3 Numerical solutions to the equations

The current chapter summarizes various topics related to the numerical solution of the model equations, for resolvable scales of motion. This part of an atmospheric model that treats the resolvable scales is called the *dynamical core*, and is distinct from the representations of the subgrid-scale, parameterized physical processes. An especially important topic is how the numerical approximations that are used to solve the equations can affect the model solution. These nonphysical effects should be thoroughly understood by all model users. Even though basic concepts are described here, and examples provided, this presentation of numerical methods is far from exhaustive. A comprehensive text on this subject, such as Durran (1999), should be consulted if more depth is needed. Step-by-step derivations are frequently left to the reader.

Numerical methods used for solving the equations have naturally evolved over the last few decades, partly because of the results of research and partly because of changes in the available computational resources. Various factors are involved in the decision about the numerical methods to use for a particular modeling application, including computational efficiency (speed), accuracy, memory requirements, and code-structure simplicity. The last factor is especially important if the model is going to be used for research, especially by students. Simple methods that are not typically used in current operational models are sometimes described here for pedagogical purposes.

3.1 Overview of basic concepts

The following brief overview of concepts will help the reader to better understand the specialized material in later sections.

3.1.1 Grid-point and spectral methods for representing spatial variations of the atmosphere

The model equations are often solved at points defined by a quasi-regular, three-dimensional spatial grid. Section 3.2.1 reviews the different options for the structure of these grids. The term "quasi-regular" is used here as an acknowledgment that the points are typically not exactly equally spaced, when the grid is defined on a map projection where the Earth-distance between grid points varies from place to place. Sometimes the points are very nonuniformly spaced, for example when using latitude–longitude coordinates or with adaptive grids where the resolution is increased in areas of strong gradients.

The time axis is also defined by discrete, evenly spaced points. The time and space derivatives in the equations can be approximated using finite-difference methods (Section 3.3), which introduce nonphysical properties to the model solution, and have stability criteria that limit the time step (Section 3.4). As an example of solving an equation on a grid, the first equation of motion (Eq. 2.1) will be represented using a simple three-point centered-difference approximation in time and space, such as

$$\frac{\partial}{\partial y} f(x,y,z,t) = \frac{f^{\tau}_{i,j+1,k} - f^{\tau}_{i,j-1,k}}{y(j+1) - y(j-1)} = \frac{f^{\tau}_{j+1} - f^{\tau}_{j-1}}{2\Delta y},$$

where f is any dependent variable; τ defines a discrete point on the time axis; i, j, k define coordinates on the x, y, z space axes, respectively; and Δy is the distance between two adjacent points on the y axis. The first equation of motion, Eq. 2.16,

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - w\frac{\partial u}{\partial z} - \frac{1}{\rho}\frac{\partial p}{\partial x} + fv + Fr_x,$$

is a nonlinear, nonhomogeneous, partial differential equation that cannot be solved analytically. With the above, three-point, finite-difference approximations, it becomes the following solvable arithmetic equation

$$\frac{u^{\tau+1}_{i,j,k} - u^{\tau-1}_{i,j,k}}{2\Delta t} = -u^{\tau}_{i,j,k}\frac{u^{\tau}_{i+1,j,k} - u^{\tau}_{i-1,j,k}}{2\Delta x} - v^{\tau}_{i,j,k}\frac{u^{\tau}_{i,j+1,k} - u^{\tau}_{i,j-1,k}}{2\Delta y}$$

$$-w^{\tau}_{i,j,k}\frac{u^{\tau}_{i,j,k+1} - u^{\tau}_{i,j,k-1}}{2\Delta z} - \frac{1}{\rho^{\tau}_{i,j,k}}\frac{p^{\tau}_{i+1,j,k} - p^{\tau}_{i-1,j,k}}{2\Delta x} + fv^{\tau}_{i,j,k} + Fr^{\tau-1}_{x(i,j,k)}, \quad (3.1)$$

where Δx and Δy are often assumed to be the same, and Fr is a frictional-dissipation term. This equation is solved for $u^{\tau+1}_{i,j,k}$ on the left side, and, for each grid point, the right side is evaluated based on values of the dependent variables from the two previous time levels τ and $\tau - 1$. The other equations are similarly solved for the $\tau + 1$ values of the dependent variable.

The value of Δx is chosen so that there is a sufficient number of grid points to adequately represent the smallest meteorological feature of interest, for the particular application of the model. Section 3.4.1 on the concept of *truncation error* quantifies the accuracy associated with representing continuous functions with a finite number of points. A rule of thumb is that 10 grid points are needed to reasonably resolve a wave. So, depending on whether the purpose of the model is to simulate synoptic-scale Rossby waves or mesoscale convective complexes, the grid increment must be chosen accordingly. An alternative approach is to represent the spatial variation of the dependent variables using global or local functions, and calculate the derivatives analytically. Such approaches include the spectral and finite-element methods described in Sections 3.2.2 and 3.2.3, respectively.

There are two general types of models in terms of the spatial extent of the computational volume. If the model calculations span the sphere, the model is referred to as a *global model*. If the model applies only to a particular regional subvolume of the atmosphere, it is called a *limited-area model.*

3.1.2 The time integration

Section 3.3.1 reviews different methods for integrating the equations in time. The approach shown in Eq. 3.1, for all terms except the friction term, is known as leap-frog, or three-point centered, time differencing because the value of the derivative is calculated at a time (τ, right side of equation) that is centered between the initial ($\tau - 1$) and final ($\tau + 1$) times of the extrapolation. Figure 3.1 illustrates this time-differencing method. Note that a forward time step is required at the beginning of the integration, before the leap-frog process can be used. For the friction term, forward differencing is used, where the derivative is calculated at the point from which the extrapolation originates. This is the only method that is stable. For many differencing schemes, the time step is constrained by a limiting value of the *Courant number*, defined as $U\Delta t/\Delta x$, where U is the horizontal speed of the fastest wave on the grid, and Δx was chosen, as described earlier, to allow resolution of the relevant meteorological processes. If the time step is too long, a stability criterion is violated, nonmeteorological features grow exponentially in the solution, and floating-point overflows in the computer will cause the integration to stop. The stability requirement for the advection term represented in an Eulerian framework (the equations are solved at grid points), for the combination of space and time differencing methods used in Eq. 3.1, is called the Courant–Friedrichs–Lewy (CFL) criterion, which requires that $U\Delta t/\Delta x \leq 1$. The concept of numerical stability is further developed in Section 3.4.2.

In order to understand what controls the computational requirements for running a model, assume that a forecast is needed over a certain limited geographic area. Also, for simplicity, assume that the grid points are regularly spaced in the horizontal. The horizontal grid increment is chosen such that the features of meteorological interest are well represented by a sufficient number of points over the length of a wave. The chosen vertical distribution of points will similarly depend on the vertical structures that need to be resolved. Given the types of atmospheric waves admitted by the model equations, it is possible to estimate the fastest wave on the grid. For example, if sound waves and external gravity waves are not part of the model solution, the fastest advective wave (e.g., in a jet

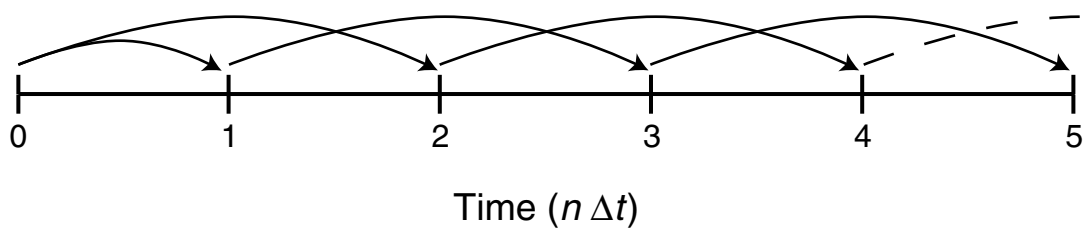

Fig. 3.1 Centered-in-space time-differencing schematic, except for the initial forward-in-time step to extrapolate from $n = 0$ to $n = 1$.

stream) can be estimated based on knowledge of the local atmosphere. Thus, the only remaining parameter in the expression for the Courant number is the time step. This must be chosen such that the numerical solution will remain stable. The cost of running the model on this grid will be based on the number of times the full set of algebraic equations, of which Eq. 3.1 is one, must be solved during the forecast period. This is linearly proportional to the number of calculation points, sometimes called nodes, in the three-dimensional grid, because the equations need to be solved at each point. The cost is also related to the fact that the equations need to be solved for each time step, at each grid point, so the number of time steps in the forecast period will also control the computational requirements. Thus, for a given area, the smaller the grid increment in the vertical and the horizontal, and the smaller the time step, the more computationally demanding the model will be to run. For a given area, a fine-mesh mesoscale model will have many more grid points than will a model designed for synoptic-scale processes, and thus it will be much more costly to run. To reveal the nonlinearity of the dependence of computational cost on resolution, assume that we want to double the horizontal resolution of a grid over a given area. This will require four times as many grid points, and because the stability criterion (e.g., based on the Courant number) needs to be satisfied, the time step will probably need to be halved. Thus, increasing the horizontal resolution by a factor of two will cause an increase in the computational expense by a factor of eight. It is for this reason that NWP research has often focussed on the development of more efficient numerical schemes for solving the equations of motion.

3.1.3 Boundary conditions

Solving the model equations represents both a boundary-value (lateral, upper, lower) problem and an initial-value problem. For a global model, there are no lateral boundaries because the computational area is naturally periodic. For limited-area models (i.e., not global), the equations cannot be solved for points on the edge of the grid because there are no points beyond the boundary to use for evaluating the derivative perpendicular to it (see Section 3.5 for a discussion of lateral-boundary conditions). The values of dependent variables at these boundary points need to be externally specified. For operational forecasting with limited-area models, the lateral-boundary values must be defined by interpolation from grid points of a previously run global forecast model. For research applications, archived, gridded regional or global analyses of observations may be used.

In addition to these lateral-boundary conditions, there are also upper and lower boundary values that must be specified with both global and limited-area models. Because the model atmosphere cannot extend to infinity as does the real atmosphere, and because we sometimes want to limit our computations to the troposphere, it is necessary to define an artificial upper-boundary condition for models. Approaches for doing that, which minimize downward reflections, are discussed in Section 3.6. Another major challenge is defining the fluxes of heat, moisture, and momentum at the land and ocean surface. Because the midlatitude planetary circulation and monsoons are driven by gradients in sensible heating at the surface, it should be obvious that models must treat this process reasonably well. In

addition, mesoscale boundary-layer circulations result from horizontally differential heating at coastlines and at other landscape boundaries, so small-scale variations in sensible-heat fluxes need to be modeled accurately as well. And, the sensible- and latent-heat fluxes compete for the solar-energy input at the surface, so the latent-heat fluxes can greatly influence boundary-layer winds and thermal properties. The modeling of land-surface processes and surface fluxes is discussed in Chapter 5 on land-surface modeling and in Section 4.4 on boundary-layer parameterizations, respectively.

3.1.4 Initial conditions

Because atmospheric modeling is an initial-value problem, the state of the dependent variables at the beginning of the integration of the equations must be specified (the left-most point in Fig. 3.1). This process is called *initializing* the model, and is discussed in Chapter 6. How well this is accomplished has important consequences regarding the accuracy of the forecast. First, except for locally forced processes (e.g., forced by orography, coastlines), it is reasonable to assume that forecast quality can generally be no better than that of the initial conditions. Second, if the mass and momentum fields are far out of balance (e.g., geostrophic) relative to what should prevail for the physical processes as rendered by the model equations, inertia–gravity waves are created by the model fields adjusting after the initialization. These waves, which have no counterpart in the atmosphere, can sometimes have sufficient amplitude to obscure real features in the model solution, at least until they have been damped or have propagated away from the source region. Lastly, if the initial conditions do not contain realistic vertical motions associated with orography or convection, or realistic mesoscale coastal or mountain-valley circulations, the model has to *spin up* these features during the forecast. The above adjustment issues led to the historical situation where forecasters did not use the forecast for the first 12–24 h after initialization.

There are two general types of initializations: *static initializations* and *dynamic initializations.* In the former, observations applicable at the initial time are objectively analyzed to the model grid, perhaps some balance constraint is applied, and the model forecast is begun. These static initializations that do not provide spun-up vertical motions and ageostrophic circulations are referred to as *cold starts*. For the dynamic initialization, the name implies that there is a dynamic constraint, typically from a model, that is aimed at ensuring that the model solution is spun up, or nearly so, at the initial time of the forecast. One approach would be to perform a static initialization, 12–24 h before the desired start time of the forecast, and run the model to allow the solution to spin up during the preforecast period. A variant of this is for the model to assimilate observations during the preforecast period of integration. Or, a common technique is to use an existing spun-up model forecast, which is valid at the initialization time, as the *first guess* for an objective analysis that incorporates observations made within some time window (e.g., perhaps ± 1 h) of the initial time. That is, the observations are used to adjust a model forecast that is valid at the initialization time. For example, Fig. 3.2 illustrates a series of forecasts of 24-h duration, where the forecasts are initialized at a 6-h interval. In this example, it would be said that the model runs with a 6-h forecast cycle. For initialization of the fourth forecast in the cycle, the 12-h forecast

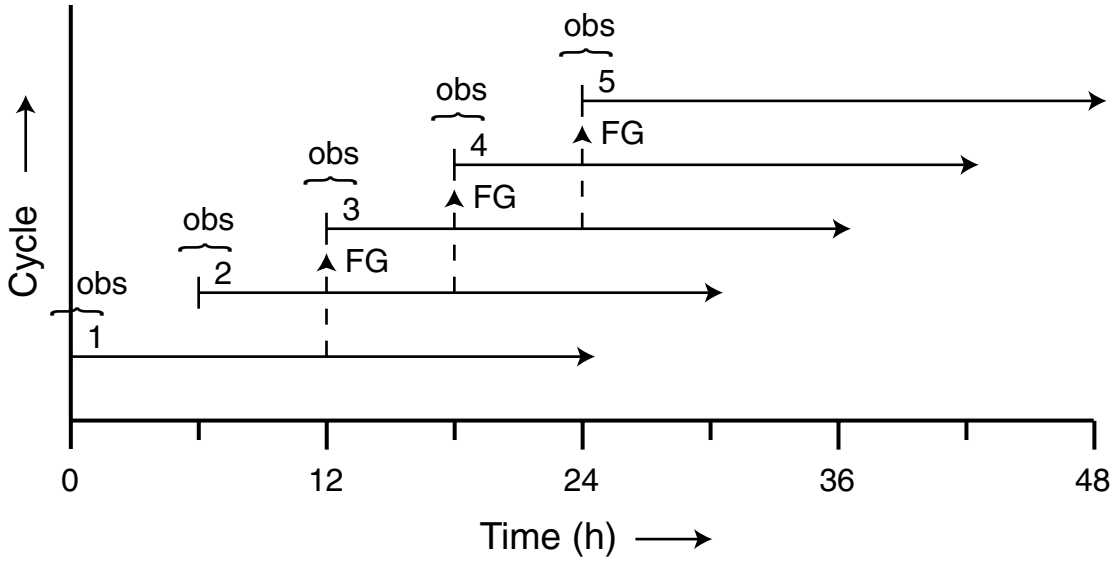

Fig. 3.2 Illustration of forecast cycling, where the model is initialized and a new forecast is launched every 6 h. The vertical dashed lines show that output from a previously initialized forecast is used as a first guess (FG) for an objective analysis that employs observations (obs) that are available within a time window around the initialization time. The cycle number is shown at the initial time of each forecast.

from the second forecast in the series (which should be spun up by that time) could be used as a first guess in an objective analysis of observations. This process by which observations are merged with an analysis at a sequence of times is called *intermittent* or *sequential data assimilation*. That is, data are incorporated into the model intermittently at the initialization times only. Alternatively, there are techniques where observations are ingested by a continuously running data assimilation system, as they become available. This is called *continuous data assimilation*. Initializations using conditions where small-scale circulations are spun up to varying degrees are called *warm starts* or *hot starts*.

Because radiosonde soundings are still the only generally available, and somewhat spatially and temporally regular, sources of three-dimensional atmospheric data over land, they are a primary source of information. Unfortunately, these profiles are many hundreds of kilometers apart, so a typical situation is one in which synoptic-scale processes may be represented reasonably, but those on the mesoscale are not. Even though there are currently many other sources of observational data from satellites, radars, commercial aircraft, etc., the reliance on radiosondes explains why operational forecast models still use 0000 UTC and 1200 UTC as two of the times when forecasts are initiated daily (most operational models employ four cycles per day, initialized at 0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC).

3.1.5 Physical-process parameterizations

The terms in Eqs 2.1–2.7 that represent the effects of turbulent mixing of heat, water vapor, and momentum; moist convection; cloud-microphysical processes; and solar and atmospheric radiation are very complex to include in a model, and often require

considerably more arithmetic operations than the rest of the terms in the equations combined. The parameterization of these physical processes is treated in Chapter 4. Parameterization involves the representation of a process in terms of its known relationship to dependent variables resolved on the model grid. For example, we cannot resolve individual turbulent eddies, but we can develop relationships between turbulence intensity and model-resolved wind shear and static stability. There are typically three reasons why we parameterize a process: we do not understand the process well enough to represent it directly through physical relationships, the process is of sufficiently fine scale that we cannot resolve it on the model grid, or the physical relationships are so complex that they would require a prohibitive amount of computing resources to treat explicitly.

3.2 Numerical frameworks

There are four different modeling frameworks described here for dealing with the space dependence in the nonlinear partial differential equations of atmospheric dynamics and thermodynamics:

- finite difference, or grid point;
- spectral;
- finite element; and
- finite volume.

This section does not focus on the details of the methods used to approximate the space derivatives in the equations, but rather on the overall approaches.

3.2.1 Spatial finite-difference/grid-point methods

Over the past half century, atmospheric scientists and oceanographers have developed numerous approaches for applying grid-point methods to the solution of the equations of fluid flow over part or all of the sphere. These methods include the use of map projections, latitude–longitude grids, and spherical geodesic grids. In each case, a procedure is defined for organizing grid points in a systematic way over the area of the sphere for which the atmosphere is to be modeled. The choice of which method to adopt in a particular modeling application depends on a variety of factors including whether the model has a limited area or global computational area, and the degree to which the code needs to be easy to modify for research purposes.

Computational grids may be classified as structured or unstructured. Traditional grids are structured in that they consist of an array of cells that are arranged in a regular pattern in two or three dimensions. In contrast, unstructured grids are defined by collections of elements, such as triangles, in an irregular pattern. This provides for greater flexibility in discretizing complex domains, and allows for the convenient use of adaptive-meshing techniques where cells can be added or subtracted. Unlike structured grids, unstructured grids require a list of mesh connectivities.

Map projections

Map projections are geometric, and therefore mathematical, relationships that transform atmospheric properties defined on a quasi-spherical surface, such as Earth's surface or a 500-hPa surface, to a flat surface, such as a geographic map, a weather map, or a model grid. Figure 3.3 shows the geometric relationships between a spherical surface and a surface on which properties defined on the sphere are projected (the image surface), for the three types of projections commonly used in atmospheric modeling. In each case, imagine a set of rays drawn from a common origin, so that the rays connect points on the sphere

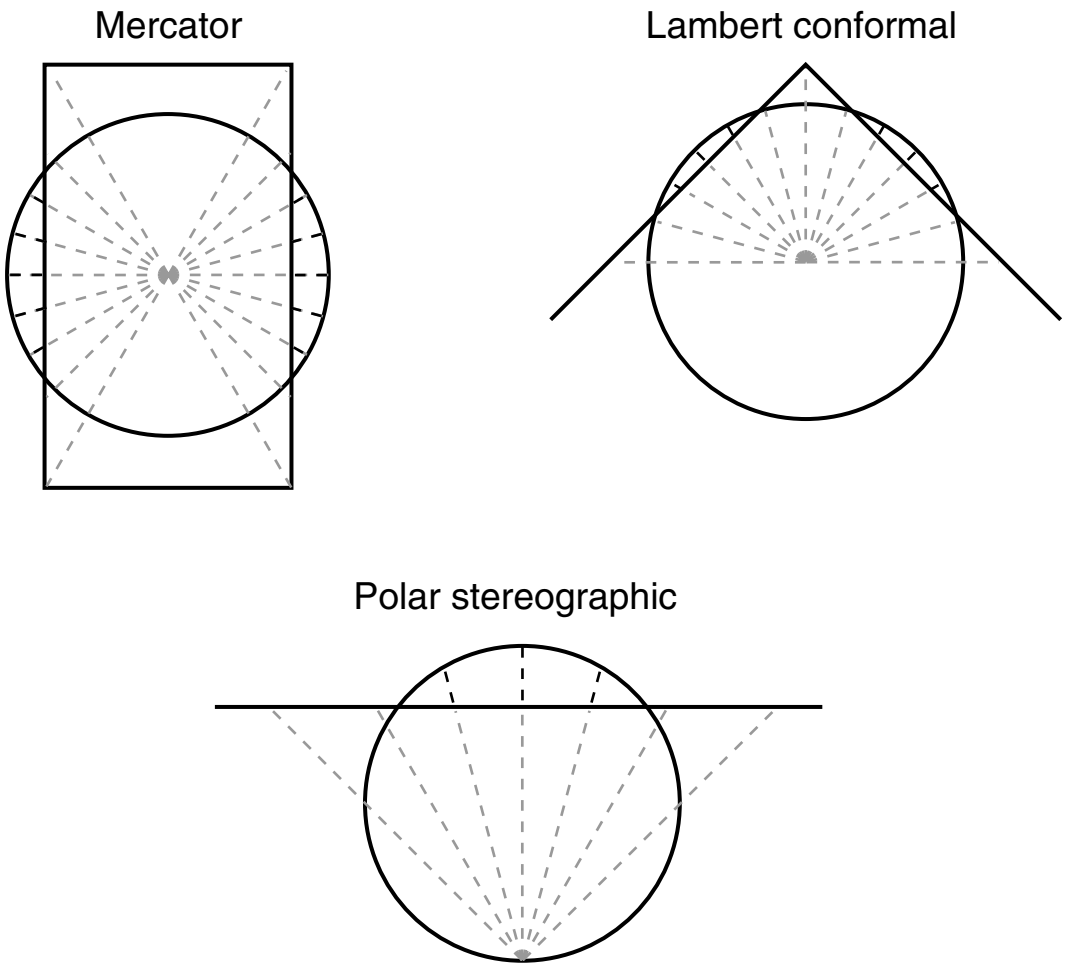

Fig. 3.3 Three map projections commonly used in atmospheric modeling. The cylinder (Mercator), right-circular cone (Lambert conformal), and plane (polar stereographic) are the surfaces on which the information on the sphere is projected. The radial lines connect points on the sphere and points on the projection surface. The axes of the cylinder and the cone, and the perpendicular to the plane, are parallel to Earth's axis of rotation. In these images, we are thus viewing Earth from over the Equator.

and points on the surface of the projection. For example, the Mercator projection is defined using a cylinder whose central axis passes through the center of the sphere. The conditions on the sphere are mapped to the cylinder, and the cylinder can be cut, opened, and made flat. Similarly, a flat surface and a cone define the projection surfaces for polar-stereographic and Lambert-conformal projections, respectively. The plane and the cylinder may be viewed as special cases of the cone, with vertex angles of 180° and 0°, respectively. In atmospheric-modeling applications, the axes of the cylinder and cone, and the perpendicular to the plane, are virtually always coincident with the axis of rotation of Earth. For each type of projection, the projection surface may intersect the sphere, as in the figure, or it may be tangent to the sphere. In the former case it is called a secant projection, and in the latter a tangent projection.

It is not possible to preserve all geometric properties on the sphere (e.g., area, shape, angles) during a projection. For example, Fig. 3.3 shows that distances and areas of high-latitude features are exaggerated with the Mercator projection, as are low-latitude features with the polar-stereographic projection. In fact, only at the lines or points of intersection of the sphere and the projection surface (the standard parallels) are all properties preserved. But, the three projections described above are all *conformal* in that they everywhere preserve angles between two curves, and the distance distortion is the same in all directions at a point.[1] For meteorological applications where preserving the angles of atmospheric features is important (e.g., the angle between isobars and wind vectors), conformal projections are desirable. The lack of distance and area preservation with conformal projections is dealt with by applying them only for latitudes where the distortion is small.

Map projections are needed for virtually all atmospheric models. For a global model that uses spherical coordinates, visualizing the output on paper or on a computer screen requires that the atmospheric conditions, and associated georeference information, such as political boundaries and natural features, be defined on one of the map projections. Mathematical transformations for each of the projections convert from spherical coordinates (latitude and longitude) to Cartesian coordinates on the projection surface. This is the process by which geographers transfer properties defined on Earth's surface to a map.

For limited-area models, which employ Cartesian coordinate systems and solve the equations on planar surfaces, the transformation between the sphere and projection surface becomes an intimate part of the modeling process and the equations themselves. In particular, observations whose locations are defined in latitude–longitude coordinates need to be applied at the appropriate coordinates of the Cartesian model grid that is defined on the projection surface. And, because of the distance distortion, the grid increment used in the finite-difference equations needs to reflect the true horizontal distance between points. Figure 3.4 shows how the grid increment is affected by the distance transformation between the sphere and a projection surface on which a computational grid is defined. The points on the computational grid defined on the projection surface are equidistant, but the

[1] If x is a horizontal displacement from a point, $\delta x_E / \delta x_G$ is the same regardless of the direction of the displacement, where E refers to Earth and G to the grid.

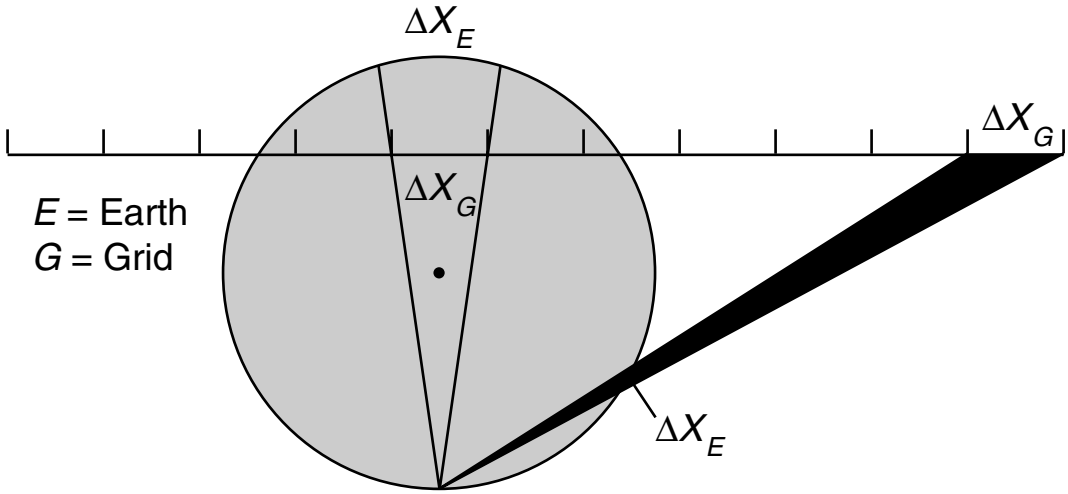

Fig. 3.4 The relationship between the grid increment on Earth and on the computational grid, for a polar-stereographic, secant projection.

physical distance represented by each increment is generally different and dependent on the location on the grid. A measure of this distortion is the map-scale factor, defined in Eq. 3.2, using the notation in Fig. 3.4, as the ratio of distance on the projection to distance on the sphere:

$$m = \frac{\Delta x_G}{\Delta x_E}. \tag{3.2}$$

The map-scale factor is the same along each latitude circle. Figure 3.5 shows this scale factor as a function of latitude for the three projections. The direction of the departure of m from unity depends on whether the image surface is above or below Earth's surface. For each point on a model grid, the map-scale factor is precalculated, and used in the equations to account for the varying distance between grid points. Equation 3.3 shows the first equation of motion, with Earth-curvature terms, for a popular limited-area model (Dudhia and Bresch 2002).

$$\frac{du}{dt} = -\frac{m}{\rho}\frac{\partial \rho'}{\partial x} + \left(f + u\frac{\partial m}{\partial y} - v\frac{\partial m}{\partial x}\right)v - ew\cos\alpha - \frac{uw}{r} + D_u. \tag{3.3}$$

Here, m is the map-scale factor given by, for example,

$$m = \frac{1 + \sin 60^\circ}{1 + \sin\phi}$$

for a polar-stereographic projection that is true at $60°$; e and f represent the full Coriolis force, where $f = 2\Omega\sin\phi$ and $e = 2\Omega\cos\phi$; D is a diffusion, or frictional-dissipation,

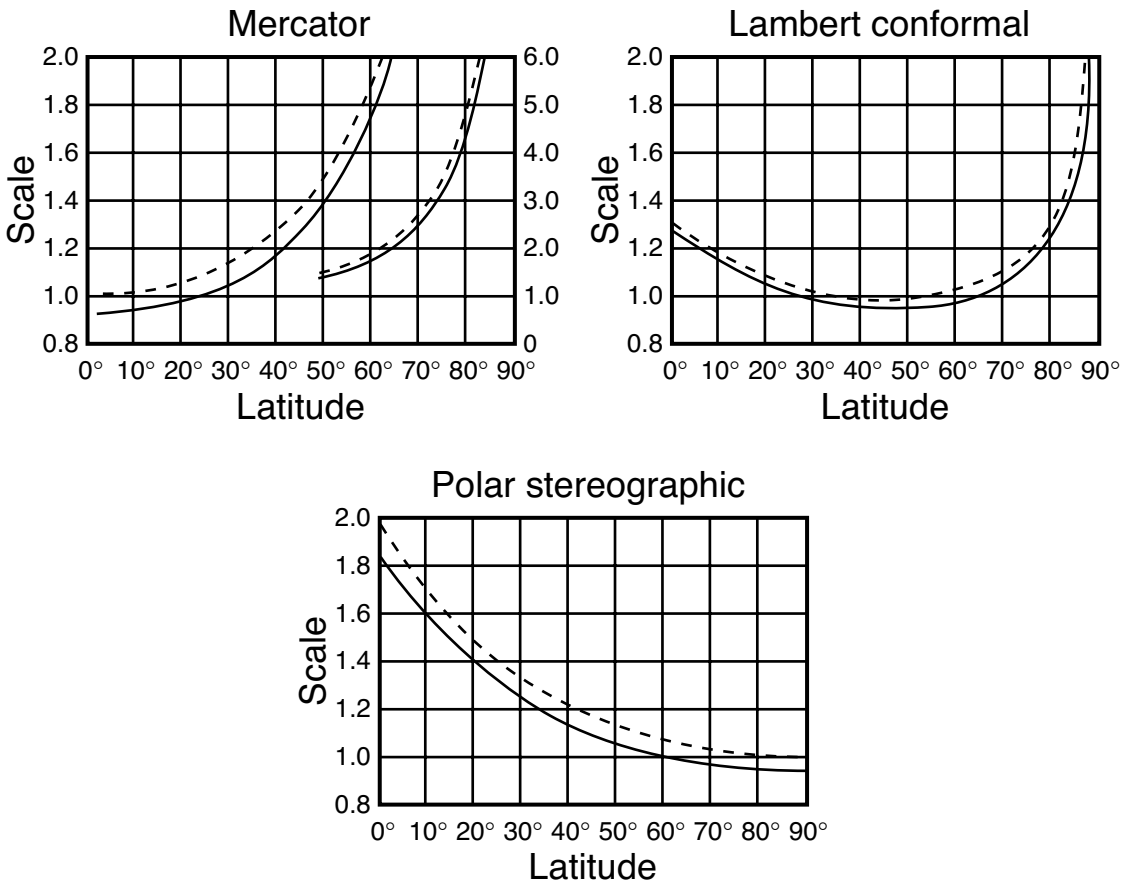

Fig. 3.5 Map-scale factors for different tangent (dashed lines) and secant (solid lines) projections as a function of latitude. For the secant projections, the conical surface intersects the sphere at 30° and 60° (north or south) latitude, the plane intersects at 60° (north or south) latitude, and the cylinder intersects at 20° (north and south) latitudes. From Saucier (1955).

term; r is the radius of Earth; ρ is density; and $\alpha(x, y)$ is the angle between the local meridian and the y axis. Figure 3.5 shows that the differential-map-scale terms in Eq. 3.3 are smallest for this projection near the poles and largest in equatorial areas.

Because of the varying effective distance between grid points, computations are really being performed on a "stretched" grid, and this leads to spatial contrasts in the numerical properties (i.e., the errors) of the solution to the equations. As will be seen later in this chapter, this means that the same wave on the grid will have different phase and group speeds depending on latitude. And there will be latitudinal differences in the conditions needed to maintain stability of the numerical solution to the equations. Thus, objectives in the decision about the best choice of map projection to use for a particular model application are to minimize (1) the departure of the map-scale factor from unity over the grid and (2) the latitudinal derivative of the map-scale factor. In general, these conditions can be

best satisfied by using the Mercator projection for grids in tropical latitudes, the polar-stereographic projection for high-latitude grids, and the Lambert-conformal projection for midlatitude grids (see Fig. 3.5).

Even when a reasonable map-projection choice is made for a particular application of a model, and the transformations are properly incorporated into the model equations and the initialization process (i.e., getting the observations in the right place on the computational grid), there are a few ways in which the properties of the projection may impact the model user. One is that the u and v velocity components in the atmosphere (defined in terms of the east–west and north–south directions on the sphere) are not the same as the u and v components on the computational grid (defined in terms of grid-point rows and columns). This issue must be dealt with when initializing the model. Another is that the time step that is chosen by the user, or automatically by the model, is based on the grid increment in order to maintain a stable solution to the equations (see Section 3.4.2). Because the true horizontal grid increment varies spatially, some areas of the grid may have sufficiently large values that stability criteria are locally violated. Evidence of this would be unrealistic-appearing (e.g., small-scale waves) model solutions in latitudes where the grid increment is the smallest.

The above discussion was in the context of using map projections to model limited areas of Earth's surface. However, there have also been methods developed for using combinations of map projections to model the entire sphere using a *composite grid*, where one of the objectives is to avoid the problems of latitude–longitude grids described later. For example, Phillips (1957a, 1962) used a combination of a Mercator projection for latitudes equatorward of a boundary latitude and stereographic projections for higher latitudes. And, Stoker and Isaacson (1975) and Dudhia and Bresch (2002) used two polar-stereographic projections that overlapped in equatorial regions. Calculations on these grids can communicate at the interfaces, thus avoiding the need for artificial lateral-boundary conditions, or the integrations can be separate. Figure 3.6 shows an example of two overlapping polar grids. The method of using two overlapping polar-stereographic projections for global simulations was compared with two spectral methods by Browning *et al.* (1989) in terms of memory requirements, execution time, and arithmetic-operation count. The results were mixed, but the conclusion was that the methods were generally competitive.

Another common approach to modeling the sphere with map projections is to circumscribe a regular polyhedron, such as a cube, by the sphere. On each face of the polyhedron is defined a regular Cartesian grid, and radials from the center of the sphere are projected through the grid points in order to map the grid to the surface of the sphere. In the case of the cube, the model equations are solved on each of the six grids, but the calculation of finite differences at the boundaries is challenging. Sadourny (1972), McGregor (1996), Rančić *et al.* (1996), Ronchi *et al.* (1996), and Purser and Rančić (1997, 1998) review the testing and properties of such polyhedral-gnomonic projections. Adcroft *et al.* (2004) describe the use of the expanded cube as the basis for the general-circulation model of the Massachusetts Institute of Technology, McGregor and Dix (2001) summarize the use of this approach in Australia's Commonwealth

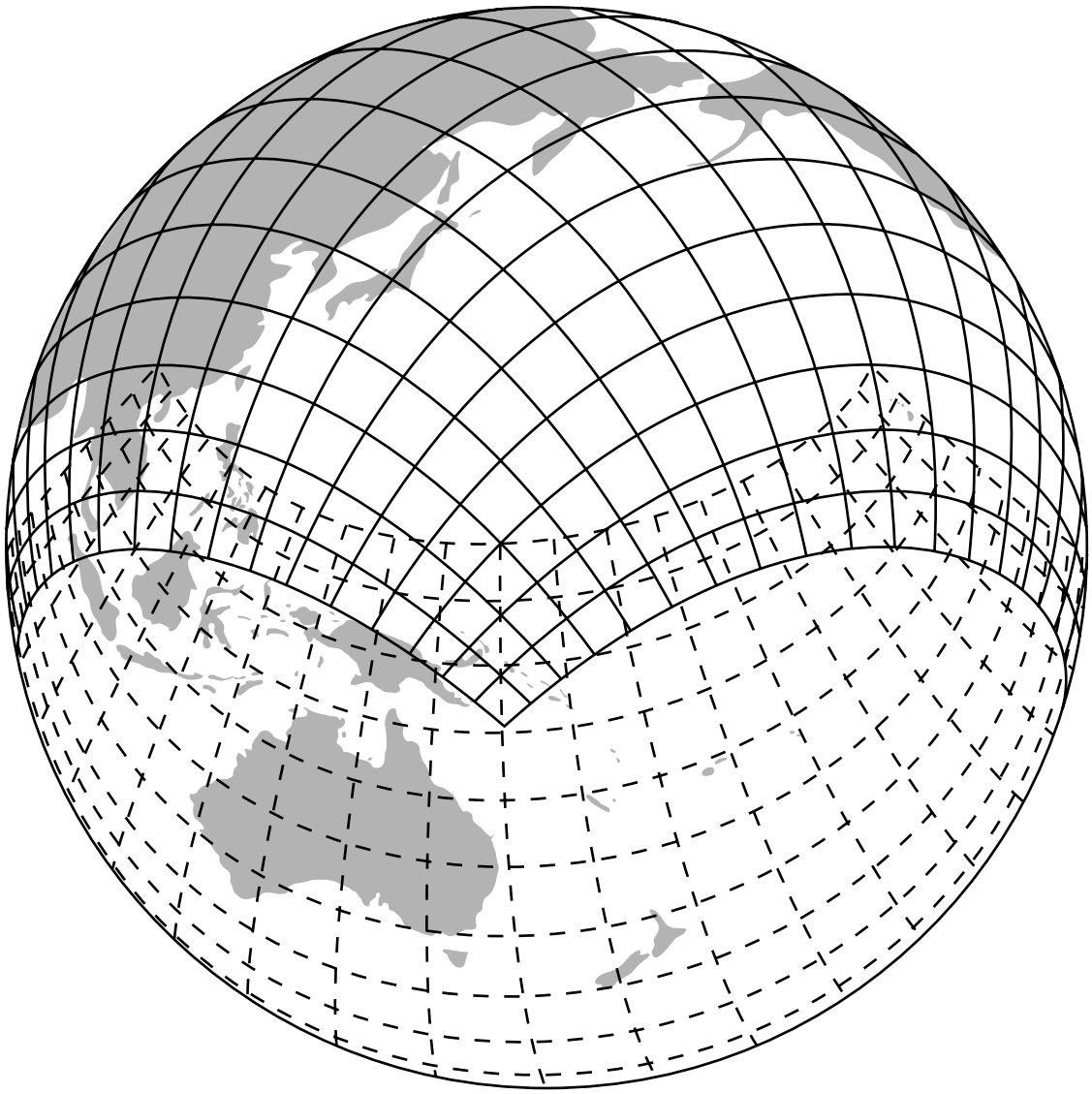

Fig. 3.6 A composite grid defined by overlapping North Polar and South Polar stereographic projections. From Williamson (2007).

Scientific and Industrial Research Organization (CSIRO) general-circulation model, and Zhang and Rančić (2007) apply the method to a version of the US National Weather Service's (NWS) Eta model. Figure 3.7 illustrates the projection of Earth's surface on the faces of an exploded cube, as well as an example of the relatively uniform distribution of grid points on the sphere. This approach produces a nonconformal projection, but additional transformations can convert it to one that has conformal properties.

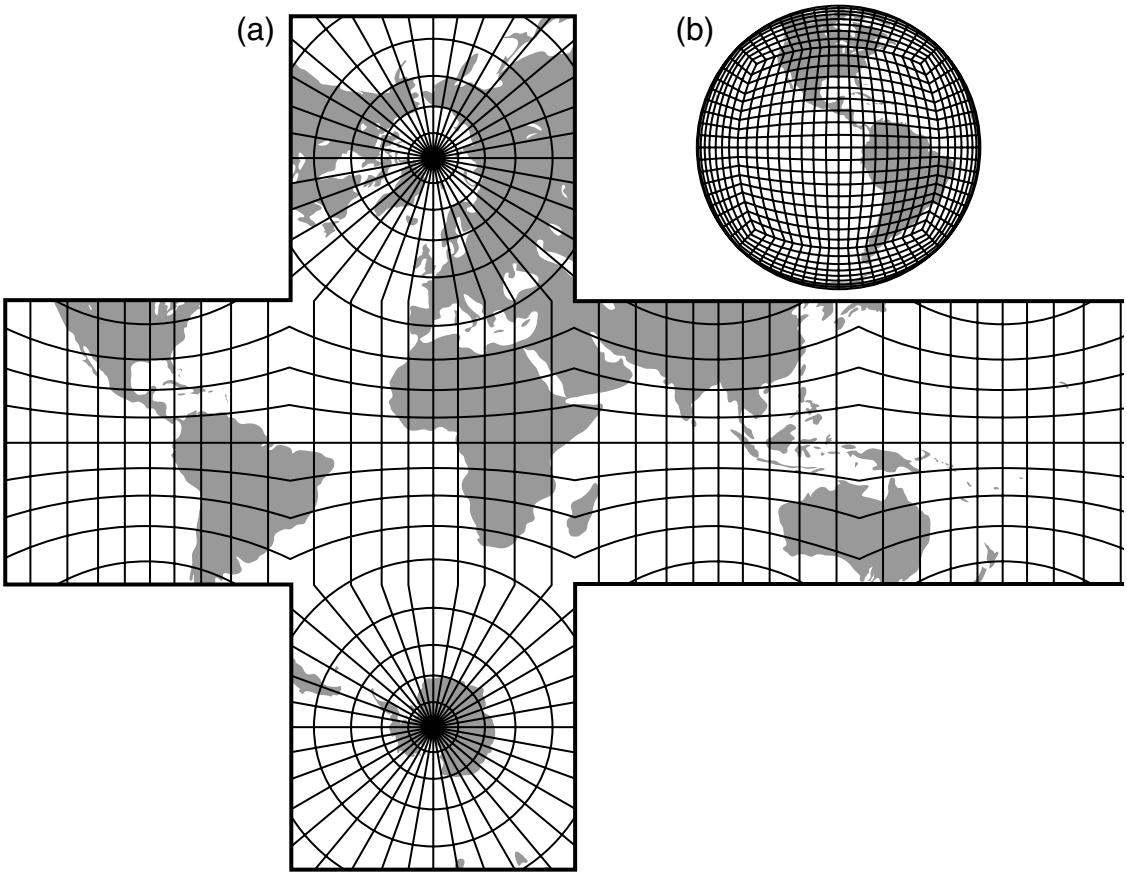

Fig. 3.7 A *cubic gnomonic projection*, used as the basis for a global model grid, is defined by establishing Cartesian grids on each face of a cube that is inscribed within a sphere. These grids are mapped to Earth's surface, producing the relatively evenly spaced grid points shown on the sphere in (b). The expanded cube with geographic and latitude–longitude references is shown in (a). Panel (b) is from Rančić *et al.* (1996).

Latitude–longitude grids

In this approach, latitude and longitude are the horizontal coordinates, and the vertical coordinate is defined along the local radial from the center of Earth. On each vertical coordinate surface, the sphere is partitioned into grid cells using increments of latitude and longitude. If these intervals are constant over the entire sphere, the longitudinal distance between grid points becomes progressively smaller as the meridians converge at the poles (Fig. 3.8).

This requires that time steps be small in order to maintain stable solutions to the equations. In addition, the existence of the singularities at the poles, where the coordinate lines (the meridians) intersect, means that calculating horizontal derivatives can be problematic. Shortening the time step near the poles produces satisfactory results, but calculations for

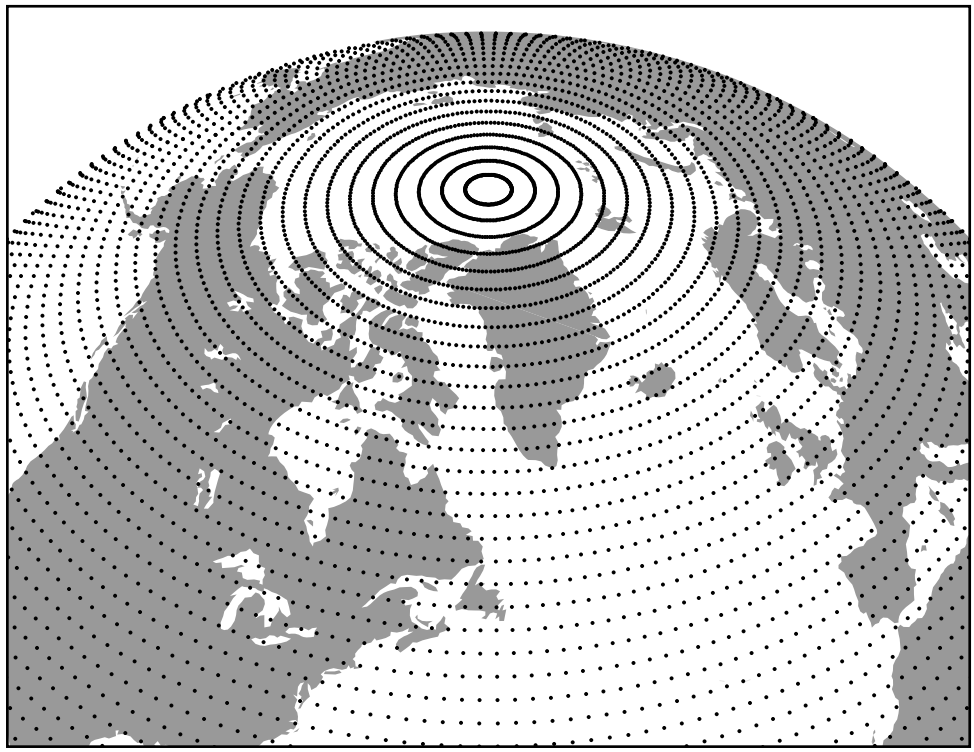

Fig. 3.8 A latitude–longitude grid shown for part of the sphere, where the points are defined at a uniform interval in each coordinate direction.

the small percent of Earth's surface area near the poles can consume more than half the computer time (Grimmer and Shaw 1967).

One approach to dealing with the small distance between grid points near the poles, and the associated impact on the time step and model performance, is to Fourier filter the variable fields in the east–west direction near the poles. This is accomplished by Fourier transforming a variable, filtering out the higher wavenumbers by truncating the series, and then inverse transforming back to physical/grid-point space to obtain a smoother field. This removal of small-scale information from the model solution effectively filters the faster modes, and permits the use of a longer time step in spite of the small longitudinal grid distance. Note that, in this approach, there is still the computational burden of solving the equations at the dense array of grid points near the poles, and the excessive density of points wastes memory. Williamson (2007) points out that this is an unsatisfying engineering approach, but it is in use today, for example in the optional finite-volume core of the National Center for Atmospheric Research's (NCAR) Community Atmosphere Model (CAM). Another method is to use larger increments of longitude as the pole is approached (Williamson and Rosinski 2000). An example of the resulting *reduced grid* is shown in Fig. 3.9.

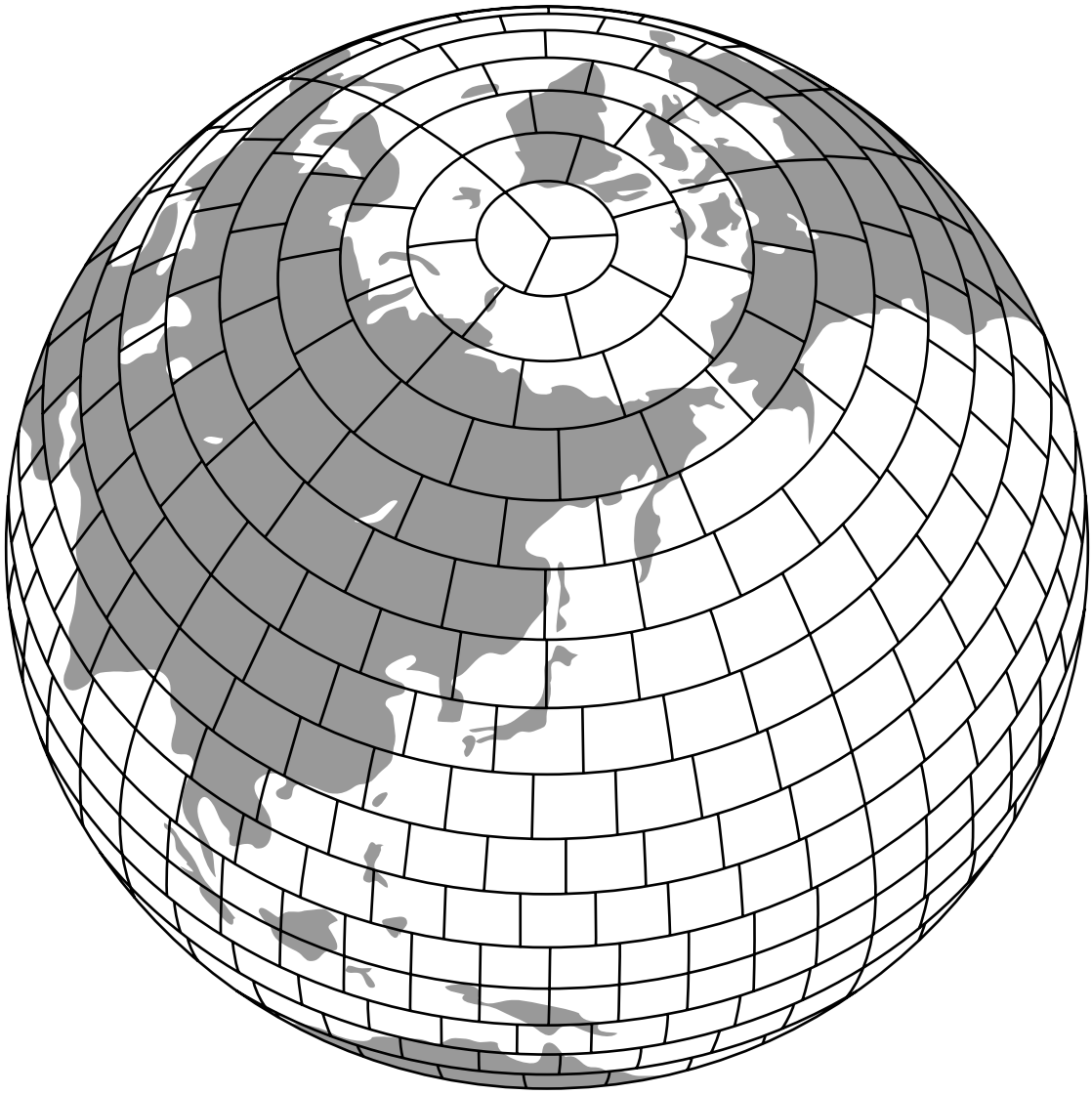

Fig. 3.9 A reduced grid, in which the longitudinal grid increment in degrees is increased with decreasing distance from the pole, where the objective is to maintain a relatively uniform physical distance between grid points. From Williamson (2007).

Spherical geodesic grids

A desirable property of a grid is uniformity in the spacing of the grid points over the sphere, or a portion of the sphere. It has been shown that map projections can produce grids that have significant variability in the Earth-distance between points (Fig. 3.5), especially when the computational domain must span a large area. Similarly, latitude–longitude grids have inherently higher resolution where the meridians converge at high

latitudes. However, geodesic grids have a nearly homogeneous distribution of points over the sphere.

In mathematics, a geodesic is the equivalent of a straight line, but on a curved surface. On a spherical surface, such as that of Earth, a geodesic is the shortest path between two points, specifically a segment of a great circle. A spherical geodesic grid is defined by spherical, equilateral triangles whose edges are geodesics. One way of defining this grid is to begin with an icosahedron, the geometric solid shown in Fig. 3.10a with 20 triangular faces (major triangles), 12 vertices, and 30 edges, where the vertices touch the surface of a sphere. The vertices may then be connected by geodesics on the sphere, producing spherical triangles. A grid may be created by dividing the major triangles into smaller ones (grid triangles) using a variety of approaches. For example, bisecting each edge of the icosahedron and connecting the bisection points produces four new equilateral triangles for each original one (Fig. 3.10b). The vertices of these new triangles can then be projected onto the sphere along a radial from the center (Fig. 3.10c), and then connected by geodesics to again produce spherical triangles (Fig. 3.10d). Even though the distances between adjacent points look uniform, they are not exactly so. A hint at the asymmetries from one part of the surface to the next can be seen in the fact that the "new" vertex facing the viewer in the upper-center (Fig. 3.10d) is surrounded by six adjoining triangles, while the "original" icosahedron vertex to its right is surrounded by only five. Williamson (1968) and Sadourny *et al.* (1968) describe another approach for dividing the major triangles into grid triangles, where the inequality in the distance between points is less than that resulting from the method just described. Figure 3.11 shows an example of the distribution of grid points over the sphere.

Some applications of spherical geodesic grids employ the above triangular cells, while others use a related grid with hexagonal cells. To obtain the latter, Voronoi cells are constructed based on the triangular grid, where such cells consist of the set of all points that are closer to a particular vertex than to any other vertex. For the twelve original vertices in the icosahedral grid (e.g., in Fig. 3.11), the Voronoi cells are pentagons. For all the rest, they are hexagons. Figure 3.12 illustrates the geometric relationship between the triangular

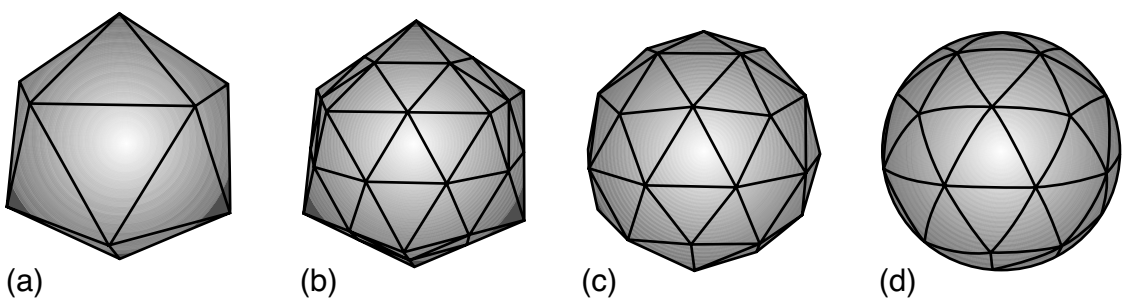

Fig. 3.10 In the generation of a spherical geodesic grid, the major triangles of the icosahedron (a) are subdivided, where (b) shows one approach. The vertices of the new triangles are projected (c) onto the sphere that is coincident with the vertices of the icosahedron. Geodesic lines are then drawn between the new vertices to generate spherical grid triangles (d).

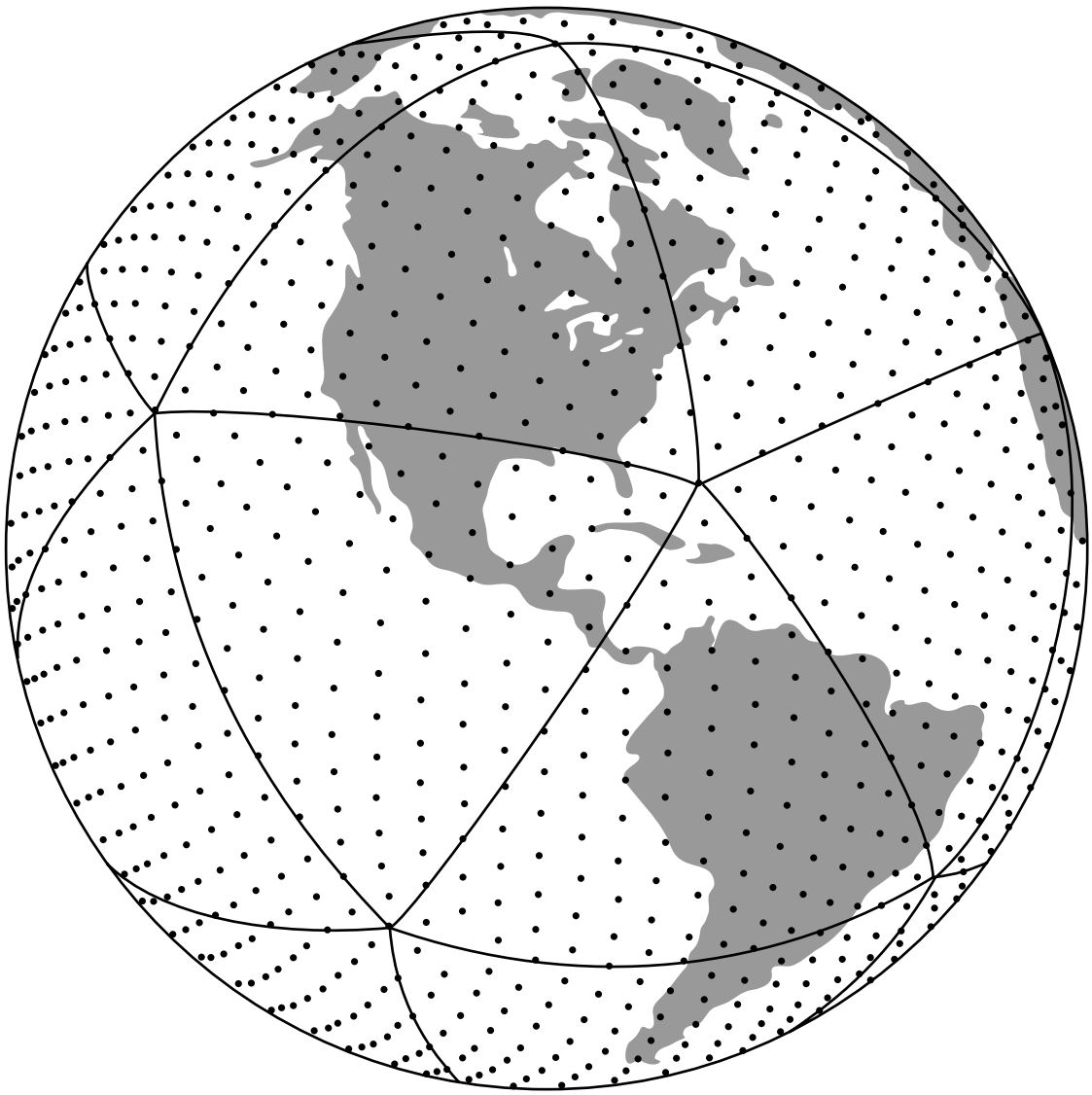

Fig. 3.11 An example of the relatively uniform distribution of grid points over the sphere, for a spherical geodesic grid, based on one method for dividing the major triangles of an icosahedron into grid triangles. Note that the horizontal resolution in this example is very coarse. Adapted from Williamson (1968).

and hexagonal grids. Randall *et al.* (2002) describe the relative merits of hexagonal, square, and triangular cells, and Weller *et al.* (2009) compare a few additional mesh-refinement methods.

Other advantages of the spherical geodesic grid include the fact that it is straightforward to selectively enhance the resolution in some areas (adding triangles), to better render fine-scale features in the vicinity of mountains or other types of small-scale local

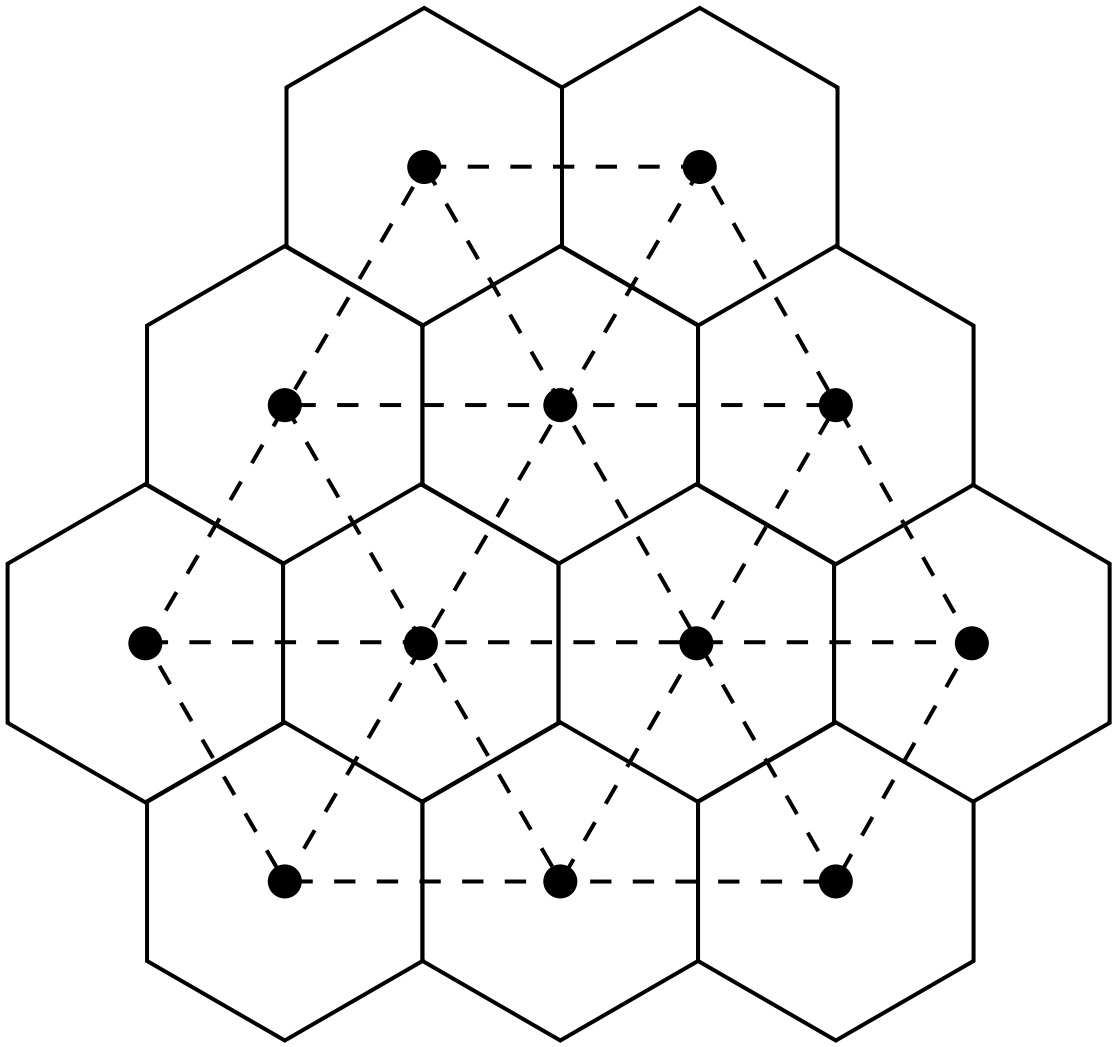

Fig. 3.12 Illustration of the geometric relationship between triangular cells and hexagonal cells, either of which can be used as the basis for a spherical geodesic grid.

forcing. A disadvantage is that the indexing of the grid points is more complex than for Cartesian or latitude–longitude grids, but Randall *et al.* (2002) show how the grid cells on the sphere can be separated into rectangular panels whose cell values can be logically organized in computer memory.

Examples of contemporary models that employ spherical geodesic grids are the Ocean–Land–Atmosphere Model (OLAM; Walko and Avissar 2008a,b), the Operational Multiscale Environment Model with Grid Adaptivity (OMEGA; Bacon *et al.* 2000), and the operational GME model of the German Weather Service (Majewski *et al.* 2002). Randall *et al.* (2002) describe the development of a coupled ocean–land–atmosphere

spherical-geodesic-grid model for climate applications. Tomita and Satoh (2004) and Satoh *et al.* (2008) describe a model that uses this method for global, cloud-resolving simulations. And, the US National Oceanic and Atmospheric Administration (NOAA) is developing the Flow-following finite-volume Icosahedral Model (FIM) for operational use. Further discussions and literature summaries about spherical geodesic grids can be found in Sadourny *et al.* (1968), Williamson (1968), Baumgardner and Frederickson (1985), Nickovic (1994), Ringler *et al.* (2000), and Ringler and Randall (2002).

Differential grid resolution across the sphere

For modeling global-scale processes, it is reasonable to desire somewhat uniform horizontal resolution over the sphere. However, in other model applications it is common to want greater resolution in certain regions. For example, in research settings a specific meteorological feature is often being studied, and the grid points and computational resources should be focussed there. For operational forecasting models, small-scale processes dominate in certain regions such as near complex terrain, making it desirable to have greater horizontal resolution there than elsewhere. Additionally, operational models are often set up to serve limited areas such as specific countries, so, again, greater resolution is needed there.

There is a variety of approaches that can be used to produce different horizontal resolutions over a three-dimensional computational volume. A common one is to embed a high-resolution limited-area model within a global model, with the global model providing lateral-boundary conditions to the limited-area model. The global model forecast or simulation is performed first. The limited-area model may consist of a single high-resolution grid, or a nest of multiple grids with grid spacings that change abruptly by a factor of perhaps three to five between adjacent grids. Figure 3.13 illustrates an example of a nest with multiple grids that is used for operational prediction over an area near the Chesapeake Bay in the eastern USA (Liu *et al.* 2008a). Section 3.5 on lateral-boundary conditions discusses the methods and limitations of this approach for obtaining higher horizontal resolution over specific areas of the sphere.

A property of some models is that all the grids in a horizontal nest must have the same vertical resolution (distribution of layers). In contrast, with others, not only can the vertical resolution vary among the grids in a nest, but vertical nesting is permitted. Figure 3.14c shows the common situation where the grids in a nest have the same vertical resolution. In contrast, Figs. 3.14a and b illustrate vertical nesting, where the inner grid not only has higher vertical resolution, but it can focus computational resources in certain vertical layers. For example, in the inner grid with higher horizontal resolution, enhanced vertical resolution can be used in the boundary layer, near the tropopause, or in layers with low-level jets, all regions with larger vertical gradients of variables. See Clark and Farley (1984) and Clark and Hall (1991, 1996) for additional information about vertical grid nesting in the Clark model.

Because wave reflections can sometimes occur at an abrupt variation in resolution, such as at the transition between grids in a nest (Davies 1983), there is a benefit to using a gradual variation in horizontal grid increment to achieve greater resolution in certain areas.

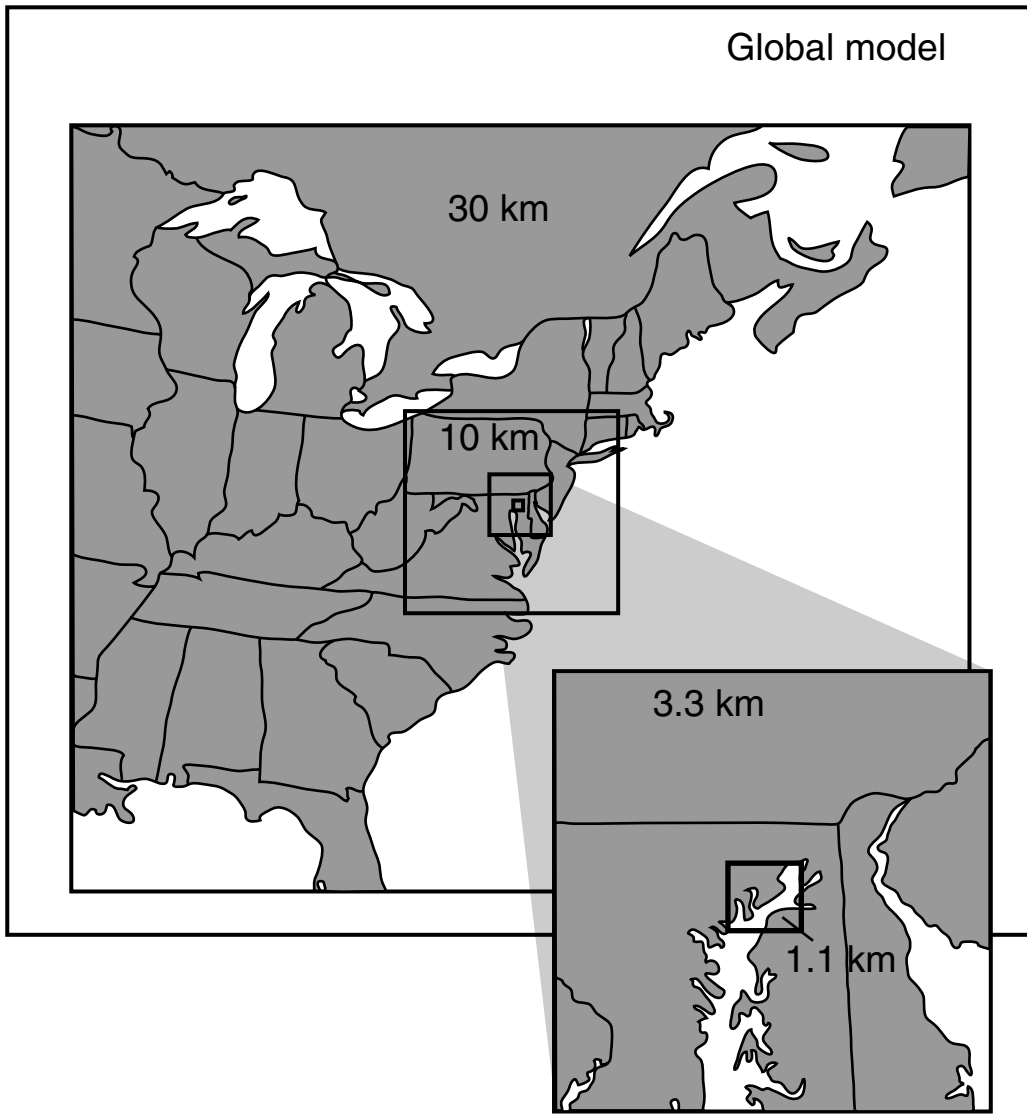

Fig. 3.13 An example of a nested-grid model used for operational prediction. The model with the two-way interacting grids is embedded within a global-model prediction. Model grid increments are indicated. From Liu *et al.* (2008a).

Such grids with a gradual change in resolution are sometimes called stretched grids. Kalnay de Rivas (1972) and Fox-Rabinovitz *et al.* (1997) discuss the truncation error associated with the approximation of derivatives on a variable-resolution mesh. They point out that such errors with a nonuniform smoothly varying mesh are equivalent to those with a uniform mesh defined by a transformation, such as associated with map projections. That is, even though ΔX_G is constant, the effective grid increment on Earth's surface, ΔX_E, varies smoothly. Virtually all models employ a stretched grid in the vertical, with higher

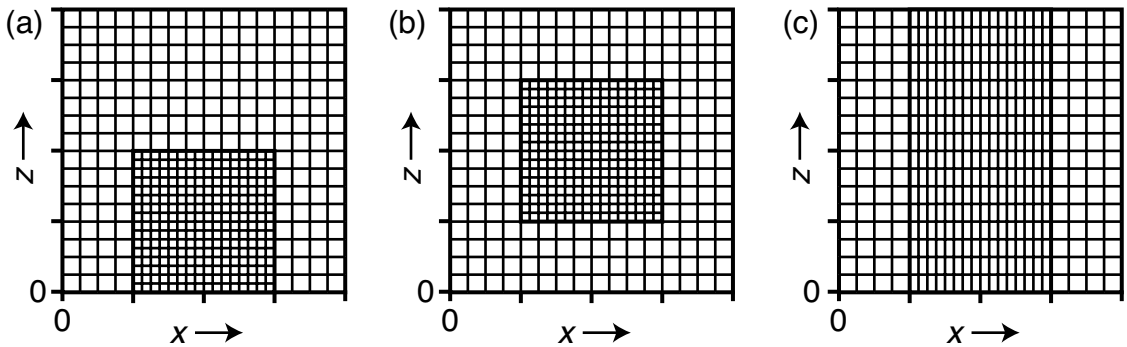

Fig. 3.14 Schematic examples of vertical grid nesting, where certain layers in the atmosphere are represented by a subgrid with greater vertical resolution. Three examples are shown for a situation where there is a grid with greater horizontal resolution that is horizontally nested within a grid having lower horizontal resolution. One example shows vertical nesting where the lower atmosphere is represented by the higher-resolution grid (a), another shows the enhanced vertical-resolution grid confined to the middle of the model atmosphere (b), and another illustrates the situation where the model does not allow vertical nesting or grid stretching (c).

vertical resolution in some layers, such as near Earth's surface. Cartesian, horizontally stretched grids for limited-area models have been investigated and used by Anthes (1970), Anthes and Warner (1978), Staniforth and Mitchell (1978), Staniforth and Mailhot (1988), and Walko *et al.* (1995b). Global-model stretched grids are sometimes motivated by the fact that there is a cost-saving associated with modeling systems that are sufficiently versatile that they can be used both for global studies and for those that focus on specific geographic regions. Examples are the Global Environmental Multiscale (GEM) model of the Meteorological Service of Canada (MSC) (Côté *et al.* 1998a,b; Yeh *et al.* 2002) and the National Aeronautics and Space Administration (NASA) Goddard Earth Observing System (GEOS) general circulation model (Suarez and Takacs 1995; Takacs and Suarez 1996; Fox-Rabinovitz *et al.* 1997, 2000). Côté *et al.* (1993) describe a global shallow-fluid model that employs a similar stretching method. This strategy uses spherical coordinates with variable resolution in both horizontal directions. For global studies, a conventional latitude–longitude grid is employed, with the singularities at the poles. When a concentration of grid points is needed for higher resolution in a particular region, the resolution is varied, as in the example shown in Fig. 3.15. The poles of the coordinate system do not necessarily coincide with Earth's poles. With increasing distance from the grid's poles, near the geographic poles in this example, the resolution in that direction increases, as shown, until becoming constant for the east–west belt around the sphere shown in gray. The resolution of the other spherical coordinate also varies, with uniform high resolution in the north–south belt through the Americas. Where the areas of highest resolution in the two horizontal dimensions overlap, a high-resolution regional grid with a uniform latitude–longitude grid increment of 0.04° exists in this example. Fox-Rabinovitz *et al.* (2006) discuss the Stretched-Grid Model Intercomparison Project (SGMIP), where a number of variable-resolution global models were used for regional-climate simulations.

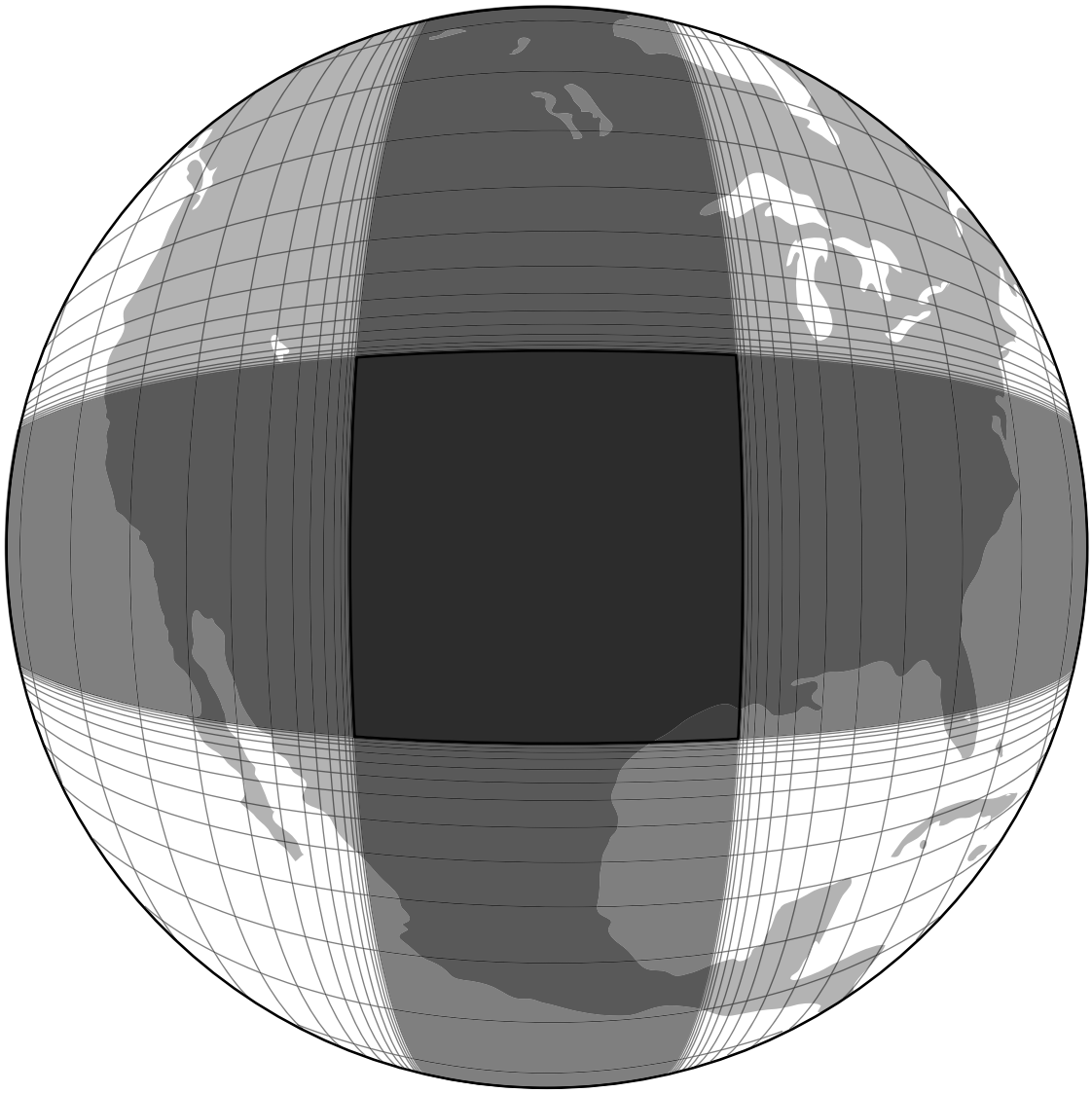

Fig. 3.15 An example of the variable-resolution horizontal grid of the operational regional configuration of the GEM model. See the text for details. Adapted from Yeh *et al.* (2002).

There are also various approaches in which the grid resolution changes during a model integration in order to better represent the evolving atmospheric processes. One method used with Cartesian limited-area models is called adaptive mesh, or grid, refinement, and is described in Berger and Oliger (1984) and Skamarock and Klemp (1993). This method uses the above concept of nested grids, but here the fine grids can change size, shape, location, and number (i.e., they can adapt) automatically based on estimates of the truncation error during a simulation or forecast. For example, fine-mesh grids are spawned

automatically to follow convective or tropical storms as they move. Other methods are described in Dietachmayer and Droegemeier (1992) and Srivastava *et al.* (2000). Also, spherical geodesic grids easily accommodate adaptive mesh refinement methods because spherical triangles can be added or removed as needed (e.g., Bacon *et al.* 2000).

Consistency of vertical and horizontal grid increments

There is evidence that model vertical and horizontal resolutions should not be specified independently. In particular, physical features that can be resolved well by the horizontal grid increment should also be resolvable by the vertical grid increment (and vice versa). If the vertical grid increment is too coarse to satisfy this criterion, the resulting truncation error will generate spurious gravity waves during the simulation and the features will be poorly rendered by the model. This problem has been most-commonly described in the context of sloping features in the atmosphere such as fronts (e.g., Snyder *et al.* 1993) or the slantwise convection resulting from conditional symmetric instability (e.g., Persson and Warner 1991). The mathematical relationship that defines consistency between the vertical and horizontal grid increments has been defined differently by different authors, but the expressions tend to be quite similar from a practical standpoint. For example, Pecnick and Keyser (1989) state that the optimal vertical grid increment is related to the horizontal grid increment by the expression

$$\Delta z_{opt} = s\Delta y\,, \tag{3.4}$$

where s is the frontal slope, Δz_{opt} is the optimal vertical grid increment, and Δy is the horizontal grid increment. For synoptic-scale fronts with typical slopes from 0.005 to 0.02, this relationship gives optimal vertical grid spacings of 0.5–2.0 km for $\Delta y = 100$ km, and 50–200 m for $\Delta y = 10$ km. Alternatively, Lindzen and Fox-Rabinovitz (1989) suggest two consistency relationships, one for quasi-geostrophic flows and another for flows that contain gravity waves near a critical layer. In both cases, the Δz_{opt} for midlatitudes is similar to that obtained from Eq. 3.4.

These consistency relationships and the associated research with hydrostatic and nonhydrostatic models suggest that decreasing the horizontal grid spacing of a model without also reducing the vertical grid spacing may not lead to an improvement, and may actually produce a worse simulation. The two-dimensional model simulations of frontogenesis presented by Pecnick and Keyser (1989) show that spurious gravity-wave structures and spurious large velocity and vorticity values result when $\Delta z > \Delta z_{opt}$. Also, Lindzen and Fox-Rabinovitz (1989) refer to instabilities, spurious amplitude growth, and other problems when an inconsistency exists between the vertical and horizontal resolutions. And, Gall *et al.* (1988) report that the intensity of erroneous waves generated at a front was diminished when the vertical grid spacing was reduced, to be consistent with the horizontal grid spacing.

Examples of the effects of using a vertical grid spacing that is insufficient to represent structures that are well resolved in the horizontal are shown in many of the above studies. As an illustration here, Fig. 3.16 shows the numerical noise in the vertical motion that results

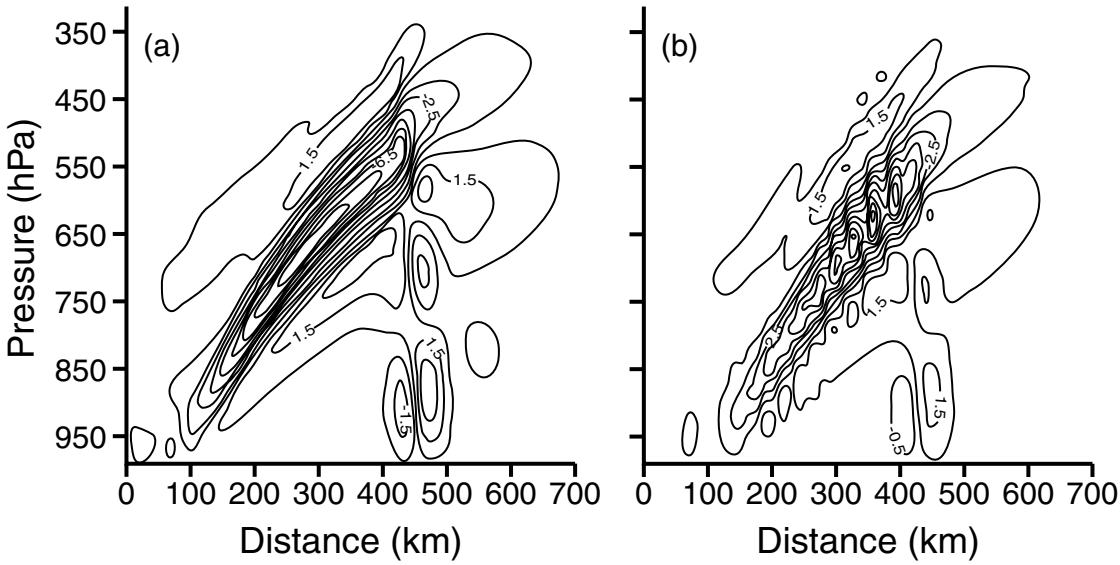

Fig. 3.16 Vertical cross sections of vertical velocity, ω (solid lines, $\mu b\ s^{-1}$) after 24-h simulations of conditional symmetric instability with a 10-km horizontal grid increment and 75 layers (a) and 25 layers (b). From Persson and Warner (1991).

from the use of an inappropriately large vertical grid spacing in a simulation of slantwise convection. Figure 3.16a shows the smooth vertical-motion field from a 24-h simulation that employed 75 layers of equal depth, and a horizontal grid increment of 10 km, while Fig. 3.16b shows the same field for an analogous simulation that employed only 25 layers of equal depth. The truncation error associated with the poor vertical resolution of the feature has created a noisy vertical motion field and associated gravity waves (not shown). In the first experiment, $\Delta z = \Delta z_{opt}$, while in the second $\Delta z = 0.33\Delta z_{opt}$. A third experiment used 25 layers, as in Fig. 3.16b, but the horizontal grid increment was increased from 10 km to 30 km. In this case, the use of both the coarser horizontal and vertical grid increments produced $\Delta z = \Delta z_{opt}$ and a smooth solution (not shown), but one with considerably less amplitude than in Fig. 3.16a because of the overall coarser resolution.

Even though these resolution-consistency studies clearly isolate the importance of this source of error for specific meteorological cases, this effect is not always responsible when model solutions appear to degrade with increasing horizontal resolution. For example, other problems such as the inappropriateness of resolution-dependent physical-process parameterizations may also be encountered as the horizontal resolution is increased in a model. And, it is well known that conventional objective measures of forecast skill, such as bias, mean-absolute error, and root-mean-square error, show lower skill for forecasts that have small-scale structures (i.e., that can result from high horizontal resolution), compared with forecasts that are smoother (Rife *et al.* 2004).

Many numerical models used for research and operational forecasting, especially those applied on the mesoscale, do not satisfy these consistency relations. For example, in the model simulations associated with Fig. 3.16, 750 vertical levels would have been required for consistency if the grid increment had been 1 km. Nevertheless, it is not clear what the

practical consequences are of not satisfying the consistency relations, especially in light of the general success of numerical model studies and forecasts where the vertical resolution has presumably been insufficient. Lindzen and Fox-Rabinovitz (1989) suggest that this success is partly attributable to the use of horizontal numerical diffusion that limits the effective horizontal resolution of features. Complicating a desire to comply with the consistency constraint is the fact that, with operational modeling, in contrast to focussed case studies, there are many features with different slopes across the computational domain. And, the vertical grid spacing in most models varies considerably with distance above Earth's surface. Furthermore, two-way interacting model grids in a grid nest have different horizontal grid increments, but they typically must utilize the same vertical grid structure. Thus, Δz_{opt} will vary considerable with time and place in the same model integration.

For a given amount of computational resources, modelers tend to maximize the horizontal resolution at the expense of the vertical resolution – the horizontal grid increment has tended to be the Holy Grail of modeling. However, for a given model application, even though there may not be computational resources available to completely satisfy the consistency criterion, the above experimental evidence suggests that modelers should not ignore this issue. Instead, a compromise should be made between the vertical and horizontal resolutions, where the sensitivity of the model solution to different choices should be evaluated using case studies of the prevailing meteorological processes in a given area. With the trend toward using an ensemble of coarser horizontal-resolution model runs, this consistency issue may, at least temporarily, become less critical.

3.2.2 Spectral methods

Early approaches to global grid-point modeling included the use of latitude–longitude grids with reduced time steps near the poles, quasi-homogeneous spherical-geodesic and cubed-sphere grids, and composite meshes. At the time that they were proposed, all of these approaches were problematic in some respect, and this resulted in the dominance of spectral modeling after the spectral-transform method (Machenhauer 1979) was introduced by Eliasen *et al.* (1970) and Orzag (1970), and implemented by Bourke (1974). This method dominated global modeling for decades, and still is widely used even though refinements to the above grid-point approaches and new computer architectures have resulted in the adoption of other options.

The spectral form of a series of meteorological equations is obtained by substituting finite expansions of the dependent variables, typically using double Fourier series or Fourier–Legendre functions (called *basis functions*) to represent the horizontal spatial variation. The orthogonality of these functions allows the derivation of a series of coupled, nonlinear, ordinary differential equations in the expansion coefficients, which are functions of time and the vertical coordinate. The equations are numerically integrated forward in time using conventional finite differencing in time and in the vertical dimension. Such models are initialized by forward transforming the standard dependent variables from physical space (grid-point values) to the transform space (expansion coefficients), and interpretable forecast fields are obtained by inverse transforming back to physical space.

Before further discussing the spectral method and its strengths and weaknesses, the one-dimensional shallow-fluid equations will be used to illustrate the approach. A Fourier series will be used as the basis function. In general, a one-dimensional field can be represented by the following series:

$$A(x) = \sum_{m=0}^{\infty} (a_m \cos mkx + b_m \sin mkx), \tag{3.5}$$

where the Fourier coefficients a_m and b_m are real, m is an integer wavenumber, $k = 2\pi/L$, and L is the domain length. If we add and subtract the two exponentials in Euler's relations

$$e^{imkx} = \cos mkx + i \sin mkx \text{ and}$$

$$e^{-imkx} = \cos mkx - i \sin mkx,$$

where $i = \sqrt{-1}$, the following expressions are obtained:

$$\cos mkx = \frac{e^{imkx} + e^{-imkx}}{2} \text{ and} \tag{3.6}$$

$$\sin mkx = \frac{e^{imkx} - e^{-imkx}}{2i}. \tag{3.7}$$

Substituting Eqs. 3.6 and 3.7 into Eq. 3.5 produces

$$A(x) = \sum_{m=0}^{\infty} \left[\left(\frac{a_m}{2} + \frac{b_m}{2i} \right) e^{imkx} + \left(\frac{a_m}{2} - \frac{b_m}{2i} \right) e^{-imkx} \right]. \tag{3.8}$$

Doing some algebraic manipulation, and defining

$$C_0 = a_0,$$

$$C_m = \frac{a_m - ib_m}{2}, \text{ and}$$

$$C_{-m} = \frac{a_m + ib_m}{2},$$

allows Eq. 3.8 to be rewritten as

$$A(x) = C_0 + \sum_{m=1}^{\infty} C_m e^{imkx} + \sum_{m=1}^{\infty} C_{-m} e^{-imkx}.$$

The index in the last summation can be multiplied by −1, allowing the terms to be combined into

$$A(x) = \sum_{m=-\infty}^{\infty} C_m e^{imkx} = \sum_{|m| \le \infty} C_m e^{imkx}.$$

Before substituting this solution into a set of meteorological equations, it will be assumed that the Fourier coefficients C_m are a function of time, and a maximum value of the wavenumber m will be defined. Recall that m represents the number of waves over the domain length, L, so defining the highest wavenumber (or smallest wavelength) in the series determines the model resolution.

The shallow-fluid equations (Eqs. 2.36–2.38) will be converted to spectral form by representing the dependent variables in terms of truncated versions of the above Fourier series. Specifically,

$$u(x,t) = \sum_{|m| \le K} U_m(t) e^{imkx}, \tag{3.9}$$

$$v(x,t) = \sum_{|m| \le K} V_m(t) e^{imkx}, \text{ and} \tag{3.10}$$

$$h(x,t) = \sum_{|m| \le K} H_m(t) e^{imkx}, \tag{3.11}$$

where $U_{-m}(t)$ is the complex conjugate of $U_m(t)$ and K is the highest permitted wavenumber. Thus, the time dependence of the dependent variables will be represented through the complex Fourier coefficients U, V, and H, and the space dependence will be represented analytically in terms of the sinusoidal variation of the sine and cosine functions embodied by the exponential. To obtain the spectral equations, substitute Eqs. 3.9–3.11 into the differential equations Eqs. 2.36–2.38, multiply each equation by e^{-ijkx}, where j is any arbitrary wavenumber, and integrate each equation with respect to x, from 0 to L. After using the following condition resulting from the orthogonality of the exponential,

$$\int_0^L e^{imkx} e^{inkx} dx = \begin{Bmatrix} 0; & m = -n, \\ L; & m = -n, \end{Bmatrix},$$

the shallow-fluid equations in spectral form are obtained. The original partial differential equations are now ordinary differential equations, where

$$\dot{U}_m = -ik \sum_{\substack{|p| \le K \\ |m-p| \le K}} (m-p) U_p U_{m-p} + f V_m - ikgm H_m,$$

$$\dot{V}_m = -ik \sum_{\substack{|p| \le K \\ |m-p| \le K}} (m-p) U_p V_{m-p} - f U_m - g \frac{\partial H}{\partial y} \delta_m, \text{ and}$$

$$\dot{H}_m = -ik \sum_{\substack{|p| \le K \\ |m-p| \le K}} (m-p) U_p H_{m-p} + V_m \frac{\partial H}{\partial y} - ik \sum_{\substack{|p| \le K \\ |m-p| \le K}} (m-p) H_p U_{m-p},$$

where $\delta_m = 1$ for $m = 0$ and $\delta_m = 0$ for $m \neq 0$. And,

$$\dot{U}_m = \frac{d}{dt} U_m(t), \text{ etc.}$$

The spatial derivatives have been evaluated analytically, leaving only the time derivatives to be approximated with finite differencing.

The basis functions used in most global spectral atmospheric models are *spherical harmonics*, a combination of sines and cosines in a Fourier series to represent the zonal variation of the dependent variables, and *associated Legendre functions* for the meridional variation. Such spherical harmonics take the form

$$\psi(\lambda,\phi) = \sum_{m=-K}^{K} \sum_{n=|m|}^{N(m)} \Psi_n^m \; Y_n^m(\lambda,\phi), \tag{3.12}$$

where

$$Y_n^m(\lambda,\phi) = e^{im\lambda} P_n^m(\sin\phi),$$

ψ is any dependent variable, λ is longitude, ϕ is latitude, m is zonal wavenumber, K is the highest zonal wavenumber, n is the order of the associated Legendre polynomial, $N(m)$ is the highest order of the associated Legendre polynomial, Ψ_n^m are spectral coefficients, Y_n^m are spherical harmonics, and $P_n^m(\sin\phi)$ are the associated Legendre functions of the first kind (which are polynomials). See, for example, Krishnamurti *et al.* (2006a) for the form of the associated Legendre polynomials. Other approaches include the use of two-dimensional Fourier expansions (Cheong 2000, 2006).

The relationship between the number of waves allowed in the meridional direction and the number of waves allowed in the zonal direction defines the type of truncation used in the model. Two types of truncation schemes are typically used in global spectral models – triangular and rhomboidal. The truncation defines the form of $N(m)$ in Eq. 3.12. For triangular truncation, $N(m) = K$ and the same number of waves is allowed in each direction. For rhomboidal truncation $N(m) = |m| + K$ and the number of meridional waves is greater than the zonal wavenumber by a constant factor. Figure 3.17 shows graphically the differences in these truncations. The triangular truncation is more commonly used today, for reasons discussed in Krishnamurti *et al.* (2006a). Triangular truncation provides a resolution that is uniform on the sphere and the same in the zonal and meridional directions. In contrast, rhomboidal truncation produces higher resolution near the poles.

Solving the nonlinear terms in models that are completely spectral is computationally intensive, and makes the process prohibitive. And, local-forcing processes (e.g., latent heat

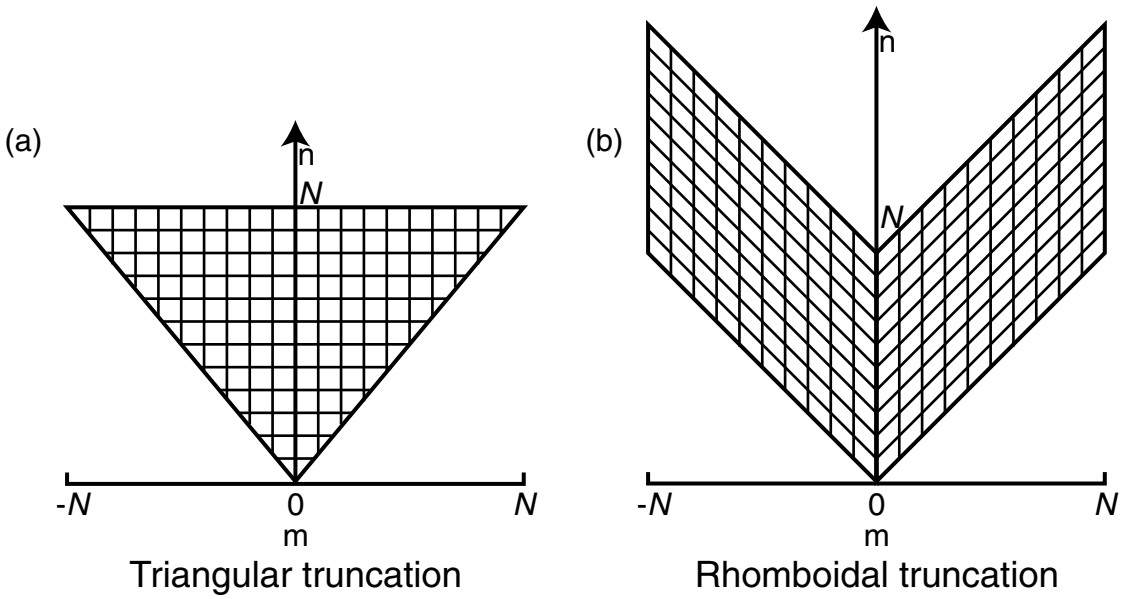

Fig. 3.17 Schematic showing the relationship between the permitted wavenumbers in the zonal (m) and meridional (n) directions for triangular and rhomboidal truncations.

release, differential surface heat fluxes), some of which are discontinuous, are only possible to represent in physical space. These problems have been resolved through the development of *pseudospectral* models that treat some processes in spectral space and others in physical, or grid-point, space. This approach is called the *transform method* because it involves transformations between spectral and physical space every time step. In particular, after extrapolations in time produce new expansion coefficients, the dependent variables are converted from spectral space to values defined on an appropriate grid using an inverse transform (e.g., Eqs. 3.9–3.11 for the one-dimensional Fourier expansion, or Eq. 3.12 for the two-dimensional expansion with spherical harmonics). The transform method works as follows, for a nonlinear term such as $u\partial u/\partial x$.

- Calculate $\partial u/\partial x$ in spectral space, and inverse transform it and u to physical space on an appropriate grid.
- Calculate $u\partial u/\partial x$ on the grid by multiplication.
- Transform $u\partial u/\partial x$ back to spectral space, providing a value for each predicted wavenumber.
- Add this number to the tendency equations for the u coefficients for each wavenumber, along with the contributions from other terms.
- Predict new values for the u coefficients for each wavenumber.

No finite-difference approximations to the derivatives are required in this procedure. A number of particular spectral-transform methods exist. For example, Swarztrauber (1996) compared the accuracy of nine methods for solving the shallow-water equations.

For alias-free solutions to the nonlinear terms, with both triangular and rhomboidal truncations, the number of grid points in the zonal direction on the transform grid must be

$3N + 1$. In the meridional direction, $(3N + 1)/2$ points are needed for triangular truncation and $5N/2$ are needed for rhomboidal truncation. The points are equally spaced in the zonal direction, but are not in the meridional direction. Legendre polynomial solutions on this Gaussian grid are exact. However, the use of simple latitude–longitude transform grids for the above purpose has the previously described pole problem (a Gaussian grid would look similar to the one in Fig. 3.8). Constructing reduced Gaussian grids, of the same type shown schematically in Fig. 3.9 for a pure latitude–longitude grid, and described in Williamson (2007), is one way of addressing this. Figure 3.18 shows an example of a reduced Gaussian grid. Hortal and Simmons (1991) showed that the use of a reduced grid in

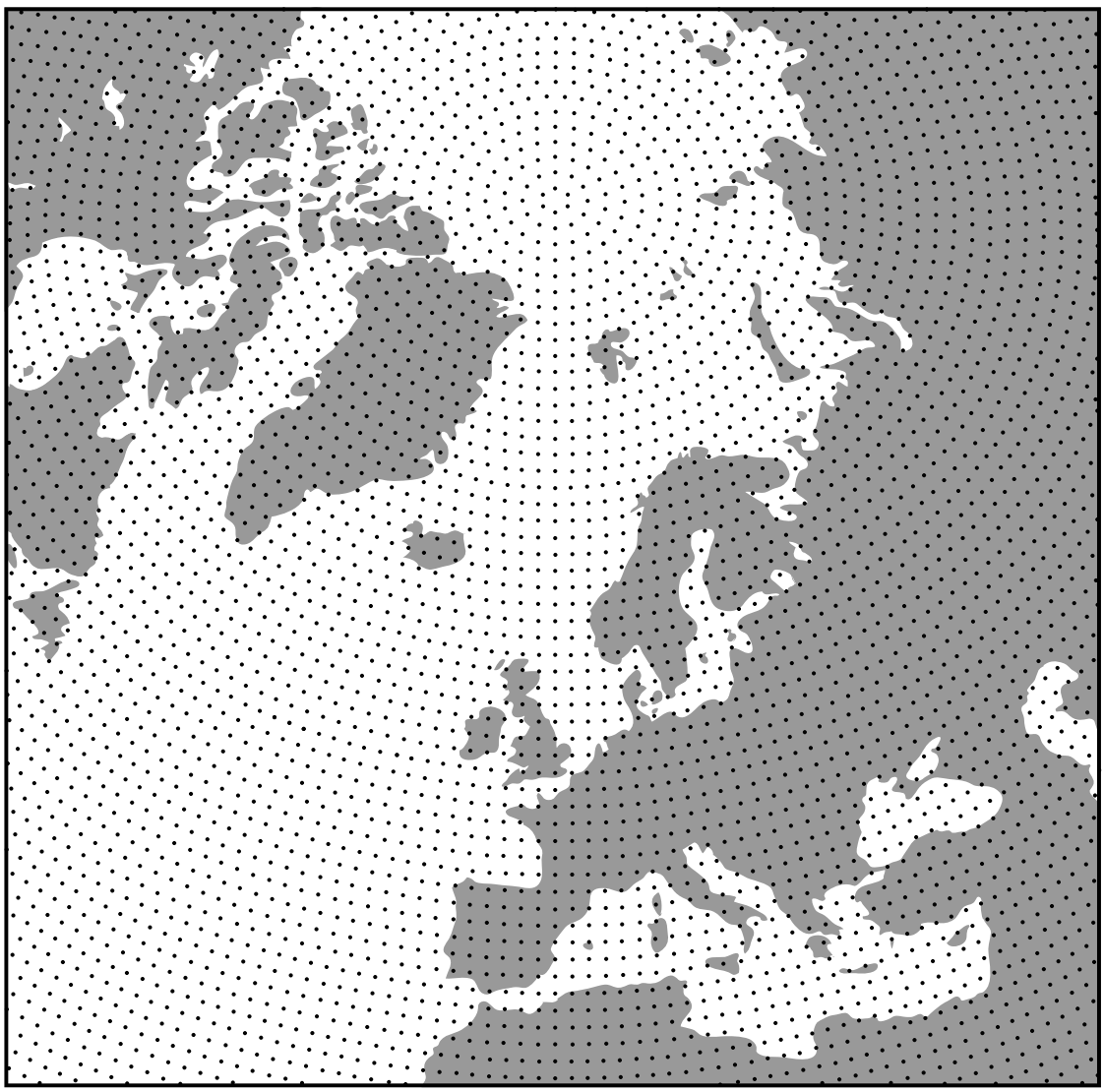

Fig. 3.18 An example of a reduced Gaussian T106 grid. From Hortal and Simmons (1991).

a spectral model produced no significant loss of accuracy relative to the use of a full grid. However, reducing the number of grid points leads to error in the calculation of nonlinear terms, and therefore some aliasing. One constraint for spectral applications of reduced grids is that the number of grid points on latitude circles must be consistent with the requirements of fast Fourier transform algorithms. The use of a relatively uniform grid increment across the sphere in reduced-grid, spectral-transform models is consistent with the fact that triangular truncation of the spherical harmonics produces a uniform resolution in spectral space.

Vertical derivatives in spectral models are typically approximated with standard finite differencing or with finite-element methods. Legendre polynomials and Laguerre polynomials have been used for vertical basis functions, but they both have significant disadvantages. Applications of the finite element approach are discussed in Béland and Beaudoin (1985), Steppeler (1987), and Hartmann (1988).

The following is a typical process by which forecasts are produced with global spectral models.

- Based on the desired resolution in physical space, a spectral truncation is chosen (e.g., triangular with $K = 42$ would be referred to as T42). The numbers of grid points in both the latitudinal and longitudinal directions are chosen, possibly based on a desire to avoid aliasing. For the chosen spectral truncation, the highest-degree Legendre polynomial is identified, and the latitudes where the roots of the polynomial occur are determined. These are the Gaussian latitudes, and are the basis for the Gaussian grid used in the transform.
- Observations of dependent variables are objectively analyzed, in physical space, to a grid.
- Gridded data representing model initial conditions are forward transformed to define expansion coefficients.
- Each time step, dependent variables are inverse transformed from spectral to physical space, and tendency contributions associated with local processes are calculated on the grid for the momentum, thermodynamic, and moisture equations. Such processes that are inherently local, with strong gradients or discontinuities that cannot be represented in spectral space include surface heat, moisture, and momentum fluxes; radiative flux divergence; latent-heat gains or losses; cloud-microphysical and convective processes; etc. Similarly, nonlinear terms are calculated on the grid according to the transform method described above. Terms with vertical derivatives may also be calculated in physical space. Tendency contributions are then transformed back to spectral space.
- Tendency contributions for terms calculated in spectral and physical space are added, and a time extrapolation is performed using standard time-differencing methods.
- At a desired frequency, dependent variables are reverse transformed to physical space and graphically displayed on a map projection to provide information on the state of the forecast and for comparison with observations.

Because modelers tend to think in terms of grid increments rather than wavenumbers for defining horizontal resolution, it is useful to have a simple conversion between global spectral resolution and the equivalent grid spacing. A few alternatives, based on different

interpretations of the meaning of spectral resolution, were given by Laprise (1992).[2] The formulae are

$$L_1 = \frac{2\pi a}{3K+1},$$

$$L_2 = \frac{\pi a}{K},$$

$$L_3 = \frac{2\sqrt{\pi} a}{K+1},$$

$$L_4 = \frac{\sqrt{2}\pi a}{K},$$

where K is as defined above, and a is the radius of Earth. The spectral resolution is expressed in terms of K and the type of truncation. If K is 799, and the truncation is triangular, the resolution is given as T799. For a spectral resolution of T799, the estimates of the equivalent grid increments are L_1 = 16.7 km, L_2 = 25.0 km, L_3 = 28.2 km, and L_4 = 35.4 km. Some advantages and disadvantages of spectral models are summarized below.

Advantages

- There is generally no aliasing of quadratic nonlinear terms, and thus no nonlinear instability.
- There is no spatial truncation error because the derivatives are treated analytically, and thus there is no numerical dispersion of waves.
- Semi-implicit time-differencing schemes are easily implemented.
- There is almost no grid (computational) diffusion.

Disadvantages

- Local-forcing processes (e.g., latent-heat release, differential surface heat fluxes), some of which are discontinuous, are only possible to represent in physical space.
- When a linear combination of waves (e.g., spherical harmonics) is used to represent a large gradient or discontinuity, spurious waves can result (Gibbs phenomenon). In the case of specific humidity, this "spectral ringing" can result in negative values, which are physically impossible. And overshooting the correct solution can lead to spurious precipitation, called *spectral rain*.
- For higher resolutions, spectral models are computationally more demanding than grid-point models.
- Spectral models do not exactly conserve mass or energy.

[2] It should not be surprising that there is lack of agreement on the meaning of "horizontal resolution" in spectral space, given that there is even disagreement about the meaning of the term in the much more intuitive physical (grid-point) space. In addition to the fact that the scales represented by a grid-point model depend on the numerical smoothing and other factors, Pielke (1991) points out that modelers often erroneously use the term resolution to refer to the grid increment.

Lastly, a few regional, that is limited-area, spectral models have been developed and employed for research and operational prediction. One of the most widely used is the US National Centers for Environmental Prediction (NCEP) Regional Spectral Model (RSM; Juang and Kanamitsu 1994, Juang *et al.* 1997, Juang 2000, Roads 2000, Juang and Hong 2001) that has been used for operational weather prediction. The high-resolution RSM is typically used in a nest with a low-resolution global spectral model, with the two spectral models having the same vertical structure and physical processes. The regional model uses a double Fourier series as the basis function, and is defined on a map projection. Multiple spectral nests are possible within the global model. The Scripps Experimental Climate Prediction Center has used the RSM, coupled with the NCEP Global Spectral Model, for seasonal predictions (Roads 2004), and Han and Roads (2004) describe its use for 10-year climate simulations. There are numerous other applications of regional spectral models. The Florida State University nested regional spectral model has been used for weather and seasonal-climate simulations (Cocke 1998, Cocke *et al.* 2007), and the Japan Meteorological Agency has used such a model for operational prediction (Tatsumi 1986). Boyd (2005) provides a table of spectral limited-area models, and Krishnamurti *et al.* (2006a) provide a summary of the modeling process.

3.2.3 Finite-element methods

Finite-element methods were first developed for engineering applications, and have been since adopted for use in some models of ocean and atmospheric processes. They are analogous to spectral modeling methods, both being special cases of the Galerkin procedure in which the dependent variables are approximated by a finite sum of spatially varying basis functions with time-dependent coefficients. For spectral modeling, global (i.e., nonlocal) basis functions are employed, where for finite-element modeling the basis functions are low-order polynomials that are zero except in a localized region. In finite-element modeling, the computational domain is divided into a number of contiguous finite subregions called elements. On each element is defined a simple function, where continuity between functions on adjacent elements is required.

The finite-element method has been used in the operational Canadian Regional Finite Element (RFE) model (Staniforth and Mailhot 1988, Benoit *et al.* 1989, Tanguay *et al.* 1989, Belair *et al.* 1994), in the ECMWF model (Burridge *et al.* 1986), and elsewhere (Staniforth and Daley 1979). Finite-element representations are sometimes used only in the vertical, where finite-difference or spectral methods are employed for the horizontal (Staniforth and Daley 1977, Beland *et al.* 1983, Beland and Beaudoin 1985, Burridge *et al.* 1986, Steppler 1987). Good summaries of the application of finite-element methods in atmospheric models can be found in Cullen (1979), Staniforth (1984), and Hartmann (1988).

3.2.4 Finite-volume methods

In contrast to grid-point models where the prognostic quantity is the value of dependent variables at grid points, with *finite-volume models* it is the integrated value of a variable over a specific finite control volume. The control volumes are typically the traditional

three-dimensional model grid cells, which leads to the fact that finite-volume methods are also referred to as cell-integrated methods. They are especially well suited for applications where it is important to conserve quantities such as mass, total energy, angular momentum, or entropy. Indeed, one of the reasons for the renewed interest in this approach is the significant lack of global mass conservation in models that use the semi-Lagrangian technique. There are two approaches in the finite-volume framework. One is the Departure Cell-Integrated Semi-Lagrangian (DCISL) scheme, and the other is the Flux-Form Semi-Lagrangian (FFSL) scheme. The DCISL and the FFSL differ in terms of how they estimate a property, for example the mass, of the cell at the trajectory's departure point in the semi-Lagrangian transport calculation. If mass conservation is the primary concern, the finite-volume method can be applied to the continuity equation, while conventional semi-Lagrangian grid-point methods are used for the other equations. Two recent examples of dynamical cores that use the finite-volume method are the European HIgh Resolution Limited Area Model (HIRLAM) and the NCAR global Community Atmospheric Model (CAM 3.0, Collins *et al.* 2006b). The HIRLAM employs the DCISL method and the CAM uses a flux-based scheme. The FIM model being developed by NOAA, mentioned earlier, also uses the finite-volume method. See Machenhauer *et al.* (2008) for a thorough summary of the use of finite-volume methods in atmospheric modeling.

3.3 Finite-difference methods

3.3.1 Time-differencing methods

Time-differencing methods can be explicit or implicit, or a combination of both. With *explicit methods*, the prognostic equation can be solved for the value of the dependent variable at the new (most-forward) time, with the new value on the left side of the equation, and the right side consisting of dependent variables defined at current or prior times. With *implicit methods*, dependent variables at the new time level appear on both sides of the equations, and the solution must be obtained iteratively. Semi-implicit techniques solve some terms in the equations explicitly, and some implicitly.

Unless the anelastic or Boussinesq equations described in Chapter 2 are used for a model, the solution contains acoustic waves and external gravity waves that both move with speeds at or near Mach number 1.[3] These meteorologically inconsequential waves can require the use of a small time step because of Courant-number constraints, and thus make the model computations inefficient. The next sections will explain how both explicit and implicit time-differencing methods deal with this problem. With *split-explicit* differencing, only the terms in the equations associated with the acoustic and external-gravity modes are computed with a short time step. The terms related to the meteorological processes use a longer time step, which is consistent with the relatively slow speed of those waves. With

[3] The Mach number is the ratio of the phase speed of a wave to that of a sound wave.

semi-implicit differencing, implicit methods, whose time step is not constrained by the Courant number, are used for the fast acoustic and gravity modes, and explicit time differencing is used for the other terms associated with the slower meteorological waves.

Explicit time differencing

There are two general types of explicit time-differencing approaches. One employs a single computational step, such as the forward-in-time or centered-in-time methods described earlier, to arrive at the new values of the dependent variables at the next time level. Another single-step scheme is the Adams–Bashforth method evaluated in Durran (1991). The other type uses multiple steps. When there are two steps, they are called *predictor–corrector schemes*. The first step is the predictor step and the second is the corrector step. Even though the multi-step methods involve a greater number of arithmetic operations, and therefore have greater computational expense, their numerical properties are superior in some respects to those of the single-step methods. In the following equations, F represents the finite-difference approximation to all the terms on the right side of any of the model prognostic equations (the forcing), θ is any dependent variable, and θ^* is an intermediate solution. Equation 3.1 above shows an example of the centered-in-time, centered-in-space, single-step explicit scheme. Equation 3.13 is a representation for any such equation:

$$\theta_j^{\tau+1} = \theta_j^{\tau-1} + 2\Delta t F_j^{\tau}. \tag{3.13}$$

There are many multi-step schemes, and some have numerous variations with different accuracies. For example, one approach is the Lax–Wendroff scheme (Lax and Wendroff 1960):

$$\theta^*_j = \frac{1}{2}(\theta^{\tau}_{j+1} + \theta^{\tau}_{j-1}) + \frac{\Delta t}{2} F_j^{\tau} \qquad \text{(predictor step)}, \tag{3.14}$$

$$\theta_j^{\tau+1} = \theta_j^{\tau} + \Delta t F^*_j \qquad \text{(corrector step)}. \tag{3.15}$$

In the first step, a forward extrapolation is made from time τ for one-half the time step. In the second, these forecast values are used to calculate the tendency for the extrapolation from τ to $\tau + 1$, which is a centered time step. Another two-step scheme is the Euler-backward (or Matsuno 1966) method. Here, we have

$$\theta^*_j = \theta_j^{\tau} + \Delta t F_j^{\tau}, \tag{3.16}$$

$$\theta_j^{\tau+1} = \theta_j^{\tau} + \Delta t F^*_j. \tag{3.17}$$

Another commonly used method, with many variations, is the Runge–Kutta scheme. One, described by Wicker and Skamarock (2002), is

$$\theta^*_j = \theta_j^{\tau} + \frac{\Delta t}{3} F_j^{\tau}, \tag{3.18}$$

$$\theta^{\dagger}_j = \theta_j^{\tau} + \frac{\Delta t}{2} F^*_j, \tag{3.19}$$

$$\theta_j^{\tau+1} = \theta_j^{\tau} + \Delta t F^{\dagger}_j . \tag{3.20}$$

This scheme is used in the community Advanced Research version of the Weather Research and Forecasting (WRF) model (Skamarock *et al.* 2008).

Another type of explicit time differencing is the so-called split-explicit method. The motivation for this approach, which is also called *time splitting*, is based on the fact that compressible, nonhydrostatic equations support sound waves (Mach number 1) and fast-moving external-gravity waves, as well as of course the meteorological waves (e.g., advective waves, Rossby waves, and internal-gravity waves that effect geostrophic adjustment) whose Mach numbers rarely exceed 0.3 even in the fastest jet-stream winds. Because the sound waves and external-gravity waves have small amplitudes and are not meteorologically significant, it is wasteful of computing resources to use the very-small time step that is needed to satisfy the Courant condition for these waves. There are a few methods to deal with this problem. One is to use split-explicit methods that integrate different terms in the equations using different time steps. The few terms associated with sound and external-gravity wave propagation are integrated with a small time step, and the rest of the terms that represent meteorological processes are integrated with a larger time step. All explicit methods have linear-stability criteria that are constrained by the Courant number. A number of nonhydrostatic models use this approach, including the WRF (Skamarock and Klemp 2008), the Mesoscale Model Version 5 (MM5, Dudhia 1993), the Lokal Modell (LM, Doms and Schättler 1997), the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS, Hodur 1997), and the Advanced Regional Prediction System (ARPS, Xue *et al.* 2000). Additional discussion of split-explicit time differencing can be found in Marchuk (1974), Klemp and Wilhelmson (1978), Wicker and Skamarock (1998), Klemp *et al.* (2007), Purser (2007), and Skamarock and Klemp (1992, 2008).

Implicit and semi-implicit time differencing

An example of an explicit treatment of a one-dimensional linear advection equation would be

$$u_i^{\tau+1} = u_i^{\tau-1} - \frac{U\Delta t}{\Delta x}(u_{i+1}^{\tau} - u_{i-1}^{\tau}) ,$$

where there are no variables defined at $\tau + 1$ on the right side of the equation. In contrast, the following form of the linear advection equation is implicit because $\tau + 1$ values of dependent variables appear on the right:

$$u_i^{\tau+1} = u_i^{\tau} - \frac{U\Delta t}{2}\left(\frac{u_{i+1}^{\tau+1} - u_{i-1}^{\tau+1}}{2\Delta x} + \frac{u_{i+1}^{\tau} - u_{i-1}^{\tau}}{2\Delta x}\right). \tag{3.21}$$

The approximation to the space derivative is represented as a time average of the derivative evaluated at the forward time and at the current time, so that it applies at $\tau + 1/2$. Such schemes applied to the full set of equations are typically unconditionally stable, and can use long time steps that are unconstrained by the Courant number. Because implicit equations need to be solved iteratively (e.g., for $u_i^{\tau+1}$), they require much more computation

per time step than do explicit equations. Specifically, three-dimensional Helmholtz equations must be solved each time step. Unfortunately, for fully three-dimensional problems, with six or seven variables and complex equations, the computational savings from the longer time step are typically more than offset by the greater computational cost per time step. Motivated by this problem, Marchuk (1965) and many others since then have developed *semi-implicit* schemes, which do provide computational advantages compared with explicit methods. With semi-implicit approaches, some terms are treated explicitly and some are treated implicitly. That is, in the finite-difference equation, implicit terms use averaging operators similar to the one above, while explicit terms have the conventional formulation. Those terms treated implicitly are typically the ones associated with processes that motivate the use of the implicit method in general; that is, those terms associated with fast-moving acoustic and external-gravity waves that would normally demand the use of a short time step. The rest of the terms, which are related to the slower meteorological processes, are treated explicitly. The time step that is stable for the explicit terms is also stable for the implicit terms because they are stable for any time step. Robert (1979) analyzed the following semi-implicit form of the shallow water equations:

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - \overline{\frac{\partial \phi}{\partial x}}^{t},$$

$$\frac{\partial v}{\partial t} = -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - \overline{\frac{\partial \phi}{\partial y}}^{t},$$

$$\frac{\partial \phi}{\partial t} = -u\frac{\partial \phi}{\partial x} - v\frac{\partial \phi}{\partial y} - (\phi - \phi_0)\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) - \phi_0\overline{\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)}^{t}.$$

Here, ϕ_0 is a mean geopotential height and the overbarred height gradient and divergence terms are the implicit time-mean expression of Eq. 3.21. The rest of the terms are treated explicitly. Robert (1979) linearized this set of equations and showed that the gravity waves have no time-step restriction and that the advection has the standard CFL time-step restriction. The shallow-fluid equations are incompressible and have no acoustic mode.

3.3.2 Space-differencing methods

Eulerian space differencing

Eulerian models calculate the transport (advection) terms in the equations at points that have fixed horizontal and vertical coordinates. This is the approach described thus far in this chapter, regardless of whether the models are entirely grid-point based or use the spectral-transform method. Equations such as the following, for any independent variable α, are solved for the values of the time derivatives at specific points, where the advection terms and the other forcing terms (F) apply at the same locations:

$$\left.\frac{\partial \alpha}{\partial t}\right|_{x,y,z} = -\vec{V}\cdot\nabla\alpha + F(x, y, z, t).$$

In all cases, the advection (and other) terms require adherence to a stability criterion that is related to the time step, regardless of whether such a short time step is needed for accuracy (truncation-error control). Because of this inherent stability limitation of Eulerian methods, and the associated computational liability, the semi-Lagrangian approach described below has become widely used.

Lagrangian and semi-Lagrangian space differencing

With purely *Lagrangian methods*, changes in the properties of individual moving parcels of air are calculated. That is, our reference frame is air parcels and not grid points. In this case, the following equation would apply, where the material derivative, or the derivative following the parcel, is on the left side:

$$\frac{d\alpha}{dt} = F(x, y, z, t) . \tag{3.22}$$

For a perfectly conserved quantity,

$$\frac{d\alpha}{dt} = 0 . \tag{3.23}$$

That is, the value of α does not change following the parcel. Integrating Eq. 3.23 in time, as part of a larger set of equations, would simply involve estimating how the conserved α field is redistributed by the wind. Such a Lagrangian forecast system could be initialized by beginning with regularly spaced parcels, and assigning values of α to parcels based on standard initialization techniques. However, after a short forecast period the parcels would become very unevenly distributed, providing unacceptable contrasts in spatial resolution. This problem motivated the development of *semi-Lagrangian* space differencing, where a completely new set of regularly spaced parcels is chosen each time step. One approach is, at each time step, to initially define the parcels at grid points, and move each parcel in space for one time step based on the prevailing velocity field. The new parcel positions will, of course, not be at grid points, but the parcels will have retained their original value of α so a new spatial distribution of α will be defined based on the irregularly spaced end-points of the trajectories. The values of α at trajectory end-points are then spatially interpolated to the original grid points, defining new regularly spaced parcels from which to begin the next time step. An alternative approach is to also begin with parcels at grid points, but calculate one-time-step back-trajectories using the same wind field as in the above process. Then the α at the back-trajectory's end-point is defined by spatial interpolation from the current grid-point values, and this value is assigned to the grid point (i.e., the value of α is the same at both ends of the trajectory). The latter approach is more common because it is more straightforward to interpolate from a regular grid to irregularly located points, than the opposite.

For the typical situation where there are forcing terms (F), and α is not conserved for each parcel (Eq. 3.22), a schematic one-dimensional finite-difference equation based on a trapezoidal integration approach would be the following, where $x_j = j\Delta x$ refers to the

position at grid points, $t^n = n\Delta t$, and $\tilde{x}^n$ is the estimate of the starting position of the trajectory at time t^n (notation based on Durran 1999):

$$\frac{\alpha(x_j, t^{n+1}) - \alpha(\tilde{x}_j^n, t^n)}{\Delta t} = \frac{1}{2}[F[x_j, t^{n+1}] + F(\tilde{x}^n, t^n)].$$

Here, the forcing is defined as an average of the values applicable at the beginning and end of the trajectory. If this is an equation for a chemical species, α, the forcing term would represent sources or sinks of that material. If α is a meteorological variable, the forcing would represent all the standard terms on the right side of the equation, other than advection of course. These terms would be calculated using standard Eulerian differencing methods on the grid. The value of F at grid point x_j would be calculated directly, and the value of F at $\tilde{x}^n$ would be defined by interpolation from adjacent grid points. Unfortunately, this form of the equation is implicit (variables defined at t^{n+1} are on both sides of the equation) and would require additional computational work in order to solve it. Alternatively, the following explicit, centered-in-time approach could be solved explicitly. The time step is only limited by the constraints that trajectories cannot cross, and trajectory end-points need to be within the grid:

$$\frac{\alpha(x_j, t^{n+1}) - \alpha(\tilde{x}_j^{n-1}, t^{n-1})}{2\Delta t} = F[\tilde{x}_j^n, t^n].$$

As a result of the pioneering work of Robert (1981, 1982) and others, the semi-Lagrangian method has become an extremely popular approach for global and limited-area modeling, where some of the reasons are:

- it is often more efficient than competing Eulerian schemes because the CFL condition associated with advection terms is avoided,
- it can be used with both grid-point and spectral-transform methods,
- it may be combined with semi-implicit methods that are applied to the pressure-gradient and velocity-divergence terms, and
- the primary source of nonlinear instability, the nonlinear advection terms, does not exist.

Conversely, there are criticisms of semi-Lagrangian methods in that they generally fail to exactly conserve energy and mass. Summaries of semi-Lagrangian methods can be found in Staniforth and Côté (1991), Durran (1999), and Williamson (2007).

Grid staggering methods

Grid staggering involves defining different dependent variables on different grids. Typically, the points in one mesh are offset from those in the other by $0.5\Delta x$. Figure 3.19 shows a one-dimensional schematic of an approach to staggering. For the unstaggered grid shown in Fig. 3.19a, calculation of an advection term such as $u\partial\theta/\partial x$ with a centered, three-point method would require differencing across a $2\Delta x$ interval. For the staggered grid in Fig. 3.19b, the derivative can be calculated by differencing across a $1\Delta x$ interval. This halves the effective grid increment for such terms, increasing the spatial resolution and decreasing the effects of truncation error on the solution. Also, a benefit of

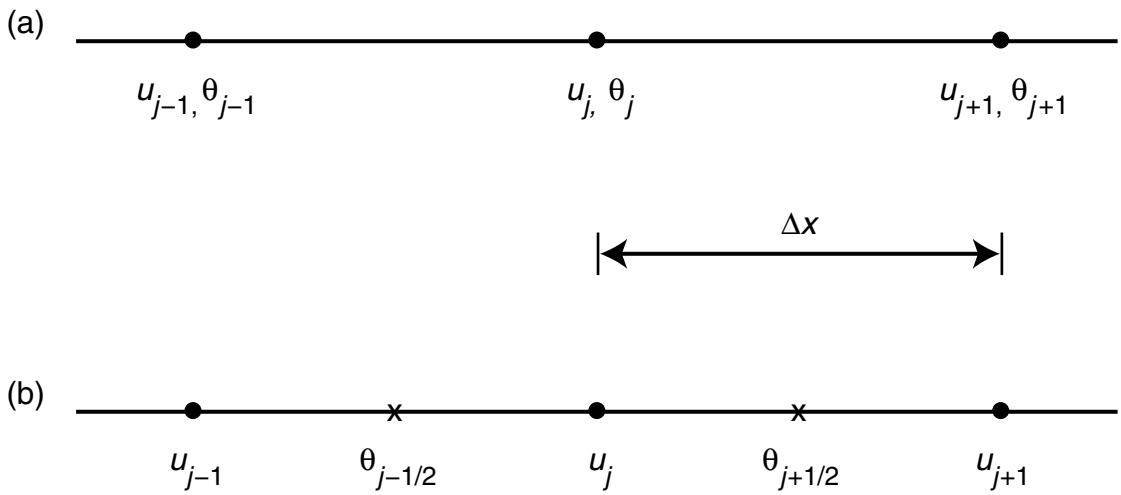

Fig. 3.19 Schematic of one-dimensional unstaggered (a) and staggered (b) grids. For the staggered grid, the mass-field variable (θ) is offset by one-half grid increment from the momentum variable (u). Adapted from Durran (1999).

staggering the horizontal and vertical velocity in hydrostatic models, described by Pielke (2002a), is that, when the vertical velocity is diagnosed from the horizontal divergence by integrating the continuity equation, lateral boundary values for the horizontal velocity have no direct impact on the vertical velocity. This is illustrated in Fig. 3.20. For

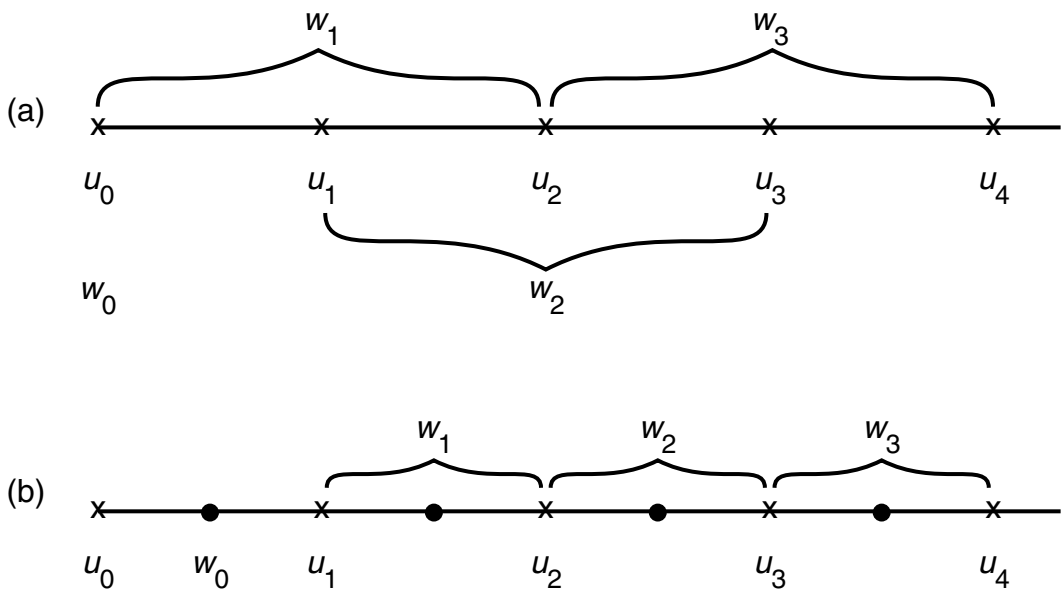

Fig. 3.20 Schematic showing horizontal and vertical velocities on a one-dimensional unstaggered grid (a) and on one type of staggered grid (b). The subscript shows the position of the grid points relative to the left boundary. Adapted from Pielke (2002a).

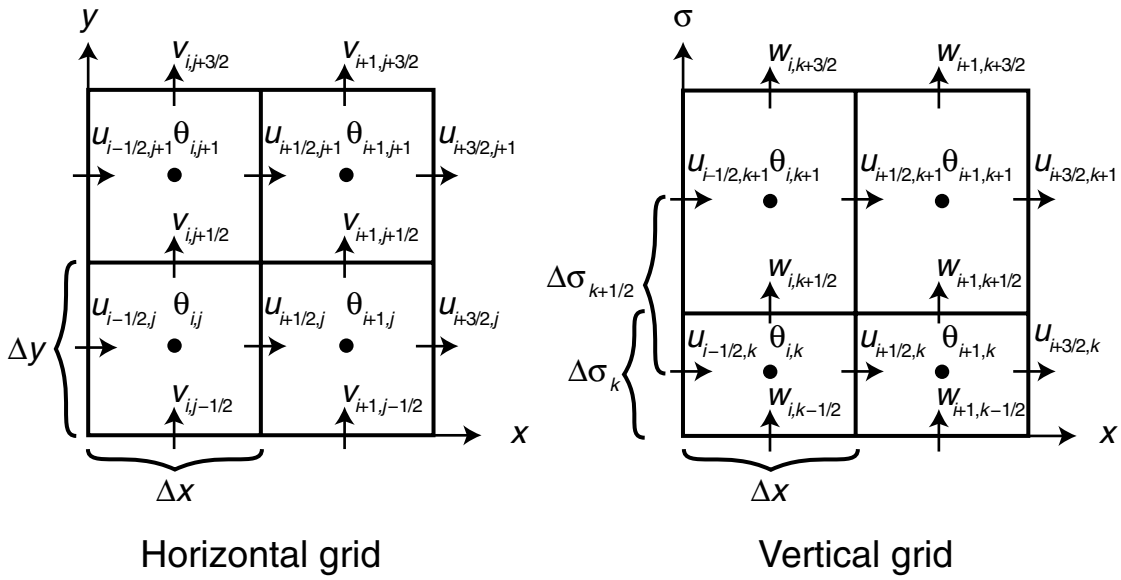

Fig. 3.21 The Arakawa-C grid-staggering method. Adapted from Skamarock *et al.* (2008).

unstaggered grids, where u and w are defined at the same points, Fig. 3.20a shows that the vertical velocity at the first interior point (w_1) is calculated using a lateral boundary value of $u(u_0)$. In contrast, for the staggered grid the vertical velocity at the first interior point is calculated using only nonboundary values (Fig. 3.20b). Figure 3.21 shows an example of a standard approach to horizontal and vertical grid staggering for a three-dimensional model. The wind components (u, v) are defined at the locations of the vectors in the figure, and the mass variables (q, p, θ) are defined at the locations defined by the dots. This is called the Arakawa-C grid, where other alternative staggering methods are described in Arakawa and Lamb (1977) and Haltiner and Williams (1980). Unfortunately, the smaller distance over which the derivatives are calculated means that the effective grid increment is smaller, and therefore the time steps need to be smaller in order for the CFL condition to be satisfied. Nevertheless, the increase in effective resolution, relative to an unstaggered grid, is gained without the use of additional grid points, which would require more computer storage and more computations. Virtually all contemporary grid-point models use staggered meshes.

3.4 Effects of the numerical approximations

This section focusses on the important subject of how the numerical methods that are employed to integrate the equations can affect the model solution in various ways. The discussion of truncation error shows how the derivatives in the equations are incorrectly estimated by finite-difference approximations. Then is described how each term in the

equations possesses criteria, based on model parameters and atmospheric conditions, that must be met in order for the model solution to be stable. The effects of the numerical methods on the phase and group speeds of meteorological waves in the model solution are described, illustrating that for some differencing schemes wave energy can even move in the wrong direction. The nonlinear interaction of waves on a grid can lead to the erroneous accumulation, through aliasing, of wave energy in small wavelengths, and this can lead to the problem of nonlinear instability. And, the concept of horizontal diffusion (the spread, and smoothing, of properties on the grid) is discussed because it is a process that can remove correct small-scale information in the model solution, and it can be used to control numerical problems in the model solution. Lastly, the strengths and weaknesses are summarized of the various vertical coordinates used in models.

3.4.1 Truncation error

Because the equations that govern atmospheric processes are differential equations, with derivatives in most of the terms, approximating the continuous space and time derivatives with finite-difference expressions represents a considerable potential source of error in the modeling process. It is straightforward to quantify this error by using Taylor's theorem, which defines a polynomial that approximates any function over an interval. A remainder term in the polynomial represents the difference between the values of the function and the approximation. The following polynomial is called *Taylor's series*, where f is any meteorological variable in the derivative terms of Eqs. 2.1–2.6, and the series can be written for any independent variable:

$$f(x) = f(a) + (x-a)\frac{\partial f(a)}{\partial x} + \frac{(x-a)^2}{2!}\frac{\partial^2 f(a)}{\partial x^2} + \frac{(x-a)^3}{3!}\frac{\partial^3 f(a)}{\partial x^3} + \cdots$$

$$\cdots + \frac{(x-a)^n}{n!}\frac{\partial^n f(a)}{\partial x^n} + R(n,x)\,. \tag{3.24}$$

It states that the value of a function, f, at any point, x, can be approximated by using the known value and derivatives at point a. In the case of an infinite series, the expression would be exact. For a series truncated at n terms, there is a residual, R, that defines the error. The truncation error will be defined here for three finite-difference approximations to the derivative: two-point, three-point, and five-point formulae.

For a two-point approximation, let $x = a + \Delta x$ in Eq. 3.24, truncate the series by dropping second-order and higher terms, and solve for the derivative, to obtain the following:

$$\frac{\partial f(a)}{\partial x} = \frac{f(a+\Delta x) - f(a)}{\Delta x}\,. \tag{3.25}$$

This is called the forward-in-space differencing formula, which has first-order accuracy because second-order and higher terms in Taylor's series were dropped. An analogous backward-in-space formula results from letting $x = a - \Delta x$.

A three-point differencing scheme can be obtained by writing Taylor's series as

$$f(a+\Delta x) = f(a) + \Delta x\frac{\partial f(a)}{\partial x} + \frac{(\Delta x)^2}{2!}\frac{\partial^2 f(a)}{\partial x^2} + \frac{(\Delta x)^3}{3!}\frac{\partial^3 f(a)}{\partial x^3} + \cdots \tag{3.26}$$

and

$$f(a-\Delta x) = f(a) - \Delta x\frac{\partial f(a)}{\partial x} + \frac{(\Delta x)^2}{2!}\frac{\partial^2 f(a)}{\partial x^2} - \frac{(\Delta x)^3}{3!}\frac{\partial^3 f(a)}{\partial x^3} + \cdots . \tag{3.27}$$

Subtracting the two series produces

$$f(a+\Delta x) - f(a-\Delta x) = 2\Delta x\frac{\partial f(a)}{\partial x} + \frac{2(\Delta x)^3}{3!}\frac{\partial^3 f(a)}{\partial x^3} + \cdots .$$

Solving for $\partial f(a)/(\partial x)$ provides

$$\frac{\partial f(a)}{\partial x} = \frac{f(a+\Delta x) - f(a-\Delta x)}{2\Delta x} - \frac{(\Delta x)^2}{3!}\frac{\partial^3 f(a)}{\partial x^3} + \cdots . \tag{3.28}$$

Truncating this equation after the first term on the right produces the following approximation, which we say has second-order accuracy because we ignore the third-order and higher terms in the series. This is called a three-point approximation to the derivative because it spans points $x - \Delta x$, x, and $x + \Delta x$:

$$\frac{\partial f(a)}{\partial x} = \frac{f(a+\Delta x) - f(a-\Delta x)}{2\Delta x}.$$

One way of calculating the effect of truncating the series on the accuracy of the derivative is to compare the value of the derivative from this approximation with the exact value. Let $f = A\cos kx$, where $k = 2\pi/L$ and L is the wavelength. The exact value of the derivative is

$$\frac{\partial f}{\partial x} = -kA\sin kx, \tag{3.29}$$

where the approximation is

$$\frac{\Delta f}{\Delta x} = \frac{A\cos k(x+\Delta x) - A\cos k(x-\Delta x)}{2\Delta x}. \tag{3.30}$$

Using trigonometric identities it can be shown that

$$\frac{\dfrac{\Delta f}{\Delta x}}{\dfrac{\partial f}{\partial x}} = \frac{\sin k\Delta x}{k\Delta x}, \tag{3.31}$$

where, as $\Delta x/L \to 0$, $k\Delta x \to 0$ and $\sin k\Delta x \to k\Delta x$. That is, as the argument of the sine function approaches zero, so does the function itself, and the ratio approaches unity. This ratio defines the truncation error because it represents the error in the finite-difference approximation to the derivative that is associated with the truncation made in Taylor's series. Thus, for a wave of length L that is defined by many grid points, the ratio of the approximation to the derivative and the exact derivative is near unity. Figure 3.22 shows this ratio for different wavelengths. For a given grid increment, the derivatives of long waves are clearly better represented than the derivatives of shorter waves. For example,

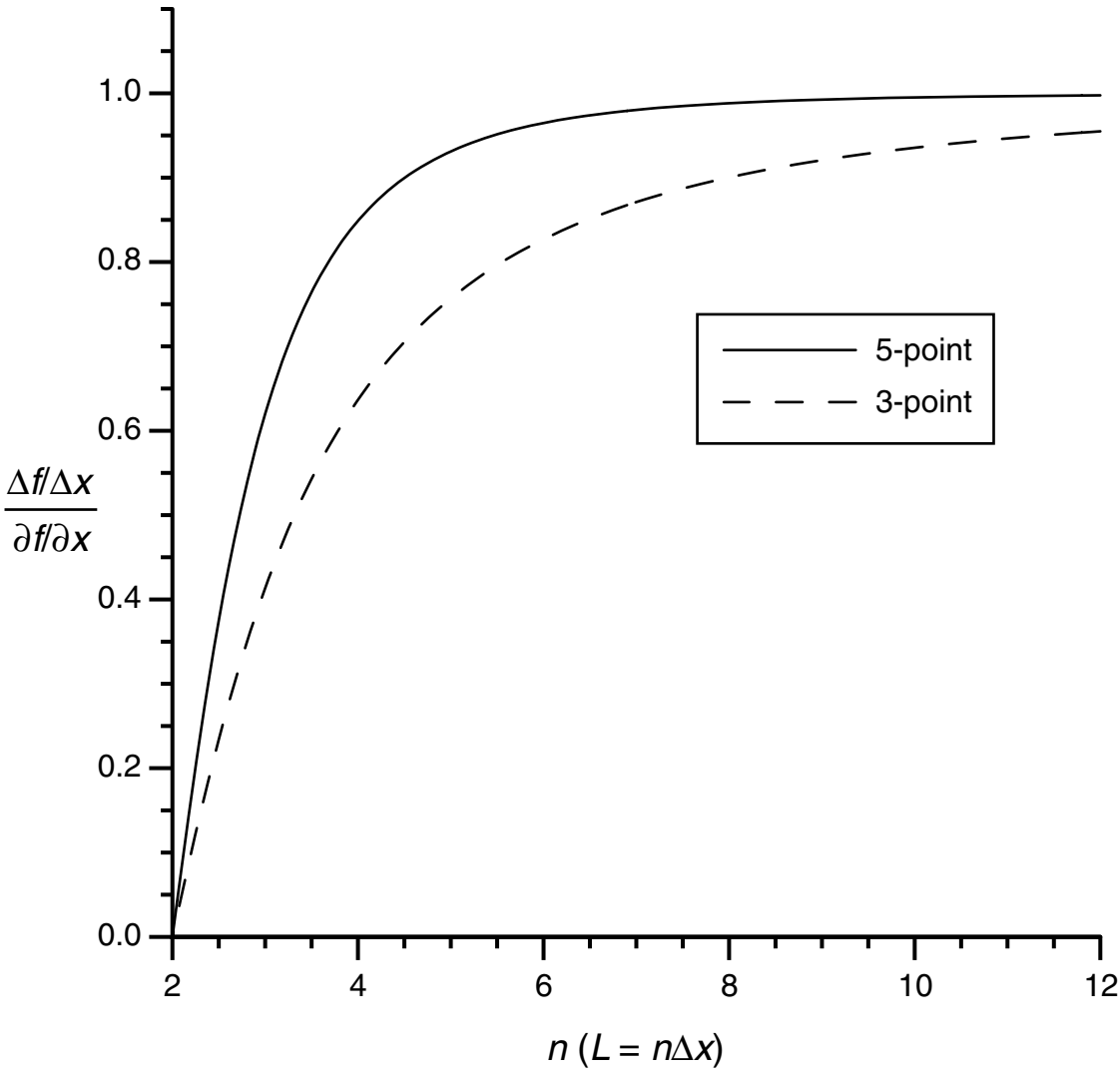

Fig. 3.22 The ratio of the value of the numerical approximation to the derivative of the cosine function and the value of the true derivative, for different numbers of grid increments per wavelength (how well the wave is resolved), for the five-point (fourth-order) and three-point (second-order) approximations.

calculating the derivative of a wave of length $6\Delta x$ with the three-point approximation underestimates the derivative by 17%. For a $10\Delta x$ wave the error is only 6%. This leads to the common qualitative statement that at least ten grid increments are needed in order to properly represent a wave.

Taylor's series can also be used to estimate the truncation error for a five-point approximations to the derivative. For example, add Eqs. 3.26 and 3.27 and solve for the second derivative to obtain

$$\frac{\partial^2 f(a)}{\partial x^2} = \frac{f(a+\Delta x)+f(a-\Delta x)-2f(a)}{(\Delta x)^2} + \cdots .$$

A third derivative can then be defined as

$$\frac{\partial^3 f(a)}{\partial x^3} = \frac{\dfrac{\partial^2 f(a+\Delta x)}{\partial x^2} - \dfrac{\partial^2 f(a-\Delta x)}{\partial x^2}}{2\Delta x}.$$

Using the above expression for the second derivative,

$$\frac{\partial^3 f(a)}{\partial x^3} = \frac{\dfrac{f(a+2\Delta x)+f(a)-2f(a+\Delta x)}{(\Delta x)^2} - \dfrac{f(a)+f(a-2\Delta x)-2f(a-\Delta x)}{(\Delta x)^2}}{2\Delta x},$$

and simplifying yields

$$\frac{\partial^3 f(a)}{\partial x^3} = \frac{f(a+2\Delta x)-f(a-2\Delta x)-2f(a+\Delta x)+2f(a-\Delta x)}{2(\Delta x)^3}.$$

Substituting this expression for the third derivative into Eq. 3.28 and simplifying produces the following expression, which is a five-point approximation to the derivative:

$$\frac{\partial f(a)}{\partial x} = \frac{1}{2\Delta x}\left[\frac{4}{3}(f(a+\Delta x)-f(a-\Delta x)) - \frac{1}{6}(f(a+2\Delta x)-f(a-2\Delta x))\right]. \qquad (3.32)$$

The truncation error is calculated as before (Eqs. 3.29–3.31) and is shown in Fig. 3.22. This is a fourth-order accurate approximation to the derivative because the fifth-order and higher terms in the series were truncated, and has smaller error than the three-point/second-order approximation, as can be seen in the figure. It is interesting that Eq. 3.31 for the truncation error of the second-order approximation, and the analogous one for the fourth-order approximation (not shown), only depend on how well the wave is resolved on the grid (n in $L = n\Delta x$) and not on x itself. However, for the two-point approximation in Eq. 3.25, the truncation error also depends on x (not shown). That is, the ratio $\Delta f/\Delta x \div \partial f/\partial x$ depends on position within the cosine wave. For a wave of length $8\Delta x$, this ratio is unity at $x = L/4$, but becomes very large in magnitude as $x \to 0, \pi$.

In summary, in a grid-point model (i.e., a nonspectral model), every time a derivative is calculated with a finite-difference approximation – which is in virtually every equation, at every grid point, at every time step – the derivative is imperfectly estimated, where the magnitude of the error depends on the sophistication of the differencing scheme and how well the wave is resolved by the grid. Such errors in the pressure-gradient terms in the momentum equations lead to errors in the geostrophic wind, errors in the divergence terms of the continuity equation result in incorrect vertical motions, errors in the gradients in advective terms lead to incorrect advective changes, etc. Also, the pressure gradient in equations expressed in terrain-following, sigma vertical-coordinate systems consists of two derivative terms (see Section 3.4.8), where each term is large and the small difference between them defines the pressure-gradient force. Over large terrain gradients, these terms become especially large, and truncation errors that do not cancel in the two terms create erroneous pressure gradients and accelerations. This problem is partly mitigated by using perturbation forms of the equations, where the derivatives apply to departures from a mean state.

3.4.2 Linear stability, and damping properties

The term *stability* in the context of atmospheric modeling is related to whether the amplitudes of waves in the numerical solution to Eqs. 2.1–2.7, or some other equation set that is the basis for a model, grow exponentially for numerical (i.e., nonphysical) reasons, quickly causing floating-point-overflow conditions that halt the integration of the equations. Many modelers will normally not need to worry about this problem because model codes often contain limits on the time step, and other parameters, that attempt to prevent the instability. Nevertheless, it is useful to know why these constraints exist, and what to expect if the linear-stability criteria are accidentally violated.

Different finite-difference approximations to the time and space derivatives in the equations of motion have different criteria for maintaining stable solutions. Some approximations are absolutely stable – that is, they are never unstable. Some are always unstable – they are called absolutely unstable – and cannot be used. Most are conditionally stable, meaning that stable solutions are obtainable for certain ranges of model parameters and meteorological conditions. Each term in Eqs. 2.1–2.7 contributes to the stability of the numerical solution of its respective equation, but the advection terms are often the most problematic. It is fortunate that the condition for stability of the linear advection equation is about the same as that of the nonlinear advection equation, allowing us to analytically calculate a useful stability criterion with the linear term. Because this kind of instability exists for linear advection, it is called a *linear instability*, contrasting it with the nonlinear instability problem that is described later. This section discusses the linear stability condition for both advection and diffusion terms.

Linear stability of an advection term

The following linear equation will be used as the basis for our analysis of the stability of the advection equation. Assume that h is a meteorological variable such as the height of a pressure surface or the depth of a shallow fluid, and that U is a mean wind speed.

This notation indicates that the terms apply at grid point j on the x axis and at time step τ:

$$\left.\frac{\partial h}{\partial t}\right|_j^{\tau} = -U\left.\frac{\partial h}{\partial x}\right|_j^{\tau}.$$

Assume harmonic solutions to this equation of the following form,

$$h = \hat{h}e^{i(kx-\omega t)}, \tag{3.33}$$

where $\hat{h}$ is the amplitude, $k = 2\pi/L$, L is wavelength, $i = \sqrt{-1}$, and $\omega = Uk$ is wave frequency. Now assume that ω is complex, so that $\omega = \omega_R + i\omega_I$. The implication of this can be seen by substitution into Eq. 3.33, producing

$$h = \hat{h}e^{\omega_I t}e^{i(kx-\omega_R t)}. \tag{3.34}$$

The assumption of a complex frequency has allowed for a wave-amplitude variation with time, such that positive ω_I is associated with exponential wave growth as time (t) increases, negative ω_I is associated with wave damping, and $\omega_I = 0$ means that the amplitude remains constant at $\hat{h}$. The value of ω_I will determine which of these situations prevails, where wave growth is associated with an unstable model solution. The second exponential defines the phase of the wave in the x direction.

For instructional purposes, we will first analyze the stability of the advection equation using forward differencing for the time derivative and backward differencing for the space derivative. The finite-difference expression is

$$\frac{h_j^{\tau+1}-h_j^{\tau}}{\Delta t} = -U\frac{h_j^{\tau}-h_{j-1}^{\tau}}{\Delta x}, \text{ or} \tag{3.35}$$

$$h_j^{\tau+1} - h_j^{\tau} = -\frac{U\Delta t}{\Delta x}(h_j^{\tau} - h_{j-1}^{\tau}), \tag{3.36}$$

where τ is the time-step number and j is the grid-point number. Expressing the assumed form of the solution in Eq. 3.34 in finite-difference form by letting $x = j\Delta x$ and $t = \tau\Delta t$ produces

$$h_j^{\tau} = \hat{h}e^{\omega_I\tau\Delta t}e^{i(kj\Delta x-\omega_R\tau\Delta t)}. \tag{3.37}$$

Substitute this form of the solution into the finite-difference expression Eq. 3.36, producing

$$e^{\omega_I\Delta t}e^{-i\omega_R\Delta t} - 1 = -\frac{U\Delta t}{\Delta x}[1-e^{ik\Delta x}]. \tag{3.38}$$

Using Euler's relations

$$e^{ix} = \cos x + i\sin x \text{ and} \tag{3.39}$$

$$e^{-ix} = \cos x - i\sin x \tag{3.40}$$

in Eq. 3.38 yields

$$e^{\omega_I\Delta t}(\cos\omega_R\Delta t - i\sin\omega_R\Delta t) = 1 + \frac{U\Delta t}{\Delta x}(\cos k\Delta x - 1 + i\sin k\Delta x).$$

In order to obtain information about whether the solution damps or amplifies, the complex equation is separated into its real and imaginary parts:

$$e^{\omega_I \Delta t} \cos \omega_R \Delta t = 1 + \frac{U\Delta t}{\Delta x}(\cos k\Delta x - 1), \tag{3.41}$$

$$e^{\omega_I \Delta t} \sin \omega_R \Delta t = -\frac{U\Delta t}{\Delta x} \sin k\Delta x. \tag{3.42}$$

Squaring both sides of each equation and adding them eliminates the real part of the frequency, leaving

$$e^{\omega_I \Delta t} = \sqrt{1 + 2\left(\frac{U\Delta t}{\Delta x}\right)(\cos k\Delta x - 1)\left(1 - \frac{U\Delta t}{\Delta x}\right)}. \tag{3.43}$$

Recall that Eq. 3.34 shows that the value of this exponential in the assumed form of the model solution controls whether the solution increases or decreases in amplitude with increasing time. That is

$$e^{\omega_I t} = e^{\omega_I \tau \Delta t} = \left(e^{\omega_I \Delta t}\right)^{\tau},$$

so the value of the solution amplifies or damps exponentially as the time-step value τ increases with the model integration.

For this particular combination of space and time differencing schemes, Eq. 3.43 shows the dependence of the exponential on wavelength and the ratio $U\Delta t/\Delta x$. Figure 3.23 indicates that the model solution damps exponentially with time for $U\Delta t/\Delta x < 1$, it does not change amplitude when this ratio is unity, and it amplifies exponentially for ratios greater than unity. Shorter wavelengths are damped more severely than are longer wavelengths. Thus, the stability criterion for this differencing scheme is $U\Delta t/\Delta x \le 1$. The ratio $U\Delta t/\Delta x$ is the previously defined CFL condition. It is also called the *Courant number*, and is described in Courant *et al.* (1928). Thus, given the chosen grid increment, and the largest wave speed that is likely to exist anywhere on the grid during the forecast (U), the time step required for stability is chosen. Note that such selective damping of short waves that are poorly resolved on a grid is sometimes considered to be an advantageous property of a differencing scheme.

A similar procedure can be used to analyze the stability criterion for the forward-in-time, centered-in-space advection equation

$$\frac{h_j^{\tau+1} - h_j^{\tau}}{\Delta t} = -U\frac{(h_{j+1}^{\tau} - h_{j-1}^{\tau})}{2\Delta x},$$

obtaining

$$e^{\omega_I \Delta t} = \sqrt{1 + \left(\frac{U\Delta t}{\Delta x}\right)^2 \sin^2 k\Delta x}.$$

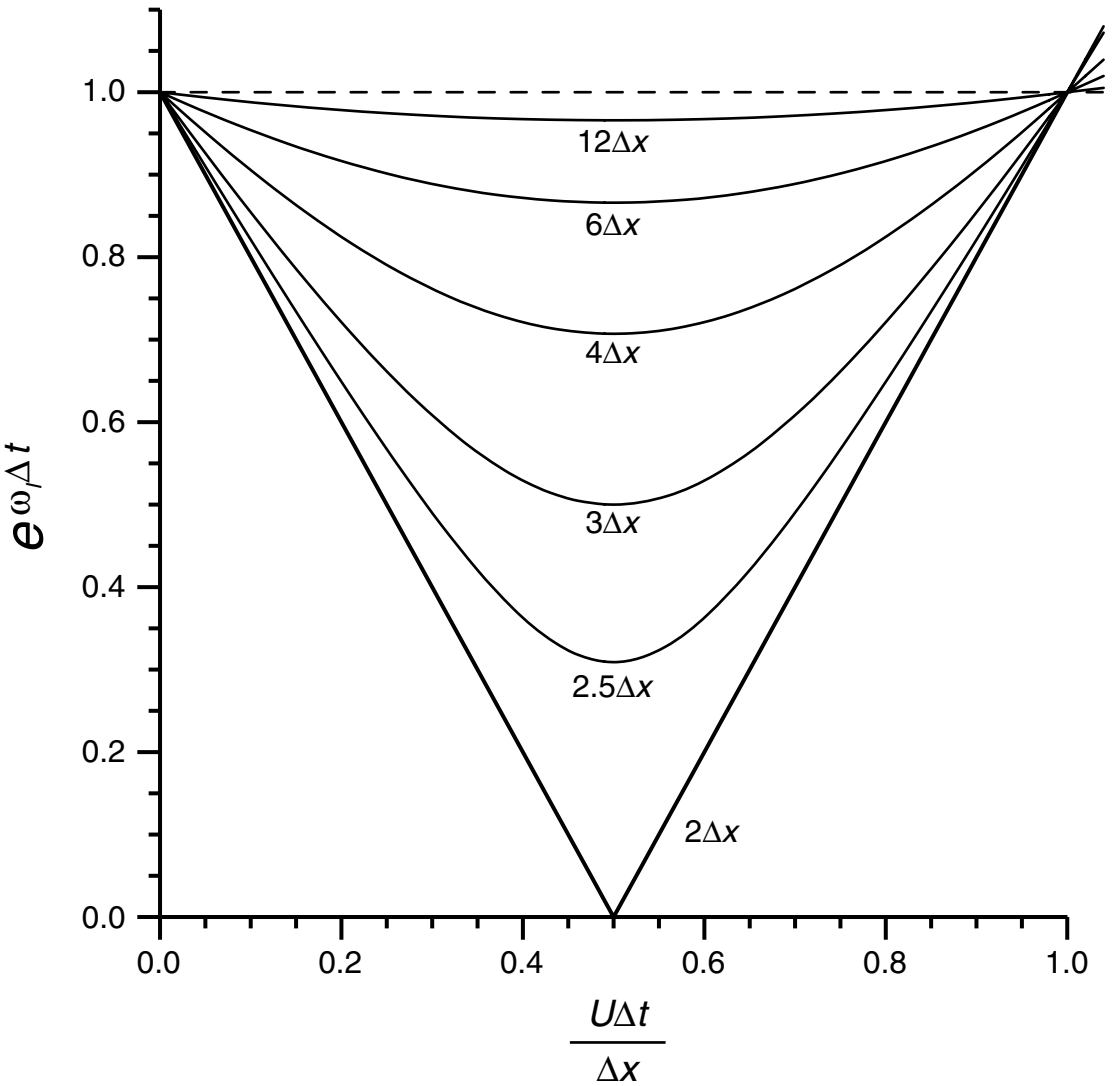

Fig. 3.23 Fractional amplification or damping each time step, of different wavelengths for the forward-in-time, backward-in-space linear advection equation, as a function of the Courant number.

The only time step that will allow the exponential to be less than or equal to unity, and guarantee a stable solution, is zero. Thus, this differencing scheme is absolutely unstable.

Now consider the stability of the three-point centered-in-space, centered-in-time, linear advection equation:

$$\frac{h_j^{\tau+1} - h_j^{\tau-1}}{2\Delta t} = -U\frac{(h_{j+1}^{\tau} - h_{j-1}^{\tau})}{2\Delta x}. \tag{3.44}$$

From the assumed form of the solution in Eq. 3.37, it is easy to obtain

$$h_{j+\beta}^{\tau} = e^{i\beta k\Delta x} h_j^{\tau} \quad \text{and} \tag{3.45}$$

$$h_j^{\tau+\upsilon} = e^{\omega_I \upsilon\Delta t}\, e^{-i\omega_R \upsilon\Delta t} h_j^{\tau} . \tag{3.46}$$

The substitution of Eq. 3.45 into the right side of Eq. 3.44, and then the use of Euler's relations (Eqs. 3.39 and 3.40), provides

$$h_j^{\tau+1} = h_j^{\tau-1} - \frac{U\Delta t}{\Delta x}(2i\sin k\Delta x) h_j^{\tau} .$$

Defining $\alpha = (2U\Delta t/\Delta x)\sin k\Delta x$ produces

$$h_j^{\tau+1} = h_j^{\tau-1} - i\alpha h_j^{\tau} \quad \text{and} \tag{3.47}$$

$$h_j^{\tau+2} = h_j^{\tau} - i\alpha h_j^{\tau+1} . \tag{3.48}$$

Using Eq. 3.47 to eliminate $h_j^{\tau+1}$ in Eq. 3.48 yields

$$h_j^{\tau+2} = (1-\alpha^2) h_j^{\tau} - i\alpha h_j^{\tau-1} . \tag{3.49}$$

The matrix form of Eqs. 3.47 and 3.49 is

$$\begin{bmatrix} h_j^{\tau+1} \\ h_j^{\tau+2} \end{bmatrix} = \begin{bmatrix} 1 & -i\alpha \\ -i\alpha & 1-\alpha^2 \end{bmatrix} \begin{bmatrix} h_j^{\tau-1} \\ h_j^{\tau} \end{bmatrix} . \tag{3.50}$$

The time step for this differencing scheme is $2\Delta t$ (Eq. 3.44), so we can represent the phase and amplitude change in the solution during that period by substituting 2 for υ in Eq. 3.46, which becomes

$$h_j^{\tau+2} = \lambda h_j^{\tau} , \tag{3.51}$$

where $\lambda = e^{2\omega_I \Delta t}\, e^{-i2\omega_R \Delta t}$.

Because λ represents the change between any two time steps, we can also write

$$h_j^{\tau+1} = \lambda h_j^{\tau-1} . \tag{3.52}$$

Substitution of Eq. 3.52 into Eq. 3.47 and Eq. 3.51 into Eq. 3.49 yields

$$0 = (1-\lambda) h^{\tau-1} - i\alpha h_j^{\tau} ,$$

$$0 = (1-\alpha^2-\lambda) h_j^{\tau} - i\alpha h^{\tau-1} .$$

Thus, Eq. 3.50 becomes

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1-\lambda & -i\alpha \\ -i\alpha & 1-\alpha^2-\lambda \end{bmatrix} \begin{bmatrix} h^{\tau-1} \\ h_j^{\tau} \end{bmatrix}.$$

This linear system of equations has a nontrivial ($h \neq 0$) solution if the determinant of the matrix of coefficients equals zero. That is

$$\begin{vmatrix} 1-\lambda & -i\alpha \\ -i\alpha & 1-\alpha^2-\lambda \end{vmatrix} = 0 \text{ , yielding}$$

$$\lambda^2 + (\alpha^2 - 2)\lambda + 1 = 0 \text{ , with the two solutions being}$$

$$\lambda = 1 - \frac{\alpha^2}{2} \pm \frac{\alpha}{2}\sqrt{\alpha^2 - 4}\,. \tag{3.53}$$

Recall that λ represents the amplitude and phase change of the wave over a centered time step, $2\Delta t$. Thus, the magnitude of λ represents the amplification or damping of the wave during the $2\Delta t$ period:

$$|\lambda| = \left| 1 - \frac{\alpha^2}{2} \pm \frac{\alpha}{2}\sqrt{\alpha^2 - 4} \right|.$$

If $\alpha^2 - 4 = 0$, $|\lambda| = 1$ for both solutions. For $\alpha^2 - 4 > 0$, $|\lambda| > 1$ for the solution with the negative sign. For $\alpha^2 - 4 < 0$, λ is complex, and solving for $|\lambda|$ using

$$|\lambda| = \sqrt{|\lambda_R|^2 + |\lambda_I|^2}$$

also yields $|\lambda| = 1$. Thus $|\lambda| = 1$ for $\alpha^2 \leq 4$, where the latter is the requirement for stability. Note that it is sufficient that only one solution for $|\lambda|$ be greater than unity when $\alpha^2 > 4$ in order to make this condition unstable. Substituting the definition for α into the stability requirement leads to

$$\frac{U\Delta t}{\Delta x} \sin k\Delta x \leq 1\,, \tag{3.54}$$

and because the sine can equal unity, the stability requirement is $U\Delta t/\Delta x \leq 1$. Note that if this condition is marginally violated, the wave of length $4\Delta x$ for which the sine function in Eq. 3.54 is unity will be the first to become unstable. Using a similar approach, it can be shown that the stability requirement for the five-point centered-in-space (see Eq. 3.32), and centered-in-time, linear advection equation

$$\frac{h_j^{\tau+1} - h_j^{\tau-1}}{2\Delta t} = -U \frac{\frac{4}{3}(h_{j+1}^{\tau} - h_{j-1}^{\tau}) - \frac{1}{6}(h_{j+2}^{\tau} - h_{j-2}^{\tau})}{2\Delta x} \tag{3.55}$$

is $U\Delta t/\Delta x \leq 0.73$. These centered-in-space and centered-in-time schemes, unlike the forward-in-time and backward-in-space method (Fig. 3.23), do not damp for any stable value of the Courant number.

As noted earlier, the advection term has one of the most restrictive linear-stability criteria of all the terms in the model equations. That is, the time step must be sufficiently small to guarantee that the Courant number ($U\Delta t/\Delta x$) is always less than unity at any time in the integration and at all locations on the grid. The grid increment in this ratio is chosen to be sufficiently small so that processes or meteorological features that are being simulated or forecasted can be defined on the grid with acceptably small truncation error. The velocity scale is a function of the prevailing meteorology, and cannot be controlled. That leaves the time step as the only free parameter that can be adjusted to maintain a stable solution. A useful geometric way of visualizing this stability criterion is that $U\Delta t$, the numerator in the ratio, is the distance traveled by an advecting feature in one time step. If this distance is greater than one grid increment (the denominator), the ratio is greater than unity and the solution is linearly unstable. Unfortunately, further analysis of this term in the context of the full equations, which contain more than advection effects, reveals that the speed in the stability criterion that must be accommodated when choosing a time step is $U + C_P$. This is the advective speed plus the phase speed of the fastest wave on the grid. If the model equations admit sound waves or external-gravity waves, this phase speed can be 300 $\mathrm{m\,s^{-1}}$. So, in choosing a stable time step, this largest wave speed must be estimated, as well as the magnitude of the advective speed in the strongest jet on the grid. Many models will make these estimations internally, and choose a "safe" time step, allowing also for the fact that a horizontally varying map-scale factor will cause the Earth distance between grid points to depart from Δx. Nevertheless, sometimes the estimates will not be sufficiently conservative, and a linear instability will occur in extreme circumstances. Such occasional stability problems may be acceptable given the fact that an extremely conservative small time step would waste computer resources.

The linear stability of the vertical-advection term also has a potentially serious constraint on the time step. This condition parallels that of the horizontal-advection term, and for the three-point approximation is $w\Delta t/\Delta z \leq 1$ for z as the vertical coordinate. Analogous to the horizontal-advection problem, the velocity in this expression is the maximum wave velocity in the vertical, which could be simply the advective velocity, or it could be the sum of the advective velocity and that of vertically propagating gravity waves and sound waves. The vertical grid increment, Δz, typically varies significantly across the depth of the model atmosphere, where smaller values are often used in the boundary layer and near the tropopause in order to resolve large vertical gradients. The vertical advective wave speed is often a maximum near the level of nondivergence, but can be locally large when convective circulations are explicitly represented in the model. Where especially small vertical grid increments and large vertical velocities coexist, the constraint on the time step may be greater than that associated with the horizontal-advection term. For example, recall the discussion in the previous chapter of techniques (Boussinesq, anelastic, hydrostatic approximations) for filtering sound waves from the model solution. The combination of sound waves propagating in the vertical at 300 m $\mathrm{s^{-1}}$, and the fact that vertical grid increments are typically much smaller than

horizontal grid increments for most models, would require an extremely small time step. Thus, there is great motivation to remove the sound waves, or to deal with them numerically in such a way that they do not represent a severe constraint on the time step for the entire equation (for example, the split-explicit time differencing described previously in Section 3.3.1).

Linear stability of an explicit horizontal-diffusion term

Even though much more will be discussed later in this chapter about explicit numerical diffusion terms, in the overall context of the diffusion or damping of model solutions, the linear stability of the following low-order diffusion term will be evaluated here. The strength of the diffusion effect is controlled by the specified magnitude of the positive diffusion coefficient, K:

$$\frac{\partial h}{\partial t} = K\frac{\partial^2 h}{\partial x^2}. \tag{3.56}$$

The term on the right side of this equation is added to the standard physical-process terms in a prognostic equation for the variable h, where the purpose is to damp poorly resolved and sometimes-erroneous small space-scale features in the model solution. If the damping is sufficiently strong, some of the problems described later related to the nonphysical behavior of small-scale energy on the grid can be mitigated. As in the analysis of the advection equation, the amplitude change of the solution depends on the value of the imaginary part of the wave frequency in the assumed form of the wave solution in Eq. 3.34. Approximating this equation with the centered-in-time, centered-in-space approach used for the advection equation,

$$h_j^{\tau+1} - h_j^{\tau-1} = \frac{2K\Delta t}{(\Delta x)^2}(h^{\tau}_{j+1} + h^{\tau}_{j-1} - 2h^{\tau}_{j}),$$

yields

$$e^{\omega_I t} = \frac{2K\Delta t}{(\Delta x)^2}(\cos k\Delta x - 1) \pm \sqrt{4\left(\frac{K\Delta t}{(\Delta x)^2}\right)^2(\cos k\Delta x - 1)^2 + 1}\,, \text{ for } \omega_R = 0\,.$$

Unless $K = 0$, the exponential is more-negative than -1, for the negative sign on the radical, and thus that solution is absolutely unstable, amplifying and changing sign (i.e., phase) each time step. If a forward-in-time, centered-in-space approximation is used,

$$h_j^{\tau+1} - h_j^{\tau} = \frac{K\Delta t}{(\Delta x)^2}(h^{\tau}_{j+1} + h^{\tau}_{j-1} - 2h^{\tau}_{j}), \text{ and}$$

$$e^{\omega_I t} = 1 + \frac{2K\Delta t}{(\Delta x)^2}(\cos k\Delta x - 1), \text{ again for } \omega_R = 0\,. \tag{3.57}$$

For infinitely long waves, the exponential equals unity, so there is no damping or amplification regardless of the value of K or Δt. For waves of length $2\Delta x$,

$$e^{\omega_I t} = 1 - \frac{4K\Delta t}{(\Delta x)^2}. \tag{3.58}$$

For $0 < K\Delta t/\Delta x^2 \le 1/4$, $0 \le e^{\omega_I t} < 1$ and the $2\Delta x$ solution damps;

for $1/4 < K\Delta t/\Delta x^2 \le 1/2$, $-1 \le e^{\omega_I t} < 0$ and the solution damps while changing phase every time step;

and for $1/2 < K\Delta t/\Delta x^2$, $e^{\omega_I t} < -1$ and the solution amplifies while changing phase each time step.

Thus, for physically realistic solutions, the linear stability criterion is $K\Delta t/\Delta x^2 \le 1/4$. Section 3.4.7 discusses other types of horizontal-diffusion terms. Note that models also typically have vertical-diffusion terms with analogous stability criteria. For thin model layers (small Δz) and a large diffusion coefficient, for example in the boundary layer, this criterion may be easily violated.

Maintaining linear stability with multiple terms in an equation

The individual analyses of the linear stability conditions for the finite-difference approximations to the advection and diffusion terms provided quantitative information about the combinations of values of model parameters (e.g., time step) and meteorological conditions required for stable, realistic solutions. For the advection term, the stability constraint was $U\Delta t/\Delta x \le 1$ for centered-in-time, and second-order centered-in-space differencing. For the second-order diffusion term, realistic solutions with forward-in-time, centered-in-space, differencing required that $K\Delta t/\Delta x^2 \le 1/4$. With both of these terms in the same equation, there are questions about how we choose our parameters appropriately to maintain stability, and how we accommodate the fact that one term employs centered-in-time differencing and the other uses forward-in-time differencing. The latter issue can be addressed by evaluating the diffusion term at the $\tau - 1$ time, and extrapolating over a $2\Delta t$ interval to the $\tau + 1$ time. Equation 3.59 shows the prognostic finite-difference equation for the two terms combined.

$$\frac{h_j^{\tau+1} - h_j^{\tau-1}}{2\Delta t} = -u_j^{\tau} \frac{(h_{j+1}^{\tau} - h_{j-1}^{\tau})}{2\Delta x} + \frac{K}{(\Delta x)^2} (h_{j+1}^{\tau-1} + h_{j-1}^{\tau-1} - 2h_j^{\tau-1}). \tag{3.59}$$

The previous separate analysis of the stability of the individual linear versions of the advection and diffusion terms was necessary for mathematical reasons, but the actual operative constraint is based on the combination of all the terms in an equation. Nevertheless, in practice, the constraints associated with the *individual* terms are considered, and the time step from the most restrictive one is employed for the integration. For example, in the case of

Eq. 3.59, let the grid increment be 25 km, and the estimated largest speed on the grid be 50 m s^{-1}. Because unity is the largest stable value of the Courant number, the maximum time step would be 500 s. In practice, the chosen time step used is typically about 20–25% less than this limit to account for the facts that (1) the estimate of the largest stable Courant number is based on a linear analysis, (2) the distance on Earth between grid points is less than Δx in some areas because of the map projection, and (3) the maximum wind speed may be incorrectly estimated. So, the actual time step used in this case would perhaps be 400 s. For the diffusion term, assume that we would like to damp 25% of the amplitude of the $2\Delta x$ wave each time step, which means that $e^{\omega_I t}$ must equal 0.75. Because Eq. 3.58 shows that the damping rate is a function of both the time step and the value of the diffusion coefficient, K, the time step from the advection equation can be used, and the value of K chosen to achieve the desired damping. Thus $K \approx 10^5$ would produce the desired damping.

3.4.3 Phase/group-speed errors

In this section it will be shown how finite-difference approximations can lead to physically unrealistic phase and group speeds. Consider first the forward-in-time and backward-in-space approximation to the advection equation discussed earlier (Eq. 3.35). The speed of an advecting feature should simply be U. To define the advective speed in the numerical solution, divide Eq. 3.42 by Eq. 3.41 to eliminate ω_I, take the inverse tangent of the resulting equation, and use the fact that $\omega_R = C_R k$, obtaining

$$C_R = \frac{1}{k\Delta t}\operatorname{atan}\left[\frac{(U\Delta t/\Delta x)\sin k\Delta x}{1+(U\Delta t/\Delta x)(\cos k\Delta x - 1)}\right].$$

This phase speed of the advective wave on the grid is a function of wavelength (in terms of the wavenumber k) and the Courant number. Figure 3.24 illustrates this relationship. For all wavelengths, Courant numbers of less than 0.5 cause the waves to move more slowly than they should, and Courant numbers of greater than 0.5 cause the waves to move at an erroneously high speed. In general, waves whose phase speed is a function of wavelength are called *dispersive* (e.g., Rossby waves). Even though the advective wave is not dispersive in nature, in the numerical solution it is, so this process is called *numerical dispersion*.

For the three-point, centered-in-space and centered-in-time advection equation, for the situation with a stable solution where $\alpha^2 - 4 \le 0$ in Eq. 3.53,

$$\lambda = e^{2\omega_I \Delta t}\, e^{-i2\omega_R \Delta t} = 1 - \frac{\alpha^2}{2} \pm \frac{i\alpha}{2}\sqrt{4-\alpha^2}.$$

For nondamping, stable solutions, the first exponential is equal to unity. Employing Euler's relation, Eq. 3.40, to rewrite the second exponential, separating the real and imaginary parts of the complex equation, recalling that $\omega_R = C_R k$, substituting the definition of α, and employing trigonometric identities yields

$$C_R = \frac{1}{k\Delta t}\operatorname{asin}\left(\pm\frac{U\Delta t}{\Delta x}\sin k\Delta x\right). \tag{3.60}$$

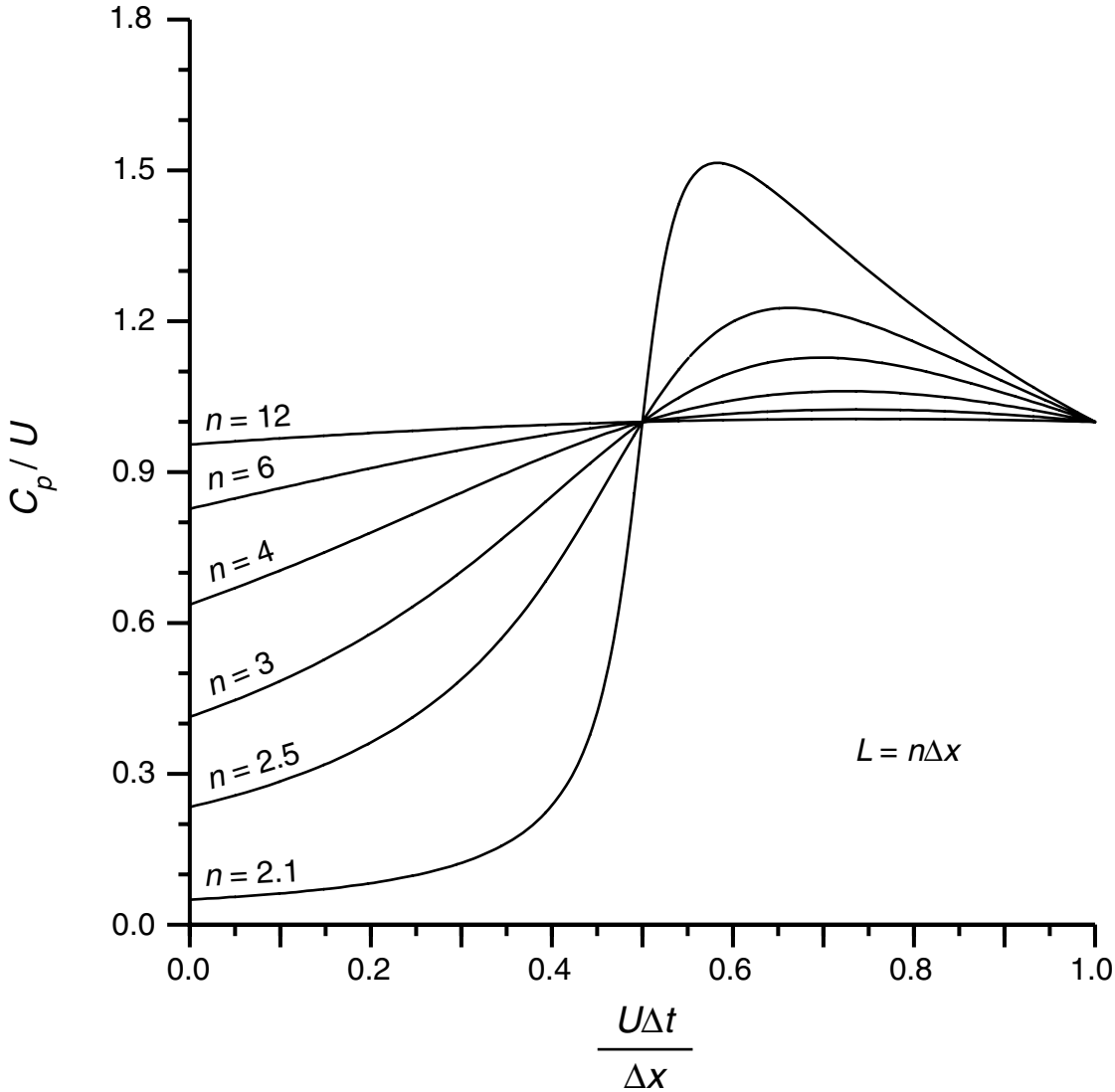

Fig. 3.24 Ratio of the numerical phase speed and the true advective speed for different wavelength features ($n\Delta x$) for the forward-in-time, backward-in-space linear advection equation, as a function of the Courant number.

There are two waves defined here. The one corresponding to the positive argument of the inverse sine is an approximation to the physical wave, moving in the correct direction but slower than the true feature being advected. The other wave, called the *computational mode* or *ghost mode*, is entirely fictitious, with no counterpart in nature, and moves in the opposite direction. The amplitude of the computational mode is typically much smaller than that of the physical mode. The phase speed of the physical mode is shown in Fig. 3.25, as a function of wavelength and the Courant number. This differencing scheme

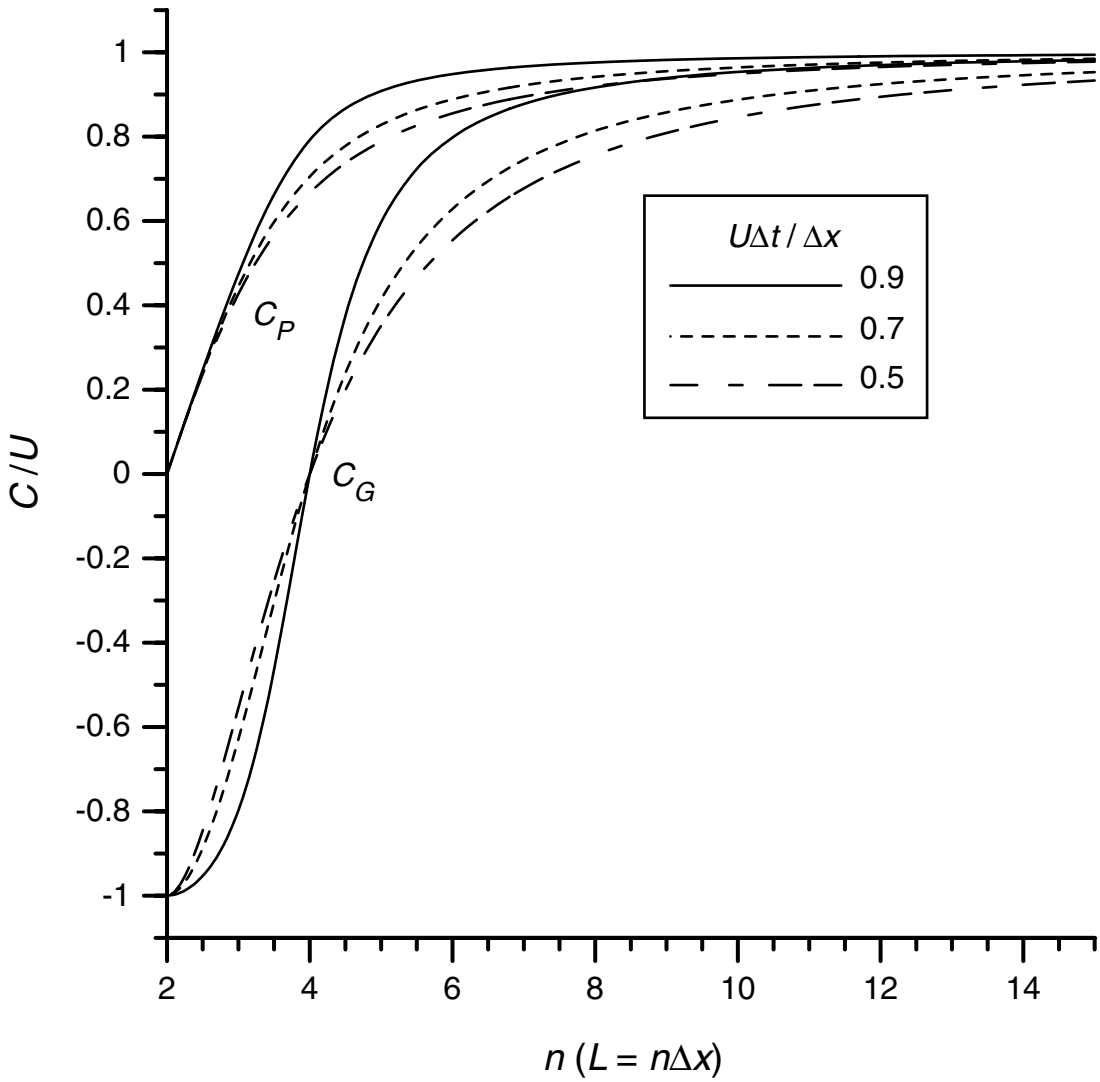

Fig. 3.25 Phase speed (C_P) and group speed (C_G) as a function of wavelengths for different values of the Courant number, for the three-point, centered-in-space and centered-in-time approximation to the linear advection equation. The phase and group speeds pertain to the physical, not the computational, mode.

is also dispersive, with the phase speeds of the longer waves better approximating the true speed, and the speed of the $2\Delta x$ wave being zero. Also, the wave speeds are more realistic for Courant numbers closer to unity.

Examples of model solutions for the three-point (second-order accuracy) centered-in-space and centered-in-time approximation to the linear advection equation (Eq. 3.44) are shown in Fig. 3.26. For the "no-diffusion" curve, the model represents only the linear advection equation defined on a one-dimensional grid that has periodic lateral-boundary

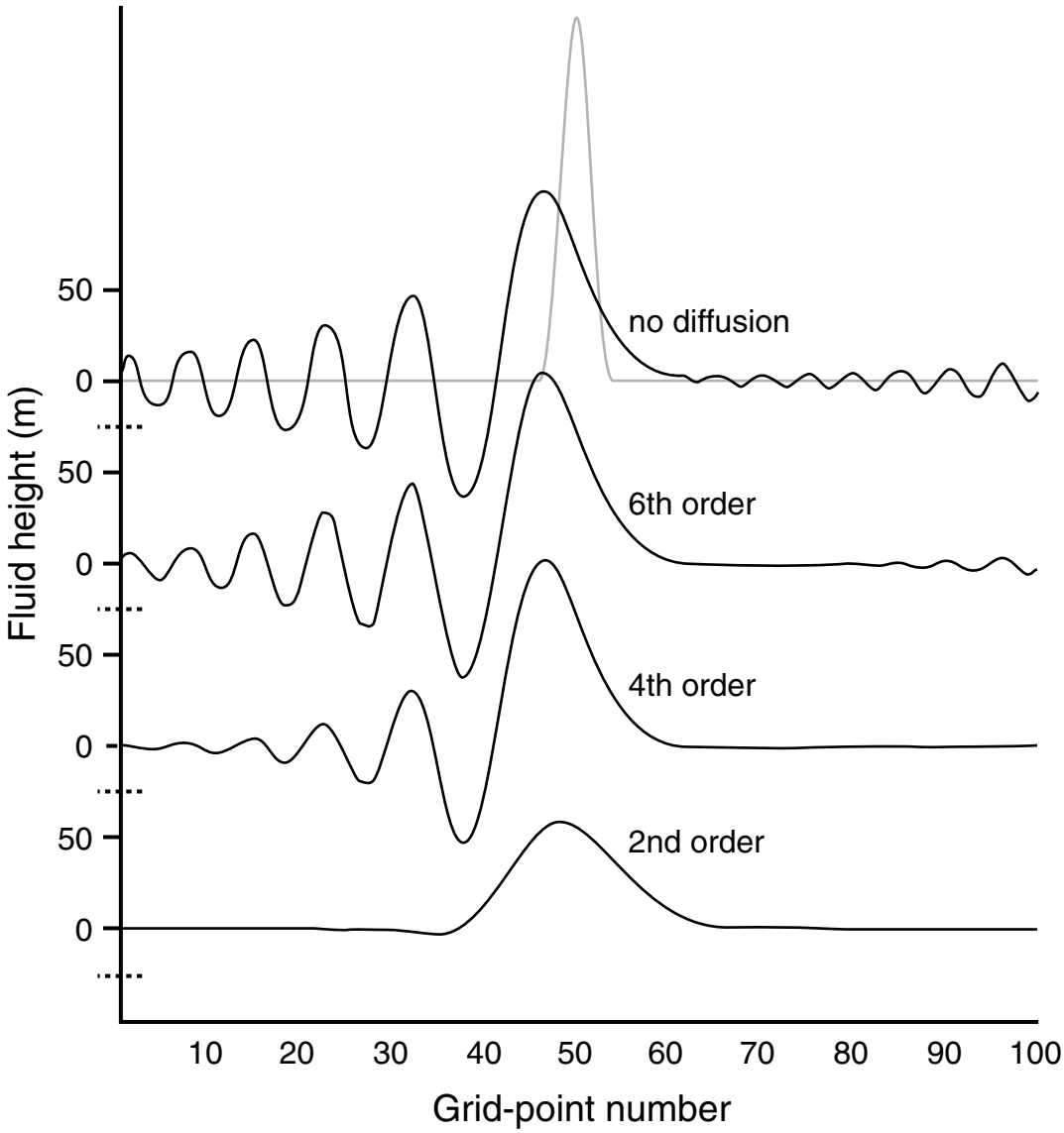

Fig. 3.26 Examples of the use of the three-point, centered-in-space and centered-in-time approximation to the linear advection equation. The initial condition was a symmetric wave in the height field, centered on grid-point 50, and is shown by the light line at the top. The exact solution is also shown by the light line, after the wave has advected over about 100 grid increments from left to right across the grid, exiting at the right boundary and entering at the left. Numerical solutions are shown with heavier lines, for no diffusion and diffusion operators of different order. The Courant number was 0.1.

conditions. That is, a wave exiting one end of the computational domain enters on the opposite end. The initial conditions were defined by a wave in the height field centered on grid-point 50 and with the shape of the lighter curve in the figure. The advective velocity

was from left to right in the figure, with a magnitude of 10 m s^{-1}. The lighter line in the figure also shows the theoretical solution for the wave after it has traversed approximately 100 grid points (100 km), exiting on the right and entering on the left. That is, the wave moved at exactly the advective speed, U, and there was no damping or amplification. In contrast, the black lines (focus on the upper one in this discussion) show the numerical solution at the same time for a Courant number of 0.1 ($\Delta t = 10$ s). The original wave, even though it appears smooth, was composed of many Fourier components with different wavelengths, which, as we have seen, have different numerical phase speeds. The longer wavelengths have speeds closer to the correct value (see Fig. 3.25), but the shorter components move at speeds that are proportional to their wavelength. The very short waves have not even exited the grid on the right at this time, having moved at less than half the correct speed. Some of the erroneous wave energy might be associated with the previously mentioned computational mode, but it is difficult to visually separate it from the poorly represented short waves in the physical mode. Clearly, this is not an especially satisfactory solution for representing the advection process. In particular, when model dependent variables change rapidly over a small distance, such as across fronts, the sharp gradient is defined by short-wavelength Fourier components. Thus, even when such physical features are realistically rendered in model initial conditions, as the short wavelengths become separated from the longer wavelengths as the feature propagates, the gradient will weaken.

To illustrate how the choice of the Courant number can affect numerical dispersion, Fig. 3.27 depicts model solutions analogous to the upper one (no diffusion) in Fig. 3.26, but for the additional Courant numbers of 0.5 ($\Delta t = 50$ s) and 0.9 ($\Delta t = 90$ s). The differences can be explained by referring to Fig. 3.25, which shows that the use of Courant numbers closer to unity produces more-correct phase speeds for the $3–10\Delta x$ wavelengths, and thus there is less energy in the erroneously slow waves. The influence of the Courant number is dependent on the specific approximation to the advection term, but this shows how great the impact can be.

The analogous model solution using the higher-accuracy, fourth-order (five-point) approximation for the derivative in the linear advection equation (Eq. 3.55) is shown in Fig. 3.28 for a Courant number of 0.1. This also shows erroneous numerical dispersion, like the three-point scheme, but it is less severe. Nevertheless, significant amplitude has been lost relative to the correct solution, and there is still wave energy in erroneous features that trail the more-correctly rendered longer wave.

In nature, the advective wave is nondispersive, and the phase speed and the group speed are equal. Given that these centered-in-space and centered-in-time solutions exhibit numerical dispersion, it is revealing to calculate the group speed (C_G) of the waves, or in other words the speed at which the wave energy propagates. In general,

$$C_G = \frac{\partial}{\partial k}(C_P k)\,. \tag{3.61}$$

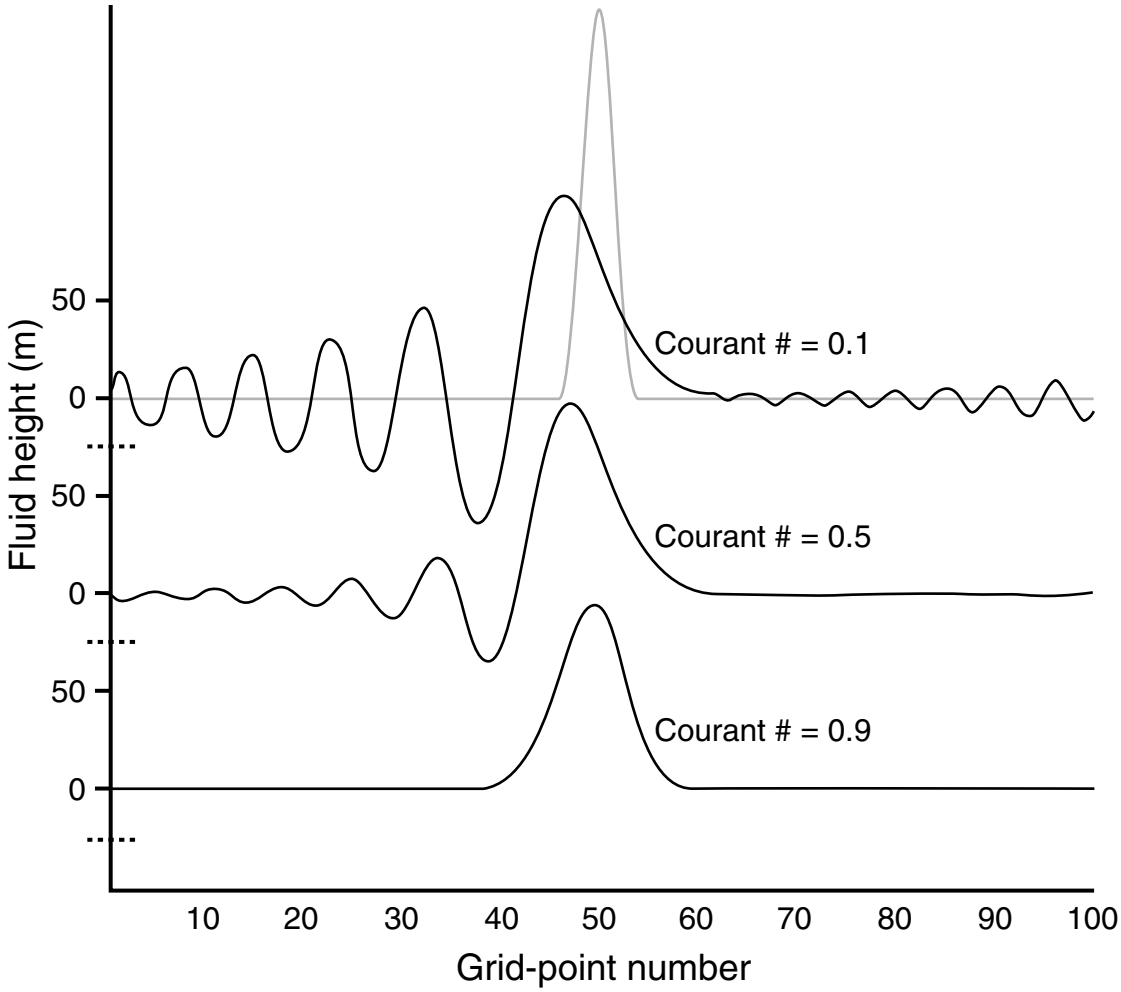

Fig. 3.27 Model solutions for the linear advective wave, without diffusion, for the different indicated values of the Courant number. Second-order, centered-in-space and centered-in-time differencing was used. The exact solution is shown by the light line at the top.

For the three-point, centered-in-space and centered-in-time approximation to the advection equation, Eq. 3.60 represents the phase speed, C_P. Substituting this expression into Eq. 3.61 and evaluating the derivative gives

$$C_G = \frac{U\cos k\Delta x}{\left[1-\left(\frac{U\Delta t}{\Delta x}\sin k\Delta x\right)^2\right]^{1/2}}.$$

This group speed is plotted in Fig. 3.25 as a function of wavelength and Courant number. For a wave of length $4\Delta x$, the group speed is zero. The $2\Delta x$ wave energy travels at the correct speed, but in the wrong direction. Thus, the energy propagation properties of the shorter waves are severely mishandled by this finite-difference approximation.

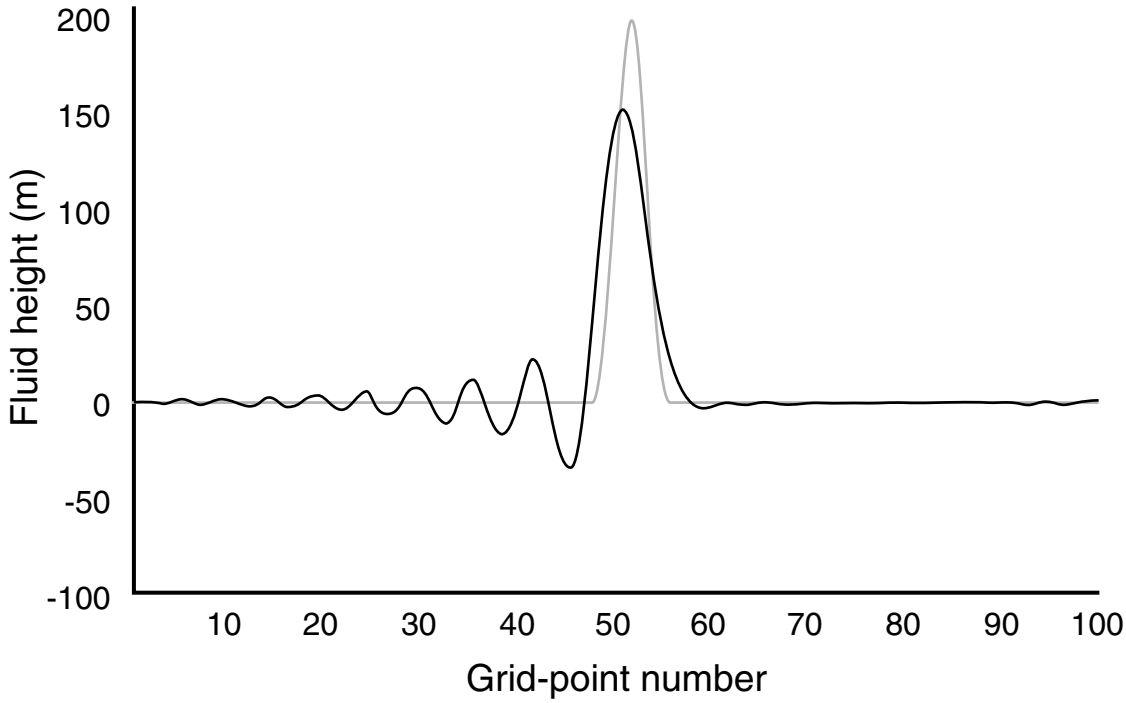

Fig. 3.28 Analogous to the no-diffusion curve in Fig. 3.26, except that the model used the fourth-order-accurate, five-point, approximation to the spatial derivative in the advection equation. No explicit diffusion was employed, and the Courant number was 0.1. The exact solution is shown by the light line at the top.

3.4.4 Properties of some example, multi-step, time-differencing schemes

Section 3.3.1 defines only a few of the many multi-step time-differencing schemes that have been used in atmospheric models. Even though Durran (1999) provides a thorough discussion of their numerical properties, a few schemes will be mentioned here. In general, these methods are popular because their stability criteria are often not as stringent as for the single-step methods that are the focus above, they can have relatively high orders of accuracy, and some very selectively damp the smaller, poorly resolved wavelengths. Figure 3.29a shows the fractional damping each time step associated with the Lax–Wendroff and Euler-backward time-differencing schemes applied to the linear-advection equation. In each case, the second-order, centered-in-space approximation is used for the space differencing (the variable F in Eqs. 3.14–3.20). Because of the desirability of selectively damping short-wavelength, poorly resolved waves, while leaving the well-resolved waves relatively undamped, the Lax–Wendroff scheme is superior in this respect. The numerical dispersion caused by these schemes is shown in Fig. 3.29b, where the Euler-backward method produces more-correct phase speeds for the better-resolved waves. An example of an advective-wave solution using a multi-step time-differencing scheme with high-order space differencing is seen in Fig. 3.30. The third-order Runge–Kutta time-differencing scheme,

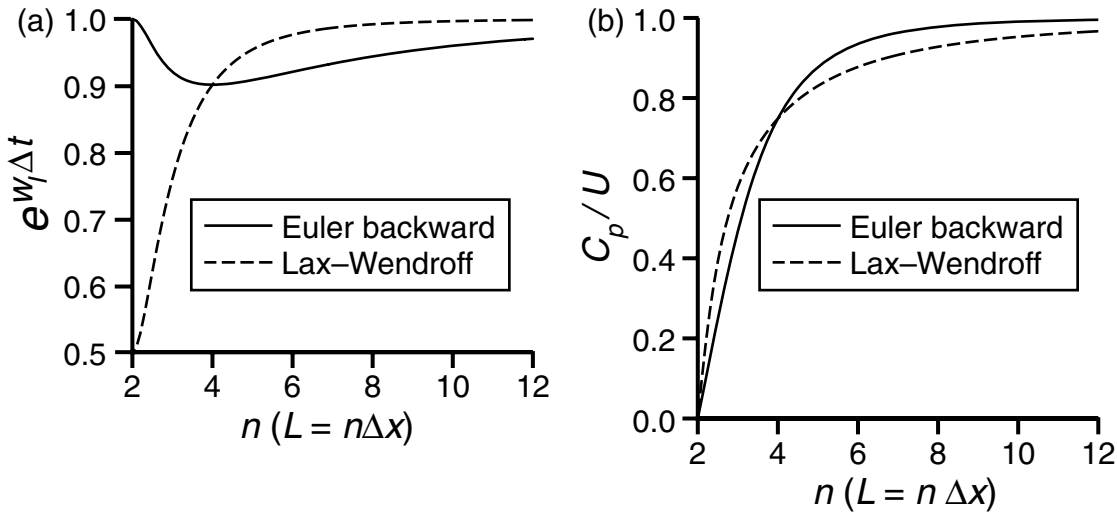

Fig. 3.29 Fractional damping each time step (a) and the ratio of the numerical phase speed and the correct phase speed (b), as a function of wavelength for the Lax–Wendroff and Euler-backward time-differencing schemes, with second-order, centered space differencing.

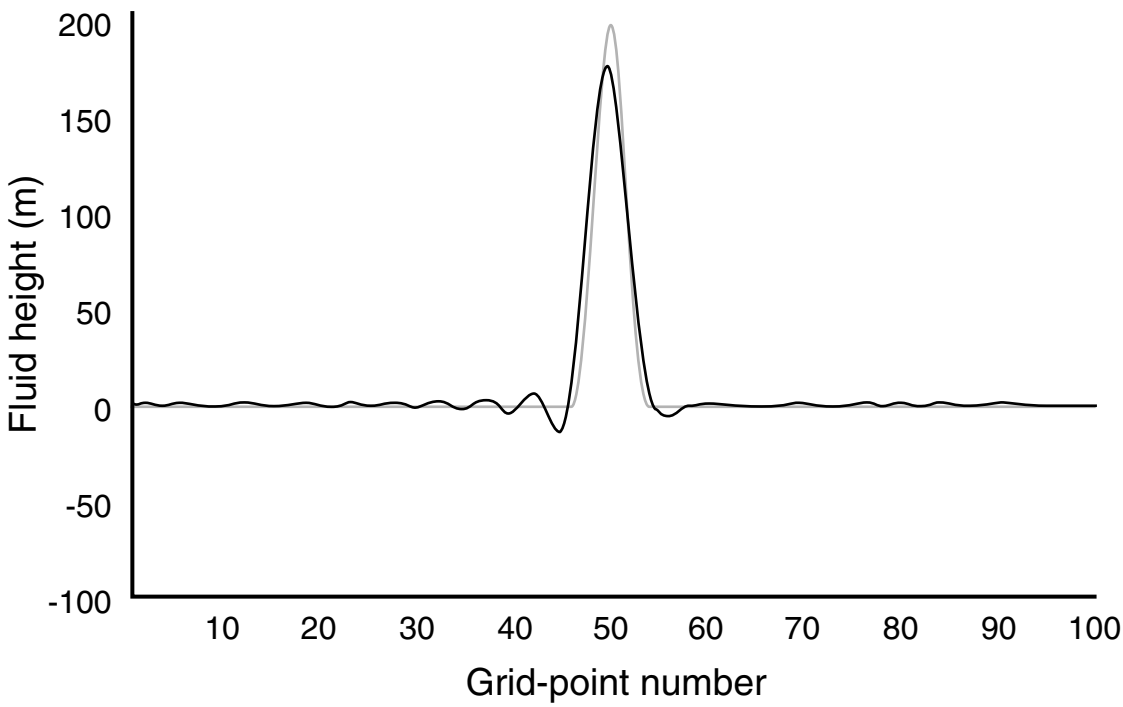

Fig. 3.30 Analogous to the no-diffusion curve in Fig. 3.26, except that third-order Runge–Kutta time differencing and sixth-order space differencing were used. There was no explicit diffusion employed. The exact solution is shown by the light line at the top.

shown in Eqs. 3.18–3.20 and employed in the dynamic core of the community Advanced Research WRF (ARW) model (Skamarock *et al.* 2008), is combined with sixth-order space differencing. Compared to the wave solution shown in Fig. 3.28 resulting from single-step, centered-in-time differencing and fourth-order centered-in-space differencing, there is less numerical dispersion and the amplitude of the primary wave is better preserved. The ARW offers second- through sixth-order options for the space differencing, where the default is the fifth-order option because the odd-order schemes have desirable implicit damping properties, whereas the even-order schemes do not. But, regardless of the approach for the space-differencing, the Runge–Kutta time differencing contributes some damping of its own.

3.4.5 Aliasing

Aliasing is a process by which two waves represented on a model grid interact through a nonlinear term in the equations to produce fictitious waves, resulting in an erroneous redistribution of energy (amplitude) in the wave spectrum, and possibly even leading to an instability that is fatal to the model integration. This process can be illustrated with a simple, nonlinear advection term:

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x}.$$

For simplicity, assume that u can be represented mathematically as a sum of cosine waves, such as

$$u = \sum_{m=0}^{\infty} a_m \cos k_m x, \text{ where } k_m = \frac{2\pi}{L} m,$$

L is the length of the computational grid, and k_m is a wavenumber.[4] Differentiating this expression with respect to x, as in the advection term, we obtain

$$\frac{\partial u}{\partial x} = -\sum_{m=0}^{\infty} a_m k_m \sin k_m x,$$

and multiplying by $-u$ gives

$$-u\frac{\partial u}{\partial x} = -(a_0 + a_1 \cos k_1 x + a_2 \cos k_2 x + \ldots + a_m \cos k_m x + \ldots) \times$$
$$(a_1 k_1 \sin k_1 x + a_2 k_2 \sin k_2 x + \ldots + a_n k_n \sin k_n x + \ldots).$$

The result of any two waves, k_m and k_n, interacting is

$$a_n a_m k_n \sin k_n x \cos k_m x.$$

[4] Note that both k_m and m are wavenumbers. The m represents the number of waves on the domain of length L, and is nondimensional. The k_m is 2π divided by the wavelength (L/m), has dimensions of inverse distance, and is sometimes referred to as a rotational wavenumber.

But

$$\sin x \cos y = \frac{\sin(x+y) + \sin(x-y)}{2},$$

and therefore the interaction product is

$$a_n a_m k_n [\sin(k_n + k_m)x + \sin(k_n - k_m)x] =$$

$$a_n a_m k_n \left[\sin\frac{2\pi}{L}(n+m)x + \sin\frac{2\pi}{L}(n-m)x\right].$$

Thus, when wavenumbers m and n interact, they produce two waves, one with wavenumber $n+m$ and one with wavenumber $n-m$. This is no problem in continuous space where all wavenumbers are possible, but it can be in the discrete (grid-point) space of a model. For example, assume a one-dimensional grid having j_{max} intervals (see Fig. 3.31), where j_{max} is an even number. Table 3.1 shows the range of wavenumbers and wavelengths on this grid, where the longest complete wave that can be represented is defined by the domain length, L, and the shortest wave is defined in terms of the grid increment, Δx.

Now consider what happens when resolvable wavenumbers m and n interact to yield a wavenumber that is larger than what is permitted by the grid (i.e., a wavelength of less than $2\Delta x$). This would result from the $m+n$ interaction product rather than the $m-n$ product, so $m+n > j_{max}/2$. A way of defining $m+n$ without the inequality is $m+n = j_{max} - s$, where $s < j_{max}/2$. So the wave resulting from the problematic $m+n$ interaction would be

$$\sin\frac{2\pi}{L}(m+n)x = \sin\frac{2\pi}{j_{max}\Delta x} \cdot (j_{max} - s) \cdot (j\Delta x)$$

$$= \sin 2\pi \frac{j_{max} - s}{j_{max}} j = \sin\left(2\pi j - \frac{2\pi s j}{j_{max}}\right).$$

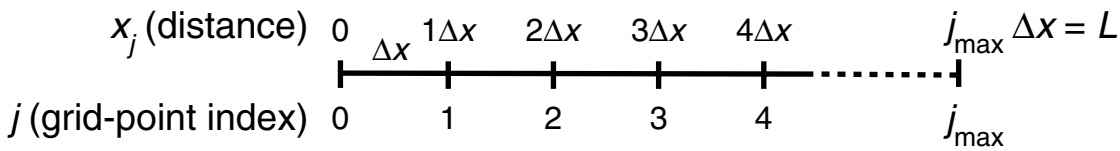

Fig. 3.31 A one-dimensional grid with j_{max} grid intervals and a length of L.

Table 3.1 Corresponding wavenumbers and wavelengths on a grid with j_{max} points

Wavenumber	Wavelength
1	$j_{max}\,\Delta x$ (longest)
2	$j_{max}\,\Delta x/2$
.	.
.	.
.	.
$j_{max}/4$	$4\,\Delta x$
$j_{max}/2$	$2\,\Delta x$ (shortest)

But $\sin(x-y) = \sin x \cos y - \cos x \sin y$, so that

$$\sin\left(2\pi j - \frac{2\pi sj}{j_{max}}\right) = \sin 2\pi j \cos\frac{2\pi sj}{j_{max}} - \cos 2\pi j \sin\frac{2\pi sj}{j_{max}},$$

where the sine function in the first term on the right side is equal to zero, and the cosine function in the second term is equal to unity. Thus,

$$\sin\frac{2\pi}{L}(m+n)x = \sin\frac{2\pi sj}{j_{max}} = \sin\frac{2\pi s}{L}x.$$

So the unresolvable wavenumber shows up on the grid as wavenumber s, such that $s = j_{max} - (m+n)$. For example, say $m = \frac{1}{2}j_{max}$ (a $2\Delta x$ wave) and $n = \frac{1}{4}j_{max}$ (a $4\Delta x$ wave), so $m+n = \frac{3}{4}j_{max}$ (a $\frac{4}{3}\Delta x$ wave) and $m-n = \frac{1}{4}j_{max}$ (a $4\Delta x$ wave). But the $\frac{4}{3}\Delta x$ wave is unresolvable, and the aliasing produces energy in the $4\Delta x$ wavelength ($s = j_{max} - \frac{3}{4}j_{max} = \frac{1}{4}j_{max}$).

To illustrate all possible interactions, assume $j_{max} = 24$. Any interaction that produces a wavenumber greater than 12 will result in aliasing. The erroneous redistribution of energy on this grid is illustrated in Fig. 3.32.

Not only does this aliasing process cause energy to be incorrectly located in the wrong scales, resulting in errors in the model solution, it can also result in the model solution becoming unstable and stopping the numerical integration process. This is called a

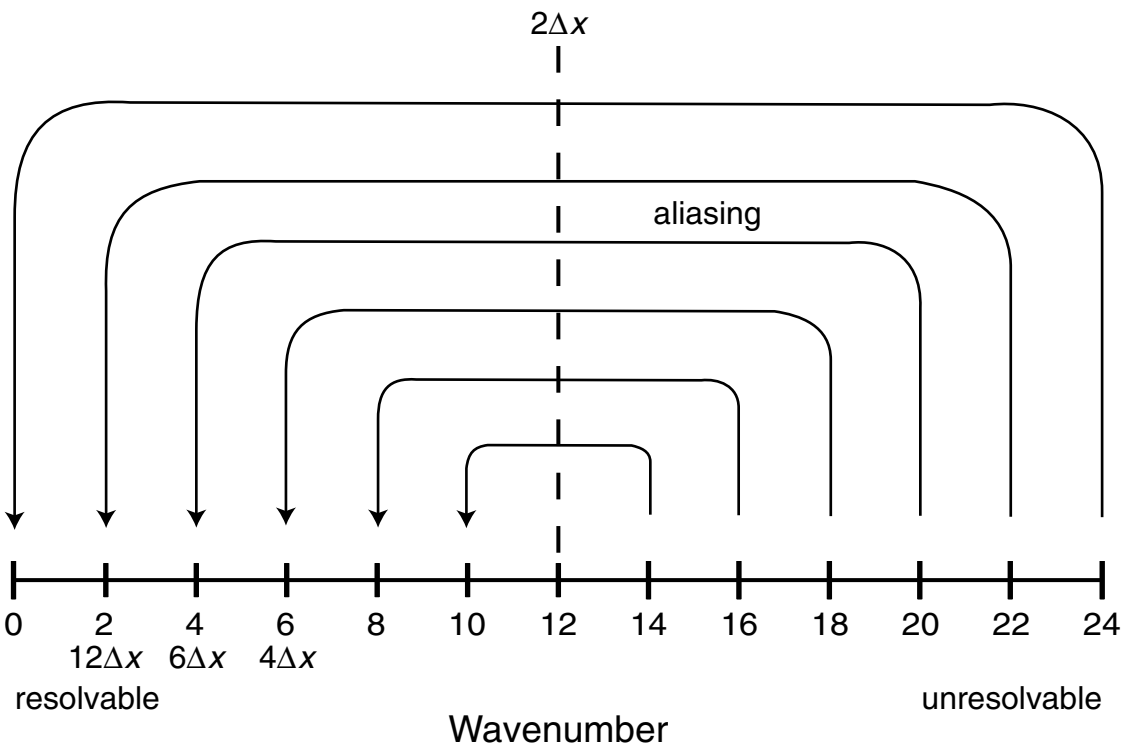

Fig. 3.32 Illustration of how nonlinear interactions on a 24-point grid produce aliasing, resulting in energy being erroneously placed in the wrong wavelengths. Interactions that produce unresolvable wavelengths on the right result in the energy being "folded" over the smallest resolvable wavenumber 12 ($2\Delta x$) to the resolvable side of the scale.

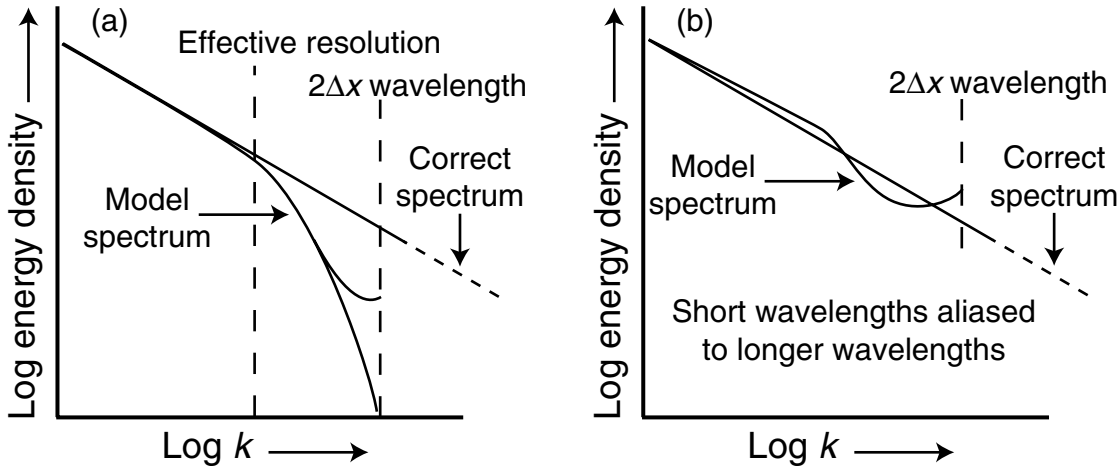

Fig. 3.33 Schematic of the correct (straight line) and model kinetic-energy spectra. In (a), two examples are shown of the normal damping of the kinetic energy at the high-wavenumber end of the model spectrum by the model diffusion (see Section 3.4.7). In this situation, any energy aliased to the high wavenumbers has been controlled by the diffusion. In (b) is shown a model solution where aliasing has added erroneous energy that remains in the high wavenumbers. Adapted from Skamarock (2004).

nonlinear instability and will be discussed in the following section. Aliasing is also mentioned in the discussion of spectral modeling, wherein each wave interaction is analytically treated, and those interactions that would result in unresolvable products are not permitted. This is one of the advantages of the spectral-modeling approach.

Figure 3.33 shows the impact of aliasing on a model's kinetic-energy spectrum. The plot on the left shows a normal spectrum, where aliasing is not a significant problem. The filtering of the high wavenumber (short wavelength) part of the model solution, through processes described in Section 3.4.7 below, causes a loss of kinetic energy in the segment of the spectrum between the $2\Delta x$ wavelength and the effective resolution of the model. This damping is desirable because these wavelengths are poorly represented on the grid. The right panel shows the spectrum when aliasing has added erroneous energy in the resolved part of the spectrum, overwhelming the desirable damping illustrated in the left panel, and impacting the model's representation of physical processes.

3.4.6 Nonlinear instability

The discussion of computational instability in Section 3.4.2 was based on a linear differential equation, and adhering to the appropriate stability criterion is sufficient to avoid problems of that type. For nonlinear equations, there is a similar criterion, but even when this condition is satisfied it is possible for another type of instability to develop in the numerical solution. As shown in Chapter 2, primitive-equation models are based on nonlinear equations, and the source of the problem is the aliasing that was just described. The symptom of *nonlinear instability* that results from aliasing is a rapid buildup of energy in

the 2–4Δx wavelengths in the model solution, after the model has been integrated for a long period of time. The cause of this can be inferred from the list of wave interactions that are associated with aliasing on the 24-point grid shown in Fig. 3.32 (see Problem 1 at the end of this chapter). In particular, there are 42 combinations of wavenumbers m and n that result in aliasing, and 30 of these interactions produce energy in the 2–4Δx range. Couple this with the fact that every aliasing interaction involves at least one wave in this range, and it is clear that such an uncontrolled accumulation of energy can lead to numerical problems. The energy accumulation in these short wavelengths can be controlled by using a differencing scheme that selectively damps the short waves, or scale-selective diffusion (dissipation) terms can be added to the equations (see next section). Alternatively, spectral or semi-Lagrangian methods can be used for the nonlinear terms so that the interactions are treated analytically. Continued integration of a model that has a kinetic-energy spectrum like that in Fig. 3.33b could lead to nonlinear instability.

3.4.7 Diffusion: real, explicit numerical, implicit numerical, grid

The diffusion processes described here all have the effect of spatially spreading features in the heat, moisture, and momentum fields of the modeled fluid. This can have the effect of damping the amplitude of perturbations in the variables, so sometimes they are referred to as damping processes. Because the diffusion or damping is scale selective, the methods may also be considered filters. There is obviously real (physical) diffusion, or mixing, in the atmosphere, caused by turbulence, and this needs to be represented in some realistic way. In addition, however, a nonphysical, scale-selective, diffusion or damping process is incorporated in all models, through explicit terms in the predictive equations or implicitly through the use of damping differencing schemes. The purpose is to "clean up" unrealistic features in the model solution associated with lateral-boundary noise, computational modes, and erroneous shortwave energy from numerical wave dispersion. Lastly, even if none of the above diffusion processes were incorporated in the model, there would still be the spatial spread of information about the model variables through the vertical and horizontal finite differencing. The terminology used in the literature for referring to the different types of diffusion is not standard, so the reader should be cautious.

Physical diffusion

The atmosphere contains turbulence that smooths out, or diffuses, structures in the momentum, thermal, and moisture fields, in all three coordinate directions. Where gradients exist, turbulent fluxes transport properties, such that the amplitudes of maxima or minima in physical fields are reduced. Because of the intensity of shear- and buoyancy-driven turbulence in the boundary layer, and the typically strong gradients there, planetary boundary-layer parameterizations are needed to represent this important physical process. Likewise, turbulent mixing can be important elsewhere, in the free atmosphere, such as near jets in the wind field and in the vicinity of moist convection, and models need to be able to treat the associated mixing in a realistic way. This is the "real" diffusion, or mixing, that must be

represented in a physically faithful model of the atmosphere. Chapter 4 discusses the representation of this *physical diffusion* by turbulence parameterizations.

Explicit numerical diffusion

In the previous section on aliasing, it was mentioned that one way of controlling the artificial accumulation of energy in short wavelengths, and the resulting nonlinear instability, is through the use of diffusion terms, on the right side of the predictive equations, that are explicitly designed to damp these short wavelengths. In addition to controlling this instability, damping the short wavelengths also improves model solutions like those shown in Figs. 3.27 and 3.28, where numerical dispersion causes short wavelengths to be erroneously separated from the physical solution. A challenge is sufficiently damping the erroneous component of the model solution while not damping the physically realistic part.

There are a few different mathematical forms, shown in Eq. 3.62, that are used for the term that explicitly controls shortwave amplitudes:

$$\frac{\partial h}{\partial t} = (-1)^{n/2+1} K_n \nabla^n h. \tag{3.62}$$

Here, K is the *diffusion coefficient*, h is any dependent variable, and $n = 0, 2, 4, 6$ indicates the order of the term. The right side for zero-order ($n = 0$) damping is $-K_0 h$. This produces a non-scale-selective relaxation, and is typically applied near lateral and upper boundaries. The second-order term has the form of the Laplacian, and is the equivalent of Eq. 3.56 except that it has two horizontal space dimensions. Recall that a term with this form appears in the physical heat-diffusion equation, where higher values are always transferred down gradient toward regions with smaller values. The fact that the change in the property depends solely on the sign and magnitude of the curvature (the second derivative) means that new extrema are not added to a field. This second-order term is less scale-selective than the higher-order ones. Equation 3.57 represents the amount of damping per time step for the forward-in-time, centered-in-space finite-difference scheme for the one-dimensional problem. Equation 3.63 shows analogous equations for the fourth- and sixth-order diffusion as well. The upper equation is the same as Eq. 3.57, and represents the damping per time step for second-order diffusion, and the middle and lower equations apply to fourth- and sixth-order diffusion, respectively:

$$e^{\omega_I t} = 1 - K\Delta t \begin{bmatrix} (2 - 2\cos k\Delta x)/(\Delta x)^2 \\ (6 - 8\cos k\Delta x + 2\cos 2k\Delta x)/(\Delta x)^4 \\ (20 - 30\cos k\Delta x + 12\cos 2k\Delta x - 2\cos 3k\Delta x)/(\Delta x)^6 \end{bmatrix}. \tag{3.63}$$

The amount of damping at different wavelengths for these three diffusion-operator options is important because it is necessary to filter the poorly resolved, small scales, especially in the 2–$4\Delta x$ range, without greatly damaging the amplitudes of the better-resolved length scales. Figure 3.34 shows the amount of damping per time step for the second-, fourth-, and sixth-order terms as a function of wavelength. For each curve, the values of K and Δt have been chosen so that the $2\Delta x$ wave is completely removed, each time step. In

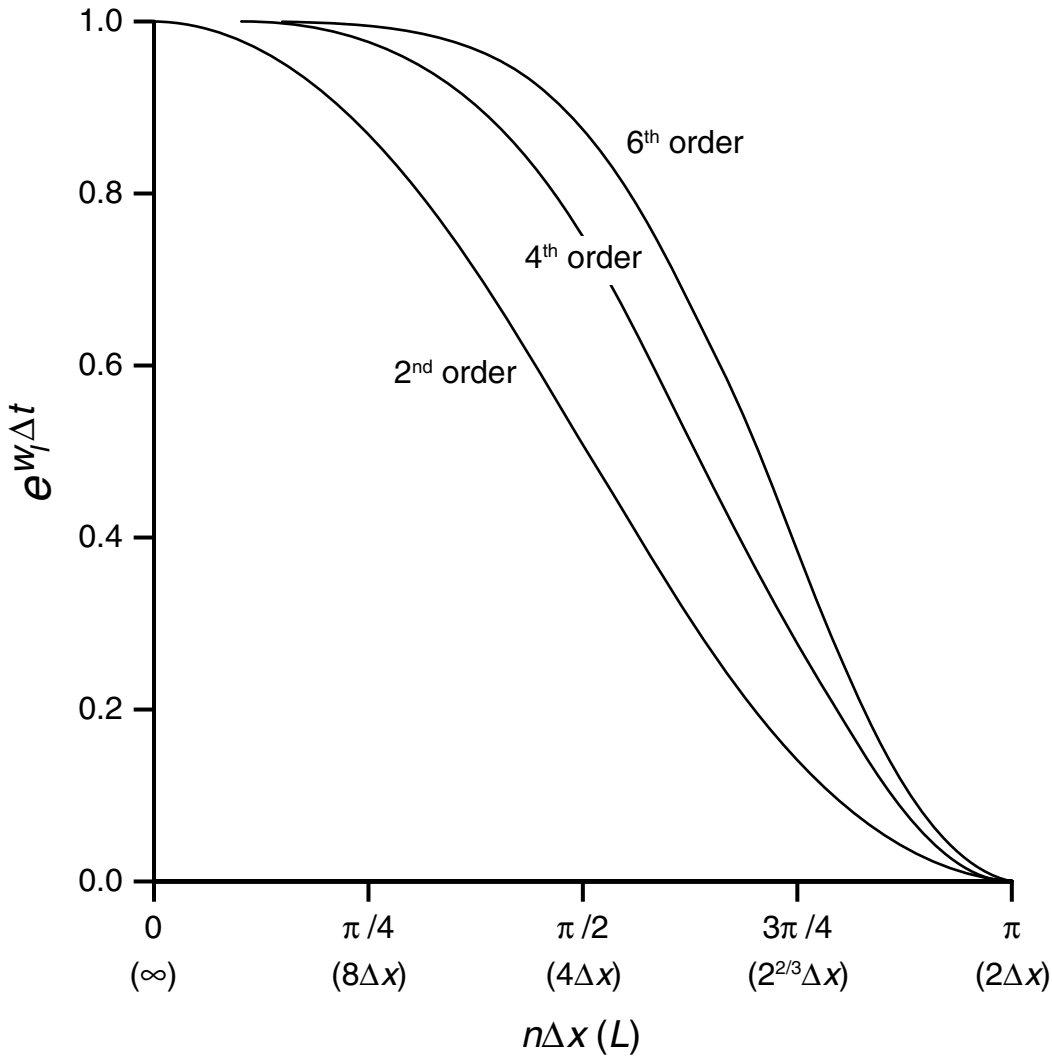

Fig. 3.34 Damping rate per time step (Eq. 3.63) for second-, fourth-, and sixth-order diffusion, for $K\Delta t$ chosen such that there is 100% damping of the $2\Delta x$ wave each time step. The abscissa is wavelength ($n\Delta x$, bottom) and wavenumber (top).

practice, it would be unusual to use such a large diffusion coefficient, but such an assumption allows us to normalize the curves to reveal relative damping rates. For example, for the reasonably well-resolved $8\Delta x$ wave, second-order diffusion removes ~15% of the amplitude each time step, whereas the higher-order approaches remove 1–2% or less. For the $4\Delta x$ wave, second-order diffusion removes about twice the amplitude per time step as does the fourth-order diffusion.

Figure 3.35 shows the result of the damping in Eq. 3.62, over multiple time steps, by the second- and sixth-order diffusion terms ($n = 2$ and $n = 6$, respectively). A square-wave, even though it possesses first-order discontinuities that would typically not exist for most variables in the atmosphere, is chosen for the initial shape of the feature to be diffused.

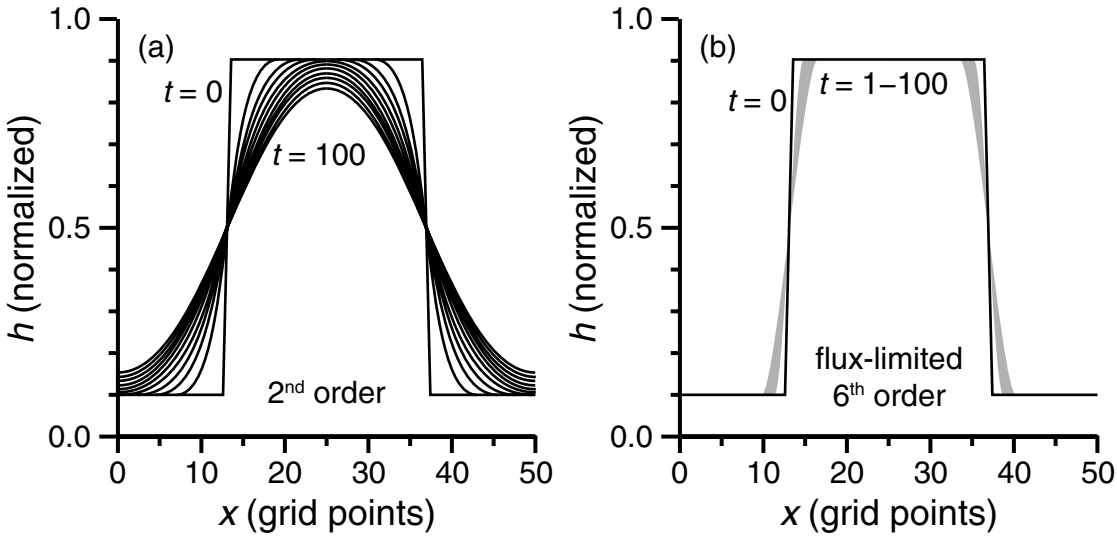

Fig. 3.35 Damping of a square wave by second-order diffusion (a, $n = 2$ in Eq. 3.62), and sixth-order diffusion with a flux limiter to prevent added noise from Gibbs phenomenon (b, $n = 6$ in Eq. 3.62). The curves in panel (a) correspond to the solution for the wave at intervals of $10\Delta t$. The gray area in panel (b) defines the envelope of the ten $10\Delta t$ -interval curves. As with Fig. 3.34, $K\Delta t$ was chosen such that there is 100% damping of the $2\Delta t$ wave each time step. Adapted from Xue (2000).

Figure 3.35a shows that 100 time steps of the second-order diffusion, in addition to damping the small scales that make up the corners of the square wave, suppresses the amplitude of the $25\Delta x$ main wave by ~15%. Higher-order diffusion terms, even though they are more scale-selective, can introduce new local extrema, or noise, near large gradients in the model solution. This effect is known as the Gibbs phenomenon. In these schemes, the diffusive flux is not necessarily down gradient, leading to the nonphysical artifacts. One remedy to this problem is described in Xue (2000), wherein diffusive fluxes are set to zero whenever they are in the same direction as the gradient. Figure 3.35b is analogous to 3.35a, but pertains to the sixth-order diffusion with this flux limiter.

Figure 3.26 provides an additional illustration of the effect of diffusion operators of different order. Here, the one-dimensional, shallow-fluid, second-order, advection equation, with a Courant number of 0.1 that produced the "no diffusion" solution, is rerun with different-order diffusion terms. Even though the fourth-order-advection term used for Fig. 3.28 is more realistic, the second-order approach is employed in this illustration because it is easier to visualize the effects of the diffusion on the different wavelengths. For each experiment, the diffusion coefficient was chosen such that 10% of the $2\Delta x$ wave amplitude was damped each time step. The second-order diffusion damps all wavelengths, including the main wave. Fourth-order diffusion is more selective in its damping, affecting shorter waves the most. The sixth-order term only touches the very-shortest waves, especially those on the right side of the domain that have not yet exited the grid. Further discussions and examples of the use of diffusion terms or filters to remove small-scale wave energy can be found in Shapiro (1970, 1975), Raymond and Garder (1976, 1988), Raymond (1988), Durran (1999), Xue (2000), and Knievel *et al.* (2007).

It is worth noting that the diffusion term should be calculated on a quasi-horizontal surface, and not on a model's constant sigma or potential temperature vertical-coordinate surfaces. For example, the temperature on a terrain-following sigma surface will typically be a minimum over a mountain. Thus, diffusion of temperature (from high to low values) in the thermodynamic equation, if calculated on the sigma surface, will produce temperature increases over the mountain. This temperature rise over the elevated terrain will result in the development of an erroneous thermally direct wind circulation. Thus, for each grid point, the variable being diffused should be vertically interpolated from the model-coordinate surface to the horizontal surface passing through the grid point. The value of the diffusion term calculated on the horizontal surface should then be used in the tendency equation.

To illustrate the effect of this diffusion on the spatial spectrum of a model variable, Fig. 3.36 shows the kinetic-energy spectrum for a WRF-model forecast having a 10-km grid

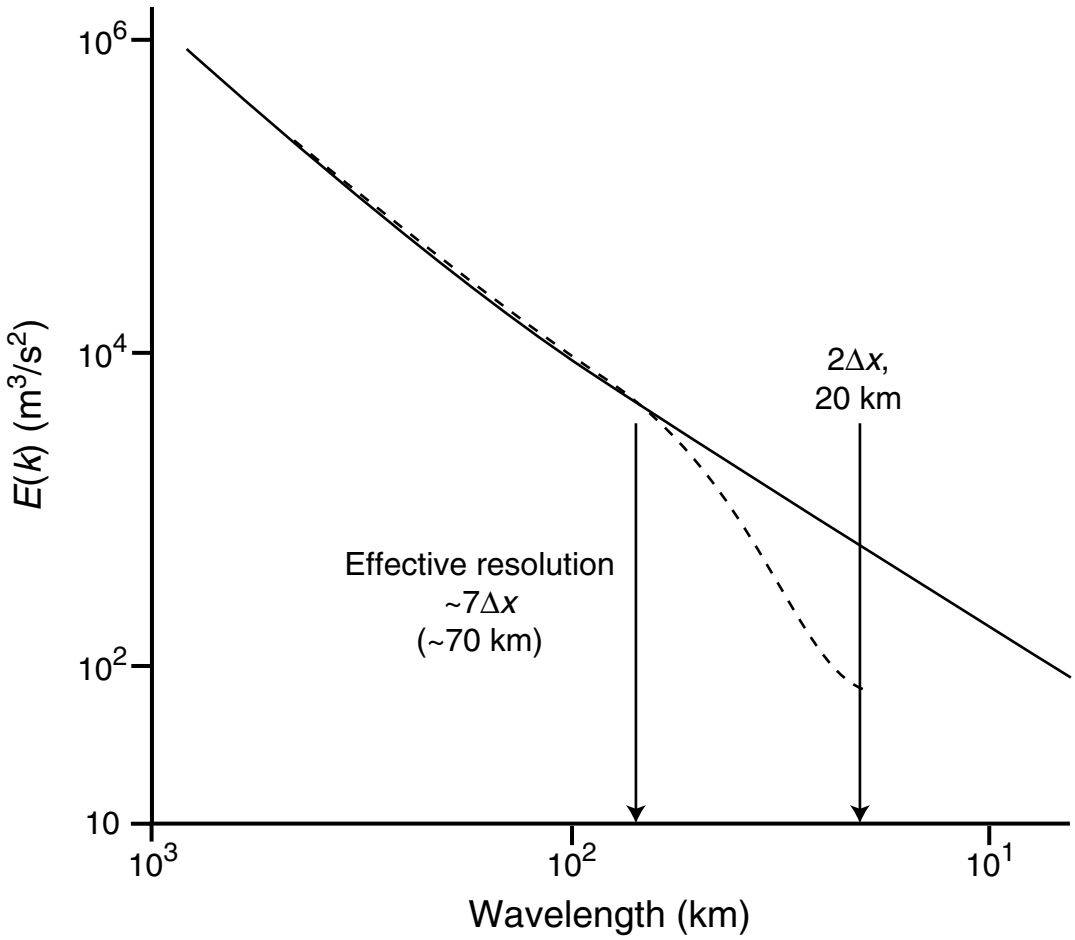

Fig. 3.36 The effect of diffusion on the kinetic-energy spectrum for a WRF-model forecast having a 10-km grid increment. The expected slope of $k^{-5/3}$ is shown as a reference, and is reproduced by the model for wavelengths above $7\Delta x$. But the energy between the $2\Delta x$ and $7\Delta x$ wavelengths has been damped by the diffusion, resulting in an effective resolution of 70 km, not 20 km. Adapted from Skamarock (2004).

increment. The expected slope of $k^{-5/3}$ is shown as a reference, and is reproduced by the model for wavelengths above $7\Delta x$. But the energy between the $2\Delta x$ and $7\Delta x$ wavelengths has been damped by the diffusion, resulting in an effective resolution of 70 km, not 20 km. See Frehlich and Sharman (2008) for an additional analysis of effective model resolution.

Implicit numerical diffusion

Some finite-difference schemes, such as the forward-in-time and backward-in-space method described in Section 3.4.2, and the odd-order Runge–Kutta time-differencing schemes mentioned in Section 3.4.4, selectively damp certain wavelength bands (e.g., Fig. 3.23). If the damping is controllable and sufficiently scale-selective, this is a desirable property of the differencing scheme, and an explicit numerical diffusion term may not be needed to damp poorly resolved shortwave energy.

Grid diffusion

This process results from the fact that the model variables at each grid point are affected each time step by the variables at neighboring grid points, through terms with spatial derivatives. The grid increment and the time step thus define the rate at which grid diffusion causes the spread of atmospheric properties through every nonzero term in the equations that involves a spatial finite-difference expression. Naturally, processes in the real atmosphere such as advection, turbulent diffusion, and inertia–gravity wave motion cause information to spread spatially, but grid diffusion is nonphysical, ubiquitous, and can act rapidly. For example, consider a model with a 25-km grid increment. If we assume a maximum wave speed of 50 m s^{-1}anywhere on the grid, and if we conservatively require that the Courant number be less than 0.7, the time step would be 350 s. Also assume that the wind speed in the boundary layer is 5 m s^{-1}. With a three-point finite-difference approximation to an advection term, where the tendency at each grid point uses information that is one grid increment away, the information propagates at a speed of $\Delta x/\Delta t$, or over 70 m s^{-1}. This fictitious propagation is over 10 times faster than the speed of the transfer of information by advection in the boundary layer. Even though this process is nonphysical, the resulting smoothing or mixing is sometimes used to represent the real diffusion in the atmosphere. Unfortunately, the effect is not controllable in terms of its overall strength and its ability to selectively damp small scales.

3.4.8 Numerical implications of the choice of the model vertical coordinate

The following sections provide a brief summary of the numerical implications of the use of the historically most-common choices for the vertical coordinate in NWP models. See Sundqvist (1979) for more information.

Height above sea level

At face value, this would seem like an intuitively appealing coordinate. In particular, the coordinate surfaces are fixed relative to Earth's surface, and, unlike some other options, the

pressure-gradient force in the momentum equations is represented by one term, the gradient of pressure. There are significant problems, however. Because the height surfaces are penetrated by orographic features at low levels, there are areas of grid points on the coordinate surfaces where atmospheric properties are undefined. This makes it virtually impossible to properly calculate derivatives on the constant-z surfaces at the grid points located next to these voids, and it is impossible to employ spectral methods to define the horizontal variation of model variables. Lastly, grid-point-model codes typically scan systematically through the rows and columns of grid points in the matrix of points, and it becomes very cumbersome to interrupt this process with breaks in the pattern where the points do not exist.

There is an approach to the use of z coordinates, called the volume-fraction or shaved-grid-cell method (Adcroft *et al.* 1997, Steppeler *et al.* 2002, Walko and Avissar 2008b), that avoids some of the above disadvantages. Computations for cells that are partially embedded below the terrain surface are modified to account for the kinematic effect of the barrier. Even with this approach, a disadvantage relative to terrain-following vertical-coordinate systems is that employing high vertical resolution immediately above the surface requires the use of many thin model layers when there is a large variation in topographic height, thus increasing the computational expense. Another disadvantage is that the grid cells that intersect topography have different properties than the rest, and thus they must be treated differently in the numerical algorithms.

Pressure

Pressure is the variable that radiosondes use to define vertical position when observed values of dependent variables are transmitted to the ground station. So, in some sense, it may be reasonable to use this as the vertical coordinate in a model in which the radiosonde observations must be assimilated. However, this coordinate has virtually the same problems as does the height system. But the difficulty with orography interrupting the surfaces is even more problematic because the heights of the pressure surfaces change with time, and thus does the pattern of the grid points that are masked. During the integration, grid points will appear and disappear, and it is very difficult to assign realistic physical properties to grid points that are only temporarily part of the calculations.

Potential temperature

Under hypothetical adiabatic conditions, the potential temperature (θ) of a parcel does not change as it moves, and the parcel remains on θ surfaces. That is, θ surfaces are *material surfaces*, and when these surfaces are the vertical coordinate surfaces, the vertical motion ($d\theta/dt$) is zero. Even though both real and model atmospheric processes are close to being adiabatic, outside of the boundary layer and where phase changes are not consuming or releasing latent heat, they are never perfectly adiabatic because radiative flux divergences are never zero. So, over those large volumes where $d\theta/dt$ is small,

vertical advection is also small, and artificial grid diffusion in the vertical is small. This reduced artificial vertical spread of moisture and other scalar quantities in θ coordinates leads to their more-realistic transport.

Because this variable is obviously linearly related to temperature, where temperature gradients are largest the model potential-temperature surfaces are more-tightly packed. This means that there is more vertical resolution in the model where it is needed most in order to represent large gradients, for example along quasi-horizontal frontal surfaces and near the tropopause. Figure 3.37 illustrates a cross section of a front in both pressure (a) and θ (b) coordinates. In θ coordinates (Fig. 3.37b), the strong wind shear in the frontal zone (shaded area) spans one-quarter of the vertical extent of the cross section, whereas in pressure coordinates (Fig. 3.37a) this shear is concentrated within a narrow region in the vertical. Also, the fact that the coordinate surfaces approximate material surfaces implies that horizontal gradients will be smaller than when the coordinate surfaces cut across fronts. Thus, the truncation error associated with horizontal and vertical derivatives will be smaller. Because isentropic surfaces intersect Earth's surface, with or without orography, potential temperature has disadvantages similar to those of the pressure and height coordinates. Also, near the strongly heated surface of Earth, potential temperature can decrease with height (superadiabatic lapse rates) in a shallow layer, below where it displays its normal increase with height. Thus, in a model with this vertical coordinate, any lapse rate that approaches the adiabatic value during the forecast must be artificially adjusted to a

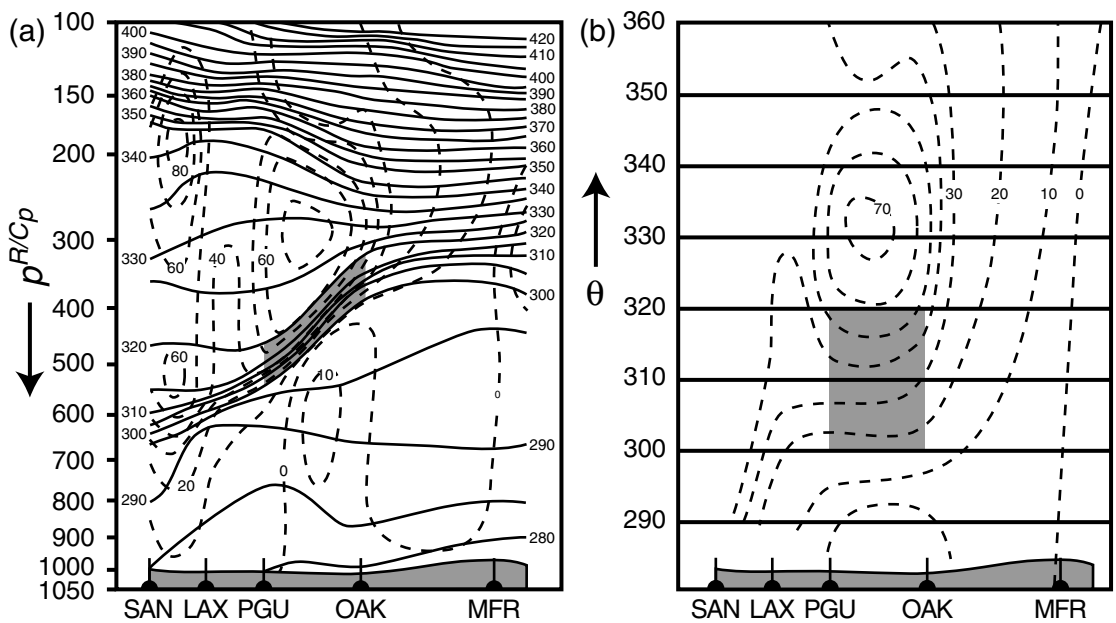

Fig. 3.37 Cross section of a front in pressure (a) and isentropic (b) coordinates, where the horizontal axis is north–south along the coast of the western USA. The gray area in the two cross sections spans the same volume of atmosphere. The solid lines are isentropes and the dashed lines are isotachs. From Benjamin (1989), based on Shapiro and Hastings (1973) and Bleck and Shapiro (1976).

more-stable one to avoid the situation where potential temperature is the same at two places on the vertical scale. Another problem that is in common with sigma coordinates (see below) is that the pressure-gradient term appears as the horizontal derivative of the Montgomery potential, which consists of the sum of two terms ($C_P T + gz$). The horizontal derivatives of these individual terms can be large, and the pressure gradient is represented by a small difference between the two large derivatives. Thus, noncancelling truncation errors in the derivatives can produce large errors in the pressure gradient.

Sigma-p

The so-called sigma coordinate systems are terrain following and thus avoid the above noted problems of the height, pressure, and potential-temperature coordinates that intersect the land or water surface. The pressure-based sigma coordinate (Phillips 1957b, Gal-Chen and Somerville 1975) is defined as

$$\sigma = \frac{p - p_t}{p_s - p_t},$$

where p_t is a constant pressure chosen for the top of the model, p_s is the surface pressure, and p is local pressure at any point in the column. If the top of the model is defined to be at the top of the atmosphere, we simply have $\sigma = p/p_s$. For $p = p_s$, the condition at the surface, $\sigma = 1$ everywhere. For $p = p_t$, $\sigma = 0$ everywhere. Thus, over any column of model atmosphere, $0 < \sigma < 1$. Because surface pressure and local pressure are functions of time, the vertical positions of sigma-coordinate surfaces will change. Figure 3.38 shows a vertical cross section of sigma surfaces in a model of the eastern USA, where the model top is located at 500 hPa.

As noted earlier, the pressure gradient in equations expressed in sigma vertical-coordinate systems consists of two terms. In the sigma-p system, one contains the derivative of the surface pressure, p^*, and the other contains the derivative of the geopotential height of the sigma surface, as shown in the following pressure-gradient term from the first equation of motion.

$$\frac{\partial p^* u}{\partial t} \propto -m p^* \left(\frac{RT}{p^* + \frac{p_t}{\sigma}} \frac{\partial p^*}{\partial x} + \frac{\partial \phi}{\partial x} \right).$$

Each term is potentially very large, and the small difference between them represents the pressure gradient force. Where there are large terrain-elevation gradients, these individual terms become especially large, and truncation errors that do not cancel in the two terms create erroneous pressure gradients and accelerations. This issue is partly addressed by defining a base state condition, and by using perturbation forms of the equations where the derivatives apply to departures from a mean state. See Mesinger *et al.* (1988) for a discussion of the history and shortcomings of this coordinate.

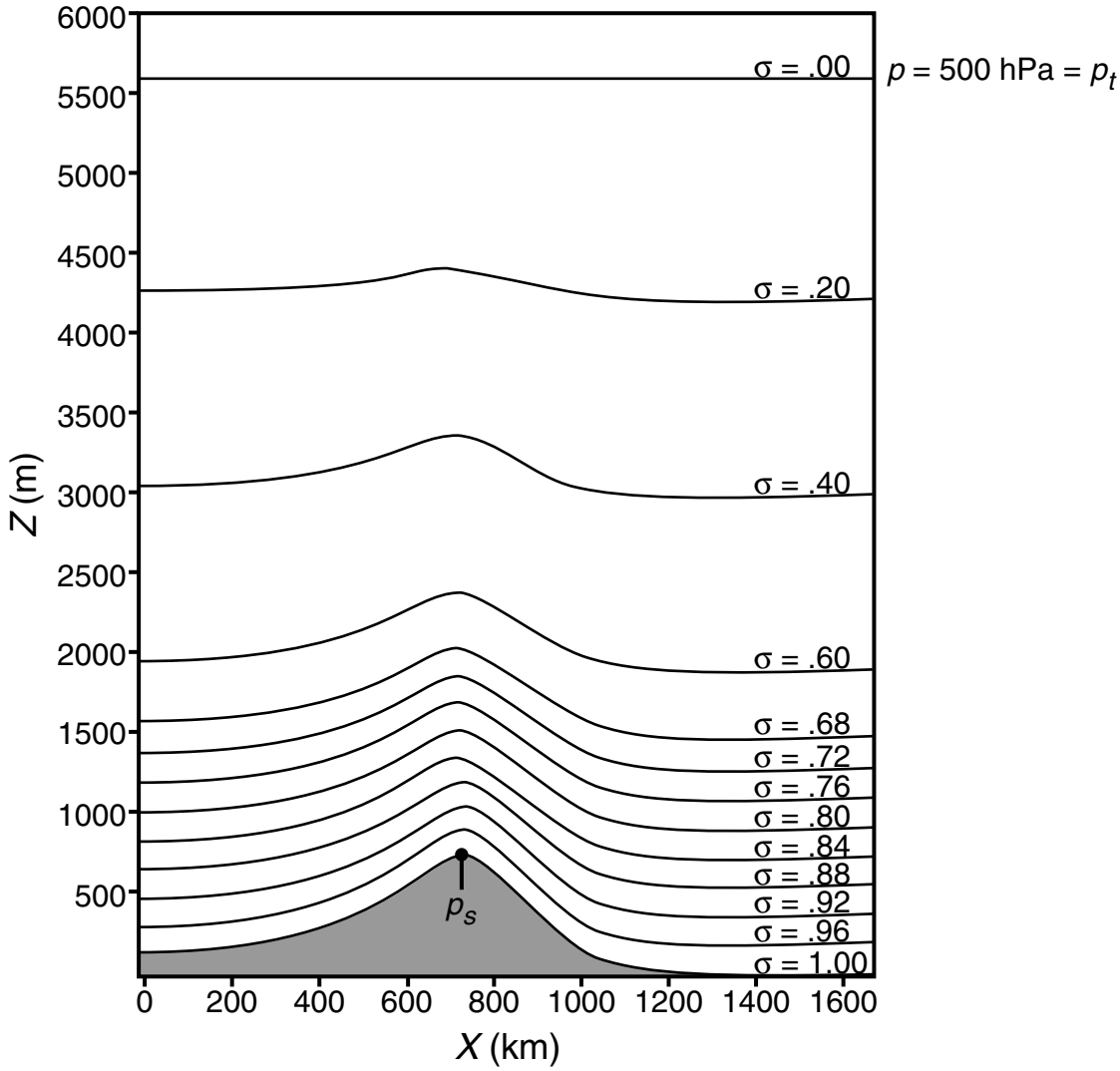

Fig. 3.38 Cross section of sigma surfaces for a model grid over the eastern USA. The model top is at 500 hPa. Adapted from Warner *et al.* (1978).

Sigma-z

Where the above sigma-p coordinate system is normalized by the pressure depth of the model atmospheric column, the sigma-z system (Kasahara 1974) is normalized by the physical depth of the atmosphere. Specifically

$$\sigma = \frac{z_t - z}{z_t - z_s},$$

where z_t is the constant height chosen for the top of the model, z_s is the surface height, and z is local height at any point in the column. Obviously the heights of these coordinate surfaces do not vary with time. Like the sigma-p system, the coordinate ranges from 0 to 1 through the depth of the model atmosphere.

Hybrid isentropic-sigma

This approach involves the use of terrain-following sigma coordinates in the lower troposphere and isentropic coordinates above. It retains the advantages of the isentropic representation, but avoids the previously mentioned major shortcoming that occurs in the boundary layer. A variety of these hybrid schemes has been developed. Benjamin *et al.* (2004b) should be consulted for more information.

Step-mountain (eta)

A vertical coordinate described by Mesinger *et al.* (1988), Black *et al.* (1993), Black (1994), and Wyman (1996) is the step-mountain coordinate, also known as the eta coordinate. Figure 3.39 shows a vertical cross section of the coordinate surfaces. The approach

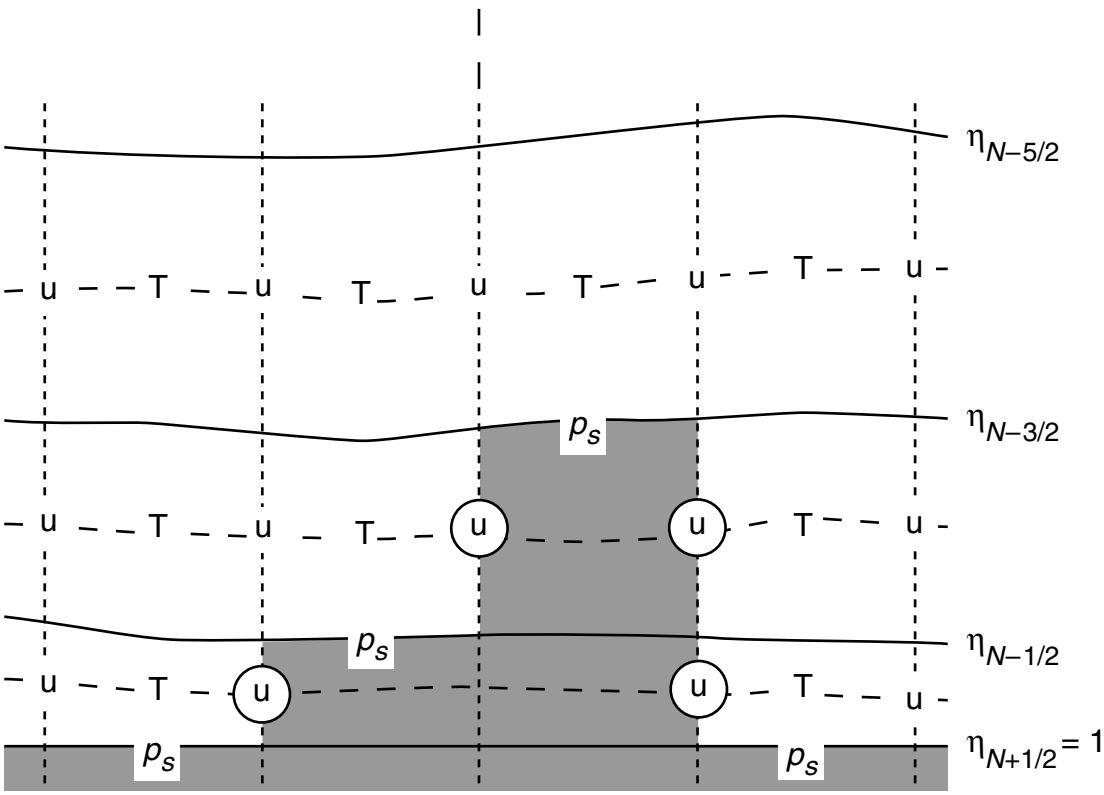

Fig. 3.39 Cross section of the three lower model levels for the step-mountain coordinate system, showing where variables are defined. The shaded area represents the land surface. From Mesinger *et al.* (1988).

was motivated in order to avoid the problems of the sigma system that are associated with steep orographic slopes. As seen in the figure, the orography is constructed from the three-dimensional grid boxes of the model, with the surface elevation being defined with a discrete set of values. At the vertical surfaces, which are essentially internal hard boundaries, the normal component of the velocity (circled in the figure) is zero. The coordinate surfaces are quasi-horizontal. The eta coordinate is defined as

$$\eta = \frac{p - p_t}{p_s - p_t}\eta_s = \eta_s\sigma,$$

where σ is the definition of the sigma coordinate provided above, p_t is the pressure at the model top, p_s is the pressure at the surface, and

$$\eta_s = \frac{p_{rf}(z_s) - p_t}{p_{rf}(0) - p_t},$$

which is the value of eta at Earth's surface. The reference pressures, p_{rf}, correspond to the pressures at the interface between model layers. For $p_t = 0$, the eta-coordinate's definition simplifies to

$$\eta = \eta_s\sigma = \frac{p_{rf}(z_s)}{p_{rf}(0)}\sigma.$$

For flat terrain ($z_s = 0$), the eta coordinate is identical to the sigma coordinate. See the above references for example simulations using this coordinate.

3.4.9 Time smoothers and filters

Propagating disturbances can be damped by both space and time smoothers. The explicit numerical diffusion operators described above in Section 3.4.7 are intended to damp or smooth small-scale disturbances in terms of the spatial variability. Other operators smooth in the time dimension. Again, propagating disturbances may be smoothed in either way. In particular, there are situations where, with centered time-differencing methods, the model solutions at odd and even time steps can depart from each other. This *separation of the solution* results from the fact that, after the initial forward time step, the leapfrog differencing allows even and odd time steps to affect each other only through the derivative. That is, the leap is from even-to-even and odd-to-odd time steps. This $2\Delta t$ oscillation can easily be damped with a time smoother. One of the most popular is described by Asselin (1972),

$$\alpha^\tau = (1 - \beta)\alpha^\tau + \frac{\beta}{2}(\alpha^{\tau+1} + \alpha^{\tau-1}),$$

where α is any dependent variable and a typical value of β is 0.1. It can be applied at every time step, or intermittently.

3.5 Lateral-boundary conditions

The values of dependent variables must be specified (i.e., not internally calculated) at the lateral edges of the computational grids of Limited-Area Models (LAMs; that is, everything but global models). Even though some global models used for weather or climate prediction are capable of resolving mesoscale processes, for the foreseeable future there will be a need to embed even-higher resolution LAM grids within the coarser models. Thus, the challenges of dealing with Lateral-Boundary Conditions (LBCs) will need to be addressed. The LBCs should have the following properties.

- Meteorological features should propagate from coarse- to fine-mesh grids without significant distortion.
- Inertia–gravity waves should propagate through the boundary, especially longer-length waves that are related to important physical processes such as geostrophic adjustment. Shorter waves may be damped on outflow, but they should not be reflected.
- The LBCs should not allow artificial dynamic/numerical feedbacks between grids that can cause a catastrophic termination of the model integration.

Note that there are numerous references that describe various kinds of evidence of the potentially serious effect of LBC error on LAM forecasts (e.g., Miyakoda and Rosati 1977, Oliger and Sundstrom 1978, Gustafsson 1990, Mohanty *et al.* 1990, and Warner *et al.* 1997). Much of the following analysis of LBC effects is based on Warner *et al.* (1997).

3.5.1 Sources of LBC error

Because the negative impacts of LBCs on LAM solutions are inevitable, our objective should be to understand the nature of the problems well and learn how to mitigate their effects. The LBC's negative influences can be attributed to at least six factors.

- *Low resolution of LBC data* – Open LBCs (see Section 3.5.3) are defined based on forecasts from coarser-resolution models or analyses of observations, depending on whether the LAM is being used for operational or research applications. In either case, the horizontal, vertical, and temporal resolution of the boundary information is generally poorer than that of the LAM, and thus the boundary values interpolated to the LAM grid at every time step have the potential of degrading the quality of the solution.
- *Errors in the meteorology of the LBCs* – Even if the LBC-data resolution is hypothetically similar to that of the LAM, and there is little interpolation error, the quality of the LBC data may be erroneous for other reasons, especially if they are based on other model forecasts. That is, the forecast that provides the LBCs may simply be wrong in some important respect having nothing to do with its resolution. In any case, these errors will be transmitted to the LAM domain at the grid interface.
- *Lack of interactions with larger scales* – Specified LBCs determine the computational-domain-scale structure of the meteorological fields. But, these longer wavelengths

cannot interact with the model solution on the interior. This limited spectral interaction can affect the evolution of the LAM forecast because the LAM solution cannot feed back to the large scales.

- *Noise generation* – The specific LBC formulation used can produce transient, nonmeteorological, inertia–gravity modes on the LAM domain. Even though these modes are thought to not interact strongly with the meteorological solution, they are superimposed on the physically realistic fields and can complicate the interpretation of the forecast.
- *Physical-process parameterization inconsistencies* – The physical-process parameterizations may, sometimes out of necessity, be different for the LAM and the coarser-resolution model providing the LBCs. The resulting inevitable differences in the solution at the boundary may cause spurious gradients and feedbacks between the two grids, which can influence the solution on the LAM domain.
- *Phase- and group-speed contrasts* – Earlier in this chapter it was shown that some differencing schemes can cause phase- and group-speed errors whose magnitude depends on how well a wave is resolved on the grid. Thus, as a wave passes between computational areas with different grid increments, waves can be stretched or compressed. Browning *et al.* (1973) refer to a numerical refraction effect resulting from the phase-speed differences, that causes "unexpectedly large errors" on the coarse mesh of two-way interacting grids.

3.5.2 Examples of LBC error

At least four general types of studies have been performed, from which we can gain insight into LBC error. One involves the application of model computational domains of different size, and from these simulations a direct determination is made of the effect of the proximity of the lateral boundaries on some measure of the quality of the simulation. Another type can be grouped into the general category of mesoscale predictability studies wherein a control simulation is first performed with a LAM. Then, perturbations are imposed on the model initial conditions or LBCs, and the differences between the model solutions with and without the perturbations are analyzed and ascribed to specific factors, including the LBCs. A third category of study uses an adjoint model from which actual LBC-sensitivity fields are produced directly. Relevant studies from which we can gain insight are described here. And a fourth type is the Big-Brother–Little-Brother experiment, which is discussed in more detail in Chapter 10.

Domain-size sensitivity studies

One of the first studies of the effects of defining LAM LBCs with a coarser-resolution forecast was that of Baumhefner and Perkey (1982). A LAM (Valent *et al.* 1977) with a 2.5° latitude–longitude (lat–lon) grid was embedded within, and obtained its LBCs from, a 5° lat–lon hemispheric model (Washington and Kasahara 1970). Both models used the same vertical grid structure and physical-process parameterizations. The LBC "error" was

first assessed by comparing the solution from this nested system with that from a non-nested, 2.5° lat–lon version of the hemispheric model. Figure 3.40 shows the midtropospheric pressure error (difference between the LAM and hemispheric-model solutions) associated with the LBCs for a 48-h forecast period. Large pressure errors with amplitudes of 5–10 hPa propagate rapidly onto the forecast domain at middle and high latitudes, primarily from the west and north boundaries, with speeds of 20°–30° lon day^{-1}. Comparison of this error distribution with the location of synoptic disturbances (not shown) shows that the error maxima are associated with areas in which significant changes are taking place at the boundaries. The fairly inactive large-scale meteorological conditions in the subtropics and tropics generate very little LBC error. For LAM simulations in which the LBCs were provided by a 2.5° lat–lon hemispheric model (i.e., the LAM and hemispheric models had the same horizontal resolution), errors were also large and had a similar distribution, indicating that significant LAM errors in these regions resulted from

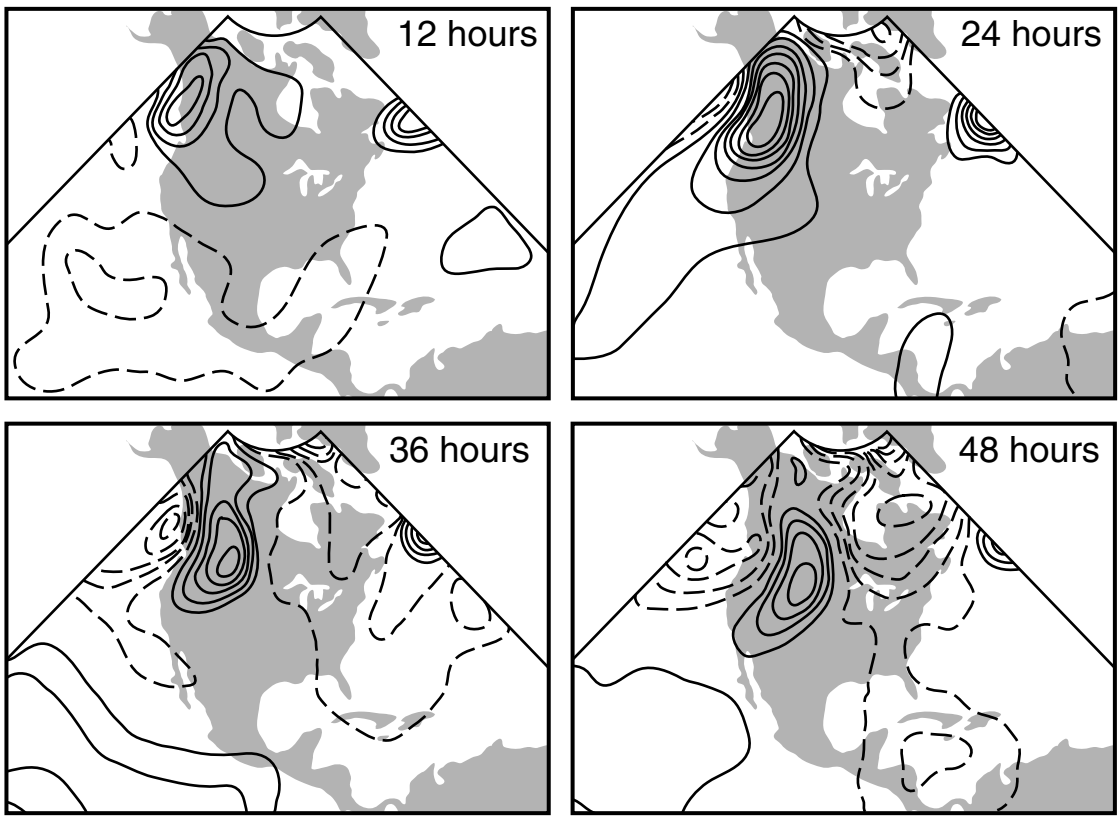

Fig. 3.40 For different simulation lead times, pressure difference at 6 km ASL (about 500 hPa) between simulations from a 2.5° lat–lon hemispheric model and a limited-area model with the same resolution embedded within a 5° lat–lon hemispheric model. The differences are associated with boundary-condition effects. The area delineated is that of the LAM domain. The isobar interval is 1 hPa, and negative values are dashed. Adapted from Baumhefner and Perkey (1982).

the LBC formulation itself and not just the quality of the LBC data. Figure 3.41 summarizes the Root-Mean-Square (RMS) error growth in 500-hPa heights on the limited-area domain associated with the use of LBCs from the 2.5° (dotted curve) and 5° (dashed

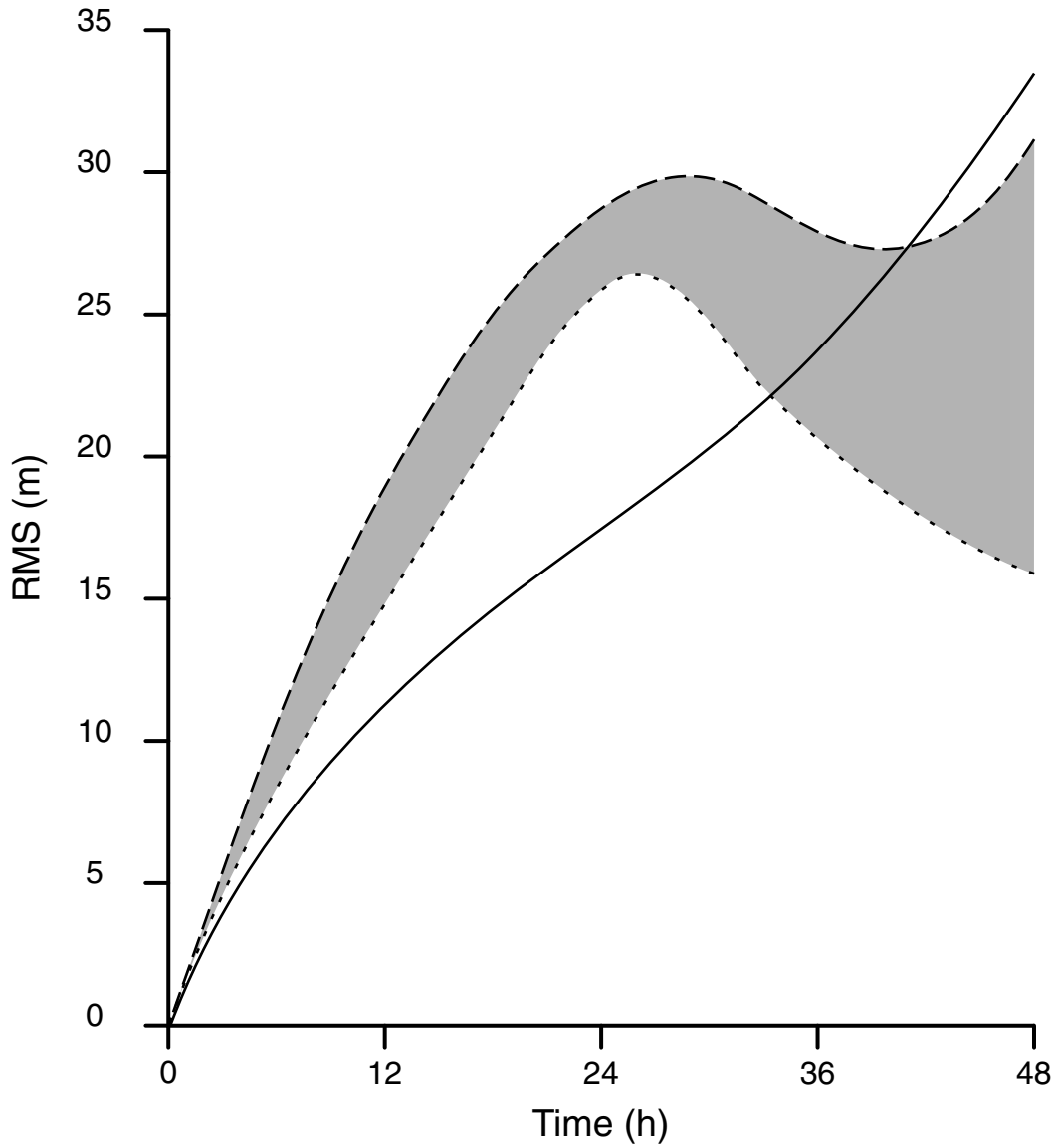

Fig. 3.41 RMS 500-hPa height differences (m), where the solid line shows the difference between 5° and 2.5° hemispheric simulations over the area of the LAM grid, the dashed line shows the difference between the 2.5° hemispheric simulation and that from the 2.5° LAM whose LBCs are provided by the 5° hemispheric simulation, and the dotted line shows the difference between the 2.5° hemispheric simulation and that from the 2.5° LAM whose LBCs are provided by the 2.5° hemispheric simulation. The abscissa is forecast hours. From Baumhefner and Perkey (1982).

curve) lat–lon hemispheric models. It is revealing that the error growth in the LAM is similar whether or not the LBC information was defined by a model of the same or worse resolution. The solid curve shows the difference between the 2.5° and 5° hemispheric simulations over the area of the LAM domain, and represents the error that is associated with the use of the 5° unbounded grid compared to the 2.5° unbounded grid. The most rapid error growth is during the first 24 h for both the 2.5° and 5° LBCs. The fact that the error associated with the 2.5° LBCs decreases after 24 h probably indicates that some of it is related to rapidly propagating and damped transients generated at the lateral boundaries early in the simulation. In contrast, when the 5° LBCs are used there is a continuing propagation of coarse-resolution information that causes the error to generally increase throughout the forecast.

This, of course, is not true forecast error because observations are not being used as a reference. However, it is sobering to see that, when the hemispheric 2.5° simulation is used as a reference, the hemispheric 5° simulation shows smaller error than do either of the 2.5° LAM simulations containing the LBC error. That is, when using the 2.5° hemispheric solution as a standard, higher accuracy is obtained by using only the coarse hemispheric model rather than the coarse hemispheric model with an embedded higher-resolution LAM. In another experiment (not shown), where the computational domain was extended by 20° lon at the east and west boundaries, the center of the domain was protected from LBC contamination for a longer period, but by 48 h the high central latitudes were contaminated from both the east and the west by error propagating inward at about 30° lon day^{-1}. Baumhefner and Perkey (1982) state that "these experiments lead to the not too surprising conclusion that boundary locations should be determined from the forecast time frame selected and the typical boundary error propagation rate." Comparison of model-simulation error defined based on observed conditions for the 2.5° hemispheric model and the 2.5° LAM embedded within the 5° hemispheric model revealed that the LBCs increased the total simulation error at high latitudes by up to 50% after 24 h. That is, the total error growth from all non-LBC sources is about twice that which is related to the LBCs. Naturally, the relative contribution of the LBCs to the total error depends greatly on the overall predictive skill of the model. It is noteworthy that similar results were obtained using two totally different algorithms for specifying the LBCs.

A well-controlled demonstration of the domain-size problem is described by Treadon and Petersen (1993), who performed a series of experiments with 80- and 40-km grid-increment versions of the US NWS Eta model (Black *et al.* 1993, Black 1994) with a winter and summer case. While maintaining the same resolution and physics, they progressively reduced the area coverage and documented the impact on forecast skill. The "control simulation" utilized the full computational domain of the Eta model, while experimental simulations used domains that were progressively smaller, with each having approximately one-half of the area coverage of the next larger domain (Fig. 3.42). In each case, US NWS global spectral, T126, previous-cycle forecasts were used for LBCs. For a winter cyclogenesis case, the 80- and 40-km grid-increment models with the full domain produced reasonably accurate forecasts. However, the forecast on the smallest domain, which had its lateral boundaries close to the area affected by the storm, had 500-hPa RMS

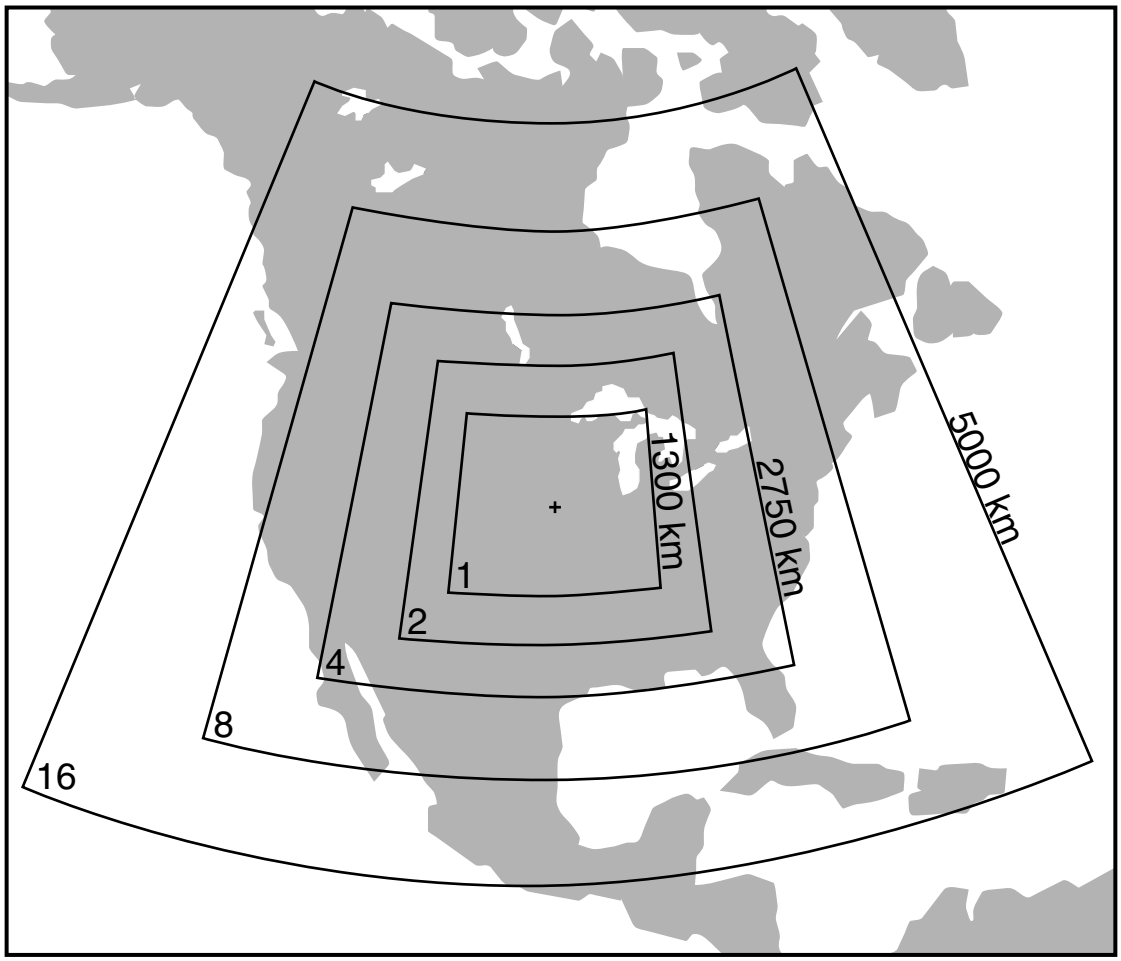

Fig. 3.42 Five integration domains of the 80-km grid-increment Eta model used in the domain-size sensitivity study. The grid number corresponds to the factor by which the grid area is larger than that of the smallest grid. From Treadon and Petersen (1993).

height errors that were twice as large (relative to data analyses) as those of the forecast on the full domain, by only 12 h into the forecast period. In addition, the surface low-pressure center was much weaker than observed, and was erroneously placed, in the forecast on the smallest domain. For a summer case, with much weaker flow over the small domains, the error growth was qualitatively similar to that of the winter case. Again, RMS 500-hPa height errors were more than twice as large on the smallest domain than they were on the largest domain by the 36-h forecast time (Fig. 3.43). An example is shown in Fig. 3.44 of the rapid influence that the LBCs can have at upper levels, even when the cross-boundary flow is weak to moderate. For this summer case, Fig. 3.44 illustrates two 12-h simulations of 250-hPa isotachs from the 40-km grid-increment Eta model. Figure 3.44a shows a

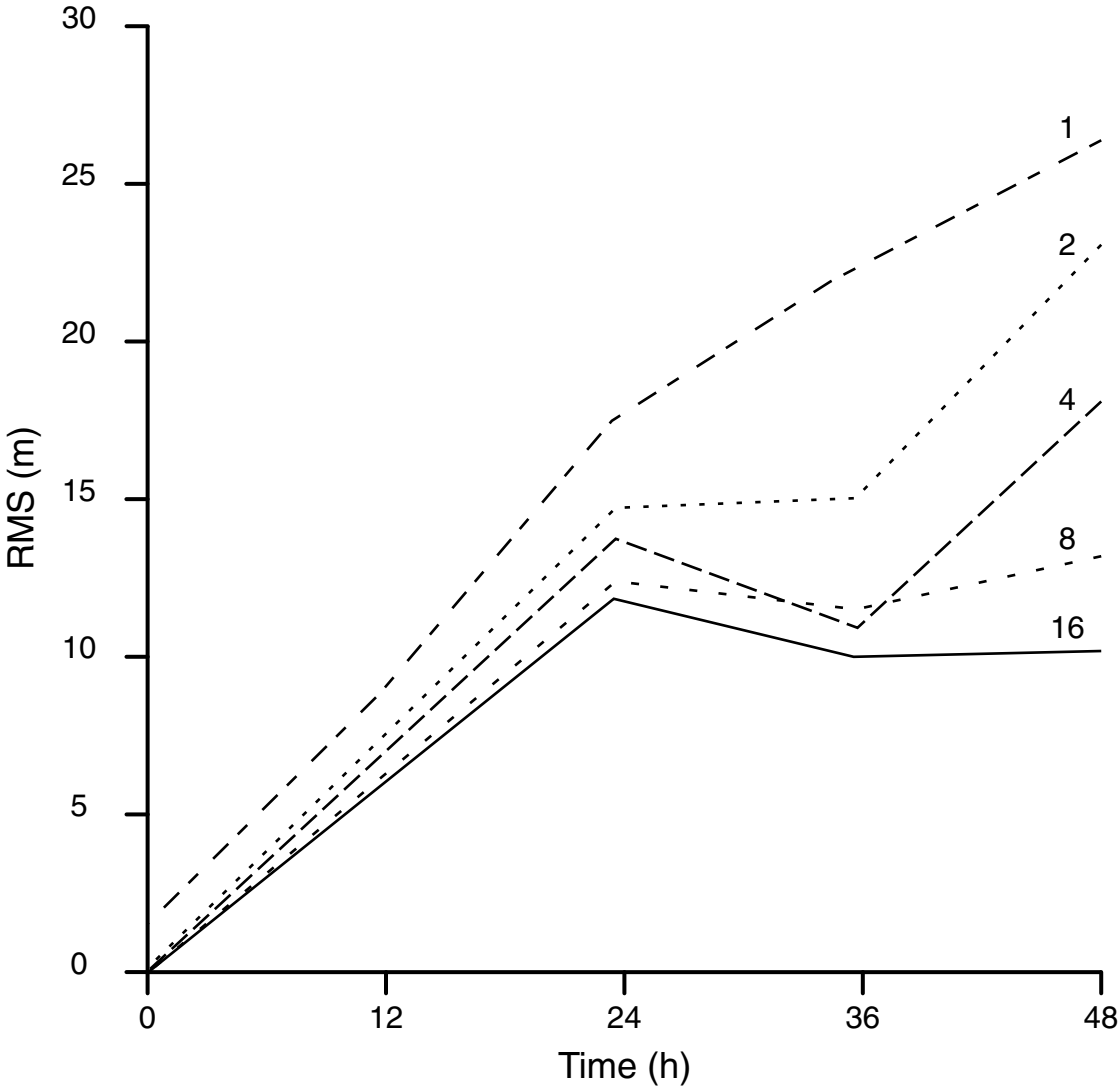

Fig. 3.43 Temporal evolution of the RMS 500-hPa height errors (relative to data analyses) associated with the use of each of the five computational grids in Fig. 3.42, for a forecast initialized at 0000 UTC 3 August 1992. For each experiment, the errors were calculated for the same area of the innermost grid (number 1). The abscissa is forecast time. The grid numbers correspond to those defined in Fig. 3.42. From Treadon and Petersen (1993).

strong narrow jet streak simulated on the largest domain, while Fig. 3.44b shows that the same feature on the smallest domain (with the same resolution) has been considerably smoothed.

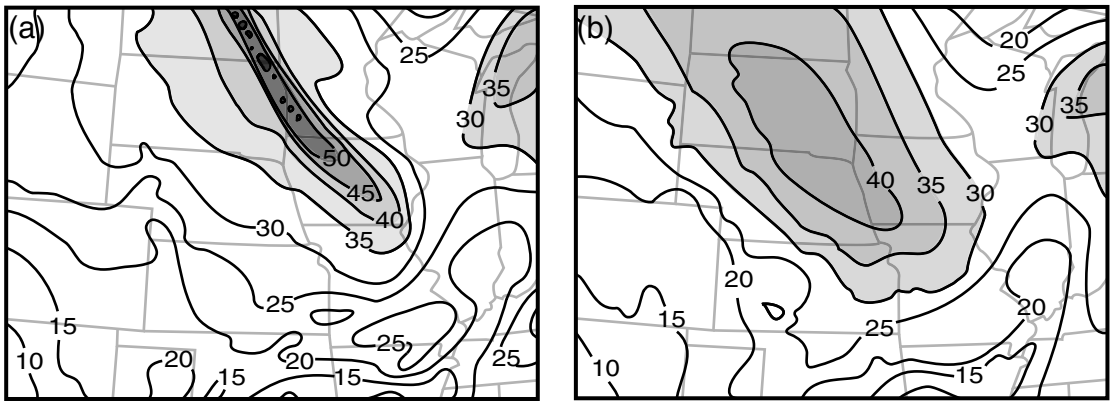

Fig. 3.44 Twelve-hour simulations of 250-hPa isotachs (m s^{-1}) from the 40-km grid-increment Eta model initialized at 1200 UTC 3 August 1992, based on experiments that used the largest computational domain (a) and the smallest (b). Both maps apply for the area of the smallest grid. The isotach interval is 5 m s^{-1}. From Treadon and Petersen (1993).

Mesoscale predictability studies

Predictability studies with mesoscale LAMs have demonstrated that error growth is much different than what has been documented for global models (Anthes *et al.* 1985, Errico and Baumhefner 1987, Vukicevic and Paegle 1989, Warner *et al*. 1989). When small perturbations (errors) are added to the initial conditions (but not the boundary conditions) of a LAM, the simulation from the perturbed initial state and that from the unperturbed (control) initial state do not diverge as they would with an unbounded model. The perturbed atmosphere on the domain interior is advected out of the domain at the outflow boundaries, and the use of identical LBCs in the two simulations causes unperturbed atmosphere to be swept in at the inflow boundaries.

In a predictability study that is revealing of LBC effects, Vukicevic and Errico (1990) used a relatively coarse resolution version of The Pennsylvania State University–NCAR Mesoscale Model Version 4 (MM4) with a grid increment of 120 km for a 96-h simulation of Alpine cyclogenesis. The LBCs were defined for MM4 using data analyses, and simulations from the NCAR global Community Climate Model Version 1 (CCM1) that was initialized at the same time as the LAM. In one experiment, a control simulation was first performed with MM4, and then the initial conditions were perturbed and the model was again integrated. The LBCs were based on analyses of data and were thus "forecast-error free" and the same for both simulations. Figure 3.45 shows the 96-h 500-hPa geopotential-height differences between the two simulations. The largest differences between the two simulations are on the eastern, downwind half of the domain because the identical LBC data strongly influence the model solutions on the western half.

To gain further insight about LBC effects on LAM solutions, an additional experiment used normal (control) and perturbed-initial-condition CCM1 forecasts to define the LBCs of a corresponding pair of MM4 forecasts that had initial conditions that were identical

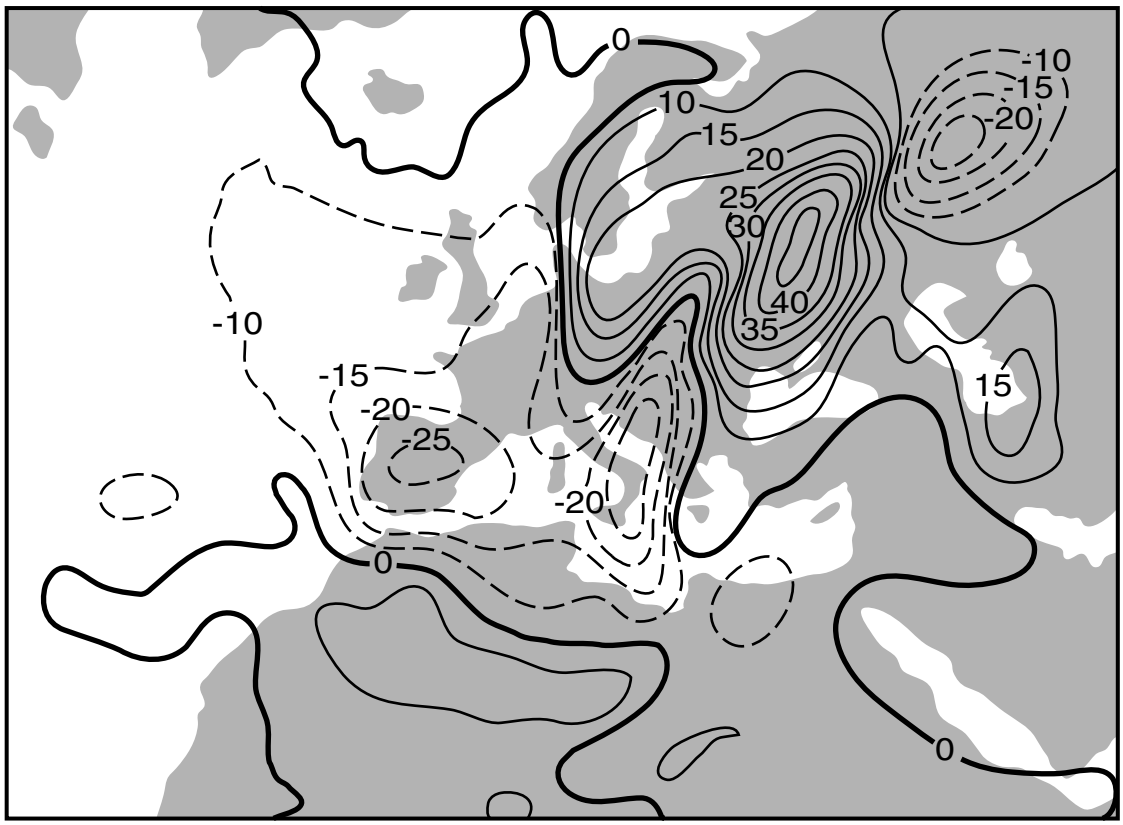

Fig. 3.45 The 500-hPa 96-h geopotential-height difference between a control simulation with MM4 and a parallel one with perturbed initial conditions. The contour interval is 5 m. The LBCs were identical and based on analyses of observations. From Vukicevic and Errico (1990).

and equal to those of the control CCM1 simulation. The perturbed CCM1 initial conditions were defined so as to emulate expected operational measurement errors. Thus, this experimental design has considerable relevance to operational forecasting with a LAM because it isolates the effects of normal errors in a coarse-mesh forecast on the dynamical evolution of a LAM forecast for which it provides LBCs. Figure 3.46 shows the 500-hPa geopotential-height difference in the two 6-h LAM solutions, where differences of over 10 m appear near the domain center over Europe. During this short time, high-frequency transient modes resulting from the LBC formulation have contaminated the entire domain. It is important to recognize that the LAM domain employed here has perhaps four times the area of many LAMs, and thus the LBC error effects would normally be felt on considerably shorter time scales. Based on these results, Vukicevic and Errico (1990) state that "medium range forecasts with nested limited-area models may not significantly reduce RMSEs relative to the same forecasts performed with global models."

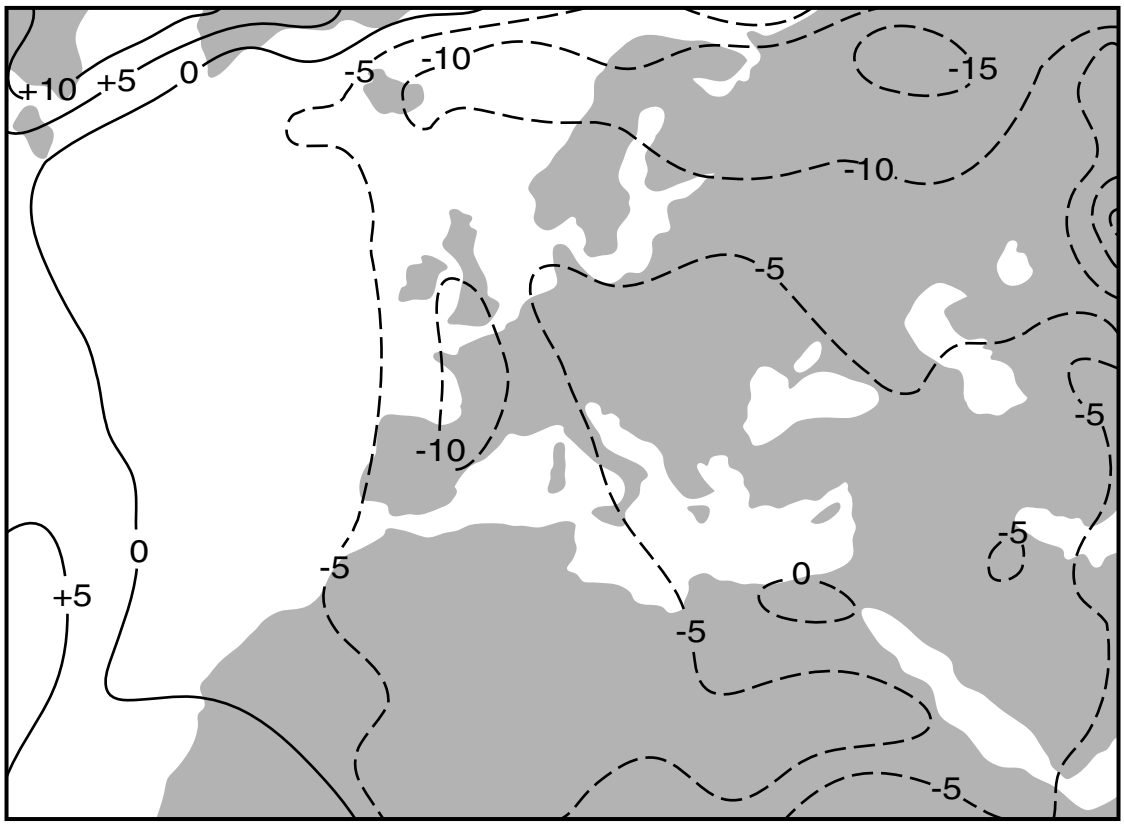

Fig. 3.46 The 500-hPa geopotential height difference between a 6-h control simulation with MM4 and a parallel one with perturbed LBCs. The contour interval is 5 m, where negative values are dashed. Adapted from Vukicevic and Errico (1990).

Adjoint sensitivity studies

Variational techniques employing an adjoint model have been used to investigate the sensitivity of LAM forecasts to initial conditions and boundary conditions. The adjoint operator produces fields that indicate the quantitative impact on a particular aspect of the forecast of any small, but arbitrary, perturbation in initial conditions, boundary conditions, or model parameters. This approach has an advantage over the traditional types of predictability studies discussed above in that the resulting dependencies are not sensitive to the specific perturbations applied to the initial or boundary conditions. For a more in-depth discussion of this technique, the reader should consult Hall and Cacuci (1983), Errico and Vukicevic (1992), and Errico (1997).

Errico *et al.* (1993) applied this approach to investigate the sensitivity of LAM simulations to conditions on the domain interior and LBCs. A dry version (no moisture variables) of the MM4 model and its adjoint were employed, where the model had a grid increment of 50 km and 10 computational layers. The LBCs were provided by linear temporal interpolation between 12-h T42 analyses from the European Centre for Medium-Range

Weather Forecasts (ECMWF). The sensitivity was tested in 72-h simulations of both a summer and a winter case. A number of aspects of the simulations were investigated relative to their sensitivity to initial and boundary conditions. We will concentrate on the influence of the LBCs on the 72-h relative vorticity at the 30 grid points on each computational level that are within 150 km of the center of the domain.

Figure 3.47a shows the sensitivity of the 72-h relative vorticity in a small column in the center of the domain to perturbations of the initial 400-hPa *v*-component of the wind on the domain interior for the winter case. (For further discussion of the sensitivity metric, see Errico *et al.* 1993.) For comparison, Fig. 3.47b illustrates the sensitivity of the same 72-h vorticity to the *v*-component of the wind on the lateral boundaries. The LBC-sensitivity metric extends over four rows and columns of grid points near the boundary because the LBC formulation in this model is such that LBCs are defined at all four points closest to the boundary. The isopleth intervals differ greatly between Figs. 3.47a and 3.47b (see caption). The LBC and grid-interior sensitivities are only in the upwind directions to

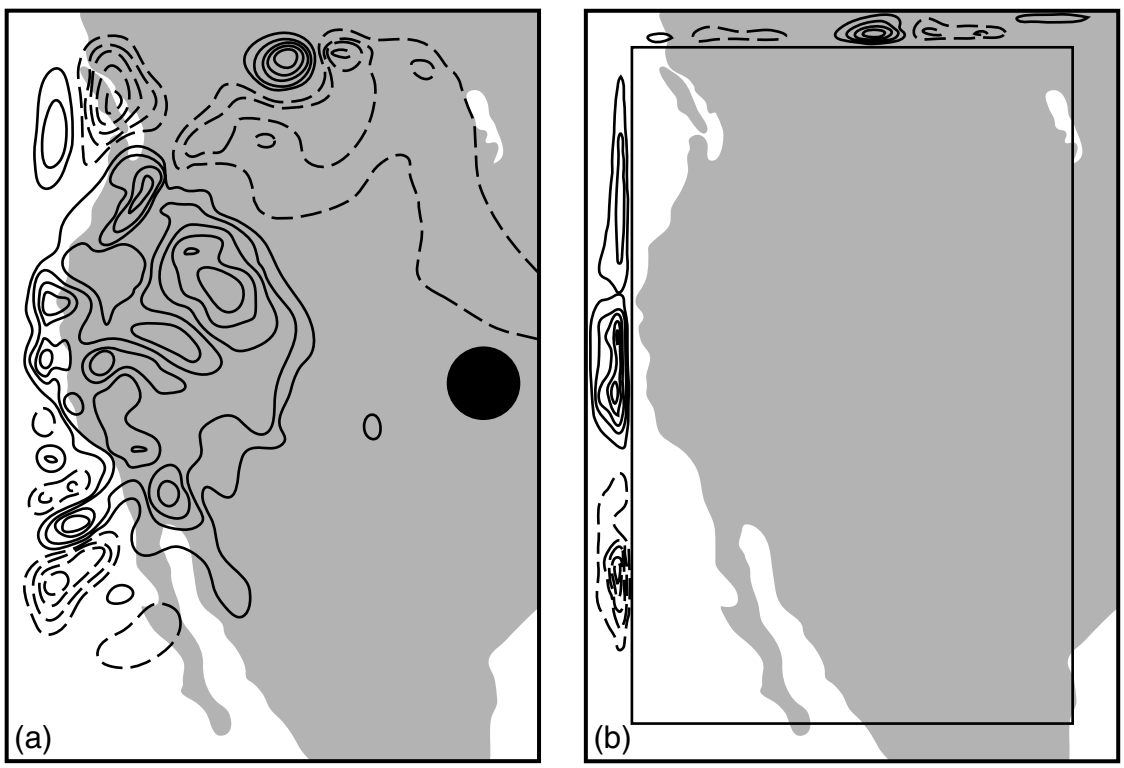

Fig. 3.47 Sensitivity of the 72-h relative vorticity in a limited volume in the center of the domain (black circle, panel a) to perturbations of the 400-hPa *v*-component of the wind on (a) the domain interior and (b) the lateral boundaries at the initial time for the winter case. For panel "a" ("b"), the maximum absolute value is 1.4 units (8 units), and the isopleth interval is 0.25 unit (1 unit). Only the western one-half of the computational domain is shown. Adapted from Errico *et al.* (1993).

Table 3.2 Maximum values of the metric of the sensitivity of the 72-h relative vorticity near the center of the domain to the 400-hPa *v*-wind component on the lateral boundaries and on the domain interior. Values are shown pertaining to the sensitivity of the 72-h vorticity to the *v*-component perturbations at four times during the simulation.

	Simulation time (h)			
	0	24	48	60
Lateral boundary sensitivity	8	40	150	52
Interior sensitivity	1.4	18	76	93

Source: From Warner *et al.* (1997).

the west and north. Table 3.2 summarizes the maximum value of the sensitivity metric on the domain interior and on the lateral boundaries at four times during the simulation, and indicates that, as expected, the sensitivity of the 72-h vorticity to conditions on the domain interior is less for early times of the simulation. That is, the 72-h vorticity simulation tends to "forget" the impact of the perturbations as these conditions become more temporally removed. In terms of the effect on the 72-h simulation, the 48-h LBCs are more important than those at other times because the 24-h difference (between 48 h and 72 h) is the time required for the LBC signal to propagate to the center of the domain at this level. It is interesting that the 72-h forecast is less sensitive to initial condition ($t = 0$) perturbations (1.4 units) than it is to LBC perturbations at any time (8–150 units). The results for lower levels in the model (i.e., perturbations below 400 hPa) with weaker winds are qualitatively similar except that it naturally requires more time for LBC effects to penetrate to the center of the domain. For the summer case, the weaker wind speeds cause a factor-of-two slower propagation of the sensitivity.

Big-Brother–Little-Brother experiments

In these experiments, a high-resolution model whose grid spans a large area is used to generate a reference simulation. This is the *Big-Brother simulation*. Then, using the identical model, another simulation is performed for a sub-area within the reference-simulation's grid. Lateral-boundary conditions are provided based on a data set that results from filtering all but the larger scales from the Big-Brother solution. This is the *Little-Brother simulation*. Because the experimental conditions in the two simulations are exactly the same, except for the presence of the LBCs in the Little-Brother simulation, differencing the two model solutions over the area of the smaller grid isolates the effect of the LBCs. See the discussion in Section 10.4 for additional information and references about this type of experiment. An example illustrating LBC effects that have been isolated using this method is shown in Fig. 3.48. Shown is the computational-domain-averaged precipitation rate for the area of the small grid, based on both the Big-Brother and Little-Brother simulations. The Canadian Regional Climate Model (Caya and Laprise 1999) was employed here in a test of regional climate modeling methods. The existence of the LBCs in this case had very little effect on the average precipitation rate.

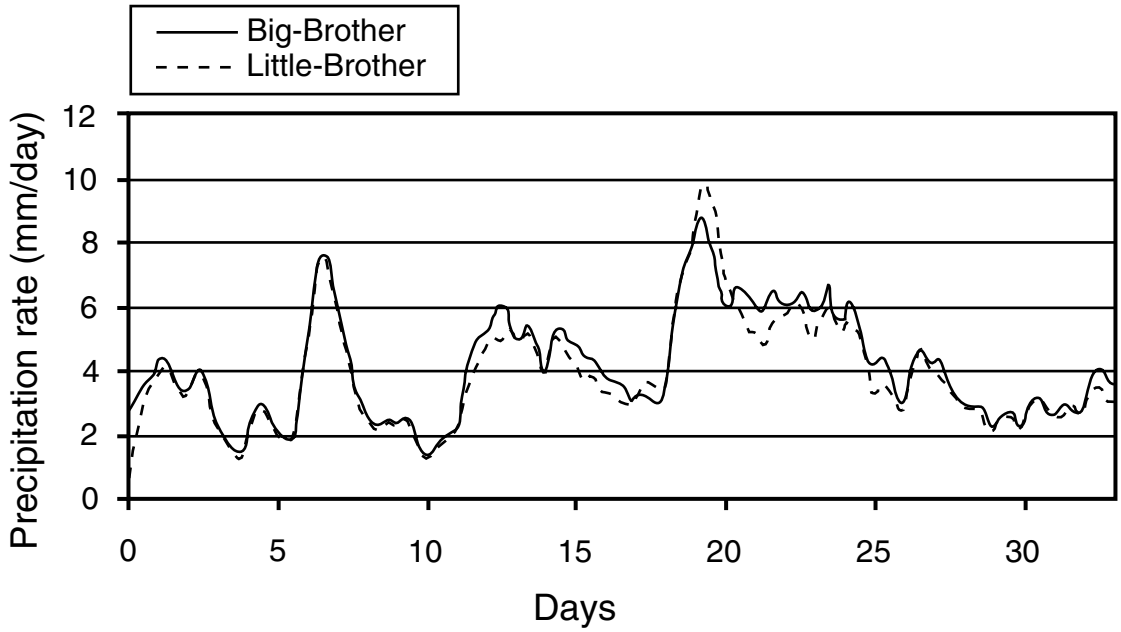

Fig. 3.48 Spatially averaged precipitation rate for the area of the small grid in a Big-Brother–Little-Brother-experiment, for both the Big-Brother and Little-Brother simulations. Adapted from Denis *et al.* (2002).

3.5.3 Types of LBC formulations

Open, or free, LBCs allow values of variables to be externally specified based on forecasts from a larger model grid (e.g., a global model), or from gridded analyses of data. There are two approaches for defining LBCs from a coarser-resolution grid. One involves the simultaneous integration of the LAM and a coarser-mesh model within which it is embedded, where the information flow between the domains is in both directions. See Harrison and Elsberry (1972), Phillips and Shukla (1973), and Staniforth and Mitchell (1978) for a discussion of such techniques. In the other approach, LBCs are prescribed based on the output from a previous integration of a coarser-mesh model or an analysis of observations. The development of these techniques is described in Shapiro and O'Brien (1970), Asselin (1972), Kesel and Winninghoff (1972), and Anthes (1974). The first approach is called *two-way interactive nesting*, and the latter is called *one-way*, or *parasitic*, *nesting*. In both cases, meteorological information from the coarser-mesh domain must be able to enter the fine-mesh domain, and inertia–gravity and other waves must be able to freely exit the fine-mesh domain. With the two-way interacting boundary conditions, the information from the fine mesh can affect the solution on the coarse mesh, which can feed back to the fine mesh. An example of the desirability of this approach is provided in Perkey and Maddox (1985), who use numerical experiments to show that a convective-precipitation system can influence its large-scale environment, which can then feed back to the mesoscale. Note that LAMs that employ a two-way interacting nested grid system must generally obtain LBCs for their coarsest-resolution domain from a previously run global model or from

analyses of observations. Thus, whether or not a two-way interacting nesting strategy is employed, the use of a one-way interacting interface condition is almost always necessary.

For the interface condition between domains of a two-way interacting nest, a variety of approaches are successfully used for interpolating the coarser-grid solution to the finer grid, and for filtering the finer-grid solution that is fed back to the coarser grid (Clark and Farley 1984, Zhang *et al.* 1986, Clark and Hall 1991). For one-way interacting grids, techniques are common that filter or damp small scales in the fine-mesh solution near the boundary (Perkey and Kreitzberg 1976, Kar and Turco 1995). For example, in the Perkey and Kreitzberg (1976) approach, a wave-absorbing or sponge zone near the lateral boundary prevents internal reflection of outward-propagating waves through an enhanced diffusion as well as truncation of the time derivatives. In these approaches, the fine grid is forced with large-scale conditions through a relaxation or diffusion term (Davies 1976, 1983, Davies and Turner 1977).

It is intuitive that two-way interactive nesting should provide for better model solutions on the finest grid than does one-way, parasitic nesting, simply because upscale effects can feed back to the fine mesh. This has, in fact, been demonstrated, for example by Clark and Farley (1984) for forced gravity-wave flow, and as noted earlier by Perkey and Maddox (1985) for convection. However, there are sometimes practical reasons for one-way nesting. For example, for operational nested modeling systems, one-way nesting allows the coarse-grid forecast to be completed first, and the products made available quickly to forecasters while the more computationally intensive calculations are taking place for the finer grids. And, in situations where significant computer-memory limitations exist, it is sometimes essential to limit calculations to one grid at a time.

For simple research or educational models, *periodic*, or *cyclic*, LBCs may be employed. Here, the grid points near one edge of the domain are coupled with those near the opposite edge, so that features that exit at one boundary enter at the other. This is illustrated in Fig. 3.49 for a three-point horizontal differencing scheme applied on a one-dimensional grid. At each time step, after the extrapolation in time is performed for grid points 2 through j_{max-1}, the values of the variables at the penultimate points are used to redefine the values at the corresponding edge points. For a model that uses five-point horizontal

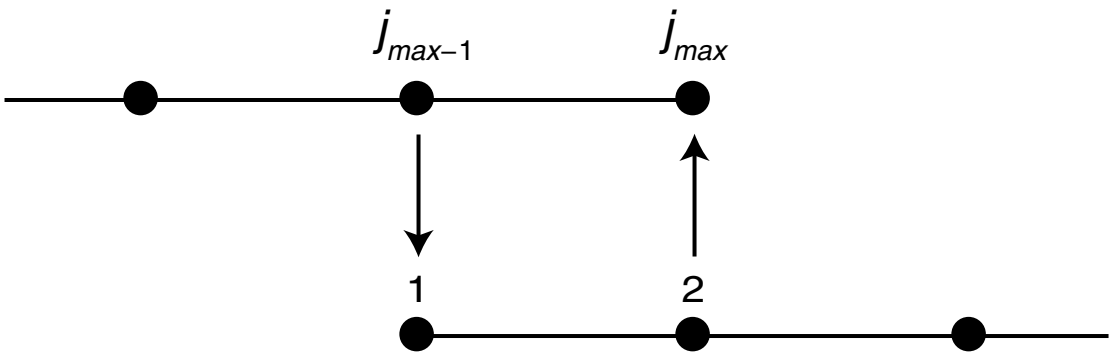

Fig. 3.49 Transfer of information between grid points at the edges of a computational grid that employs cyclic, or periodic, LBCs.

differencing, one more overlap point would be required near each edge. In a model with two horizontal dimensions, LBCs can be periodic in both directions. Or, a periodic LBC can be employed in only one direction, and an impervious wall boundary condition can be assumed on the other two edges, in which case the model is called a channel model.

3.5.4 Some practical recommendations

The studies described in the previous section, and others, are synthesized into the following recommendations for how LBC effects can be minimized in any LAM application.

(1) Utilize a lateral-boundary *buffer zone*

The LBC errors that reach the central part of a LAM grid are sometimes so large as to render the LAM forecast to be of no more value than that of the coarser-mesh model that produces the LBCs. In this situation, if enough computational resources are available, the lateral boundaries can be distanced from the central part of the LAM grid so that LBC errors do not penetrate to this region during a forecast with the desired duration. Alternatively, a standard domain area can be employed and the forecast duration can be limited so as to prevent penetration of the LBC errors into the central area of meteorological interest. Table 3.3 illustrates the domain sizes and forecast-duration limits that are necessary in order to minimize LBC impacts for different meteorological regimes and forecast-area length scales.

To illustrate the implications of the need for an LBC "buffer zone", a typical LAM configuration will be assumed, and the useful length of the forecast will be calculated. It will be assumed that the lateral boundaries are removed in each direction from the area of meteorological interest (having length scale L) by a distance equal to one-half L. For example, if the computational domain has 100 grid points in each direction, the inner protected area of meteorological interest on the model domain is represented by the central subset of 50×50 points. Most modelers would agree that this is a reasonable compromise, even though there are three times as many computational points in the buffer-zone region outside the area of interest than there are in it. This seemingly large computational "overhead" is generally accepted as unavoidable. The useful period of the forecast is defined here as the time required for LBC influences to advect to the central forecast area.[5] Table 3.3 shows the useful forecast periods (entry a) for four different computational areas with different scales (rows), and for four different meteorological regimes (columns). Average midtropospheric wind speeds (S in Table 3.3) are used in the advection-time calculation for midlatitude winter and summer regimes, and for the tropical regime. For the midlatitude-uncoupled regime, it is assumed that there is weak vertical coupling and that the dominant meteorological processes are forced by lower-tropospheric effects. The smallest domain has the size of a large city (row 1, metropolitan area), the next larger one spans an area equivalent to the coverage of a typical weather radar (row 2, radar-range area), the next larger one covers about a

[5] For simplicity, it is assumed here that the advective speed represents the speed with which LBC error penetrates inward on the LAM domain. However, LBC errors may be propagated by nonadvective waves such as inertia–gravity or Rossby waves.

Table 3.3 For four different computational areas (rows) and four different meteorological regimes (columns): [a]useful duration of forecasts for a standard domain; [b]width of buffer zone required (in units of L, defined in column 2) for forecasts of "standard" duration (defined in column 3); and [c]ratio of buffer-zone grid points to central forecast-area grid points for forecasts of "standard" duration.

			Meteorological regimes			
Forecast domain size	Interior forecast-area length scale (L)	"Standard" forecast duration	Winter mid lat $S = 30$ m s^{-1} (~60 kt)	Summer mid lat $S = 15$ m s^{-1} (~30 kt)	Tropical $S = 8$ m s^{-1} (~15 kt)	Mid lat uncoupled $S = 5$ m s^{-1} (~10 kt)
Metropolitan area	50 km	6 h	[a]14 min [b]13.0 L [c]724	[a]28 min [b]6.5 L [c]194	[a]52 min [b]3.5 L [c]63	[a]1.4 h [b]2.2 L [c]27
Radar-range area	500 km	18 h	[a]2.3 h [b]3.9 L [c]76	[a]4.6 h [b]1.9 L [c]23	[a]8.7 h [b]1.0 L [c]8	[a]13.9 h [b]0.6 L [c]4
Regional area	2000 km	36 h	[a]9.3 h [b]1.9 L [c]23	[a]18.5 h [b]1.0 L [c]8	[a]34.7 h [b]0.5 L [c]3	[a]55.6 h [b]0.3 L [c]1.7
Continental area	5000 km	72 h	[a]23.1 h [b]1.6 L [c]16	[a]46.3 h [b]0.8 L [c]6	[a]86.8 h [b]0.4 L [c]2	[a]138.9 h [b]0.3 L [c]1.3

Source: From Warner *et al.* (1997).

quarter of a typical continent (row 3, regional area), and the largest one covers an entire continent (row 4, continental area). For the metropolitan-area domain, the forecast is hardly more than a "nowcast", regardless of the regime (entry a, useful forecast length). The radar-range and regional domains are of a scale that might be appropriate for regional weather prediction for small to moderate size countries, but unless they are in the tropics the forecast period is generally limited to considerably less than one day. Only for continental domains can useful forecasts have durations beyond a day.

Also shown in Table 3.3 is the lateral boundary displacement (entry b), in units of L, required to produce a forecast of "standard" duration (column 3) without LBC-error penetration to the domain interior. In addition, for each of these extended domains is computed the ratio of the number of buffer-zone grid points to the number of interior forecast-area grid points (entry c), which serves as a metric of the computational overhead resulting from the need for a buffer zone. If the buffer-zone width is increased for the small domains to allow for forecasts with a longer, more operationally useful, duration, the computational overhead generally becomes quite large. For example, to obtain a 6-h forecast in winter with the metropolitan area domain could require an overhead factor of between 500 and 1000. Often it is possible to take advantage of an asymmetry in the speed/direction

climatology of the prevailing advecting wind, and increase the width of the buffer zone in the direction of stronger prevailing flow. Using available computational resources wisely by asymmetrically protecting the domain interior is recommended, but this will likely only permit an increase in the useful duration of the forecast by less than 50% compared to the use of a symmetric buffer zone with the same number of grid points. It has been implied that the LBC error is sufficiently large that it overwhelms the forecast accuracy when the error penetrates to the domain interior. However, there are measures that can be taken to control the amplitude of the LBC errors, and some current LBC formulations may not be especially damaging to the model solution.

(2) Minimize interpolation error with the lateral-boundary data

The actual magnitudes of LBC errors will depend on a number of factors including the quality of the coarse-mesh forecast that is producing the LBCs and the magnitude of the error associated with the spatial and temporal interpolation from the coarse mesh to the LAM domain at the lateral boundaries. The interpolation error can be reduced through the frequent passing of LBC information from the coarse-mesh model to the LAM. For example, passage of a fast-moving mesocyclone through the boundary may be missed entirely if LBC data are updated only every six hours.

(3) Use compatible numerics and physics with the LAM and the model providing the LBCs

The use of reasonably consistent physical-process parameterizations (convection, cloud microphysics, turbulence, and radiation) on the two grids will minimize unrealistic gradients that can develop at the interface and propagate onto the LAM domain through advection and inertia–gravity waves. For example, Warner and Hsu (2000) show how parameterized convection on an outer grid can strongly influence resolved convection on an inner grid through LBC-forced mass-field adjustments.

(4) Employ well-tested and effective LBC formulations

Many LBC formulations for meteorological models are inherently ill-specified mathematically, and thus engineering approaches have been devised to minimize the potentially serious numerical problems that can develop. The LBC formulation used should be sufficiently well tested and designed so that it does not generate significant-amplitude, inertia–gravity waves that can move toward the central area of the domain at much greater than advective speeds. Even though some of the examples presented earlier demonstrate that this error can be significant, the use of appropriately engineered LBC algorithms can often limit the amplitude of this mode of error propagation to acceptable levels.

(5) Allow for effects of data assimilation on LBC impact

The use of a preforecast FDDA period can have both a positive and negative effect on the LBC influence, whether continuous or intermittent assimilation techniques are utilized. On the one hand, the preforecast integration period will allow LBC error to propagate

closer to the domain center by the start of the forecast. Conversely, the data assimilated during the period will partially correct for errors of LBC origin that are within the influence region of the observations.

(6) Account for importance of local forcing

If strong local forcing mechanisms prevail within the fine mesh, and dominate the local meteorology, the forecast quality may not be as strongly affected by LBC errors as it would otherwise be. For example, the time of onset of a coastal-breeze circulation is more strongly correlated with local thermodynamic effects than with specific characteristics of the large-scale flow field and its LBC-related errors.

(7) Avoid strong forcing at the lateral boundaries

Strong dynamic forcing at the lateral boundaries can create numerical problems with many LBC formulations. Even though it is not possible to avoid the passage of transient high-amplitude meteorological phenomena through the boundaries, it is possible to avoid collocating lateral boundaries with known regions of strong surface forcing such as associated with steep orography and surface temperature gradients. Locating large terrain gradients near or at lateral boundaries is one of the most common ways in which LBCs can cause the catastrophic failure of a model integration.

(8) Utilize interactive grid nests when possible

When a LAM cannot influence the solution of the coarser-mesh model that provides its boundary values, the scale interactions of the LAM-resolved waves and those on the large scale are prevented. In addition, the use of a two-way interactive interface can, but will not necessarily, reduce the development of spurious gradients at the boundaries. Thus, interactive boundaries should be employed where possible, rather than one-way-specified boundaries.

(9) With any new model application, perform sensitivity studies to determine the LBC influences

After considering the experiences described in the last section, it should be clear that LBC sensitivity studies should be performed for any new application of a LAM, especially if the aforementioned recommendations regarding the buffer-zone width are not taken literally. These sensitivity studies should include the testing of the dependence of forecast accuracy on buffer-zone width, the sensitivity of the forecast quality to different LBC formulations, and a comparison of the LAM skill to that of other operational modeling systems that have unbounded domains. A practical test for any LAM application is to compare the solution over the limited area with that from a model with equivalent resolution integrated over a much larger domain (Yakimiw and Robert 1990). If the LAM is to be used operationally, the forecasts naturally should be evaluated for LBC sensitivity over a wide range of events within all seasons.

3.6 Upper-boundary conditions

Artificial upper-boundary conditions are required in all atmospheric models because the model atmospheres do not extend to infinity. Indeed, for some historical applications the upper boundary has been located within the troposphere in order to save computational resources. An example of this approach is that Lavoie (1972) placed the upper model boundary, the "lid", at the top of the boundary layer. Pielke (2002a) describes the location of the upper boundary in various historical model applications.

Upward-propagating internal-gravity waves, for example generated by mountains or by deep convective storms, can extend to great heights in the atmosphere. Commonly used upper-boundary conditions (e.g., rigid lid, free surface) completely reflect these waves, which is a problem because no such reflection happens in nature, and erroneous downward-propagating waves contaminate the model solution. There are a number of approaches for preventing this from happening. One involves the use in the model of a gravity-wave absorbing layer, or sponge layer, immediately below the model top, to prevent the wave from reaching the top and reflecting. Such wave absorption can be produced by employing a greatly enhanced, artificial horizontal and/or vertical diffusion (viscosity), where the viscosity increases from the standard value at the bottom of the layer to a maximum at the top boundary. A particular disadvantage of this approach is that the absorbing layer may need to be thick, spanning a large number of model layers and thus involving a large computational overhead. The overall effectiveness of the absorption depends on the wavelength of the gravity wave, the thickness of the absorbing layer, and the distribution of viscosity in the layer. Note that using a shallow absorbing layer with a very large, but computationally stable, viscosity will not be effective because large gradients in viscosity will also cause wave reflections. Klemp and Lilly (1978) defined the entire upper half of their computational domain as the absorbing layer. Figure 3.50 shows a two-dimensional model solution for idealized flow over a maximum in the orography, with and without the use of a viscous damping layer. The Gaussian obstacle had a 5-km half-width, and an amplitude of 1 km. Shown is the vertical motion field in the lowest 10 km of the 50-km-deep model. The model is described in Sharman and Wurtele (1983). The damping spanned the 20 km below a rigid lid that defined the model top. Without the damping, the reflected waves produce considerable noise in the troposphere, over 40 km below the model top. The waves in the solution for the experiment with the absorbing layer could be a result of imperfect damping, or more likely they could be a consequence of wave reflections from the lateral boundaries. An alternative approach for damping the waves before they reach the upper boundary is to use a Raleigh damping layer, again below the model top, where model variables are relaxed toward a predetermined reference state. For example, the Rayleigh damping term in a prognostic equation would be like

$$\frac{\partial \alpha}{\partial t} = \tau(z)(\alpha - \bar{\alpha}),$$

where α is any dependent variable, $\bar{\alpha}$ is the reference value of that variable, and $\tau(z)$ increases upward within the damping layer and defines its vertical structure (e.g., see

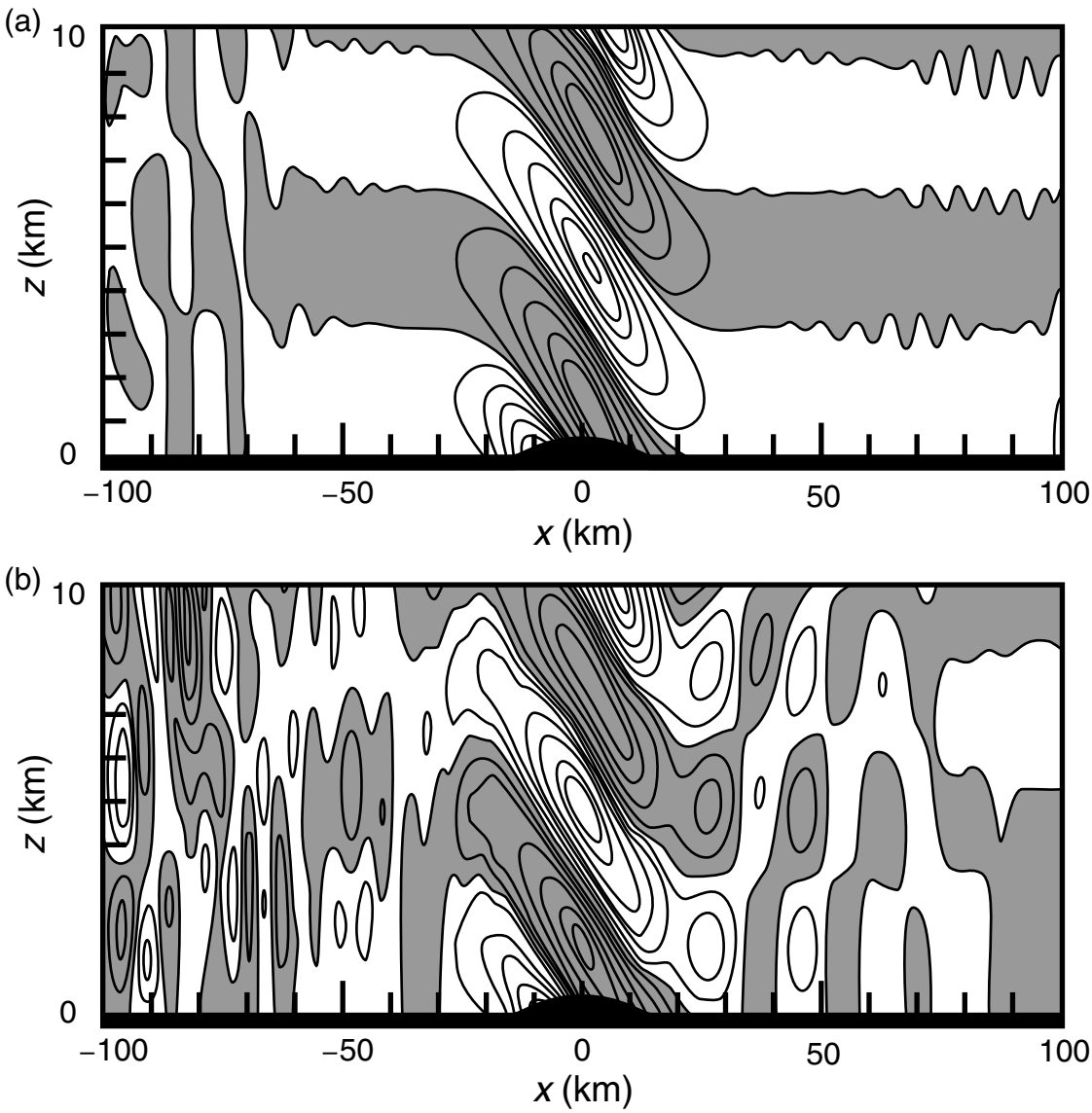

Fig. 3.50 Vertical motion over the lowest 10 km of the atmosphere from two-dimensional model simulations of flow over elevated orography (see the black shading at the bottom), with a viscous layer at the top of the model (a) and without a viscous layer (b). The flow was from left to right, and downward motion is shaded. The damping layer spanned the upper 20 km of the 50-km-deep model. Provided by Robert Sharman.

Durran and Klemp 1983). Israeli and Orzag (1981) compare the Rayleigh-damping and viscous-damping methods.

A completely different approach, which does not rely on an absorbing layer, involves the use of a radiation boundary condition. Here, values of variables at the boundary are modified during the integration to minimize wave reflection. Clearly the term radiation

refers to the fact that waves are intended to radiate through the boundary, and not reflect from it. These approaches are discussed in Durran (1999) and Klemp and Durran (1983), and compared with the sponge approaches in Israeli and Orzag (1981).

3.7 Conservation issues

The various numerical approaches used in atmospheric models possess inherent properties that determine the extent to which they conserve mass, energy, and other quantities. Even though we might like to see a model have the same conservation properties as the continuous equations and the real atmosphere, there are many factors that enter into the choice of numeric methods, such as the inherent damping of small-scale energy, the correct rendering of phase speeds, and numerical efficiency. That said, systematic leaks in mass or energy that are manifested as slow artificial drifts in the model mean state may be tolerable for short-term forecasts, but definitely would not be for integrations on climate time scales. Thus, serious consideration needs to be given to the degree to which spurious sources and sinks of physical quantities are acceptable for a particular model application. Thuburn (2008) contains a summary of conservation issues for weather-prediction and climate models, and suggests that we can expect accurate solutions from models provided that the time scale for artificial numerical sources is long compared to the time scale for the true physical sources.

The conservation of mass is arguably the most absolute conservation property, given that true physical sources are irrelevant. And, unlike other quantities, mass is conserved for diabatic and frictional processes. If mass is not conserved, it affects the surface pressure distribution, and in turn the circulations. Sometimes when models do not conserve mass, a nonphysical, so-called mass-fixer is used each time step to correct for changes in the total global mass, but where the mass is added or removed in the correction is arbitrary. Furthermore, if total mass is not conserved, neither are the various constituents such as water vapor or long-lived chemical species. Thuburn (2008) argues that, at least for long climate simulations, there is a very strong argument for requiring dynamical cores to conserve total mass, and therefore the mass of constituents. He also discusses the situations in which it is important to conserve momentum, angular momentum, potential enstrophy, energy (kinetic, and available and unavailable potential), entropy, and potential vorticity.

3.8 Practical summary of the process for setting up a model

This section is meant to summarize how our knowledge of the numerical processes discussed in this chapter should serve as a guide for setting up a model. There are additional factors that must be considered, such as the appropriateness of physical-process parameterizations, but these are reviewed in other chapters. It is assumed in the following that the

time step is internally determined by the model, and that there is no choice in the methods used to solve the equations (e.g., spectral versus grid-point approaches, explicit versus semi-implicit time differencing, etc.).

- Based on a knowledge of the purpose for using the model and the prevailing meteorology in the geographic area to be modeled, determine the physical processes that must be simulated or forecasted.
- Choose a horizontal grid increment that is sufficiently small to resolve all the processes to be represented on the grid.
- Define a vertical distribution of grid points that adequately defines anticipated important vertical structures (e.g., boundary-layer gradients, low-level jets, the tropopause) and, if possible, ensure reasonable compatibility of the vertical grid increment with the horizontal increment.
- For limited-area models, choose the map projection that is most suitable for the range of latitudes represented by the model grid. View a graphic of the map-scale factor at each grid point when setting up the model grid to confirm the degree to which it departs from unity.
- Compare the model solution with observations, and quantify the skill. If the model is to be used as a general research or operational-forecasting tool, numerous cases should be chosen from all seasons. Just because the model has been reported to be accurate for other locations and configurations, do not assume that this step can be avoided.
- For limited-area models, perform tests to define the sensitivity of the accuracy of the model solution to different locations of the boundaries and different domain sizes.
- Perform tests to determine the sensitivity of the model accuracy to the vertical and horizontal grid increments.

Section 10.1 provides additional practical guidance for applying models to perform research case studies.

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

Durran, D. R. (1999). *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*. New York, USA: Springer.

Krishnamurti, T. N., H. S. Bedi, V. M. Hardiker, and L. Ramaswamy (2006). *An Introduction to Global Spectral Modeling*. New York, USA: Springer.

Staniforth, A., and N. Wood (2008). Aspects of the dynamical core of a nonhydrostatic deep-atmosphere, unified weather and climate-prediction model. *J. Comput. Phys*., **227**, 3445–3464.

Williamson, D. L. (2007). The evolution of dynamical cores for global atmospheric models. *J. Meteor. Soc. Japan*, **85B**, 241–269.

World Meteorological Organization (1979). *Numerical Methods Used in Atmospheric Models*, *Volume II*. Global Atmospheric Research Programme, GARP Publication Series No. 17. Geneva, Switzerland: World Meteorological Organization.

PROBLEMS AND EXERCISES

1. For the 24-point grid referenced in Fig. 3.32, list all the combinations of interacting wavenumbers that produce aliasing, and the erroneous wavenumbers that results from the interactions.
2. Derive an expression for the ratio of the five-point numerical approximation to the derivative and the analytic solution for the derivative, analogous to what is shown for the three-point approximation in Section 3.4.1.
3. Explain graphically, or in words, why the truncation error for the forward-in-space differencing formula in Eq. 3.25 is dependent on position within the wave, in addition to how well the wave is resolved on the grid.
4. Given that Fig. 3.27 shows that the use of Courant numbers close to unity produces more-realistic solutions than do smaller values, for the centered-in-space and centered-in-time approximation to the advection term, why can't we use sufficiently large time steps to ensure the prevalence of these large Courant numbers?
5. Prove the orthogonality of the exponential function.
6. Using the programming language of your choice, construct a one-dimensional model based on the shallow-fluid equations (Chapter 2) with three-point time and space differencing and no explicit diffusion. Assume periodic lateral-boundary conditions, and perform the following experiments.
   - Simulate an advective wave and a gravity wave.
   - Choose a time step that violates the linear stability criterion, and output the model solution each time step.
   - Add an explicit diffusion term and show its effect on the model solution for different diffusion coefficients.
   - Alter the time step to evaluate how the use of different Courant numbers affects the model solution.
   - For the same initial conditions, evaluate the effect of horizontal resolution on the model solution.
7. Some studies with LAMs (e.g., Alpert *et al.* 1996) suggest that the quality of simulations decreases as lateral boundaries become too distant from or too close to the area of meteorological interest. Provide possible explanations for this situation.
8. Given the typical spacing of radiosonde soundings (~400 km), explain the types of meteorological processes that can be adequately resolved by them in the initial conditions.
9. Regarding Fig. 3.44, explain why the jet streak simulated on the small domain is so much smoother than the one simulated on the larger domain.