Acta Numerica (2007), pp. 67–154
doi: 10.1017/S0962492906290019

Modelling atmospheric flows

Mike Cullen
Met Office,
Fitzroy Road,
Exeter EX1 3PB, UK
E-mail: mike.cullen@metoffice.gov.uk

This article demonstrates how numerical methods for atmospheric models can be validated by showing that they give the theoretically predicted rate of convergence to relevant asymptotic limit solutions. This procedure is necessary because the exact solution of the Navier–Stokes equations cannot be resolved by production models. The limit solutions chosen are those most important for weather and climate prediction. While the best numerical algorithms for this purpose largely reflect current practice, some important limit solutions cannot be captured by existing methods. The use of Lagrangian rather than Eulerian averaging may be required in these cases.

CONTENTS

1. Introduction

This article is a review of the mathematical basis of the numerical methods used in production atmosphere models. Many of the results can also be applied to ocean models. A recent review of ocean modelling issues is given by Higdon (2006). Atmospheric models are routinely used in weather prediction for time-scales of a few hours up to a few seasons, and for

climate predictions for hundreds of years ahead. These predictions require modelling not just of the atmosphere, but also of the ocean and the rest of the 'Earth system' such as the vegetation. They also require modelling of the interactions between the atmosphere, ocean and land surface. This is difficult because of the very different time-scales of the separate systems.

In atmospheric predictions, the state of the system at an initial time has to be determined from observations. The techniques used are usually called 'data assimilation', where information from new observations is blended with a first guess computed from observations made at previous times. The data assimilation problem is not discussed in this article. A good introduction is given in Kalnay (2003). However, knowledge of the dynamics of the system is very important in data assimilation, and some discussion of how the knowledge is used is given. In addition, modern data assimilation methods such as 'four-dimensional variational data assimilation' (see Courtier, Thépaut and Hollingsworth (1994)) require a model trajectory to be fitted to the observations. Practical ways of doing this require a linearization of the numerical algorithm used in the model, together with an adjoint. The ability to create an accurate linearization of the numerical method used in the nonlinear model is thus important, and is discussed in the article. However, there are many other numerical issues associated with the data assimilation problem which are not discussed.

In production atmosphere models, the affordable resolutions are many orders of magnitude coarser than those required to solve the equations accurately. As discussed in Section 2.2, this would require a grid-length of the order of 1 mm. In practice, the equations are averaged in space and time, and thus depend on implicit or explicit sub-grid modelling assumptions. The problem is therefore to show that the numerical solution stays close to the averaged solution. It is, of course, not possible to write down a set of equations which describe the averaged behaviour exactly; and it is also not possible to estimate the difference between the solution of the averaged equations and the average of the true solution.

The article concentrates on numerical methods for the averaged equations governing the atmosphere, rather than the choice of sub-grid models. The choice of averaging methods and the design of sub-grid models are very large subjects, and exploit detailed observational and modelling studies of the small-scale behaviour. Only those issues which cannot be separated from the design of numerical methods are discussed. Some more information about this very diverse subject is given in Garratt (1992) and Smith (1997). The atmospheric circulation is forced directly by radiation and indirectly through boundary fluxes. The effects of phase changes involving water vapour is very important. In this article, we assume very simple representations of these processes. In production models, a large proportion of the computer code and execution time is spent on modelling them.

The ideal in a mathematical analysis of the numerical methods used in prediction models is to prove that the numerical solution stays close to the exact solution for large times. While this can be attempted if the sub-grid model is assumed to be exact, such estimates are of limited use because of the unknown error of the sub-grid model itself. It is therefore more useful to analyse the accuracy of the approximation to asymptotic limit solutions of the governing equations, whose accuracy is known and which can be well represented using the space- and time-averaging scales that can be afforded. This is the method used in this article. Production numerical models can describe a large variety of asymptotic regimes of the governing equations; and ideally it is necessary to demonstrate the accuracy of the numerical method in all of them. In practice only the most important regimes are considered.

While this article describes how to optimize numerical methods for the most important regimes being modelled; this is only one aspect of the subject of numerical methods for the atmosphere. More comprehensive descriptions of other aspects of the subject can be found in Durran (1998).

In Section 2 we introduce some of the most important asymptotic regimes and discuss their properties relevant to numerical approximation. In Section 3, we discuss the properties of numerical methods in representing these regimes with reference to methods used in current operational models. In Section 4, we illustrate these procedures using the Met Office Unified Model. We demonstrate that the numerical solutions converge to solutions of asymptotic limit equations appropriate to large scales in three cases which are simple enough for accurate numerical solutions to be possible, but still physically relevant. We also demonstrate that the solutions converge to asymptotic limit solutions governing smaller-scale behaviour.

2. Asymptotic limits of the equations of motion

2.1. Basic equations

The starting point is that the atmosphere can be regarded as a fluid continuum which obeys the basic physical laws of dynamics and thermodynamics. The study of fluid dynamics recognizes that fluids exhibit a wide range of different behaviour under different circumstances. These are characterized as asymptotic regimes by identifying dimensionless parameters that control the flow, and choosing appropriate ranges of values of these parameters. A comprehensive survey is given in Batchelor (1967).

A number of simplifying assumptions are universally made when modelling the atmosphere. Current research is exploring whether some of these should be relaxed as the availability of more powerful computers enables more accurate solutions. A more detailed account of these issues is given in White (2002).

The Earth is assumed to rotate with angular velocity $\boldsymbol{\Omega}$ on an axis through the coordinate poles. The acceleration due to gravity and the centrifugal acceleration due to the Earth's rotation are combined, and act normally to geopotential surfaces. The geopotential surfaces are then approximated by spherical surfaces. The equations are defined in spherical polar coordinates (λ, ϕ, r), with origin at the centre of the Earth. The Earth's surface is then assumed to be a spherical surface with radius a with perturbations due to orography. It is defined by the equation $r = r_0(\lambda, \phi)$. The combined gravitational and centrifugal acceleration are assumed to be towards the origin, with a constant magnitude g.

The atmosphere is assumed to consist of a compressible ideal gas with pressure, density and temperature p, ρ, T which are functions of position and time. It contains a mixing ratio q of water vapour. It moves with a vector velocity $\mathbf{u} = (u, v, w)$. The evolution is described by the compressible Navier–Stokes equations, the first law of thermodynamics and the equation of state for an ideal gas, all written in a frame of reference rotating with the Earth's angular velocity. These are

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + 2\boldsymbol{\Omega} \times \mathbf{u} + \frac{1}{\rho}\nabla p + g\hat{\mathbf{r}} &= \nu\nabla^2\mathbf{u} + \frac{1}{3}\nu\nabla(\nabla \cdot \mathbf{u}),\\
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho\mathbf{u}) &= 0,\\
C_v\frac{\mathrm{D}T}{\mathrm{D}t} - \frac{RT}{\rho}\frac{\mathrm{D}\rho}{\mathrm{D}t} &= \kappa_h\nabla^2 T + S_h + LP, \qquad (2.1)\\
\frac{\mathrm{D}q}{\mathrm{D}t} &= \kappa_q\nabla^2 q + S_q - P,\\
p &= \rho RT.
\end{aligned}$$

Here the Lagrangian derivative $\mathrm{D}/\mathrm{D}t$ is a shorthand for $\partial/\partial t + \mathbf{u} \cdot \nabla$ and $\hat{\mathbf{r}}$ is a unit vector in the radial direction. R is the gas constant and C_v the specific heat of air at constant volume. ν is the kinematic viscosity. All of these are assumed constant. S_h and S_q are the total heat and moisture sources. P is the rate of conversion of water vapour to liquid water or ice, with L the associated latent heat. $\rho\kappa_h$ and κ_q are the thermal conductivity and moisture diffusivity, also assumed constant. In production atmospheric modelling the thermodynamic parameters are allowed to be functions of atmospheric composition. The true viscosity and thermal conductivity are invariably superseded by sub-grid models. The representation of phase changes and forcing terms is very complex, but only the leading order effects will be discussed in this article.

These equations form a system of seven equations for the unknowns $(\mathbf{u}, p, \rho, T, q)$. The obvious physical boundary conditions are that $\mathbf{u} = 0$ at $r = r_0$ and that $p, \rho \to 0$ as $r \to \infty$. While the no-slip condition at

the Earth's surface is standard, the issue of the correct mathematical upper boundary condition for an unbounded atmosphere is open. Fluxes of heat and moisture are specified at the lower boundary. Fluxes of momentum are discussed in the next section.

Suppose there are no dissipation and source terms, so that the right-hand side terms of the first four equations of (2.1) vanish. Solve equations (2.1) in a closed time-independent region Γ with boundary conditions $\mathbf{u} \cdot \mathbf{n} = 0$, where $\mathbf{n}$ is a vector pointing outward from the boundary. Then the energy integral

$$E = \int_\Gamma \rho \left(\frac{1}{2}(u^2 + v^2 + w^2) + C_v T + gr \right) r^2 \cos\phi \, \mathrm{d}\lambda \, \mathrm{d}\phi \, \mathrm{d}r \tag{2.2}$$

is conserved. The requirement that the upper boundary be rigid can be removed most conveniently, while retaining energy and angular momentum conservation, by reformulating the equations in 'mass' coordinates (Wood and Staniforth 2003).

It is convenient to rewrite the first law of thermodynamics in terms of the potential temperature $\theta = T(p/p_{\mathrm{ref}})^{-R/C_p} \equiv T/\Pi$, where C_p is the specific heat of air at constant pressure, p_{ref} is a constant reference pressure equal to a typical pressure at the Earth's surface, and Π is the Exner pressure. This gives, noting $R = C_p - C_v$,

$$\frac{\mathrm{D}\theta}{\mathrm{D}t} = \frac{1}{C_p \Pi}(\kappa_h \nabla^2 T + S_h + LP). \tag{2.3}$$

This form of the equation is particularly useful in situations where the right-hand side terms can be neglected. The equation of state can now be rewritten as

$$p_{\mathrm{ref}} \Pi^{\frac{1}{\gamma - 1}} = \rho R \theta \tag{2.4}$$

where $\gamma = C_p/C_v$.

We can also rewrite the momentum equations by using the definition of θ and the equation of state (the last equation of (2.1)). After some algebraic manipulations we obtain

$$\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + 2\boldsymbol{\Omega} \times \mathbf{u} + C_p \theta \nabla \Pi + g\hat{\mathbf{r}} = \nu \nabla^2 \mathbf{u} + \frac{1}{3}\nu \nabla(\nabla \cdot \mathbf{u}). \tag{2.5}$$

In the absence of dissipation and source terms, equations (2.1) imply a conservation law for the *Ertel potential vorticity*

$$Q = \frac{1}{\rho}(\nabla \times \mathbf{u} + 2\boldsymbol{\Omega}) \cdot \nabla \theta \tag{2.6}$$

in the form

$$\frac{\mathrm{D}Q}{\mathrm{D}t} = 0. \tag{2.7}$$

2.2. Methods of averaging

As noted in the Introduction, the equations have to be averaged before they are solved numerically. This article only discusses a few of the key issues in the choice of averaging method. A much fuller discussion is given by Ferziger (1998). The averaging is much coarser than that which is usual in large eddy modelling. The kinematic viscosity of air is about $10^{-5}\,\mathrm{m^2\,s^{-1}}$ near the Earth's surface. The resolution required for a direct numerical simulation has to be chosen to make the Reynolds number based on the grid-length O(1). This is difficult to estimate. An estimate based on dissipation rates given by Gill (1982, p. 79) gives a resolution of about 1 mm. The resolution used in even the finest-scale production models is about 1 km. Thus the averaging has to reduce the resolution required by a factor of 10^9. In practice, therefore, the general ideas of turbulence modelling based on the maintenance of an inertial range have to be supplemented by specialized sub-grid models, which use additional knowledge about the behaviour of the atmosphere on scales which cannot be resolved.

The averaging has to be carried out in time as well as space. The ratio of the space- and time-averaging scales is important. The optimal choice will depend on the type of motion being modelled, and is quite different for sound waves and for solutions which move with the flow speed; see Browning and Kreiss (1994). It is therefore difficult to separate the issue from the identification of important asymptotic limits, which is discussed in the following sections.

A basic issue is the choice between Eulerian and Lagrangian averaging. In principle, the state of the atmosphere can either be described in terms of space-time averages at particular locations, or in terms of the behaviour of fluid parcels of a finite size. The latter is very appealing in terms of the observed physics. Descriptive accounts of meteorology often talk in terms of 'air-masses' with particular characteristics. The boundaries between different air-masses can be quite sharp, and spatial averaging would not then give satisfactory results. The mathematical theory underlying Lagrangian averaging is set out in Andrews and McIntyre (1978). That paper illustrates that it is technically much harder to work with than Eulerian averaging.

The coarseness of the averaging required means that the optimum technique may well depend on the asymptotic regime being modelled. For instance, we would expect Eulerian averaging to be appropriate to describe flow over hills, since the hills are fixed in space. However, Lagrangian averaging would be more appropriate for treating moving air-mass boundaries. In the rest of the article, we therefore discuss the optimum averaging for particular regimes alongside the discussion of the numerical method. In particular, it is possible to build the averaging into the

numerical method, thus obtaining an implicit turbulence model as discussed in Fureby and Grinstein (2002). However, many production models use an analytic turbulence model, which is added to the equations.

In designing averaging methods, it is essential to note that the prognostic variables in equations (2.1) are observed to be bounded quantities, and therefore their space-time averages will be smooth on the averaging scale. If analytic Eulerian averaging is used to produce modified versions of equations (2.1), it is essential that their solutions can be proved to be smooth on the averaging scale. The modified equations will therefore need to include terms which prevent the growth of unresolvable gradients in the prognostic variables. However, some terms in the governing equations will become insignificant in the presence of the averaging, and can be removed. Thus there is never any point in including the real kinematic viscous or thermal conduction terms. Some other approximations which are often made for coarser averaging scales are noted in the next section. Alternatively, if implicit Eulerian averaging is included in the numerical methods, it is essential that the computed solutions vary smoothly on the averaging scale.

A particular issue with the averaging is the treatment of the no-slip boundary conditions. The real viscous sub-layers are much too thin to resolve explicitly, even if the vertical grid is stretched to give increased resolution near the boundaries. Specialized sub-grid models are therefore used to represent the momentum fluxes near the boundaries. A review of this topic can be found in Garratt (1992).

Another issue is the treatment of instabilities which occur in the real system on scales much too small to be resolved in production models. If these instabilities are not removed by the sub-grid modelling, unstable circulations are likely to develop on resolved scales which may be orders of magnitude larger than the real ones. This can lead to results which are too inaccurate to be useful. It is normal to develop specialized sub-grid models to deal with this. In the atmosphere, such instabilities are often triggered by moisture phase changes, and thus correspond to significant weather events.

2.3. Important asymptotic limits

In this section we discuss some of the important asymptotic regimes for production atmosphere and ocean models. A more comprehensive treatment is given in textbooks such as Pedlosky (1987) and Gill (1982).

Figure 2.1 shows a diagram due to Smagorinsky (1974) which plots atmospheric phenomena as a function of horizontal scale and time-scale. The phenomena traditionally associated with extra-tropical weather forecasting are grouped along the diagonal, indicating that their horizontal scale L is typically proportional to $T^{\frac{3}{2}}$, where T is their time-scale, so they can all be characterized by a horizontal velocity $U = LT^{-1}$ between about

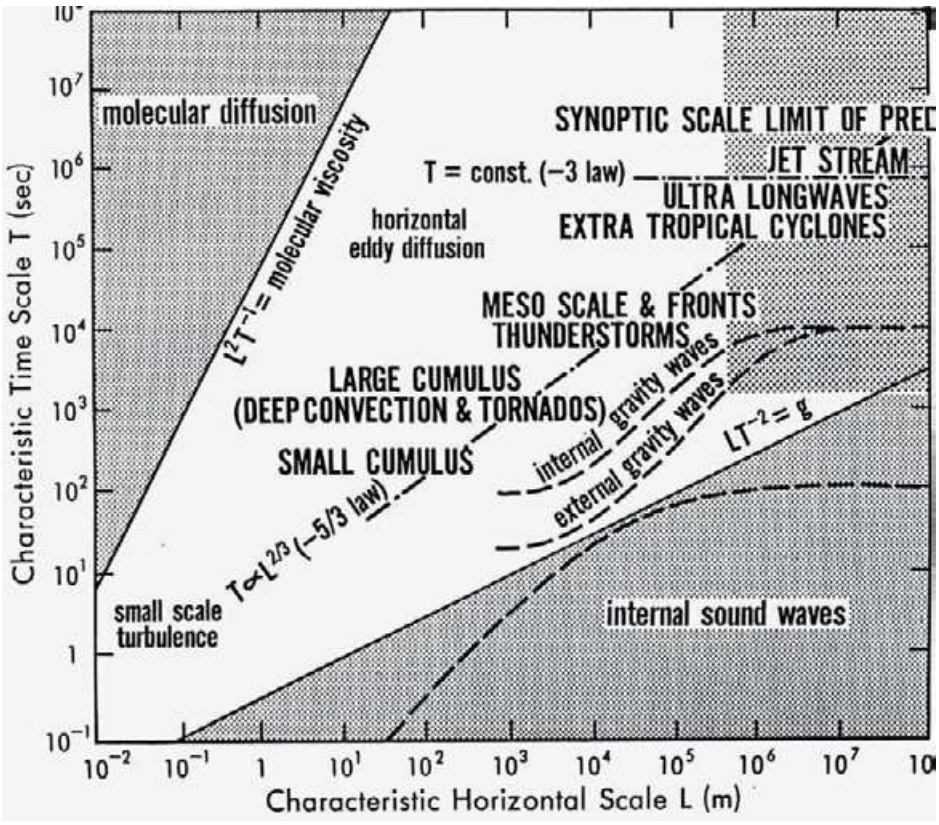

Figure 2.1. Typical space- and time-scales of atmospheric phenomena (following Smagorinsky (1974)).

1 and 10 m s^{-1}. There are also other classes of motion, such as gravity waves, which have time-scales shorter by an order of magnitude for a given length-scale, and sound waves with a time-scale several orders of magnitude shorter. Molecular diffusion acts on much smaller horizontal scales, as discussed earlier.

Though not shown in Figure 2.1, the typical vertical scale is also important. The effect of surface friction dominates in a boundary layer, which has a thickness of order 1 km. The troposphere, which contains nearly all the moisture, and where most weather systems are confined, is about 10 km deep. Above the troposphere is the stratosphere which is much more stably stratified. It is therefore necessary to consider a range of aspect ratios H/L, where H is the vertical scale. The requirement that the vertical velocity W has a similar time-scale to the horizontal velocity gives that $W/U \simeq H/L$.

Large horizontal scales

We now consider some of the important limits in more detail. We first consider large horizontal scales, where 'large' will be specified below but is intended to correspond to the scale of extra-tropical weather systems. This means that the viscous and thermal conductivity terms can be neglected.

The first step is to recognize that the acceleration due to gravity, g, is much larger than the acceleration of the air in a weather system. However, this acceleration is largely compensated by a vertical pressure gradient. We therefore define a time-independent reference state at rest, which satisfies equations (2.1), with uniform potential temperature θ_0 and with pressure p_0 and density ρ_0 depending only on the radial coordinate. This is given by an Exner pressure $\Pi_0(r)$ satisfying

$$\begin{aligned} C_p\theta_0\frac{\mathrm{d}\Pi_0}{\mathrm{d}r} + g &= 0, \\ \theta_0 &= \text{constant}, \\ \Pi_0 &= 1 \text{ at } r = a. \end{aligned} \tag{2.8}$$

Subtract this state from (2.5). The equation becomes

$$\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + 2\boldsymbol{\Omega}\times\mathbf{u} + C_p\theta\nabla\Pi' - g\frac{\theta'}{\theta_0}\hat{\mathbf{r}} = 0 \tag{2.9}$$

where $\theta' = \theta - \theta_0$ and $\Pi' = \Pi - \Pi_0$.

We next separate the cases where the horizontal scale is comparable to the radius of the Earth (so that the spherical geometry has to be considered), and smaller horizontal scales which can be studied in plane geometry. In the large-scale case, the next step is to recognize that the atmosphere is 'shallow'. The restriction of the vertical scale to the depth of the troposphere means that the aspect ratio for such flows is less than 0.01 and therefore the ratio of the vertical to the horizontal velocities is also less than 0.01. Under these conditions it can be shown that only the horizontal components of the Coriolis force $2\boldsymbol{\Omega}\times\mathbf{u}$ need be considered, and that the radial coordinate r can be replaced by a wherever it appears undifferentiated. This approximation is discussed in detail by White, Hoskins, Roulstone and Staniforth (2005). The result is that the components of the Coriolis force can be written as $(-fv, fu, 0)$ where the Coriolis parameter $f = 2\Omega\sin\phi$. The shallow atmosphere approximation is very accurate, and is used in many operational weather forecasting and climate models.

The next step in analysing large-scale flow is to show that it is hydrostatic. In the vertical component of equation (2.9), the term $g\theta'/\theta_0$ is typically about $1\,\mathrm{m\,s^{-2}}$, since the horizontal variations of θ are typically about 10% of the mean value. Since horizontal and vertical velocities have similar timescales, $\mathrm{D}w/\mathrm{D}t \simeq (H/L)\mathrm{D}u/\mathrm{D}t$. Given the aspect ratio of 0.01 discussed above, if $\mathrm{D}w/\mathrm{D}t \simeq 1\,\mathrm{m\,s^{-2}}$, then $\mathrm{D}u/\mathrm{D}t \simeq 100\,\mathrm{m\,s^{-2}}$. This is far larger

than observed. The implication is that the vertical component of equation (2.9) can be replaced by a statement of hydrostatic balance.

$$C_p\theta\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_0} = 0. \tag{2.10}$$

This approximation is very accurate on large scales, and is also used in many weather forecasting and climate models. Its relation to other possible large-scale approximations is discussed by White *et al.* (2005).

We summarize the resulting reduced version of (2.1), (2.3), (2.4) and (2.5), writing ∇_r for the horizontal components of the gradient operator and $\mathbf{u}_r = (u, v)$:

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}_r}{\mathrm{D}t} + (-fv, fu) + C_p\theta\nabla_r\Pi' &= 0,\\
C_p\theta\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_0} &= 0,\\
\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= 0,\\
\frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p\Pi}(S_h + LP),\\
\frac{\mathrm{D}q}{\mathrm{D}t} &= S_q - P,\\
p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta.
\end{aligned} \tag{2.11}$$

The hydrostatic approximation means that there is no explicit evolution equation for w. The next step is to deduce w from the other equations. The hydrostatic relation (2.10) and the equation of state together give two constraints between the thermodynamic variables Π, ρ and θ. Consistency of the separate evolution equations for ρ and θ yields 'Richardson's equation' (see White (2002)):

$$\gamma\frac{\partial}{\partial r}\left\{p\left(\frac{\partial w}{\partial r} + \nabla_r\cdot\mathbf{u} - \frac{1}{C_p\Pi}(S_h + LP)\right)\right\} = \frac{\partial p}{\partial r}\nabla_r\cdot\mathbf{u} - \frac{\partial\mathbf{u}_r}{\partial r}\cdot\nabla_r p. \tag{2.12}$$

It is also shown in White (2002) that equations (2.11) with $S_h = P = 0$ yield a potential vorticity conservation law of the form (2.7) with potential vorticity

$$Q = \frac{1}{\rho}(\hat{\mathbf{r}}\times\nabla\mathbf{u}_r + f\hat{\mathbf{r}})\cdot\nabla\theta. \tag{2.13}$$

Classification of large-scale flows

We now classify the various types of large-scale flow further. The approximations made in the previous subsubsection filter sound waves which have a component propagating in the vertical. We now carry out an analysis of

equations (2.10) to (2.12) linearized about a state of rest with a basic state θ which varies with r. This enables us to identify the different types of motion and the important flow parameters. These are then used to define asymptotic regimes.

As in Thuburn, Wood and Staniforth (2002*a*), a simple analytic treatment requires a choice of basic state where the squared speed of sound $c^2 = C_p\theta\Pi(\gamma-1)$ and the Brunt–Väisälä frequency $N = \sqrt{\frac{g}{\theta}\frac{\partial\theta}{\partial r}}$ are constants. Set $(p,\rho,\theta) = (p_1(r),\rho_1(r),\theta_1(r))$ with $C_p\theta_1\partial\Pi_1/\partial r + g = 0$. Setting $(p,\rho,\theta) = (p_1,\rho_1,\theta_1) + (p',\rho',\theta')$ in (2.11) and (2.12) yields the system

$$\begin{aligned}
\frac{\partial u}{\partial t} - fv + \frac{C_p\theta_1}{a\cos\phi}\frac{\partial\Pi'}{\partial\lambda} &= 0,\\
\frac{\partial v}{\partial t} + fu + \frac{C_p\theta_1}{a}\frac{\partial\Pi'}{\partial\phi} &= 0,\\
\frac{\partial\theta}{\partial t} + w\frac{\partial\theta_1}{\partial r} &= 0,\\
C_p\theta_1\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_1} &= 0,\\
(\gamma-1)\frac{\partial p_1}{\partial r}\nabla_r\cdot\mathbf{u} + \gamma\frac{\partial p_1}{\partial r}\frac{\partial w}{\partial r} + p_1\gamma\frac{\partial}{\partial r}\nabla_r\cdot\mathbf{u} &= 0.
\end{aligned} \tag{2.14}$$

This can be analysed in plane geometry by writing $x = a\lambda\cos\phi, y = a\phi, z = r$ and considering solutions proportional to $\exp^{\mathrm{i}(kx+ly-\omega t)}$. Write the vertical structure function as $Z(z)$. Following Thuburn *et al.* (2002*a*), this can be shown to yield non-trivial solutions if $\omega = 0$ or

$$-Zc^2\frac{(k^2+l^2)}{(\omega^2-f^2)} - Z + \frac{c^2}{N^2}\left(\frac{\mathrm{d}}{\mathrm{d}z} + \frac{N^2}{g}\right)\left(\frac{\mathrm{d}}{\mathrm{d}z} + \frac{g}{c^2}\right)Z = 0. \tag{2.15}$$

Equation (2.15) represents the dispersion equation for inertia-gravity waves. The boundary conditions on w imply that

$$\left(\frac{\mathrm{d}}{\mathrm{d}z} + \frac{g}{c^2}\right)Z = 0 \tag{2.16}$$

at $z = 0, z_{\mathrm{top}}$. This is an eigenvalue problem for $1+c^2\frac{(k^2+l^2)}{(\omega^2-f^2)}$. ω can then be calculated for given horizontal wave-numbers. This equation is identical in form to equation (4.9) of Thuburn *et al.* (2002*a*) with a different constant multiplying Z. The solutions therefore take the form given in Thuburn *et al.*'s equations (4.13) and (4.15). There are external modes, with

$$Z \propto \exp\left(-\frac{gz}{c^2}\right), \tag{2.17}$$

and internal modes

$$Z \propto \big(\Gamma \sin(mz) - m\cos(mz)\big) \exp\left(-\frac{\gamma g z}{2c^2}\right), \tag{2.18}$$

where $\Gamma = (1/2)(g/c^2 - N^2/g)$.

Consider the external modes first. Then

$$\left(\frac{\mathrm{d}}{\mathrm{d}z} + \frac{g}{c^2}\right) Z = 0$$

and ω is given by

$$\omega^2 = f^2 + c^2(k^2 + l^2). \tag{2.19}$$

This corresponds to a horizontally propagating sound wave, modified by rotation. The sound wave speed is $c = \sqrt{\gamma p_1/\rho_1}$ which is of order $300\,\mathrm{m\,s^{-1}}$. The associated external Froude number Fr is the ratio of the typical horizontal velocity U to c, and is usually less than 0.1 in the atmosphere. The frequency associated with the horizontal velocity is U/L. This is much less than ω if either $U \ll c$, so that Fr is small, or $U \ll fL$, so that the Rossby number $\mathrm{Ro} = U/fL$ is small. The two conditions coincide for a length-scale equal to the external *Rossby radius of deformation* L_R, in this case equal to c/f. This is about 3000 km in the atmosphere, corresponding to a wave-length of 20 000 km, or a planetary wave-number 2.

For the internal waves, the eigenvalue equation takes the form

$$c^2 \frac{(k^2 + l^2)}{(\omega^2 - f^2)} + 1 + \Lambda(m) = 0,$$

where $\Lambda(m)$ is a quadratic polynomial in m with leading term c^2m^2/N^2. For large m, this gives

$$\omega^2 \simeq f^2 + N^2 \frac{(k^2 + l^2)}{m^2}. \tag{2.20}$$

Given a characteristic horizontal velocity U, the associated frequency is U/L. Equation (2.20) shows that this frequency is significantly less than the inertia-gravity wave frequency ω if either $U/L \ll NH/L$, so the internal Froude number $U/(NH) \ll 1$, or $U/L \ll f$, so the Rossby number $U/(fL) \ll 1$. The two conditions coincide if $H/L = f/N$. This defines a characteristic aspect ratio. Equivalently, we can state that the horizontal scale is equal to the internal Rossby radius NH/f. If $f = 10^{-4}\,\mathrm{s^{-1}}$, which is characteristic of the extra-tropics, $N = 10^{-2}\,\mathrm{s^{-1}}$, which is characteristic of the troposphere, and $H = 10\,\mathrm{km}$, which is the tropospheric depth, then the internal Rossby radius is 1000 km, corresponding to a wavelength of 6000 km. The aspect ratio f/N is 0.01. In the stratosphere $N = 3 \times 10^{-1}\,\mathrm{s^{-1}}$, giving an internal Rossby radius of 3000 km and $f/N = 0.001$. If $U = 10\,\mathrm{m\,s^{-1}}$, then the frequency U/L is much less than the inertia-gravity wave frequency in the troposphere if either $H \gg 1\,\mathrm{km}$ or $L \gg 100\,\mathrm{km}$.

The aspect ratio f/N is important when understanding the behaviour of atmospheric flows with frequencies less than that given by (2.20). Large extra-tropical disturbances tend to develop with aspect ratios close to f/N because this configuration allows efficient conversion of potential to kinetic energy, as first established by Charney (1948). If the aspect ratio is less than f/N, the potential vorticity defined by (2.13) is approximately $\frac{1}{\rho} f\hat{\mathbf{r}} \cdot \nabla\theta$ and is thus primarily a measure of static stability, together with the variation of the planetary vorticity $f\hat{\mathbf{r}}$. The energy of disturbances is primarily potential rather than kinetic energy. If the aspect ratio is greater than f/N, but smaller than 1, which is required for the validity of the shallow atmosphere hydrostatic equations, then the potential vorticity is approximately $\frac{1}{\rho}(\hat{\mathbf{r}} \times \nabla\mathbf{u}_r + f\hat{\mathbf{r}}) \cdot \nabla\theta_1$, where $\theta_1(r)$ is a reference state such as used in (2.14). Thus potential vorticity variations are determined by vorticity variations. The energy of disturbances is primarily kinetic energy. These behaviours are analysed by Gill (1982), and also hold for disturbances independent of height when the threshold becomes the external Rossby radius L_R. The case of aspect ratios of order 1 is discussed in the next subsection. We will show in Section 3 that the behaviour of the solutions of (2.11) is qualitatively different as the aspect ratio changes through f/N. This is important in considering optimal numerical methods or data assimilation techniques.

If the frequency of forcing terms is comparable to the inertia-gravity wave frequency, then the inertia-gravity wave response will be important. This happens on large scales in the tropics where f tends to zero. Since f changes rapidly away from zero away from the equator, tropical waves tend to have a very asymmetric structure with a larger scale along the equator (Gill 1982), which determines the response to localized forcing on time-scales less than a few days. Tidal motions occur because the frequency (2.17) determined by the horizontal sound speed is comparable to the diurnal frequency of radiative forcing, so a resonant response is possible.

We now show how this analysis can be exploited in the nonlinear equations (2.11) and (2.12) by deriving a second-order wave equation for the evolution of the horizontal divergence. This will be needed when applying the analysis to production models. The horizontal momentum equations and the thermodynamic equation from (2.11) can be rewritten in component form as

$$\begin{aligned} \frac{\partial u}{\partial t} - fv + \frac{C_p\theta}{a\cos\phi}\frac{\partial \Pi'}{\partial \lambda} &= A_1, \\ \frac{\partial v}{\partial t} + fu + \frac{C_p\theta}{a}\frac{\partial \Pi'}{\partial \phi} &= A_2, \\ \frac{\partial \theta}{\partial t} + w\frac{\partial \theta}{\partial r} &= A_3. \end{aligned} \tag{2.21}$$

where A_1, A_2 and A_3 represent all the remaining terms. The choice of terms

retained on the left-hand side of equations (2.21) is suggested by the linear analysis above.

The first two of equations (2.21) can be combined into an equation for the evolution of the horizontal divergence $\Delta = \nabla_r \cdot \mathbf{u}_r$. The terms resulting from differentiating f and θ are transferred to the right-hand side:

$$\frac{\partial \Delta}{\partial t} - \frac{f}{a\cos\phi}\left(\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \phi}(u\cos\phi)\right) + C_p\theta\nabla_r^2\Pi' = A_4, \tag{2.22}$$

where $\nabla_r^2 \equiv \nabla_r \cdot (\nabla_r)$. Now calculate the second time derivative, substituting for $\partial u/\partial t$ and $\partial v/\partial t$ using the first two equations of (2.21), and again transferring some terms to the right-hand side:

$$\frac{\partial^2 \Delta}{\partial t^2} + f^2\Delta + C_p\theta\nabla_r^2\frac{\partial \Pi'}{\partial t} = A_5. \tag{2.23}$$

The final step is to differentiate with respect to r, to use the hydrostatic relation in the form $C_p\theta\partial\Pi/\partial r + g = 0$ and the time-independence of the basic state to substitute for $\partial^2\Pi'/\partial r\partial t$ in terms of $\partial\theta/\partial t$, and to use the third equation of (2.21) for $\partial\theta/\partial t$. This gives

$$\frac{\partial^2}{\partial t^2}\left(\frac{\partial \Delta}{\partial r}\right) + f^2\frac{\partial \Delta}{\partial r} + N^2\nabla_r^2 w = A_6, \tag{2.24}$$

where the Brunt–Väisälä frequency $N = \sqrt{\frac{g}{\theta}\frac{\partial\theta}{\partial r}}$. Using (2.12) it is possible to eliminate w in favour of Δ by integrating in r, thus obtaining the desired wave equation for the horizontal divergence. It takes the generic form

$$\frac{\partial^2 \Delta}{\partial t^2} + \mathbf{L}\Delta = A \tag{2.25}$$

where $\mathbf{L}$ is a positive definite linear operator. This operator will reduce to that derived from (2.14) if the atmospheric state is given by the reference state used in (2.14) and plane geometry is assumed.

If $\mathbf{L}\Delta$ is larger than the right-hand side terms A, then equation (2.25) describes forced linear inertia-gravity waves. If $\mathbf{L}\Delta$ and A are of similar magnitude, and the natural frequencies of the waves are large compared with those contained in A, then the response to the 'forcing' terms A can be expressed as the 'slow' equation

$$\mathbf{L}\Delta = A. \tag{2.26}$$

This equation will apply in the cases identified above where ω as given by (2.19) or (2.20) is much larger than U/L. When this happens $\mathbf{L}$ will have large eigenvalues compared with the frequencies implied by A.

Smaller-scale behaviour

Now consider smaller-scale behaviour. We characterize this by assuming an aspect ratio of O(1), so that the horizontal scale is of order 10 km. The hydrostatic approximation was justified by the assumption of a small aspect ratio, so is no longer applicable. For typical wind speeds, this horizontal scale implies a time-scale less than f^{-1}, so that the Rossby number $\mathrm{Ro} > 1$. The internal Froude number Fr may be greater or less than 1 according to the vertical scale and the strength of the stratification. This scale is, however, still sufficiently large for the viscous and thermal conductivity terms in (2.1) to be neglected. On these scales we can use Cartesian coordinates (x, y, z). Therefore consider equations (2.1) with (2.3), (2.4) and (2.5), omit the viscous and conductivity terms, and subtract the reference state (2.8) to give (2.9). We also omit the moisture equation and other forcing terms. This yields the equations

$$\begin{aligned} \frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + 2\boldsymbol{\Omega} \times \mathbf{u} + C_p\theta\nabla\Pi' - g\frac{\theta'}{\theta_0}\hat{\mathbf{r}} &= 0, \\ \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= 0, \\ \frac{\mathrm{D}\theta}{\mathrm{D}t} &= 0, \\ p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta. \end{aligned} \tag{2.27}$$

As in the large-scale case, we first study the behaviour of (2.27) linearized about the hydrostatic reference state defined before (2.14). We assume that rotation is unimportant on this scale, so the equations are

$$\begin{aligned} \frac{\partial\mathbf{u}}{\partial t} + C_p\theta_1\nabla\Pi' - g\frac{\theta'}{\theta_1}\hat{\mathbf{r}} &= 0, \\ \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho_1\mathbf{u}) &= 0, \\ \frac{\partial\theta}{\partial t} + w\frac{\partial\theta_1}{\partial z} &= 0, \\ \frac{1}{\gamma-1}\frac{\Pi'}{\Pi_1} &= \frac{\rho'}{\rho_1} + \frac{\theta'}{\theta_1}. \end{aligned} \tag{2.28}$$

We assume rigid upper and lower boundary conditions. These equations are a subset of those analysed by Thuburn, Wood and Staniforth (2002*b*). Consider solutions proportional to $\exp^{\mathrm{i}(kx+ly+mz)}$. The dispersion relation for the external mode can then be deduced from Thuburn *et al.*'s equation (4.4) as

$$\omega^2 = c^2(k^2 + l^2). \tag{2.29}$$

This corresponds to a horizontally propagating sound wave. The dispersion

relation for the internal modes is deduced from Thuburn *et al.*'s equation (4.5) as

$$(\omega^2 - c^2(k^2+l^2))(\omega^2 - N^2) - c^2\left(m^2 + \frac{1}{4}\left(\frac{g}{c^2} - \frac{N^2}{g}\right)^2\right)\omega^2 = 0. \quad (2.30)$$

This equation describes pairs of oppositely propagating gravity and sound waves. The sound waves propagate with speed c. For small vertical scales, the gravity wave dispersion relation is approximately $\omega^2 m^2 = N^2(k^2+l^2)$.

Observed atmospheric winds are always smaller than the speed of sound, usually by an order of magnitude. It is thus safe only to consider the case where $U \ll c$. The internal Froude number is defined as the ratio of the wind speed to the internal gravity wave speed, this is given approximately by $\mathrm{Fr} = U/NH$. This can be either greater or less than 1 according to the circumstances. Observations of flow over hills show that lee waves occur if $\mathrm{Fr} > 1$ and flow blocking, possibly with downstream hydraulic jumps, occurs if $\mathrm{Fr} < 1$ (Gill 1982). Since the qualitative nature of the solutions to (2.27) is different in these cases, the choice of optimal numerical methods and data assimilation techniques may also be different, as discussed in Section 3.

We now seek an analogue of the derivation of (2.26) from (2.25) by exploiting the knowledge that $U \ll c$. By analogy with the derivation of (2.21), rewrite equations (2.27) in the form

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + C_p\theta\nabla\Pi' - g\frac{\theta'}{\theta_0}\hat{\mathbf{k}} &= B_1, \\ \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= 0, \\ \frac{\partial \theta}{\partial t} + w\frac{\partial \theta_1}{\partial z} &= B_3, \\ p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta, \end{aligned} \quad (2.31)$$

where $\hat{\mathbf{k}}$ is a unit vector in the $z-$ direction. Take the divergence of the momentum equations and transfer the terms in $\partial\theta/\partial r$ and $\partial\theta'/\partial r$ to the right-hand side to give

$$\frac{\partial(\nabla\cdot\mathbf{u})}{\partial t} + C_p\theta\nabla^2\Pi' = B_4. \quad (2.32)$$

Differentiating the last equation of (2.31) with respect to time, the result can be written

$$\frac{\rho}{\gamma-1}\frac{\partial \Pi}{\partial t} - \Pi\frac{\partial \rho}{\partial t} = B_5. \quad (2.33)$$

Combining (2.33), (2.32) and the continuity equation from (2.31) gives

$$\frac{\partial^2(\nabla\cdot\mathbf{u})}{\partial t^2} - C_p\theta\Pi(\gamma-1)\nabla^2(\nabla\cdot\mathbf{u}) = B_6. \quad (2.34)$$

This is a forced second-order wave equation for $\nabla \cdot \mathbf{u}$. Now $C_p\theta\Pi(\gamma-1) = \gamma R\theta\Pi = \gamma p/\rho = c^2$, so we see that if $U \ll c$, (2.34) can be approximated by the slow equation

$$-c^2\nabla^2(\nabla \cdot \mathbf{u}) = B_6. \tag{2.35}$$

Note that this is an equation for the three-dimensional divergence, rather than the two-dimensional divergence which is determined by (2.26). This is an analogue of (2.26) which can be used on small scales.

2.4. Use of asymptotic limit solutions

We assume that the governing equations can be averaged, as discussed in Section 2.2. The viscous and thermal conductivity terms can be removed, and additional terms inserted on the right-hand sides of the equations representing the sub-grid model. After subtracting the reference state (2.8) to give (2.9), equations (2.1) with (2.3), (2.4) and (2.5) can be written as

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + 2\boldsymbol{\Omega} \times \mathbf{u} + C_p\theta\nabla\Pi' - g\frac{\theta'}{\theta_0}\hat{\mathbf{r}} &= F_{\mathbf{u}},\\
\frac{\partial\rho}{\partial t} + \nabla \cdot (\rho\mathbf{u}) &= 0,\\
\frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p\Pi}(S_h + LP) + F_\theta,\\
\frac{\mathrm{D}q}{\mathrm{D}t} &= S_q - P + F_q,\\
p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta.
\end{aligned} \tag{2.36}$$

Suppose that the averaging is on a horizontal scale L, a vertical scale H and a time-scale T. The solutions of the equations will then vary smoothly on these space- and time-scales, either in an Eulerian or a Lagrangian sense. The sub-grid terms may be introduced implicitly by the method of discretization. We can then apply analyses such as those set out in Section 2.3 on the assumption that the solutions of the equations only contain scales larger than L, H and T.

For example, if $L \gg 10\,\mathrm{km}$, the aspect ratio of the solutions must be small, and the shallow atmosphere hydrostatic equations, (2.11) will be satisfied accurately on all scales that are permitted by the averaging. They can then be used in the predictive equations. The right-hand side of (2.11) should now include the sub-grid model terms, giving

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}_r}{\mathrm{D}t} + (-fv, fu) + C_p\theta\nabla_r\Pi' &= F_{\mathbf{u}_r},\\
C_p\theta\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_0} &= F_w,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) &= 0, \\
\frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p \Pi}(S_h + LP) + F_\theta, \\
\frac{\mathrm{D}q}{\mathrm{D}t} &= S_q - P + F_q, \\
p_{\mathrm{ref}} \Pi^{\frac{1}{\gamma - 1}} &= \rho R \theta.
\end{aligned} \tag{2.37}$$

Note that the sub-grid term must even appear in the hydrostatic relation, since it is this term that ensures that $\mathrm{D}w/\mathrm{D}t$ is always small. Equations (2.37) are widely used in operational prediction models, though the F_w term is not normally included. For instance, the model currently used by the European Centre for Medium Range Weather Forecasts (ECMWF) uses these equations with an averaging scale of about 50 km.

Now consider the approximation of (2.25) by (2.26), which filters inertia-gravity waves. This is not valid on all scales permitted by current production models. This is because the typical T is less than 1 hour, while the inertial period $2\pi f^{-1}$ is at least 12 hours. However, the period of some inertia-gravity waves is much less than 1 hour. Thus (2.26) will be satisfied on some, but not all scales that are represented. This is discussed in detail in Section 3.2. Selective application of (2.26) is useful in data assimilation. It is also useful for ensuring that time-varying forcing is included in a way which ensures the response to it is on an appropriate time-scale. A further application is in validating model solutions.

The inclusion of the sub-grid terms is important when considering what conservation properties are important. Equation (2.11) conserves the potential vorticity (2.13) if the source and sink terms are zero. However, the inclusion of the sub-grid terms to give (2.37) will destroy potential vorticity conservation. Thus there is no longer a justification for enforcing it in discrete models.

In order to demonstrate that the sub-grid terms are correct, it is necessary to ensure that (2.36) or (2.37) have solutions which are smooth on the scales L, H and T. For example, Cao and Titi (2005) analyse a slightly simplified form of equations (2.37), and show that they can be solved if viscosity is included. Their proof depends critically on the viscosity, as it shows that the problem can be reduced to the three-dimensional Burgers equation $\mathrm{D}\mathbf{u}/\mathrm{D}t = 0$. This describes colliding particle trajectories and can only be solved if there is sufficient viscosity to prevent such collisions. Equations (2.37) can thus be solved if the sub-grid term $F_{\mathbf{u}_r}$ incorporates sufficient viscosity. A similar result was obtained by Lions, Temam and Wang (1992*a*), where the importance of including the term F_w in equations similar to (2.37) is demonstrated. With horizontal averaging scales of order 100 km, this level of viscosity may lead to inaccurate solutions, since it is

unlikely that the averaged behaviour of the atmosphere is like that of a fluid with a low enough Reynolds number to give smooth behaviour on this scale.

The limitation of the averaging approach is that, because of the nonlinearity of the equations, it is not possible to write down sub-grid models which will ensure that the solution of the modified governing equations will accurately represent the average of the solution of the original equations. An alternative approach is to prove that the behaviour of the unaveraged governing equations stays close to that of suitable asymptotic limit equations for 'long' time periods, typically of order the length of time that is being modelled. This requires proving the existence and uniqueness of solutions to the asymptotic limit equations for sufficiently large times to describe the phenomena of interest. The results can then be extended to numerical approximations to the governing equations. Typically, the proof of existence depends on particular conservation properties of the limit equations, and some analogue of these conservation properties should then be enforced on the numerical approximation to the governing equations.

In choosing useful asymptotic limit equations, it is desirable that existence of solutions can be proved without introducing extra regularization which is not appropriate for the regime being considered. If such terms have to be included, they may degrade the accuracy and usefulness of the resulting estimate. An example is provided by the viscosity that has to be included to solve (2.37). We show in Section 3.6 that this is probably much larger than the real viscosity, and will lead to solutions that suggest that large-scale circulations are more dissipative than they really are.

3. Numerical methods for particular asymptotic regimes

3.1. Introduction

In this section we consider a selection of the most important asymptotic regimes of the atmosphere introduced in the previous section. It does not attempt to be comprehensive, since almost all types of fluid flow can occur somewhere in the atmosphere. We discuss the structure of the appropriate limit equations, and identify the conservation properties which would be required to integrate the limit equations for long periods. We then discuss how these could be applied in integrations of the basic equations (2.1). We expect that it will be necessary to do this in order to ensure that the numerical solution of (2.1) will stay as close to the limit solution as it should.

3.2. Regimes with 'fast' inertia-gravity waves

Consider the case where the advection frequency U/L is much less than the inertia-gravity wave frequency ω calculated in (2.15). This is often called the 'balanced' regime, but this term is ambiguous and so better avoided.

We derive limit equations called the nonlinear balance equations from the shallow atmosphere hydrostatic equations (2.11) by using $\varepsilon = U/(\omega L)$ as the small parameter. For internal waves with small vertical scale, where ω is given by (2.20), ε can be related to the Rossby and internal Froude numbers by $\varepsilon^2 = (\mathrm{Ro}^{-2} + \mathrm{Fr}^{-2})^{-2}$. A similar equation applies for external waves with ω given by (2.19). There are many versions of this type of analysis in the literature; for instance, see McWilliams, Yavneh, Cullen, and Gent (1999), Warn, Bokhove, Shepherd and Vallis (1995), Lynch (1989), Holm (1996), Ford, McIntyre and Norton (2000) and Mohebalhojeh and Dritschel (2001).

Equations (2.11) can be rewritten in terms of the vertical component of the vorticity, ζ, and the horizontal divergence, Δ, defined by

$$\begin{aligned} \zeta &= \frac{1}{a\cos\phi}\left(\frac{\partial v}{\partial\lambda} - \frac{\partial u\cos\phi}{\partial\phi}\right), \\ \Delta &= \frac{1}{a\cos\phi}\left(\frac{\partial u}{\partial\lambda} + \frac{\partial v\cos\phi}{\partial\phi}\right). \end{aligned} \tag{3.1}$$

The horizontal momentum equations become

$$\frac{\partial\zeta}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial(\zeta+f)}{\partial\lambda} + \frac{v}{a}\frac{\partial(\zeta+f)}{\partial\phi} + w\frac{\partial\zeta}{\partial r} + (\zeta+f)\Delta + \frac{1}{a\cos\phi}\frac{\partial w}{\partial\lambda}\frac{\partial v}{\partial r} - \frac{1}{a}\frac{\partial w}{\partial\phi}\frac{\partial u}{\partial r} = 0, \tag{3.2}$$

$$\frac{\partial\Delta}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial\Delta}{\partial\lambda} + \frac{v}{a}\frac{\partial\Delta}{\partial\phi} + w\frac{\partial\Delta}{\partial r} + \Delta^2 - 2J(u,v) + \frac{1}{a\cos\phi}\frac{\partial w}{\partial\lambda}\frac{\partial u}{\partial r} + \frac{1}{a}\frac{\partial w}{\partial\phi}\frac{\partial v}{\partial r} + \nabla_r\cdot C_p\theta\nabla_r\Pi' - \nabla_r\cdot(fv,-fu) = 0, \tag{3.3}$$

where ∇_r again denotes a horizontal derivative and $(u,v) \equiv \mathbf{u}_r$. We now follow the steps used to derive (2.24), but making explicit some of the most important terms in A_6. This gives

$$\begin{aligned} -\frac{\partial^2}{\partial t^2}\frac{\partial^2 w}{\partial r^2} - f\frac{\partial}{\partial r}\left((\zeta+f)\frac{\partial w}{\partial r}\right) - \frac{g}{\theta}\nabla_r^2\left(w\frac{\partial\theta}{\partial r}\right) = \\ \frac{g}{\theta}\nabla_r^2\left(\mathbf{u}_r\cdot\nabla_r\theta - \frac{1}{C_p\Pi}(S_h + LP)\right) \\ + \frac{\partial}{\partial r}\left(-f\mathbf{u}_r\cdot\nabla_r(\zeta+2f) + 2\frac{\partial}{\partial t}J(u,v)\right) + \text{remainder}. \end{aligned} \tag{3.4}$$

The linearization of this equation describes inertia-gravity waves with the frequency ω calculated in (2.15). Consider a regime with velocity-scales

and length-scales U and L, such that $\varepsilon = U/(L\omega) \ll 1$, and depth-scale H such that $H \ll L$, so that (3.2) and (3.4) are valid. Then the argument that was used to derive (2.26) from (2.25) shows that equation (3.4) can be approximated by

$$f\frac{\partial}{\partial r}\left((\zeta+f)\frac{\partial w}{\partial r}\right) + \nabla_r^2(N^2 w) = \tag{3.5}$$
$$-\frac{g}{\theta}\nabla_r^2\left(\mathbf{u}_r\cdot\nabla_r\theta - \frac{1}{C_p\Pi}(S_h + LP)\right)$$
$$-\frac{\partial}{\partial r}\left(-f\mathbf{u}_r\cdot\nabla(\zeta+2f) + 2\frac{\partial}{\partial t}J(u,v)\right) + \text{remainder}.$$

It can be shown that it is also consistent to neglect many but not all of the terms in 'remainder'. Equation (2.26) shows that $\Delta = \mathbf{L}^{-1}A$ and is thus small under our assumptions. The second equation of (3.2) can then be approximated by

$$-2J(u,v) + \frac{1}{a\cos\phi}\frac{\partial w}{\partial\lambda}\frac{\partial u}{\partial r} + \frac{1}{a}\frac{\partial w}{\partial\phi}\frac{\partial v}{\partial r} + \nabla_r\cdot C_p\theta\nabla_r\Pi' - \nabla_r\cdot(fv,-fu) = 0. \tag{3.6}$$

Equation (3.5) is exactly of the form (2.26), and is thus used to define $\mathbf{L}$. The system of equations comprising (3.1), (3.2), the diagnostic relations (2.10) and (2.12), the definition of Π', and equations (3.5) and (3.6) form eight equations for $u, v, w, \zeta, \Delta, \theta, p, \Pi'$. These form the nonlinear balance equations. After some manipulations, further consistent approximations, and the use of potential temperature as a coordinate, these reduce to the equations used in McWilliams *et al.* (1999), subject to changes in notation. It is also shown there that the resulting equations retain the potential vorticity conservation law (2.13).

Solvability of these equations is shown by McWilliams *et al.* (1999) to depend on the conditions

(i) $\partial\theta/\partial r$ does not change sign,

(ii) $\zeta + f$ does not change sign,

(iii)
$$(\zeta+f)^2 - \left(\frac{\partial^2\psi}{\partial x^2} - \frac{\partial^2\psi}{\partial y^2}\right)^2 - 4\left(\frac{\partial^2\psi}{\partial x\partial y}\right)^2 > 0, \tag{3.7}$$

where the stream-function ψ is defined by $\nabla_r^2\psi = \zeta$, with ζ defined by (3.1). The product of conditions (i) and (ii) requires that the potential vorticity does not change sign. Since (2.13) still holds, spontaneous violations of condition (i) are only possible if both (i) and (ii) are violated simultaneously. Condition (iii) is liable to spontaneous violations as it is not a constant of

the motion. Computations (McWilliams and Yavneh 1998) show that these do indeed occur. If the initial data has small Rossby number, the velocity gradients have magnitude $U/L \ll f$ and condition (3.7) will be satisfied. If the initial data has small internal Froude number $U/(NH)$ but large Rossby number, (3.7) may not be satisfied even though $H \ll L$. The solvability conditions result from writing the equations as a fully coupled system. If the Froude number is small, but the Rossby number is large, the eigenvalue of the system associated with $\partial\theta/\partial r$ is much larger than the others. A good approximation to the system can then be solved by decoupling it, as described in Section 3.5.

The nonlinear balance equations approximate equations (2.11) to an accuracy $\mathrm{O}(\varepsilon^2)$. This was achieved by neglecting the first and second time derivatives of Δ in (3.3) and (3.4). In Mohebalhojeh and Dritschel (2001) it is shown that approximations of any polynomial order to (2.11) could be generated by neglecting successively higher time derivatives in deriving approximations to equations (3.3) and (3.4). If the resulting equations could be solved for long time periods, it would be possible to prove that solutions of (2.11) with suitable initial data remained close to the solutions of the nonlinear balance equations for long times, and the inertia-gravity waves would thus be restricted to very small amplitudes. The apparent lack of solvability means that there is no such thing as a solution of the nonlinear balance equations for long times which can be compared with the solution of the original equations, and so no such conclusion can be drawn. In McWilliams and Yavneh (1998) it is suggested that spontaneous violations of condition (3.7) correspond to local bursts of inertia-gravity wave activity.

The discussion in Section 2.4 suggests that it would be of interest to regularize the equations by including sub-grid terms in a way that ensured solvability. The approximations made in deriving the nonlinear balance equations are derived in an Eulerian frame. Start from the averaged shallow atmosphere hydrostatic equations (2.37), which requires assuming a horizontal averaging scale $L \gg 10\,\mathrm{km}$. Assume additionally that L is large enough for ε to be small. Following the derivation used to obtain equation (3.4) gives

$$
\begin{aligned}
&-\frac{\partial^2}{\partial t^2}\frac{\partial^2 w}{\partial r^2} - f\frac{\partial}{\partial r}\left((\zeta+f)\frac{\partial w}{\partial r}\right) - \frac{g}{\theta}\nabla_r^2\left(w\frac{\partial\theta}{\partial r}\right) = \qquad (3.8)\\
&\quad \frac{g}{\theta}\nabla_r^2\left(\mathbf{u}_r\cdot\nabla_r\theta - \frac{1}{C_p\Pi}(S_h + LP)\right)\\
&\quad + \frac{\partial}{\partial r}\left(-f\mathbf{u}_r\cdot\nabla_r(\zeta+2f) + 2\frac{\partial}{\partial t}J(u,v)\right) + \frac{\partial^2}{\partial t\partial r}(\nabla\cdot F_{\mathbf{u}_r}) + \text{remainder}.
\end{aligned}
$$

A sub-grid term $\frac{\partial^2}{\partial t\partial r}(\nabla\cdot F_{\mathbf{u}_r})$ has been introduced.

The assumption $\varepsilon \ll 1$ means that (3.8) can be approximated by

$$f\frac{\partial}{\partial r}\left((\zeta+f)\frac{\partial w}{\partial r}\right)+\nabla_r^2\big(N^2 w\big)= \tag{3.9}$$
$$-\frac{g}{\theta}\nabla_r^2\left(\mathbf{u}_r\cdot\nabla_r\theta-\frac{1}{C_p\Pi}(S_h+LP)\right)$$
$$-\frac{\partial}{\partial r}\left(-f\mathbf{u}_r\cdot\nabla_r(\zeta+2f)+2\frac{\partial}{\partial t}J(u,v)\right)-\frac{\partial^2}{\partial t\partial r}(\nabla\cdot F_{\mathbf{u}_r}),$$

which is derived by including the sub-grid term in (3.5). Similarly, a sub-grid term has to be added to the right-hand side of (3.6), giving

$$-2J(u,v)+\frac{1}{a\cos\phi}\frac{\partial w}{\partial\lambda}\frac{\partial u}{\partial r}+\frac{1}{a}\frac{\partial w}{\partial\phi}\frac{\partial v}{\partial r} \tag{3.10}$$
$$+\nabla_r\cdot C_p\theta\nabla_r\Pi'-\nabla_r\cdot(fv,-fu)=\nabla\cdot(F_{\mathbf{u}_r}).$$

Since (3.7) is satisfied under these conditions, the modified form (3.9) and (3.10) of the nonlinear balance equations can be solved, and the solutions will be close to those of equations (2.37).

Sub-grid terms will now appear in the potential vorticity equation. Potential vorticity conservation will then no longer apply. An example is given by Ziemianski and Thorpe (2003), though with a different interpretation. In their computations using a numerical solution of (2.1), large sources of potential vorticity appear. These are in a frontal zone, where there is organized unresolved small-scale activity in the solution, and so potential vorticity, which is nonlinear, would not be conserved under the Eulerian averaging implied by the model. This is represented by the sub-grid term in the potential vorticity equation and would be likely to explain the results shown in that paper. The effect of Lagrangian averaging on potential vorticity conservation for a frontal zone is discussed in Section 4.3.

Now, given $L \gg 10\,\mathrm{km}$, we achieve the condition that ε is small by choosing a sufficiently large vertical averaging scale H. This choice of L is not sufficient to ensure that (3.7) is satisfied. In this case (3.9) and (3.10) will hold, but the time evolution cannot be calculated using the nonlinear balance equations. An alternative procedure which does work is described in Section 3.5.

The discussion above shows that the diagnostic conditions (3.9) and (3.10) are useful, but can only be applied selectively. This is exploited in the semi-implicit method of time integration introduced by Robert (1981). This is widely used in operational models, such as the ECMWF model (Ritchie, Temperton, Simmons, Hortal, Davies, Dent and Hamrud 1995), and the Met Office Unified Model (UM) (Davies, Cullen, Malcolm, Mawson, Staniforth, White and Wood 2005). Robert's motivation was to increase the efficiency of

operational models by not resolving the time evolution of fast inertia-gravity waves. Thus the time-step is chosen such that $U\delta t \leq 1$, where U is the maximum velocity for which advection has to be treated accurately. Inertia-gravity waves will not then be resolved in time in cases where $\varepsilon \ll 1$. In such cases, the solution will satisfy the nonlinear balance conditions (3.9), (3.10) accurately. We therefore seek a time discretization of equations (2.37) which ensures that the solution satisfies the nonlinear balance conditions whenever $\varepsilon \ll 1$, while noting that this condition will not be satisfied everywhere. We illustrate the method by writing a time discretization of equation (2.21) as

$$\begin{aligned} \delta_t u - f\overline{v}^t + \frac{C_p\overline{\theta}^t}{a\cos\phi}\frac{\partial\overline{\Pi}'^t}{\partial\lambda} &= \overline{A}_1^t, \\ \delta_t v + f\overline{u}^t + \frac{C_p\overline{\theta}^t}{a}\frac{\partial\overline{\Pi}'^t}{\partial\phi} &= \overline{A}_2^t, \\ \delta_t\theta + \overline{w}^t\frac{\partial\overline{\theta}^t}{\partial r} &= \overline{A}_3^t, \end{aligned} \tag{3.11}$$

where $\delta_t u = (u(t+\delta t) - u(t))/\delta t$ and $\overline{v}^t = (v(t+\delta t) + v(t))/2$.

We can now go through the same steps as were used to derive (2.25) to give a second-order equation for $\Delta(t+\delta t)$. The result is

$$\delta_{tt}\Delta + \overline{\mathbf{L}\Delta}^{tt} = \overline{A}^{tt}. \tag{3.12}$$

Assuming A is proportional to $\mathrm{e}^{\mathrm{i}\nu t}$, and replacing $\mathbf{L}$ by its eigenvalue ω^2, there is a solution of (3.12) proportional to $\mathrm{e}^{\mathrm{i}\nu t}$ which is

$$-\sin^2(\tfrac{1}{2}\nu\delta t)\Delta + \delta t^2\omega^2\cos^2(\tfrac{1}{2}\nu\delta t)\Delta = A\delta t^2\cos^2(\tfrac{1}{2}\nu\delta t). \tag{3.13}$$

For non-zero A, since $\omega\delta t \gg 1$, this reduces to $\omega^2\Delta = A$ as desired. This is because of the consistent time averaging on both sides of the equation. If an equation of the form (3.12) were derived from a three-time-level approximation to (3.11), then the time averaging of A would not be required for numerical stability, and (3.13) would reduce to $\omega^2\cos^2(\frac{1}{2}\nu\delta t)\Delta = A$, thus degrading the accuracy of the balanced solution. The frequency of free inertia-gravity waves (obtained by setting $A = 0$), is reduced to $\omega\cos(\omega\delta t)/\sin(\omega\delta t) \ll \omega$. The inertia-gravity waves in the solution are thus no longer accurately treated.

This illustrates that, if $\omega^2\delta t^2 \gg 1$, (3.12) can be approximated by

$$\overline{\mathbf{L}\Delta}^{tt} = \overline{A}^{tt} \tag{3.14}$$

which is a discretization of (2.26). This approximation is exactly that used to go from (3.4) to (3.5), which defines the nonlinear balance equations. Moreover, the presence of the first term on the left-hand side of (3.12) will ensure solvability if δt is small enough, even though (3.14) may not be

solvable. The use of (3.12) therefore achieves the aim of enforcing nonlinear balance selectively where $\omega^2\delta t^2 \geq 1$.

It is important that the spatial approximations maintain the condition $|\mathbf{L}|\delta t^2 \geq 1$. Some numerical discretizations will allow the eigenvalues of $\mathbf{L}$ to become zero on the smallest spatial scales. Examples of this are given in Section 3.3, and detailed analysis is given in Durran (1998). It is important to note that the operator $\mathbf{L}$ which appears in (2.26) is a variable-coefficient operator. This is necessary to ensure that the nonlinear balance equations are valid on large scales, where the coefficients may have O(1) variations. Most implementations of the semi-implicit method have used a constant-coefficient operator, which will thus not ensure consistency with the balanced solution. The UM uses a variable-coefficient operator for this reason. If equation (3.14) is solved iteratively, with a constant-coefficient approximation to $\mathbf{L}$ as a preconditioner, then the accuracy of the balanced solution will be degraded if the convergence of the iteration is insufficient.

The enforcement of balance can be strengthened by using a decentred time averaging of the terms on the left-hand side of (3.11). If $\mathbf{L}$ is time-independent, this yields

$$\delta_{tt}\Delta + \mathbf{L}(\alpha^2\Delta(t+\delta t) + 2\alpha(1-\alpha)\Delta(t) + (1-\alpha)^2\Delta(t-\delta t) = \overline{A}^{tt}, \quad (3.15)$$

as an approximate replacement for (3.14). The condition $\alpha \geq \frac{1}{2}$ is required for stability. The decentering will not change the potential vorticity conservation law, except for the change to the advection of the potential vorticity by the divergent wind. The inertia-gravity waves which are not well-resolved in time will be damped, but this is desirable as they are no longer travelling at the correct speed.

When considering the effect of decentred time integration in production atmospheric models, note that invariably $f\delta t \ll 1$, so that the rotational contribution to ω as defined in (2.20) will not be significant on scales where $\omega\delta t \gg 1$. In the global forecast version of the UM, the horizontal grid-length is about 40 km, and the time-step 15 minutes. The Courant number $U\delta t/\delta x$ will thus be 1 for a speed U of about $45\,\mathrm{m\,s^{-1}}$. The condition $\omega\delta t \geq 1$ is satisfied for the external gravity wave which has speed approximately $300\,\mathrm{m\,s^{-1}}$ if the spatial wave-length is less than 300 km, and thus decentering will suppress the waves with wave-lengths up to about 1000 km. Internal gravity waves with speeds greater than $45\,\mathrm{m\,s^{-1}}$ will be suppressed on progressively smaller scales. An example of the effect of decentering with explicit time integration is described and analysed by Fox-Rabinowitz (1996), while the effect of decentred implicit time-differencing was pointed out by A. Staniforth (personal communication); see Section K.5.6 of Staniforth, White, Wood, Thuburn, Zerroukat and Cordero (2002).

While decentering is useful in preventing rapidly oscillating solutions, because $f\delta t \ll 1$ it is far short of selecting out the meteorologically significant

motions which are rotation-dominated. The model will thus resolve many other motions, including external gravity waves with a wave-length greater than about 1000 km. The use of decentred time integration may be particularly useful when forcing terms with high time-frequency are added to the equations, since their effect is then time-averaged to fit the time-resolution of the model. It may also be useful in four-dimensional variational data assimilation, where observation increments are added in a way which may not be compatible with the diagnostic relations (3.5) and (3.6). This was the motivation for the paper by Fox-Rabinowitz (1996).

Equation (3.11) represents a fully implicit time-discretization of (2.21). The time-average on the right-hand side generates a nonlinear implicit problem for the values at time $t + \delta t$. As noted by Staniforth, the procedure introduced by Fox-Rabinowitz (1996) formed the first two iterations of an implicit procedure for solving this system. It should also be noted that the Heun advection scheme used in the original UM, Cullen and Davies (1990) takes the same form. We illustrate by displaying the iterations for solving the first equation of (3.11):

$$\begin{aligned} \frac{(u^* - u(t))}{\delta t} - f\overline{v}^* + \frac{C_p\theta(t)}{a\cos\phi}\frac{\partial\overline{\Pi}'^*}{\partial\lambda} &= A_1(t), \\ \delta_t u - f\overline{v}^t + \frac{C_p\overline{\theta}^*}{a\cos\phi}\frac{\partial\overline{\Pi}'^t}{\partial\lambda} &= \overline{A}_1^*, \end{aligned} \tag{3.16}$$

where the notation $\overline{v}^*$ indicates $\frac{1}{2}(v(t)+v^*)$. Each iteration requires solution of an implicit problem of the form (3.12). Yeh, Côté, Gravel, Méthot, Patoine, Roch and Staniforth (2002) used this scheme in the Canadian GEM model, and Cullen (2001) demonstrated its use in the ECMWF model. In four-dimensional variation data assimilation, it is necessary to integrate a linearized version of the model forward in time. An obvious method is to use the first iteration of the procedure (3.16) as the basis for the linear model. The right-hand side terms will now represent perturbations to A_1, and therefore be linear in the evolution variables, Lawless, Nichols and Ballard (2003).

An example using the UM is shown in Figures 3.1 and 3.2. The model solves equations (2.1) by a semi-implicit scheme as described in Davies *et al.* (2005). The balance constraint is applied by using backward time-weighting in the semi-implicit scheme as in (3.12). The operational UM has $\alpha = 0.7$ and the modified version illustrated has $\alpha = 1$. Figure 3.1 shows that the decentering has a very small effect on even the small-scale structure of the potential vorticity Q, while Figure 3.2 shows a significant reduction in the divergence tendency $\partial\Delta/\partial t$, which should be small if $\varepsilon \ll 1$. The larger values in the operational forecast probably mainly reflect the effect of insufficiently constrained initial data and insufficiently smooth representations of

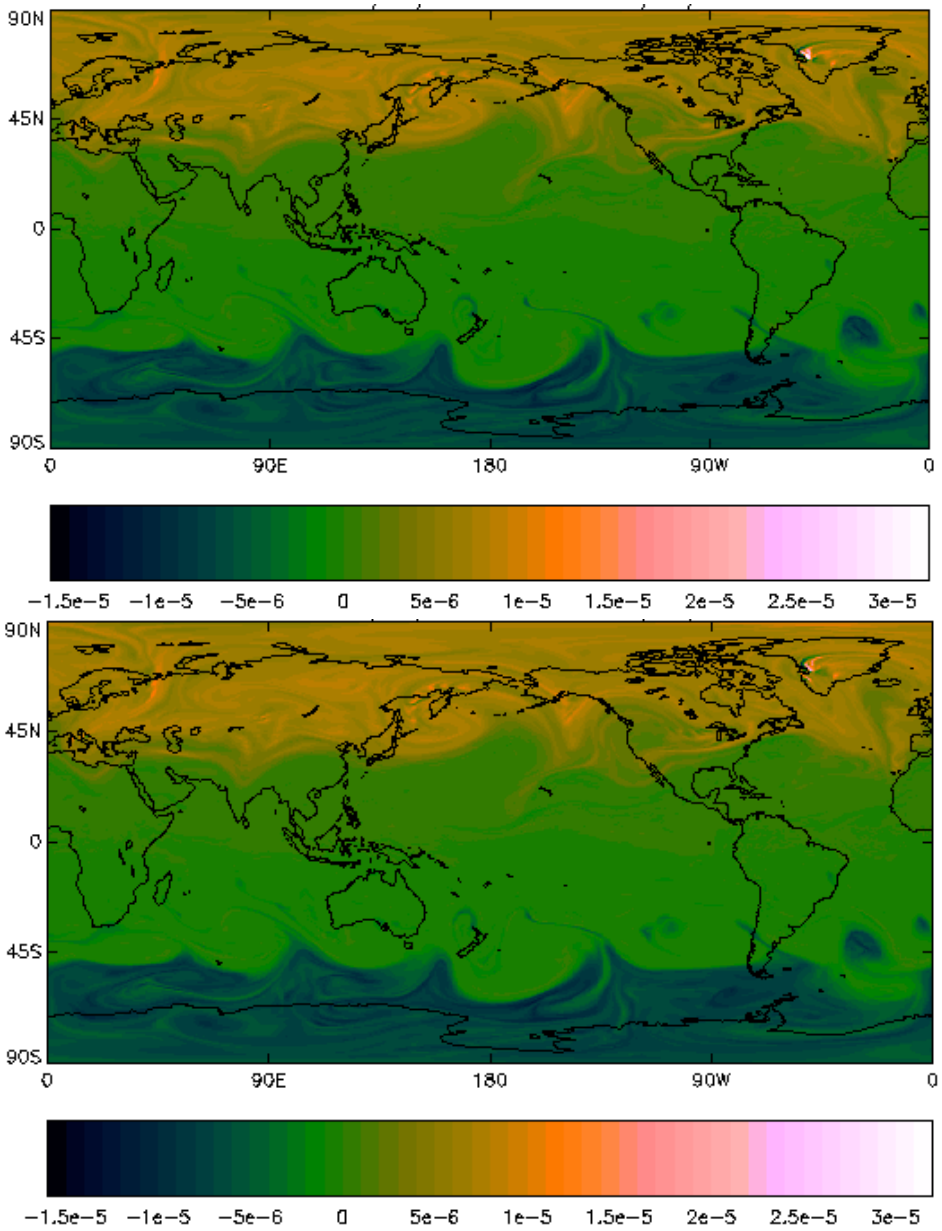

Figure 3.1. Ertel potential vorticity at $\theta = 330\,\mathrm{K}$ predicted by the UM for 0900UTC 24 January 2006 from data at 0900UTC on 22 January 2006. *Top panel*: $\alpha = 0.7$. *Bottom panel*: $\alpha = 1.0$.

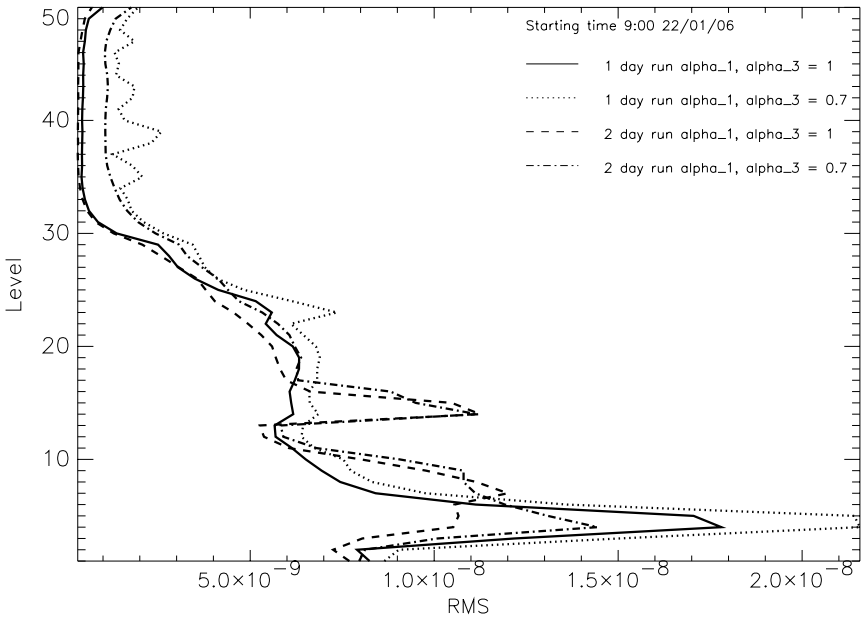

Figure 3.2. Vertical profile of r.m.s. divergence tendency (s^{-2}) plotted against model level from the same forecasts as shown in Figure 3.1.

the forcing terms. However, it is also shown by Mohebalhojeh and Dritschel (2004) that unrealistic divergence tendencies can be created by a poor choice of prognostic variables.

The constraints (3.4) and (3.5) are enforced in a scale-dependent way in other methods of initialization. The bounded derivative method of Browning and Kreiss (1994) is a systematic way of doing this, based on the analysis above. Digital filter methods (Lynch 1997) apply time-filtering, but allow a longer time-scale than the time-step to be used to define the frequency cut-off. Nonlinear normal-mode initialization (Machenhauer 1977) enforces (3.14) for all inertia-gravity waves with sufficiently small m, thus applying a speed constraint rather than a frequency constraint. A typical value is about $50\,\mathrm{m\,s^{-1}}$. These methods are usually modified to ensure that the diurnal tidal signal is not filtered out.

Since the nonlinear balance equations, including the sub-grid terms, can be written as prediction of potential vorticity, Q, together with diagnostic relations to recover the other variables, it is natural to try and exploit this structure in numerical methods. It was shown by Mohebalhojeh and Dritschel (2004) that in order to achieve the best approximations to the nonlinear balance solutions, the model should be formulated in variables which

separate the potential vorticity from variables describing inertia-gravity waves. Recent work by D. Devlin and D. Dritschel (personal communication) suggests that decentering is more effective at preserving nonlinear balance when these variables are used. Exploiting this requires inversion of the operator $\mathbf{L}$ in order to make the change of variables. This is difficult because of the scale-dependence of the inertia-gravity wave frequency ω, equation (2.15). This means that $\mathbf{L}$ will have a large condition number which may be made worse by a poor choice of discretization, as discussed in Section 3.3. It is therefore expensive to invert it, which is why these transformations are not yet in widespread use. The potential benefit of using these variables for data assimilation is illustrated by Cullen (2003).

A number of the lessons for numerical methods have been pointed out in this section, in particular the need to preserve large values of $|\mathbf{L}|$ in the spatial discretization, the need to maintain the condition $\mathbf{L}\Delta = A$ in the large time-step limit, and the need to preserve accuracy in the advection of potential vorticity. These issues will be analysed further in Section 3.3 in a constant-coefficient context. It is not, however, possible to identify what properties are needed to preserve long-time accuracy for cases where $\epsilon \ll 1$ from this analysis because of the non-existence of long-time solutions to the nonlinear balance equations. In order to make further progress, we split the regime $\varepsilon \ll 1$ into parts where $H/L \simeq f/N$, $H/L < f/N$, and $H/L > f/N$, as discussed in Section 2.3. These cases are discussed respectively in Sections 3.3, 3.4 and 3.5. It is then possible to find solvable systems of equations appropriate to each case and thus possible in principle to prove long-time estimates for the errors in numerical methods.

3.3. The quasi-geostrophic regime

This regime is defined by the requirement that the Rossby and internal Froude numbers Ro and Fr are small and equal. It was introduced by Charney (1948). This assumption requires the aspect ratio to be small, typically 10^{-2} in the troposphere and 10^{-3} in the stratosphere. Under these conditions $\varepsilon = \sqrt{2}\,\mathrm{Ro} = \sqrt{2}\,\mathrm{Fr}$ and is thus small, so the nonlinear balance approximation can be made. It is thus appropriate to start from (3.1), (3.2), (2.10), (2.12), (3.5) and (3.6).

We seek a system of equations which is valid in this regime and can be solved for large times. The quasi-geostrophic equations are thus derived as a leading-order approximation in ε which is chosen so that energy is conserved, which aids the proof of solvability. A higher-order version is derived in Bourgeois and Beale (1994), but this cannot be solved for arbitrarily large times. The equations are given below. The justification for the various approximations is quite complex and details of the analysis are given by Pedlosky (1987). In particular, in some terms θ is replaced by a hydrostatic

reference state value $\theta_1(r)$, chosen such that the reference value N_0^2 of N^2 is a constant as in Section 2.3. Also f is replaced by a reference value f_0 except where it is differentiated. Thus we obtain

$$\begin{aligned}
\frac{\partial \zeta}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial(\zeta+f)}{\partial\lambda} + \frac{v}{a}\frac{\partial(\zeta+f)}{\partial\phi} + f_0\Delta = 0,\\
\frac{\partial\theta}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial\theta}{\partial\lambda} + \frac{v}{a}\frac{\partial\theta}{\partial\phi} + w\frac{\partial\theta_1}{\partial r} = \frac{1}{C_p\Pi}(S_h + LP),\\
C_p\theta_1\frac{1}{a\cos\phi}\frac{\partial}{\partial\lambda}\Pi' - f_0 v = 0,\\
C_p\theta_1\frac{1}{a}\frac{\partial}{\partial\phi}\Pi' + f_0 u = 0,\\
C_p\theta_1\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_1} = 0,\\
C_p\theta_1\frac{\partial\Pi_1}{\partial r} + g = 0,\\
\frac{\partial}{\partial r}\left(\frac{w}{\theta_1}\right) + \frac{\Delta}{\theta_1} = 0.
\end{aligned} \tag{3.17}$$

Here ζ and Δ are still defined by (3.1) and $\theta' = \theta - \theta_1$, $\Pi' = \Pi - \Pi_1$. Consistency between equations (3.17) implies that w satisfies

$$\begin{aligned}
f_0^2\frac{\partial^2}{\partial r^2}\left(\frac{w}{\theta_1}\right) + N_0^2\nabla_r^2\left(\frac{w}{\theta_1}\right) = \\
-\frac{g}{\theta_1^2}\nabla_r^2\left(\mathbf{u}_r\cdot\nabla_r\theta - \frac{1}{C_p\Pi}(S_h+LP)\right) + f_0\frac{\partial}{\partial r}\left(\frac{1}{\theta_1}\mathbf{u}_r\cdot\nabla(\zeta+f)\right).
\end{aligned} \tag{3.18}$$

In the absence of source terms, these equations after some manipulations imply the potential vorticity conservation law

$$\begin{aligned}
\frac{\partial Q}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial Q}{\partial\lambda} + \frac{v}{a}\frac{\partial Q}{\partial\phi} = 0,\\
Q = \frac{1}{\theta_1}\left(N_0^2(\zeta+f) + f_0\frac{\partial}{\partial r}\left(\frac{g\theta'}{\theta_1}\right)\right).
\end{aligned} \tag{3.19}$$

The horizontal velocities and potential temperature can be calculated from Q by first using the third, fourth and fifth equations of (3.17) to give a Poisson equation for Π' given Q:

$$f_0 Q = \frac{N_0^2 f f_0}{\theta_1} + C_p\left(N_0^2\nabla_r^2\Pi' + f_0^2\frac{\partial^2\Pi'}{\partial r^2}\right). \tag{3.20}$$

The horizontal velocities are then calculated from the third and fourth equations of (3.17) and the potential temperature from the fifth equation. w can then be calculated from the final equation.

Equation (3.20) can be solved for Π' given suitable boundary conditions. Neumann boundary conditions on Π' imply that θ' is given on horizontal boundaries and u or v on lateral boundaries. The proofs of solvability in Bourgeois and Beale (1994), which use plane geometry, assume periodic boundary conditions in the horizontal and constant values of θ' at the upper and lower boundaries. Other proofs are given in Majda (2003) and Bennett and Kloeden (1981, 1982).

The proofs cited above show that solvability of equations (3.17) depends on the fact that equation (3.19) cannot generate any values of Q outside the range of the initial values. The constant-coefficient elliptic equation (3.20) is used to find u, v and θ' from Q, and this is shown in Bourgeois and Beale (1994) to give a bound on the gradients of (u, v) in terms of bounds on Q and its derivatives:

$$\| \nabla \mathbf{u} \|_{L^\infty} + \| \nabla \theta' \|_{L^\infty} \leq C \| Q \|_{L^\infty} \left(1 + \log^+ \frac{\| Q \|_{H^s}}{\| Q \|_{L^\infty}} \right) \tag{3.21}$$

where $s \geq 2$. The L^∞ norm is the maximum norm, and H^s measures the L^2 norm of all derivatives up to order s. The $+$ superscript indicates that only positive values are used. C is a constant. Precise definitions are given in Bourgeois and Beale (1994). The maximum value of Q is independent of time, so it is necessary to control the mean-square value of the derivatives of Q. It is shown by differentiating (3.19) that the rate of increase of the gradient of Q is controlled by the gradients of $\mathbf{u}$, which can be estimated from (3.21). This gives an estimate for any time t that

$$\| Q(t) \|_{H^s} \leq C \| Q(0) \|_{H^s} \exp\bigl(C(\mathrm{e}^{Ct\|Q(0)\|_{H^s}} - 1)\bigr). \tag{3.22}$$

This can be calculated for any t given the initial data for Q, and provides the required long-time estimate.

In order to ensure that a numerical approximation to (3.17) remains bounded for large times, in other words is nonlinearly stable, it is necessary to be able to derive finite-dimensional analogues of the estimates (3.21) and (3.22).

We first have to maintain the bound on Q by its initial values. If Q is a prognostic variable, this can be achieved by using a monotonicity-preserving advection scheme. There is a very large literature on this topic, though much of it deals with enforcing monotonicity in high-speed flow problem with shocks. The method has the disadvantage that monotonicity enforcement inevitably involves diffusion and thus energy loss. This can be limited by use of sufficiently selective filters. For instance, recent work by Zerroukat, Wood and Staniforth (2005) demonstrates an appropriate selective filter for cases where non-monotonicity can only be generated by errors in upstream interpolation. The bound on Q can also be maintained by a fully Lagrangian

method such as one based on contour dynamics, as in Mohebalhojeh and Dritschel (2004). Such methods are not inherently diffusive.

The prescription above could be followed in solutions of (2.1) or (2.11) if the potential vorticity is used as a prognostic variable as in Mohebalhojeh and Dritschel (2004). This will be a good approximation to $\mathbf{Q}$ in the quasi-geostrophic regime. We will show how a similar requirement may be achievable without introducing new prognostic variables in Section 3.4.

Another approach is to note that (3.19), together with the non-divergence of the advecting velocity (u, v) resulting from the third and fourth equations of (3.17), implies that all moments $\int Q^n$ are conserved. A numerical method can only conserve a finite number of moments, and a popular method is to conserve only the first and second moments, following Sadourny (1975). This method has the disadvantage that, even if the distribution of Q is well-resolved at the initial time, it may not stay well-resolved. Conservation of the integral of Q^2 may thus result in variance being incorrectly retained in resolved scales.

The derivation of (3.21) involves controlling the second derivatives of Π' resulting from the solution of the elliptic equation (3.20). This requires the approximation to the elliptic operator to satisfy a discrete maximum principle. This is a standard requirement in numerical solution of elliptic equations, and is discussed, for instance, by Ganzha (1996, p. 216). It leads to the standard five-point discrete approximation to the Laplacian in two dimensions. The condition will be satisfied if the three-dimensional Laplacian in (3.20) is approximated by a seven-point stencil. However, if the derivation of (3.20) from (3.17) is carried out at the discrete level, a seven-point stencil is only obtained if u is held at grid-points staggered from Π in the ϕ-direction, v at grid-points staggered from Π in the λ-direction, and θ at grid-points staggered from Π in the r-direction. This is illustrated in Figure 3.3. This horizontal staggering is referred to as the Arakawa D-grid (Arakawa and Lamb 1977), and the vertical staggering as the Charney–Phillips grid (Arakawa and Konor 1996). If any of the other grids shown in Figure 3.3 is used to solve (3.17), the solution of (3.20) is likely to oscillate in space, leading to instability. While it may be possible to control this using filtering, significant inaccuracy will result.

We now need to consider the conditions for a numerical solution of (3.17) to stay close to a solution of (2.11). In Section 2.2 we showed that this required the solution of (3.12) to stay close to that of (3.14). Equation (3.13) showed that this depends on the condition $\omega^2 \delta t^2 \gg 1$. When the spatial discretization is considered, the eigenvalue ω^2 of $\mathbf{L}$ will be approximated by an eigenvalue ω_h; so the condition becomes $\omega_h^2 \delta t^2 \gg 1$. In the quasi-geostrophic case, this can be analysed because $\mathbf{L}$ is a constant-coefficient operator. Such analyses are given in Arakawa and Lamb (1977) and Durran (1998) for various spatial discretizations. If the horizontal grids B or D

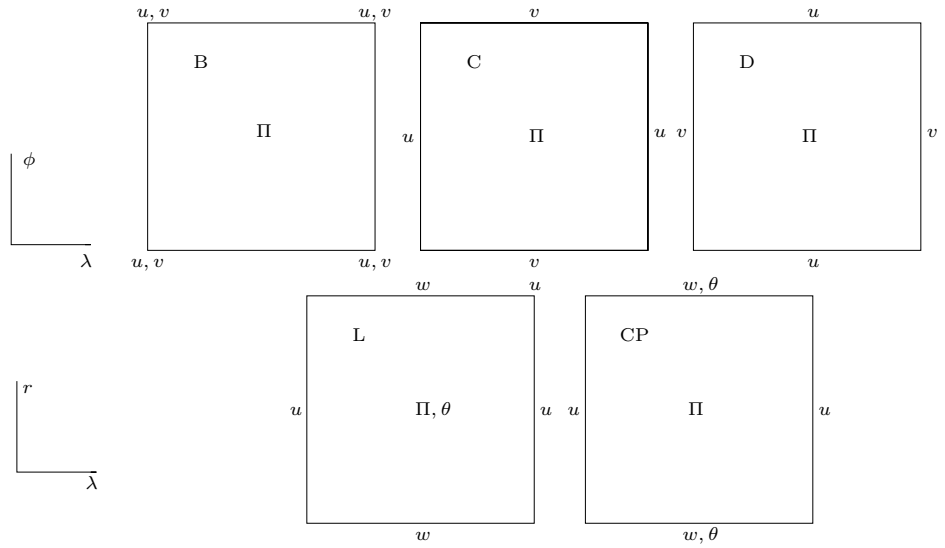

Figure 3.3. *Top line*: Various horizontal staggerings of variables for solving (3.17). These are referred to as Arakawa grids B, C and D. *Bottom line*: Two vertical staggerings, the Lorenz (L) and Charney–Phillips (CP).

shown in Figure 3.3 are used, the smallest-scale inertia-gravity waves become stationary so that $\omega_h = 0$, and the condition for (3.17) to be accurate will be violated. Only grid C gives satisfactory behaviour. In the vertical, the required condition is satisfied if the Charney–Phillips grid is used (Thuburn 2006).

There is thus a clash between the requirements on horizontal staggering of variables between the need to ensure that the discrete $\mathbf{L}$ is large when it should be, and the need to obtain stable solutions of (3.20). The Charney–Phillips grid, however, is optimal in the vertical on both counts. This clash indicates that the use of semi-implicit time integration is not sufficient to ensure accurate treatment of the limit solution, but spatial differencing issues must also be considered. Various methods have been used to address the issue. In Lin and Rood (1997) the C- and D-grids are both used for the parts of the calculation for which they are suited. However, interpolation between the grids at some point is unavoidable. A compromise horizontal arrangement called the B-grid is shown in Figure 3.3. Analysis, such as in Bryan (1989), shows that the B-grid is superior when the horizontal grid-length of the model is greater than the Rossby radius, but the C-grid is superior when the grid-length is less than the Rossby radius. In the atmosphere the internal Rossby radius for the most energetic parts of the flow is more like 1000 km, and is well-resolved. Thus the C-grid is more appropriate. The importance of preventing incorrectly small values of ω_h

is greatest in semi-implicit schemes, where the left-hand side of (3.12) has to be inverted. In explicit schemes, the problems can be managed by using filtering procedures. Thus the original Met Office Unified Model (Cullen and Davies 1990) used split-explicit finite differencing on the B-grid in the horizontal and the Lorenz grid in the vertical, while the current version (Davies *et al.* 2005) uses semi-implicit time integration on a C-grid in the horizontal and the Charney–Phillips grid in the vertical.

The difficulties in establishing an optimal horizontal representation can be resolved by using vorticity and divergence, rather than u and v, as prognostic variables. This is expensive in finite-difference methods, because of the need to solve elliptic equations to recover the velocity components. However it is a natural choice when using a spectral horizontal representation, which is very popular for global models. The elliptic problems then become trivial in spectral space. Under the assumptions made in this section, the operator $\mathbf{L}$ is separable in the horizontal and vertical. It has coefficients that depend only on the vertical. If equation (3.12) is discretized in the vertical, $\mathbf{L}$ becomes a matrix. Transforming variables to diagonalize this matrix means that the solution procedure can be decoupled into a set of constant-coefficient two-dimensional problems for each vertical mode: see Durran (1998, p. 387). A spectral model is very well-suited to solving these. The same transformation will reduce (3.20) to a set of two-dimensional problems. In practice, the scale-dependence of ω shown in (2.15) means that $\mathbf{L}$ has a large condition number. Standard iterative solvers will find it difficult to invert $\mathbf{L}$ accurately. The method of transforming the variables, followed by a solution in spectral space, will invert it to machine precision, which may be an important advantage. It should be noted that this method can be used to solve semi-implicit discretizations of (2.1): it does not depend on the approximations made in deriving (3.17). The spectral method is thus well-suited to maintaining the accuracy of (3.17) and to solving (3.20).

The advantages of spectral methods discussed above have to be balanced against the computational overhead of the transforms between grid-point and spectral space. They also maintain quadratic conservation properties, such as those discussed by Sadourny (1975). However, they are not well-suited to maintaining the bounds on Q, since there is no way of enforcing a monotonicity property. The combination of a spectral representation with semi-Lagrangian advection, which can be made to enforce monotonicity, has therefore become popular, as in the ECMWF model (Ritchie *et al.* 1995).

3.4. Large-scale flows

For this purpose we consider large-scale as a horizontal scale large relative to the Rossby radius, so that $\mathrm{Ro} < \mathrm{Fr}$, and we also assume Ro to be small. It is then appropriate to make the shallow atmosphere and hydrostatic assump-

tions, as in (2.11), and the geostrophic approximation, but the assumptions that the static stability and Coriolis parameter are close to reference values are not made. In particular, this means that the geostrophic wind is not non-divergent, so it is better to work directly from (2.11) rather than introduce the vorticity and divergence equations (3.1). The semi-geostrophic equations introduced by Hoskins (1975) are then appropriate, but note that f is not assumed constant. The system is

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}_g}{\mathrm{D}t} + (-fv, fu) + C_p\theta\nabla_r\Pi' &= 0, \\
\frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p\Pi}(S_h + LP), \\
\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= 0, \\
\frac{\mathrm{D}q}{\mathrm{D}t} &= S_q - P, \\
p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta, \\
\frac{C_p\theta}{a\cos\phi}\frac{\partial}{\partial\lambda}\Pi' - fv_g &= 0, \\
\frac{C_p\theta}{a}\frac{\partial}{\partial\phi}\Pi' + fu_g &= 0, \\
C_p\theta\frac{\partial\Pi'}{\partial r} - g\frac{\theta'}{\theta_0} &= 0.
\end{aligned} \tag{3.23}$$

In these equations, $\mathbf{u}_g = (u_g, v_g)$ is the geostrophic wind, defined by the sixth and seventh equations. Comparing (3.23) with (2.11) shows that the momentum has been approximated by its geostrophic value, but no other approximations have been made. In particular, the trajectory is not approximated, so that $\frac{\mathrm{D}}{\mathrm{D}t} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla$. It can be shown that the resulting equations conserve energy in the absence of forcing terms (Cullen, Norbury, Purser and Shutts 1987). The conserved energy is

$$E = \int \rho\Big(\tfrac{1}{2}(u_g^2 + v_g^2) + C_vT + gr\Big)a^2\cos\phi\,\mathrm{d}r\,\mathrm{d}\lambda\,\mathrm{d}r. \tag{3.24}$$

This differs from the energy (2.2) conserved by equations (2.1) by the replacement of the kinetic energy by its geostrophic value.

Most analyses of (3.23) have been carried out for constant f, which allows the equations to be solved by the geostrophic coordinate transformation, Hoskins (1975). However, this assumption is inappropriate for large-scale flow. An analysis of the spherical case is given by Cullen, Douglas, Roulstone and Sewell (2005). A review of the mathematical theory of these equations is given by Cullen (2006).

Equations (3.23) differ in an important respect from the nonlinear balance equations considered in Section 3.2. They contain two different velocity fields, representing the trajectory and the momentum respectively. This looks strange as a direct approximation to (2.11). However, it arises naturally if the equations are considered as Lagrangian averages of (2.11). The analysis of Andrews and McIntyre (1978) shows that the trajectory should represent the Lagrangian-averaged velocity, but the momentum-like quantity is a 'pseudo-momentum' which includes the effects of the waves excluded by the averaging. Such equations arise naturally when deriving asymptotic limit equations from Hamilton's principle. A systematic review is given in Holm, Marsden and Ratiu (2002). This idea has been exploited in sub-grid models, with the aim of modelling unresolved motions without introducing energy dissipation. Examples are the Gent–McWilliams parametrization of oceanic eddies (Gent and McWilliams 1996), and the 'alpha' model of turbulence (Foias, Holm and Titi 2001). Equations (3.23) can be interpreted as a Lagrangian average of (2.11) under the assumption that the pseudo-momentum is geostrophic. This is not the same as assuming that the momentum is geostrophic, but still assumes rotation-dominated flow.

The structure of these equations can be understood by rewriting them in the form used by Schubert (1985). This gives

$$\mathbf{Q}\begin{pmatrix} u \\ v \\ w \end{pmatrix} + C_p\theta\frac{\partial}{\partial t}\nabla\Pi' = \begin{pmatrix} f^2u_g \\ f^2v_g \\ \frac{g}{C_p\Pi\theta}(S_h + LP) \end{pmatrix}, \tag{3.25}$$

$$\mathbf{Q} = \begin{pmatrix} f^2 + \frac{f\theta}{a\cos\phi}\frac{\partial}{\partial\lambda}\left(\frac{v_g}{\theta}\right) + \frac{fu_g\tan\phi}{a} & \frac{f\theta}{a}\frac{\partial}{\partial\phi}\left(\frac{v_g}{\theta}\right) & f\theta\frac{\partial}{\partial r}\left(\frac{v_g}{\theta}\right) \\ -\frac{f\theta}{a\cos\phi}\frac{\partial}{\partial\lambda}\left(\frac{u_g}{\theta}\right) + \frac{fv_g\tan\phi}{a} & f^2 - \frac{f\theta}{a}\frac{\partial}{\partial\phi}\left(\frac{u_g}{\theta}\right) & -f\theta\frac{\partial}{\partial r}\left(\frac{u_g}{\theta}\right) \\ \frac{g}{a\theta\cos\phi}\frac{\partial\theta}{\partial\lambda} & \frac{g}{a\theta}\frac{\partial\theta}{\partial\phi} & \frac{g}{\theta}\frac{\partial\theta}{\partial r} \end{pmatrix}.$$

Equation (3.25) can be rewritten as an elliptic equation for $\partial\Pi/\partial t$ by using the last three equations of (3.23) to give

$$\frac{1}{1-\gamma}\frac{\rho}{\Pi}\frac{\partial\Pi}{\partial t} + \frac{\rho C_p\theta}{g}\frac{\partial}{\partial r}\left(\frac{\partial\Pi}{\partial t}\right) + \nabla\cdot\left[C_p\rho\theta\mathbf{Q}^{-1}\nabla\left(\frac{\partial\Pi}{\partial t}\right)\right] = \nabla\cdot\rho\mathbf{Q}^{-1}\begin{pmatrix} f^2u_g \\ f^2v_g \\ \frac{g}{\theta}(S_h + LP) \end{pmatrix}. \tag{3.26}$$

This is an elliptic equation for $\partial\Pi/\partial t$ if $\mathbf{Q}$ is positive definite. Since $\mathbf{Q}$ is purely a function of Π, this is a constraint on the pressure field. It is shown by Shutts and Cullen (1987) that this corresponds to *symmetric stability*, that is, stability of the resulting flow to parcel displacements, neglecting perturbation pressures arising from the displacement. They do this

by showing that a geostrophic and hydrostatic state with $\mathbf{Q}$ positive definite corresponds to a minimum energy state with respect to such parcel displacements. The semi-geostrophic approximation is only relevant for such flows, since flows which are unstable in this sense will evolve on a faster time-scale than f^{-1}. The manifestation of this in the horizontal is inertial stability, and in the vertical is static stability. Shutts and Cullen also show that the neglect of perturbation pressures is justified in the large-scale regime where semi-geostrophic theory is appropriate.

If (3.26) can be solved, the trajectory $\mathbf{u}$ can be deduced from (3.25). However, it is not obvious that the positive-definiteness of $\mathbf{Q}$ can be maintained during the time evolution. In order to study this, following Cullen and Feldman (2006), we introduce the Lagrangian map $F(t, \mathbf{x})$, where $\mathbf{x} = (\lambda, \phi, r)$ is a shorthand for the three-dimensional space coordinates. This map takes initial parcel positions to positions at time t, so that in particular $F(0, \mathbf{x}) = \mathbf{x}$. The map is assumed to conserve mass. This means that the Jacobian of $F(t, \mathbf{x})$ is $\rho(t, \mathbf{x})/\rho(0, \mathbf{x})$. We use the notation $F(t, \mathbf{x})\#\rho(0, \mathbf{x}) = \rho(t, \mathbf{x})$ as shorthand for this property. Now write equations (3.23), omitting the moisture equation and forcing terms, in the Lagrangian form

$$\begin{aligned} \partial_t Z(t, \mathbf{x}) - f\partial_t F(t, \mathbf{x}) &= fJZ, \\ F(t, \mathbf{x})\#\rho(0, \mathbf{x}) &= \rho(t, \mathbf{x}), \\ p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta, \\ Z(0, \mathbf{x}) &= f^{-1}C_p\theta^0\nabla\Pi^0(\mathbf{x}). \end{aligned} \tag{3.27}$$

Here, $Z(t, \mathbf{x}) = f^{-1}C_p\theta\nabla\Pi'(t, F(t, \mathbf{x}))$, the first equation of (3.27) corresponds to the first two equations of (3.23) and the second equation corresponds to the third equation of (3.23).

We now use this formulation to define an energy minimizer with respect to infinitesimal parcel displacements, assumed to take place over a virtual time δ. Given a state $\tilde{\Pi}(0, \mathbf{x}), \tilde{u}(0, \mathbf{x}), \tilde{v}(0, \mathbf{x}), \tilde{\theta}(0, \mathbf{x})$, with $\tilde{\rho}(0, \mathbf{x}), \tilde{T}(0, \mathbf{x})$ calculated from $\tilde{\Pi}(0, \mathbf{x})$ using the equation of state and the thermodynamic relations, define the energy $\tilde{E}$ by

$$\tilde{E} = \int \tilde{\rho}\left(\frac{1}{2}(\tilde{u}^2 + \tilde{v}^2) + C_v\tilde{T} + gr\right) \mathrm{d}\mathbf{x}. \tag{3.28}$$

Define the displacement by the Lagrangian map $F(\delta, \mathbf{x})$. The assumption of no perturbation pressure due to the displacement means that

$$\begin{aligned} Z(\delta, F(\delta, \mathbf{x})) - Z(0, F(0, \mathbf{x})) &= f(F(\delta, \mathbf{x}) - F(0, \mathbf{x})), \\ F(\delta, \mathbf{x})\#\tilde{\rho}(0, \mathbf{x}) &= \tilde{\rho}(\delta, \mathbf{x}), \\ Z(0, \mathbf{x}) &= (\tilde{v}, -\tilde{u}). \end{aligned} \tag{3.29}$$

We calculate the change to the energy by substituting $\tilde{\rho}(\delta, \mathbf{x}), \tilde{u}(\delta, \mathbf{x}), \tilde{v}(\delta, \mathbf{x})$

and $\tilde{T}(\delta, \mathbf{x})$ into (3.28). Then Theorem 4.1 of Cullen (2006) states that $\tilde{E}$ is stationary with respect to these variations if $Z = f^{-1}C_p\tilde{\theta}\nabla\tilde{\Pi}$. Thus a solution of (3.27) corresponds to a stationary energy state. The requirement that the solution is a minimizer, rather than just stationary, comes from the physical consistency discussed above. The mathematical theory of the semi-geostrophic equations shows that it is always possible to find an energy minimizer, and thus a physically consistent solution of the equations. Other possible solutions are ignored. The effect is that it is always possible to find a time evolution such that $\mathbf{Q}$ remains positive definite, given initial data such that $\mathbf{Q}$ is positive definite. One result of this is that there can be no horizontal pressure gradients along the equator, as discussed by Cullen *et al.* (2005). Details of the theorems are reviewed in Cullen (2006).

The situation is thus different from the nonlinear balance equations, (3.1), (3.2), (2.10), (2.12), (3.5) and (3.6). These also reduce to a variable-coefficient equation but whose ellipticity cannot be guaranteed in the time evolution. There are no solvability issues with the quasi-geostrophic equations because they reduce to solving a constant-coefficient elliptic equation, (3.20). However, no system of limit equations valid on large scales can be reduced to a constant-coefficient elliptic problem.

Another feature of the solutions of (3.23) is that they can be discontinuous. This does not conflict with the idea that the equations represent a Lagrangian average. There is no reason why air parcels with different physical properties should not come close together in a (near-)discontinuity. An example is discussed in Section 4.3, where the ability to capture such a discontinuity in a conventional numerical method is explored.

The condition for positive-definiteness of $\mathbf{Q}$ is harder to maintain in the vertical if moist effects are allowed for. This is because we can write the LP term that appears in the second equation of (3.23) as

$$LP = L\frac{\mathrm{d}q_{\mathrm{sat}}}{\mathrm{d}t} \simeq L\frac{\mathrm{d}q_{\mathrm{sat}}}{\mathrm{d}T}\frac{\mathrm{d}T}{\mathrm{d}t} \simeq L\Gamma w \tag{3.30}$$

where Γ now represents the moist adiabatic lapse rate (Houghton 2002, p. 21). Since this term is proportional to w, it should be transferred to the left-hand side of (3.25). The effect is to reduce the diagonal term in $\mathbf{Q}$ and make it easier to violate the condition for $\mathbf{Q}$ to be positive definite. Though no rigorous mathematical treatment has been given in this case, computations by Holt (1990) suggest that (3.25) can still be solved for general data. These computations demonstrate that the solutions may involve discontinuous mass transport when the condition $q \geq q_{\mathrm{sat}}$ is only satisfied in parts of the domain. This is still compatible with the interpretation of the equations as a Lagrangian average of (2.1). Representing this type of process in a sub-grid model to be used with equations (2.1) has been a major research challenge; see Smith (1997).

A priori estimates of the difference between solutions of (3.23) and (2.11) or (2.1) given in Cullen (2006) show that the difference is $O(\mathrm{Ro}_L(\mathrm{Ro}/\mathrm{Fr})^2)$ for small Ro_L. Ro_L is the Lagrangian Rossby number defined as the ratio of $\mathrm{D}/\mathrm{D}t$ and f. It is small if the rate of change of wind direction following a fluid particle is small compared with f. At latitude 45° that requires the direction of a trajectory to change by less than 45° in 24 hours. The effect of this is illustrated in Figure 3.4. Trajectories are plotted from a UM forecast for a very active period of weather in January 2005. Making allowance for the Mercator projection, the condition on the trajectory direction is only violated in the trajectory starting furthest west off the Californian coast, and in the lowest level trajectory as it crosses the Rockies. This suggests that the approximation will be valid most of the time for scales represented by this model, which has an averaging scale of about 120 km.

We now consider conventional finite-difference approximations to (3.26). In the case of small disturbances, $\mathbf{Q}$ is well approximated by the diagonal matrix

$$\begin{pmatrix} f^2 & 0 & 0 \\ 0 & f^2 & 0 \\ 0 & 0 & N^2 \end{pmatrix}.$$

Equation (3.26) is naturally solved at pressure points. Then u_g and v_g are naturally held at the positions occupied by u and v on the D-grid in Figure 3.3 and θ will be held as indicated on the Charney–Phillips grid in Figure 3.3; u, v and w are naturally held on the C-grid and Charney–Phillips grid. As in the quasi-geostrophic case, this arrangement is not ideal in the horizontal, because the accurate treatment of the right-hand side requires u_g at v_g-points and *vice versa*.

In general, the structure of equation (3.26) suggests that it could be incorporated in a semi-implicit discretization of (2.1), analogous to the derivation of equation (3.12). This is difficult to achieve, because of the presence of the two different velocity fields $\mathbf{u}$ and $\mathbf{u}_g$. In addition, it is found that in circumstances where (3.23) has discontinuous solutions, conventional finite-difference methods attempt to maintain smoothness by allowing $\mathbf{Q}$ not to be positive definite. This then results in computational instability (see Cullen (2006, Section 5.3.3)). In such cases, it is not even clear if the Eulerian form of advection used to write (3.25) makes sense.

To make further progress, we therefore need to base the discretization on the Lagrangian formulation of the equations (3.27), which means finding the Lagrangian map $F(t, \mathbf{x})$ by using all the constraints in equations (3.27). This requires methods where advection is treated in a Lagrangian manner. When solving (2.11) in circumstances when (3.23) is accurate, we then seek a Lagrangian map F close to that which satisfies (3.27).

We first have to establish a condition on Π such that the matrix $\mathbf{Q}$ derived from it is positive definite. Equations (3.23) and the definition of the

reference state, (2.8), show that

$$C_p \nabla \Pi = \left(\frac{f v_g}{\theta}, -\frac{f u_g}{\theta}, -\frac{g}{\theta} \right). \tag{3.31}$$

Then it can be seen from (3.25) that, in the case of Cartesian geometry where f is constant, $\mathbf{Q}$ is the matrix of second derivatives of Π with all terms multiplied by θ and the additional terms f^2 on the diagonal. If variations

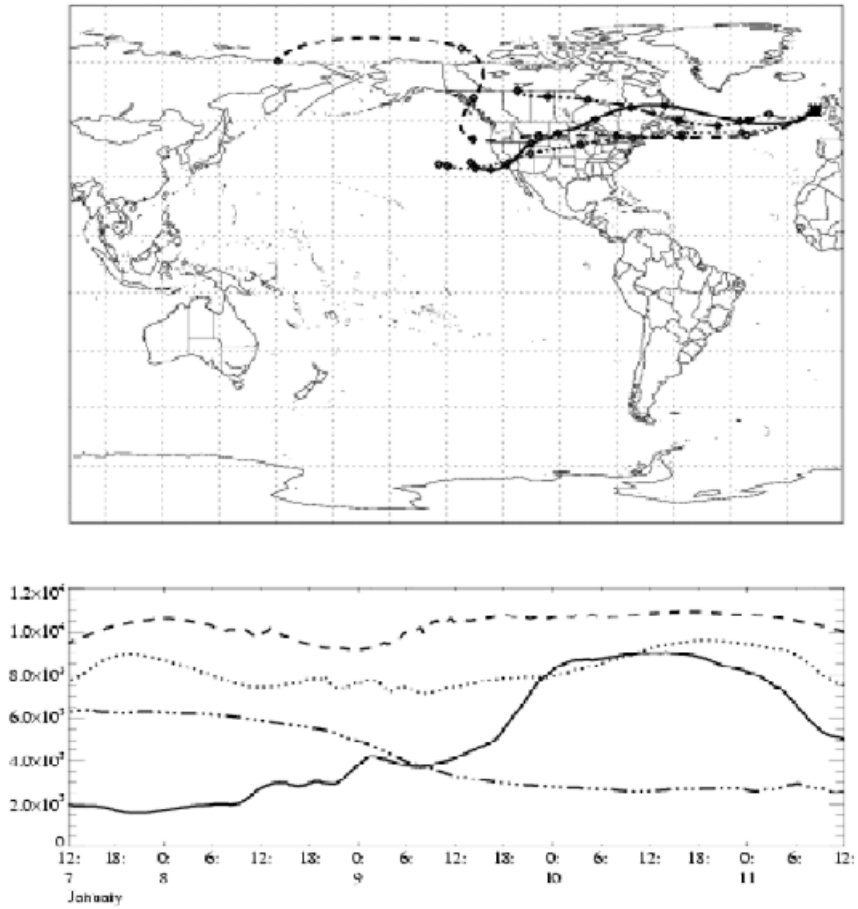

Figure 3.4. *Top*: Back trajectories from Mace Head for 12UTC on 11 January 2005. The total time covered is 96 hours, with points marked at 12 hour intervals. The trajectories were computed from 3-hourly data. *Bottom*: The vertical position of the trajectories (m). Source: Atmospheric Dispersion Group, Met Office.

of the multiplying factor θ are ignored, the condition for $\mathbf{Q}$ to be positive definite is that $C_p\Pi + \frac{1}{2}(x^2 + y^2)$ is convex, where x and y are Cartesian coordinates. When variations of θ are included, the equivalent condition is derived by Shutts and Cullen (1987). In spherical geometry, the condition is derived by Cullen *et al.* (2005), and is referred to as *involutivity*. In both these cases, it can be shown that the condition prevents oscillatory behaviour of Π, in the same way that a convex function cannot oscillate.

Another important feature of the solutions of (3.27) is that only integrals of fluid properties are well-defined. This allows the solutions to be singular. Physically it means that the behaviour of infinitesimally small volumes of the fluid is irrelevant. Thus we choose a function φ which is very smooth, and define a 'weak' solution of (3.27) on a time interval $(0, \tau)$ by requiring that, for any $\varphi \in C^\infty([0,\tau] \times \Omega : \mathcal{R}^3)$,

$$\int_{\Omega\times(0,\tau)} \left[Z(t,\mathbf{x}) \cdot \partial_t\varphi(t,\mathbf{x}) - fF(t,\mathbf{x})\partial_t\varphi(t,\mathbf{x}) + fJZ(t,\mathbf{x}) \cdot \varphi(t,\mathbf{x})\right] \mathrm{d}t\,\mathrm{d}\mathbf{x} + \int_\Omega f^{-1}\theta^0 \nabla\Pi^0(\mathbf{x}) \cdot \varphi(0,\mathbf{x})\,\mathrm{d}\mathbf{x} = 0. \tag{3.32}$$

Suppose that we are given initial data $\Pi^0(\mathbf{x})$, assumed to be an involutive bounded non-negative function defined on an open set $\Omega \subset \mathcal{R}^3$. Let ρ^0 and θ^0 be calculated from Π^0 by the hydrostatic relation. In the case where f is constant, Cullen and Feldman (2006) proved that we can find for any $t \in (0,\tau)$ a Lagrangian map $F_t = F(t,\mathbf{x}) : \Omega \to \Omega$ satisfying $F_t \# \rho^0 = \rho(t,\mathbf{x})$ and such that (3.32) is satisfied. This map also has an inverse F_t^* such that $F_t^* \circ F_t(\mathbf{x}) = \mathbf{x}$ for almost all $\mathbf{x}$ (*i.e.*, it may not exist for infinitesimal masses of fluid). This property means that the Lagrangian map F satisfies the 'flow' property that $F_{t_1+t_2} = F_{t_1} \circ F_{t_2}$. This is required for the solution to make physical sense, and not to depend on an arbitrary discretization of the time interval. This proof has not yet been extended rigorously to the case of variable f. Cullen *et al.* (2005) describe the formal arguments which show that the result is expected to hold in that case.

Now consider numerical methods for approximating the Lagrangian map. The most widely used Lagrangian-based methods are the semi-Lagrangian methods introduced by Robert (1982). A review of these is given by Staniforth and Côté (1991). They are used in both the UM and the ECMWF model, as well as many others. They have the advantage that the data are mapped onto an Eulerian grid at each time-step, thus allowing computations of terms other than advection to be carried out easily. They also have the advantage of being unconditionally stable for time integration unless the trajectories cross; see Durran (1998, Chapter 6). This was the original reason for their introduction. Enforcement of the condition $F(t,\mathbf{x})\#\rho(0,\mathbf{x}) = \rho(t,\mathbf{x})$ requires the discrete Lagrangian map to be

mass-preserving. This is the motivation for conservative semi-Lagrangian methods such as Zerroukat, Wood and Staniforth (2004). However, such methods are hard to use and have not yet been implemented in production models. They are discussed in more detail in Section 3.6. The proof that (3.32) can be solved depends on the fact that a Lagrangian map cannot create new values of the quantity to which it is applied, as in the advection of Q in the quasi-geostrophic case. In the present case the condition has to be applied separately to the momentum components and the potential temperature. This again favours the use of monotone interpolation schemes, as discussed in Section 3.3. However, the remapping to grid-points every time-step means that strictly monotone schemes will be too diffusive, and more selective schemes such as that of Zerroukat *et al.* (2005) are needed.

The trajectory is derived implicitly by the solution procedure. This suggests that an implicit calculation of the trajectory departure points is desirable. An iteration like (3.16) can be used for this purpose, as in Yeh *et al.* (2002) and Cullen (2001). In Cullen (2001), the increased accuracy of doing this is demonstrated. In White (2003) the dynamical consistency resulting from this choice is discussed. The application of this scheme to the UM is described in Diamantakis, Davies and Wood (2007). An example is shown in Figure 3.5. The top two panels use the operational version of the UM with the departure point calculation extrapolated in time to obtain second-order accuracy. They show that the vertical motion is greatly exaggerated in a run with a 30-second time-step as compared to a run with a 10-second time-step. If the departure-point calculation is iterated once in time, as in Cullen (2001), then the results shown in the bottom panel are obtained using a 40-second time-step. The results are close to that obtained with the operational scheme and a 10-second time-step, showing that the time iteration is beneficial and cost-effective.

The proof of existence of solutions to (3.23) depends critically on the involutivity of Π, which prevents oscillatory behaviour and thus gives considerable stability to the time evolution, as demonstrated in Cullen (2002*a*). This is why large-scale disturbances persist. However, on smaller scales, where semi-geostrophic theory is not applicable, there is no reason why solutions of (2.1) or (2.11) should obey this condition and oscillatory and unstable behaviour is readily observed. Section 4.4 illustrates such a case, where the instability is triggered by orography. The direct solution of (2.1) is quite successful, though it has clearly not converged. This is, however, a rather simple case. In real cases, satisfactory solutions to (2.1) in regions where $\mathbf{Q}$ is not positive definite can usually only be obtained by using a horizontal grid-length of no more than 1 km. Operational averaging scales are normally much greater than this, except for local models. In order to get satisfactory performance with coarser averaging scales, the sub-grid model has to enforce the involutivity of Π in order to prevent unrealistic unstable

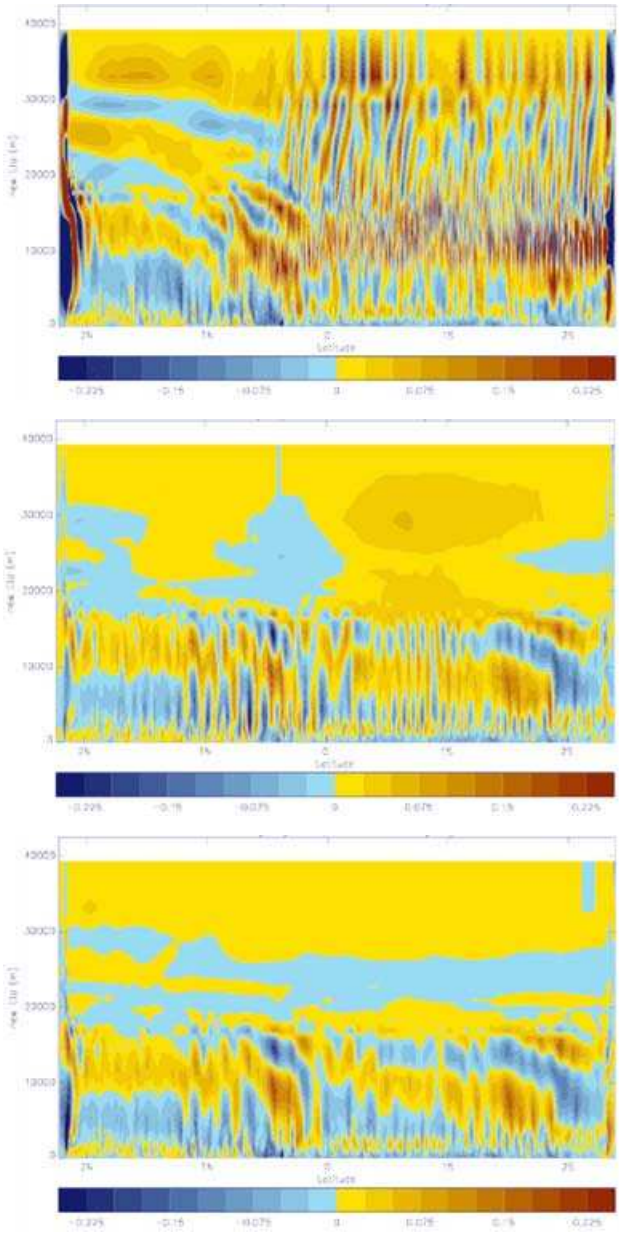

Figure 3.5. Cross-sections of vertical velocity ($\mathrm{m\,s^{-1}}$) from UM integration over the Alpine region plotted against height (m) and latitude relative to the centre of the domain (degrees). *Top*: Explicit departure-point calculation, time-step 30 s, decentering parameter $\alpha = 0.7$. *Middle*: As above with time-step 10 s and $\alpha = 0.6$. *Bottom*: Iterated departure-point calculation, time-step 40 s, $\alpha = 0.6$.

circulations developing. Since typical time-averaging scales in the atmosphere are much less than f^{-1}, but greater than N^{-1}, it is usual to enforce only the condition that $\partial\theta/\partial r \geq 0$, modified for the effects of moisture as shown in (3.30). Recent collections of papers in this area are given in ECMWF (2002) and ECMWF (2005).

The long-term behaviour of solutions to (3.23) was studied in the shallow water case by Cullen (2002*a*). The cascades of vorticity to small scales that are a feature of classical two-dimensional turbulence, Leith and Kraichnan (1972), are inhibited, and disturbances persist for long times. The long-lived anomalous circulations observed in the extra-tropical atmosphere thus represent the natural dynamics of rotation-dominated flows. This is therefore the most appropriate regime in which to study the requirements for maintaining long-time accuracy in numerical solutions.

This behaviour means that, if flows on a scale larger than the Rossby radius are initially well-resolved, then they will stay well-resolved. Numerical methods such as those of Sadourny (1975) may then be beneficial in preserving long-term accuracy. In particular, it is tempting to suggest that schemes which conserve $\int Q^2$, where Q is the Ertel potential vorticity, would be appropriate. However, fronts can form in three-dimensional flow. As demonstrated in Section 4.3, these are associated with apparent potential vorticity sources in an Eulerian sense, even though potential vorticity is still conserved in a Lagrangian sense. These solutions could not develop if schemes like those of Sadourny (1975) were used. It is therefore a major research question how to preserve long-term accuracy in this regime.

3.5. Small-scale flows with $\varepsilon \ll 1$

In this case we assume that the aspect ratio may be O(1), but that the horizontal extent is sufficiently small for plane geometry to be used. The assumption that $\varepsilon \ll 1$ for such aspect ratios means that U/NH has to be small. The solvability condition for the nonlinear balance equations is not necessarily satisfied, as noted in Section 3.2. The difficulty is caused by the flow-dependent coefficients in the elliptic problem that has to be solved. If only small-scale flows are considered, the flow-dependent coefficients are not required. Consider the equation for the evolution of θ:

$$\frac{\partial\theta}{\partial t} + \mathbf{u}_r.\nabla_r\theta + w\frac{\partial\theta}{\partial r} = \frac{1}{C_p\Pi}(S_h + LP). \tag{3.33}$$

If the right-hand side terms of (3.33) are ignored, then it can be shown that the assumption $U/NH \ll 1$ means that $\nabla_r\theta \ll \frac{H}{L}\frac{\partial\theta}{\partial r}$. We can therefore assume a hydrostatic reference state $\theta_1(r)$ with a reference value N_0^2 of N^2 as in Section 3.3. Comparison of the terms in equation (3.33) then shows that $w/\mathbf{u}_r \ll H/L$, so that the vertical advection can be neglected in the momentum equations of (2.11). Majda (2003, Chapter 6) shows that an

energetically consistent system can then be obtained by approximating the pressure gradient term and replacing the continuity equation by $\nabla_r \cdot \mathbf{u} = 0$. The resulting system is

$$\begin{aligned} \frac{\mathrm{D}_r \mathbf{u}_r}{\mathrm{D}t} + (-fv, fu) + C_p\theta_1 \nabla_r \Pi' &= 0, \\ C_p\theta \frac{\partial \Pi'}{\partial r} - g\frac{\theta'}{\theta_0} &= 0, \\ \nabla_r \cdot (\mathbf{u}) &= 0, \\ \frac{\mathrm{D}_r \theta}{\mathrm{D}t} + w\frac{\partial \theta_1}{\partial r} &= \frac{1}{C_p \Pi}(S_h + LP), \\ \Pi &= (p/p_{\mathrm{ref}})^{-R/C_p}, \\ p &= \rho R \theta \Pi. \end{aligned} \tag{3.34}$$

The first and third equations in (3.34) reduce to the equations for two-dimensional incompressible flow. The nonlinear balance relation, (3.6), reduces to

$$-2J(u, v) + \nabla_r \cdot C_p\theta_1 \nabla_r \Pi' - \nabla_r \cdot (fv, -fu) = 0. \tag{3.35}$$

These equations can be solved by integrating the first equation of (3.34) forward in time, with the pressure gradient terms determined implicitly by using (3.35) to enforce the incompressibility constraint, the third equation of (3.34). Equation (3.35) together with the hydrostatic relation (the second equation of (3.34)) can be used to determine Π and θ. It is then possible to derive w from the evolution equation for θ.

Solvability of these equations follows from the solvability of the equations for two-dimensional incompressible flow; *e.g.*, Chemin (2000). This depends on the fact that the vorticity ζ defined by (3.1) is bounded by its initial values, and that the elliptic equation (3.35) for Π' satisfies second derivative estimates analogous to (3.22). This ensures that the horizontal velocity fields stay smooth for smooth initial data. However, the vorticity typically cascades to small scale, so an initially well-resolved solution will not stay well-resolved. The remaining steps are explicit calculations.

A discrete version of the first requirement can be achieved in numerical models by using ζ as a prognostic variable, and advecting it with a monotonicity-preserving scheme. The alternative method, based on Sadourny (1975), of enforcing conservation of ζ^2 has the disadvantage of forcing the variance of ζ to be conserved on resolved scales, which is inappropriate in the presence of cascades of variance to small scales. The second requirement can be met if the elliptic equation (3.35) for Π' satisfies a discrete maximum principle. As in (3.19), this favours the use of the D-grid for the velocity components. The explicit calculation of θ from Π needs the

Charney–Phillips grid, as the discrete version of this calculation on the Lorenz grid is impossible because of the computational mode (Arakawa and Konor 1996).

The behaviour of this system is analysed in Majda (2003). The difficulty is that the first two equations of (3.34) evolve independently at each value of r. There is thus no control over the vertical scale of the solutions. In particular, values of $\partial\Pi/\partial r$ may become very large, leading to unrealistic values of θ and hence w. The problem is that the solutions violate the assumption $U/NH \ll 1$ as H is reduced. This is called 'zigzag' instability. In Majda (2003) it is shown that the problem can be resolved either by vertical viscosity, which cannot be physically justified in this strongly stratified case, or by rotation. As the vertical scale reduces, it may be that the condition $NH = fL$ becomes valid while the Froude number is still small. In that case the quasi-geostrophic equations derived in Section 3.3 apply, and the vertical coupling is restored.

As will be illustrated in Section 4.2, the assumption $\nabla_r \cdot \mathbf{u}_r = 0$ is very restrictive. A scheme with higher-order accuracy, but avoiding the solvability condition (3.7), involves iterating the solution procedure above. Thus, the calculated w is used via Richardson's equation (2.12) to generate a value of $\nabla_r \cdot \mathbf{u}_r$. This is then used along with ζ to calculate $\mathbf{u}$. In Ford *et al.* (2000) such an iterative procedure is developed in a rather more general context. Numerical solutions described by McIntyre and Norton (2000) show its effectiveness in shallow water integrations. These did not use sufficiently high resolution for (3.7) to be violated. The convergence of such an iteration in cases where (3.7) is violated is uncertain.

The solution procedure above is very attractive for numerical modelling, because it only involves solving two-dimensional rather than three-dimensional elliptic problems. However, the instability of this regime means that such methods will be of limited applicability.

3.6. Small-scale flows with $\varepsilon \gg 1$

In this case we assume that the aspect ratio may be O(1), but that the horizontal extent is sufficiently small for plane geometry to be used. We thus use Cartesian coordinates (x, y, z) and assume that f is a constant. Equations (2.27), with the moisture equation and forcing terms restored, become

$$\begin{aligned} \frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + f(-v, u, 0) + C_p\theta\nabla\Pi' - g\frac{\theta'}{\theta_0}\hat{\mathbf{r}} &= 0, \\ \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= 0, \\ \frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p\Pi}(S_h + LP), \end{aligned} \tag{3.36}$$

$$\frac{\mathrm{D}q}{\mathrm{D}t} = S_q - P,$$
$$p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} = \rho R\theta.$$

We showed in Section 2.3 that the characteristic velocity $U \ll c$ for all cases of interest. This means that equation (2.34) can be approximated by (2.35). Energetically consistent approximations to (3.36) that achieve this include the anelastic approximations of Ogura and Phillips (1962) and Lipps and Hemler (1982). In these approximations, the continuity equation is replaced by the constraint $\nabla \cdot (\rho_0 \mathbf{u}) = 0$, where ρ_0 is given by either the reference state (2.8) or the reference state used in (3.34). This reference state has also to be used in the momentum equations in order to retain energetic consistency. Issues with the effect of this approximation on convergence to the anelastic limit are discussed below. It is assumed that the scale is still large enough for the real viscous and conductive terms to be neglected.

With either choice of reference state, these approximations reduce equations (3.36) to

$$\begin{aligned}
\frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} + f(-v, u, 0) + C_p\nabla(\theta_0\Pi') - g\frac{\theta'}{\theta_0}\hat{\mathbf{r}} &= 0,\\
\nabla \cdot (\rho_0 \mathbf{u}) &= 0,\\
\frac{\mathrm{D}\theta}{\mathrm{D}t} &= \frac{1}{C_p\Pi}(S_h + LP),\\
\frac{\mathrm{D}q}{\mathrm{D}t} &= S_q - P,\\
p_{\mathrm{ref}}\Pi^{\frac{1}{\gamma-1}} &= \rho R\theta.
\end{aligned} \tag{3.37}$$

The conserved energy density is

$$\rho_0\left(\frac{1}{2}\mathbf{u}^2 + gr + C_p\Pi_0\theta\right). \tag{3.38}$$

Since, in effect, this approximation makes the speed of sound infinite, it is not valid on large scales where the finite sound speed matters; see Davies, Staniforth, Wood and Thuburn (2003).

In the case where $f = g = 0$, (3.37) shows that $\nabla \cdot \theta_0\nabla\Pi$ is of order $|\mathbf{u}|^2$. Thus the variations in Π will be of the same order as the variations in θ, with an additional variation of $\mathrm{O}(|\mathbf{u}|^2)$. The equation of state then shows that ρ will have the same order of magnitude of variations as Π. In the special case of uniform θ, this means that the error in using the reference state value of ρ is of order (Mach number)^2. This is the case where convergence of compressible flow to incompressible flow has been proved (Majda 1984). There will also be an error related to the departures

in θ from its assumed reference value. In general, this error will increase with the size of the variations in θ in the initial data. It will not be related to the Mach number. The error will be reduced in strongly stratified flows where θ remains close to a reference state, even though that reference state has large variations in θ. This was the case treated in Section 3.5 where the dynamics is hydrostatic and can be described by (3.34).

We therefore consider the case where $U/(NH)$ is large, so that the stratification is not dominant. This will be particularly true of the circulations in convective clouds, for instance. We avoid the difficulties caused by the nonlinearity by using a Lagrangian form of the equations, as in Section 3.4. Thus write $\mathbf{x} = (x, y, z)$ and define a Lagrangian map $F(t, \mathbf{x})$ with $F(0, \mathbf{x}) = \mathbf{x}$. The conserved mass density of a fluid element is now $\rho_0(z)$. The continuity equation is expressed by the condition $F(t, \mathbf{x})\#\rho_0 = \rho_0$. Define $Z(t, \mathbf{x}) = (-fv(t, F(t, \mathbf{x})), fu(t, F(t, \mathbf{x})), -g\theta'(t, F(t, \mathbf{x}))/\theta_0)$ and $Y(t, \mathbf{x}) = C_p\nabla(\theta_0\Pi)(t, F(t, \mathbf{x}))$. The Lagrangian form of (3.37) and the initial conditions is then

$$\begin{aligned}
\partial_t F(t, \mathbf{x}) &= \mathbf{u}(t, F(t, \mathbf{x})), \\
\partial_t \mathbf{u}(t, F(t, \mathbf{x})) + Z(t, \mathbf{x}) + Y(t, \mathbf{x}) &= 0, \\
F(t, \mathbf{x})\#\rho_0 &= \rho_0, \\
\partial_t \theta(t, F(t, \mathbf{x})) &= \frac{1}{C_p\Pi(t, F(t, \mathbf{x}))}(S_h + LP), \\
\partial_t q(t, F(t, \mathbf{x})) &= S_q - P, \\
\mathbf{u}(0, \mathbf{x}) &= \mathbf{u}^0(\mathbf{x}), \\
\theta(0, \mathbf{x}) &= \theta^0(\mathbf{x}), \\
q(0, \mathbf{x}) &= q^0(\mathbf{x}), \\
F(0, \mathbf{x}) &= \mathbf{x}.
\end{aligned} \tag{3.39}$$

As in Section 3.4, the omission of viscosity and thermal conductivity means that there is no reason why the solutions should be smooth. Air parcels with different properties can 'tangle' as much as desired. We therefore again seek weak solutions, where the behaviour of infinitesimally small fluid volumes is ignored, and work only with integrals. We thus multiply (3.39) by smooth test functions φ and ϖ and seek solutions in the sense that, for any $\varphi \in C^\infty([0, \tau] \times \Omega : \mathcal{R}^3)$, $\varpi \in C^\infty([0, \tau] \times \Omega : \mathcal{R})$:

$$\begin{aligned}
\int_{\Omega\times(0,\tau)} \big[F(t, \mathbf{x}) \cdot \partial_t\varphi(t, \mathbf{x}) + \mathbf{u}(t, F(t, \mathbf{x})) \cdot \varphi(t, \mathbf{x})\big]\, \mathrm{d}t\, \mathrm{d}\mathbf{x} \\
+ \int_\Omega \mathbf{x} \cdot \varphi(0, \mathbf{x})\, \mathrm{d}\mathbf{x} = 0,
\end{aligned} \tag{3.40}$$

$$\int_{\Omega\times(0,\tau)} \big[\mathbf{u}(t,F(t,\mathbf{x}))\cdot\partial_t\varphi(t,\mathbf{x}) + Z(t,\mathbf{x})\cdot\varphi(t,\mathbf{x}) + Y(t,\mathbf{x})\cdot\varphi(t,\mathbf{x})\big]\,\mathrm{d}t\,\mathrm{d}\mathbf{x} + \int_\Omega \mathbf{u}^0(\mathbf{x})\cdot\varphi(0,\mathbf{x})\,\mathrm{d}\mathbf{x} = 0,$$

$$\int_{\Omega\times(0,\tau)} \left[\theta(t,F(t,\mathbf{x}))\partial_t\varpi(t,\mathbf{x})\frac{1}{C_p\Pi}(S_h + LP)\varpi(t,\mathbf{x})\right]\mathrm{d}t\,\mathrm{d}\mathbf{x} + \int_\Omega \theta^0(\mathbf{x})\varpi(0,\mathbf{x})\,\mathrm{d}\mathbf{x} = 0,$$

$$\int_{\Omega\times(0,\tau)} \big[q(t,F(t,\mathbf{x}))\partial_t\varpi(t,\mathbf{x}) - (S_q - P)\varpi(t,\mathbf{x})\big]\,\mathrm{d}t\,\mathrm{d}\mathbf{x} + \int_\Omega q^0(\mathbf{x})\varpi(0,\mathbf{x})\,\mathrm{d}\mathbf{x} = 0.$$

The existence of such solutions is an open question. However, a time-discretized version of (3.37) can be solved by a method based on Brenier (1991). It is described in Cullen (2002*b*, Section 3.3). Given data at $t = 0$, construct a 'solution' at time δt by the following procedure. The forcing terms are omitted. Define a first guess Lagrangian map F^+ by

$$\begin{aligned}
x_c(\mathbf{x}) &= x + \frac{v^0(\mathbf{x})}{f},\\
y_c(\mathbf{x}) &= y - \frac{u^0(\mathbf{x})}{f},\\
F^+(\delta t,\mathbf{x}) &= \Big(x_c + (x-x_c)\cos(f\delta t) + (y-y_c)\sin(f\delta t),\\
&\qquad y_c + (y-y_c)\cos(f\delta t) - (x-x_c)\sin(f\delta t),\\
&\qquad z + w^0(\mathbf{x})\delta t + \frac{1}{2}g\frac{\theta'}{\theta_0}\delta t^2\Big),\\
u(\delta t,\mathbf{x}) &= u^0(\mathbf{x})\cos(f\delta t) + v^0(\mathbf{x})\sin(f\delta t),\\
v(\delta t,\mathbf{x}) &= v^0(\mathbf{x})\cos(f\delta t) - u^0(\mathbf{x})\sin(f\delta t),\\
w(\delta t,\mathbf{x}) &= w^0(\mathbf{x}) + g\frac{\theta'(\mathbf{x})}{\theta_0}\delta t,\\
\theta'(\delta t,\mathbf{x}) &= \theta^0(\mathbf{x}),\\
q(\delta t,\mathbf{x}) &= q^0(\mathbf{x}).
\end{aligned} \tag{3.41}$$

The forcing terms can easily be included if they can be expressed in terms of $\mathbf{u}, \theta$ and q. Otherwise they have to be iterated, and a rigorous treatment may not be possible.

We now correct this first guess to enforce the condition that $F^+ \# \rho_0 = \rho_0$. This is done by using the *polar factorization* theorem of Brenier (1991). This states that under suitable non-degeneracy conditions, we can uniquely write

$$F^+(\delta t, \mathbf{x}) = F(\delta t, \mathbf{x}) \circ \nabla \Upsilon(\mathbf{x}), \tag{3.42}$$

where $F(\delta t, \cdot) \# \rho_0 = \rho_0$ and Υ is convex. If the term $C_p \nabla(\theta_0 \Pi)$ had been included in (3.41), and it took constant values on trajectories, F^+ would have been incremented by $\frac{1}{2}\delta t^2 C_p \nabla(\theta_0 \Pi)$. We can thus identify the projection $\nabla \Upsilon$ defined in (3.42) with the mapping $I + \frac{1}{2}\delta t^2 C_p \nabla(\theta_0 \Pi)$. Note that the presence of the identity map, which can be written as $\nabla(\frac{1}{2}\mathbf{x}^2)$, will ensure convexity of Υ as $\delta t \to 0$, whatever (bounded) value of Π is required for the projection.

In their proof that the viscous hydrostatic equations could be solved, Cao and Titi (2005) showed that equations (2.11) could be reduced to the 'pressure-less' equations which are solved to define the first guess trajectory (3.41). In these equations trajectories will usually intersect. They then used the theory of the three-dimensional Burgers equation to show that the equations could be solved by including viscosity. In the present case, we have shown that the pressure gradient term can be used to construct a trajectory that satisfies the incompressibility condition over a finite time interval. No viscosity is required. This is consistent with the analyses in Lions, Temam and Wang (1992*b*) which show that the non-hydrostatic equations (3.37) have more regular solutions than equations (2.11) when viscosity is included in both.

In the present case, it is possible that the equations can be solved without viscosity. That would require proving the existence of a sufficiently regular limit as $\delta t \to 0$. This has not yet been achieved, and existence of solutions to (3.37) remains open but possible, given that computations have so far failed to provide convincing evidence that solutions break down in finite time (Kerr 1993). This issue is discussed at length in Majda and Bertozzi (2002). If the limit of this solution procedure does not exist, it implies that trajectories lose their identity and energy dissipation is inevitable.

This solution procedure depends critically on the anelastic approximation. The projection method is analogous to that used to prove existence of solutions to the incompressible semi-geostrophic equations by Benamou and Brenier (1998). The extension of their result to the compressible case by Cullen and Maroofi (2003) suggests that the solution procedure for the anelastic equations can be generalized to at least slightly compressible flows. This would be sufficient for the atmospheric case.

In a grid-based model, the natural discrete representation of this solution procedure is the semi-Lagrangian method discussed in Section 3.4. The remapping to grid-points every time-step in this method will inevitable involve loss of identity of the trajectories, so it reasonable to infer that such a

method will converge as $\delta t \to 0$, though it is not clear whether convergence will be to an inviscid energy-conserving solution or to a viscous, dissipative solution. Domaradzki, Xiao and, Smolarkiewicz (2003) illustrate how a limiting value of the dissipation can be estimated from numerical solutions. Nonlinear stability will require enforcing the constraint that transport cannot create new values of any of the primary variables (u, v, w, θ', q). This suggests the use of monotonicity-preserving schemes for all the primary variables. If this is done, the method will remain stable at a given resolution irrespective of any energy cascade to smaller scales. Thus the numerical method incorporates an implicit turbulence model, as discussed by Fureby and Grinstein (2002), but based on Lagrangian averaging. No other dissipation need be included. In the ECMWF model, monotonicity-preserving schemes are used for all variables. In the UM they are used for scalars. In the UM no other dissipation needs to be included in the spatial representation, except at the poles where the grid is highly distorted. In the ECMWF model only small amounts of extra dissipation are used. Specialized subgrid models are used in regions where there is organized small-scale flow such as the atmospheric boundary-layer and regions of convective cloud.

The other issue is the discrete version of the projection algorithm (3.42). This is needed to enforce the continuity equation. In order to maintain consistency with (3.39), the continuity equation should be written in fully Lagrangian form as

$$\frac{\rho_0(z_d)\partial(x_d, y_d, z_d)}{\rho_0(z)\partial(x, y, z)} = 1, \tag{3.43}$$

where $\mathbf{x}_d$ denotes the departure point associated with arrival point $\mathbf{x}$. Equation (3.43) is most naturally calculated with all components of $\mathbf{u}$ held together at the vertices of grid volumes, giving a three-dimensional version of the B-grid shown in Figure 3.4. This leads to a 27-point stencil for the Poisson equation for Π which will be difficult to invert. This problem does not arise if the conservative scheme is written as a product of one-dimensional schemes, as in Lin and Rood (1997) or Zerroukat *et al.* (2004). If an Eulerian form of the constraint is used, it takes the form of a discretization of the equation $\nabla \cdot (\rho_0 \mathbf{u}) = 0$ from (3.37). This is naturally expressed on the C-grid shown in Figure 3.4, with Π calculated by solving a discrete Poisson equation on a 7-point stencil. This satisfies a discrete maximum principle, as noted in Section 3.3, and is the basis of the scheme used in the UM (Davies *et al.* 2005).

3.7. Summary

In this section, we have demonstrated how numerical methods can be chosen optimally for different asymptotic regimes. The examples used were primarily nonlinear regimes, so emphasis is placed on maintaining Lagrangian

conservation properties by using spatial discretizations that preserve monotonicity. Since in most cases the asymptotic limit solutions are derived by solving equations with time derivatives removed, the use of selective time filtering is an important method of staying close to the asymptotic solutions. Where the limit equations are derived by removing Lagrangian time derivatives, as in Section 3.4, the time filtering needs to act in a Lagrangian sense.

The regimes surveyed here are not comprehensive. In some cases the behaviour is well described by linear dynamics. These cases have been extensively analysed in the literature (Arakawa and Lamb 1977, Durran 1998). The use of monotonicity-preserving advection schemes is not optimal when the dynamics is nearly linear and the data smooth. Long-term accuracy is then better preserved by methods such as spectral methods, which have no computational damping. This is another example of the dependence of the choice of optimal numerical methods on the regime being modelled.

4. Examples using the Met Office model

4.1. The numerical scheme in the Unified Model

The issues described in Section 3 are illustrated using various simplified configurations of the Met Office Unified Model (UM) described by Davies *et al.* (2005). We summarize the equations and discretization in this section.

The equations used are the compressible Euler equations (2.1). The momentum equation is solved in the form (2.8) and the thermodynamic equation in the form (2.5). Sub-grid modelling terms, represented as $\mathbf{S}_u$, which include turbulent diffusion are added to the horizontal momentum equations. The equations are formulated in the spherical polar coordinates (λ, ϕ, r) introduced in Section 2.1, with a lower boundary condition $\mathbf{u} = 0$ and $\partial\mathbf{u}/\partial r$ at $r = r_0(\lambda, \phi)$ and an upper boundary condition $w = 0$ at $r = r_T$. This assumes that the terms $\mathbf{S}_u$ contain higher-order terms near the surface consistent with a no-slip lower boundary condition, but that $\mathbf{S}_u$ does not contain second-order vertical derivatives near $r = r_T$. For convenience in applying the lower boundary condition, the radial coordinate r is replaced by a terrain-following vertical coordinate η, such that $\eta = 0$ at $r = r_0$ and $\eta = 1$ at $r = r_T$. The transformation from r to η is described by Davies *et al.* (2005). All variables in the equations, including r and w, are then considered as functions of (λ, ϕ, η). The terms which account for the effects of moisture are not considered in these tests and are thus omitted. For convenience, the resulting equations are summarized in component form:

$$\frac{\mathrm{D}}{\mathrm{D}t} \equiv \frac{\partial}{\partial t} + \frac{u}{r\cos\phi}\frac{\partial}{\partial\lambda} + \frac{v}{r}\frac{\partial}{\partial\phi} + \dot{\eta}\frac{\partial}{\partial\eta},$$

$$\dot{\eta}\frac{\partial r}{\partial\eta} = w - \frac{u}{r\cos\phi}\frac{\partial r}{\partial\lambda} - \frac{v}{r}\frac{\partial r}{\partial\phi},$$

$$
\begin{aligned}
&\frac{\mathrm{D}u}{\mathrm{D}t} - \frac{uv\tan\phi}{r} + \frac{uw}{r} - 2\Omega\sin\phi v + 2\Omega\cos\phi w \\
&\qquad + \frac{C_p\theta}{r\cos\phi}\left(\frac{\partial\Pi}{\partial\lambda} - \frac{\partial\Pi}{\partial\eta}\left(\frac{\partial r}{\partial\eta}\right)^{-1}\frac{\partial r}{\partial\lambda}\right) = \mathbf{S}_u, \\
&\frac{\mathrm{D}v}{\mathrm{D}t} + \frac{u^2\tan\phi}{r} + \frac{vw}{r} + 2\Omega\sin\phi u + 2\Omega\sin\phi w \\
&\qquad + \frac{C_p\theta}{r}\left(\frac{\partial\Pi}{\partial\phi} - \frac{\partial\Pi}{\partial\eta}\left(\frac{\partial r}{\partial\eta}\right)^{-1}\frac{\partial r}{\partial\phi}\right) = \mathbf{S}_v, \\
&\frac{\mathrm{D}w}{\mathrm{D}t} - \frac{(u^2+v^2)}{r} + 2\Omega\cos\phi u + C_p\theta\frac{\partial\Pi}{\partial\eta}\left(\frac{\partial r}{\partial\eta}\right)^{-1} + g = \mathbf{S}_w, \\
&\frac{\partial}{\partial t}\left(r^2\rho\frac{\partial r}{\partial\eta}\right) + \nabla\cdot\left(r^2\rho\frac{\partial r}{\partial\eta}(u,v,\dot\eta)\right) = 0, \\
&\frac{\mathrm{D}\theta}{\mathrm{D}t} = S_h, \\
&\Pi = (p/p_{\mathrm{ref}})^{-R/C_p}, \\
&p = \rho R\theta\Pi.
\end{aligned}
\tag{4.1}
$$

The discretization of the equations is described by Davies *et al.* (2005). The variables are held on the C-grid defined in Figure 3.4 in the horizontal and the Charney–Phillips grid in the vertical, with θ held at w-points and ρ at pressure points. It is shown in Thuburn (2006) that, given the form of the pressure gradient terms used in (4.1), this results in an optimal discretization of (4.1) in the vertical. The equations are solved by a semi-implicit scheme in which the factor $\nabla\Pi$ in the pressure gradient terms and the components of the Coriolis term proportional to $\sin\phi$ are integrated implicitly. The advection terms, including the metric terms that arise in spherical polar coordinates, are integrated explicitly by a vector semi-Lagrangian method based on that of Bates, Moorthi and Higgins (1993).

This scheme encompasses many of the features noted in Section 3 as desirable for accurate representation of important asymptotic limits. Research is being carried out into implicit calculation of the semi-Lagrangian departure points, as illustrated in Section 3.4, and into conservative semi-Lagrangian schemes, as discussed in Sections 3.4 and 3.6.

4.2. Validation for shallow water flow

The first set of tests is to demonstrate that the correct behaviour is obtained for shallow water flow. If the differences between the solutions of the exact equations and the asymptotic limit equations agree with the theoretical predictions, it is a validation of the experimental procedure, the numerical methods, and the correctness of the coding.

The tests are carried out on a sphere of radius a, to avoid boundary issues, but with the rotation terms replaced by a constant value 2Ω. This is to ensure that the Rossby radius is uniform over the domain. This formulation was used by Cullen (2002*a*). The water depth is written as h and the velocity as $\mathbf{u} = (u, v)$. The shallow water version of equations (4.1) is then

$$\begin{aligned} \frac{\mathrm{D}}{\mathrm{D}t} &\equiv \frac{\partial}{\partial t} + \frac{u}{a\cos\phi}\frac{\partial}{\partial\lambda} + \frac{v}{a}\frac{\partial}{\partial\phi}, \\ \frac{\mathrm{D}u}{\mathrm{D}t} - \frac{uv\tan\phi}{a} - 2\Omega v + \frac{g}{a\cos\phi}\frac{\partial h}{\partial\lambda} &= 0, \\ \frac{\mathrm{D}v}{\mathrm{D}t} + \frac{u^2\tan\phi}{a} + 2\Omega u + \frac{g}{a}\frac{\partial h}{\partial\phi} &= 0, \\ \frac{\partial h}{\partial t} + \nabla\cdot(h\mathbf{u}) &= 0. \end{aligned} \tag{4.2}$$

The tests are carried out with a shallow water version of the UM code, described in Mawson (1996), and the shallow water semi-geostrophic model of Mawson (1996). The semi-geostrophic approximation to (4.2) replaces the momentum (u, v) by (u_g, v_g), where

$$2\Omega u_g = -\frac{g}{a}\frac{\partial h}{\partial\phi}, \qquad 2\Omega v_g = \frac{g}{a\cos\phi}\frac{\partial h}{\partial\lambda}. \tag{4.3}$$

The equations can be rewritten in an analogous way to (3.25), with a matrix $\mathbf{Q}$ given by

$$\mathbf{Q} = \begin{pmatrix} 4\Omega^2 + \frac{2\Omega}{a\cos\phi}\frac{\partial v_g}{\partial\lambda} + \frac{2\Omega u_g\tan\phi}{a} & \frac{2\Omega}{a}\frac{\partial v_g}{\partial\phi} \\ -\frac{2\Omega}{a\cos\phi}\frac{\partial u_g}{\partial\lambda} + \frac{2\Omega v_g\tan\phi}{a} & 4\Omega^2 - \frac{2\Omega}{a}\frac{\partial u_g}{\partial\phi} \end{pmatrix}. \tag{4.4}$$

We also use an incompressible version of the UM code, in which the final equation of (4.2) is replaced by

$$\nabla\cdot\mathbf{u} = 0. \tag{4.5}$$

The experiments shown here were designed to test the effect of varying Ro/Fr for fixed Ro. Since the Froude number for shallow water flow is $U/\sqrt{gH}$ and so Ro/Fr is proportional to $\sqrt{gH}$ for a fixed length-scale, this was achieved by using the same perturbation depth field for all runs, but varying the mean value h_0 from 12800 m down to 400 m. The perturbation depth field is the same as that used in Cullen (2002*a*). It includes components on zonal wave-numbers ranging from 3 to 20, with amplitudes such that the total depth h is always positive and the matrix $\mathbf{Q}$, (4.4), calculated from h using the geostrophic relations is positive definite. The resulting depth field is shown in Figure 4.1. The amplitude of the variations is about ± 340 m. The horizontal velocity had a maximum value of about $16\,\mathrm{m\,s^{-1}}$. The gravity wave speed varied from $360\,\mathrm{m\,s^{-1}}$ to $65\,\mathrm{m\,s^{-1}}$. The results

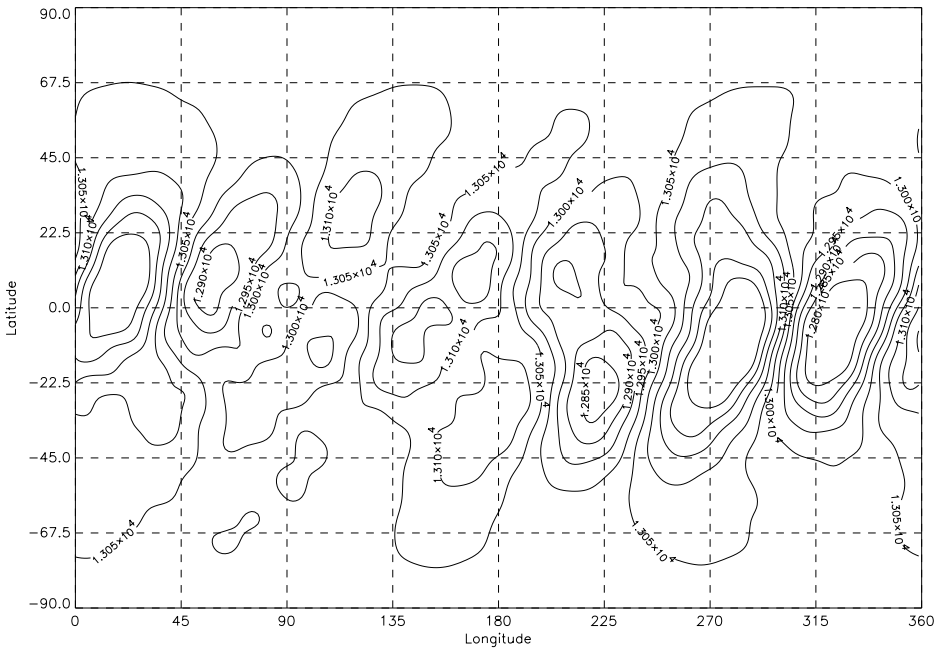

Figure 4.1. Initial depth data gh, units $\mathrm{m}^2\,\mathrm{s}^{-2}$ for shallow water experiments. Contour interval $50\,\mathrm{m}^2\,\mathrm{s}^{-2}$.

quoted by Reiser (2000, p. 65) suggest a gravity wave speed of $140\,\mathrm{m\,s}^{-1}$ as giving the best match of shallow water solutions to the evolution of the barotropic part of the atmospheric flow (*i.e.*, that part whose direction is independent of height). The average Rossby number was thus about 0.1, and the average Froude number ranged from 0.05 to 0.3. The Rossby radius L_R ranged from 3600 km (about wave-number 2) to 650 km (about wave-number 11). Thus for the largest value of L_R nearly all the variance was on scales less than L_R, and for the smallest value much of the variance was on scales larger than L_R. As discussed in Section 3.4, it is therefore expected that the semi-geostrophic model should be accurate for the small values of mean depth, but deteriorate significantly for larger values. The reverse is expected to happen for the incompressible model.

The data are initialized for the semi-geostrophic model using the procedure described by Mawson (1996). This calculates an initial u and v consistent with the given h and the semi-geostrophic approximation to (4.2). These data are passed to the shallow water model. The vorticity calculated from u and v using (3.1) is used to initialize the incompressible equations. While the incompressible solution itself is insensitive to the mean depth, the initialization procedure means that the initial vorticity is different in

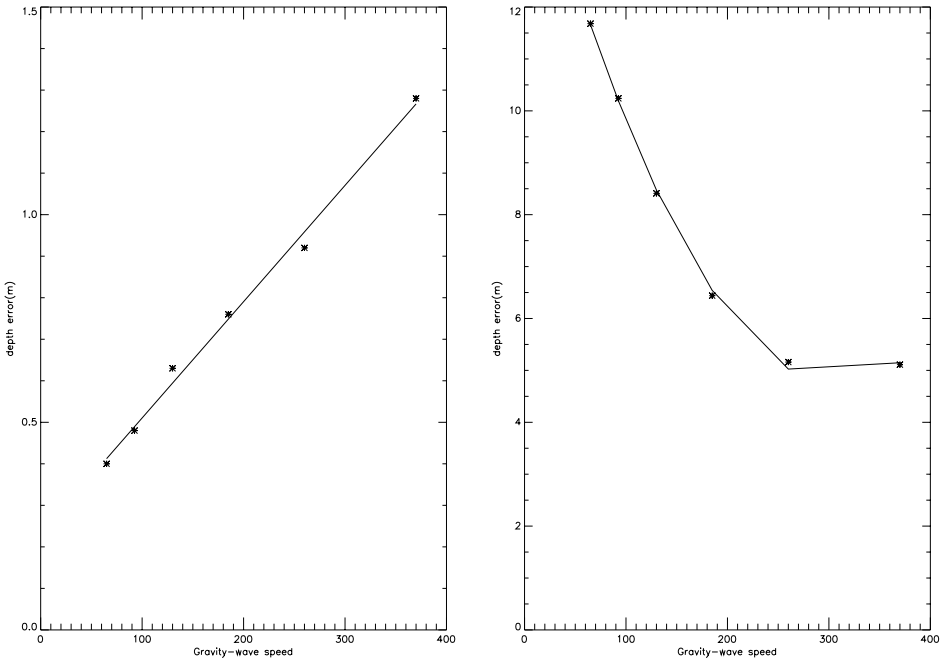

Figure 4.2. *Left*: Root-mean-square depth differences (m) between semi-geostrophic model and shallow water model with best linear fit shown. *Right*: Root-mean-square depth differences between incompressible model and shallow water model with best fit by a quadratic function. Both calculated after 48 hours and plotted against gravity-wave speed ($\mathrm{m\,s^{-1}}$).

experiments with different mean depths. The results were found to be very sensitive to the initialization procedure.

The resolution for the experiments was a latitude–longitude grid with 288 points around latitude circles and 193 points between the poles. The models are run for two days. As shown in Cullen (2002*a*), this is sufficiently short for the vorticity distribution to remain well resolved in the case of large mean depth. The results are compared in terms of r.m.s. errors of the primary prognostic fields, thus using the trajectory u, v from the semi-geostrophic model rather than the geostrophic winds.

The results for the depth errors plotted against gravity-wave speed are shown in Figure 4.2. The gravity wave speed is proportional to Fr^{-1}. The wind errors are shown in Figure 4.3. According to Cullen (2000), the expected error for the semi-geostrophic depth field is the greater of $\mathrm{O(Ro(Ro/Fr)^2)}$ and $\mathrm{O(Ro^2)}$ for small Ro and $\mathrm{Ro} < \mathrm{Fr}$. The error is O(Ro) for small Ro with $\mathrm{Ro} \geq \mathrm{Fr}$ and O(Fr) for small Fr. The expected error for

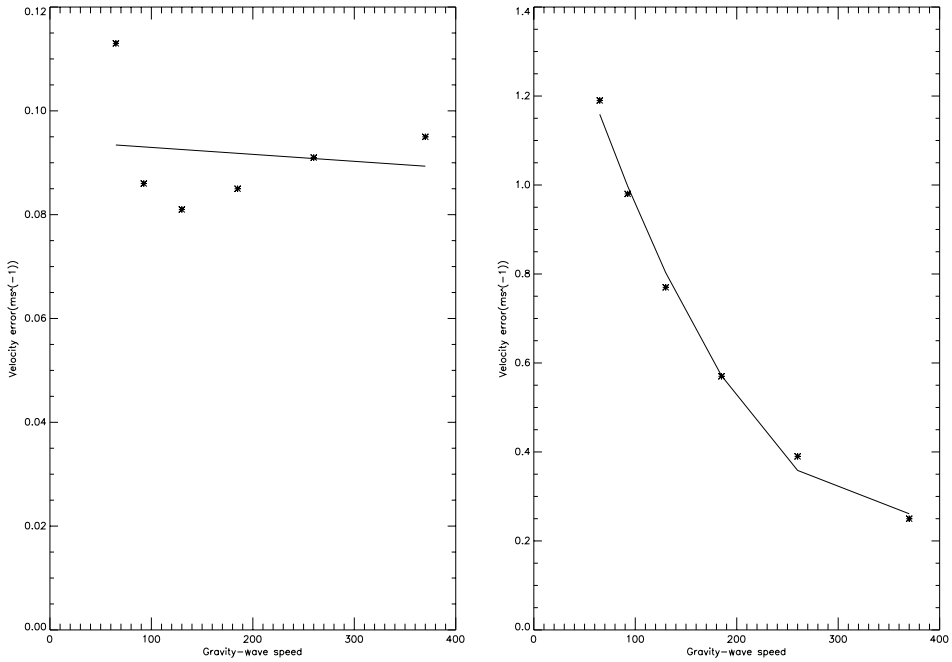

Figure 4.3. As Figure 4.2 for root-mean-square wind errors ($\mathrm{m\,s^{-1}}$).

the incompressible model is $\mathrm{O(Fr^2)}$. In the present experiment, Ro/Fr is O(1), but the smallest Fr is 0.05.

Figure 4.2 shows a linear decrease of the depth error in the semi-geostrophic model as Fr increases. This is in the expected direction, but there is no quadratic convergence, unlike that shown by Cullen (2000). This reflects the fact that Ro/Fr =O(1) in these experiments; the smaller values used in Cullen (2000) could not be reached using these data. The error in the incompressible model exhibits the expected quadratic behaviour, but it does not asymptote to zero. The errors in the incompressible model are much larger than those in the semi-geostrophic model. Iterated calculations allowing non-zero divergence, as described in Section 3.5, would have much smaller errors.

Figure 4.3 shows that the trajectory errors in the semi-geostrophic model are essentially independent of Fr. They are very small, about $0.1\,\mathrm{m\,s^{-1}}$. The errors in the incompressible model are quadratic in Fr, and asymptote to about $0.2\,\mathrm{m\,s^{-1}}$. These residual errors may result from initialization differences or numerical errors. To put this in perspective, if the shallow water and incompressible models were started with zero divergence, and vorticity calculated from geostrophic velocities derived from the semi-geostrophic model, the errors after 48 hours were of the order of $3\,\mathrm{m\,s^{-1}}$.

These results show that the numerical models are behaving in the expected way. We now supplement this by illustrating how schemes that are not constructed according to the principles set out in Section 3 may not give the correct asymptotic behaviour. We thus analyse discrete approximations to a linearized version of (4.2) in plane geometry, which allows the effect of different discretizations to be demonstrated. It is necessary to consider forced equations so that we can demonstrate convergence to the asymptotic limit for prescribed forcing. The equations are linearized about a state of rest with $h = h_0$. The equations solved are thus

$$\begin{aligned} \frac{\partial u}{\partial t} - 2\Omega v + g\frac{\partial h}{\partial x} &= A_u, \\ \frac{\partial v}{\partial t} + 2\Omega u + g\frac{\partial h}{\partial y} &= A_v, \\ \frac{\partial h}{\partial t} + h_0 \nabla \cdot \mathbf{u} &= A_h. \end{aligned} \tag{4.6}$$

We assume a domain $(-\pi L, \pi L) \times (-\pi L, \pi L) \subset \mathbb{R}^2$ and periodic boundary conditions. Assume forcing and thus solutions proportional to $\mathrm{e}^{\mathrm{i}(kx+ly-\nu t)}$. The inertia-gravity wave frequency ω is then $\sqrt{gh_0(k^2+l^2)+4\Omega^2}$. We assume that the forcing is non-resonant, so that $\nu \neq \omega$ for any choice of k and l that is an integer multiple of L^{-1}. The Froude number Fr is now the ratio of the forcing frequency ν to the gravity wave frequency $\sqrt{gh_0(k^2+l^2)}$. The limit equations as $\mathrm{Fr} \to 0$ are given by replacing the third equation of (4.6) by (4.5), noting that taking this limit for fixed forcing requires $A_h/h_0 \to 0$. We then eliminate h to obtain

$$\begin{aligned} \frac{\partial \zeta}{\partial t} &= \frac{\partial A_v}{\partial x} - \frac{\partial A_u}{\partial y}, \\ \zeta &= \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \\ \nabla \cdot \mathbf{u} &= 0. \end{aligned} \tag{4.7}$$

Write the solution of (4.6) as $(\hat{u}, \hat{v}, \hat{h}) \exp^{\mathrm{i}(kx+ly-\nu t)}$ and the forcing as $(\hat{A}_u, \hat{A}_v, \hat{A}_h) \exp^{\mathrm{i}(kx+ly-\nu t)}$. Consider, for example, the case where $\hat{A}_v = \hat{A}_h = 0$. Then we can show that

$$\begin{pmatrix} \hat{u} \\ \hat{v} \\ \hat{h} \end{pmatrix} = \frac{\mathrm{i}}{\nu\left(\frac{\nu^2}{gh_0} - (k^2+l^2) - \frac{4\Omega^2}{gh_0}\right)} \begin{pmatrix} \frac{\nu^2}{gh_0} - l^2 \\ -\frac{2\mathrm{i}\Omega\nu}{gh_0} - kl \\ \frac{1}{g}(-2\mathrm{i}\Omega l + k\nu) \end{pmatrix} \hat{A}_u. \tag{4.8}$$

The limit solution for small Fr, using (4.7) and inferring $\hat{h}$ by back sub-

stitution, is given by

$$\begin{pmatrix}\hat{u}\\ \hat{v}\\ \hat{h}\end{pmatrix} = \frac{\mathrm{i}}{-\nu(k^2+l^2)}\begin{pmatrix}-l^2\\ -kl\\ \frac{1}{g}(-2\mathrm{i}\Omega l + k\nu)\end{pmatrix}. \tag{4.9}$$

Since $\mathrm{Fr}^2 \propto 1/(gh_0)$, we can see from (4.8) and (4.9) that the solution of (4.6) converges to the limit solution at a rate $\mathrm{O}(\mathrm{Fr}^2)$. The accuracy of the limit solution increases as k and l increase, consistent with the dependence on Ro/Fr demonstrated above.

Now consider a discretization of (4.6) using the C-grid and implicit time integration as used in the UM. This can be written using the notation introduced in Section 3.2 as

$$\begin{aligned}\delta_t u - 2\Omega\overline{v}^t + g\delta_x\overline{h}^t &= \overline{A}_u^t,\\ \delta_t v + 2\Omega\overline{u}^t + g\delta_y\overline{h}^t &= \overline{A}_v^t,\\ \delta_t h + h_0(\delta_x\overline{u}^t + \delta_y\overline{v}^t) &= \overline{A}_h^t.\end{aligned} \tag{4.10}$$

Assuming grid-lengths $\delta x, \delta y$, write

$$\tilde{k} = 2\sin(\tfrac{1}{2}k\delta x)/\delta x, \qquad \tilde{l} = 2\sin(\tfrac{1}{2}l\delta y)/\delta y.$$

Write $\tilde{\nu} = 2\tan(\frac{1}{2}\nu\delta t)$ and write $\tilde{A}_u$ for the spatially discretized version of the forcing term A_u, with similar notation for the other forcing terms. Then the discrete version of (4.8) becomes

$$\begin{pmatrix}\hat{u}\\ \hat{v}\\ \hat{h}\end{pmatrix} = \frac{\mathrm{i}}{\tilde{\nu}\big(\frac{\tilde{\nu}^2}{gh_0} - (\tilde{k}^2+\tilde{l}^2) - \frac{4\Omega^2}{gh_0}\big)}\begin{pmatrix}\frac{\tilde{\nu}^2}{gh_0} - \tilde{l}^2\\ -\frac{2\mathrm{i}\Omega\tilde{\nu}}{gh_0} - \tilde{k}\tilde{l}\\ \frac{1}{g}(-2\mathrm{i}\Omega\tilde{l} + \tilde{k}\tilde{\nu})\end{pmatrix}\tilde{A}_u \tag{4.11}$$

and (4.9) becomes

$$\begin{pmatrix}\hat{u}\\ \hat{v}\\ \hat{h}\end{pmatrix} = \frac{\mathrm{i}}{-\tilde{\nu}(\tilde{k}^2+\tilde{l}^2)}\begin{pmatrix}-\tilde{l}^2\\ -\tilde{k}\tilde{l}\\ \frac{1}{g}(-2\Omega\tilde{l} + \tilde{k}\tilde{\nu})\end{pmatrix}\tilde{A}_u. \tag{4.12}$$

Since $k\delta x, l\delta y, \nu\delta t$ are all less than or equal to π for any resolved function, $\tilde{k}$, $\tilde{l}$ and $\tilde{\nu}$ are all non-zero. Therefore convergence of (4.11) to (4.12) will occur at the predicted rate $\mathrm{O}(\mathrm{Fr}^2)$.

If the spatial discretization is instead performed on the B-grid defined in Figure 3.3, then

$$\tilde{k} = 2\sin(\tfrac{1}{2}k\delta x)\cos(\tfrac{1}{2}l\delta y)/\delta x, \qquad \tilde{l} = 2\sin(\tfrac{1}{2}l\delta y)\cos(\tfrac{1}{2}k\delta x)/\delta y.$$

These are both smaller than their values on the C-grid, so the accuracy of the approximation to the limit solution will be degraded. $\tilde{k}$ and $\tilde{l}$ are both

zero if $k\delta x = l\delta y = \pi$. In this case, (4.11) reduces to

$$\begin{pmatrix}\hat{u}\\ \hat{v}\\ \hat{h}\end{pmatrix} = \frac{\mathrm{i}}{\tilde{\nu}\left(\frac{\tilde{\nu}^2}{gh_0} - \frac{4\Omega^2}{gh_0}\right)} \begin{pmatrix}\frac{\tilde{\nu}^2}{gh_0}\\ -\frac{2\mathrm{i}\Omega\tilde{\nu}}{gh_0}\\ 0\end{pmatrix} \tilde{A}_u = \frac{\mathrm{i}}{(\tilde{\nu}^2 - 4\Omega^2)} \begin{pmatrix}\tilde{\nu}\\ -2\mathrm{i}\Omega\\ 0\end{pmatrix} \tilde{A}_u. \quad (4.13)$$

This represents a forced inertial oscillation. This cannot be a solution of (4.7) because the latter equation has no dependence on Ω. The B-grid analogue of (4.7) becomes completely degenerate in this case and is satisfied by any $(\hat{u}, \hat{v})$, leading to accumulation of noise and possible computational instability. We thus see that an inappropriate choice of discretization can lead to the limiting behaviour of the solution being incorrect. Note that in the analytic case, the manipulations leading to (4.7) do not make sense if $k = l = 0$ and so in this case (4.7) is not the correct limit of (4.6) as $\mathrm{Fr} \to 0$. The B-grid discretization creates this situation unphysically because of numerical errors.

4.3. Validation for baroclinic waves and fronts

The tests described in this subsection use the Eady model of frontogenesis (Gill 1982, p. 556) and the Boussinesq incompressible forms of (4.14) and (4.15). This allows the fundamental mechanism by which extra-tropical weather systems evolve to be studied in two-dimensional vertical slice geometry. Thus all variables are assumed to be independent of one horizontal direction except for basic state variations in pressure and potential temperature which are in hydrostatic balance. We use Cartesian coordinates (x, y) in the horizontal. Since we will use the same model for tests of flow over ridges, we retain the terrain-following coordinate from (4.1). Write

$$\Pi = \Pi_1(y, \eta) + \Pi(x, \eta), \quad \theta = \theta_1(y) + \theta(x, \eta),$$

with $C_p\theta_1(\partial\Pi_1/\partial\eta)(\partial z/\partial\eta)^{-1} + g = 0$. Equations (4.1) then reduce to

$$\frac{\mathrm{D}}{\mathrm{D}t} \equiv \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + \dot{\eta}\frac{\partial}{\partial \eta},$$
$$\dot{\eta}\frac{\partial z}{\partial \eta} = w - u\frac{\partial z}{\partial x},$$

$$\frac{\mathrm{D}u}{\mathrm{D}t} - fv + C_p\theta\left(\frac{\partial \Pi}{\partial x} - \frac{\partial \Pi}{\partial \eta}\left(\frac{\partial z}{\partial \eta}\right)^{-1}\frac{\partial z}{\partial x}\right) = 0,$$
$$\frac{\mathrm{D}v}{\mathrm{D}t} + fu + C_p\theta\frac{\partial \Pi_1}{\partial y} = 0,$$
$$\frac{\mathrm{D}w}{\mathrm{D}t} + C_p\theta\frac{\partial \Pi}{\partial \eta}\left(\frac{\partial z}{\partial \eta}\right)^{-1} + g = 0, \quad (4.14)$$

$$\frac{\partial}{\partial t}\left(\rho\frac{\partial z}{\partial \eta}\right)+\nabla\cdot\left(\rho\frac{\partial z}{\partial \eta}(u,\dot\eta)\right)=0,$$
$$\frac{\mathrm{D}\theta}{\mathrm{D}t}+v\frac{\partial\theta_1}{\partial y}=0,$$
$$\Pi=(p/p_{\mathrm{ref}})^{-R/C_p},$$
$$p=\rho R\theta\Pi.$$

The use of vertical slice geometry means that the semi-geostrophic model is the natural approximation when rotation is important. This is because all quantities are assumed independent of y, and the frontal scaling introduced by Hoskins (1975) is appropriate. The semi-geostrophic approximation to equations (4.14) is

$$\begin{aligned}
-fv+C_p\theta\left(\frac{\partial\Pi}{\partial x}-\frac{\partial\Pi}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}\frac{\partial z}{\partial x}\right)&=0,\\
\frac{\mathrm{D}v}{\mathrm{D}t}+fu+C_p\theta\frac{\partial\Pi_1}{\partial y}&=0,\\
C_p\theta\frac{\partial\Pi}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}+g&=0,\\
\frac{\partial}{\partial t}\left(\rho\frac{\partial z}{\partial\eta}\right)+\nabla\cdot\left(\rho\frac{\partial z}{\partial\eta}(u,\dot\eta)\right)&=0,\\
\frac{\mathrm{D}\theta}{\mathrm{D}t}+v\frac{\partial\theta_1}{\partial y}&=0,\\
\Pi=(p/p_{\mathrm{ref}})^{-R/C_p},&\\
p=\rho R\theta\Pi.&
\end{aligned}\tag{4.15}$$

The first of these equations states that v is geostrophic, showing that these equations are consistent with (3.23) under the assumption of no y-dependence. As with the derivation of (3.26), these equations can be solved by reducing them to a single elliptic equation for $\partial\Pi/\partial t$. The method is described in detail in Cullen (2007). However, it is not clear if this form of the equations has well-posed solutions.

Cullen and Maroofi (2003) showed that these equations are well-posed if written in geostrophic and isentropic coordinates and then solved as a transport equation for the mass density. The coordinates are defined by

$$X=x+f^{-1}v,\ Z=\theta.\tag{4.16}$$

The second and fifth of equations (4.15) then become

$$\begin{aligned}
\frac{\mathrm{D}X}{\mathrm{D}t}+C_p\theta\frac{\partial\Pi_1}{\partial y}&=0,\\
\frac{\mathrm{D}Z}{\mathrm{D}t}+f(X-x)\frac{\partial\theta_1}{\partial y}&=0.
\end{aligned}\tag{4.17}$$

The UM equations (4.14) are solved by the methods described in Section 4.1. Thus u is staggered in the horizontal from Π, but, because of the slice geometry, v is held at the same points as Π. The algorithm for solving (4.15) uses the same vertical arrangement of variables. However, in order to represent geostrophic balance, v is staggered in the horizontal from Π. The matrix $\mathbf{Q}$, defined as in (3.25), which appears in the elliptic equation (3.26) for $\partial\Pi/\partial t$, is evaluated at pressure points. This means that the finite differences on the diagonal are calculated over a single grid-length, but the off-diagonal terms involve additional averaging. This has the desirable property of increasing the diagonal dominance of the system. The reduction of the problem to a single equation for Π is needed in order to ensure that v and θ represent an exactly geostrophic and hydrostatic state. As discussed in Section 3.4, physically relevant solutions of SG are characterized by the matrix $\mathbf{Q}$ being positive definite. This condition is also required for (3.26) to be solvable. This condition is not naturally enforced by the numerical method. It is therefore necessary to correct the data at the end of each timestep so that $\mathbf{Q}$ is positive definite. This is done by a variational method described in Cullen (2007).

Previous work on this problem by Cullen and Roulstone (1993) used a geometric algorithm to solve the Boussinesq incompressible form of (4.15) and showed that the solutions represent a sequence of energy-conserving life-cycles in which disturbances develop and decay. The solutions are discontinuous when the disturbances are fully developed and are shown to be predictable for many life-cycles. These results were confirmed by a more efficient geometric algorithm due to R. J. Purser; see Cullen (2006, Section 5.3.2).

We illustrate the main points of the solutions obtained using the geometric algorithm. The equations are solved on a domain $\Gamma : [-L, L] \times [0, H]$ in the (x, z) plane with periodic boundary conditions in x and rigid wall conditions $w = 0$ on $z = 0, H$. The initial data are represented on a set of fluid elements, on each of which X and Z as defined in (4.16) are constant. Two resolutions are used, of 21×13 and 40×26 elements. Solutions of the Boussinesq incompressible form of (4.15) for initial data of the form

$$\theta' = N_0^2 \theta_0 z/g + B \sin\bigl(\pi(x/L + z/H)\bigr), \tag{4.18}$$

where N_0^2 and B are positive constants, are illustrated. The data used are taken from Nakamura (1994), so that

$$L = 1000\,\mathrm{km}, \quad H = 10\,\mathrm{km}, \quad N_0^2 = 2.5 \times 10^{-5}\,\mathrm{s}^{-2}, \quad f = 10^{-4}\,\mathrm{s}^{-1},$$
$$g = 10\,\mathrm{m\,s}^{-1}, \quad \theta_0 = 300\,\mathrm{K}, \quad \partial\theta_1/\partial y = 3 \times 10^{-6}\,\mathrm{m}^{-1}\,\mathrm{K}.$$

These data correspond to an unstable mode of the linearized equations derived from (4.15). As discussed in Gill (1982, p. 556), if the isentropes have a negative slope $\mathrm{d}x/\mathrm{d}z$, then v_g will increase with z, and the evolution

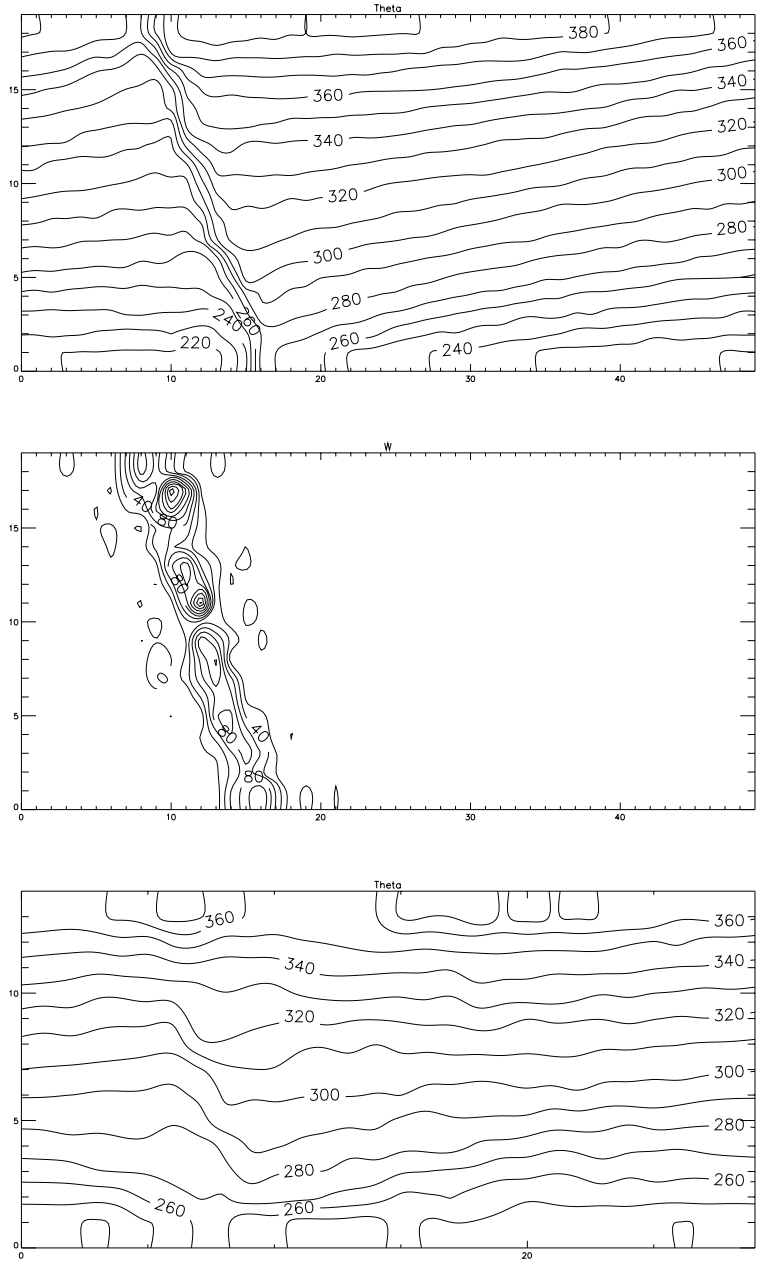

Figure 4.4. *Top*: Plot of potential temperature after 8 days (degrees K, contour interval 10 K). *Middle*: Plot of potential vorticity scaled by $f^2 N_0^2$ after 8 days (contour interval 20). *Bottom*: Plot of potential temperature after 15 days (degrees K, contour interval 10 K). From Cullen (2006).

equation for θ' will increase the vertical gradient of θ, giving a positive feedback. It represents conversion of potential energy from the infinite reservoir implied by the imposed basic state $\partial\theta_1/\partial y$ into kinetic energy.

We show the solutions after 8 days in Figure 4.4, at which point there is strong frontogenesis, as illustrated by the potential temperature plot. The potential vorticity at the same time is also shown. There is a large increase in the value in the frontal zone. The semi-geostrophic potential vorticity is defined as the determinant of the matrix $\mathbf{Q}$ defined in (3.25). As discussed in Cullen (2006, Chapter 3), it represents the Jacobian of the transformation between physical coordinates and the coordinates (X, Z) defined in (4.16). During the time evolution, the area occupied by the fluid is conserved in both (x, z) and (X, Z) coordinates. However, the shape of the region occupied by the fluid in (X, Z) space becomes highly distorted, as illustrated schematically in Figure 4.5. The total potential vorticity calculated in (x, z) space would be the ratio of the area of the convex hull of Σ to the area of Γ. This ratio will not be conserved during the time evolution.

The potential vorticity is initially equal to the uniform value $f^2 N_0^2$. At the front it becomes a Dirac mass, which is represented in Figure 4.4 as large values by the plotting software. The irregularities are due to the use of piecewise constant data and thus the irregularity of the boundaries between elements. The small negative values are artifacts of the plotting.

The formation of the fronts at the upper and lower boundaries destroys the normal mode property of the initial data, and the vertical shear in the basic state reverses the slope of the isentropes by day 10. The front is then destroyed. The mean potential vorticity remains larger than its initial value, represented by layers of enhanced static stability near the upper and lower boundaries as shown in Figure 4.4. Further life-cycles then take place, with

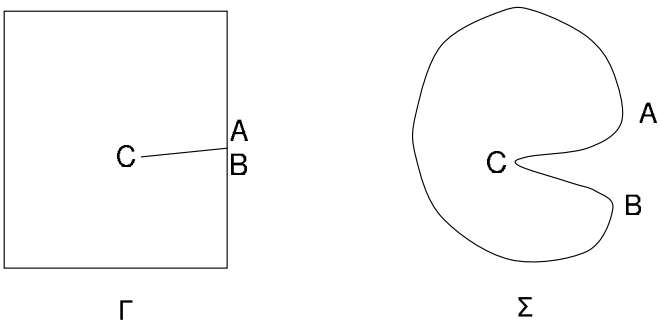

Figure 4.5. Mapping a region in (X, Z) space (Σ) to a rectangular region in (x, z) (Γ). The points A, B and C indicate corresponding points in Γ and Σ. From Cullen (2006).

the disturbances confined to the region of uniform static stability in the centre of the domain.

A graph of the kinetic energy against time is shown in Figure 4.6. This shows that after 8 days the maximum kinetic energy is reached. It then reduces, and smaller amplitude life-cycles follow. The periodic oscillations continue to day 30. The graphs for the two resolutions are almost identical, showing that the solution is highly predictable, despite the formation of fronts. The Boussinesq incompressible form of (4.15) has a natural period equal to $2Lf\theta_1(gH\partial\theta_1/\partial y)$, the length of time between when features at the upper and lower boundaries come back into phase under the action of the basic state wind. With the data chosen, the difference in the basic state wind between the boundaries is $10\,\mathrm{m\,s^{-1}}$, giving a period of about 2.3 days. This is much shorter than the period observed, reflecting the fact that the vertical shear is impeded during the growth phase. It also shows that the prediction of the same period by two different discretizations is a non-trivial achievement.

Solutions using conventional numerical methods have been obtained by Nakamura and Held (1989) and Nakamura (1994). They lose predictability after a single life-cycle. Nakamura and Held (1989), using solutions of the Boussinesq incompressible form of (4.14), show that the loss in predictability is associated with the need to include a form of artificial viscosity in order to capture the frontogenesis. The solution after the first life-cycle

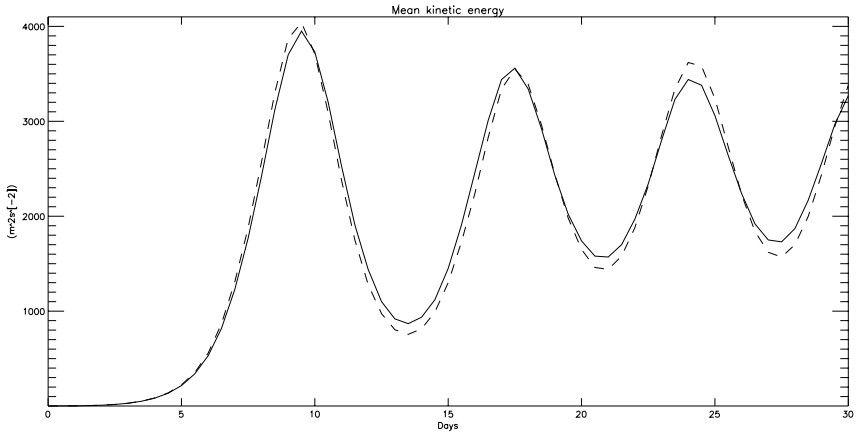

Figure 4.6. Graph of domain-averaged meridional kinetic energy $\mathrm{m^2\,s^{-2}}$ against time for the solution of the Boussinesq incompressible form of (4.15) by the geometric method. Solid line = 40×26 elements, dashed line = 21×13 elements. From Cullen (2006).

is strongly dependent on the form of artificial viscosity used. Nakamura (1994) describes solutions of the Boussinesq incompressible form of (4.15) using Eulerian numerical methods based on potential vorticity, and using artificial viscosity to capture the front. There is a large increase in potential vorticity generated by the artificial viscosity, which again results in layers of enhanced static stability near the boundaries.

In this section we illustrate solutions obtained using the UM discretization scheme and a solution of the semi-geostrophic equations in real variables, based on (4.15). These are compared qualitatively with the solutions of the Boussinesq incompressible equations discussed above, but the additional approximations mean that quantitative comparisons are not appropriate. The integration domain is the same as that used above, with the same flow parameters. In order to ensure that the initial potential vorticity, as defined by (2.6), is uniform, N_0^2 is set to $2.5 \times 10^{-5}\rho_0(0)/\rho_1(z)$, where $\rho_1(z)$ is an initial reference profile in hydrostatic balance calculated for an isentropic state using (2.8). The basic state $\partial\Pi_1/\partial y$ is chosen to vary linearly in z, so that the initial u in geostrophic balance with it varies from -5 to $5\,\mathrm{m\,s^{-1}}$ over the depth of the domain, as in the Boussinesq incompressible case. The basic state $\partial\theta_1/\partial y$ is calculated from $\partial\Pi_1/\partial y$ using the hydrostatic relation.

The mean meridional kinetic energy from the two integrations is illustrated in Figure 4.7. The mean potential vorticity, scaled by its initial value, is shown in Figure 4.8. The semi-geostrophic integration produces an earlier growth of the initial disturbance. The UM integration gives an initial growth on the same time-scale as the results of Nakamura (1994). After the initial growth, the semi-geostrophic integration reduces the meridional kinetic energy to very small values while the UM retains a value close to the maximum reached. Both integrations show a large increase in potential vorticity as the disturbance grows, and retain high values for the rest of the integration. Further life-cycles occur in the UM integration, but with much smaller amplitude than those given by the geometric model and even those shown by Nakamura (1994).

Further study of the results (diagrams not shown) shows that, after the initial growth, both solutions are dominated by a wave in the v-field which is almost independent of height and a geostrophically balanced pressure. The potential temperature signal is rather weak. In the UM, the wave then propagates slowly through the domain. It is completely different from the solutions shown in Nakamura (1994). However, the compressible formulation of the UM means that it has solutions not available to the models of Nakamura (1994). After the initial life-cycle, the semi-geostrophic solution becomes almost independent of height, and the amplitude of the wave in the v-field then slowly decays. The increased potential vorticity is represented by layers of enhanced static stability at both boundaries. The latter feature can also be seen in the geometric model solutions shown in Figure 4.4.

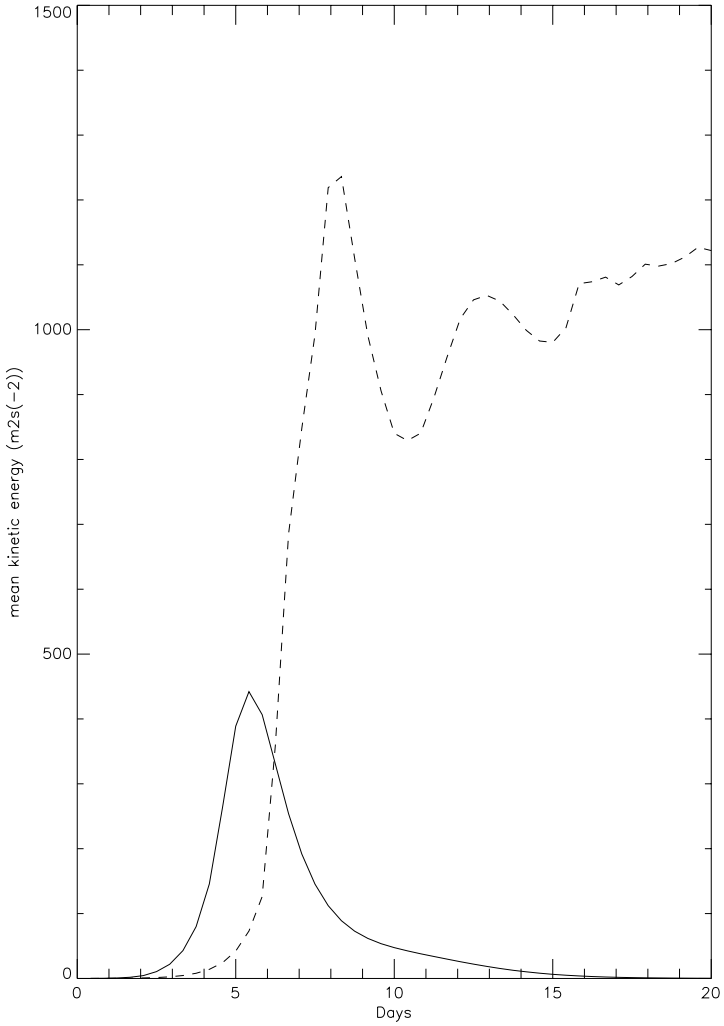

Figure 4.7. Graph of domain-averaged meridional kinetic energy $\mathrm{m}^2\,\mathrm{s}^{-2}$ against time. Solid line = solution of (4.15), dashed line = solution of (4.14).

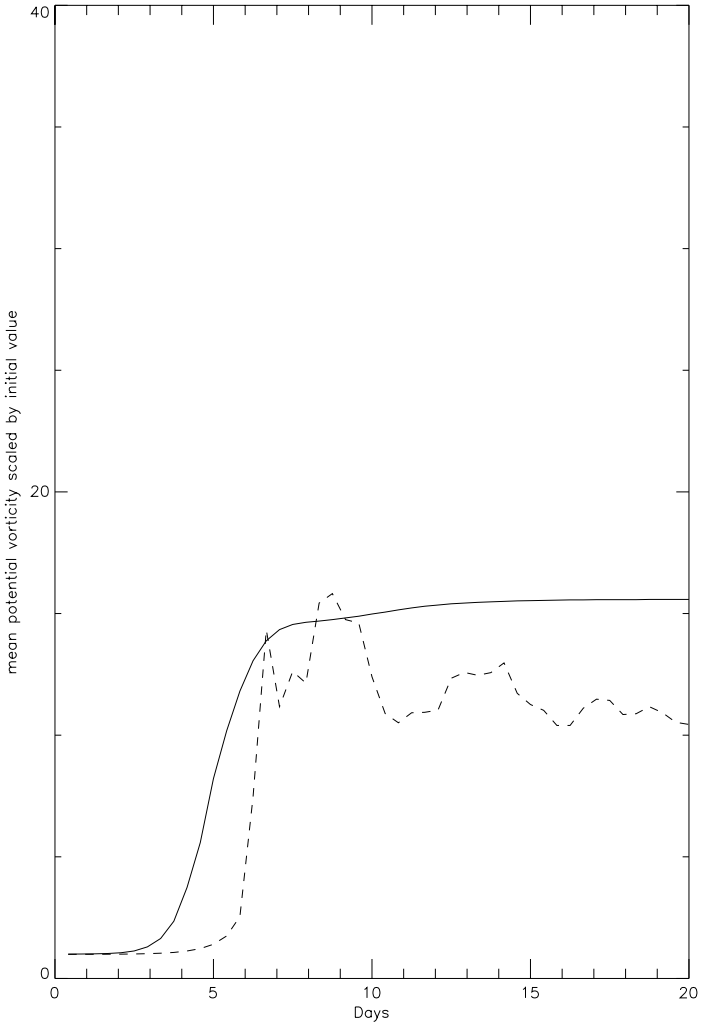

Figure 4.8. Graph of domain-averaged potential vorticity, scaled by its initial value, against time. Solid line = solution of (4.15), dashed line = solution of (4.14).

The symmetry of the solution in the vertical shown from the geometric model is lost because of the large basic-state density variation.

These results suggest that neither model is able to capture the qualitative behaviour shown in the geometric model. The reason is that the state left after the initial life-cycle is strongly dependent on the numerical algorithm. The geometric method is fully Lagrangian, and thus much more naturally suited to the problem. Fully Lagrangian methods have not yet been considered as practical for production models, though the contour advection method of Mohebalhojeh and Dritschel (2004) has been successful in idealized problems. The normal formulation of this method would be unsuitable in the present case because of the lack of potential vorticity conservation integrated over the physical domain. It would be necessary to apply the method in geostrophic and isentropic coordinates, which is only practicable even in principle for a semi-geostrophic model.

4.4. Validation for large-scale flow over a ridge

In this section we demonstrate in vertical-slice geometry that solutions of the UM equations (4.14) converge to the solutions of the semi-geostrophic (SG) equations (4.15) at the expected rate in the presence of orography. These experiments are described in more detail in Cullen (2007), where the tests are carried out in a wider range of cases.

The UM equations (4.14) and SG equations (4.15) are solved as described in Section 4.3. The forcing pressure gradient $\partial\Pi_0/\partial y$ is chosen to be uniform in z. The initial horizontal pressure gradient is set to zero. Because of the orography, the values of Π have to be calculated so that

$$\frac{\partial\Pi}{\partial x} - \frac{\partial\Pi}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}\frac{\partial z}{\partial x} = 0 \tag{4.19}$$

in a discrete sense. This can only be achieved if $\partial^2\Pi/\partial z^2$ is uniform, because otherwise the finite difference expressions for the terms in (4.19) do not cancel up to rounding error. We therefore use initial and boundary pressure data satisfying this condition, and modify the finite differencing in Davies *et al.* (2005) to extrapolate θ to the lower boundary assuming hydrostatic balance and constant $\partial^2\Pi/\partial z^2$. In the SG model, the matrix $\mathbf{Q}$ defined in (3.25) is calculated in (x, z) coordinates rather than (x, η) coordinates, and the positive-definiteness condition is enforced in these coordinates. The matrix is then transformed to (x, η) coordinates as described in Cullen (2007).

Assume a length-scale l, which will be the half-width of the ridge, and a geostrophic velocity u_g in the x-direction. The Froude number Fr is defined as $u_g/(Nh)$, where N is the Brunt–Väisälä frequency and h is the ridge

height. The Rossby number Ro is defined as u_g/fl. The Rossby radius $L_R = NH/f$. The error estimates for SG theory were given in Section 4.2. We test the limit Ro $\to$ 0 for fixed L_R by letting $u_g \to 0$ and leaving all other parameters fixed. In Cullen (2007) the limit Ro $\to$ 0 with Ro =O(Fr^2) was also tested.

We choose a domain of length 2000 km and height 10 km. The ridge is centred at x =750 km, and has width 150 km and height 2400 m. The ridge profile is an isolated cosine-squared hill. The Brunt–Väisälä frequency is 0.01 and the Coriolis parameter 10^{-4}. The ridge is thus slightly narrower than the radius of deformation. Since u_g is determined from $\partial\Pi_0/\partial y$, which is height-independent, u_g increases slowly with z because of the increase of θ. Values of u_g at $z = 0$ are chosen to range from 0.625 m s^{-1} to 10 m s^{-1}, giving Ro ranging from 0.083 to 1.33.

The integrations start from very simple initial data with uniform u and zero w. A rapid initial adjustment will take place to create an approximately non-divergent flow. A further adjustment will take place on the inertial time-scale to establish approximate geostrophic balance of the wind parallel to the ridge. The experiments are therefore run for several inertial periods to allow this to happen. The solutions that are verified will be close to steady-state. Since it is unlikely that either model actually gives steady-state solutions to this problem, there will be a sampling issue in the error calculations. Results are therefore shown for two times, corresponding to 1.875 and 2.5 days for $u_g = 10$ m s^{-1}. Two spatial resolutions are also shown for the UM integrations, of 121 $\times$ 61 and 201 $\times$ 121 points respectively. The SG solutions are only shown for 121 $\times$ 61 resolution, since they have essentially converged by this point.

We note that the SG solution is independent of u_g in the sense that the pressure, and fields derived directly from it, only depend on u_gT where T is the period of integration. The velocities u, $\dot{\eta}$ and w will be proportional to u_g. Thus, in order to compare results with different values of u_g, we choose a total integration time inversely proportional to u_g. This means that the same SG states can be used for comparison of UM runs with all values of u_g.

The limit solution for $u_g = 0$ has the same p, v_g and θ as the solution for finite u_g, and has $u = \dot{\eta} = 0$. Since this represents an exactly geostrophic and hydrostatic state, it is also a solution of the UM equations. Convergence should therefore be possible. The limit solution depends on the history of the problem. If the limit $u_g \to 0$ is taken with negative values of u_g, the SG solution will be reflected in $x = 0$, and thus be different.

The main characteristics of the solution are discussed in Cullen (2006, Chapter 6) and Shutts (1998). Some of these can be seen from the cross-section of v from the UM shown in Figure 4.9. A particular feature is the blocking of cold air near the surface on the upstream side of the ridge and the associated barrier jet. The effect is to extend the influence of the ridge

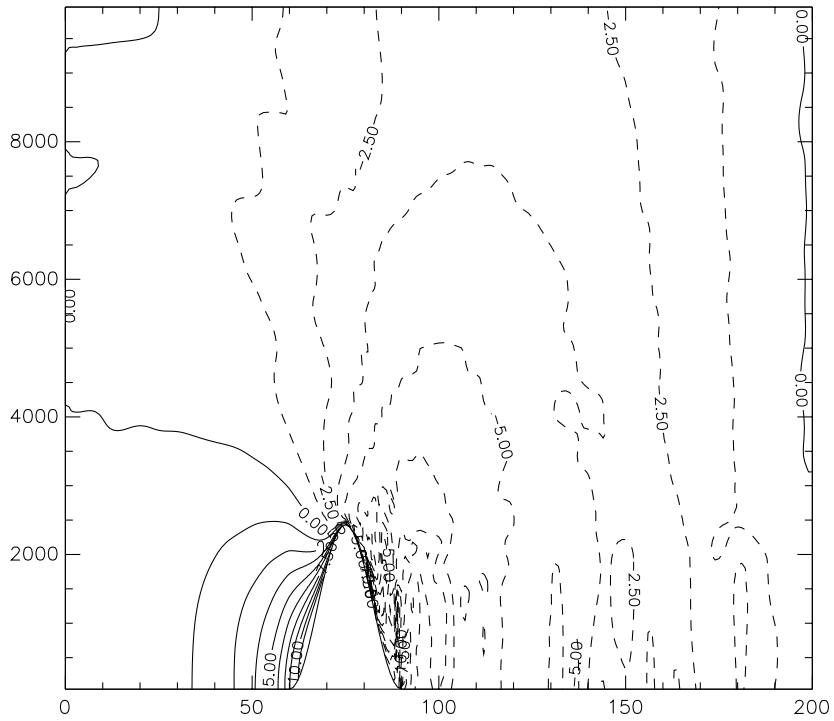

Figure 4.9. UM solution for v at 40 days (solid), with $\mathrm{Ro} = 0.083$ and 201×121 grid, plotted against height (m) and horizontal grid-point number. Units $\mathrm{m\,s^{-1}}$, contour interval $1.25\,\mathrm{m\,s^{-1}}$. From Cullen (2007).

upstream for a horizontal distance L_R. The solution has a pressure force acting on the ridge, which represents the orographic drag. Air trapped on the upstream side of the ridge will have v increasing with time, as u is constrained to be zero. Therefore the slopes of the isentropes, which are related to $\partial v/\partial z$, will increase with time and trapped air will reach the top of the ridge with a large v. At this point, there will be a large negative value of $f + \partial v/\partial x$, so that the air parcel will be unstable and 'jump' downstream to a stable position. In reality, there would be a rapid down-slope wind not described by semi-geostrophic theory. In the UM solutions, we therefore expect to see a region of inertial instability downstream of the mountain, together with a downslope wind. The overturning associated with inertial instability is clearly seen in Figure 4.9. This would be expected to relax

to a stable state as given by the SG model, but the relaxation could take many inertial periods. In the real three-dimensional system the relaxation is likely to be faster as three-dimensional turbulence will be generated in the inertially unstable region.

Figure 4.10 shows the r.m.s. differences in v. These exhibit the expected linear convergence, with little difference between the two UM resolutions. However, the best fit lines do not actually go to zero. This is likely to reflect the failure of the UM to relax to a symmetrically stable state at very small Ro as discussed above. It is also shown in Cullen (2007) that the need to enforce positive-definiteness of $\mathbf{Q}$ in the SG model means that the maximum of the barrier jet occurs below the top of the ridge. This is because the variational adjustment which is used to enforce this condition does not know about the flow direction, and therefore makes adjustments both upstream and downstream of the ridge. In the UM, the barrier jet does reach the top of the ridge, as shown in Figure 4.9. This contributes to the error for small Ro.

Figure 4.11 shows the convergence of the drag. Again the expected linear convergence is achieved, with little sensitivity to UM resolution. There is less of an issue at small Ro than with the v-field. This is because the drag measures the total amount of flow blocking, which is determined by large-scale aspects of the problem. It is insensitive to the details of how the energy is dissipated downstream of the ridge.

Figure 4.12 shows the convergence in θ. This is less satisfactory, though the error does reduce with Ro. The convergence is better using the higher resolution UM data, suggesting that numerical errors may be significant. Difference fields (not shown) indicate that the errors are mainly upstream of the ridge, where the SG model has greater flow blocking, and in the inertially unstable region downstream. These are related to the same issues as those discussed in the context of the errors in v.

The test of the limit $\mathrm{Ro} \to 0$ with $\mathrm{Ro} = \mathrm{O}(\mathrm{Fr}^2)$ in Cullen (2007) shows a similar story. The expected higher-order rate of convergence is obtained, but the errors do not asymptote to zero. The error in the θ field is the most sensitive to numerical errors. In this limit, the SG solution becomes smooth, so issues with the removal of inertial instability are not relevant. The residuals thus probably represent accumulated computational errors.

Overall, this problem suggests reasonable agreement with theory. The residual errors are related to the technical difficulty of maintaining inertial stability in the SG model, and the issue of how fast and on what scales the UM should relax to an inertially stable solution. There are also issues in converting an accurate solution of the evolution problem into an accurate solution of the steady-state problem. Tests of accuracy in the initial evolution would be of little practical relevance because of the large transient motions set up by the initial conditions chosen.

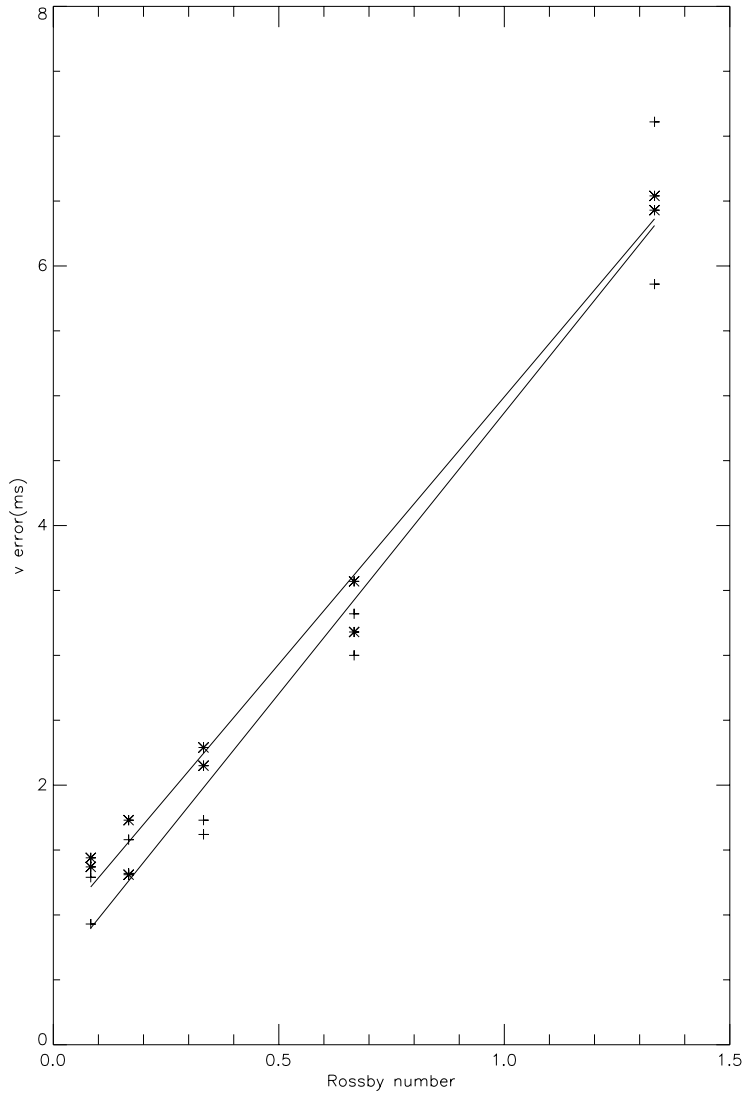

Figure 4.10. Convergence of r.m.s. differences of v ($\mathrm{m\,s^{-1}}$) between UM and SG solutions for fixed L_R, plotted against Rossby number. SG solutions on 121×61 grid. Thin line = best linear fit in Ro to differences between UM solutions on 121×61 grid and SG solutions. Individual differences plotted as + symbols, using two verification times for each Ro. Thick line = best linear fit in Ro to differences between UM solutions on 201×121 grid and SG solutions. Individual differences plotted as asterisks, using two verification times for each Ro. From Cullen (2007).

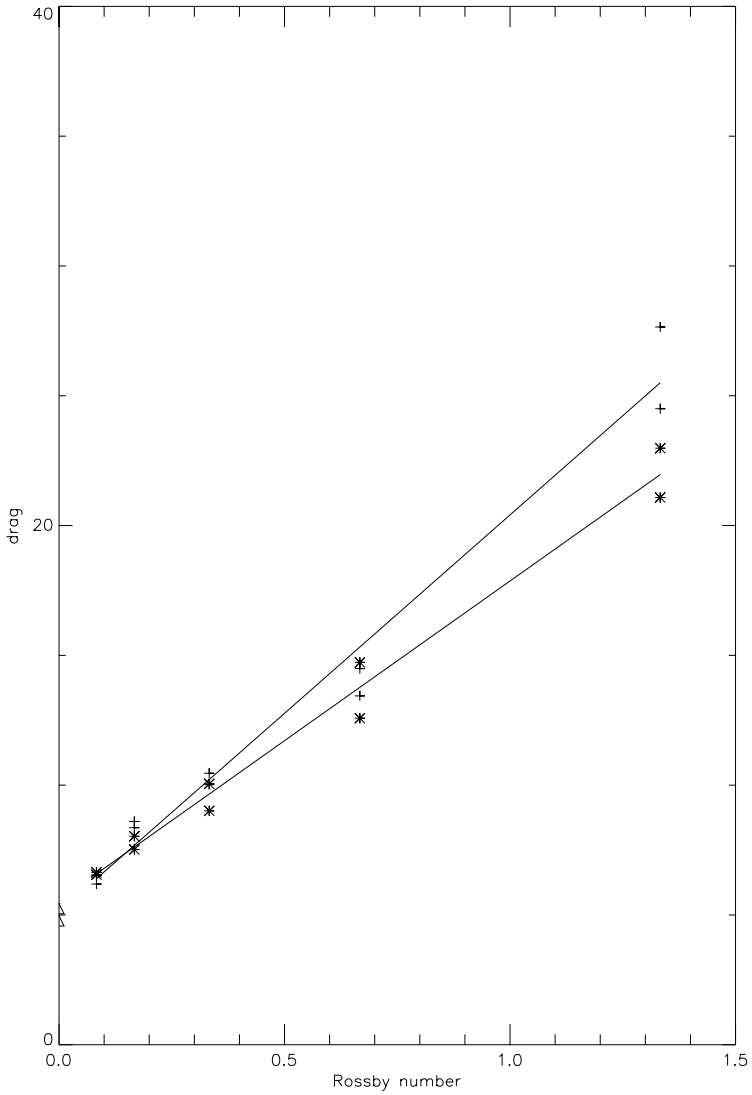

Figure 4.11. Convergence of drag on the ridge between UM and SG solutions for fixed L_R, plotted against Rossby number. The drag is measured per unit length in the y-direction, units are $10^5\,\mathrm{Pa\,m^{-1}}$. SG values are plotted as triangles at Ro = 0. Thin line = best linear fit in Ro to UM solutions on 121×61 grid. Thick line = best linear fit in Ro to UM solutions on 201×121 grid. Individual differences plotted as + signs (low resolution) and asterisks (high resolution), using two verification times for each Ro. From Cullen (2007).

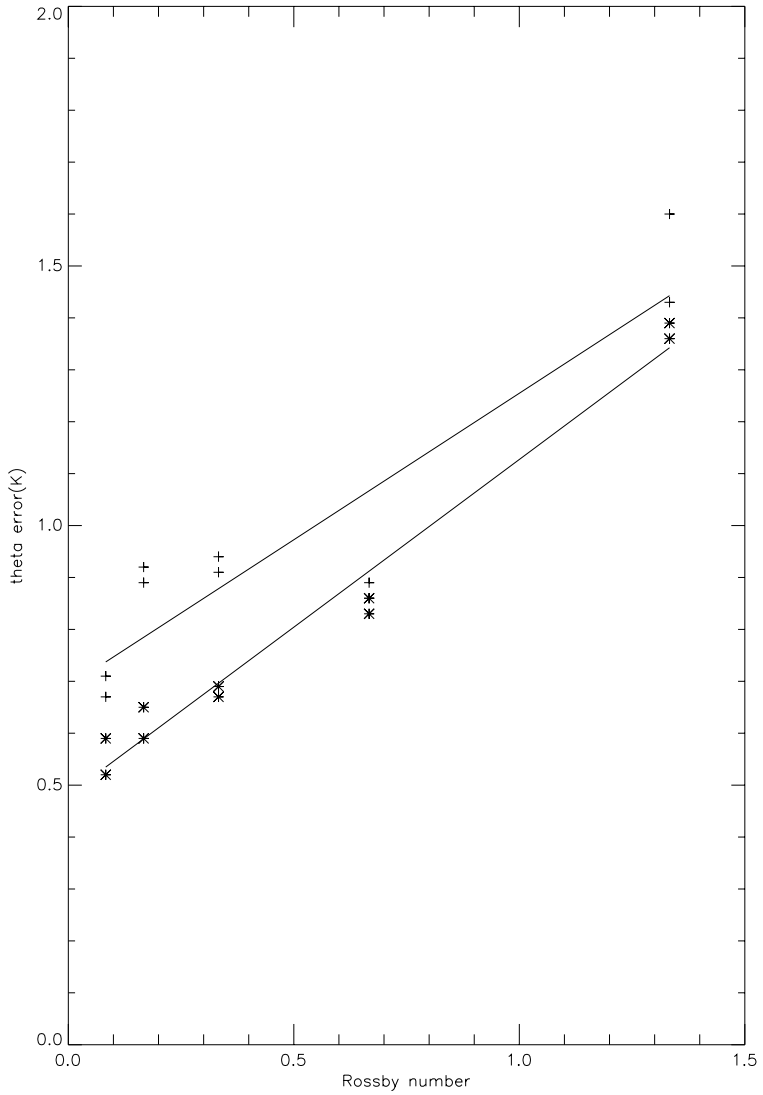

Figure 4.12. Convergence of r.m.s. differences of θ (deg K) between UM and SG solutions for fixed L_R, plotted against Rossby number. Resolutions and plotting conventions as in Figure 4.10. From Cullen (2007).

4.5. Validation for small-scale flow over a ridge

In this subsection we compare UM solutions for the flow over a ridge problem treated in Section 4.4 with solutions of an anelastic version of the UM. The aim is to demonstrate convergence of the UM to an anelastic solution at the predicted rate. The UM equations for this problem are again (4.14). The anelastic version is obtained by first defining initial data satisfying the hydrostatic equation and equation of state:

$$\begin{aligned} C_p\theta(0,\cdot)\frac{\partial\Pi(0,\cdot)}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}+g=0,\\ \Pi(0,\cdot)=(p(0,\cdot)/p_{\rm ref})^{-R/C_p},\\ p(0,\cdot)=\rho R\theta(0,\cdot)\Pi(0,\cdot). \end{aligned} \tag{4.20}$$

These data are chosen to have zero horizontal pressure gradient, as in Section 4.4. However, there is a non-zero gradient along the terrain-following coordinate surfaces. The resulting $\rho(0,\cdot)$ is used as a reference state, as in the derivation of (3.37). The equations are then as follows:

$$\begin{aligned} \frac{\mathrm{D}u}{\mathrm{D}t}-fv+C_p\theta\left(\frac{\partial\Pi}{\partial x}-\frac{\partial\Pi}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}\frac{\partial z}{\partial x}\right)=0,\\ \frac{\mathrm{D}v}{\mathrm{D}t}+fu+C_p\theta\frac{\partial\Pi_1}{\partial y}=0,\\ \frac{\mathrm{D}w}{\mathrm{D}t}+C_p\theta\frac{\partial\Pi}{\partial\eta}\left(\frac{\partial z}{\partial\eta}\right)^{-1}+g=0,\\ \nabla\cdot\left(\rho(0,\cdot)\frac{\partial z}{\partial\eta}(u,\dot\eta)\right)=0,\\ \frac{\mathrm{D}\theta}{\mathrm{D}t}+v\frac{\partial\theta_1}{\partial y}=0. \end{aligned} \tag{4.21}$$

Given data at time t, the Exner pressure Π is calculated from a Poisson equation derived by enforcing the continuity equation at time $t+\delta t$. This is a slight modification of the Helmholz equation derived by Davies *et al.* (2005) as part of the semi-implicit integration scheme in the UM. The resulting pressure is purely diagnostic and may not represent a realistic thermodynamic pressure in places where the anelastic approximation is inaccurate. It would be possible to derive a density from it using the equation of state, and use this density in the continuity equation (the fourth equation of (4.21)). However, this could cause unrealistic behaviour when the pressure is inaccurate, which is why a reference density has to be used in (3.37) to ensure energetic consistency.

The tests are carried out using the same geometry as in Section 4.4. The length-scale is reduced by a factor of 100 to give a small-scale regime. The ridge then has an O(1) aspect ratio. The length of the domain is now 20 km, and the width of the ridge is 1.5 km. Values of u_g at $z = 0$ were chosen from $10\,\mathrm{m\,s^{-1}}$ up to $80\,\mathrm{m\,s^{-1}}$, giving Mach numbers up to 0.25. As in Section 4.4, the experiments are run for a time inversely proportional to u_g. It is not useful to measure the errors in the initial evolution because of rapid transients, so the errors have to be measured after a period of evolution. The aim is to allow time for the adjustment to a quasi non-divergent solution, but to use a time short enough for geostrophic adjustment to have a negligible effect. Results are thus shown for times corresponding to 27 and 36 minutes for $u_g = 10\,\mathrm{m\,s^{-1}}$. Two spatial resolutions were used, of 121×61 and 201×121 points.

In the first set of experiments, N^2 is set to 0.01, as in Section 4.4. Figure 4.13 shows the errors in the u field. These are consistent with quadratic convergence except for very small Mach number where the differences asymptote to a non-zero value. This does not appear to be a function of spatial resolution, though the differences at larger Mach number are sensitive to the resolution.

Figure 4.14 shows the differences in the θ-field. These are more sensitive to resolution, suggesting that numerical errors are also contributing. The results are again consistent with a second-order increase in error with Mach number, starting from a non-zero basic value.

In order to establish whether the residual error at small Mach number is due to the effect of the non-constant θ in the initial data, we repeat the experiments with $N^2 = 0.0001$. Retention of a non-zero value of N^2 allows the effect on the convergence of θ to be estimated. Figure 4.15 shows the errors in the u field, using the same scale as Figure 4.13. These are now consistent with quadratic convergence to zero even for the smallest Mach numbers plotted, and show very little sensitivity to resolution.

Figure 4.16 shows the differences in the θ-field. The scale is magnified by a factor of 100, to allow for the reduced variations in θ in the initial data. There is still a residual non-zero value of the error for small Mach number, but it is 3 times smaller than in the case with N^2 =0.01. The results remain quite sensitive to the resolution, as in the more strongly stratified case.

Overall the results are consistent with the error estimate in Section 3.6. The non-convergence for non-uniform θ means that the use of anelastic equations for quantitative predictions should be treated with caution, as should numerical methods which employ a reference state, unless the effects of the reference state are removed by iteration as discussed in Section 3.2.

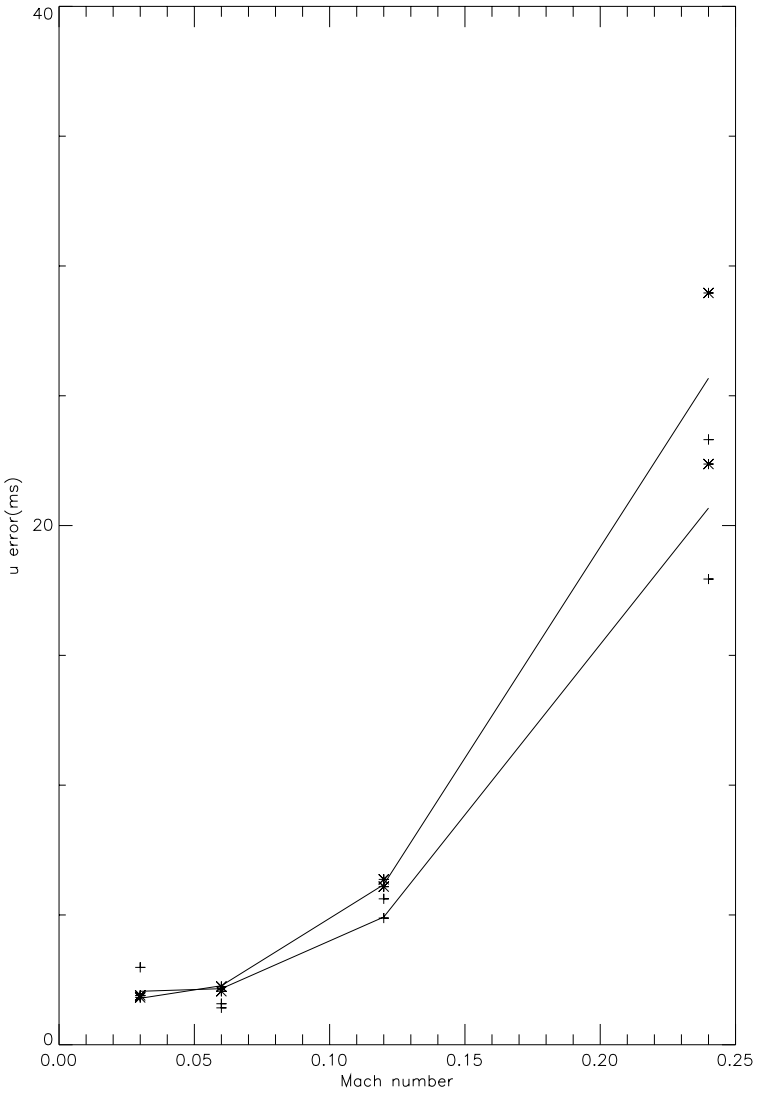

Figure 4.13. Convergence of r.m.s. differences of u ($\mathrm{m\,s^{-1}}$) between UM and anelastic solutions, plotted against Mach number Ma. Thin line = best fit by quadratic polynomial in Ma to differences between UM and anelastic solutions on 121×61 grid. Individual differences plotted as + symbols, using two verification times for each Ma. Thick line = best quadratic fit in Ma to differences between UM solutions on 201×121 grid and anelastic solutions. Individual differences plotted as asterisks, using two verification times for each Ma.

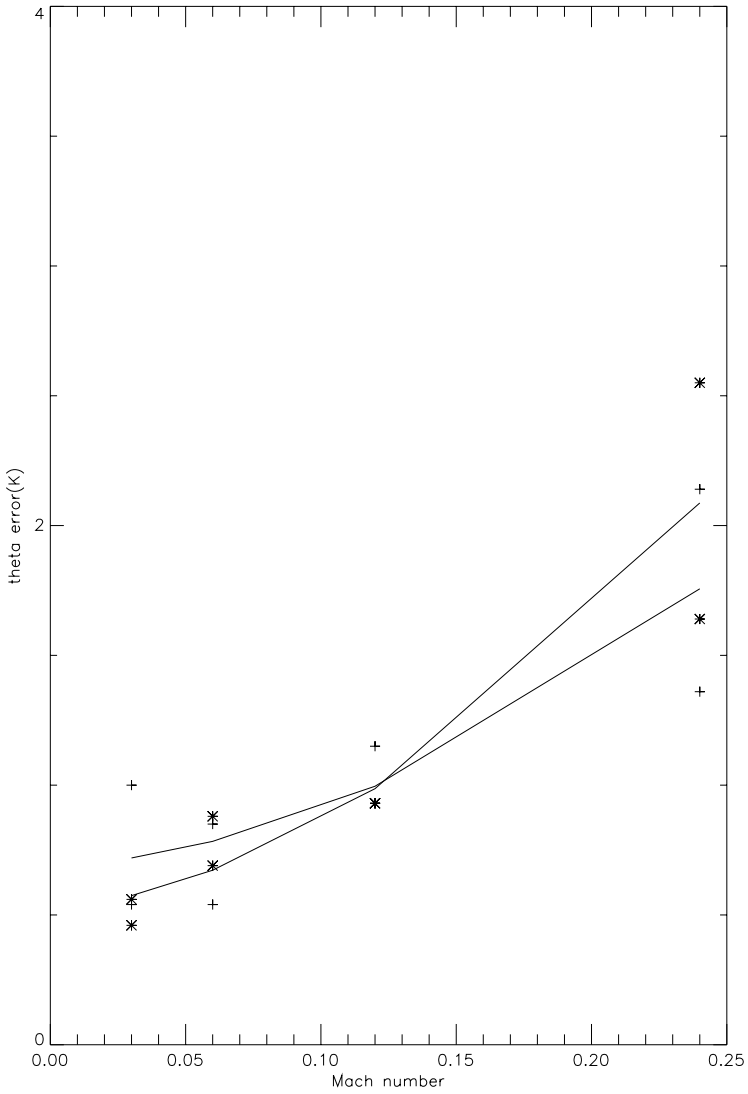

Figure 4.14. Convergence of r.m.s. differences of θ (deg K) between UM and anelastic solutions, plotted against Mach number. Resolutions and plotting conventions as in Figure 4.13.

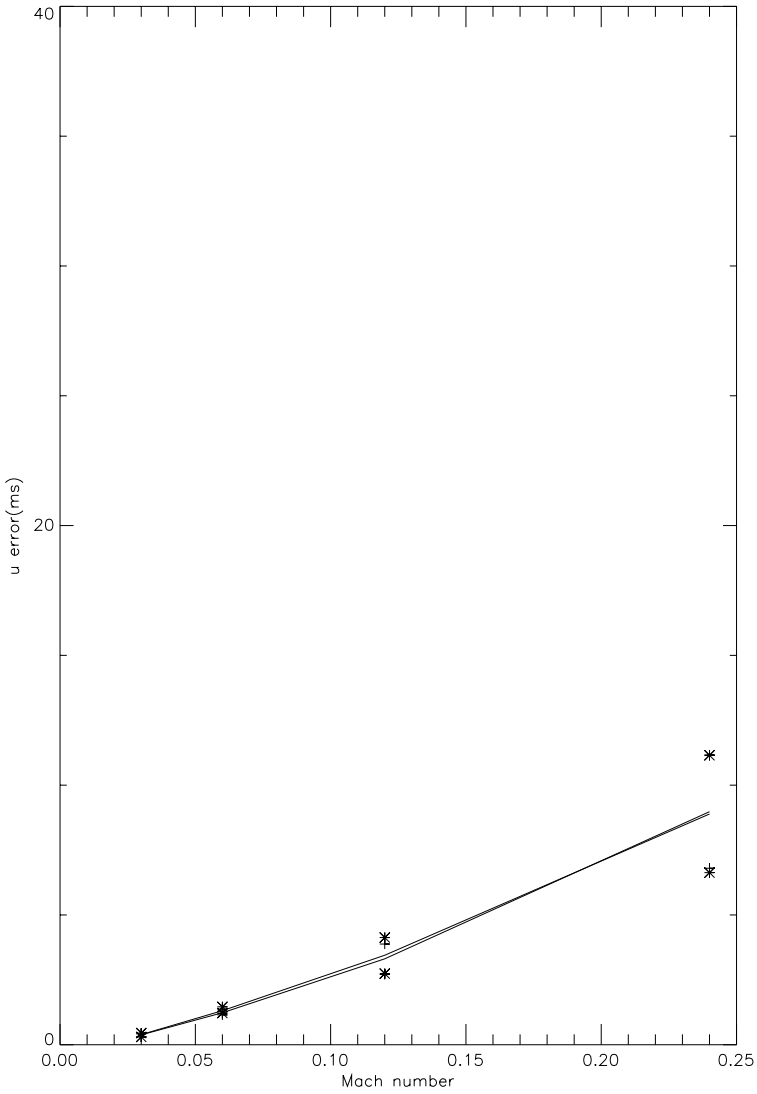

Figure 4.15. Convergence of r.m.s. differences of u ($\mathrm{m\,s^{-1}}$) between UM and anelastic solutions, with $N^2 = 10^{-4}$, plotted against Mach number Ma. Resolutions and plotting conventions as in Figure 4.13.

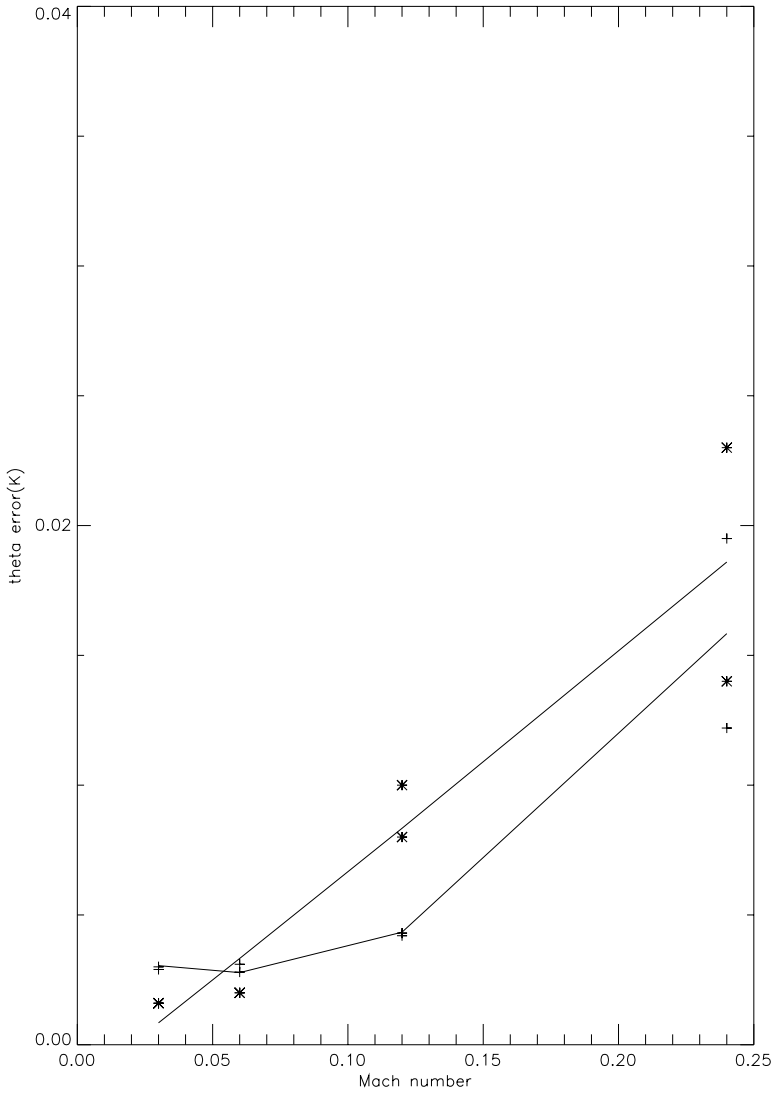

Figure 4.16. Convergence of r.m.s. differences of θ (deg K) between UM and anelastic solutions, with $N^2 = 10^{-4}$, plotted against Mach number. Resolutions and plotting conventions as in Figure 4.13.

5. Discussion

The theme of this paper is that numerical methods for atmosphere models need to be designed and evaluated using asymptotic limit solutions, rather than exact solutions of the governing equations. This is because production models fall short of adequately resolving the exact solution by several orders of magnitude. A related theme is that substantial research is needed on methods of averaging the equations. This has not been reviewed comprehensively in this paper but some key pieces of work have been discussed. An important issue is the choice between Eulerian and Lagrangian averaging.

A number of the most important asymptotic limit solutions have been described and their validity illustrated in computations. The emphasis has been on nonlinear regimes, since analysis of linear regimes is well represented in the existing literature. The related optimal numerical methods have been discussed. Not surprisingly, these mostly reflect current practice. The use of asymptotic limit solutions allows systematic justification for these choices. It should also allow long-term error estimates to be made for suitable numerical methods. There has been little work on that topic so far.

The numerical tests, with the exception of the Eady problem, show the expected convergence rates but to non-zero residuals. Some of these residuals represent dependencies on the initial data or additional parameters defining the flow. The others appear to represent a combination of numerical errors, initialization issues, and uncertainty in predicting turbulent flows. If these residuals are only significant for values of the dimensionless parameters (Rossby and Mach numbers in the examples presented) which are smaller than typical values in real data, then they will not be very significant. An issue with the larger-scale tests, which requires further study using observations, is to determine how prevalent symmetric instability actually is in the atmosphere and ocean on various scales, and whether this is correctly reflected in the solutions given by production models.

The failure of either model to predict the Eady solution after the initial life-cycle is disappointing, since this is the example most relevant to long-term prediction of weather systems. The results suggest that the Lagrangian conservation laws enforced by the geometric model are important in retaining long-term accuracy. However, it is difficult to see how they can be enforced in a production model. This remains an outstanding research challenge. The example does, however, suggest the importance of Lagrangian averaging in describing the physics of fronts.

Acknowledgements

The author would like to thank Nigel Wood for critically reading the manuscript and providing a lot of useful input. He would also like to thank John Thuburn for useful comments, and Marek Wlasak for providing Figures 3.1, 3.2 and 4.1.

REFERENCES

D. G. Andrews and M. E. McIntyre (1978) 'An exact theory of nonlinear waves on a Lagrangian-mean flow', *J. Fluid Mech.* **89**, 609–646.

A. Arakawa and C. S. Konor (1996) 'Vertical differencing of the primitive equations based on the Charney–Phillips grid in hybrid $\sigma - p$ vertical coordinates', *Mon. Weather Rev.* **124**, 511–528.

A. Arakawa and V. R. Lamb (1977) 'Computational design of the basic dynamical processes of the UCLA general circulation model', *Methods in Comput. Phys.* **17**, 173–265.

G. K. Batchelor (1967) *An Introduction to Fluid Dynamics*, Cambridge University Press.

J. R. Bates, S. Moorthi and R. W. Higgins (1993) 'A global multilevel atmospheric model using a vector semi-Lagrangian finite-difference scheme, Part I: Adiabatic formulation', *Mon. Weather Rev.* **121**, 244–263.

J.-D. Benamou and Y. Brenier (1998) 'Weak existence for the semi-geostrophic equations formulated as a coupled Monge–Ampère/transport problem', *SIAM J. Appl. Math.* **58**, 1450–1461.

A. F. Bennett and P. E. Kloeden (1981) 'The quasi-geostrophic equations: Approximation, predictability and equilibrium spectra', *Quart. J. Roy. Meteorol. Soc.* **107**, 121–136.

A. F. Bennett and P. E. Kloeden (1982) 'The periodic quasi-geostrophic equations: Existence and uniqueness of strong solutions', *Proc. Roy. Soc. Edinburgh* **91A**, 185–203.

A. J. Bourgeois and J. T. Beale (1994) 'Validity of the quasi-geostrophic model for large-scale flow in the atmosphere and ocean', *SIAM J. Math. Anal.* **25**, 1023–1068.

Y. Brenier (1991) 'Polar factorization and monotone rearrangement of vector-valued functions', *Comm. Pure Appl. Math.* **44**, 375–417.

G. Browning and H.-O. Kreiss (1994) 'Splitting methods for problems with different time-scales', *Mon. Weather Rev.* **122**, 2614–2622.

K. Bryan (1989) The design of numerical models of the ocean circulation. In *Oceanic Circulation Models: Combining Data and Dynamics* (*Proc. NATO ASI*, Les Houches, France, February 1988), pp. 465–500.

C. S. Cao and E. S. Titi (2005) Global well-posedness of the three-dimensional viscous primitive equations of large-scale ocean and atmosphere dynamics. arXiv.math AP/0503028.

J. G. Charney (1948) 'On the scale of atmospheric motions', *Geofys. Publ.* **17**, 1–17.

J.-Y. Chemin (2000) *Perfect Incompressible Fluids*, Oxford University Press.

P. Courtier, J.-N. Thépaut and A. Hollingsworth (1994) 'A strategy for operational implementation of 4D-Var, using an incremental approach', *Quart. J. Roy. Meteorol. Soc.* **120**, 1367–1387.

M. J. P. Cullen (2000) 'On the accuracy of the semi-geostrophic approximation', *Quart. J. Roy. Meteorol. Soc.* **126**, 1099–1115.

M. J. P. Cullen (2001) 'Alternative implementations of the semi-Lagrangian semi-implicit scheme in the ECMWF model', *Quart. J. Roy. Meteorol. Soc.* **127**, 2787–2802.

M. J. P. Cullen (2002*a*) 'Large scale non-turbulent dynamics in the atmosphere', *Quart. J. Roy. Meteorol. Soc.* **128**, 2623–2640.

M. J. P. Cullen (2002*b*) New mathematical developments in atmosphere and ocean dynamics, and their application to computer simulations. In *Large-Scale Atmosphere-Ocean Dynamics* (J. Norbury and I. Roulstone eds), Vol. I, Cambridge University Press, pp. 202–287.

M. J. P. Cullen (2003) 'Four dimensional variational data assimilation: A new formulation of the background error covariance matrix, based on a potential vorticity representation', *Quart. J. Roy. Meteorol. Soc.* **129**, 2777–2796.

M. J. P. Cullen (2006) *A Mathematical Theory of Large-Scale Atmospheric Flow*, Imperial College Press.

M. J. P. Cullen (2007) 'Semi-geostrophic solutions for flow over a ridge', *Quart. J. Roy. Meteorol. Soc.*, to appear.

M. J. P. Cullen and T. Davies (1990) 'A conservative split-explicit scheme with fourth-order horizontal advection', *Quart. J. Roy. Meteorol. Soc.* **117**, 993–1002.

M. J. P. Cullen and M. Feldman (2006) 'Lagrangian solutions of semi-geostrophic equations in physical space', *SIAM J. Math. Anal.* **37**, 1371–1395.

M. J. P. Cullen and H. Maroofi (2003) 'The fully compressible semi-geostrophic system from meteorology', *Arch. Rat. Mech. Anal.* **167**, 309–336.

M. J. P. Cullen and I. Roulstone (1993) 'A geometric model of the nonlinear equilibration of two-dimensional Eady waves', *J. Atmos. Sci.* **50**, 328–332.

M. J. P. Cullen, R. J. Douglas, I. Roulstone and M. J. Sewell (2005) 'Generalized semi-geostrophic theory on a sphere', *J. Fluid Mech.* **531**, 123–157.

M. J. P. Cullen, J. Norbury, R. J. Purser and G. J. Shutts (1987) 'Modelling the quasi-equilibrium dynamics of the atmosphere', *Quart. J. Roy. Meteorol. Soc.* **126**, 735–757.

T. Davies, M. J. P. Cullen, A. J. Malcolm, M. H. Mawson, A. Staniforth, A. A. White and N. Wood (2005) 'A new dynamical core for the Met Office's global and regional modelling of the atmosphere', *Quart. J. Roy. Meteorol. Soc.* **131**, 1759–1782.

T. Davies, A. Staniforth, N. Wood and J. Thuburn (2003) 'Validity of anelastic and other equation sets as inferred from normal-mode analysis', *Quart. J. Roy. Meteorol. Soc.* **129**, 2761–2775.

M. Diamantakis, T. Davies and N. Wood (2007) An iterative time-stepping scheme for the Met Office's semi-implicit semi-Lagrangian non-hydrostatic model', *Quart. J. Roy. Meteorol. Soc.*, to appear.

J. A. Domaradzki, Z. Xiao and P. K. Smolarkiewicz (2003) 'Effective eddy viscosities in implicit large eddy simulations of turbulent flows', *Phys. Fluids* **15**, 3890–3893.

D. Durran (1998) *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, Springer.

ECMWF (2002) *Key Issues in the Parametrisation of Sub-Grid Physical Processes* (*Proc. ECMWF Seminar*, September 2001).

ECMWF (2005) *Representation of Sub-Grid Processes Using Stochastic–Dynamic Models* (*Proc. ECMWF Workshop*, June 2005).

J. H. Ferziger (1998) Direct and large eddy simulations of turbulence. In *Numerical Methods in Fluid Mechanics* (A. Vincent, ed.), Vol. 16 of *CRM Proceedings and Lecture Notes*, Université de Montreal, AMS.

C. Foias, D. D. Holm and E. S. Titi (2001) 'The Navier–Stokes-alpha model of fluid turbulence', *Physica D* **152**, 505–519.

R. Ford, M. E. McIntyre and W. A. Norton (2000) 'Balance and the slow quasi-manifold: Some explicit results', *J. Atmos. Sci.* **57**, 1236–1254.

M. S. Fox-Rabinowitz (1996) 'Diabatic dynamical initialization with an iterative time integration scheme as a filter', *Mon. Weather Rev.* **124**, 1544–1557.

C. Fureby and F. F. Grinstein (2002) 'Large eddy simulation of high-Reynolds number free and wall-bounded flows', *J. Comput. Phys.* **181**, 68–97.

V. G. Ganzha (1996) *Numerical Solutions for Partial Differential Equations*, CRC Press.

J. R. Garratt (1992) *The Atmospheric Boundary Layer*, Cambridge University Press.

P. R. Gent and J. C. McWilliams (1996) 'Eliassen–Palm fluxes and the momentum equation in non-eddy-resolving ocean circulation models', *J. Phys. Oceanog.* **26**, 2539–2546.

A. E. Gill (1982) *Atmosphere-Ocean Dynamics*, Academic Press.

R. L. Higdon (2006) Numerical modelling of ocean circulation. In *Acta Numerica*, Vol. 15, Cambridge University Press, pp. 385–470.

D. D. Holm (1996) 'Hamiltonian balance equations', *Physica D* **98**, 379–414.

D. D. Holm, J. E. Marsden and T. Ratiu (2002) The Euler–Poincaré equations in geophysical fluid dynamics. In *Large-Scale Atmosphere-Ocean Dynamics* (J. Norbury and I. Roulstone, eds), Vol. II, Cambridge University Press, pp. 251–300.

M. W. Holt (1990) 'Semi-geostrophic moist frontogenesis in a Lagrangian model', *Dyn. Atmos. Ocean.* **14**, 463–481.

B. J. Hoskins (1975) 'The geostrophic momentum approximation and the semi-geostrophic equations', *J. Atmos. Sci.* **32**, 233–242.

J. Houghton (2002) *The Physics of Atmospheres*, 3rd edn, Cambridge University Press.

E. Kalnay (2003) *Atmospheric Modelling, Data Assimilation and Predictability*, Cambridge University Press.

R. M. Kerr (1993) 'Evidence for a singularity of the three-dimensional incompressible Euler equations', *Phys. Fluids A* **5**, 1725–1746.

A. S. Lawless, N. K. Nichols and S. P. Ballard (2003) 'A comparison of two methods for developing the linearization of a shallow-water model', *Quart. J. Roy. Meteorol. Soc.* **129**, 1237–1254.

C. E. Leith and R. E. Kraichnan (1972) 'Predictability of turbulent flows', *J. Atmos. Sci.* **29**, 1041–1052

M. J. Lighthill (1978) *Waves in Fluids*, Cambridge University Press.

S.-J. Lin and R. B. Rood (1997) 'An explicit flux-form semi-Lagrangian shallow water model on the sphere', *Quart. J. Roy. Meteorol. Soc.* **123**, 2477–2498.

J.-L. Lions, R. Temam and S. Wang (1992*a*) 'New formulations of the primitive equations of atmosphere and applications', *Nonlinearity* **5**, 237–288.

J.-L. Lions, R. Temam and S. Wang (1992*b*) 'On the equations of the large-scale ocean', *Nonlinearity* **5**, 1007–1053.

F. Lipps and R. Hemler (1982) 'A scale analysis of deep moist convection and some related numerical calculations', *J. Atmos. Sci.* **29**, 2192–2210.

P. Lynch (1989) 'The slow equations', *Quart. J. Roy. Meteorol. Soc.* **115**, 201–219.

P. Lynch (1997) 'The Dolph–Chebyshev window: A simple optimal filter', *Mon. Weather Rev.* **125**, 1976–1982.

B. Machenhauer (1977) 'On the dynamics of gravity oscillations in a shallow water model with applications to normal-mode initialization', *Contrib. Atmos. Phys.* **50**, 253–271.

R. J. McCann (2001) 'Polar factorization of maps on Riemannian manifolds', *Geom. Funct. Anal.* **11**, 589–608.

M. E. McIntyre and W. A. Norton (2000) 'Potential vorticity inversion on a hemisphere', *J. Atmos. Sci.* **57**, 1214–1235.

J. C. McWilliams and I. Yavneh (1998) 'Fluctuation growth and instability associated with a singularity of the balance equations', *Phys. Fluids* **10**, 2587–2596.

J. C. McWilliams, I. Yavneh, M. J. P. Cullen and P. R. Gent (1999) 'The breakdown of large-scale flows in rotating, stratified fluids', *Phys. Fluids* **10**, 3178–3184.

A. J. Majda (1984) *Compressible Fluid Flow, and Systems of Conservation Laws in Several Space Variables*, Vol. 53 of *Applied Mathematical Sciences*, Springer, New York.

A. J. Majda (2003) *Introduction to PDEs and Waves for the Atmosphere and Ocean*, Vol. 9 of *Courant Lecture Notes*, AMS.

A. J. Majda and A. L. Bertozzi (2002) *Vorticity and Incompressible Flow*, Cambridge University Press.

M. H. Mawson (1996) 'A shallow water semi-geostrophic model on a sphere', *Quart. J. Roy. Meteorol. Soc.* **122**, 267–290.

A. R. Mohebalhojeh and D. G. Dritschel (2001) 'Hierarchies of balance conditions for the f-plane shallow water equations', *J. Atmos. Sci.* **58**, 2411–2426.

A. R. Mohebalhojeh and D. G. Dritschel (2004) 'Contour advective semi-Lagrangian algorithms for many-layer primitive-equation models', *Quart. J. Roy. Meteorol. Soc.* **130**, 347–364.

N. Nakamura (1994) 'Nonlinear equilibration of two-dimensional Eady waves: Simulations with viscous geostrophic momentum equations', *J. Atmos. Sci.* **51**, 1023–1035.

N. Nakamura and I. M. Held (1989) 'Nonlinear equilibration of two-dimensional Eady waves', *J. Atmos. Sci.* **46**, 3055–3064.

Y. Ogura and N. A. Phillips (1962) 'Scale analysis for deep and shallow convection in the atmosphere', *J. Atmos. Sci.* **19**, 173–179.

J. Pedlosky (1987) *Geophysical Fluid Dynamics*, Springer.

H. Reiser (2000) The development of numerical weather prediction in Deutsche Wetterdienst. In *Proc. 50th Anniversary of Numerical Weather Prediction*, Deutsche Meteorologische Gesellschaft, pp. 51–80.

H. Ritchie, C. Temperton, A. J. Simmons, M. Hortal, T. Davies, D. W. Dent and M. Hamrud (1995) 'Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model', *Mon. Weather Rev.* **123**, 489–514.

A. Robert (1981) 'A stable numerical integration scheme for the primitive meteorological equations', *Atmos. Ocean* **19**, 35–46.

A. Robert (1982) 'A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations', *J. Meteorol. Soc. Japan* **60**, 319–324.

R. Sadourny (1975) 'The dynamics of finite-difference models of the shallow-water equations', *J. Atmos. Sci.* **32**, 680–689.

W. H. Schubert (1985) 'Semi-geostrophic theory', *J. Atmos. Sci.* **42**, 1770–1772.

G. J. Shutts (1998) 'Idealized models of the pressure drag force on mesoscale mountain ridges', *Contrib. Atmos. Phys.* **71**, 303–313.

G. J. Shutts and M. J. P. Cullen (1987) 'Parcel stability and its relation to semi-geostrophic theory', *J. Atmos. Sci.* **44**, 1318–1330.

J. Smagorinsky (1974) Global atmospheric modelling and the numerical simulation of climate. In *Weather and Climate Modification*, Wiley, New York, pp. 633–686.

R. K. Smith, ed., (1997) *The Physics and Parameterization of Moist Atmospheric Convection*, Kluwer, Dordrecht.

A. Staniforth and J. Côté (1991) 'Semi-Lagrangian integration scheme for atmospheric models: A review', *Mon. Weather Rev.* **119**, 2206–2223.

A. Staniforth, A. A. White, N. Wood, J. Thuburn, M. Zerroukat and E. Cordero (2002) Unified Model Documentation Paper No. 15: The Joy of UM 6.1: Model Formulation. Unpublished documentation, available from: www.metoffice.gov.uk/research/nwp/publications/papers/unified_model/index.html.

J. Thuburn (2006) 'Vertical discretizations giving optimal representation of normal modes: Sensitivity to the form of the pressure gradient term', *Quart. J. Roy. Meteorol. Soc.* **132**, 2809–2826.

J. Thuburn, N. Wood and A. Staniforth (2002*a*) 'Normal modes of deep atmospheres I: Spherical geometry', *Quart. J. Roy. Meteorol. Soc.* **128**, 1771–1792.

J. Thuburn, N. Wood and A. Staniforth (2002*b*) 'Normal modes of deep atmospheres II: $f - F$-plane geometry', *Quart. J. Roy. Meteorol. Soc.* **128**, 1793–1806.

T. Warn, O. Bokhove, T. G. Shepherd and G. K. Vallis (1995) 'Rossby-number expansions, slaving principles, and balance dynamics', *Quart. J. Roy. Meteorol. Soc.* **121**, 723–739.

T.-K. Wee and Y.-H. Kuo (2004) 'Impact of a digital filter as a weak constraint in MM5 4D-Var: An observing system simulation experiment', *Mon. Weather Rev.* **132**, 543–559.

A. A. White (2002) The equations of meteorological dynamics and various approximations. In *Large-Scale Atmosphere-Ocean Dynamics* (J. Norbury and I. Roulstone, eds), Vol. I, Cambridge University Press, pp. 1–100.

A. A. White (2003) 'Dynamical equivalence and the departure-point equation in semi-Lagrangian numerical models', *Quart. J. Roy. Meteorol. Soc.* **129**, 1317–1324.

A. A. White, B. J. Hoskins, I. Roulstone and A. Staniforth (2005) 'Consistent approximate models of the global atmosphere: Shallow, deep, hydrostatic, quasi-hydrostatic and non-hydrostatic', *Quart. J. Roy. Meteorol. Soc.* **131**, 2081–2108.

N. Wood and A. Staniforth (2003) 'The deep atmosphere Euler equations with a mass-based vertical coordinate', *Quart. J. Roy. Meteorol. Soc.* **129**, 1289–1300.

K.-S. Yeh, J. Côté, S. Gravel, A. Méthot, A. Patoine, M. Roch and A. Staniforth (2002) 'The CMC–MRB Global Environmental Multiscale (GEM) model, Part III: Non-hydrostatic formulation', *Mon. Weather Rev.* **130**, 339–356.

M. Zerroukat, N. Wood and A. Staniforth (2004) 'SLICE-S: A Semi-Lagrangian Inherently Conserving and Efficient scheme for transport problems on the sphere', *Quart. J. Roy. Meteorol. Soc.* **130**, 2649–2664.

M. Zerroukat, N. Wood and A. Staniforth (2005) 'A monotonic and positive-definite filter for a Semi-Lagrangian Inherently Conserving and Efficient (SLICE) scheme', *Quart. J. Roy. Meteorol. Soc.* **131**, 2923–2936.

M. J. Ziemianski and A. J. Thorpe (2003) 'Nonlinear balanced models for stratified fluids conserving Ertel–Rossby PV', *Quart. J. Roy. Meteorol. Soc.* **129**, 139–156.