentropic and not on a horizontal surface. That distribution must be compared with the distribution of PV on the same isentropic surface in the reference state; and here, as in section 3, the reference state is horizontally uniform, and practically the same as the actual state far from the anomaly.

In the real atmosphere one generally sees what look like various superpositions of the structures whose idealized forms are exemplified in Figs. 15 and 16. For instance one often sees what looks qualitatively like the vertical superposition of a cyclonic upper air anomaly (Fig. 15(a)) with an anticyclonic (cold) surface anomaly (Fig. 16(b)). Figure 8 is a case in point. One reason is probably the fact that advection in the lower troposphere may often take place in roughly the same sense as upper air advection, as least as far as the meridional component is concerned. For instance the formation of a cutoff cyclone may involve positive upper air IPV advection simultaneously with low-level cold advection. This can account for the formation of structures like that seen in Fig. 8. These structures, and their opposite-signed counterparts exemplified by Fig. 14, are consistent with the traditional view of air masses and their recognition by low-level thermal properties and tropopause height. In the case of Fig. 14 it is clear from Figs. 13(b) and 11(c) that both negative upper air IPV advection and warm low-level advection must have contributed to the formation of the structure. We note that frictional convergence and divergence is another process which may create somewhat similar low-level structures.

In real synoptic situations one will usually need to take account of several IPV and surface θ anomalies in neighbouring locations aloft, near the surface, or both, advecting each other via their induced wind fields. We shall see in section 6 that certain kinds of real and idealized cyclogenesis can be usefully thought of in this way.

The approximate superposition principle employed here, and in section 4, is strictly true only within the confines of quasi-geostrophic theory. But it is often a useful qualitative guide in practice. As we shall see shortly, that theory shows that one can indeed think of the interior and surface IPV anomalies as qualitatively like the interior and surface charge distributions of a problem in electrostatics, and the induced flow structures as analogous to the (superposable) induced electric fields, with the Montgomery potential or the geopotential height playing the role of the electrostatic potential. For the purpose of this analogy it must be understood that the vertical coordinate is scaled according to (33a) (and that the Coriolis parameter is being assumed not to vary too much), and that the non-Boussinesq effects indicated by (33c, d) are not overwhelmingly important. In applications where quasi-geostrophic theory is not sufficiently accurate, we can still think of the geopotential height field induced by an IPV or surface θ anomaly as somewhat

analogous to an electrostatic potential, but in a nonlinear medium whose dielectric properties are locally altered by the induced field. Such local nonlinear effects are apt to be important in regions of very high baroclinicity such as fronts.

(*b*) *Quasi-geostrophic theory*

It was quasi-geostrophic theory, in the form developed in the landmark paper of Charney and Stern (1962), a decade after Kleinschmidt's early publications, which provided the first clear way of expressing the invertibility principle which was also sufficiently general to apply directly to time-dependent, non-circular motion with variable f. Charney and Flierl (1981) and Eliassen (1984) give interesting surveys of the history of the theory, whose development began in the late 1940s. The equations were usually written in terms of height or pressure rather than isentropic coordinates, and consequently led to an approximate conservation principle for a quantity q which differed from P in significant aspects. Today q is usually called the quasi-geostrophic potential vorticity, although Charney and Stern called it, perhaps more appropriately, the pseudopotential vorticity. In the absence of diabatic and frictional effects it is conserved, in an approximate sense, not following an air parcel but, rather, following the horizontal geostrophic flow, ignoring all vertical motion. It can be written in pressure coordinates in the form

$$q = f + \mathbf{k}.\nabla_p \times \mathbf{v} + f_0 \frac{\partial}{\partial p}\left(\frac{\theta'}{d\theta_{\mathrm{ref}}/dp}\right) \tag{38}$$

where f_0 is a constant standard value of f, $\theta_{\mathrm{ref}}(p)$ is the reference potential temperature distribution, and θ' the deviation from it. Despite its name q is not, except in special circumstances, an approximation to the full potential vorticity P. Charney and Stern elegantly clarified the general relationship between q and P by pointing out that q is a quantity whose variation on a constant altitude or pressure surface is approximately proportional to the variation of P on an isentropic surface. More precisely, they derived a result (*op. cit.*, Eq. (2.31)), which can be written in pressure coordinate form as

$$\left(\frac{\partial}{\partial t}\right)_\theta P \simeq -g\frac{d\theta_{\mathrm{ref}}}{dp}\left(\frac{\partial}{\partial t}\right)_p q \quad \text{and} \quad \nabla_\theta P \simeq -g\frac{d\theta_{\mathrm{ref}}}{dp}\nabla_p q. \tag{39a, b}$$

Adding (39a) to **v**.(39b) we see that the rate of change of P following an air parcel is approximately proportional to the rate of change of q following an isobaric trajectory.

Charney and Stern furthermore expressed q in terms of the geostrophic streamfunction

$$\psi' = f_0^{-1}\{\phi - \phi_{\mathrm{ref}}(p)\} \tag{40}$$

where $\phi_{\mathrm{ref}}(p)$ is the reference geopotential by linearizing the hydrostatic relation $\partial\phi/\partial p = -R\theta$, for small deviations at constant p, to give

$$f_0 \partial\psi'/\partial p = -R\theta' \tag{41}$$

where $R = R(p)$ is the quantity defined in (20). Substituting this and the geostrophic approximation $\mathbf{v} = \mathbf{k}\times\nabla\psi'$ into (38), and introducing the reference-state q field,

$$q_{\mathrm{ref}} = f, \tag{42}$$

we can write the result in the form (*op. cit.*, Eq. 2.25b),

$$\boxed{\mathcal{L}_{\mathrm{g}}(\psi') = q - q_{\mathrm{ref}}} \tag{43}$$

where the linear operator $\mathscr{L}_g(\ )$ is defined by

$$\mathscr{L}_g(\psi') = \nabla_h^2 \psi' + f_0^2 \frac{\partial}{\partial p}\left(\mathcal{N}^{-2}\frac{\partial \psi'}{\partial p}\right),$$

∇_h^2 being the usual two-dimensional horizontal Laplacian, and

$$\mathcal{N}^2 = -Rd\theta_{ref}(p)/dp = N_{ref}^2/g^2\rho_{ref}^2,$$

another measure of the static stability.

This result of Charney and Stern's made it obvious that, if q is regarded as kno so that the PV anomaly, defined as the right-hand side of (43), is known, then ψ' ca deduced by inverting the three-dimensional, Laplacian-like operator $\mathscr{L}_g$, just as we invert the two-dimensional Laplacian of barotropic (two-dimensional aerodynamic) fl In cases where the lower boundary condition is simply one of constant potential t perature at $p = p_0 = 1000$ mb, we can take $\theta' = 0$ there so that, from (41), the bound conditions for inverting (43) take the homogeneous (unforced) form

$$\boxed{\partial \psi'/\partial p = 0 \quad \text{at } p = 0, p_0.}$$

If, on the other hand, the potential temperature is not uniform on the lower bound a well-known mathematical device enables us to express this, whilst retaining the ho geneous boundary condition (46), by adding a contribution

$$-f_0^2\left(\mathcal{N}^{-2}\frac{\partial \psi'}{\partial p}\right)_{p=p_{0-}} \delta(p - p_0) \quad = \quad \left\{\frac{f_0\theta'}{-d\theta_{ref}/dp}\right\}_{p=p_{0-}} \delta(p - p_0),$$

to q on the right-hand side of (43), where $\delta(\ )$ is the Dirac delta function and suffix $p = p_{0-}$ means evaluation just above the boundary, i.e. just above the imp discontinuity in $\partial\psi'/\partial p$. The Dirac delta function $\delta(x)$ represents a quantity concentr at the origin $x = 0$ (a classical example being surface charge density in electrostatics) is defined such that

$$\int_{-\varepsilon}^{\varepsilon} \delta(x)\, dx = 1$$

for any positive number ε. The device (47), due to Bretherton (1966a), is one wa expressing mathematically the idea suggested by the insets in Fig. 16 but in a form rela more directly to q than to P. It allows us to think of the surface temperature distribu as part of the PV distribution. Thus, in an approximate but beautifully simple way, boundary value problem (43), (46) expresses invertibility under geostrophic bala relative to the standard reference state assumed in quasi-geostrophic theory, in w the static stability is horizontally uniform, but f is allowed to vary provided that its cha over the horizontal scale L of the motion is small in comparison with f_0. There is no restriction to axisymmetric, steady motion as there was with Eqs. (29) and (32): theory applies to general, time-dependent flow.

Bretherton further noted that under adiabatic conditions, with $\omega = Dp/Dt =$ $p = p_0$,

$$\left(\frac{\partial}{\partial t} + \mathbf{v}.\nabla_p\right)\theta' = 0, \quad \text{i.e.} \quad \left(\frac{\partial}{\partial t} + \mathbf{v}.\nabla_p\right)\left(\frac{f_0\theta'}{d\theta_{ref}/dp}\right) = 0 \quad \text{at} \quad p = p_0$$

so that, even in the extended sense implied by (47), q is, indeed, still conserved o isobaric trajectory.

Once ψ' is determined from (43) and (46) the geopotential, potential temperature and horizontal velocity fields may all be calculated directly from (40), (41) and the geostrophic relation $\mathbf{v} = \mathbf{k}\times\nabla\psi'$. The vertical velocity field may be found from the adiabatic thermodynamic equation, or alternatively, and strictly diagnostically, from the quasi-geostrophic omega equation (e.g. Eliassen 1984):

$$\mathcal{N}^2\nabla_h^2\omega + f_0^2(\partial^2\omega/\partial p^2) = g(\psi'), \tag{49}$$

together with the boundary conditions $\omega = 0$ at $p = 0, p_0$. Here $g(\psi')$ is a nonlinear function which may be written in various ways, some of which are related to the discussion in section 4. Details, including frictional and diabatic contributions, are given in the appendix.

In terms of the height-like coordinate (35) and the static stability $\tilde{N}^2$ defined by (34), the operator $\mathcal{L}_g$ can be shown to take a form resembling more closely the height coordinate form originally derived by Charney and Stern:

$$\mathcal{L}_g(\psi') = \nabla_h^2\psi' + \frac{f_0^2}{\rho_s}\frac{\partial}{\partial z}\left(\frac{\rho_s}{\tilde{N}^2}\frac{\partial\psi'}{\partial z}\right), \tag{50}$$

where

$$\rho_s(z) = \{\theta_0 R(p)\}^{-1} = (p_0/\kappa c_p\theta_0)(p/p_0)^{1/\gamma}, \tag{51}$$

a standard density distribution (which would be identical to the reference density distribution if the reference atmosphere were neutrally stable). The appropriate form of (47) is then

$$f_0^2\left(\tilde{N}^{-2}\frac{\partial\psi'}{\partial z}\right)_{z=0_+}\delta(z) = \left\{\frac{f_0\theta'}{d\theta_{\mathrm{ref}}/dz}\right\}_{z=0_+}\delta(z), \tag{52}$$

by virtue of (35) and the second of (34). Notice incidentally how quasi-geostrophic theory, which in effect is a 'weak-anomaly' theory, treats both the anticyclonic, cold-dome anomaly exemplified in Fig. 16(b) and the cyclonic surface anomaly exemplified in Fig. 16(a) as if they were concentrated at the surface, even though P has (isentropically) anomalous values throughout the finite, stippled region in the case of Fig. 16(b).

Bretherton also pointed out that the same mathematical device can be used even in the presence of smoothed, large-scale topography of height h (a function of horizontal position) provided that the approximation

$$\theta_{\mathrm{ref}}(h) - \theta_{\mathrm{ref}}(0) \simeq h(d\theta_{\mathrm{ref}}/dz)_{z=0} \tag{53}$$

is acceptable where θ_{ref} is now expressed for convenience as a function of z instead of p, as in the second of (52). The approximation (53) can be shown by standard scaling arguments to be formally consistent with the other approximations used in quasi-geostrophic theory. Then under adiabatic conditions (48) is replaced, again using z as vertical coordinate in place of p, by the statement that total potential temperature is conserved for an air parcel following the topography,

$$(\partial/\partial t + \mathbf{v}.\nabla_z)(\theta' + h\,d\theta_{\mathrm{ref}}/dz) = 0 \quad \text{at} \quad z = 0. \tag{54}$$

Here it is a self-consistent approximation, according to quasi-geostrophic theory, to evaluate all quantities, including $\mathbf{v}$ and θ', at $z = 0$ rather than at $z = h$. This is another 'weak-anomaly' approximation. With all these approximations it follows that the problem can once again be given the simple form (43), (46) provided that we add to q a contribution

$$\left\{\frac{f_0\theta'}{d\theta_{\mathrm{ref}}/dz} + f_0 h\right\}_{z=0+}\delta(z) \tag{55}$$

in place of (52), and to the reference distribution q_{ref} a contribution

$$f_0 h \delta(z) \tag{56}$$

so that the *anomaly* in q is still proportional to θ', and therefore to $\partial \psi'/\partial z$, just above $z = 0$, and q is still conserved on isobaric trajectories.

The quasi-geostrophic approximations are inaccurate, of course, for quantitative purposes, in many situations of interest. In particular the assumption of small deviations from the reference stratification is very inaccurate indeed near the tropopause, a point well illustrated in Figs. 8–10 and Fig. 15. However, the link with quasi-geostrophic theory makes evident the central role of potential vorticity in many processes such as instability, Rossby wave propagation, quasi-two-dimensional turbulence, critical-layer processes, and so on, all of which are usually studied theoretically by means of quasi-geostrophic theory. The theory has remained important, not on account of quantitative accuracy, but rather because it incorporates the fundamental qualitative insights arising from PV conservation and the invertibility principle, which do appear to carry over to more accurate representations of the real atmosphere. It also correctly suggests, through results like (54) and (55), that it is the potential temperature at the ground which is the dynamically significant quantity there, for the purposes of IPV thinking.

Perhaps the most basic weakness of quasi-geostrophic theory in its standard pressure coordinate form, as just outlined, arises from the fact that, in order to construct the large-scale distribution of q from that of P by integrating the Charney–Stern relation (39b), it is necessary to assume that each isentropic surface is close to a given isobaric surface. This may be a good local approximation in some cases, but the approximation is seldom uniformly valid over large horizontal distances. This is the price paid for the theoretical convenience of a simple description based on isobaric surfaces which, among other things, permits the approximate application of the lower boundary condition at a constant value p_0 of the pressure coordinate p, or equivalently at $z = 0$. This simplification of the lower boundary condition is a necessary device in the construction of many standard theoretical solutions representing dynamical processes of interest.

Similar formulations in isentropic coordinates, comprising generalizations of (32), can overcome the problem of horizontal nonuniformity at the inevitable cost of complicating the lower boundary condition. Such formulations have been used as the bases of numerical models in a number of important papers beginning with Danielsen and Diercks (1967), Eliassen and Raustein (1968, 1970), and Bleck (1973, 1974), following a suggestion by Charney and Phillips (1953). For more recent developments in the use of isentropic coordinates in numerical models, see for instance Uccellini *et al.* (1979), Bleck (1984), and references therein.

Many other ways of improving the standard quasi-geostrophic theory have been proposed, some of which attempt to tackle the problem of horizontal nonuniformity whilst retaining isobaric or sigma coordinates to simplify the lower boundary condition, sometimes at the cost of losing the PV-conservation principle, and most of which seek to improve the accuracy of the balance condition in some way or other. Recent theoretical work is reported, for instance, in the papers by Gent and McWilliams (1984, and refs.) and Williams and Yamagata (1984, and refs.). Amongst the many papers on this topic, a particularly important milestone is the recent pair of papers by Salmon (1983, 1985). They show, for the first time as far as we are aware, how improvements on quasi-geostrophic theory may be systematically constructed in such a way that exact analogues of PV as well as energy conservation are retained. The method used appears to have far-reaching implications for refining the balance and invertibility concepts beyond their present state of development. It goes to the heart of the matter by exploiting the fundamental connection between conservation relations and symmetries.

In the next sub-section we review the simplest IPV-conserving improvement to quasi-geostrophic theory, namely semi-geostrophic theory, together with a significant extension to it discovered by Salmon. Then, in section 5(*d*), we briefly discuss a rather different approach to expressing the invertibility principle, using the concepts of nonlinear normal-mode initialization. This promises an advantage by way of computationally uniform behaviour across the tropics.

(*c*) *Semi-geostrophic theory and Salmon's generalization*

Semi-geostrophic theory improves on quasi-geostrophic theory while retaining isobaric coordinates, by:

(i) improving on the geostrophic balance condition in the wind velocities used to advect potential temperature, momentum, and potential vorticity (the so-called geostrophic momentum approximation introduced by Eliassen (1948)); and

(ii) alleviating the horizontal nonuniformity problem by means of a quasi-Lagrangian coordinate transformation (Hoskins 1975).

This particular choice of improvements may seem arbitrary at first sight, but PV conservation is retained, and the coordinate transformation puts the equations into a simple form which allows insights from quasi-geostrophic theory to be carried over into situations in which the quasi-geostrophic approximations themselves are poor, including cases of extreme horizontal nonuniformity. For example, the theory handles frontogenetic situations far more accurately than quasi-geostrophic theory (e.g. Hoskins and West 1979), and also shows that much of the extra accuracy can be thought of as coming simply from a geometrical distortion of the corresponding quasi-geostrophic solution. The theory can describe large-amplitude distortions of the tropopause such as those exemplified by Figs. 8, 9 and 15 (Hoskins 1982, and refs.). However, the theory as originally conceived does not permit f to vary. It is this important limitation that has recently been removed by the work of Salmon (1985) already referred to.

Assuming constant f to begin with, and denoting the geostrophic wind by $\mathbf{v}_g = (u_g, v_g, 0)$, we introduce a transformation of the horizontal independent variables,

$$x = X - v_g/f, \qquad y = Y + u_g/f \tag{57}$$

and write Z for z with the understanding that $\partial/\partial Z$ means differentiation at constant X, Y (whereas $\partial/\partial z$ implies constant x, y). It can be shown that the geostrophic flow and potential temperature are all related to a streamfunction $\Psi(X, Y, Z)$ in the transformed, XYZ space, which we shall call 'semi-geostrophic space', by

$$(u_g, v_g, \theta) = (-\partial\Psi/\partial Y, \partial\Psi/\partial X, g^{-1} f\theta_0 \partial\Psi/\partial Z). \tag{58}$$

Here θ is the full potential temperature, not the deviation from the reference state, and θ_0 is a constant, representative value of θ. Following Hoskins' (1975) derivation of the semi-geostrophic equations we can show that the twisting terms disappear in the transformed system, so that the semi-geostrophic approximation to P is simply

$$P = \rho_s^{-1} \zeta_a \partial\theta/\partial Z, \tag{59}$$

where ρ_s is defined by (51), and the (isobaric) absolute geostrophic vorticity

$$\zeta_a = \frac{f}{1 - (1/f)(\partial v_g/\partial X - \partial u_g/\partial Y)}$$

to a good approximation under the conditions for which semi-geostrophic theory is applicable (Hoskins and Draghici 1977). Thus Ψ is related to P by the equation

$$N_P^2(\partial^2\Psi/\partial X^2 + \partial^2\Psi/\partial Y^2) + f^2\,\partial^2\Psi/\partial Z^2 = g\theta_0^{-1}\rho_s P, \tag{60}$$

where $N_P^2 = gf^{-1}\theta_0^{-1}\rho_s P$. Again Ψ is related to P by a simple Laplacian-like operator (although the forcing function also appears on the left, in N_P, just as P does on the left of (29)). As in quasi-geostrophic theory, it is possible to define a boundary contribution to P (and N_P) associated with the θ distribution at the bottom boundary. The contribution to P can be expressed as

$$f^3 \frac{\theta_0}{g}\left[\frac{1}{\rho_s}\frac{\partial\Psi}{\partial z}\Big/\left(f - \frac{\partial^2\Psi}{\partial x^2} - \frac{\partial^2\Psi}{\partial Y^2}\right)\right]_{Z=0+} \delta(Z) = \left[\frac{1}{\rho_s}\zeta_a\theta\right]_{Z=0+}\delta(Z). \tag{61}$$

With this convention, invertibility is expressed by solving (60) subject to $\partial\Psi/\partial Z = 0$ on $Z = 0$. It should be noticed that in this formulation the function $P(X, Y, Z)$ implicitly supplies both the reference-state information and the IPV information.

In the interior, differentiation of (60) with respect to Z, after division by N_P^2, gives

$$\partial^2\theta/\partial X^2 + \partial^2\theta/\partial Y^2 + \frac{\partial}{\partial Z}\{(f^2/N_P^2)(\partial\theta/\partial Z)\} = 0 \qquad (Z > 0), \tag{62}$$

from which θ may be determined, although it is simpler to note that (58) gives both θ and the horizontal geostrophic velocities once Ψ is found.

For frictionless, adiabatic motion the conservation of P and θ both take the same form

$$\left(\frac{\partial}{\partial t} + u_g\frac{\partial}{\partial X} + v_g\frac{\partial}{\partial Y} + W\frac{\partial}{\partial Z}\right)(P, \theta) = 0, \tag{63}$$

the vertical velocity $W = DZ/Dt$ now entering both equations. Advection of P and θ is performed by the horizontal geostrophic velocities *in semi-geostrophic*, XYZ space. This gives a more accurate representation of advection in physical, xyz space, improvement (i), which is important near jets and fronts. As noted by Hoskins and Draghici (1977), the vertical velocity W may be obtained from Ψ (or P) by solving an omega equation in semi-geostrophic space.

In order to extend the theory to the case of a variable Coriolis parameter $f(x, y)$, while retaining conservation laws analogous to (63), the key step is to replace (57) by

$$x = X - v_g/f(X, Y), \quad y = Y + u_g/f(X, Y) \tag{64}$$

and to use $f(X, Y)$ as the Coriolis parameter in the definition of the geostrophic wind (Salmon 1985). In other words, one uses the value of the Coriolis parameter at the transformed, and not at the physical, position of an air parcel. The proof that the resulting equations possess the required conservation properties is a straightforward extension of the analysis given by Salmon (1985), to whose penetrating discussion the reader is referred for more detail.

An alternative approach to the geostrophic momentum equations is to use isentropic coordinates. In the case considered here f may be allowed to vary slightly about a standard value f_0. Salmon's method would presumably allow the restriction on f to be relaxed still further. In $xy\theta$ space, the geostrophic flow and pressure field are given by

$$(f_0 u_g, f_0 v_g, \Pi) = (-\partial M/\partial y, \partial M/\partial x, \partial M/\partial\theta) \tag{65}$$

where $\Pi(p)$ is the Exner function defined by (19). The potential vorticity is approximated by

$$P = -g\frac{f + \partial v_g/\partial x - \partial u_g/\partial y}{\partial p/\partial\theta}. \tag{66}$$

Substitution from (65) gives, on rearranging,

$$(1/f_0)(\partial^2 M/\partial x^2 + \partial^2 M/\partial y^2) + \tilde{P}\partial^2 M/\partial\theta^2 = -f, \tag{67}$$

where $\tilde{P}$ is the quantity defined in (34). If the Boussinesq approximation is not made, it should be noted that the definition (34) gives

$$\tilde{P} = (p_0/\kappa g c_p^{1/\kappa})(\partial M/\partial\theta)^{1/(\gamma-1)} P.$$

There is then a weak nonlinearity in the elliptic equation (67), because of the factor involving $(\partial M/\partial\theta)^{1/(\gamma-1)}$. Suitable boundary conditions in the horizontal are easily specified but the usual vertical boundary condition of θ specified on a given pressure $(\partial M/\partial\theta)$ or height $(M - \theta\partial M/\partial\theta)$ surface makes numerical solution awkward except in the case of uniform θ on the boundaries. Methods for overcoming this problem have been proposed by Charney and Phillips (1953), Danielsen and Diercks (1967), Eliassen and Raustein (1968, 1970), Bleck (1973, 1974, 1984), Uccellini *et al.* (1979) and others. Once M has been determined the other variables follow from (19) and (65). The ageostrophic flow which is needed for a more accurate representation of horizontal advection can be determined from an extension of the omega equation analysis described in Hoskins and Draghici (1977).

It can be shown that for the case of a circularly symmetric distribution of PV on an f plane, differentiation of (67) with respect to r gives an approximation to the exact equation (29) in which terms like v/r are neglected in comparison with f. This is consistent with the present geostrophic approximation to the full gradient wind balance.

The work of Bleck (1973) cited earlier presented results from a quasi-geostrophic, isentropic coordinate forecast model in which P and boundary θ were advected geostrophically and then an equation like (67) solved for M. In cases of intense development, which were underestimated by the standard forecast models, spectacular success was obtained even though the numerical resolution of the model was modest by today's standards. In a subsequent paper (Bleck 1974) many more forecasts were investigated using an isentropic coordinate model including ageostrophic horizontal advection, hence resembling the semi-geostrophic models just described, as well as a primitive equation model in isentropic coordinates. In those cases it was concluded that a standard operational forecast model showed the best overall performance. However, considering the very great sophistication of the operational model in comparison with the experimental isentropic models developed by an individual investigator, Bleck's results were an important achievement at the time.

Differentiation of (67) with respect to θ gives an equation for Π:

$$(1/f_0)(\partial^2\Pi/\partial x^2 + \partial^2\Pi/\partial y^2) + \frac{\partial}{\partial\theta}(\tilde{P}\partial\Pi/\partial\theta) = 0. \tag{68}$$

The four simple elliptic equations, (60) and (62) for Ψ and θ in XYZ space, and (67) and (68) for M and Π in $xy\theta$ space, along with those for the static stability variables $\partial\theta/\partial Z$ and $\partial\Pi/\partial\theta$ obtained from the vertical derivatives of (62) and (68), may be used for qualitative discussion of the fields induced by any PV (or boundary θ) anomaly except near the equator, and to justify statements (i)–(iv) near the end of section 3. Within an isolated positive PV anomaly, for instance, there is a tendency for the θ-surface separation to approach the small value which would be associated with its PV if there were no change in ζ_a (statement (iii) near the end of section 3). However, the closer together the θ surfaces are in the anomaly region, the more they must be separated in the surrounding region, for a given reference state of finite horizontal extent. The smoothing properties of the inverse elliptic operators make for a monotonic transition between the two regions.

The balance condition now implies a positive ζ_a anomaly. The result is a compromise in which, in both regions, the θ surfaces are less distorted than their PV would imply if there were no change in ζ_a. The sense of the isentropic slopes between the two regions corresponds to a cyclonic circulation everywhere (statement (i) of section 3). The perturbations all decay away from the region, the vertical length scale in semi-geostrophic space being fL/N_P where L is the horizontal length scale of the flow structure in semi-geostrophic space, cf. (33a). For a weak anomaly the scale in physical space is practically the same.

Finally, we note that solutions of the two-dimensional semi-geostrophic equations for situations comprising two-dimensional analogues of those shown in Fig. 15 (calculations by E. Caetano Neto, personal communication, reproduced in Figs. 4.24, 4.25 of Robertson 1984) show that two-dimensional troughs and ridges in the tropopause have very similar static stability and circulation characteristics associated with them, as expected from our qualitative arguments and from Gill's two-dimensional solutions cited in section 3.

(*d*) *Inversion by nonlinear normal-mode initialization*

With the exception of the circular-vortex equation (29), for circularly symmetric IPV anomalies in an idealized, horizontally uniform reference state with constant f, none of the foregoing theories, even Salmon's, can express the balance condition (i) of section 1(*d*) in a way that retains any useful accuracy in the tropics. One way of expressing the invertibility principle for time-dependent motions on the full sphere may be to use the concepts of normal-mode initialization pioneered by Dickinson, Williamson, Baer and Machenhauer (see the review by Daley (1981)).

The basic idea is to express the fields of wind, temperature, etc., as a superposition of modes belonging to a complete set of atmospheric linear normal modes, and then to set to zero the coefficients of all modes not belonging to a subset considered representative of balanced motion. The accuracy of the 'linear balance' thus realized is then improved by iteration on the nonlinear, topographic, and forcing terms in the equations and boundary conditions. The choice of the subset of modes retained generally involves some arbitrariness (e.g. Tribbia 1979), reflecting the fact that 'balanced motion' has no self-evidently unique meaning in the tropics. In the case of the IPV inversion problem, however, there is only one natural choice: the relevant subset consists of the Rossby and westward-propagating Rossby–gravity modes, along with the steady zonal flows. The Kelvin and inertio–gravity modes are omitted. There is no possibility of including the Kelvin mode, since (unlike any of the other linear modes) it is invisible on an IPV map.

In specialist language, the resulting inversion computation would be called an 'IPV-constrained nonlinear normal-mode initialization' (Daley, *op. cit.*). The smoothing property of the inversion operator would be expressed by the fact that the normal-mode expansion coefficients would tend to fall off more rapidly for the wind and geopotential height fields than for the PV field. The reference state would enter the computation as the standard atmosphere whose normal modes are used.

In practice it will seldom be necessary, of course, actually to solve the inversion problem, since in practice the wind and height fields are already available. It is conceptually important, however, to know that the inversion can in principle be carried out, and under what circumstances, and to know how serious are the errors made by ignoring oscillations about 'balance', however the latter is defined. These questions pose a formidable challenge to researchers working on the mathematical aspects of the subject. Present indications are that the errors will often turn out to be much less serious than the errors inherent in quasi-geostrophic theory. It should be noted that 'less serious' may

not necessarily mean 'small'. Oscillations about balance may have amplitudes which are far from negligible in the data, but may still not seriously affect the evolution of the balanced part of the motion, since the nonlinear coupling between balanced and unbalanced motion may still be small.

6. Rossby waves and shear instabilities

From now on it will be assumed that we are dealing with balanced dynamical phenomena such that the invertibility principle holds to sufficient accuracy for the purpose at hand. To the same order of accuracy, whatever it may be, the dynamical evolution can then be described solely in terms of the IPV (and surface θ) distributions, their induced wind fields, and their advective, frictional and diabatic rates of change. In the following sections it will be shown how this viewpoint, which we have referred to as 'IPV thinking'—and which represents nothing more than a logical extension of Rossby's original arguments about wave propagation—can give useful insights into a number of different time-dependent situations.

We begin with adiabatic, frictionless Rossby waves and shear instabilities, which provide some of the simplest and thoroughly studied illustrations of time-dependent interactions between IPV anomalies. In each case the anomalies arise from air-parcel displacements across a pre-existing IPV gradient or surface θ gradient. The air-parcel displacements are themselves caused by the wind fields induced by neighbouring IPV or surface θ anomalies. The insights gained are by no means limited to the linear propagation and growth mechanisms characteristic of small amplitude waves (sections 6(*a*)–(*c*)). They also illuminate the ways in which nonlinear saturation may take place at large wave amplitude (section 6(*d*)), and lead to a sharper perception of relationships between the idealized theoretical models and the behaviour of the real atmosphere (sections 6(*d*), (*e*)). They remind us, furthermore, that the basic wave and instability phenomena do not depend, in any crucial way, on the restrictive assumptions of the quasi-geostrophic theories often used to model them. Among the published lines of evidence on this last point we may mention the elegant instability study by Eliassen (1983).

(*a*) *Rossby wave propagation and the scale effect*

For the sake of definiteness we consider the usual undisturbed basic state in which the IPV gradient is directed northwards (*y* direction) so that the contours on each IPV map are zonal (parallel to the *x* axis). If this basic state is slightly disturbed, adiabatically

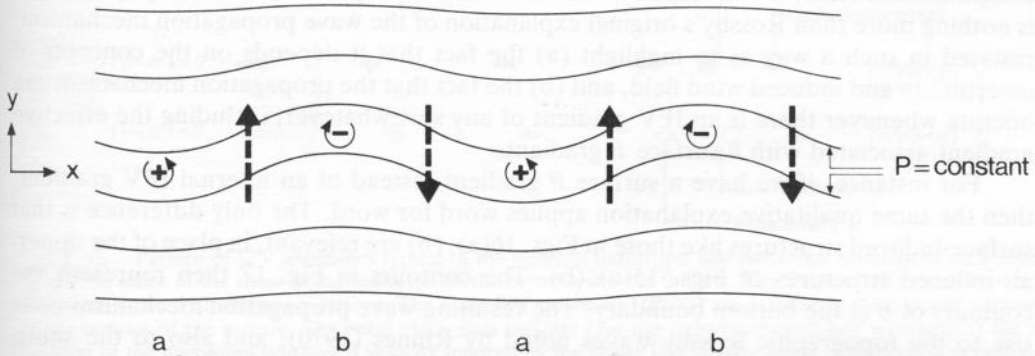

Figure 17. IPV map for a simple, *x*-periodic Rossby wave. The + and − signs respectively indicate the centres of the cyclonic and anticyclonic IPV anomalies due to southward and northward air-parcel displacements across a basic northward IPV gradient. The heavy, dashed arrows indicate the sense and relative phase of the induced velocity field (see text) which causes the westward propagation of the phase of the pattern.

and frictionlessly, then the IPV contours, which coincide with material contours, will undulate, following the material motions. The simplest case is that of a periodic, sinusoidal undulation, like that sketched in Fig. 17. If we think of the total flow as consisting of basic state plus disturbance, then the disturbance consists of a row of IPV anomalies of alternating sign, as suggested in Fig. 17 by the + and − signs enclosed by circular arrows, + meaning a cyclonic anomaly and − an anticyclonic anomaly.

To see that this pattern of IPV anomalies will tend to propagate relative to the basic state, one need only picture the qualitative nature of the wind field it induces. For this purpose we take advantage of the approximate superposability of the flow structures induced by isolated IPV anomalies, of which an idea can immediately be gained from the illustrations in Figs. 15 and 16 (although the anomalies should be considered to be weaker, with less distortion of the tropopause, if linear theory is to apply). Consider for example a case in which the anomalies are imagined to be concentrated mainly on those isentropic surfaces which lie in the region of strong basic IPV gradients near the tropopause. Then the induced flow structure, when viewed in a longitude–height cross-section, will look qualitatively like Figs. 15(a) and (b) placed side by side in overlapping positions, the whole pattern being extended periodically in the x direction as suggested by the labelling a, b, a, b, ... in Fig. 17. The winds are added vectorially everywhere, and it can be seen, in particular, that the induced wind fields tend to reinforce halfway between each pair of anomalies. The total induced north–south wind field is a quarter wavelength out of phase with the displacement field and has the sense implied by the heavy, dashed arrows in Fig. 17 (statement (i) near the end of section 3). That is, the maximum northward velocity tends to be located a quarter wavelength westward of the maximum northward displacement, and similarly for the maximum southward velocity and displacement. If one now pictures how this velocity field will advect the IPV contours in Fig. 17, one can see that the wave pattern will propagate westward relative to the basic flow.

The well-known fact that the relative westward phase speed increases with the spatial scale of the pattern is a manifestation of the scale effect already mentioned below Eq. (33d). The scale effect is most easily deducible from scaling arguments applied to the approximate form of the inversion operator defined by (44). The larger the spatial scale of an IPV anomaly pattern of given strength, the stronger the induced velocity field. The whole picture can of course be made quantitative, using the standard techniques of linear theory, for any given zonal basic state and for any given mathematical expression of the invertibility principle, from quasi-geostrophic theory onwards, depending on how much computational effort is considered worthwhile for the purpose. Qualitatively speaking it is nothing more than Rossby's original explanation of the wave propagation mechanism, restated in such a way as to highlight (a) the fact that it depends on the concepts of invertibility and induced wind field, and (b) the fact that the propagation mechanism can operate whenever there is an IPV gradient of any sort whatever, including the effective gradient associated with a surface θ gradient.

For instance, if we have a surface θ gradient instead of an internal IPV gradient, then the same qualitative explanation applies word for word. The only difference is that surface-induced structures like those in Figs. 16(a), (b) are relevant, in place of the upper-air-induced structures of Figs. 15(a), (b). The contours in Fig. 17 then represent the contours of θ at the bottom boundary. The resulting wave propagation mechanism gives rise to the topographic Rossby waves noted by Rhines (1970), and also to the short, stable waves arising in the Eady model of baroclinic instability, which are trapped near one or other of the two horizontal boundaries assumed in that model. The causes of the basic-state surface θ gradients in the two cases are, respectively, sloping topography, and

thermal wind shear at the boundary (Bretherton 1966a). Various combinations of these effects have been demonstrated in laboratory experiments by, for instance, Hide and Mason (1975, and refs.).

(*b*) *Barotropic and baroclinic shear instabilities*

An extension of the foregoing picture can now be used, following Lighthill (1963, p. 93) and Bretherton (1966b), to grasp in a direct, intuitive, yet in principle precise, way the physical nature of the simplest linear wave instabilities of a basic zonal shear flow. This in turn will give insights into the circumstances under which certain types of cyclogenesis may or may not be expected to occur, and into the limitations of idealized instability models. The simplest instabilities—by which we mean those with the simplest spatial structures—are also, in many cases, those with the fastest growth rates.

These simplest instabilities are all characterized by a pattern of IPV anomalies of the general sort shown schematically in Fig. 18. The pattern can be thought of as a pair of Rossby waves propagating side by side, or one above the other, depending on whether a barotropic or a baroclinic instability is in question. In the latter, baroclinic case, with the vertical (z) axis lying in the plane of the paper, it may be useful to picture the upper Rossby wave as being represented, again schematically, by a copy of Fig. 17 intersecting the paper with the positions of the + and − signs brought into correspondence. The angle of intersection is not exactly a right angle, because Fig. 17 represents an isentropic surface. The lower Rossby wave can be visualized in a similar way except that, since the IPV gradient is negative, the + and − signs must be exchanged in Fig. 17 and the velocity arrows reversed, as implied by the bottom row of signs in Fig. 18.

Viewed in a reference frame moving with the zonal phase speed c of the disturbance, each Rossby wave propagates against, and is held stationary by, the local basic flow. From Rossby's argument, this is dynamically possible if the sign of the basic IPV gradient is positively correlated with that of the relative zonal flow ($\bar{u} - c$), i.e. both signs positive, as in the top half of Fig. 18, or both signs negative, as in the bottom half. This is the simplest configuration consistent with the Rayleigh–Kuo and Fjørtoft necessary conditions for instability, and their generalizations (e.g. Charney and Stern 1962; Miles 1964; Pedlosky 1964; Blumen 1968; Eliassen 1983). It will be noticed that if the basic zonal flow $\bar{u}$ has a continuous profile then a steering level or critical line will be present, where $\bar{u} - c = 0$. We shall assume at first that the basic IPV gradient vanishes in some region containing the critical line, and return to the more general case afterwards. Moreover, for expository

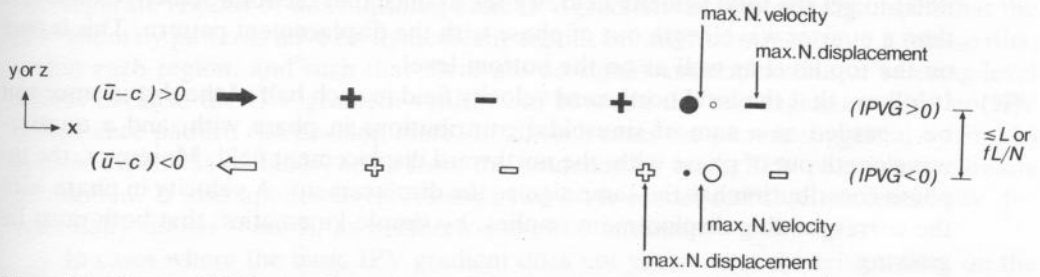

Figure 18. Pattern of IPV anomalies typical of the simplest barotropic and baroclinic instabilities on a zonal shear, the sense of the shear being indicated by the broad arrows at the left. The frame of reference has been chosen to move with the wave pattern. The basic IPV gradients (IPVG) are positive at the top and negative at the bottom of the picture. The solid black dots indicate (for the right-hand-most half wavelength) the x-location of the maximum northward velocity induced by the upper IPV pattern alone. The open dots indicate the same thing for the lower IPV pattern. The transverse dimension is northward (y) in the case of the barotropic instability, and upward (z) in the case of the baroclinic instability. Intermediate, 'mixed instability' cases also exist, to which precisely the same kind of pattern is relevant when viewed in tilted coordinates. For the fastest growing instabilities, the phase shift between the two IPV patterns is usually larger than that shown.

purposes we shall restrict attention at first to patterns whose spatial scale is such that, if the induced velocity field associated with each Rossby wave in Fig. 18 did not affect the other, then their phase propagation would be somewhat too weak to hold them stationary against the basic zonal flow.

The essence of the instability mechanism is that the induced velocity fields do, however, overlap significantly. That is why, in order to get a strong baroclinic instability of horizontal scale L, say, and simple spatial structure, the vertical separation between the two rows of IPV anomalies has to be of the order of one Rossby height fL/N or less. Similarly, the horizontal separation for a strong barotropic instability of simple spatial structure has to be of order L or less. The overlapping of the induced velocity fields has the following consequences, under the assumed conditions:

(i) Inasmuch as the IPV anomaly patterns are less than a quarter wavelength out of phase with each other, the case shown in Fig. 18, each half helps the other to propagate against the basic zonal flow. That is, the contributions to the northward velocity induced by each IPV pattern partially reinforce each other, making the phase of each pattern propagate upstream faster than it would in isolation. This is how the patterns hold themselves stationary against the basic flow, under the assumed conditions.

(ii) Because of this interdependence between the two counterpropagating Rossby waves, their relative phase tends to lock on to a configuration like that shown. For if the IPV patterns were each to shift slightly downstream, i.e. the upper pattern towards the right and the lower towards the left, so as to be more nearly in phase, then each half would help the other to propagate still more strongly, moving the patterns back upstream towards their original relative positions. Conversely, if the patterns were shifted upstream, so as to be more out of phase, then propagation would be weakened, and advection by the basic zonal flow would tend to restore the original phase relation.

(iii) Just as in Fig. 17, the northward velocity induced by the upper IPV pattern alone is a quarter wavelength out of phase with that pattern. The large black dot in Fig. 18 marks the position of the northward velocity maximum induced by the upper IPV pattern alone, for the right-hand-most wave period. This is *less than a quarter wavelength* out of phase with the bottom IPV pattern, and therefore with the bottom displacement pattern, as indicated by the position of the small black dot directly below. If we add the velocities induced by the bottom IPV pattern (open dots) to get the total velocity field, we see at once that the total velocity is also less than a quarter wavelength out of phase with the displacement pattern. This is true on the top level as well as on the bottom level.

(iv) It follows that the total northward velocity field in each half of the disturbance can be regarded as a sum of sinusoidal contributions in phase with, and a quarter-wavelength out of phase with, the northward displacement field. Moreover, the in-phase contribution has the same sign as the displacement. A velocity in phase with the corresponding displacement implies, by simple kinematics, that both must be growing.

The instability mechanism just described can be summarized in one sentence, by saying that

> 'The induced velocity field of each Rossby wave keeps the other in step, and makes the other grow.' (69)

These two effects of the induced velocity field are associated respectively with its in-quadrature and in-phase contributions. The pure, exponentially-growing normal mode

of linear instability theory describes a situation in which the two IPV anomaly patterns have locked on to each other and settled down to a common phase speed c, such that the rates of growth which each induces in the other are precisely equal, allowing the shape of the pattern as a whole to become precisely fixed, and the growth of all disturbance quantities precisely exponential.

Cases in which the spatial scale is sufficiently large that each wave in isolation would propagate *faster* than the basic zonal flow can be understood in essentially the same way. The main changes needed are in statement (i) of the foregoing, where 'help' is replaced by 'hinder', 'faster' by 'slower', and so on. Whereas in the 'helping' case the phase shift between the two IPV patterns is less than 0·25 of a wavelength, as shown in Fig. 18, in the 'hindering' case the phase shift lies between 0·25 and 0·5 of a wavelength. The relative phase tends to lock on just as before, and the summarizing statement (69) remains true.

In fact this latter case is usually the one which exhibits the largest growth rates, as would generally be expected from the fact that a larger phase shift between the two IPV anomaly patterns enables the total induced velocity to be more nearly in phase with the displacement, tending to give a larger growth rate. This is exemplified by the Eady problem (e.g. Gill 1982, Fig. 13.4c), where the fastest-growing mode has an IPV phase shift of 0·37 of a wavelength. (Note that the IPV anomalies are entirely in the form of surface θ anomalies, in the quasi-geostrophic description at least, and that the signs are reversed at the top boundary: there, warm is anticyclonic and cold is cyclonic.) Another example is given in Fig. 13.7 of the same reference, which reproduces Rayleigh's original barotropic instability calculation for the simplest unbounded, unstable shear-layer profile, with three constant-IPV regions side by side, the profile crudely approximating a continuous, tanh-like function by means of three straight-line segments. Such profiles can also be looked on as quasi-geostrophic models of the baroclinic instability of an internal vertical shear layer. In Rayleigh's problem the IPV phase shift for the fastest-growing mode amounts to 0·32 of a wavelength, again greater than 0·25.

The analyses just cited also confirm that the phase shifts in the northward velocity and geopotential height anomaly patterns are substantially less than those in the corresponding IPV anomaly patterns (respectively 0·25 and 0·18 of a wavelength in the two examples), as suggested by Fig. 18. This can be looked upon as another consequence of the smoothing property of the inversion operator.

As Bretherton's (1966b) analysis showed, the picture developed so far applies accurately and without qualification to basic flows of the kind just mentioned, having two separate regions of nonvanishing basic IPV gradients of opposite sign, such that the IPV anomaly patterns have no dynamically significant internal structure, e.g. phase tilts, within each region, and such that there are no significant critical-line or steering-level effects because the IPV gradient vanishes in between the two regions so that no IPV disturbance pattern can develop there. In all such cases the picture suggested by Fig. 18 applies exactly as it stands, apart from the tendency for larger IPV phase shifts already mentioned. It also applies to the Phillips two-layer model of baroclinic instability, for essentially similar reasons, as Bretherton showed in detail.

In cases where the basic IPV gradient does not vanish near the critical line, on the other hand, the foregoing picture needs some modification. How much depends on the circumstances. For a fast-growing instability, the modification will evidently be slight if the IPV gradient is sufficiently weak at the critical line. However, in cases like the Charney baroclinic-instability problem, the results of many published studies show that the IPV gradient at the critical line is not effectively weak. Nevertheless, it turns out, perhaps surprisingly, that the qualitative picture for the fastest-growing instability is still much the same in its essentials, provided we recognize that the 'upper' IPV anomaly

pattern in Fig. 18 is now distributed over a deep layer including the steering level. Its induced fields can still be kept *conceptually* separate from those of the lower anomaly pattern (which, as in the Eady problem, consists of surface θ anomalies); and the foregoing arguments then apply almost word for word, even though the upper and lower IPV anomaly patterns are no longer cleanly separated in space, and even though the 'upper' IPV pattern does, now, have its own internal phase structure. It tilts in the opposite sense to, but less strongly than, the main phase shift relative to the lower anomaly pattern. With realistic parameter values, the main IPV phase shift for the fastest-growing mode varies with altitude from half a wavelength at the lower surface to about a third of a wavelength aloft (R. Pierrehumbert, personal communication).*

For the fastest-growing Charney mode, the 'upper' IPV anomaly pattern is concentrated mainly in the lower half of the troposphere. Its maximum amplitude, as measured by the quantity q defined by (38), occurs near the steering level between 600 and 700 mb. Above that the disturbance amplitude of q falls off quite rapidly with height: at 200 mb it is about a third of its maximum value, and at 100 mb, about a fifth; the precise values depend on the precise choice of parameters. The flow fields induced by the 'upper' pattern alone are therefore intermediate in character between those suggested by Figs. 15 and 16, being centred neither at the tropopause nor at the ground but, rather, somewhat below mid-troposphere. When the fields induced by the surface pattern are added, the large main phase shift leads to considerable cancellation in the lower troposphere.

Instabilities on more realistic, jet-like zonal flows are very like the Charney mode in all these respects (Edmon *et al.* 1980; Hoskins 1983; Hoskins and McIntyre 1985). There is some evidence that the early stages of such instabilities may be observable in the southern hemisphere (Randel and Stanford 1985). The physical situation is one in which upper air IPV anomalies at altitudes typical of the tropopause seem to play a comparatively minor role in the early stages (but cf. section 6(d)). Such altitudes appear to be too far above the ground, in comparison with the Rossby height fL/N, and upper air winds too strong, to permit phase-locking and sustained growth of a synoptic-scale disturbance whose amplitude is initially small.

(*c*) *Lateral and vertical Rossby wave propagation*

It will prove useful to return briefly to the case of simple Rossby wave propagation on a one-signed IPV gradient, in the light of the experience gained in discussing Fig. 18.

It will have been noticed that Fig. 17 depicts a disturbance of limited extent in the y direction, its IPV anomalies being concentrated near $y = 0$. Such a state of things may well be only temporary if there are sufficiently strong basic IPV gradients outside the central region shown. Owing to the smoothing property of the inversion operator, the induced wind field of the IPV anomaly pattern will extend northward and southward beyond the anomaly pattern itself (unless constraining sidewalls are present, as in some

* The sense of the internal phase tilt within the 'upper' IPV pattern is largely governed by the fact that, at the steering level itself, the northward velocity must be *exactly* in phase with the northward displacement, as can be seen from the kinematics of a growing disturbance of small amplitude viewed in the $c = 0$ reference frame. The internal phase tilt becomes relatively larger for modes with relatively small growth rates and is important for the detailed way in which the instability balances its IPV and Eliassen–Palm fluxes (e.g. Bretherton 1966a, Eq. (4); Edmon *et al.* 1980, Eq. (4.1)). These facts can be used as the basis for an alternative way of describing Charney's instability which becomes rigorously applicable in the limit of small growth rate. Different versions of it are based on the ideas of 'critical-layer instability' (Bretherton 1966a), and 'over-reflective instability' (Gill 1965). The first version was put into quantitative form, including explicit formulae for the growth rate and a proof of the convergence of successive approximations, by McIntyre (1970, 1972), and the second has been developed very extensively in a recent series of papers by Lindzen and his collaborators (Schoeberl and Lindzen 1984, and refs.).

laboratory and theoretical models). This will in turn cause more distant IPV contours to undulate, engendering further IPV anomalies. The process can continue indefinitely if basic IPV gradients are large enough to support it out to successively greater distances in the y direction, leading to the well-known northward or southward propagation of Rossby wave activity. Vertical propagation works similarly. The quantitative meaning of having a 'large enough' basic IPV gradient depends, like the induced wind field itself, on the spatial scale of the waves. It also depends on the difference between the zonal phase speed c and the basic zonal flow $\bar{u}(y, z)$ against which the phase has to propagate, and for given $(\bar{u} - c)$ can be elegantly expressed in terms of the quasi-geostrophic refractive index introduced by Charney and Drazin (1961) and Matsuno (1970). This has recently found extensive use in theoretical discussions of stationary planetary wave propagation in the stratosphere.

The precise nature of the lateral and vertical propagation mechanism can be appreciated more clearly from a slight adaptation of Fig. 18, with the understanding that there is to be some judicious interpolation between the IPV patterns explicitly shown. If we imagine that the relative zonal flow $(\bar{u} - c)$ is towards the right everywhere, and that the basic IPV gradient is positive everywhere, then we have a qualitatively correct picture of what happens near the leading edge of the disturbance as it penetrates towards positive y or z, with the top half of the figure representing the initially weak IPV anomaly pattern at the leading edge. This situation is shown in Fig. 19.

The essential point is that the IPV pattern at the leading edge is being made to grow (at a rate which can be related to the appropriate 'group velocity' component) by the induced velocity field of the lower IPV pattern, where, we suppose, the disturbance has already reached full strength. Just as with the instability, the growth depends crucially on having a phase tilt in the sense shown. But the effect of the upper pattern upon the lower pattern is now the opposite to what it was before. With the changed sign configuration in the lower pattern, as shown in Fig. 19, the velocity is now more than a quarter wavelength out of phase with the displacement so that the effect of the upper pattern on the lower pattern is to induce it to *decay*. This is how an isolated wave packet propagates northwards, or upwards as the case may be, and contrives to leave little or no disturbance behind it.

In view of the typical large-scale IPV distribution in the atmosphere (Figs. 1, 2(a)–(d)), we might expect to find some tendency for the tropopause, where IPV gradients are very strongly concentrated in the real atmosphere, to act as a Rossby waveguide. Although refractive index distributions tend to be less strongly structured, because of their dependence on $(\bar{u} - c)$, the expectation is to some extent borne out by objective

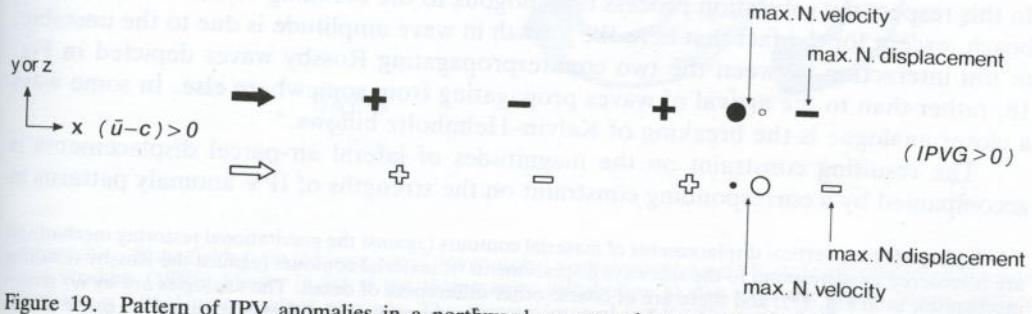

Figure 19. Pattern of IPV anomalies in a northward or upward propagating Rossby wave, represented schematically in the same way as in Fig. 18 to show the group-velocity mechanism (see text, section 6(*c*)). A similar picture can be used to understand the mechanism of 'critical-layer' absorption (e.g. Killworth and McIntyre 1985, section 2).

diagnostic studies of eddy statistics measuring the net flux of wave activity. This net eddy flux is often found to emerge from apparent sources in the middle and high-latitude troposphere and to split into two branches, the first going along the tropopause, mainly equatorwards, and the second going upwards into the stratosphere (e.g. Matsuno 1970; Sato 1980; Kanzawa 1984; Miles and Chapman 1984; Al-Ajmi *et al.* 1985).

(*d*) *The nonlinear saturation of baroclinic instabilities*

The use of IPV thinking to understand Rossby waves and related shear instabilities is by no means restricted to small wave amplitude, in the sense required by the standard linear theories of these phenomena upon which the foregoing discussion was based. IPV thinking can also give considerable qualitative insight into what happens when wave amplitudes become large for whatever reason, and, in particular, insight into the strongly nonlinear stages of a fast-growing baroclinic instability. These stages are inaccessible to analytical theory, whether linear or 'weakly nonlinear'. For the sake of brevity we merely sketch the main points here; more extensive discussions, with detailed diagnostics, may be found elsewhere (Hoskins and McIntyre 1985). Various aspects of the picture have been noted previously by, for example, Edmon *et al.* (1980), Hsu (1980), Davies (1981), Dunkerton *et al.* (1981), McIntyre (1980, 1982), Lindzen and Schoeberl (1982), Schoeberl (1982), Hoskins (1983), McIntyre and Palmer (1983, 1984) and Schoeberl and Lindzen (1984), and important precedents can be found in classical observational studies of occluding cyclones and other phenomena (e.g. Namias 1940, 1983; Berggren *et al.* 1949).

The amplitudes of lateral air-parcel displacements, and therefore of IPV anomaly strengths, are always limited in reality by constraints of an essentially kinematical nature which are neglected by linear theory. When such constraints come into play one may appropriately speak of 'nonlinear saturation'. One such constraint comes simply from the fact that there is never an infinite amount of room available for lateral air-parcel displacements to grow without bound in the way predicted, for example, by linear instability theory for a baroclinically unstable shear flow. Sooner or later the predicted displacements become comparable to the width scale of one of the regions of basic IPV gradients upon which unstable growth depends, or comparable to the width scale of the linear disturbance structure. The situation predicted by linear theory cannot then correspond, even qualitatively, to physical reality, one symptom of this being that material contours shaped initially like those in Fig. 17 begin to intersect one another. (In the case of Fig. 17 this would happen if the contour displacement amplitudes were to grow by a factor of about exp(1·1).) In reality, material and IPV contours which look like those of Fig. 17 in the early stages of growth are always found, in cases of practical interest, to deform irreversibly into complicated shapes rather than continuing to expand sideways. In this respect the saturation process is analogous to the breaking of waves on an ocean beach, except for the fact that here the growth in wave amplitude is due to the unstable, *in situ* interaction between the two counterpropagating Rossby waves depicted in Fig. 18, rather than to the arrival of waves propagating from somewhere else. In some ways a closer analogue is the breaking of Kelvin–Helmholtz billows.*

The resulting constraint on the magnitudes of lateral air-parcel displacements is accompanied by a corresponding constraint on the strengths of IPV anomaly patterns in

* In both analogies, vertical displacements of material contours (against the gravitational restoring mechanism) are considered to correspond to the sideways displacements of material contours (against the Rossby restoring mechanism, as in Fig. 17); and there are of course other differences of detail. The analogies are by no means superficial. Their basic relevance can be seen for instance from the hypotheses required to prove finite-amplitude 'nonacceleration theorems' in the general theory of wave, mean–flow interaction (Andrews and McIntyre 1978, §5; McIntyre 1980, §3; McIntyre and Palmer 1983, 1984). One of the necessary hypotheses is that the relevant material contours undulate without deforming irreversibly, i.e. that the waves do not 'break' in the sense suggested.

the region concerned. Indeed the effective strengths of the IPV anomalies may actually diminish, even in the absence of frictional and diabatic processes, when the deformation of material and IPV contours becomes severe, since the scale effect comes into play as the IPV contours wrap up. The wrapping-up process tends to produce ever smaller scales in the IPV distribution (the 'enstrophy cascade' effect), leading to destructive interference in the inversion operation as far as the original, relatively large, wave scale is concerned.

For the fastest-growing baroclinic instabilities on realistic basic jet-like flows, like the standard wave-6 case of Simmons and Hoskins (1980), it is found from numerical simulations that nonlinear saturation takes place in the manner described, and that it occurs first of all at the surface and in an adjacent layer containing the steering level. These are the places where linear instability theory predicts the largest lateral displacements. The layer is about half a density-scale-height deep. The wind, temperature and pressure fields resemble those observed in the classical occlusion process. The low-level isotherms give an indication of the shapes of the relevant material contours. Sharp surface fronts develop, in the familiar way, and these can be looked on as one manifestation of the 'cascade' to smaller scales; no such fronts are present in the initial conditions used in the simulations. By way of illustration, Fig. 20(a) shows the near-surface isotherms for day 6 of Simmons and Hoskins' standard case. Their shapes are to be contrasted with the simple, wavy shapes sketched in Fig. 17.

Low-level saturation leads on, in these cases, to a further stage of evolution describable as free Rossby wave propagation into the upper troposphere; it is an example of the *nonlinear radiation* discussed, for example, by McIntyre and Weissman (1978). Efficient upward radiation is possible because refractive index values turn out to be large

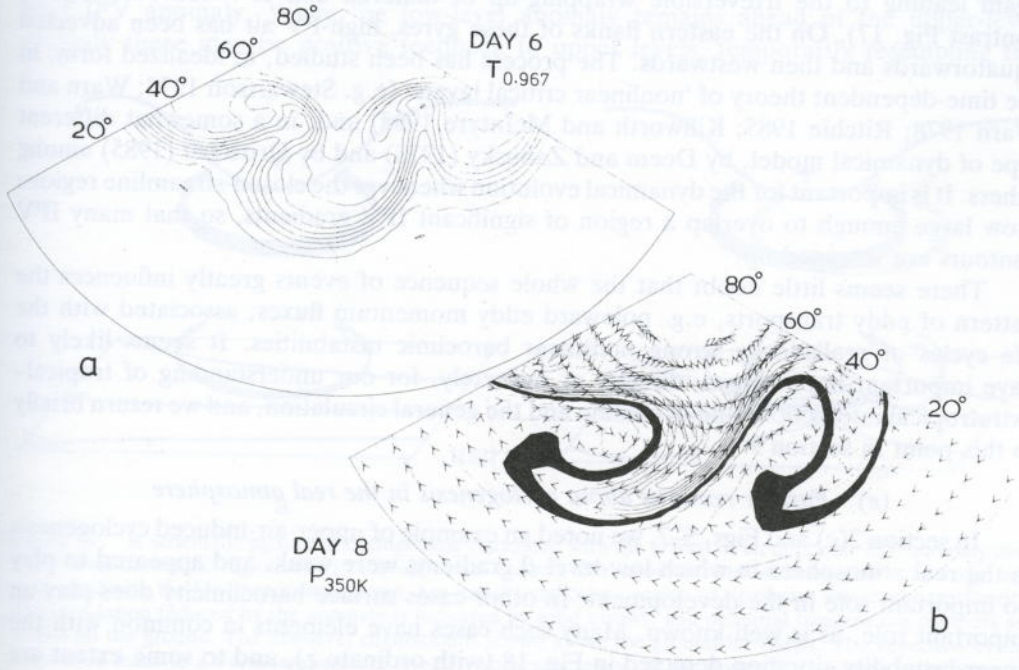

Figure 20. Some fields from the basic zonal wavenumber-six baroclinic wave life-cycle experiment of Simmons and Hoskins (1980). Shown in (a) is the temperature distribution at day 6 on the lowest 'sigma' surface, pressure = 0·9673 times surface pressure. The contour interval is 5 K. The latitudinal domain is from 0° to 90°N, and two full horizontal wavelengths are shown. In (b) the day-8 350 K IPV map is given, with contour interval 0·8 PV units. The regions with values 2·4–3·2 units have been blacked in. Also shown by arrows is the wind field on this isentropic surface in a frame of reference moving with the waves. The scale is such that the arrows nearest the equator represent a speed of 16 m s^{-1}.

enough to support free propagation of a wave-6 disturbance with matching phase speed c. Nonlinear radiation begins as soon as IPV and surface potential temperature anomalies in the saturating lower layer cease growing as fast as IPV anomalies located higher up, in the middle troposphere. Above the saturating layer, IPV contours in the main jet region have wavy shapes more like those in Fig. 17, and displacement amplitudes are still growing.

As this upper region becomes increasingly free of the influence of the lower layer, the disturbance there becomes increasingly free to propagate upward. (Some of it may also propagate downward, but if so will tend to be reflected back up.) The simulations show that the disturbance does indeed penetrate upward and that it then concentrates near the jet maximum at the tropopause, and, in most cases, appears to be guided along the tropopause mainly towards the equatorward flank of the jet (e.g. Edmon *et al.* 1980, Fig. 3). There, a second major saturation event takes place, distinctly later than the first, low-level event.

In this second saturation event, the analogy to ocean surf is closer than before and the situation is very similar indeed to that of the breaking mid-stratospheric Rossby waves discussed, for instance, by Hsu (1980), McIntyre (1980, 1982), McIntyre and Palmer (1983, 1984), Al-Ajmi *et al.* (1985), Clough *et al.* (1985) and Leovy *et al.* (1985). Here the lateral inhomogeneity of the relative zonal mean wind $\bar{u} - c$ is a crucial factor. The saturation process is illustrated in Fig. 20(b), which presents the 350 K IPV map for day 8 of the standard case. The wind vectors on the 350 K surface are drawn relative to the wave, i.e. in the frame of reference in which $c = 0$. Equatorwards of 45°, the flow seen in this frame of reference is dominated by anticyclonic closed-streamline regions, again leading to the irreversible wrapping-up of material and IPV contours (again, contrast Fig. 17). On the eastern flanks of these gyres, high-PV air has been advected equatorwards and then westwards. The process has been studied, in idealized form, in the time-dependent theory of 'nonlinear critical layers' (e.g. Stewartson 1978; Warn and Warn 1978; Ritchie 1985; Killworth and McIntyre 1985) and, in a somewhat different type of dynamical model, by Deem and Zabusky (1978) and by Dritschel (1985) among others. It is important for the dynamical evolution whenever the closed-streamline regions grow large enough to overlap a region of significant IPV gradients, so that many IPV contours are wrapped up.

There seems little doubt that the whole sequence of events greatly influences the pattern of eddy transports, e.g. poleward eddy momentum fluxes, associated with the life cycles of realistically strong nonlinear baroclinic instabilities. It seems likely to have important implications, directly or indirectly, for our understanding of tropical–extratropical interactions and of climate and the general circulation, and we return briefly to this point in section 9.

(*e*) *Further remarks about cyclogenesis in the real atmosphere*

In section 2(*c*) and Figs. 5–7, we noted an example of upper-air-induced cyclogenesis in the real atmosphere in which low-level θ gradients were weak, and appeared to play no important role in the development. In other cases surface baroclinicity does play an important role, as is well known. Many such cases have elements in common with the linear-instability situation depicted in Fig. 18 (with ordinate z), and to some extent are describable in similar terms.

However, pre-existing upper air IPV anomalies of large amplitude are often involved, like the cyclonic anomaly seen in Fig. 5(b). Therefore the cyclogenetic situation may not depend upon the prior existence of a phase-locked, linear-instability stage, and may have a more immediate dependence on initial conditions. Theoretical idealizations of different

aspects of such situations have been proposed by Farrell (1982), building on the work of Kelvin (Thomson 1887) and Orr (1907), and by Simmons and Hoskins (1979) in the context of downstream development. In some observed cases (e.g. Uccellini *et al.* 1984, 1985; Young *et al.* 1985) the 'upper air' anomaly is one which had previously been advected down sloping isentropic surfaces to mid-tropospheric levels or even lower. In such cases 'upper air' may have to be understood as referring more to the origin of the anomaly, than to its actual altitude at the time of maximum surface development. Figures 5, 8 and 9 appear to be less extreme cases of the same thing, and other examples can be seen in Figs. 2 and 3 and in many of the case studies cited in section 1(*c*). The cyclonic upper air anomalies of interest evidently come in a great variety of shapes, sizes, and rates of advection, some being associated in an obvious way with prominent, large-scale upper air troughs, as in the case studies of Petterssen and Smebye (1971), and others having smaller scales which presumably go all the way down to the scales characteristic of jet streaks.

A standard cyclogenetic situation is shown schematically in Fig. 21. Suppose, in the spirit of the thought-experiment of section 4, that a cyclonic upper air IPV anomaly (which to a greater or lesser extent will be associated with a low tropopause, depending on its exact size, shape and strength) arrives over a pre-existing low-level baroclinic region, as suggested in Fig. 21(a). Thermal advection by the induced low-level circulation will tend to create a warm low-level anomaly ahead of the upper IPV anomaly (Fig. 21(b)), enhancing the effects of any low-level warm advection already present. This warm surface anomaly will induce, as in Fig. 16(a), its own cyclonic circulation. At low levels this circulation will add to the circulation induced from upper levels (Figs. 21(b), 15(a)), giving an intense low-level cyclone whose centre is a little ahead of the advancing upper-level IPV anomaly. While the low-level anomaly remains ahead of the upper-level anomaly there may be positive feedback to upper levels, temporarily resembling the

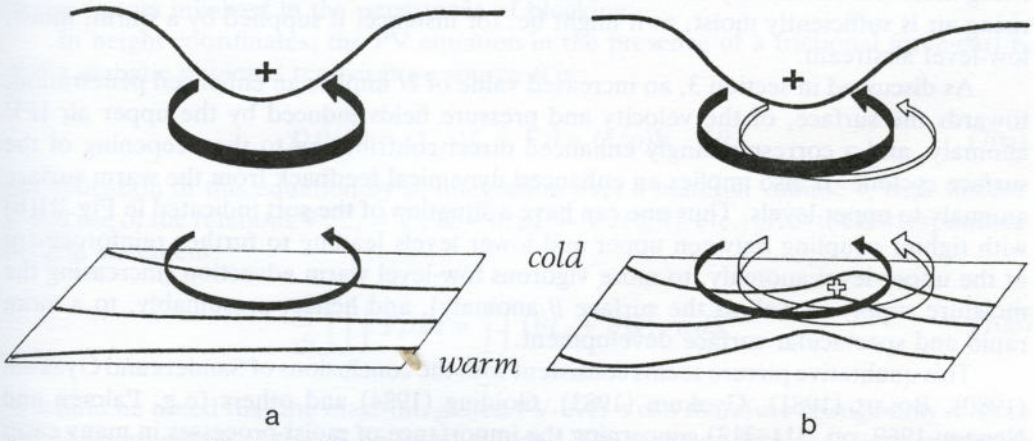

Figure 21. A schematic picture of cyclogenesis associated with the arrival of an upper air IPV anomaly over a low-level baroclinic region. In (a) the upper air cyclonic IPV anomaly, indicated by a solid plus sign and associated with the low tropopause shown, has just arrived over a region of significant low-level baroclinicity. The circulation induced by the anomaly is indicated by solid arrows, and potential temperature contours are shown on the ground. The low-level circulation is shown above the ground for clarity. The advection by this circulation leads to a warm temperature anomaly somewhat ahead of the upper IPV anomaly as indicated in (b), and marked with an open plus sign. This warm anomaly induces the cyclonic circulation indicated by the open arrows in (b). If the equatorward motion at upper levels advects high-PV polar lower-stratospheric air, and the poleward motion advects low-PV subtropical upper-tropospheric air, then the action of the upper-level circulation induced by the surface potential temperature anomaly will, in effect, reinforce the upper air IPV anomaly and slow down its eastward progression. (To this extent the situation is similar to the small-amplitude instability situation represented by Fig. 18 and described in section 6(*b*).)

small-amplitude instability situation in the 'helping' case, i.e. the case depicted in Fig. 18. This will tend to phase-lock the two anomalies and promote their mutual intensification. In particular, the upward extension of the circulation induced by the low-level warm anomaly will tend to intensify the upper-level IPV anomaly by advecting high-PV air equatorwards, on the left of the picture, and, because this advection is strongest just behind the upper-level IPV anomaly, will also tend, in effect, to slow down its advance.

It is well known that the combination of low-level warm advection and upper-level PVA or positive IPV advection can result in strong cyclonic development. Some idea of the possible magnitude of the effect, even in a dry atmosphere, can be gained by considering the possible strengths of the superposed wind fields induced by a cyclonic upper-level anomaly overlying a low-level warm anomaly. For example, a nearly-coaxial superposition of Figs. 15(a) and 16(a) gives surface winds as high as $31\,\mathrm{m\,s^{-1}}$, and a very large surface pressure anomaly of -72 mb. This of course is no more than an indication of the orders of magnitude involved, in the absence of friction, since as already explained the superposition principle holds only qualitatively for such large-amplitude anomalies.

It is of interest to note how the foregoing picture is likely to be modified by moist processes. The theory suggests what is also suggested by synoptic experience and numerical experimentation (e.g. Golding 1984), namely that moist processes can, and will on occasion, greatly enhance the surface development, even though conditions may be subcritical to moisture-driven instabilities such as CISK.

For the reasons explained in section 4, there will usually be a synoptic-scale region of rising motion within the region of reduced static stability beneath the advancing upper-level IPV anomaly (Fig. 15(a)). If condensation occurs in the rising air, as will happen if there is a sufficient supply of moisture, then the effective static stability, N, as felt by the large-scale upward motion, will be still further reduced. The effective Rossby height scale H given by the second expression in (33b) will then be increased, if the region of rising motion is sufficiently extensive horizontally—and the increase may be drastic if the rising air is sufficiently moist, as it might be, for instance, if supplied by a warm, moist, low-level airstream.

As discussed in section 3, an increased value of H implies an enhanced penetration, towards the surface, of the velocity and pressure fields induced by the upper air IPV anomaly, and a correspondingly enhanced direct contribution to the deepening of the surface cyclone. It also implies an enhanced dynamical feedback from the warm surface anomaly to upper levels. Thus one can have a situation of the sort indicated in Fig. 21(b) with tighter coupling between upper and lower levels leading to further reinforcement of the upper-level anomaly, to more vigorous low-level warm advection (increasing the moisture supply as well as the surface θ anomaly), and hence, presumably, to a more rapid and spectacular surface development.

This qualitative picture seems consistent with the conclusions of Sanders and Gyakum (1980), Bosart (1981), Gyakum (1983), Golding (1984) and others (e.g. Palmén and Newton 1969, pp. 311–312) concerning the importance of moist processes in many cases of rapid cyclonic development outside the tropics. Equally, it is consistent with the view that upper air IPV advection may also play a key role in the development (e.g. Uccellini *et al.* 1984, 1985; Young *et al.* 1985), and that the primary effect of moisture in these cases may be to *amplify* a cyclogenetic process of the general kind depicted in Fig. 21; see also Sanders and Gyakum (*op. cit.*).

7. The maintenance and dissipation of cutoff cyclones and blocking anticyclones

We now take up the question of diabatic and frictional effects for the large-amplitude

anomalies giving rise to the cutoff cyclones and blocking anticyclones which formed the subject of sections 2–4.

Once a cyclonic PV region has cut off on an IPV map in the manner of Fig. 5, it will be surrounded by its induced cyclonic circulation and, if the surface temperature is not too low, this circulation will penetrate to the ground, giving a surface cyclone. Surface frictional processes would then begin to generate a cold, anticyclonic surface anomaly (similar to the structure illustrated in Fig. 16(b)) which, if there happens to be no low-level air motion relative to the upper air anomaly, will remain in place and ameliorate the cyclonic circulation at low levels, as suggested by a qualitative superposition of Figs. 15(a) and 16(b). However, regardless of what happens at low levels, the upper-tropospheric cyclone will persist as long as the IPV anomaly persists. The structure in Fig. 16(b) is evanescent with height and cannot cancel that in Fig. 15(a) at all levels. We note, in particular, that not all the kinetic energy of the upper air part of the circulation can be destroyed by boundary layer friction alone, since the latter cannot directly affect the upper air IPV anomaly.

The anomaly could of course, be removed by simply being advected back along isentropic surfaces into the polar stratospheric reservoir. However, synoptic experience suggests that the chances of this happening in less than a week are small. In practice, diabatic processes must be crucial to the evolution and decay of such systems, which often occurs much faster than a week, depending on the circumstances. This seems to be the basic fact behind the widely differing persistence of such features under different diabatic conditions.

In a similar manner, the anticyclonic IPV anomaly associated with a blocking anticyclone (Fig. 11(d)) induces its own anticyclonic circulation. The anomaly must persist unless the low-PV air returns to the subtropical region, as actually occurs in the example of Fig. 11, or is changed *in situ* by diabatic processes. We shall see that the latter processes tend to have a slower time scale for anticyclones. This appears to be one of the factors involved in the persistence of blocking.

In height coordinates, the PV equation in the presence of a frictional force-curl $\mathbf{K}$ and a diabatic potential temperature source $\dot{\theta}$ is

$$DP/Dt = (1/\rho)\boldsymbol{\zeta}_a \cdot \nabla\dot{\theta} + (1/\rho)\mathbf{K} \cdot \nabla\theta. \tag{70a}$$

Multiplication of this equation by ρ, integration over a material volume τ, with surface S, and use of the relations $\nabla \cdot \zeta_a = \nabla \cdot \mathbf{K} = 0$, $\rho P = \nabla \cdot (\zeta_a \theta)$, etc., gives the corresponding integral statement

$$\frac{d}{dt}\iiint_{\tau} P\rho\, d\tau = \iint_{S} (\dot{\theta}\boldsymbol{\zeta}_a + \theta\mathbf{K}) \cdot \mathbf{n}\, dS. \tag{70b}$$

It should be noted that the mass-integrated PV over τ can therefore change only if there are non-zero values of $\dot{\theta}$ or $\mathbf{K}$ on its boundary S. Diabatic and frictional sources interior to the volume can only redistribute the PV; and it should be further noted that surface θ anomalies can be included in this statement if the conventions suggested by the insets to Fig. 16 are adopted, in which case $\dot{\theta}$ on the right-hand side of (70b) would be zero by definition.

For small-Rossby-number, large-Richardson-number flows, the vertical contributions to the dot product terms in (70a) dominate to give

$$DP/Dt \simeq (1/\rho)\mathbf{k} \cdot \boldsymbol{\zeta}_a(\partial\dot{\theta}/\partial z) + (1/\rho)\mathbf{k} \cdot \mathbf{K}(\partial\theta/\partial z). \tag{71}$$

In isentropic coordinates the equivalent equation is

$$DP/Dt = -\sigma^{-1}\{(f\mathbf{k} + \nabla_\theta \times \mathbf{v}) \,.\, \nabla\dot{\theta} + K_\theta\} \tag{72}$$

$$\simeq -\sigma^{-1}\{(f + \zeta_\theta)\partial\dot{\theta}/\partial\theta + K_\theta\}, \tag{73}$$

where as before σ stands for $-g^{-1}\partial p/\partial\theta$, and where K_θ is the frictional isentropic force-curl analogous to the 'isentropic vorticity' defined by (9), i.e. $K_\theta = \mathbf{k} \cdot \nabla_\theta \times \mathbf{F}$ where $\mathbf{F}$ is the friction force per unit mass. Note that here $D/Dt = \partial/\partial t + \mathbf{v} \,.\, \nabla_\theta + \dot{\theta}\partial/\partial\theta$, and that a convenient form of (73) for consideration of IPV behaviour in the presence of diabatic heating alone is

$$(\partial/\partial t + \mathbf{v} \,.\, \nabla_\theta)P \simeq P^2\,\partial(\dot{\theta}P^{-1})/\partial\theta \tag{74a}$$

or alternatively

$$(\partial/\partial t + \mathbf{v} \,.\, \nabla_\theta)(P^{-1}) \simeq -\partial(\dot{\theta}P^{-1})/\partial\theta. \tag{74b}$$

The left-hand sides are the rates of change following an isentropic trajectory, and the simple forms of the right-hand sides facilitate computation and qualitative argument. The more elegant form (74b) which results when P^{-1} is used in place of P is reminiscent of an analytical device used by Eliassen (1983).

As discussed above, the frictional force-curl term is expected to be important within the boundary layer. In the free atmosphere, Shapiro (1976) has hypothesized that mixing of potential temperature due to clear air and other small-scale turbulence could be important in producing some of the anomalously large PV values observed in the lower stratosphere on the cyclonic side of strong jet streams. The constraint (70b) on the mass-integrated PV must of course be kept in mind here, as must the possibility of advection from upstream locations.

We now return to the role of latent heat release and radiative cooling in the dissipation and structural modification of mid-latitude cutoff cyclones and blocking anticyclones. As pointed out in section 3 (statement (iv)), one of the characteristic features of cutoff cyclones is the presence of weak static stability under the 'lowered' tropopause, e.g. under the heavy curve in Fig. 8 or in Fig. 15(a). This weak static stability is consistent with the observed tendency for deep convection to be associated with such systems if there is a sufficient moisture supply from below (e.g. Pedgley 1962, p. 159; Erickson 1971). Tropospheric convective heating, with $P^{-1}\dot{\theta}$ diminishing upwards into the high-IPV anomaly above the 'lowered' tropopause, and changing sign as convective heating gives way to infrared cooling, will tend to reduce the strength of that anomaly locally.

For the example discussed in section 2(c) (Figs. 5–7), temperature and humidity soundings for Long Kesh (54°N 6°W) for 00 z on 25 September 1982 indicate conditional instability up to 450 mb, the height of the 'lowered' tropopause for the observed 12 z temperature. From the form of the right-hand sides of (74a, b) it is clear that the profile of $\dot{\theta}P^{-1}$ is important. For relatively uniform heating in the troposphere, the decrease in P^{-1} with height implies that there is some reduction in IPV values, consistent with the upward motion of air through isentropic surfaces. However, the most dramatic effect is near the tropopause. In the lower stratosphere, the PV is so large that $\dot{\theta}P^{-1}$ is effectively zero, irrespective of whether $\dot{\theta}$ is dominated by infrared cooling rates or by convective heating, for any reasonable estimates. Thus for positive $\dot{\theta}$, IPV values in the tropopause region must be subject to large rates of diminution. Taking $P^{-1}\dot{\theta} = 0$ at $\theta = 305$ K, $P =$ 2 PV units at $\theta = 300$ K, and $P = 1$ PV unit, $\dot{\theta} = 2{\cdot}5\,\mathrm{K\,d^{-1}}$ at $\theta = 295$ K gives, using centred differences in (74a), an IPV tendency at 300 K of minus 1 PV unit per day. This

is sufficient to destroy the 300 K IPV anomaly in a single day, which seems consistent with the observed behaviour of the system, Figs. 5(e), (f) in particular. By contrast, over continental interiors during wintertime the moisture supply is insufficient for much convection, and the characteristic cutoff-cyclone structure can persist much longer.

The implications of the integral constraint represented by (70b) should again be noted. Since deep convection will make $\dot{\theta} \neq 0$ only in the interior of the troposphere, away from the surface, the net effect of the convection will be to move the IPV anomaly down to lower tropospheric levels. In order to annihilate the cyclone completely, this diabatically-induced vertical redistribution of PV would have to extend all the way down to the surface, and be accompanied by the destruction of mass-integrated PV by surface friction. The need for surface friction in this hypothesized situation is clear from a consideration of the angular momentum budget of an idealized circular vortex. It can also be seen from the version of (70b) suggested by the insets to Fig. 16. If all *interior* IPV anomalies were to be destroyed by diabatic heating then in the absence of surface friction a surface θ anomaly would still remain.

The blocking anticyclone structure as typified by Figs. 14 and 15(b) is one of large tropospheric static stability, which tends to suppress convection. It is therefore the effect of radiative cooling that must primarily be considered. Using (74) with $P^{-1}\dot{\theta} = 0$ at $\theta = 335$ K, $P = 1$ PV unit at $\theta = 330$ K, and $P = 0{\cdot}5$ PV units, $\dot{\theta} = 1\,\mathrm{K\,d^{-1}}$ at $\theta = 325$ K, gives a 330 K IPV tendency of plus 1 unit per 5 days. This suggests a time scale of a week or so for diabatic processes to modify such a blocking anticyclone.

Having previously stressed the conceptual duality between the dynamics of cutoff cyclones and blocking anticyclones, in sections 2(*e*) and 3, we see now that there is no such duality for their diabatic modification, which, other things being equal, tend to be much faster for cyclones than for anticyclones. The crucial difference is the tropospheric static stability induced by the IPV anomaly. In the cutoff cyclone, deep tropospheric convection is enhanced, the latent heat release leading to an efficient diabatic decay of the upper IPV anomaly on a time scale of a few days. In the blocking anticyclone, convection is suppressed, mainly by the increased static stability. Radiative cooling gives a time scale of a week or so. Of course quasi-conservative advection and re-merging into the subtropics becomes quite likely on this time scale, as did indeed occur in the examples described in section 2 and Fig. 11.

8. Further remarks about cutoff systems and air masses

In stressing the distinction between 'cutoff' and 'non-cutoff' weather systems, synoptic meteorologists have traditionally defined them as systems with and without closed isobaric height contours at 500 mb, 300 mb or thereabouts. The distinction is undoubtedly important in practice, but its significance has never been entirely clear from a theoretical viewpoint since, for example, the addition of a uniform zonal flow which simply advects the whole system can easily change its classification from 'cutoff' to 'non-cutoff' if the criterion is taken literally, whereas the dynamics of the system would not be changed in any essential way.

The concurrent IPV and isobaric height maps shown in Figs. 3, 4, 5, 6, 11 and 12 suggest that the power of this synoptic idea can be accounted for theoretically by the fact that isobaric height maps to some extent provide a view, albeit a smoothed-out view, of the associated IPV distributions. The patterns seen in IPV maps suffer from no ambiguity of the kind just mentioned. In particular, the presence or absence of closed contours in IPV maps is a dynamically significant distinction, independent of incidental circumstances such as the frame of reference adopted.