An Introduction to Numerical Modeling of the Atmosphere

David A. Randall

Contents

Preface

Numerical modeling is one of four broadly defined approaches to the study of the atmosphere. The others are observational studies of the real atmosphere through field measurements and remote sensing, laboratory studies, and theoretical studies. Each of these approaches has both strengths and weaknesses. In particular, both numerical modeling and theory involve approximations. In theoretical work, the approximations often involve extreme idealizations, e.g., a dry atmosphere on a beta plane, but on the other hand solutions can sometimes be obtained in closed form with a pencil and paper. In numerical modeling, less idealization is needed, but no closed form solution is possible. In most cases, numerical solutions represent particular cases, as opposed to general relationships. Both theoreticians and numerical modelers make mistakes, from time to time, so both types of work are subject to errors in the old-fashioned human sense.

Theory actually plays two roles in numerical modeling. First, we use theories to design both the numerical methods and the physical parameterizations of models. Second, we use theories to interpret why the numerical results turned out as they did.

Perhaps the most serious weakness of numerical modeling, as a research approach, is that it is possible to run a numerical model built by someone else without having the foggiest idea how the model works or what its limitations are. Unfortunately, this kind of thing happens all the time, and the problem is becoming more serious in this era of "community" models with large user groups. One of the purposes of this book is to make it less likely that you, the readers, will use a model without having any understanding of how it works.

This introductory survey of numerical methods in the atmospheric sciences is designed to be a practical, "how-to" course, which also conveys sufficient understanding so that after completing the course students are able to design numerical schemes with useful properties, and to understand the properties of schemes designed by others.

We will not discuss parameterizations. That is a different subject, addressed elsewhere in the curriculum, although the actual implementation of parameterizations always involves some numerical methods, and this book can help with that part of the job.

We will not describe particular models in detail, although some mare mentioned briefly.

This book is based on my class notes. The first version of the notes, put together in 1991, was heavily based on the class notes developed by Prof. Akio Arakawa of UCLA, as they existed when I took his course in the early 1970s. Arakawa's influence is still apparent throughout the book.

The Teaching Assistants for the course have made major improvements in the material and its presentation, in addition to their help with the homework and with questions outside of class. I have learned a lot from them, and also through questions and feedback from the students.

Michelle Beckman, Amy Dykstra, and Val Hisam spent countless hours patiently assisting in the production of various versions of these notes. I am especially indebted to Claire Peters, who converted the whole book to LaTeX.

Chapter 1

What this book is about

1.1 True stories

The atmospheric science community includes a large and energetic group of researchers who devise and carry out measurements of the atmosphere. They do instrument development, algorithm development, data collection, data reduction, and data analysis.

In order to make physical sense of the data, some sort of model is needed. It might be a qualitative conceptual model, or an analytical theory, or a numerical model. Models of all three types provide a basis for understanding data, and also for making predictions about what the atmosphere will do.

Accordingly, a community of modelers is hard at work creating models, performing simulations, and analyzing the results, in part by comparison with observations. The models by themselves are just "stories" about the atmosphere. In making up these stories, however, modelers must strive to satisfy a very special and rather daunting requirement: The stories must be true, as far as we can tell; in other words, the models must be consistent with all of the relevant measurements.

1.2 Prognosis and diagnosis

Most models in atmospheric science are formulated by starting from basic physical principles, such as conservation of mass, conservation of momentum, and conservation of thermodynamic energy. Many of these equations are *prognostic*, which means that they involve time derivatives. A simple example is the continuity equation, which expresses conservation of mass:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{V}) \,. \tag{1.1}$$

Here t is time, ρ is the density of dry air, and $\mathbf{V}$ is the three-dimensional velocity vector. Prognostic variables are governed by prognostic equations. Eq. (1.1) is a prognostic equation, in which ρ is a prognostic variable. A model that uses prognostic equations is solved, in part, by time integration. Initial conditions are needed for the prognostic variables.

Any variable that is not prognostic can be called *diagnostic*. Equations that do not contain time derivatives are called diagnostic equations. The diagnostic variables of a model can be computed from the prognostic variables and the external parameters that are imposed on the model, e.g., the radius and rotation rate of the Earth. In this sense, *the prognostic variables are primary*, and the diagnostic variables are secondary.

1.3 Elementary models

The sub-disciplines of atmospheric science, e.g., geophysical fluid dynamics, radiative transfer, atmospheric chemistry, and cloud microphysics all make use of models that are essentially direct applications of physical principles to phenomena that occur in the atmosphere. These models are "elementary" in the sense that they form the conceptual foundation for other modeling work. Many elementary models were developed under the banners of physics and chemistry, but some – enough that we can be proud – are products of the atmospheric science community. Elementary models tend to deal with microscale phenomena, (e.g., the movement of a microscopic fluid particle, or the evolution of individual cloud droplets, or the optical properties of individual ice crystals) so that their direct application to practical atmospheric problems is usually thwarted by the sheer size and complexity of the atmosphere.

Analytical models produce results that consist of equations. As a simple example, consider the ideal gas law

$$p = \rho R T \ . \tag{1.2}$$

Here p is pressure, R is the gas constant, and T is temperature. Eq. (1.2) can be derived using the kinetic theory of gases, which is an analytical model; the ideal gas law can be considered as a "result" of the analytical model. Eq. (1.2) can be used to generate numbers, of course; for example, given the density and temperature of the air, and the gas constant, we can use Eq. (1.2) to compute the pressure. The ideal gas law summarizes relationships that hold over a wide range of conditions. It is sufficiently simple that we can understand what it means just by looking at it.

1.4 Numerical models

The results of a numerical model consist of (yes) numbers, which represent particular "cases" or "realizations." A realization is an particular instance of what the (model) atmosphere can do. For example, we can "run" a numerical model to create a weather forecast, which consists of a large set of numbers. To perform a new forecast, starting from a different initial condition, we have to run the model again, generating a new set of numbers. In order to see everything that the (model) atmosphere can do, we would have to run infinitely many cases. In this way, numerical models are quite different from analytical models, which can describe all possibilities in a single formula. We cannot understand what a numerical model can simulate just by looking at the computer code.

This distinction between numerical and analytical models is not cut and dried, however. Sometimes the solutions of analytical models are so complicated that we cannot understand what they mean. Then it is necessary to plot particular examples in order to gain some understanding of what the model is telling us. In such cases, the analytical solution is useful in more or less the same way that a numerical solution would be. The other side of the coin is that, in rare cases, the solution of a numerical model can represent all possibilities in form of a single (numerically generated) table or plot.

1.5 Errors

Models entail errors. All your life you have been making errors. Now, finally, you get to read a book about errors.

It is useful to distinguish between physical errors and mathematical errors. Suppose that we start from a set of equations that describes a physical phenomenon "exactly." For example, we often consider the Navier-Stokes equations to be an exact description of the fluid dynamics of air.[1] For various reasons, we are unable to obtain exact solutions to the Navier-Stokes equations (except in trivial cases). To simplify the problem, we introduce physical approximations. For example, we may use approximate physical parameterizations that can be used to determine turbulent and convective fluxes. To ensure that the models are realistic, we must rely on physical understanding of the relative importance of the various physical processes and the statistical interactions of subgrid-scale and grid-scale motions, and of course we must compare the model's formulation and its results with observations in as many ways as possible. For these reasons, the design of atmospheric models can never be a purely mathematical problem.

Models are *designed*. Creating a design is all about making choices. This book explains what choices are possible, and the advantages and disadvantages of each.

Beyond simplification, a second motivation for making physical approximations is that

[1] In reality, of course, the Navier-Stokes equations already involve physical approximations.

the approximate equations may describe the phenomena of interest more directly, omitting or "filtering" phenomena of less interest, and so yielding a set of equations that is more focused on, and more appropriate for, the problem at hand. For example, we may choose approximations that filter sound waves (e.g., the quasi-static approximation or the anelastic approximation) or even gravity waves (e.g., the quasigeostrophic approximation). These physical approximations introduce physical errors, which may or may not be considered acceptable, depending on the intended application of the model. In practice, *choosing* a set of physical approximations means choosing the equation system that is used to formulate the model.

After choosing the equation system, there there are still multiple ways to write the set of equations actually used in a model. Most atmospheric models include equations that involve time derivatives. As mentioned earlier, these are called prognostic equations. The design of a model entails *choosing* which variables to prognose. As a simple example, we could choose to predict either temperature or potential temperature. In a continuous model, exactly the same results would be obtained either way, but in a discrete model the results will almost certainly differ depending on which variable is chosen. Much more discussion of the choice of prognostic variables is given later.

Another source of error is the finite precision of the computer hardware. This *round-off error* is a property of the machine that the model is run on (and to some extent the details of how the model is coded). Round-off errors sometimes cause problems, but usually they don't, and we will not worry about them in this book.

Once we have settled on a suitable set of physical equations and prognostic variables, we must choose or invent mathematical methods to solve them. The mathematical methods are almost always approximate, which is another way of saying that they introduce errors.

This book is not a survey of numerical models. Instead, it focuses on the theory underlying formulation of numerical models of the atmosphere, and especially the mathematical errors that the models produce. We discuss where the mathematical errors come from, and how they can be identified, analyzed, and minimized. We also discuss some aspects of the physical approximations associated with particular equation sets.

1.6 The role of theory in numerical modeling

Theory enters the numerical modeling enterprise in two different ways. First, the design of atmospheric numerical models is guided by theories, from both atmospheric science and applied mathematics. Much of this book is devoted to these theories. Second, theory is used to interpret the results produced by numerical models. The results produced by a numerical model are just numbers. Theory is needed to draw physical conclusions from the results.

1.7 Discretization

Numerical models are "discrete." This simply means that they deal with a finite number of numbers. The process of approximating a continuous model by a discrete model is called "discretization."

There are multiple approaches to discretization, which are often classified by the methods used to derive them rather than by the equations that result from the derivations. Important discretization methods include finite-difference methods, finite-volume methods, finite-element methods, and spectral methods. This book emphasizes finite-volume methods, for reasons that will be explained as we go.

Galerkin methods, named for Soviet mathematician Boris Galerkin (Galerkin, 1915), include both spectral methods and finite-element methods. Galerkin methods involve expanding the fields of the model in terms of weighted sums of continuous, and therefore differentiable, basis functions, which depend on horizontal position. Simple examples include Fourier expansions and spherical harmonic expansions. In a spectral model, the basis functions are *global*, which means that each basis function is defined over the entire horizontal domain. In atmospheric science, spectral models are most often applied in the global domain, using spherical harmonics as the basis functions, but sometimes the spectral method is used in limited-area models (e.g., Moeng, 1984).

In a continuous model, infinitely many basis functions are needed to represent the spatial distribution of a model field. The basis functions appear inside a sum, each weighted by a *coefficient* that measures how strongly it contributes to the field in question. In a discrete model, the infinite set of basis functions is replaced by a finite set, and the infinite sum is replaced by a finite sum. In addition, the coefficients are defined on a discrete time grid, and almost always on a discrete vertical grid as well. Even spectral models use grid-point methods to represent the temporal and vertical structures of the atmosphere.

In a spectral model, horizontal derivatives are computed by differentiating the continuous basis functions. Although the derivatives of the individual basis functions are computed exactly, the derivatives of the model fields are only approximate because they are represented in terms of finite (rather than infinite) sums.

Finite-element methods are similar to spectral methods, except that each basis function is defined over a "patch" of the domain, rather than globally. The finite-element method can be viewed as a way of deriving grid-point methods.

Even after we have decided on a discretization method, there are many additional choices to make.

If we adopt a grid-point method, then we have to *choose* the shapes of the grid cells. Possibilities include rectangles, triangles, and hexagons. There are advantages and disadvantages to each of these possibilities.

Having settled on the shapes of the grid cells, we must *choose* where to locate the prognostic and diagnostic variables on the grid. There are excellent reasons to locate different variables in different places. This is called "staggering." It is also possible (but less common) to stagger the variables in time, i.e., to predict different variables at different time levels (e.g., Eliassen, 1956; Phillips, 1959b). We will discuss in some detail the strengths and weaknesses of various staggering schemes.

For any given grid shape and staggering, we can devise numerical schemes that are more or less accurate. Many types of "accuracy" will be discussed in this book. More accurate schemes have smaller errors, but less accurate schemes are usually simpler and faster. Again, we have to make *choices*.

1.8 Physically based design of mathematical methods

Throughout this book, I will try to persuade you that physical considerations should play a primary role in the design of the mathematical methods that we use in our models. There is a tendency to think of numerical methods as one realm of research, and physical modeling as a completely different realm. This is a mistake. The design of a numerical model should be guided, as far as possible, by our understanding of the physical processes represented by the model. This book emphasizes that very basic and inadequately recognized point.

Here is an example. To an excellent approximation, the mass of dry air does not change as the atmosphere goes about its business. This physical principle is embodied in the continuity equation, (1.1). With appropriate boundary conditions, we can show that

$$\int_{\mathrm{WA}} \nabla \cdot (\rho \mathbf{V}) dx^3 = 0 \,. \tag{1.3}$$

This is a mathematical fact, rather than a physical principle. Together, Eqs. (1.1) and (1.3) imply that

$$\frac{d}{dt}\left(\int_{\mathrm{WA}} \rho dx^3 \right) = 0 \,. \tag{1.4}$$

In these two equations, "WA" stands for whole atmosphere. Equation (1.4) is a statement of global mass conservation; in order to obtain (1.4), we had to use (1.3), which is a property of the divergence operator with suitable boundary conditions.

In a numerical model, we replace (1.1) by an approximate discrete equation; examples are given later. The approximate form of (1.1) entails a discrete approximation to the

divergence operator. The approximation inevitably involves errors, but because we are able to choose or design the approximations, we have some control over the nature of the errors. We cannot eliminate the errors, but we can refuse to accept certain kinds of errors. For example, we can refuse to accept any failure to conserve the total mass of the atmosphere. This means that we can *choose* to design our model so that a discrete analog of (1.4) is satisfied *exactly*.

In order to derive an analog of (1.4), we have to enforce an analog of (1.3); this means that we have to choose an approximation to the divergence operator that "behaves like" the exact divergence operator in the sense that the global integral (or, more precisely, a global sum approximating the global integral) is exactly zero. This can be done, quite easily. You may be surprised to learn that some models do *not* conserve mass.

There are many additional examples of important physical principles that can be enforced exactly by designing suitable approximations to differential and/or integral operators, including conservation of energy and conservation of potential vorticity. In practice, it is only possible to enforce a few such principles exactly. We must *choose* which principles to enforce, guided by our understanding of the physics.

1.9 What's the worst that could happen?

Many things can go wrong with a numerical model. Here we list some of the most basic types of bad behavior:

1. *Numerical instability.* As you probably know, models sometimes "blow up," or "crash" or "bomb." Such problems arise from numerical instability, in which small errors are amplified until they become catastrophic. It is possible to design a model that is numerically stable provided that certain conditions are met.

2. *Computational modes in time.* The continuous equations of a model predict a unique evolution through time. We say that there is a single solution. The temporally discrete equations may have more than one solution. The "extra" solutions are spurious and unphysical. There is a simple way to avoid them.

3. *Computational modes in space.* The solution of a spatially discrete model can contain spurious "noise" that persists through time because it does not interact with the physically meaningful part of the solution. Problems of this type can be minimized or avoided altogether by careful design of the discrete equations.

4. *Conservation errors.* With the continuous equations, mass, energy, momentum, potential vorticity, and other quantities are either "strictly" conserved or else conserved by an important subset of the physical processes represented in the model. Numerical errors can interfere with conservation. For example, errors in the discrete continuity equation can cause the total mass of the model to gradually diminish as a simulation

proceeds. Some types of conservation are easy to achieve, but others are more challenging. In practice, it may be necessary to choose to conserve one quantity while tolerating conservation errrors in another.

5. *Compuational dispersion.* In the continuous world, when a spatially varying intensive variable is advected by a uniform current, all wavelengths are advected at the same speed. Unfortunately, numerical advection schemes typically advect short wavelengths more slowly than longer wavelengths, causing the spatial structure of the intensive variable to "come apart" as time progresses. This problem is called computational dispersion. It can be minimized, but typically at the cost of computational dissipation of the short wavelengths.

6. *Negative concentrations.* The concentrations (i.e., mass fractions) of water vapor and other trace constituents are, of course, non-negative. Numerical errors can spuriously create negative concentrations, which are physically impossible and can cause problems in the physical parameterizations. Computational dispersion is a common cause (but not the only cause) of negative concentrations. There are several ways to avoid the problem, some better than others.

Much of this book is about how to avoid these and other potential stumbling blocks.

1.10 The utility of numerical models

A serious practical difficulty in the geophysical sciences is that it is usually impossible or at least impractical (perhaps fortunately) to perform controlled experiments using the Earth.[2] Even where laboratory experiments are feasible, as with some micrometeorological phenomena, it can be difficult to draw definite conclusions, because in the real atmosphere all physical processes interact in complicated and uncontrollable ways. Until the beginning of numerical modeling in the 1950s, the development of atmospheric science had to rely entirely upon observations of the natural atmosphere, which is an uncontrolled synthesis of many mutually dependent physical processes. Such observations can't provide direct tests of theories, which are inevitably highly idealized.

Numerical modeling is a powerful tool for studying the atmosphere through an experimental approach. Comparisons with observations must be made for evaluation of the model results. Once such comparisons have given us sufficient confidence that the model behaves like the real atmosphere, we can use it (with caution) as a substitute for the real atmosphere. Numerical experiments with such models can lead to discoveries that would not have been possible with observations alone.

Models can also be used as purely experimental tools. For example, we could perform an experiment to determine how the general circulation of the atmosphere would change if

[2] Current discussions of geoengineering are very relevant here!

the Earth's mountains were removed, and of course this was done long ago (e.g., Manabe and Terpstra, 1974).

Simpler numerical models are also very useful for studying individual phenomena, insofar as they can be isolated. Examples are models of tropical cyclones, baroclinic waves, and clouds. Simulations with these models can be compared with observations or with simpler models empirically derived from observations, or with simple theoretical models.

Numerical modeling has brought a maturity to atmospheric science. Theories, numerical simulations and observational studies have been developed in parallel since the 1950s. Observational and theoretical studies can guide the design of numerical models. It is also true that numerical simulations can suggest theoretical ideas, and can be used to design and utilize efficient observational systems.

We do not attempt, in this book, to present general rigorous mathematical theories of numerical methods; such theories are a focus of applied mathematics. Instead, we concentrate on practical aspects of the numerical solution of the specific differential equations of relevance to atmospheric modeling.

We deal mainly with "prototype" equations that are simplified or idealized versions of equations that are actually encountered in atmospheric modeling. The various prototype equations are used in dynamics, but many of them are also used in other branches of atmospheric science, such as cloud physics or radiative transfer. They include the "advection equation," the "oscillation equation," the "decay equation," the "diffusion equation," and others. We also use the shallow water equations to explore some topics, including wave propagation. Vertical coordinate systems and vertical discretization get chapters of their own, because of the powerful effects of gravity and the importance of the atmosphere's lower boundary. Emphasis is placed on time-dependent problems, but we also briefly discuss numerical methods for solving boundary-value problems.

1.11 Where we are going

Chapter 2 summarizes the equations of an atmospheric dynamical core, and discusses some of their key properties. Much of the chapter will be a review for most readers of this book.

Chapter 3 introduces the basics of finite differences. You will learn how to measure the "truncation error" of a finite-difference approximation to a derivative, and then how to design schemes that have the desired truncation error, for a differential operator of interest, in (possibly) multiple dimensions, and on a (possibly) non-uniform grid. Chapter 5 is a survey of some widely used time-differencing schemes. Chapter 6 discusses numerical schemes for solving two important prototype equations in which the only derivatives are with respect to time. The chapter also introduces the concept of numerical instability.

Chapter 7 is the first of several chapters dealing with scalar advection, which is the process by which scalar properties of the air are carried along with the air as it moves. Chapter

8 focuses on a particular advection scheme, which is called the upstream scheme. Chapter 9 discusses a family of "forward-in-time" advection schemes, of which the upstream scheme is a member. Chapter 10 generalizes to the case of two-dimensional advection. Chapter 4 introduces grids that are not based on rectangles. Chapter 11 introduces finite-volume methods, which can be used for advection and also other things, and introduces the concept of energy-conserving schemes. Chapter 12 discusses the important problem of computational dispersion, and methods designed to minimize it. Chapter 13 discusses modern Eulerian schemes. Chapter 14 discusses Lagrangian and semi-Lagrangian advection.

We then change the subject, away from advection. Chapter 15 presents methods for rapidly finding solutions of systems of linear equations. Chapter 16 discusses methods used to solve the diffusion equation. Chapter 17 introduces the shallow-water equations. Chapter 19 applies the shallow-water equations the problem of wave propagation, with an emphasis on inertia-gravity waves, and includes a detailed discussion of staggered grids. Chapter 20 discusses boundary conditions for advection and wave propagation.

The next several chapters deal with modeling the vertical structure of the atmosphere. Chapter 21 discusses sound waves and systems of equations that filter them. Chapter 22 is an introduction to vertical coordinate systems, and includes a discussion of non-hydrostatic models. Chapter 23 continues the discussion of vertical coordinate systems, with a focus on quasi-static models. Chapter 24 discusses vertical differencing in quasi-static models, including the important subject of energy conservation.

Chapter 18 presents issues that arise in trying to conserve both momentum and energy in a one-dimensional shallow-water model. This is preparation for Chapter 26, which focuses on the special issues that arise with momentum advection, including the issues of energy and enstrophy conservation.

Chapter 28 discusses finite-difference methods for use with spherical geometry. Chapter 29 gives an introduction to spectral methods. Chapter 30 introduces finite-element methods.

Chapter 31 presents a closing discussion.

Several appendices round out the book.

Many of the topics covered in this book are presented with some historical background. A comprehensive overview of the history of Earth System Modeling, including numerical modeling of the atmosphere, can be found in Randall et al. (2019).

Chapter 2

The basic equations in vector form

2.1 Introduction

This chapter presents a short summary of the "exact" equations of an atmospheric model. The equations are presented in vector form, without using a coordinate system. Most of this will be a review of ideas that you have encountered in previous work.

2.2 The equation of motion

Newton's second law can be written as $\mathbf{F} = m\mathbf{a}$, where $\mathbf{F}$ is the sum of all forces acting on a particle, m is the mass of the particle, $\mathbf{a}$ is the acceleration of the particle as seen in an inertial frame of reference. Both $\mathbf{F}$ and $\mathbf{a}$ are vectors. Because Newton's second law describes what happens to a moving particle, it is a *Lagrangian* statement.

To apply Newton's second law to fluid motion, we replace the mass by the density, ρ, which is the mass of the fluid per unit volume, and write

$$\rho \frac{D_a \mathbf{V}_a}{Dt} = \mathbf{F} \,, \tag{2.1}$$

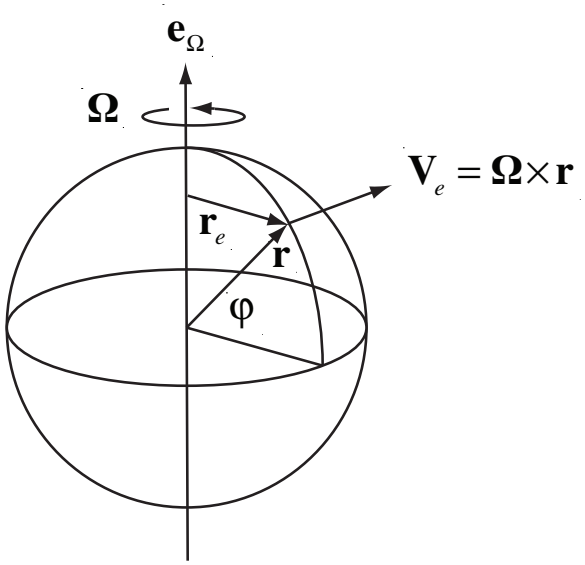
where $\mathbf{V}_a$ represents a fluid particle's velocity as seen in the inertial frame, D_a/Dt represents the time-rate of change following the particle, and $D_a\mathbf{V}_a/Dt$ is the particle's acceleration. We have put the force on the right-hand side of (2.1), which is conventional in atmospheric science and fluid dynamics generally.

2.2.1 Converting to a rotating frame of reference

We want to use Newton's second law to describe the acceleration of the air as seen in a *non-inertial* reference frame that rotates with the Earth, because that is the reference frame

that we experience as riders on the Earth. The Earth's angular velocity is a vector, denoted by $\mathbf{\Omega}$, that points towards the celestial North Pole, as shown in Fig. 2.1. The magnitude of $\mathbf{\Omega}$ is $2\pi/86,164\text{s} = 7.292 \times 10^{-5}\text{s}^{-1}$, corresponding to a sidereal day of length 86,164 s, which is slightly shorter than the "nominal" length of day, 86,400 s. In the discussion below, we assume that $\mathbf{\Omega}$ is independent of time, which is very nearly true.

Figure 2.1: Sketch defining notation used in the text.



Let $\mathbf{A}$ be an arbitrary vector. The Lagrangian time-rate-of-change of $\mathbf{A}$ is $D_a\mathbf{A}/Dt$ as seen in an inertial frame of reference, and $D\mathbf{A}/Dt$ as seen in the rotating frame of reference. The two time-rates-of-change are related by

$$\frac{D_a\mathbf{A}}{Dt} = \frac{D\mathbf{A}}{Dt} + \mathbf{\Omega}\times\mathbf{A}\,. \tag{2.2}$$

Let $\mathbf{r}$ be a position vector that points from the Earth's center to a particle of air. See again Fig. 2.1. The "absolute" velocity of the particle, as seen in an inertial reference frame, is denoted by

$$\mathbf{V}_a \equiv \frac{D_a\mathbf{r}}{Dt}\,, \tag{2.3}$$

and the "relative" velocity of the particle, as seen in the rotating frame, is

$$\mathbf{V} \equiv \frac{D\mathbf{r}}{Dt}\,, \tag{2.4}$$

Applying (2.2) to $\mathbf{r}$, and using (2.3) and (2.4), we can write

$$\mathbf{V}_a = \mathbf{V} + \boldsymbol{\Omega} \times \mathbf{r} \,. \tag{2.5}$$

As an aside,

$$\mathbf{V}_e \equiv \boldsymbol{\Omega} \times \mathbf{r} = (\Omega r \cos\varphi)\ \mathbf{e}_\lambda \tag{2.6}$$

is the eastward velocity (as seen in the inertial frame) that a particle at radius $\mathbf{r}$ and latitude φ experiences due to the Earth's rotation (Fig. 2.1). Note that "east" means "the direction towards which the planet is rotating." At the Earth's Equator, the magnitude of $\mathbf{V}_e$ is about a thousand miles per hour.

Next, we apply (2.2) to $\mathbf{V}_a$:

$$\frac{D_a \mathbf{V}_a}{Dt} = \frac{D\mathbf{V}_a}{Dt} + \boldsymbol{\Omega} \times \mathbf{V}_a \,. \tag{2.7}$$

Eq. (2.5) can be used to eliminate $\mathbf{V}_a$ on the right-hand side of (2.7). Using (2.4) again, the result can be written as

$$\begin{aligned} \frac{D_a \mathbf{V}_a}{Dt} =& \frac{D}{Dt}(\mathbf{V} + \boldsymbol{\Omega} \times \mathbf{r}) + \boldsymbol{\Omega} \times (\mathbf{V} + \boldsymbol{\Omega} \times \mathbf{r}) \\ =& \frac{D\mathbf{V}}{Dt} + 2\boldsymbol{\Omega} \times \mathbf{V} + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) \ . \end{aligned} \tag{2.8}$$

Here $-2\boldsymbol{\Omega} \times \mathbf{V}$ represents the Coriolis acceleration, whose direction is perpendicular to $\mathbf{V}$, and $-\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})$ represents the centrifugal acceleration, which is perpendicular to the Earth's axis of rotation.

With this preparation, we can express Newton's statement of momentum conservation in the rotating coordinate system:

$$\rho \left[\frac{D\mathbf{V}}{Dt} + 2\boldsymbol{\Omega} \times \mathbf{V} + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) \right] = \mathbf{F} \,. \tag{2.9}$$

2.2.2 Forces

To proceed, we have to specify the forces that comprise **F**. These include gravity, the pressure-gradient force, and friction. We write

$$\rho\left[\frac{D\mathbf{V}}{Dt}+2\boldsymbol{\Omega}\times\mathbf{V}+\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r})\right]=\rho\mathbf{g}_a-\nabla p-\nabla\cdot\mathbf{S}\,. \tag{2.10}$$

Here p is the pressure, and the acceleration due to gravity is

$$\mathbf{g}_a\equiv-\nabla\phi_a\,, \tag{2.11}$$

where ϕ_a is the gravitational potential. The stress tensor is $\mathbf{S}$ (see Appendix A). The dimensions of $\mathbf{S}$ are density times velocity squared, e.g., (kg m^{-3}) (m s^{-1})^2 = kg m^{-1} s^{-2}. Note that $\nabla\cdot\mathbf{S}$ is a vector.

The pressure-gradient force can be written in many different ways. One particularly useful form is

$$-\frac{1}{\rho}\nabla p=-\theta\nabla\Pi\,, \tag{2.12}$$

You are asked to prove this in a problem at the end of this chapter.

2.2.3 Apparent gravity

It is customary to simplify (2.10) as follows: The centrifugal acceleration can be written as

$$\begin{aligned}\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r})&=-\Omega^2\mathbf{r}_e\\&=-\nabla\left[\frac{1}{2}|\boldsymbol{\Omega}\times\mathbf{r}|^2\right]\,,\end{aligned} \tag{2.13}$$

where $\mathbf{e}_\Omega$ is a unit vector pointing toward the celestial north pole, and $\mathbf{r}_e$ is the vector shown in Fig. 2.1. Eq. (2.13) tells us that the centrifugal acceleration is perpendicular to the axis

of the Earth's rotation. The second line of (2.13) shows that the centrifugal acceleration is the gradient of a "centrifugal potential," which is given by $\frac{1}{2}|\boldsymbol{\Omega}\times\mathbf{r}|^2$. This allows us to define an "apparent" gravity that combines the effects of true gravity and the centrifugal acceleration:

$$\begin{aligned}\mathbf{g} \equiv& \mathbf{g}_a + |\boldsymbol{\Omega}|^2\mathbf{r}_e \\ =& -\nabla\phi \ ,\end{aligned} \tag{2.14}$$

where

$$\phi \equiv \phi_a + \frac{1}{2}|\boldsymbol{\Omega}\times\mathbf{r}|^2 \ . \tag{2.15}$$

is called the "geopotential." Using (2.13) and (2.15), we can rewrite (2.10) in the relatively simple and (I hope) familiar form

$$\rho\left(\frac{D\mathbf{V}}{Dt} + 2\boldsymbol{\Omega}\times\mathbf{V}\right) = -\rho\nabla\phi - \nabla p - \nabla\cdot\mathbf{S}. \tag{2.16}$$

Because of the centrifugal acceleration, and also because of spatial inhomogeneities of true gravity, due to inhomogeneities of the Earth's mass distribution, the geopotential isosurfaces are not quite spherical. In particular, the centrifugal acceleration causes the geopotential isosurfaces to bulge outward in low latitudes, and pull inward near the poles. Because of the centrifugal acceleration, the shape of a geopotential isosurface is (approximately) an "oblate spheroid," whose radius is about 20 km greater at the equator than at the poles. For further explanation, see Staniforth (2022).

We can *define* exact hydrostatic balance as a motionless state in which the pressure is uniform along isosurfaces of the geopotential, so that

$$\nabla_\phi p = 0 \tag{2.17}$$

and (2.16) reduces to

$$0 = -\nabla\phi - \alpha\nabla p \ . \tag{2.18}$$

Throughout the rest of this book, we will assume that

$$\mathbf{g} \equiv -\nabla\phi \cong \mathbf{g}_a \cong -g\mathbf{e}_r \, . \tag{2.19}$$

Here $\mathbf{e}_r$ is a unit vector pointing away from the center of the Earth (Fig. 2.1). The approximation in (2.19) is justified because the centrifugal acceleration is small compared to $\mathbf{g}_a$ and because the non-radial components of $\mathbf{g}_a$ are very small compared to its radial component. When we use (2.19) with an approximate spatially constant value of g (which is what we usually do) we must also approximate the shape of the Earth as a sphere, with topographical mountains and valleys. Staniforth (2022) gives a much more detailed discussion.

2.3 The continuity equation

The continuity equation can be expressed in these two coordinate-free forms:

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\mathbf{V} \, , \tag{2.20}$$

$$\frac{D\alpha}{Dt} = \alpha\nabla\cdot\mathbf{V} \, . \tag{2.21}$$

2.4 The thermodynamic energy equation

The thermodynamic energy equation can be written in these three ways:

$$\frac{D}{Dt}(c_v T) + p\frac{D\alpha}{Dt} = LC - \nabla\cdot\mathbf{R} + \delta \, , \tag{2.22}$$

$$\frac{D}{Dt}(c_p T) - \omega\alpha = LC - \nabla\cdot\mathbf{R} + \delta \, , \tag{2.23}$$

$$\frac{D\theta}{Dt} = \frac{LC - \nabla \cdot \mathbf{R} + \delta}{\Pi} \,. \tag{2.24}$$

Here T is temperature; c_v is the specific heat of air at constant volume; c_p is the specific heat of air at constant pressure; LC is the rate of latent heat release, where L is the latent heat of condensation and C is the condensation rate; $-\nabla \cdot \mathbf{R}$ is the radiative heating rate per unit mass; and δ is the rate of kinetic energy dissipation per unit mass. The quantity $c_v T$ is called the internal energy per unit mass, and $c_p T$ is called the enthalpy per unit mass. We also define

$$\theta \equiv \frac{c_p T}{\Pi} \tag{2.25}$$

as the potential temperature, and

$$\Pi \equiv c_p \left(\frac{p}{p_0} \right)^{\kappa} \,, \tag{2.26}$$

as the Exner function, where

$$\kappa \equiv \frac{R}{c_p} \,, \tag{2.27}$$

and R is the specific gas constant for dry air. Note that θ and T are related by

$$c_p T = \Pi \theta \,. \tag{2.28}$$

We will use

$$R = c_p - c_v \,. \tag{2.29}$$

Finally, we need the equation of state, which can be written as

$$p\alpha = RT \ . \tag{2.30}$$

Of the three forms of the thermodynamic equation, (2.24) is the simplest, but (2.22) and (2.23) include terms that explicitly represent the conversion between thermodynamic energy and mechanical energy. Further discussion is given later, starting in Chapter 22.

2.5 Moisture conservation

Conservation of water vapor is expressed by

$$\frac{Dq_v}{Dt} = -C \ , \tag{2.31}$$

where q_v is the water vapor mixing ratio. Conservation of latent energy is expressed by

$$\frac{D}{Dt}\left(Lq_v\right) = -LC \ , \tag{2.32}$$

where L is the latent heat of condensation.

2.6 Segue

This chapter has presented the dynamical equations of an atmospheric model in continuous form. The next chapter deals with the first step in discretizing the continuous equations: We must formulate discrete approximations to time and space derivatives.

2.7 Problems

1. Prove Eq. (2.12).
2. Prove Eq. (2.13).

Chapter 3

Finite-difference approximations to derivatives

3.1 Finite-difference quotients

Consider the derivative df/dx, where $f = f(x)$, and x is the independent variable (which looks like space but could be either space or time). Finite-difference methods represent the continuous function $f(x)$ by a set of values defined at a finite number of discrete points in a specified (spatial or temporal) region. The discrete points form a "grid." [1] We will use the simple one-dimensional (hereafter 1D) grid shown in Fig. 3.1 to introduce some basic ideas; multidimensional grids will be discussed later. The interval Δx, shown in the figure, is called the grid spacing. For the time being, we assume that the grid spacing is uniform, and that $x_0 = 0$; then $x_j = j\Delta x$, where j is the "index" used to identify the grid points. Note that x is defined only at the grid points denoted by the integers j, $j+1$, etc.

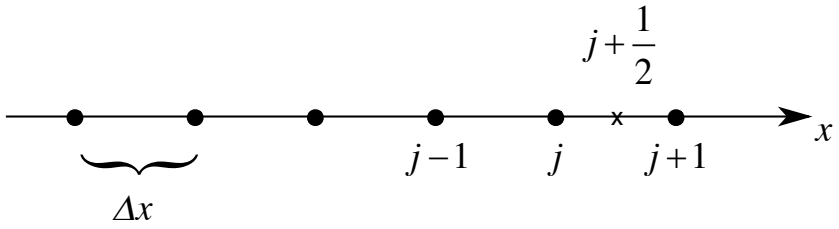


Figure 3.1: An example of a one-dimensional grid, with uniform grid spacing Δx. The grid points are denoted by the integer index j. Half-integer grid points can also be defined, as shown by the x at $j+\frac{1}{2}$.

Using the notation $f_j = f\left(x_j\right) = f\left(j\Delta x\right)$, we can define

$$\text{the } \textit{forward difference} \text{ at the point } j \text{ by } f_{j+1} - f_j \tag{3.1}$$

[1] Sometimes the words "mesh" or "lattice" are used in place of "grid."

the *backward difference* at the point j by $f_j - f_{j-1}$, and (3.2)

the *centered difference* at the point $j + \frac{1}{2}$ by $f_{j+1} - f_j$. (3.3)

Note that f itself is not defined at the point $j + \frac{1}{2}$. From (3.1) - (3.3) we can define the following "finite-difference quotients:"

the forward-difference quotient at the point j:

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1} - f_j}{\Delta x} \; ; \tag{3.4}$$

the backward-difference quotient at the point j:

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_j - f_{j-1}}{\Delta x} \; ; \tag{3.5}$$

and the centered-difference quotient at the point $j + \frac{1}{2}$:

$$\left(\frac{df}{dx}\right)_{j+\frac{1}{2},\,\text{approx}} = \frac{f_{j+1} - f_j}{\Delta x} \; . \tag{3.6}$$

In addition, the centered-difference quotient at the point j can be defined by

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1} - f_{j-1}}{2\Delta x} \; . \tag{3.7}$$

Since (3.4) and (3.5) employ the values of f at two points, and give an approximation to df/dx at one of the same points, they can be called *two-point approximations*. On the other hand, (3.6) and (3.7) are *three-point approximations*, because the approximation to df/dx is defined at a location different from the locations of the two values of f on the right-hand side of the equation.

When x is time, the time point is frequently referred to as a "level." In that case, (3.4) and (3.5) can be referred to as two-level approximations and (3.6) and (3.7) as three-level approximations.

Here we consider the discrete dependent variables to be defined at points. This is called the finite-difference method. Later we will consider schemes in which the discrete dependent variables represent averages over grid cells. These are called finite-volume schemes.

How accurate are the finite-difference approximations introduced above? "Accuracy" can be defined in many ways, as we shall see, so it is important to clearly state what we mean here. One commonly used measure of accuracy is *truncation error*, which refers to the truncation of an infinite series expansion. As an example, consider the forward difference quotient

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1}-f_j}{\Delta x} \equiv \frac{f\left[(j+1)\,\Delta x\right]-f\left(j\Delta x\right)}{\Delta x}\,. \tag{3.8}$$

Expand f in a Taylor series about the point x_j, as follows:

$$\begin{aligned} f_{j+1} =& f_j + \Delta x\left(\frac{df}{dx}\right)_j + \frac{(\Delta x)^2}{2!}\left(\frac{d^2f}{dx^2}\right)_j \\ &+ \frac{(\Delta x)^3}{3!}\left(\frac{d^3f}{dx^3}\right)_j + \cdots + \frac{(\Delta x)^{n-1}}{(n-1)!}\left(\frac{d^{n-1}f}{dx^{n-1}}\right)_j + \cdots \end{aligned} \tag{3.9}$$

The expansion (3.9) can be derived without any assumptions or approximations except that the referenced derivatives exist (Arfken, 1985). When they do, the expansion is exact if all terms of the sum are kept. This means that if we know the function and all of its derivatives at a single point, we can (in principle) calculate the value of the function anywhere else in the domain. That's pretty amazing.

Eq. (3.9) can be rearranged to

$$\left(\frac{df}{dx}\right)_j = \frac{f_{j+1}-f_j}{\Delta x} + \varepsilon\,, \tag{3.10}$$

where

$$\varepsilon \equiv \frac{\Delta x}{2!}\left(\frac{d^2 f}{dx^2}\right)_j + \frac{(\Delta x)^2}{3!}\left(\frac{d^3 f}{dx^3}\right)_j + \cdots + \frac{\Delta x^{n-2}}{(n-1)!}\left(\frac{d^{n-1} f}{dx^{n-1}}\right)_j + \cdots \tag{3.11}$$

is the truncation error, mentioned above. If Δx is small enough, the leading term on the right-hand side of (3.11) will be the largest part of the error. The lowest power of Δx that appears in the truncation error is called the "*order of accuracy*" of the corresponding difference quotient. For example, the leading term of (3.11) is "of order" Δx, abbreviated as $\mathscr{O}(\Delta x)$, and so we say that (3.10) is a first-order approximation or an approximation of first-order accuracy. Obviously (3.5) is also first-order accurate.

Just to be as clear as possible, a first-order scheme for the first derivative has the form $(df/dx)_{j,\text{ approx}} = (df/dx)_j + \mathscr{O}(\Delta x)$, where $(df/dx)_{j,\text{approx}}$ is an *approximation* to the first derivative, and $(df/dx)_j$ is the *true* first derivative. Similarly, a second-order accurate scheme for the first derivative has the form $(df/dx)_{j,\text{ approx}} = (df/dx)_j + O\left[(\Delta x)^2\right]$, and so on for higher orders of accuracy.

Similar analyses show that (3.6) and (3.7) are of second-order accuracy. For example, we can write

$$f_{j-1} = f_j + \left(\frac{df}{dx}\right)_j (-\Delta x) + \left(\frac{d^2 f}{dx^2}\right)_j \frac{(-\Delta x)^2}{2!} + \left(\frac{d^3 f}{dx^3}\right)_j \left[\frac{-(\Delta x)^3}{3!}\right] + \cdots \quad . \tag{3.12}$$

Subtracting (3.12) from (3.9) gives

$$f_{j+1} - f_{j-1} = 2\left(\frac{df}{dx}\right)_j (\Delta x) + \frac{2}{3!}\left(\frac{d^3 f}{dx^3}\right)_j \left[(\Delta x)^3\right] + \cdots \text{odd powers only} , \tag{3.13}$$

which can be rearranged to

$$\left(\frac{df}{dx}\right)_j = \frac{f_{j+1} - f_{j-1}}{2\Delta x} - \left(\frac{d^3 f}{dx^3}\right)_j \frac{\Delta x^2}{3!} + \mathscr{O}\left[(\Delta x)^4\right] . \tag{3.14}$$

Similarly,

$$\left(\frac{df}{dx}\right)_{j+\frac{1}{2}} \cong \frac{f_{j+1}-f_j}{\Delta x} - \left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!} + O\left[(\Delta x)^4\right] . \tag{3.15}$$

The ratio of the error of (3.14) to the error of (3.15) is approximately given by

$$\frac{\left(\frac{d^3f}{dx^3}\right)_j \frac{\Delta x^2}{3!}}{\left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!}} = \frac{4\left(\frac{d^3f}{dx^3}\right)_j}{\left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}}} \cong 4 . \tag{3.16}$$

This shows that the error of (3.14) is about four times as large as the error of (3.15), even though both finite-difference quotients have second-order accuracy. The point is that the *"order of accuracy" tells how rapidly the error changes as the grid is refined, but it does not tell how large the error is for a given grid size.* It is possible for a scheme of low-order accuracy to give a more accurate result than a scheme of higher-order accuracy, if a finer grid spacing is used with the low-order scheme.

Suppose that the leading (and dominant) term of the error has the form

$$\varepsilon \cong C(\Delta x)^p , \tag{3.17}$$

where C is a constant. From (3.17) we see that $\ln(\varepsilon) \cong p\ln(\Delta x) + \ln(C)$, and so

$$\frac{d[\ln(\varepsilon)]}{d[\ln(\Delta x)]} \cong p . \tag{3.18}$$

This means that if we plot $\ln(\varepsilon)$ as a function of $\ln(\Delta x)$ (i.e., plot the error as a function of the grid spacing on "log-log" paper), we will get a nearly straight line whose slope is p. This is a simple way to determine *empirically* the order of accuracy of a finite-difference quotient. Of course, in order to carry this out in practice it is necessary to know the error of the finite-difference approximation, which means that the exact derivative must be known. For that reason, this empirical approach is usually implemented by using an analytical "test function."

Can you think of a way to use the empirical approach even when the exact derivative is not known?

3.2 Groping towards higher accuracy

Suppose that we write

$$\frac{f_{j+2}-f_{j-2}}{4\Delta x}=\left(\frac{df}{dx}\right)_j+\frac{1}{3!}\left(\frac{d^3f}{dx^3}\right)_j(2\Delta x)^2+\cdots\text{even powers only}\,. \tag{3.19}$$

Here we have written a centered difference using the points $j+2$ and $j-2$ instead of $j+1$ and $j-1$, respectively. It should be clear that (3.19) is second-order accurate, although for any given value of Δx the error of (3.19) is expected to be larger than the error of (3.14). We can combine (3.14) and (3.19) with a weight, w, so as to obtain a "hybrid" approximation to $(df/dx)_j$:

$$\begin{aligned}\left(\frac{df}{dx}\right)_j=&w\left(\frac{f_{j+1}-f_{j-1}}{2\Delta x}\right)+(1-w)\left(\frac{f_{j+2}-f_{j-2}}{4\Delta x}\right)\\&-\frac{w}{3!}\left(\frac{d^3f}{dx^3}\right)_j(\Delta x)^2-\frac{(1-w)}{3!}\left(\frac{d^3f}{dx^3}\right)_j(2\Delta x)^2+\mathscr{O}\left[(\Delta x)^4\right]\,.\end{aligned} \tag{3.20}$$

Inspection of (3.20) shows that we can force the coefficient of $(\Delta x)^2$ to vanish by choosing

$$w+(1-w)\,4=0,\text{ which implies that }w=4/3\,. \tag{3.21}$$

With this choice, (3.20) reduces to

$$\left(\frac{df}{dx}\right)_j=\frac{4}{3}\left(\frac{f_{j+1}-f_{j-1}}{2\Delta x}\right)-\frac{1}{3}\left(\frac{f_{j+2}-f_{j-2}}{4\Delta x}\right)+\mathscr{O}\left[(\Delta x)^4\right]\,. \tag{3.22}$$

This is a fourth-order accurate approximation to the first derivative.

The derivation given above, in terms of a weighted combination of two second-order approximations, can be interpreted as a linear *extrapolation* of the value of the finite-difference expression to a smaller grid size, as illustrated in Fig. 3.2. Both extrapolation and interpolation use weights that sum to one. In the case of interpolation, both weights lie between zero and one, while in the case of extrapolation one of the weights is larger than one and the other weight is negative. The concepts of extrapolation and interpolation will be discussed in more detail later in this chapter.

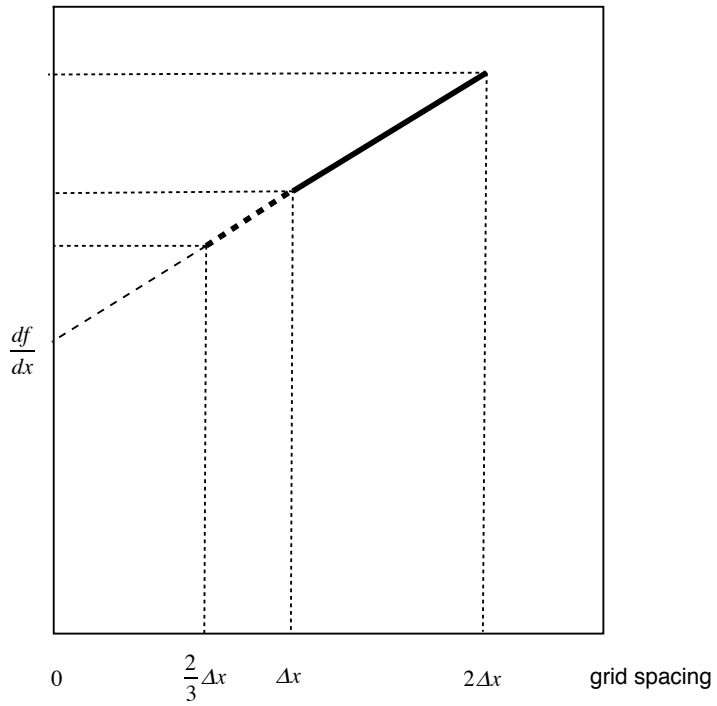


Figure 3.2: Schematic illustrating an *interpretation* of the fourth-order scheme given by (3.22) in terms of an extrapolation from two second-order schemes. The fourth-order scheme with grid spacing Δx produces the same accuracy as a second-order scheme with grid spacing $(2/3)\,\Delta x$.

3.3 A systematic approach

3.3.1 A family of schemes

There are more systematic ways to construct schemes of any desired order of accuracy. One such approach is presented here. Suppose that we write a finite-difference approximation to $(df/dx)_j$ in the following somewhat generalized form:

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'} f\left(x_j + j'\Delta x\right). \tag{3.23}$$

Here the $a_{j'}$ are coefficients or "weights," which are undetermined at this point. In writing (3.23), we have assumed for simplicity that Δx is a constant; this assumption will be relaxed soon. The index j' in (3.23) is a counter that is zero at our "home base," which is grid point j. For $j' < 0$ we count to the left, and for $j' > 0$ we count to the right. According to (3.23), our finite-difference approximation to $(df/dx)_j$ has the form of a weighted sum of values of $f(x)$ at various grid points in the neighborhood of point j. There is no guarantee that

Eq. (3.23) actually gives an approximation to $(df/dx)_j$, *but we can design a scheme by choosing suitable expressions for the* $a_{j'}$. In most schemes, all but a few of the $a_{j'}$ will be zero, so that the sum (3.23) will actually involve only a few (non-zero) terms. Every finite-difference approximation that we have considered so far does indeed have the form (3.23), but you should be aware that there are (infinitely many!) schemes that do *not* have this form; a few of them will be discussed later.

Introducing a Taylor series expansion, we can write

$$f\left(x_j + j'\Delta x\right) = f_j^0 + f_j^1\left(j'\Delta x\right) + f_j^2\frac{\left(j'\Delta x\right)^2}{2!} + f_j^3\frac{\left(j'\Delta x\right)^3}{3!} + \cdots . \tag{3.24}$$

Here we introduce a new short-hand notation: f_j^n is the (exact) nth derivative of f, evaluated at the point j. Using (3.24), we can rewrite (3.23) as

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'}\left[f_j^0 + f_j^1\left(j'\Delta x\right) + f_j^2\frac{\left(j'\Delta x\right)^2}{2!} + f_j^3\frac{\left(j'\Delta x\right)^3}{3!} + \cdots\right]. \tag{3.25}$$

Let's consider what happens to each of the various "pieces" of the right-hand side of (3.25) when $\Delta x \to 0$. The first piece is

$$\frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'} f_j^0 = \left(\frac{f_j^0}{\Delta x}\right)\sum_{j'=-\infty}^{\infty} a_{j'} . \tag{3.26}$$

As Δx goes to zero, this expression will blow up unless

$$\sum_{j'=-\infty}^{\infty} a_{j'} = 0 . \tag{3.27}$$

Therefore, any viable scheme has to satisfy (3.27).

The second piece of the right-hand side of (3.25) is

$$\frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'} f_j^1\left(j'\Delta x\right) = f_j^1\sum_{j'=-\infty}^{\infty} a_{j'} j' . \tag{3.28}$$

Since f_j^1 is the exact first derivative that we are trying to approximate, we must require that

$$\sum_{j'=-\infty}^{\infty} a_{j'} j' = 1 \,. \tag{3.29}$$

To summarize, we have shown that in order to have at least first-order accuracy, we must choose the $a_{j'}$s so that

$$\sum_{j'=-\infty}^{\infty} a_{j'} = 0 \text{ and } \sum_{j'=-\infty}^{\infty} j' a_{j'} = 1 \,. \tag{3.30}$$

In order to satisfy these two equations, we need at least two unknowns. In other words, the scheme must use at least two non-zero values of the $a_{j'}$s, so a finite-difference approximation to the first derivative must use at least two grid points. This is intuitively obvious.

To achieve at least second-order accuracy, we must impose an additional requirement, namely that the third piece of the right-hand side of (3.25) vanishes. This leads to

$$\sum_{j'=-\infty}^{\infty} \left(j'\right)^2 a_{j'} = 0 \,. \tag{3.31}$$

In general, to approximate the first derivative with at least nth-order accuracy (with uniform grid spacing), we must require that

$$\sum_{j'=-\infty}^{\infty} \left(j'\right)^m a_{j'} = \delta_{m,1} \text{ for } 0 \le m \le n \,. \tag{3.32}$$

Here $\delta_{m,1}$ is the Kronecker delta, which is used for notational convenience. In order to satisfy (3.32), we must solve a system of $n+1$ linear equations for the $n+1$ unknown coefficients $a_{j'}$.

According to (3.32), a scheme of nth-order accuracy can be constructed by satisfying $n+1$ equations. As mentioned above, because (3.30) involves two equations, a first-order scheme has to involve at least two grid points, i.e., there must be at least two non-zero values of $a_{j'}$. Pretty obvious, right? A second-order scheme must involve at least three grid points.

Note that we could make a first-order scheme that used fifty grid points if we wanted to – but then, why would we want to? A scheme that is parsimonious in its use of grid points is called "*compact*." Compact schemes are simpler to code and run faster.

Here is an example. Still assuming a uniform grid, a first order scheme for f_j^1 can be constructed using the points j and $j+1$ as follows. From (3.30), we get $a_0 + a_1 = 0$ and $a_1 = 1$. It follows that we must choose $a_1 = -1$. Substituting into (3.23), we find that the scheme is given by $f_j^1 \cong \left[f\left(x_{j+1}\right) - f\left(x_j\right)\right]/\Delta x$, i.e., it is the same as the one-sided forward difference discussed earlier. In a similar way, we can also construct a first-order scheme using the points j and $j-1$, with another one-sided and familiar result.

If we choose the points $j+1$ and $j-1$, imposing the two requirements for first-order accuracy, i.e., (3.30), will actually give us the centered second-order scheme, i.e., $f_j^1 \cong \left[f\left(x_{j+1}\right) - f\left(x_{j-1}\right)\right]/\left(2\Delta x\right)$ because (3.31) is satisfied "accidentally" or "automatically." The scheme obtained from the *two* requirements needed for first-order accuracy actually satisfies the *three* requirements for second order accurary. This "miracle" happens because the scheme is symmetrical about the point j. If we choose the three points $j-1$, j, and $j+1$, and require second-order accuracy, we get exactly the same centered scheme, because a_0 turns out to be zero. As discussed below, this does not happen with a nonuniform grid.

The list of points that we use in constructing a scheme is called the "*stencil*" for the scheme. When we choose *not* to include certain points in the stencil, in essence we are preemptively setting the coefficients of those points to zero.

It should now be apparent that (3.32) can be used to construct schemes of arbitrarily high accuracy, simply by allocating enough grid points and solving the resulting system of linear equations for the $a_{j'}$. Schemes that use a lot of grid points involve lots of arithmetic, so there is a "law of diminishing returns" at work. As a rough rule of thumb, it is usually not worth the effort to go beyond 5th-order accuracy.

3.3.2 A generalization for use with nonuniform grids

We now work out a generalization of the family of schemes described by (3.23), for the case of (possibly) non-uniform grid spacing. Eq. (3.23) is replaced by

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} b_{j,j'} f(x_{j'}). \tag{3.33}$$

Note that, since Δx is no longer a constant, the factor of $1/\Delta x$ that appears in (3.23) has been omitted in (3.33). Also, in order to avoid notational confusion, we have replaced the symbol $a_{j'}$, which has one subscript, by $b_{j,j'}$, which has two. The subscript j is included on $b_{j,j'}$ because on a non-uniform grid the numerical values of the coefficients are different

for different values of j, i.e., at different places. Naturally, it is going to turn out that $b_{j,j'} \sim 1/\Delta x$.

Using this new notation, (3.24) is replaced by

$$f\left(x_{j'}\right) = f_j^0 + f_j^1 \left(\Delta x\right)_{j,j'} + f_j^2 \frac{\left(\Delta x\right)_{j,j'}^2}{2!} + f_j^3 \frac{\left(\Delta x\right)_{j,j'}^3}{3!} + \cdots . \tag{3.34}$$

Here $\left(\Delta x\right)_{j,j'} \equiv x_{j'} - x_j$ takes the place of $j'\Delta x$ in (3.24). Note that $\left(\Delta x\right)_{j,0} = 0$, and $\left(\Delta x\right)_{j,j'} < 0$ for $j' < 0$. Substitution of (3.34) into (3.33) gives

$$\left(\frac{df}{dx}\right)_j = \sum_{j'=-\infty}^{\infty} b_{j,j'} \left[f_j^0 + f_j^1 \left(\Delta x\right)_{j,j'} + f_j^2 \frac{\left(\Delta x\right)_{j,j'}^2}{2!} + f_j^3 \frac{\left(\Delta x\right)_{j,j'}^3}{3!} + \cdots \right] . \tag{3.35}$$

To get first-order accuracy from (3.35), we must require that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} = 0 \text{ and } \sum_{j'=-\infty}^{\infty} b_{j,j'} \left(\Delta x\right)_{j,j'} = 1, \text{ for all } j . \tag{3.36}$$

Compare with (3.30). It may appear that when we require first-order accuracy by enforcing (3.36), the leading term of the error in (3.35), namely $\sum_{j'=-\infty}^{\infty} b_{j,j'} f_j^2 \frac{\left(\Delta x\right)_{j,j'}^2}{2!}$, will be of order $\left(\Delta x\right)^2$, but actually it is of order Δx because, as mentioned above and shown below, $b_{j,j'} \sim 1/\Delta x$.

To achieve second-order accuracy with (3.35), we must require, in addition to (3.36), that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} \left(\Delta x\right)_{j,j'}^2 = 0 \text{ for all } j . \tag{3.37}$$

Eq. (3.37) is the requirement that the first-order part of the error vanishes, so that we have at least second-order accuracy. In general, to have at least nth-order accuracy, we must require that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} \left(\Delta x\right)_{j,j'}^{m} = \delta_{m,1} \text{ for all } j, \text{ and for } 0 \leq m \leq n \, . \tag{3.38}$$

This is a generalization of Eq. (3.32).

As an example, consider the first-order accurate scheme using the points j and $j+1$. Since we are using only those two points, the only non-zero coefficients are $b_{j,0}$ and $b_{j,1}$, and they must satisfy the two equations corresponding to (3.36), i.e., $b_{j,0} + b_{j,1} = 0$ and $b_{j,1} = 1/\left(\Delta x\right)_{j,1}$. It follows that $b_{j,0} = -1/\left(\Delta x\right)_{j,1}$. Referring back to (3.33), we see that the scheme is $f_j^1 \cong \left[f\left(x_{j+1}\right) - f\left(x_j\right)\right] / \left(\Delta x\right)_{j,1}$, which, not unexpectedly, has the same form as the result that we obtained for the case of the uniform grid. This shows that the non-uniformity of the grid is irrelevant when we use only two points for first-order accuracy.

To obtain a second-order accurate approximation to f_j^1 on a nonuniform grid, using the three points $j-1$, j, and $j+1$, we must require, from (3.36) that

$$b_{j,-1} + b_{j,0} + b_{j,1} = 0 \quad \text{and} \quad b_{j,-1} \left(\Delta x\right)_{j,-1} + b_{j,1} \left(\Delta x\right)_{j,1} = 1 \, , \tag{3.39}$$

which suffices for first-order accuracy, and additionally from (3.37) that

$$b_{j,-1} \left(\Delta x\right)_{j,-1}^2 + b_{j,1} \left(\Delta x\right)_{j,1}^2 = 0 \, . \tag{3.40}$$

Important: Note that

$$\left(\Delta x\right)_{j,-1} \equiv x_{j-1} - x_j = -\left(\Delta x\right)_{j-1,1} \, . \tag{3.41}$$

The solution of the system (3.39) – (3.40) can be written as

$$b_{j,-1} = \frac{\left(\Delta x\right)_{j,1}}{\left(\Delta x\right)_{j,-1} \left[\left(\Delta x\right)_{j,1} - \left(\Delta x\right)_{j,-1}\right]} \, , \tag{3.42}$$

$$b_{j,0} = -\left[\frac{\left(\Delta x\right)_{j,1} + \left(\Delta x\right)_{j,-1}}{\left(\Delta x\right)_{j,1} \left(\Delta x\right)_{j,-1}}\right] \, , \tag{3.43}$$

$$b_{j,1} = \frac{-\left(\Delta x\right)_{j,-1}}{\left(\Delta x\right)_{j,1}\left[\left(\Delta x\right)_{j,1} - \left(\Delta x\right)_{j,-1}\right]} \ . \tag{3.44}$$

For the special case of uniform grid-spacing this reduces to the centered second-order scheme discussed earlier.

Here is a simple and very practical question: Suppose that we use a scheme that has second-order accuracy on a uniform grid, but we go ahead and apply it on a non-uniform grid. (People do this all the time!) What happens? As a concrete example, consider the scheme

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f\left(x_{j+1}\right) - f\left(x_{j-1}\right)}{x_{j+1} - x_{j-1}} \ . \tag{3.45}$$

By inspection, we have

$$b_{j,-1} = \frac{-1}{x_{j+1} - x_{j-1}} \ , \tag{3.46}$$

$$b_{j,0} = 0 \ , \tag{3.47}$$

$$b_{j,1} = \frac{1}{x_{j+1} - x_{j-1}} \ . \tag{3.48}$$

Eqs. (3.46) - (3.48) do satisfy both of the conditions in (3.39), so that *the scheme does have first-order accuracy, even on the non-uniform grid*. Eq. (3.42) - (3.44) are not satisfied, however, so it appears that second-order accuracy is lost.

This argument is a bit too hasty, however. Intuition suggests that, if the grid-spacing varies slowly enough, the scheme given by (3.45) should be almost as accurate as if the grid-spacing were strictly constant. Intuition can never prove anything, but it can suggest ideas. Let's pursue this idea to see if it has merit. Define $\left(\Delta x\right)_{j+1/2} \equiv x_{j+1} - x_j > 0$ for

all j, and let Δ be the grid spacing at some reference grid point. We write the centered second-order scheme appropriate to a uniform grid, but apply it on a non-uniform grid:

$$\begin{aligned}\left(\frac{df}{dx}\right)_{j,\text{ approx}} =& \frac{f_{j+1}-f_{j-1}}{x_{j+1}-x_{j-1}} \\ =& \left[\frac{f_j + f_j^1 (\Delta x)_{j+1/2} + \frac{1}{2!} f_j^2 (\Delta x)_{j+1/2}^2 + \mathscr{O}\left(\Delta^3\right)}{(\Delta x)_{j+1/2} + (\Delta x)_{j-1/2}}\right] \\ & - \left[\frac{f_j - f_j^1 (\Delta x)_{j-1/2} + \frac{1}{2!} f_j^2 (\Delta x)_{j-1/2}^2 + \mathscr{O}\left(\Delta^3\right)}{(\Delta x)_{j+1/2} + (\Delta x)_{j-1/2}}\right] \\ =& f_j^1 + \frac{1}{2!} f_j^2 \left[\frac{(\Delta x)_{j+1/2}^2 - (\Delta x)_{j-1/2}^2}{(\Delta x)_{j+1/2} + (\Delta x)_{j-1/2}}\right] + \mathscr{O}\left(\Delta^2\right) \\ =& f_j^1 + \frac{1}{2!} f_j^2 \left[(\Delta x)_{j+1/2} - (\Delta x)_{j-1/2}\right] + \mathscr{O}\left(\Delta^2\right) \ .\end{aligned} \tag{3.49}$$

This shows that there is indeed a "first-order term" in the error, as expected, but notice that it is proportional to $\left[(\Delta x)_{j+1/2} - (\Delta x)_{j-1/2}\right]$, which is the difference in the grid spacing between neighboring points, i.e., it is a "difference of differences." If the grid spacing varies slowly enough, the first-order part of the error can be smaller than the second-order part. For example, suppose that $(\Delta x)_{j+1/2} = (\Delta x)_{j-1/2}(1+\alpha)$, where $\alpha \ll 1$. Then

$$\left[(\Delta x)_{j+1/2} - (\Delta x)_{j-1/2}\right] = \alpha (\Delta x)_{j-1/2} \ . \tag{3.50}$$

3.3.3 A further generalization to higher-order derivatives

Next, we observe that (3.33) can be generalized to derive approximations to higher-order derivatives of f. For example, to derive approximations to the second derivative, f_j^2, on a (possibly) non-uniform grid, we write

$$\left(\frac{d^2 f}{dx^2}\right)_{j,\text{ approx}} = \sum_{j'=-\infty}^{\infty} c_{j,j'} f\left(x_{j'}\right) \ . \tag{3.51}$$

Obviously, it is going to turn out that $c_{j,j'} \sim 1/(\Delta x)^2$. Substitution of (3.34) into (3.51) gives

$$\left(\frac{d^2f}{dx^2}\right)_{j,\,\text{approx}} = \sum_{j'=-\infty}^{\infty} c_{j,j'} \left[f_j^0 + f_j^1 (\Delta x)_{j,j'} + f_j^2 \frac{(\Delta x)_{j,j'}^2}{2!} + f_j^3 \frac{(\Delta x)_{j,j'}^3}{3!} + \cdots \right] . \quad (3.52)$$

A first-order accurate approximation to the second derivative is ensured if we enforce the *three* conditions

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} = 0 \, , \quad \sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'} = 0 \, , \text{ and } \sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^2 = 2! \, , \text{ for all } j \, . \quad (3.53)$$

To achieve a second-order accurate approximation to the second derivative, we must also require that

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^3 = 0 \, , \text{ for all } j. \quad (3.54)$$

In general, to have an nth-order accurate approximation to the second derivative, we must require that

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^m = (2!) \, \delta_{m,2} \text{ for all } j, \text{ and for } 0 \le m \le n+1 \, . \quad (3.55)$$

We thus have to satisfy $n+2$ equations to obtain an nth-order accurate approximation to the second derivative, whereas we had to satisfy only $n+1$ equations to obtain an nth-order accurate approximation to the first derivative.

Earlier we showed that, in general, a 1D second-order approximation to the first derivative must involve a minimum of three grid points, because three conditions must be satisfied [i.e., (3.39) and (3.40)]. Now we see that, in general, a 1D second-order approximation to the second derivative must involve four grid points, because four conditions must be satisfied, i.e., (3.53) and (3.54). Five points may be preferable to four, from the point of view of symmetry. In the special case of a uniform grid, three points suffice.

At this point, you should be able to see ("by induction") that on a (possibly) non-uniform grid, an nth-order accurate approximation to the lth derivative of f takes the form

$$\left(\frac{d^l f}{dx^l}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} d_{j,j'} f\left(x_{j'}\right) , \tag{3.56}$$

where the coefficients $d_{j,j'}$ satisfy the $n+l$ equations

$$\sum_{j'=-\infty}^{\infty} (\Delta x)^m_{j,j'} d_{j,j'} = (l!)\,\delta_{m,l} \text{ for } 0 \le m \le n+l-1 . \tag{3.57}$$

In general, to satisfy $n+l$ requirements a minimum of $n+l$ points will be needed. You could write a computer program that would automatically generate the coefficients for a compact nth-order-accurate approximation to the lth derivative of f, using $n+l$ points on a nonuniform grid of your choice.

What is the meaning of (3.56) – (3.57) when $l = 0$? Recall that $0! = 1$.

3.3.4 Extension to two dimensions

The approach presented above can be further generalized to multi-dimensional problems. We will illustrate this using the two-dimensional Laplacian operator. The Laplacian appears, for example, in the diffusion equation with a constant diffusion coefficient, which is

$$\frac{\partial f}{\partial t} = K\nabla^2 f , \tag{3.58}$$

where t is time and K is a constant positive diffusion coefficient. Chapter 16 is entirely devoted to solving equations similar to (3.58).

Earlier in this chapter, we discussed one-dimensional differences that could represent either space or time differences, but the following discussion of the Laplacian is unambiguously about space differencing.

Consider a fairly general finite-difference approximation to the Laplacian, of the form

$$\left(\nabla^2 f\right)_{j,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} e_{j,j'} f_{j'} . \tag{3.59}$$

Here we use one-dimensional indices even though we are on a two-dimensional grid. The grid is not necessarily rectangular, and can be non-uniform. The subscript j is the "*name*" of a particular grid point ("home base" for this calculation), whose coordinates are $\left(x_j, y_j\right)$. Similarly, the subscript j' is the name of a grid point *in the neighborhood of point* j, and possibly including j itself, whose coordinates are $\left(x_{j'}, y_{j'}\right)$. In practice, a method is needed to identify (and record for later use) the stencil of grid points in the neighborhood of each j. In other words, for each point j we need to create an array that identifies the neighbors of j that will be used to compute the Laplacian at j.

The *two-dimensional* Taylor series can be written, using Cartesian coordinates, as

$$\begin{aligned} f_{j'} = f_j &+ \left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y} \right] f \\ &+ \frac{1}{2!} \left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y} \right]^2 f \\ &+ \frac{1}{3!} \left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y} \right]^3 f \\ &+ \frac{1}{4!} \left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y} \right]^4 f + \cdots . \end{aligned} \tag{3.60}$$

This can be expanded in gruesome detail as

$$\begin{aligned} f_{j'} \quad &= f_j \\ &+ \left[(\Delta x)_{j,j'} f_x + (\Delta y)_{j,j'} f_y \right] \\ &+ \tfrac{1}{2!} \left[(\Delta x)^2_{j,j'} f_{xx} + 2(\Delta x)_{j,j'} (\Delta y)_{j,j'} f_{xy} + (\Delta y)^2_{j,j'} f_{yy} \right] \\ &+ \tfrac{1}{3!} \left[(\Delta x)^3_{j,j'} f_{xxx} + 3 (\Delta x)^2_{j,j'} (\Delta y)_{j,j'} f_{xxy} + 3(\Delta x)_{j,j'} (\Delta y)^2_{j,j'} f_{xyy} + (\Delta y)^3_{j,j'} f_{yyy} \right] \\ &+ \tfrac{1}{4!} \left[(\Delta x)^4_{j,j'} f_{xxxx} + 4 (\Delta x)^3_{j,j'} (\Delta y)_{j,j'} f_{xxxy} + 6 (\Delta x)^2_{j,j'} (\Delta y)^2_{j,j'} f_{xxyy} \right. \\ &\left. + 4 (\Delta x)_{j,j'} (\Delta y)^3_{j,j'} f_{xyyy} + (\Delta y)^4_{j,j'} f_{yyyy} \right] + \cdots \end{aligned} \tag{3.61}$$

Here we use the notation

$$(\Delta x)_{j,j'} \equiv x_{j'} - x_j \text{ and } (\Delta y)_{j,j'} \equiv y_{j'} - y_j \, , \tag{3.62}$$

and it is understood that all of the derivatives are evaluated at the point j. Notice the "cross terms" that involve products of $(\Delta x)_{j,j'}$ and $(\Delta y)_{j,j'}$, and the corresponding cross-derivatives. A more general form of (3.61) is

$$f(\mathbf{r}+\mathbf{a}) = f(\mathbf{r}) + \sum_{n=1}^{\infty} \frac{1}{n!} (\mathbf{a}\cdot\nabla)^n f(\mathbf{r}) \ , \tag{3.63}$$

where $\mathbf{r}$ is a position vector, and $\mathbf{a}$ is a displacement vector (Arfken, 1985, p. 309). In (3.63), the operator $(\mathbf{a}\cdot\nabla)^n$ acts on the function $f(\mathbf{r})$. You should confirm for yourself that the general form (3.63) is consistent with (3.61). Eq. (3.63) has the advantage that it does not make use of any particular coordinate system. Because of this, it can be used to work out the series expansion using *any* coordinate system, e.g., Cartesian coordinates or spherical coordinates or cylindrical coordinates, by using the form of ∇ in that coordinate system.

Substituting from (3.61) into (3.59), we find that

$$\begin{aligned}
&(f_{xx}+f_{yy})_{j,\text{approx}} \cong \\
&\sum_{j'=-\infty}^{\infty} e_{j,\,j'} \Big\{ f_j + \Big[(\Delta x)_{j,j'} f_x + (\Delta y)_{j,j'} f_y \Big] \\
&\quad + \frac{1}{2!} \Big[(\Delta x)^2_{j,j'} f_{xx} + 2(\Delta x)_{j,j'} (\Delta y)_{j,j'} f_{xy} + (\Delta y)^2_{j,j'} f_{yy} \Big] \\
&\quad + \frac{1}{3!} \Big[(\Delta x)^3_{j,j'} f_{xxx} + 3(\Delta x)^2_{j,j'} (\Delta y)_{j,j'} f_{xxy} + 3(\Delta x)_{j,j'} (\Delta y)^2_{j,j'} f_{xyy} + (\Delta y)^3_{j,j'} f_{yyy} \Big] \\
&\quad + \frac{1}{4!} \Big[(\Delta x)^4_{j,j'} f_{xxxx} + 4(\Delta x)^3_{j,j'} (\Delta y)_{j,j'} f_{xxxy} + 6(\Delta x)^2_{j,j'} (\Delta y)^2_{j,j'} f_{xxyy} \\
&\quad + 4(\Delta x)_{j,j'} (\Delta y)^3_{j,j'} f_{xyyy} + (\Delta y)^4_{j,j'} f_{yyyy} \Big] + \cdots \Big\} \ .
\end{aligned} \tag{3.64}$$

Note that we have expressed the Laplacian on the left-hand side of (3.64) in terms of Cartesian coordinates. The reason for this is that we are going to use the special case of Cartesian coordinates $(x,\, y)$ as an example, and in the process we are going to to "match up terms" on the left and right sides of (3.64). *The use of Cartesian coordinates in* (3.64) *does not limit its applicability to Cartesian grids*. In other words, we can use a Cartesian coordinate system to analyze the scheme even if we are not going to code the model up using a Cartesian grid. Eq. (3.64) can be used to analyze the truncation errors of a finite-difference Laplacian on *any planar grid*, regardless of how the grid points are distributed.

To have first-order accuracy, we need

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'} = 0, \text{ for all } j, \tag{3.65}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)_{j,j'} = 0, \text{ for all } j, \tag{3.66}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta y)_{j,j'} = 0, \text{ for all } j, \text{ and} \tag{3.67}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)^2_{j,j'} = 2!, \text{ for all } j, \tag{3.68}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)_{j,j'}(\Delta y)_{j,j'} = 0, \text{ for all } j, \tag{3.69}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta y)^2_{j,j'} = 2!, \text{ for all } j. \tag{3.70}$$

Note the 2!s on the right-hand side of (3.68) and (3.70). Inspection of Eqs. (3.68) – (3.70) shows that, as expected, $e_{j,j'}$ is of order Δ^{-2}, where Δ is a shorthand for "Δx or Δy." Therefore, the quantities inside the sums in (3.66)-(3.67) are of order Δ^{-1}, and those inside the sums in (3.68)-(3.70) are of order one. This is why (3.68)-(3.70) are needed, in addition to (3.65)-(3.67), in order to obtain first-order accuracy.

Eq. (3.65) implies (with (3.59)) that a constant field has a Laplacian of zero, as it should. That's nice.

So far, (3.65)-(3.70) involve six equations. This means that to ensure first-order accuracy on a possibly nonuniform two-dimensional grid, six grid points are needed. There are exceptions to this rule. If we are fortunate enough to be working on a highly symmetrical

grid, it is possible that the conditions for second-order accuracy can be satisfied with a smaller number of points. For example, if we satisfy (3.65)-(3.70) on a square grid, we will get second-order accuracy "for free," and, as you will show when you do the homework, it can be done with only five points. More generally, with a non-uniform grid, we must also satisfy the following four additional conditions to achieve second-order accuracy:

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)^3_{j,j'} = 0, \text{ for all } j, \tag{3.71}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)^2_{j,j'}(\Delta y)_{j,j'} = 0, \text{ for all } j, \tag{3.72}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta x)_{j,j'}(\Delta y)^2_{j,j'} = 0, \text{ for all } j, \tag{3.73}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,\,j'}(\Delta y)^3_{j,j'} = 0, \text{ for all } j. \tag{3.74}$$

In general, a total of ten conditions, i.e., Eqs. (3.65)-(3.74), must be satisfied to ensure second-order accuracy for the Laplacian on a non-uniform two-dimensional grid.

If we were working in more than two dimensions, we would simply replace (3.61) by the appropriate multi-dimensional Taylor series expansion. The rest of the argument would be parallel to that given above, although of course the requirements for second-order accuracy would be more numerous.

3.4 More about the Laplacian

3.4.1 Approximations to the Laplacian on rectangular grids

Consider a 3x3 block of nine-points on a rectangular grid, as shown in Fig. 3.3. Because the grid has a high degree of symmetry, is possible to obtain second-order accuracy with just five points, and in fact this can be done in two different ways, corresponding to the two five-point stencils shown by the grey boxes in the figure. Based on their shapes, one of the stencils can be called "+", and the other one "x". We assume a grid spacing d in both the x

and y directions (i.e., square grid cells), and use a two-dimensional indexing system, with counters i and j in the x and y directions, respectively.

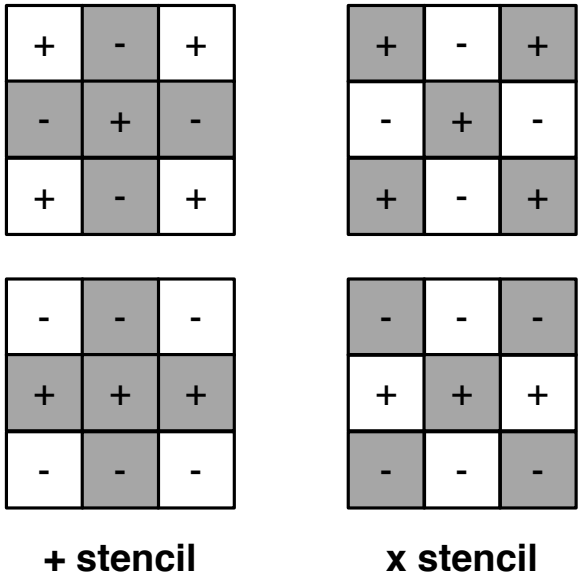


Figure 3.3: The grey shading shows two five-point stencils that can be used to create second-order Laplacians on a rectangular grid. In the upper two panels, the plus and minus symbols represent an input function that has the form of a checkerboard. In the lower two panels the plus and minus symbols represent an input function that has the form of a "rotated checkerboard," which can also be viewed as a set of horizontal stripes. (Similarly, the checkerboard in the top two panels can be viewed as "diagonal stripes.")

Using the methods explained above, it can be shown that the second-order finite-difference Laplacians are given by

$$\left(\nabla^2 f\right)_{i,\,j,\text{approx}} \cong \frac{f_{i,\,j+1} + f_{i-1,j} + f_{i,\,j-1} + f_{i+1,j} - 4f_{i,\,j}}{d^2} \text{ with the + stencil,} \tag{3.75}$$

and

$$\left(\nabla^2 f\right)_{i,\,j,\text{approx}} \cong \frac{f_{i+1,j+1} + f_{i-1,j+1} + f_{i-1,j-1} + f_{i+1,j-1} - 4f_{i,\,j}}{\left(\sqrt{2}d\right)^2} \text{ with the x stencil.} \tag{3.76}$$

Inspection of Figure 3.3 shows that *the Laplacian based on the x stencil cannot "see" a checkerboard pattern in the function represented on the grid, as shown by the plus and*

minus symbols in the top two panels. What I mean by this is that the scheme returns a zero for the Laplacian even when a checkerboard is present. A diffusion equation that uses this form of the Laplacian cannot smooth out a checkerboard, because it doesn't know that the checkerboard is there. That's bad.

The + stencil does have some issues, however. It under-estimates the strength of the "rotated checkerboard" (which can also be called "stripes") shown in the bottom two panels of Fig. 3.3, while the x stencil feels it more strongly.

A more general second-order Laplacian uses all nine points, and can be obtained by writing a weighted sum of the two Laplacians given by (3.75) and (3.76). What principle would you suggest for choosing the values of the weights? The nine-point stencil will involve more arithmetic than either of the two five-point stencils, but the benefit may justify the cost in this case.

3.4.2 Integral properties of the Laplacian

Here comes a digression, which deals with something other than the order of accuracy of finite-difference schemes.

For the continuous Laplacian on a periodic domain or a closed domain with zero normal derivatives on its boundary, we can prove the following two important properties:

$$\int_A \left(\nabla^2 f\right)\, dA = 0\,, \tag{3.77}$$

$$\int_A f\left(\nabla^2 f\right)\, dA \leq 0\,. \tag{3.78}$$

Here the integrals are with respect to area, over the entire domain; volume integrals can be used in the same way for the case of three dimensions. Eqs. (3.77) and (3.78) hold for any sufficiently differentiable function f. For the diffusion equation, (3.58), Eq. (3.77) implies that diffusion does not change the area-averaged value of f, and the inequality (3.78) implies that diffusion reduces the area-average of the square of f.

The finite-difference requirements corresponding to (3.77) and (3.78) are

$$\sum_{\text{all } j}\left[\left(\nabla^2 f\right)_{j,\text{approx}} A_j\right] \cong \sum_{\text{all } j}\left[\left(\sum_{j'=-\infty}^{\infty} e_{j,\,j'} f_{j'}\right) A_j\right] = 0\,, \tag{3.79}$$

and

$$\sum_{\text{all } j}\left[f_j\left(\nabla^2 f\right)_{j,\text{approx}}A_j\right] \cong \sum_{\text{all } j}\left[f_j\left(\sum_{j'=-\infty}^{\infty} e_{j,\,j'}f_{j'}\right)A_j\right] \leq 0\,, \tag{3.80}$$

where A_j is the area of grid-cell j, and the $e_{j,j'}$ are the weights introduced in Eq. (3.59). Suppose that we want to satisfy (3.79)-(3.80) *regardless of the numerical values assigned to* the various $f\left(x_j,y_j\right)$. This may sound impossible, but it can be done by suitable choice of the $e_{j,j'}$.

As an example, consider what is needed to ensure that (3.79) will be satisfied for an arbitrary distribution of $f\left(x_j,y_j\right)$. Each value of $f\left(x_j,y_j\right)$ will appear more than once in the double sum after the first equals sign in (3.79). We can "collect the coefficients" of each value of f, and require that the sum of the coefficients is zero:

$$\begin{aligned}\sum_{\text{all } j}A_j\left[\sum_{j'=-\infty}^{\infty} e_{j,\,j'}f_{j'}\right] &= \sum_{\text{all } j'}\left[\sum_{\text{all } j} e_{j,\,j'}f_{j'}A_j\right]\\ &= \sum_{\text{all } j'}\left[f_{j'}\left(\sum_{\text{all } j} e_{j,\,j'}A_j\right)\right]\\ &= 0\,.\end{aligned} \tag{3.81}$$

The first step, shown on the right-hand side of the top line of (3.81), is to change the order of summation. This allows us to pull $f_{j'}$ out of the inner sum, as shown on the second line. The last equality above is simply a re-statement of the requirement (3.79). The only way to satisfy the requirement for an arbitrary distribution of $f\left(x_j,y_j\right)$ is to write

$$\sum_{\text{all } j} e_{j,\,j'}A_j = 0 \text{ for each } j'\,. \tag{3.82}$$

If the $e_{j,\,j'}$ satisfy (3.82) then (3.79) will also be satisfied.

Similar (but more complicated) ideas were used by Arakawa (1966), in the context of energy and enstrophy conservation with a finite-difference vorticity equation. This will be discussed in Chapter 26.

A finite-difference scheme with a property similar to (3.78) is discussed in Chapter 16.

3.5 Segue

This chapter demonstrates that it is straightforward to design finite-difference schemes to approximate a derivative of any order, with any desired order of accuracy, on irregular grids of any shape, and in multiple dimensions. It's a done deal.

The schemes can also be designed to satisfy rules based on properties of the differential operators that they approximate; for example, we showed that it is possible to guarantee that the area-average of a two-dimensional finite-difference Laplacian vanishes on a periodic domain.

Finally, we pointed out that some finite-difference schemes suffer from an inability to recognize small-scale, "noisy" modes on the grid, such as checkerboard patterns.

This is a lot, but we have not yet gotten to the main subject of this book, which is how to find approximate numerical solutions to differential equations. Chapter 5 is the first part of that story. But before that, we briefly consider grids that are not rectangular.

3.6 Problems

1. Prove that a finite-difference scheme with errors of order n gives exact first derivatives for polynomial functions of degree n or less. For example, a first-order scheme gives exact first derivatives for linear functions.

2. Choose a simple differentiable function $f(x)$ that is not a polynomial. Find the exact numerical value of df/dx for a particular value of x, say $x = x_1$. Then choose

   (a) a first-order scheme, and

   (b) a second-order scheme

   to approximate $(df/dx)_{x=x_1}$. Find the total error by subtracting the exact derivative from the approximate derivative. For each case, plot the log of the absolute value of the total error of these approximations as a function of $\ln(\Delta x)$. By inspection of the plot, verify that for sufficiently small Δx the logs of the absolute errors of the schemes decrease along almost straight lines, with the expected slopes (which you should estimate from the plots), as Δx decreases. For sufficiently large values of Δx, the logs of the absolute errors will depart from straight lines, and (depending on the function you choose) the second-order scheme may even give errors larger than the first-order scheme. Extend your plot to include values of Δx that are large enough to show the departures from straight lines.

3. Using Cartesian coordinates (x, y), the Laplacian can be written as $\nabla^2 f = f_{xx} + f_{yy}$. Consider a second Cartesian coordinate system, (x', y'), that is rotated, with respect to the first, by an angle θ. Prove by direct calculation that $f_{xx} + f_{yy} = f_{x'x'} + f_{y'y'}$. This demonstrates that, for this example, the Laplacian is invariant with respect to

rotations of a Cartesian coordinate system. In fact, it can be shown that the Laplacian takes the same numerical value no matter what coordinate system is used. The meaning of the Laplacian is independent of coordinate system, and the Laplacian can be defined without using a coordinate system. See Appendix A for further discussion.

4. Both the + and x stencils for the Laplacian allow second-order accuracy on a uniform square grid. Is it possible to combine the + and x stencils with weights so as to achieve fourth-order accuracy? Prove your answer.

5. Consider a two-dimensional Cartesian grid in which both Δx and Δy are spatially uniform, but $\Delta x \neq \Delta y$. Find the simplest second-order accurate scheme for the Laplacian.

6. Prove Eq. (3.78).

Chapter 4

Why be square?

4.1 Tiling the plane

As is well known, only three convex regular polygons tile the plane: equilateral triangles, squares, and hexagons. A proof is as follows: Suppose that a regular convex polygon has N sides, as illustrated in Figure 4.1 for the case of hexagons. The interior angles, α, must sum to 2π, so

$$N\alpha = 2\pi \, . \tag{4.1}$$

Next, suppose that M polygons come together at each vertex. The exterior angles, β, must sum to 2π, so

$$M\beta = 2\pi \, . \tag{4.2}$$

Since the interior angles of a triangle must sum to π, we can write

$$\alpha + 2\,(\beta/2) = \pi \, ,$$

or

$$\alpha + \beta = \pi \, . \tag{4.3}$$

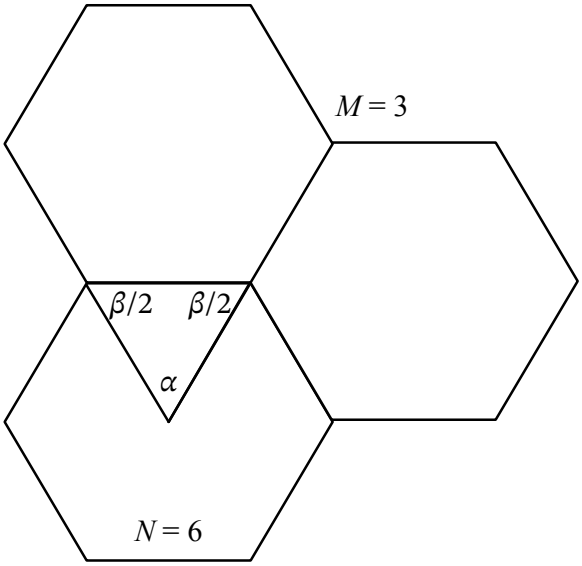


Figure 4.1: A diagram illustrating tiling the plane with convex regular polygons for the case of a hexagon, with $N = 6$ and $M = 3$. The interior angles of the hexagon are called α, and the exterior angles are called β.

Combining (4.1) - (4.3), we find that

$$M = \frac{2N}{N-2} \, . \tag{4.4}$$

Trying different values of N in (4.4), we find that M is an integer only for $N = 3$, 4, or 6. This completes the proof.

4.2 Symmetries

Fig. 4.2 shows portions of planar grids made from equilateral triangles, squares, and hexagons. On the triangular grid and the square grid, some of the neighbors of a given cell lie directly across cell walls, while others lie across cell vertices. As a result, finite-difference operators constructed on these grids tend to use "wall neighbors" and "vertex neighbors" in different ways. For example, the simplest second-order finite-difference approximation to the gradient, on a square grid, uses only "wall neighbors;" vertex neighbors are ignored. It is certainly possible to construct finite-difference operators on square grids (and triangular grids) in which information from all nearest-neighbor cells is used. An example is the Arakawa Jacobian, as discussed by Arakawa (1966) and in Chapter 26. The essential anisotropies of these grids remain, however, and are inevitably manifested in the forms of the finite-difference operators.

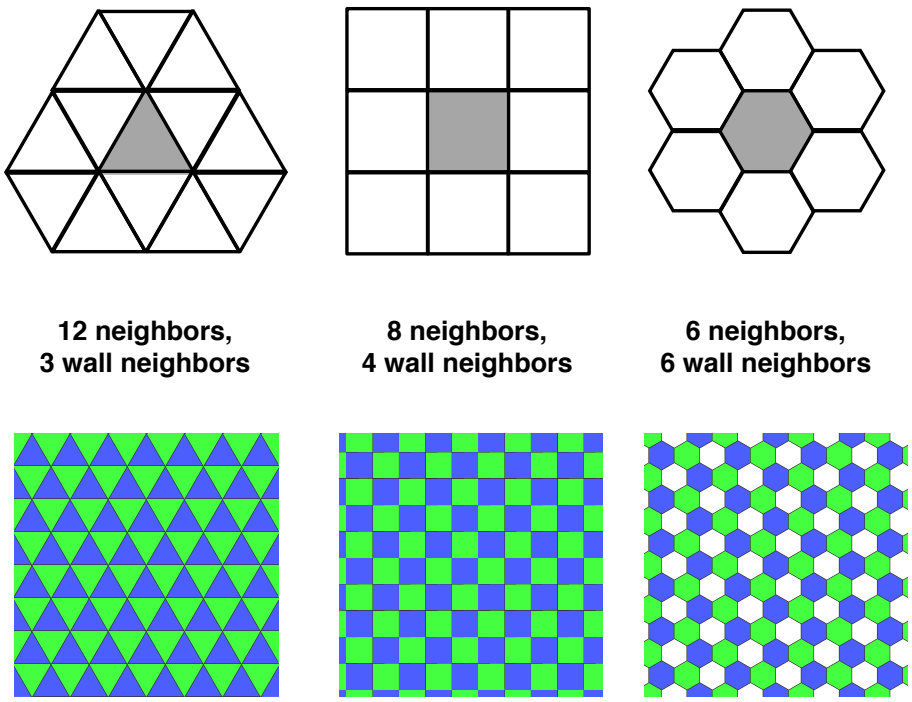


Figure 4.2: The upper row of figures shows small sections of grids made up of equilateral triangles (left), squares (center), and hexagons (right). These are the only regular polygons that tile the plane. The hexagonal grid has the highest symmetry. For example, all neighboring cells of a given hexagonal cell are located across cell walls. In contrast, with either triangles or squares, some neighbors are across walls, while others are across corners. The lower row of figures shows the "checkerboards" associated with each grid. The triangular and quadrilateral checkerboards have two colors, while the hexagonal checkerboard has three colors.

In contrast, hexagonal grids have the property that *all* neighbors of a given cell lie across cell walls; there are no "vertex neighbors." In this sense, hexagonal grids are quasi-isotropic. As a result, the most natural finite-difference Laplacians on hexagonal grids treat all neighboring cells in the same way; they are as symmetrical and isotropic as possible.

4.3 Problems

1. For this problem, use equations (3.65)-(3.74), which state the ten requirements for second-order accuracy of a finite-difference approximation to the Laplacian.

   (a) Consider a plane filled with perfectly hexagonal cells. The dependent variable is defined at the center of each hexagon. *Find a second-order accurate scheme for the Laplacian that uses just the central cell and its six closest neighbors, i.e., just seven cells in total.* To standardize the notation, let d be the distance

between grid-cell centers, as measured across cell walls. Write your scheme in terms of d, as was done in (3.75) and (3.76). Important hint: You can drastically simplify the problem by creatively taking advantage of the grid's high degree of symmetry.

(b) Consider the "checkerboard" on the the hexagonal grid, as shown in Fig. 4.2. Let the three numerical values on the checkerboard be -1, 0, and +1. Does your scheme give a non-zero Laplacian for the checkerboard?

2. Consider a grid of equilateral triangles. Prove that, despite the grid's symmetry, it is not possible to create a second-order accurate scheme for the Laplacian by using a stencil that consists of just four cells, namely "home base" and the three surrounding cells that lie *across the walls* of home base.

3. (a) Invent a way to "index" the points on a hexagonal grid with periodic boundary conditions. As a starting point, I suggest that you make an integer array like this: NEIGHBORS(I,J). The first subscript would designate which point on the grid is "home base" i.e., it would be a one-dimensional counter that covers the entire grid. The second subscript would range from 0 to 6 (or, alternatively, from 1 to 7). The smallest value (0 or 1) would designate the central point, and the remaining 6 would designate its six surrounding neighbors. The indexing problem then reduces to generating the array NEIGHBORS(I,J), which need only be done once for a given grid.

(b) Set up a hexagonal grid to represent a square domain with periodic boundary conditions. It is not possible to create an *exactly* square domain using a hexagonal grid, but you can set up a domain that is *approximately* square, with 100 points in one direction and the appropriate number of points (you get to figure it out) in the other direction. The total number of points in the approximately square domain will be very roughly 8000. This (approximately) square domain has periodic boundary conditions, so it actually represents one "patch" of an infinite domain. Because the domain is not exactly square, the period in the x-direction cannot be exactly the same as the period in the y-direction, but it can be fairly close. Make sure that the boundary conditions are truly periodic on your grid, so that no discontinuities occur. Fig. 4.3 can help you to understand how to define the computational domain and how to implement the periodic boundary conditions.

(c) Using the tools created under parts a) and b) above, write a program to compute the Laplacian. Choose a continuous doubly periodic test function. Plot the test function to confirm that it is really doubly periodic on your grid. Also attach plots of both your approximate Laplacian and the exact Laplacian.

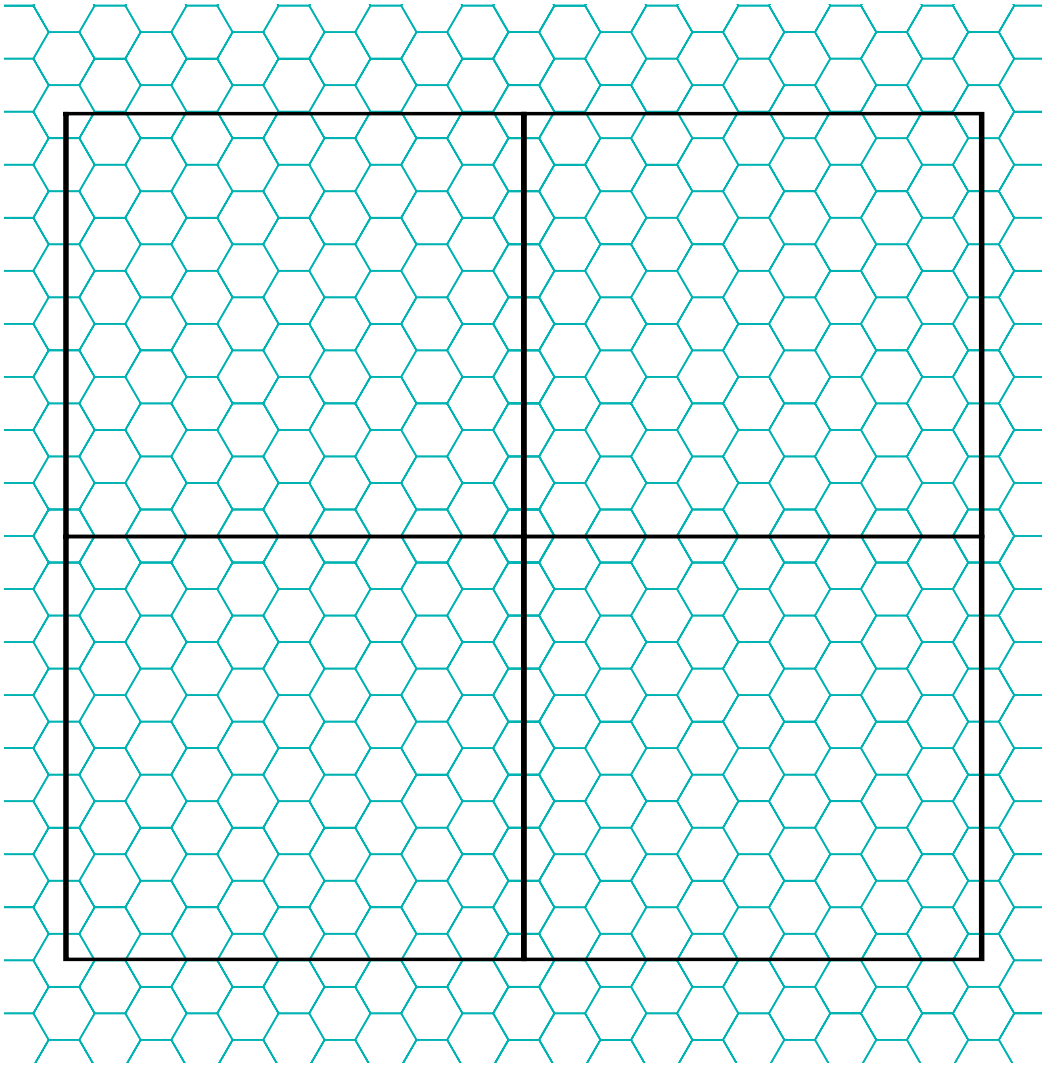


Figure 4.3: Sketch showing four copies of a periodic, nearly square domain on a (low-resolution) hexagonal grid. All four corners of each domain are located in the centers of hexagonal cells. The vertical edges of each domain run through vertical stacks of cells. Obviously the picture could be rotated by 90° (or any other angle) without changing the idea.

Chapter 5

Some time-differencing schemes

5.1 Introduction

We have already analyzed the accuracy of finite-difference quotients. Now we analyze the accuracy of *finite-difference schemes*, which are defined as finite-difference equations that approximate, term-by-term, differential equations. The solution of an accurate scheme is a useful approximation to the solution of the differential equation.

Using the methods presented in Chapter 3, we can find approximations to each term of a differential equation, and we have already seen that the error of such an approximation can be made as small as desired, almost effortlessly. This is not our goal, however. *Our goal is to find an approximation to the solution of the differential equation*. You might think that if we have a finite-difference equation, F, that is constructed by writing down a good approximation to each term of a differential equation, D, then the solution of F will "automatically" be a useful approximation to the solution of D. Wouldn't that be nice? Unfortunately, it's not true.

In this chapter, we deliberately side-step the complexities of space differencing so that we can focus on the problem of time differencing in isolation. Consider an arbitrary first-order ordinary differential equation of the form:

$$\frac{dq}{dt} = f\left[q(t),\, t\right]. \tag{5.1}$$

The example below may help to make it clear how it is possible and what it means for f to depend on both $q(t)$ and t:

$$\frac{dq}{dt} = -\kappa q + a\sin\left(\omega t\right). \tag{5.2}$$

Here the first term on the right-hand side represents decay towards zero (assuming that κ is positive), and the second term represents a temporally periodic external forcing (e.g., from the sun) that can drive q away from zero.

In this chapter we do not specify $f[q(t),\, t]$, so the discussion is "generic." Starting in Chapter 6 we will consider particular choices of $f[q(t),\, t]$. Keep in mind that, in practice, $f[q(t),\, t]$ could be very complicated.

Let $n\Delta t$ be the current time, and $(n+1)\,\Delta t$ be the future time for which we want to make a prediction by taking a time step. Suppose that we integrate (5.1) with respect to time, from $(n-m)\,\Delta t$ to $(n+1)\,\Delta t$. Here we assume that m is either zero or a positive integer. Fig. 5.1 shows the meaning of m in a graphical way. For larger values of m, the domain of integration "reaches further back" in time, before the current time, $t = n\Delta t$. We assume for now that $n \geq m$, so that $q[(n-m)\,\Delta t]$ is known. This means that the initial time, for which $n = 0$, is before $t = (n-m)\,\Delta t$. This will be the case except close to the initial condition; further discussion is given later. Integration of (5.1) gives

$$q[(n+1)\,\Delta t] - q[(n-m)\,\Delta t] = \int_{(n-m)\Delta t}^{(n+1)\Delta t} f(q,\, t)\; dt\;. \tag{5.3}$$

Equation (5.3) is still "exact;" no finite-difference approximations have been introduced.

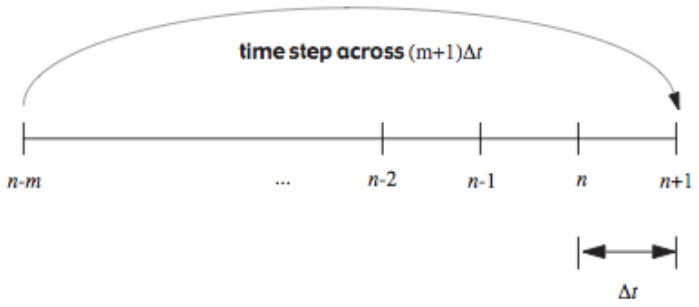


Figure 5.1: In Eq. (5.4), we use a weighted combination of f^{n+1}, f^{n}, f^{n-1}, ... f^{n-l} to compute an average value of $f \equiv dq/dt$ over the time interval $(1+m)\,\Delta t$.

5.2 A family of schemes

With a finite-difference-scheme, q and f are defined only at discrete time levels. We will use the symbol q^{n+1} as a shorthand for $q[(n+1)\Delta t]$, f^{n+1} in place of $f\{q[(n+1)\,\Delta t],\ (n+1)\Delta t\}$, and so on.

The integral on the right-hand side of (5.3) can be approximated using the values of f at the discrete time levels. Eq. (5.3), divided by $(1+m)\,\Delta t$, can be approximated by

$$\frac{q^{n+1}-q^{n-m}}{(1+m)\,\Delta t} \cong \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1} + \alpha_{n-2} f^{n-2} + \ldots \alpha_{n-l} f^{n-l}, \qquad (5.4)$$

where l can be zero or a positive integer. *A family of schemes is defined by* (5.4). In Chapter 3, we discussed families of schemes that can be used to approximate finite-difference operators.

Take a minute to look carefully at the form of (5.4), which is a slightly modified version of an equation discussed by Baer and Simons (1970). The left-hand side is a "time step" across a time interval of $(1+m)\,\Delta t$, as illustrated in Fig. 5.1. The right-hand side consists of a weighted sum of instances of the function f, evaluated at various time levels. The first time level, $n+1$, is in "the future." The second, n, is in "the present." All of the remaining time levels are in "the past." Time level $n-l$ is furthest back in the past; this is essentially the definition of l. We assume that $n \geq l$.

In designing the scheme, we get to choose the values of l and m. It is possible to have $l > m$ or $l = m$ or $l < m$. Viable schemes can be constructed with all three possibilities, and examples are given below.

Having chosen l and m, we next get to choose the numerical values of $l+2$ parameters, namely, β, and the $l+1$ values of α.

Here is some important and widely used terminology: A scheme is called "*implicit*" if it has $\beta \neq 0$, and "explicit" if it has $\beta = 0$. As discussed later, implicit schemes have very nice properties for some important choices of f. On the other hand, implicit schemes can be complicated because the "unknown" or "future" value of q, namely q^{n+1}, appears on the right-hand-side of the equation, as an argument of the possibly complicated function f^{n+1}.

Not all time-differencing schemes fit into the family defined by (5.4). We will discuss some that don't later in this chapter.

5.3 Discretization error

We need a way to measure the error of the time-differencing schemes defined by (5.4). To start, let's define some notation. Let $q(t)$ denote the (exact) solution of the differential equation, so that $q(n\Delta t)$ is the value of the exact solution at time $n\Delta t$. We use the notation q^n to denote the "exact" solution of a finite-difference equation, at the same time. In general, $q^n \neq q(n\Delta t)$. We wish that they were equal!

Because q^n is defined only at discrete times, it is not differentiable, so we can't substitute q^n into the differential equation to see how well it works. We can, however, substitute

the *true solution*, $q(t)$, and the corresponding $f[q(t),\, t]$, into the finite-difference equation (5.4). Since $q(t)$ and $f[q(t),\, t]$ are differentiable, we can expand the various terms using Taylor series around $t = n\Delta t$. The result is

$$
\begin{aligned}
&\frac{1}{(1+m)\,\Delta t}\left\{\left[q+(\Delta t)\,q'+\frac{(\Delta t)^2}{2!}q''+\frac{(\Delta t)^3}{3!}q'''+\frac{(\Delta t)^4}{4!}q''''+\ldots\right]\right.\\
&\qquad\left.-\left[q-(m\Delta t)\,q'+\frac{(m\Delta t)^2}{2!}q''-\frac{(m\Delta t)^3}{3!}q'''+\frac{(m\Delta t)^4}{4!}q''''\ldots\right]\right\}\\
&=\beta\left[f+(\Delta t)\,f'+\frac{(\Delta t)^2}{2!}f''+\frac{(\Delta t)^3}{3!}f'''+\ldots\right]\\
&+\alpha_n f\\
&+\alpha_{n-1}\left[f-(\Delta t)\,f'+\frac{(\Delta t)^2}{2!}f''-\frac{(\Delta t)^3}{3!}f'''+\ldots\right]\\
&+\alpha_{n-2}\left[f-(2\Delta t)\,f'+\frac{(2\Delta t)^2}{2!}f''-\frac{(2\Delta t)^3}{3!}f'''+\ldots\right]\\
&+\alpha_{n-3}\left[f-(3\Delta t)\,f'+\frac{(3\Delta t)^2}{2!}f''-\frac{(3\Delta t)^3}{3!}f'''+\ldots\right]\\
&+\ldots\\
&+\alpha_{n-l}\left[f-(l\Delta t)\,f'+\frac{(l\Delta t)^2}{2!}f''-\frac{(l\Delta t)^3}{3!}f'''+\ldots\right]\\
&+\varepsilon,
\end{aligned}
\tag{5.5}
$$

where ε is called the *discretization error*[1] and a prime denotes a time derivative. If the true solution satisfied the finite-difference equation exactly, then we would have $\varepsilon = 0$. Using

$$
f = q',\; f' = q'',\; \text{etc.},
\tag{5.6}
$$

and collecting powers of Δt, we can rearrange (5.5) to obtain an expression for the discretization error:

[1] Recall that the error in the approximation of a derivative is called the truncation error. The error in the approximation of a differential equation is called the discretization error.

$$\begin{aligned}\varepsilon =& q'\left[1-(\beta+\alpha_n+\alpha_{n-1}+\alpha_{n-2}+\alpha_{n-3}+\cdots+\alpha_{n-l})\right]\\ &+\Delta t q''\left[\frac{1}{2}\left(\frac{1-m^2}{1+m}\right)-\beta+\alpha_{n-1}+2\alpha_{n-2}+3\alpha_{n-3}+\cdots+l\alpha_{n-l}\right]\\ &+\frac{(\Delta t)^2}{2!}q'''\left[\frac{1}{3}\left(\frac{1+m^3}{1+m}\right)-\beta-\alpha_{n-1}-4\alpha_{n-2}-9\alpha_{n-3}-\cdots-l^2\alpha_{n-l}\right]\\ &+\frac{(\Delta t)^3}{3!}q''''\left[\frac{1}{4}\left(\frac{1-m^4}{1+m}\right)-\beta+\alpha_{n-1}+8\alpha_{n-2}+27\alpha_{n-3}+\cdots+l^3\alpha_{n-1}\right]+\cdots\ .\end{aligned} \tag{5.7}$$

Because of the substitutions based on (5.6), the function f does not appear in (5.7). Each line on the right-hand side of (5.7) goes to zero "automatically" as $\Delta t \to 0$, *except for the first line*, which does not involve Δt at all. In order to ensure that the error decreases to zero as $\Delta t \to 0$, we have to choose the parameters of the scheme so that the first line is zero. This condition can be written as

$$1 = \beta+\alpha_n+\alpha_{n-1}+\alpha_{n-2}+\alpha_{n-3}+\ldots+\alpha_{n-l}\ . \tag{5.8}$$

Eq. (5.8) simply means that the sum of the coefficients on the right-hand side of (5.4) is equal to one, which means that we can interpret the right-hand side as an "average f." Eq. (5.8) ensures at least first-order accuracy. It is sometimes called the "*consistency condition*." A scheme that satisfies the consistency condition is said to be a consistent scheme.

When (5.8) is satisfied, the expression for the discretization error reduces to

$$\varepsilon = \Delta t q''\left[\frac{1}{2}\left(\frac{1-m^2}{1+m}\right)-\beta+\alpha_{n-1}+2\alpha_{n-2}+3\alpha_{n-3}+\ldots+l\alpha_{n-l}\right]+\mathscr{O}\left[(\Delta t)^2\right]\ . \tag{5.9}$$

Recall that we started with $l+3$ free parameters. After satisfying Eq. (5.8) *we still have* $l+2$ *degrees of freedom to work with*. In particular, we can require that the coefficient of Δt in (5.7) is also zero. This will give us a second-order scheme, i.e., one in which the error, ε, goes to zero like $(\Delta t)^2$. Having required second-order accuracy in this way, we still have $l+1$ degrees of freedom. Obviously this process can be continued, giving higher and higher accuracy, as long as the value of l is large enough. Examples are given below.

It is also possible to use the degrees of freedom to impose conditions other than increased Taylor-series accuracy. Examples are given in later chapters.

In summary, the order of accuracy of a time-differencing scheme based on (5.4) can be made at least as high as $l+3$ by appropriate choices of the coefficients. One of these coefficients is β. Recall that $\beta = 0$ (by definition) for explicit schemes. Generally, then, the accuracy of an explicit scheme can be made at least as high as $l+2$.

With the approach outlined above, schemes of higher-order accuracy (i.e., smaller discretization error) are made possible by bringing in more time levels. There are other ways to obtain schemes of higher accuracy; this will be discussed later.

In summary: *Truncation error*, which was discussed in Chapter 3, measures the accuracy of an approximation to a differential operator or operators. It is a measure of the accuracy with which a term of a differential equation has been approximated. In contrast, *discretization error* measures the accuracy with which the *solution* of the differential equation has been approximated. Reducing the truncation error to acceptable levels is usually easy. Reducing the discretization error can be much harder.

In the remainder of this chapter, we survey a number of time-differencing schemes, without (yet) specifying any particular form of f. In this analysis, we can determine the order of accuracy of each scheme, but we cannot yet test the numerical stability of the schemes, because the stability properties depend on the particular choice of f. In Chapter 6, we will consider two particular choices for f, and discuss the numerical stability of a few schemes for each choice.

5.4 Explicit schemes

In this section, we discuss some explicit schemes (i.e., with $\beta = 0$) that are members of the family discussed in section 5.2.

$m = 0$, $l = 0$ (The forward scheme or Euler scheme)

For this simple case we have $\alpha_n \neq 0$, but all of the other α's are zero. This is simply a choice, among many possible choices. The consistency condition, (5.8), forces us to choose $\alpha_n = 1$. The scheme given by (5.4) then reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = f^n \ . \tag{5.10}$$

The discretization error is $\Delta t q''/2 + \mathscr{O}\left(\Delta t^2\right) = \mathscr{O}\left(\Delta t\right)$. Therefore, the scheme has first-order accuracy.

$m = 0$, $l > 0$ *(Adams-Bashforth schemes)*

One way to get better accuracy is to increase l while keeping $m = 0$. For $l = 1$, the scheme given by (5.4) reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = \alpha_n f^n + \alpha_{n-1} f^{n-1} \ , \tag{5.11}$$

the consistency condition, (5.8), reduces to

$$\alpha_n + \alpha_{n-1} = 1 \ , \tag{5.12}$$

and the discretization error is

$$\varepsilon = \Delta t q'' \left(\alpha_{n-1} + \frac{1}{2} \right) + \mathscr{O}\left(\Delta t^2\right) \ . \tag{5.13}$$

If we choose $\alpha_{n-1} = -1/2$ and $\alpha_n = 3/2$, we get a scheme with second-order accuracy, which is called the second-order Adams-Bashforth scheme. It effectively uses the weights 3/2 and -1/2 to extrapolate from f^n and f^{n-1} to estimate an "effective" value of f at time level $n + 1/2$.

Although the right-hand side of (5.11) involves two different values of f, we only have to evaluate f once per time step, provided that we *save* one "old" value of f for use on the next time step. Additional memory has to be allocated to save the old value of f, but often this is not a problem. Note, however, that something special will have to be done on the first time step, because when $n = 0$ the time level $n - 1$ is "before the beginning" of the computation. This will be discussed later.

In a similar way, we can obtain more accurate Adams-Bashforth schemes by using larger values of l, and choosing the α's accordingly. Table 5.1 shows the results for $l = 1$, 2, and 3. The paper by Durran (1991) gives an interesting discussion of the third-order Adams-Bashforth scheme. We can think of the forward scheme as the "first-order Adams-Bashforth scheme," with $l = 0$.

$m = 1$, $l = 0$ *(The leapfrog scheme)*

l	α_n	α_{n-1}	α_{n-2}	α_{n-3}	discretization error
1	$3/2$	$-1/2$			$\mathscr{O}\left(\Delta t^2\right)$
2	$23/12$	$-4/3$	$5/12$		$\mathscr{O}\left(\Delta t^3\right)$
3	$55/24$	$-59/24$	$37/24$	$-9/24$	$\mathscr{O}\left(\Delta t^4\right)$

Table 5.1: Examples of Adams-Bashforth Schemes ($\beta = m = 0,\ l > 0$)

We can also get better accuracy by increasing m while keeping $l = 0$. The famous "leap-frog" scheme is given by

$$\frac{q^{n+1} - q^{n-1}}{2\Delta t} = f^n \,, \tag{5.14}$$

and illustrated in Fig. 5.2. From (5.7) we can immediately see that the discretization error is $\frac{\Delta t^2}{6} q''' + \mathscr{O}\left(\Delta t^4\right)$, so the leapfrog scheme is second-order accurate. This is higher than $l + 1 = 1$, i.e., it is better than would be expected from the general rule, stated earlier, for explicit schemes. The leapfrog scheme was widely used for decades, but it has fallen out of favor due to issues discussed below.

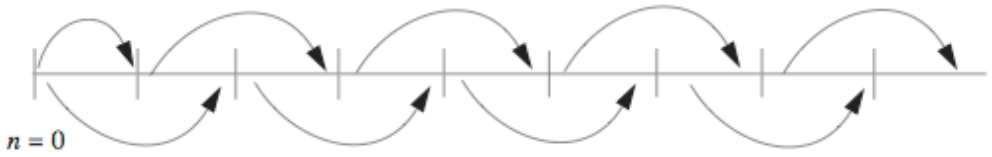


Figure 5.2: The leapfrog scheme.

A minor issue is that the leapfrog scheme is not "self-starting," because when we are sitting at $n = 0$ we don't know q at $n = -1$, so we can't step from $n = -1$ to $n = +1$. This means that a special procedure is needed for the first time step. *A similar problem arises with any scheme that involves more than two time levels.* We really need two initial conditions to solve the finite-difference equation, even though only one initial condition is needed to solve the differential equation. One is the "physical" initial condition that is also needed for the differential equation. The second initial condition arises because of the form of the finite-difference scheme itself, and has nothing to do with the physics. It can be called a "computational" initial condition.

As an example, consider the special case $f \equiv 0$. Then (5.14) gives

$$q^{n+1} - q^{n-1} = 0\,, \tag{5.15}$$

which is, of course, the right answer. The initial condition at $n = 0$ will determine the solution for all *even* values of n. To obtain the solution for *odd* values of n, we have to give a second initial condition at $n = 1$. Since we are analyzing the case $f \equiv 0$, the obvious thing to do is set $q^1 = q^0$. But suppose that we perversely make the two initial conditions different, i.e., set $q^1 \neq q^0$. Then an oscillation will occur, as shown in Fig. 5.3. This oscillatory solution is an example of a "computational mode in time." *Such computational modes in time can occur with all schemes that use more than two time levels*, including the already-discussed second- and third-order Adams-Bashforth schemes. A computational mode does not correspond to any solution of the continuous equation. Discretization error compares the solution of the scheme with the continuous solution. Because there is no continuous solution corresponding to a computational mode, the concept of discretization error does not apply to computational modes. Further discussion is given in Chapters 6 and 19.

As discussed later, computational modes can also arise from space differencing.

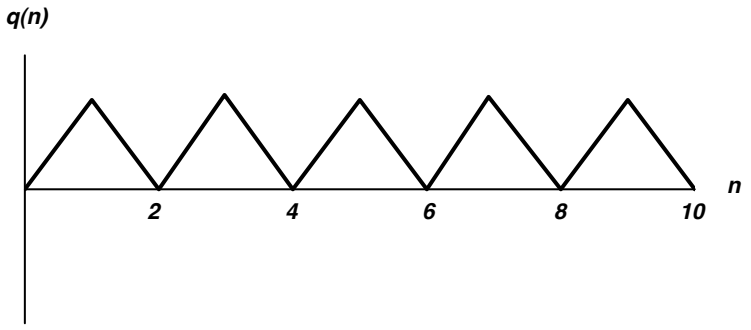


Figure 5.3: An oscillatory solution that arises with the leapfrog scheme for $dq/dt = 0$, in case the two initial values of q are not the same.

$\underline{m = 1,\, l = 1}$

Here there is no gain in accuracy. The highest accuracy (second order) is obtained for $\alpha_{n-1} = 0$, which gives the leapfrog scheme.

$\underline{m = 1,\, l > 1}$ *(Nystrom schemes)*

For $m = 1$, we can increase the order of accuracy by allowing $l > 1$. These are called Nystrom schemes.

$\underline{m > 1}$

Schemes with $m > 1$ are not of much interest and will not be discussed here.

5.5 Implicit schemes

For implicit schemes, with $\beta \neq 0$, we can achieve accuracy at least as high as $l+2$. We consider several special cases:

$\underline{m=0,\, l=0}$

Eq. (5.4) reduces to

$$\frac{q^{n+1}-q^n}{\Delta t} = \beta f^{n+1} + \alpha_n f^n \,. \tag{5.16}$$

In this case, the consistency condition reduces to $\alpha_n + \beta = 1$. The discretization error is $\Delta t q'' \left(\frac{1}{2} - \beta\right) + \mathscr{O}\left(\Delta t^2\right)$. For the special case $\beta = 1,\ \alpha_n = 0$, we get the "backward implicit scheme," which has first-order accuracy, and can be said to correspond to $l = -1$. For $\beta = \alpha = \frac{1}{2}$, we get the "trapezoidal implicit" scheme, which has second-order accuracy, as expected from the general rule for implicit schemes. The trapezoidal implicit scheme has some very nice properties, which will be discussed later. It can be difficult to use, however.

$\underline{m=0,\, l>0 \text{ \textit{(Adams-Moulton schemes)}}}$

These are analogous to the Adams-Bashforth schemes, except that $\beta \neq 0$. Table 5.2 summarizes the properties of the Adams-Moulton schemes, for $l = 1$, 2, and 3. For $l = 0$ (not listed in the table), the maximum accuracy (2nd order) is obtained for $\beta = 0$, in which case we just recover the leapfrog scheme.

l	β	α_n	α_{n-1}	α_{n-2}	α_{n-3}	truncation error
1	$5/12$	$8/12$	$-1/12$			$\mathscr{O}\left(\Delta t^3\right)$
2	$9/24$	$19/24$	$-5/24$	$1/24$		$\mathscr{O}\left(\Delta t^4\right)$
3	$251/720$	$646/720$	$-264/720$	$106/720$	$-19/720$	$\mathscr{O}\left(\Delta t^5\right)$

Table 5.2: Adams-Moulton Schemes

$\underline{m=1,\, l=1 \text{ \textit{(Milne corrector)}}}$[2]

[2]If there is a "Milne corrector," then there must be "Milne predictor." (See Section 5.6 for an explanation of this terminology.) In fact, the Milne predictor is an explicit scheme with

$$m=3,\ l=3,\ \alpha_n = 2/3,\ \alpha_{n-1} = -1/3,\ \alpha_{n-2} = 2/3,\ \alpha_{n-3} = 0.$$

Eq. (5.4) reduces to

$$\frac{q^{n+1}-q^{n-1}}{2\Delta t} = \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1} \,, \tag{5.17}$$

so that the consistency condition is

$$\beta + \alpha_n + \alpha_{n-1} = 1. \tag{5.18}$$

The discretization error is

$$\varepsilon = \Delta t q'' \left(-\beta + \alpha_{n-1}\right) + \frac{\Delta t^2}{2!} q''' \left(\frac{1}{3} - \beta - \alpha_{n-1}\right) + \frac{\Delta t^3}{3!} q'''' \left(-\beta + \alpha_{n-1}\right) + \mathscr{O}\left(\Delta t^4\right). \tag{5.19}$$

The choices $\beta = 1/6,\ \alpha_n = 4/6,\ \alpha_{n-1} = 1/6$, give fourth-order accuracy. This is again more than would be expected from the general rule.

$\underline{m = 1, l = 2}$

Here there is no gain in accuracy. The highest accuracy is obtained for $\alpha_{n-2} = 0$, so that the scheme reduces to the Milne corrector.

5.6 Iterative schemes

Iterative schemes are not new, but they have become increasingly popular. The idea is to obtain q^{n+1} through an iterative, multi-step procedure, which involves multiple evaluations of the function f. In a two-step iterative scheme, the first step is called the "predictor," and the second step is called the "corrector."

Iterative schemes have several advantages:

- Iterative schemes are two-time-level schemes (with $m = 0$ and $l = 0$), which means that they have no computational modes in time. This may be the main reason for their increasing popularity.
- As shown below, iterative schemes can give higher accuracy, without increasing m.

- It may be possible to take longer time steps than with non-iterative schemes.

Non-iterative schemes, such as those discussed earlier in this chapter, involve only a single evaluation of f for each time step. A disadvantage of iterative schemes is their computational expense, which increases because of the multiple evaluations of f. This drawback may be tolerable if the scheme is stable with a longer time step.

Iterative schemes may or may not fit into the family of schemes discussed earlier in this chapter, depending on what $f\left[q\left(t\right),\, t\right]$ is.

Here is a simple example of an iterative scheme. Change (5.16) by replacing $f^{n+1} \equiv f\left[q^{n+1},\, (n+1)\,\Delta t\right]$ by $\left(f^{n+1}\right)^* \equiv f\left[\left(q^{n+1}\right)^*,\, (n+1)\,\Delta t\right]$. In the predictor step, $\left(q^{n+1}\right)^*$ is obtained using the Euler scheme:

$$\frac{\left(q^{n+1}\right)^* - q^n}{\Delta t} = f^n\,. \tag{5.20}$$

Think of $\left(q^{n+1}\right)^*$ as a "provisional" value of q^{n+1}. We then complete the time step by obtaining the "final" value of q^{n+1} using the corrector step:

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta^* \left(f^{n+1}\right)^* + \alpha_n f^n\,. \tag{5.21}$$

Here β^* and α_n are coefficients that can be chosen to make the scheme do what we want. This is the corrector step. When $\beta^* = 1,\ \alpha_n = 0$, Eq. (5.21) is an imitation of the backward (implicit) scheme, and is called the Euler-backward scheme or Matsuno[3] scheme (Matsuno, 1966). When $\beta^* = \frac{1}{2},\ \alpha_n = \frac{1}{2}$, Eq. (5.21) is an imitation of the trapezoidal implicit scheme and is called the Heun scheme. The Matsuno scheme has first-order accuracy, and the Heun scheme has second-order accuracy.

Note that (5.21) does not "fit" into the framework (5.4), because $\left(f^{n+1}\right)^*$ does not appear on the right-hand side of (5.4), and in general $\left(f^{n+1}\right)^*$ cannot be written as a linear combination of the fs that do appear there.

Also note that the Heun scheme is explicit, and does not require the past history (does not require $l > 0$). Despite this, it has second order accuracy, because of the iteration. This illustrates that *iteration can increase the order of accuracy*.

A famous example of an iterative scheme is the fourth-order Runge-Kutta scheme, which is given by:

[3] Yes, this is the same Matsuno who is known for his work on equatorial waves and stratospheric sudden warmings.

$$
\begin{aligned}
q^{n+1} &= q^n + \Delta t\left(k_1 + 2k_2 + 2k_3 + k_4\right)/6\,, \\
k_1 &= f\left(q^n, n\Delta t\right), \quad k_2 = f\left[q^n + k_1\Delta t/2, \left(n+\tfrac{1}{2}\right)\Delta t\right]\,, \\
k_3 &= f\left[q^n + k_2\Delta t/2, \left(n+\tfrac{1}{2}\right)\Delta t\right], \quad k_4 = f\left[q^n + k_3\Delta t, (n+1)\,\Delta t\right]\,.
\end{aligned}
\tag{5.22}
$$

Each of the ks in (5.22) can be interpreted as an approximation to f. The ks have to be evaluated successively, which means that the function f has to be evaluated four times to take one time step. None of the four fs can be "re-used" on the next time step. For this reason, the scheme is not very practical unless a long time step can be used. Fortunately, long time steps are often possible, depending on the form of f.

Fig. 5.4 provides a simple fortran example to illustrate more clearly how the fourth-order Runge-Kutta scheme actually works. Appendix B provides a proof that the scheme really does have the advertised fourth-order accuracy.

A more general discussion of Runge-Kutta schemes can be found on Wikipedia.

5.7 Segue

It is possible to construct time-differencing schemes of arbitrarily high accuracy by including enough time levels, and/or through iteration. Schemes of very high accuracy (e.g., tenth order) can be constructed quite easily, but highly accurate schemes involve a lot of arithmetic and so are expensive. In addition they are complicated. An alternative approach to obtain high accuracy is to use a simpler low-order scheme with a smaller time step. This also involves a lot of arithmetic (because more time steps are needed), but on the other hand the small time step makes it possible to represent the temporal evolution in more detail.

This chapter has presented a survey of some explicit, implicit, and iterative time-differencing schemes, including a discussion of their order of accuracy. Computational stability has not yet been addressed. The next chapter will introduce that subject, in the context of two simple but important ordinary differential equations.

5.8 Problems

1. What are the orders of the discretization errors of the Matsuno and Heun schemes? Prove your answer.

```
c       Initial conditions for time-stepped variables X, Y, and Z.

c       The time step is dt, and dt2 is half of the time step.

        X = 2.5
        Y = 1.
        Z = 0.

        do n=1,nsteps

c       Subroutine dot evaluates time derivatives of X, Y, and Z.

        call dot(X, Y, Z,Xdot1,Ydot1,Zdot1)

c       First provisional values of X, Y, and Z.

        X1 = X + dt2 * Xdot1
        Y1 = Y + dt2 * Ydot1
        Z1 = Z + dt2 * Zdot1

        call dot(X1,Y1,Z1,Xdot2,Ydot2,Zdot2)

c       Second provisional values of X, Y, and Z.

        X2 = X + dt2 * Xdot2
        Y2 = Y + dt2 * Ydot2
        Z2 = Z + dt2 * Zdot2
        call dot(X2,Y2,Z2,Xdot3,Ydot3,Zdot3)

c       Third provisional values of X, Y, and Z.

        X3 = X + dt * Xdot3
        Y3 = Y + dt * Ydot3
        Z3 = Z + dt * Zdot3

        call dot(X3,Y3,Z3,Xdot4,Ydot4,Zdot4)

c       "Final" values of X, Y, and Z for this time step.

        X = X + dt * (Xdot1 + 2.*Xdot2 + 2.*Xdot3 + Xdot4)/6.
        Y = Y + dt * (Ydot1 + 2.*Ydot2 + 2.*Ydot3 + Ydot4)/6.
        Z = Z + dt * (Zdot1 + 2.*Zdot2 + 2.*Zdot3 + Zdot4)/6.

        end do
```

Figure 5.4: A simple fortran program that illustrates how the fourth-order Runge-Kutta scheme works. Note the four calls to subroutine "dot."

Chapter 6

What a difference an *i* makes

6.1 Motivation 1

In this Chapter, we focus on two simple and similar-looking ordinary differential equations that are relevant to important atmospheric processes. We will explore some time-differencing schemes for the equations, and show how to test the computational stability of the schemes.

The first process is advection, in which air moves from one place to another, carrying its properties with it. Advection will be discussed in detail later, starting with Chapter 7. The advection of an intensive scalar, A, can be described by

$$\frac{\partial A}{\partial t} = -u\frac{\partial A}{\partial x} \,. \tag{6.1}$$

This is a partial differential equation. Here we restrict ourselves to one spatial dimension, x, with a spatially uniform advecting current u. Suppose that $A(x)$ consists of a single Fourier mode, with wave number k:

$$A(x) = \hat{A}(t)\, e^{ikx} \,, \tag{6.2}$$

where $\hat{A}$ is a constant, complex amplitude. Then

$$\frac{\partial A}{\partial x} = ik\hat{A}e^{ikx} \,. \tag{6.3}$$

This allows us to rewrite (6.1) as

$$\frac{d\hat{A}}{dt} = -uik\hat{A} \, . \tag{6.4}$$

This is an ordinary differential equation. We have converted the partial differential equation (6.1) into the ordinary differential equation (6.4) by using the assumed for the spatial dependence, given by (6.2). The solution of (6.4) is circular or oscillatory motion in the complex plane, with frequency $-uk$. In this chapter, we will explore time differencing schemes for the oscillation equation, (6.4), with the simplified notation shown below:

$$\frac{dq}{dt} = i\omega q \, . \tag{6.5}$$

Here the frequency is simply denoted by ω.

The point of the discussion above is that the oscillation equation, (6.5), is a "prototype" of the advection equation. You will not be surprised to hear that the oscillation equation is also relevant to wave propagation. In the discussion above, we have kept the space derivatives as continuous, so that we can focus on the time differencing.

The exact solution of (6.5) is

$$q(t) = q(0)\, e^{i\omega t} \, , \tag{6.6}$$

From (6.6), it follows that

$$q(t+\Delta t) = q(t)\, e^{i\omega \Delta t} \, . \tag{6.7}$$

The state of the system can be characterized by an amplitude and a phase. Here and frequently throughout the remainder of this book we will use Euler's formula (or rather *one* of Euler's formulas), which says that

$$\boxed{e^{i\gamma} = \cos\gamma + i\sin\gamma} \, , \tag{6.8}$$

where γ is any complex number. We will use this important equation many times, so if you don't already know it by heart you should make an effort to remember it going forward. Since $\sin^2\gamma+\cos^2\gamma=1$, we see that

$$\left|e^{i\gamma}\right| = 1\ . \tag{6.9}$$

It follows from (6.6) and (6.9) that

$$|q| = |q\,(0)| \text{ for all time}\ . \tag{6.10}$$

This means that the amplitude is independent of time. In view of (6.10), it is reasonable to say that a "good" scheme for the oscillation equation will give a solution that has a time-independent amplitude.

6.2 Computational stability

In almost all physical problems, including both (6.5) and (6.59), the true solution is bounded for all time. We now investigate the behavior of the discretization error, $q^n - q\,(n\Delta t)$, as n increases, for fixed Δt. *Does the solution remain bounded after an arbitrary number of time steps, for any initial condition*? If so, the scheme is said to be stable; otherwise it is unstable. Computational instability happens over a sequence of time steps. If you are taking only one time step, an instability (if present) does not matter.

Any scheme can be written in the form

$$q^{n+1} \equiv \lambda q^n\ , \tag{6.11}$$

where λ, which may be complex, is called the *amplification factor*. Eq. (6.11) is the definition of λ. John von Neumann proposed that the stability of a time-differencing scheme can be tested by working out the form of λ, and checking its absolute value. If $|\lambda| \leq 1$ then the scheme is stable. Otherwise it is unstable. The calculation of λ usually involves linearlization of the finite-difference equation. This is a limitation of von Neumann's method.

6.3 Time differencing schemes for the oscillation equation

6.3.1 The solution of the continuous oscillation equation

From (6.6) we can show that

$$q\left[(n+1)\,\Delta t\right] = e^{i\Omega} q\left(n\Delta t\right) \ , \tag{6.12}$$

where

$$\Omega \equiv \omega \Delta t \tag{6.13}$$

is the change in phase over the time interval Δt. Note that Eq. (6.12) describes the exact solution of the differential equation.

6.3.2 Amplitude errors and phase errors

Comparing (6.11) with (6.12), we see that the exact value of λ, based on the solution of the differential equation, is given by

$$\lambda_T = e^{i\Omega} \ . \tag{6.14}$$

As discussed above, Euler tells us that

$$|\lambda_T| = 1 \ . \tag{6.15}$$

If the $|\lambda|$ of the solution of a numerical scheme for (6.5) is not equal to one, we say that the scheme has "amplitude errors," which means that the amplitude of the numerical solution is spuriously growing or decaying. If the numerically simulated phase change per time step is not equal to Ω, we say that the scheme has "phase errors," which means that the numerically simulated oscillation is either too fast or too slow. As discussed later, some schemes have no amplitude error at all, but even those schemes do have phase errors.

We now show how to obtain useful measures of the amplitude and phase errors of a scheme, based on the values of λ and λ_T. For the solution of a finite-difference equation,

the phase change per time step can be expressed in terms of the real and imaginary parts of λ. We write

$$\begin{aligned} \lambda &= \lambda_R + i\lambda_I \\ &= |\lambda| e^{i\theta} . \end{aligned} \tag{6.16}$$

where

$$\theta = \tan^{-1}\left(\frac{\lambda_I}{\lambda_R}\right), \quad \lambda_R = |\lambda|\cos\theta, \text{ and } \lambda_I = |\lambda|\sin\theta . \tag{6.17}$$

Here θ is the change in phase per time step in the approximate numerical solution. If the numerical solution was perfect, we would have $\theta = \Omega$. Positive θ (like positive Ω) denotes counterclockwise rotation in the complex plane. For example, if $\theta = \pi/2$, it takes four time steps to complete one oscillation. This is shown schematically in Fig. 6.1, in which the ordinate represents the imaginary part of q^n. This is the case in which λ is purely imaginary.

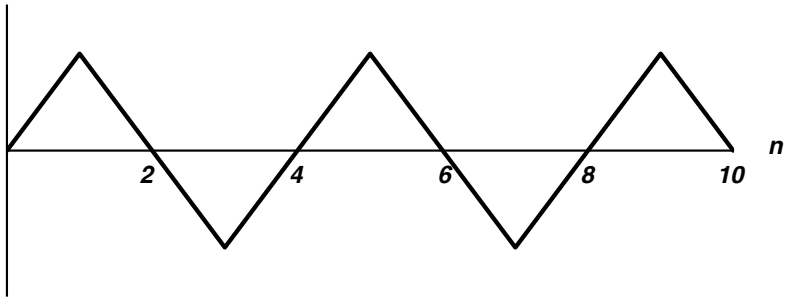


Figure 6.1: Schematic illustration of the solution of the oscillation equation for the case in which λ is pure imaginary and the phase changes by $\theta = \pi/2$ on each time step. The abscissa is the time step counter, and the vertical axis is the imaginary part of the solution. The initial condition is assumed to be real in this example.

From the preceding discussion, we see that, for the oscillation equation, the amplitude error per time step is $|\lambda| - 1$, and the phase error per time step is $\theta - \Omega$. Further discussion of amplitude and phase errors is given in Chapter 12.

6.3.3 Non-iterative two-level schemes for the oscillation equation

In principle, any of the time-differencing schemes described in Chapter 5 is a candidate for application to the oscillation equation, (6.5). Each scheme has its own strengths and weaknesses, as discussed below. Here we consider just a few examples.

A family of (possibly) implicit schemes is given by

$$q^{n+1} - q^n = i\omega\Delta t \left(\alpha q^n + \beta q^{n+1}\right) . \tag{6.18}$$

We require $\alpha + \beta = 1$ in order to guarantee consistency. The explicit Euler scheme is obtained (as a special case) with $\alpha = 1, \quad \beta = 0$; the backward implicit scheme with $\alpha = 0, \quad \beta = 1$; and the trapezoidal-implicit scheme with $\alpha = \beta = 1/2$. Eq. (6.18) can easily be solved for q^{n+1}:

$$(1 - i\Omega\beta)\, q^{n+1} = (1 + i\Omega\alpha)\, q^n , \tag{6.19}$$

or

$$\begin{aligned} q^{n+1} &= \left(\frac{1 + i\Omega\alpha}{1 - i\Omega\beta}\right) q^n \\ &\equiv \lambda q^n . \end{aligned} \tag{6.20}$$

In the second equality in (6.20) we make use of the definition of λ.

For the forward (Euler) scheme, $\alpha = 1$, $\beta = 0$, and so from (6.20) we find that

$$\lambda = 1 + i\Omega , \tag{6.21}$$

and

$$|\lambda| = \sqrt{1 + \Omega^2} > 1 . \tag{6.22}$$

This means that, for the oscillation equation, the forward scheme is *unconditionally unstable*. It's a non-starter. From (6.16) and (6.21), we see that the phase change per time step, θ, satisfies $\tan\theta = \Omega$, so that $\theta \cong \Omega$ for small Δt, as expected.

For the backward scheme, $\alpha = 0, \quad \beta = 1$, and

$$\begin{aligned}\lambda =& \frac{1}{1-i\Omega} \\ =& \frac{1+i\Omega}{1+\Omega^2} ,\end{aligned} \tag{6.23}$$

so that

$$\begin{aligned}|\lambda| =& \frac{\sqrt{1+\Omega^2}}{1+\Omega^2} \\ =& \frac{1}{\sqrt{1+\Omega^2}} < 1 .\end{aligned} \tag{6.24}$$

The backward scheme is, therefore, *unconditionally stable*. The amplitude of the oscillation decreases with time, however. This is an error, because the amplitude is supposed to be constant with time. As with the forward scheme, the phase change per time step satisfies $\tan\theta = \Omega$, so again the phase error is small for small Δt. The real part of λ (the cosine part) is always positive, which means that, no matter how big we make the time step, the phase change per time step satisfies $-\pi/2 < \theta < \pi/2$. This scheme can be used for the Coriolis terms in a model, but because of the damping it is not a very good choice.

For the trapezoidal implicit scheme, given by $\alpha = \beta = 1/2$, we find that

$$\lambda = \frac{1+i\Omega/2}{1-i\Omega/2} , \tag{6.25}$$

which leads to $|\lambda|^2 = 1$. This scheme is *unconditionally stable*, and has no amplitude error at all. The phase error per time step is small. It is a very nice scheme for the oscillation equation.

6.3.4 Iterative schemes for the oscillation equation

Now consider a family of iterative schemes for the oscillation equation, given by

$$q^{n+1*} - q^n = i\Omega q^n, \tag{6.26}$$

$$q^{n+1} - q^n = i\Omega \left(\alpha q^n + \beta^* q^{n+1*}\right) . \tag{6.27}$$

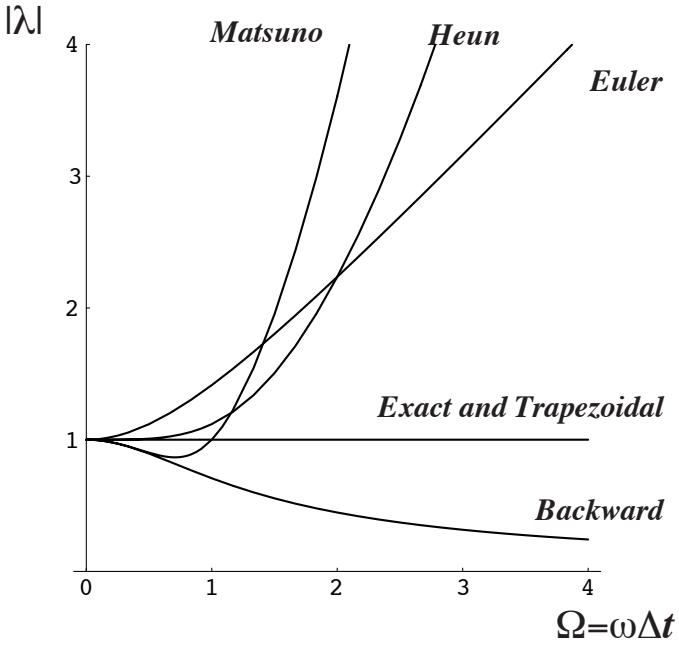


Figure 6.2: The magnitude of the amplification factor λ as a function of $\Omega \equiv \omega\Delta t$, for various difference schemes. The Euler, backward, trapezoidal, Matsuno, and Heun schemes are shown. The magnitude of the amplification factor for the trapezoidal scheme coincides with that of the true solution for all values of Ω. Caution: This does not mean that the trapezoidal scheme gives the exact solution!

Recall that $\alpha = 0, \quad \beta^* = 1$ gives the Matsuno scheme, and $\alpha = \beta^* = 1/2$ gives the Heun scheme. Eliminating q^{n+1*} between (6.26) and (6.27) for the Matsuno scheme, we find that

$$\lambda = \left(1 - \Omega^2\right) + i\Omega \ , \tag{6.28}$$

so that

$$|\lambda| = \sqrt{1 - \Omega^2 + \Omega^4} \begin{cases} > 1 & \text{for} \quad \Omega^2 > 1 \\ = 1 & \text{for} \quad \Omega^2 = 1 \\ < 1 & \text{for} \quad \Omega^2 < 1 \ . \end{cases} \tag{6.29}$$

This is, therefore, a *conditionally stable* scheme; the condition for stability is $\Omega \leq 1$. Similarly, for the Heun scheme, we find that

$$\lambda = \left(1 - \Omega^2/2\right) + i\Omega \ , \tag{6.30}$$

and

$$|\lambda| = \sqrt{(1-\Omega^2/2)^2 + \Omega^2} = \sqrt{1+\Omega^4/2} > 1 \ . \tag{6.31}$$

This shows that the Heun scheme is *unconditionally unstable* when applied to the oscillation equation, but for small Ω it is not as unstable as the forward scheme. It can be used in a model that includes some physical damping.

The results discussed above are summarized in Fig. 6.2 and Fig. 6.3.

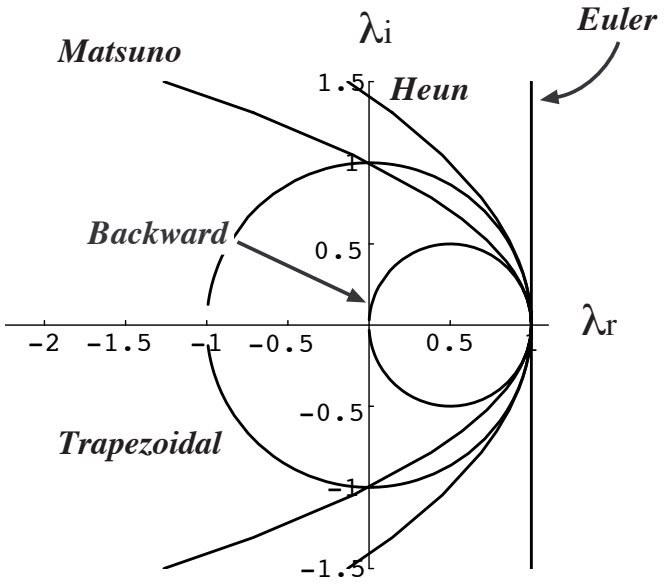


Figure 6.3: Variations of the real and imaginary components of the amplification factor, as Ω changes "parametrically." The actual values of Ω are not shown in the figure. Both the exact solution and the trapezoidal scheme lie on the unit circle.

6.3.5 The leapfrog scheme for the oscillation equation

The leapfrog scheme for the oscillation equation is

$$q^{n+1} - q^{n-1} = 2i\Omega q^n \ . \tag{6.32}$$

The leapfrog scheme was widely used in the past, but has fallen out of favor. We discuss it here to explain why. We can rewrite (6.32) as

$$q^{n+1} - 2i\Omega q^n - q^{n-1} = 0 \,, \tag{6.33}$$

and look for a solution of the form $q^{n+1} = \lambda q^n$, for all n. Then (6.33) reduces to

$$\lambda^2 - 2i\Omega\lambda - 1 = 0 \,. \tag{6.34}$$

Make sure that you understand how (6.34) follows from (6.33). The solutions of (6.34) are

$$\lambda_1 = i\Omega + \sqrt{1-\Omega^2},\ \lambda_2 = i\Omega - \sqrt{1-\Omega^2} \,, \tag{6.35}$$

i.e., there are two "modes," satisfying

$$q_1^{n+1} = \lambda_1 q_1^n, \quad q_2^{n+1} = \lambda_2 q_2^n \,. \tag{6.36}$$

One of these two solutions corresponds to the solution of the differential equation. We call it the "physical mode." The other solution is a computational mode in time. Such computational modes were already briefly discussed in Chapter 5. *The differential equation only has one solution, so the fact that the finite-difference equation has more than one solution is a problem.*

How can we tell which solution is the physical mode? Consider the limiting values of λ_1 and λ_2 as $\Omega \to 0$ (which we interpret as $\Delta t \to 0$). Notice that $\lambda_1 \to 1$, while $\lambda_2 \to -1$. We know that for the true solution $\lambda = 1$, and so we can identify q_1 as the physical mode, and q_2 as the 'computational mode.

Notice that q_2^{n+1} generally does not approach q_2^n as $\Delta t \to 0$! This means that you can't get rid of the computational mode by making the time step smaller. The computational mode arises from the *structure* of the scheme.

6.3.6 The stability of the leapfrog scheme for the oscillation equation

To evaluate the numerical stability of the leapfrog scheme as applied to the oscillation equation, consider three cases.

Case (i): $|\Omega| < 1$

In this case, the factor $\sqrt{1-\Omega^2}$ in (6.35) is real, and we find that $|\lambda_1|^2 = |\lambda_2|^2 = 1$. This means that both the physical and the computational modes are neutral, so we have phase errors but no amplitude errors. Let the phase changes per time step of the physical and computational modes be denoted by θ_1 and θ_2, respectively. Then

$$\lambda_1 = e^{i\theta_1} \text{ and } \lambda_2 = e^{i\theta_2}. \tag{6.37}$$

Comparing (6.37) with (6.35), and using Euler's formula, we find by inspection that

$$\begin{gathered} \cos\theta_1 = \sqrt{1-\Omega^2}, \quad \cos\theta_2 = -\sqrt{1-\Omega^2}, \\ \sin\theta_1 = \Omega, \quad \sin\theta_2 = \Omega\ . \end{gathered} \tag{6.38}$$

From (6.38), you should be able to see that $\theta_2 = \pi - \theta_1$. For $\Omega \to 0$, we find that $\theta_1 \cong \Omega$ (good), and $\theta_2 \cong \pi$ (bad). The apparent frequency of the physical mode is $\theta_1/\Delta t$, which is approximately equal to ω. Then we can write

$$q_1^{n+1} = e^{i\theta_1} q_1^n \tag{6.39}$$

for the physical mode, and

$$q_2^{n+1} = e^{i(\pi-\theta_1)} q_2^n \tag{6.40}$$

for the computational mode. Recall that the true solution is given by

$$q\left[(n+1)\,\Delta t\right] = e^{i\Omega} q\left(n\Delta t\right)\ . \tag{6.41}$$

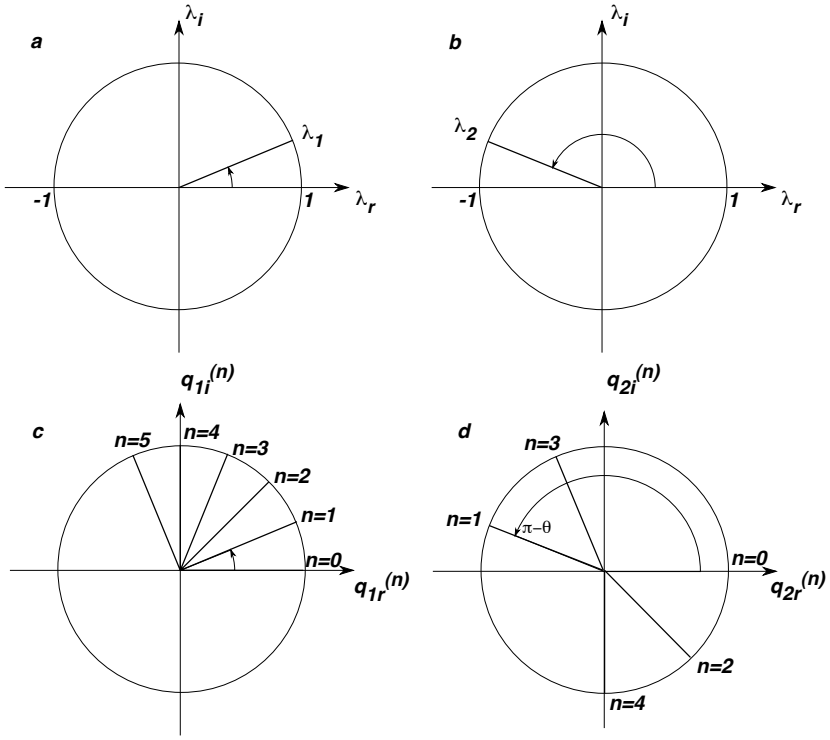


Figure 6.4: Panels a and b show the amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| < 1$. Panel a is for the physical mode, and panel b is for the computational mode. Panels c and d show solutions of the oscillation as obtained with the leapfrog scheme for $|\Omega| < 1$. Panel c is for the physical mode, and panel d is for the computational mode. In making these figures it has been assumed that $\theta_1 = \pi/8$.

Panels a and b of Fig. 6.4 respectively show plots of λ_1 and λ_2 in the complex λ-plane. The figures have been drawn for the case of $\theta_1 = \pi/8$. The absolute value of λ is, of course, always equal to 1. Panel c shows the graph of the real part of q_1^n versus its imaginary part. Recall that $q_1^n = \lambda_1^n q_1^0 = e^{in\theta_1} q_1^0$. Panel d gives a similar plot for q_2^n. Here we see that the real and imaginary parts of the computational mode of q^n both oscillate from one time step to the next. Graphs showing each part versus n are given in Figure 6.5. The physical mode looks nice. The computational mode is ugly.

Case (ii): $|\Omega| = 1$

Here $\lambda_1 = \lambda_2 = i\Omega = i$ (see (6.35)), i.e., both λ's are purely imaginary with modulus one (i.e., both are neutral), as shown in Fig. 6.6. This means that both solutions rotate through $\pi/2$ on each time step, so that the period is $4\Delta t$. The phase errors are very large; the correct phase change per time step is 1 radian, and the computed phase change is $\pi/2$ radians, which implies an error of 57%.

This illustrates that a scheme that is stable but on the verge of instability is usually

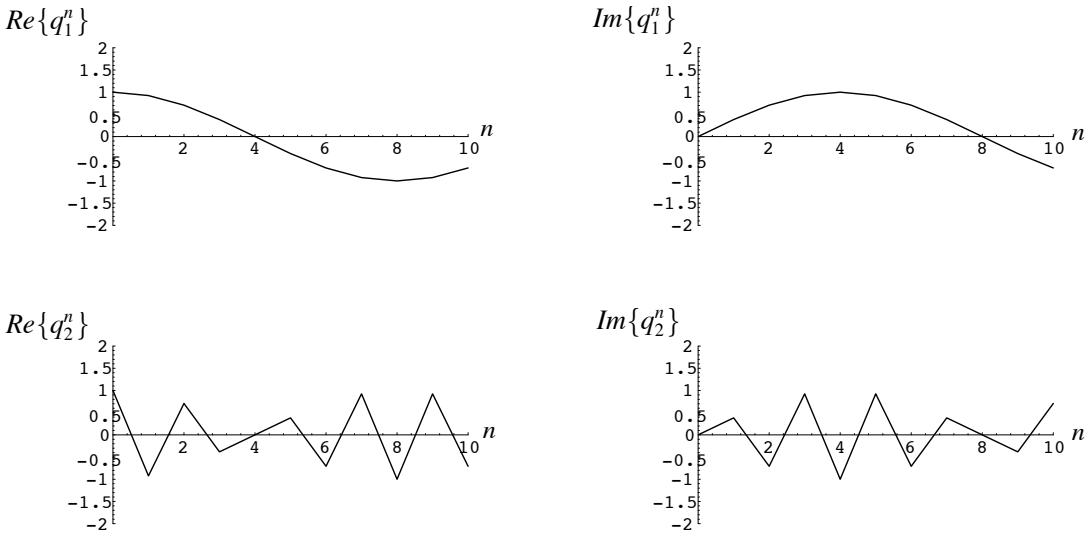


Figure 6.5: Graphs of the real and imaginary parts of the physical and computational modes for the solution of the oscillation equation as obtained with the leapfrog scheme for $|\Omega| < 1$.

subject to large discretization errors and may give a very poor solution; you should not be confident that you have a good solution just because your model does not blow up!

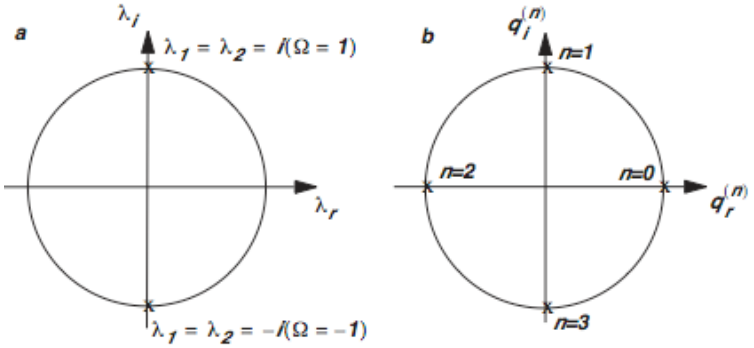


Figure 6.6: Panel a shows the amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| = 1$. Panel b shows the real and imaginary parts of the corresponding solution, for $n = 0, \quad 1, \quad 2,$ and 3.

Case (iii): $|\Omega| > 1$

Here again both λ_1 and λ_2 are purely imaginary, so again both solutions rotate by $\frac{\pi}{2}$ on each time step, regardless of the value of ω. Eq. (6.36) can be written as

$$\lambda_1 = i\left(\Omega + \sqrt{\Omega^2 - 1}\right) \text{ and } \lambda_2 = i\left(\Omega - \sqrt{\Omega^2 - 1}\right) . \tag{6.42}$$

If $\Omega > 1$ then $|\lambda_1|^2 > 1$ and $|\lambda_2|^2 < 1$, and if $\Omega < -1$, then $|\lambda_1|^2 < 1$ and $|\lambda_2|^2 > 1$. Either way, one of the modes is damped and the other amplifies. The existence of an amplifying mode means that the scheme is unstable.

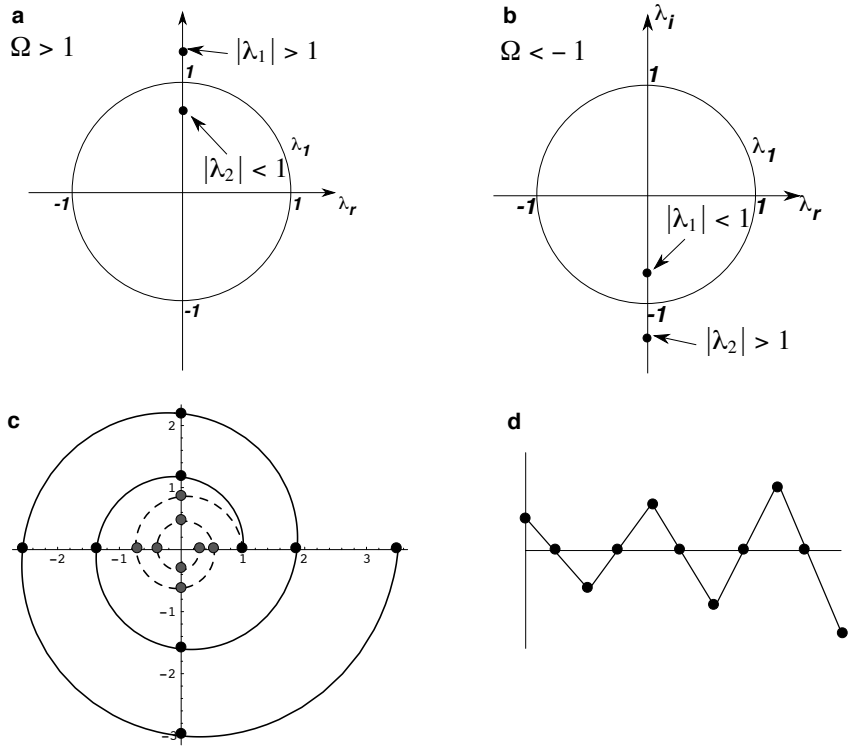


Figure 6.7: Panels a and b show the amplification factors for the oscillation equation with the leapfrog scheme, with $|\Omega| > 1$. Panel c shows the corresponding solution. The solid curve shows the unstable mode, which is actually defined only at the black dots. The dashed curve shows the damped mode, which is actually defined only at the grey dots. Panel d is a schematic illustration of the amplification of the unstable mode. Note the period of $4\Delta t$, which is characteristic of this type of instability.

Panels a and b of Fig. 6.7 give a graphical representation of λ in the complex plane, for the case $|\Omega| > 1$. Note that $\lambda_1 = \left|\Omega + \sqrt{\Omega^2 - 1}\right| e^{i\frac{\pi}{2}}$ and $\lambda_2 = \left|\Omega - \sqrt{\Omega^2 - 1}\right| e^{i\frac{\pi}{2}}$. Panel c of Fig. 6.7 shows a plot of the evolving solution, q^n, in the complex plane, for the modes corresponding to λ_1 and λ_2 for $\Omega > 1$. The phase changes by $\frac{\pi}{2}$ on each step, because λ is purely imaginary, and so the period is $4\Delta t$; this period is a recognizable characteristic of the instability of the leapfrog scheme for the oscillation equation (a clue!). Panel d of Fig. 6.7 schematically shows q^n as a function of n for the amplifying mode corresponding to λ_1 i.e., q_1 is unstable and q_2 is damped.

In summary, the centered or leapfrog scheme is second-order-accurate and gives a neutral solution when $|\Omega| \leq 1$. For $|\Omega| > 1$, which means large Δt, the scheme is unstable. In short, the leapfrog scheme is conditionally stable when applied to the oscillation equation.

The leapfrog scheme is explicit, has a higher accuracy than the general rule, and is neutral if $\Omega \leq 1$. For these reasons, it was very widely used in the past. On the other hand, the leapfrog scheme allows a computational mode in time, which is bad, and has caused the scheme to fall out of favor.

The trapezoidal implicit scheme is also neutrally stable for the oscillation equation, and it is also time-reversible, just like a real oscillator. Unfortunately, it is more difficult to use in complicated nonlinear problems.

6.3.7 The second-order Adams-Bashforth Scheme for the oscillation equation

The second-order Adams-Bashforth scheme and its third-order cousin (Durran, 1991) have some very nice properties. The second-order Adams-Bashforth scheme for the oscillation equation is

$$q^{n+1} - q^n = i\Omega \left(\frac{3}{2} q^n - \frac{1}{2} q^{n-1} \right). \tag{6.43}$$

Like the leapfrog scheme, this is a three-level scheme, so it has a computational mode. The right-hand side of (6.43) represents a linear *extrapolation* (in time) of q from q^{n-1} and q^n to $n+1/2$. It can be interpreted in terms of a scheme centered around time level $n+1/2$. The amplification factor satisfies

$$\lambda^2 - \lambda \left(1 + \frac{3}{2} i\Omega \right) + \frac{1}{2} i\Omega = 0. \tag{6.44}$$

The two solutions of (6.44) are

$$\lambda_1 = \frac{1}{2} \left(1 + \frac{3}{2} i\Omega + \sqrt{1 - \frac{9}{4}\Omega^2 + i\Omega} \right), \tag{6.45}$$

and

$$\lambda_2 = \frac{1}{2} \left(1 + \frac{3}{2} i\Omega - \sqrt{1 - \frac{9}{4}\Omega^2 + i\Omega} \right), \tag{6.46}$$

Since $\lambda_1 \to 1$ as $\Omega \to 0$, the first mode is the physical mode, corresponding to the true solution. Note, however, that $\lambda_2 \to 0$ as $\Omega \to 0$. This means that with a sufficiently short time step *the "computational" mode is damped*, which is good.

In order to find $|\lambda|$, we will make some approximations based on the assumption that $\Omega \ll 1$, because the expressions in (6.46) - (6.70) are complicated and in practice Ω is usually small. Using the binomial theorem, we can approximate λ_1 by

$$\lambda_1 \cong 1 + i\Omega - \frac{9}{16}\Omega^2 \cong 1 + i\Omega - \frac{1}{2}\Omega^2 \,, \tag{6.47}$$

so that

$$|\lambda_1| \cong \sqrt{1 + \frac{\Omega^4}{4}} \cong 1 + \frac{\Omega^4}{8} \,. \tag{6.48}$$

which is always greater than 1. The physical mode is, therefore, unconditionally unstable. If Δt (and Ω) are sufficiently small, however, the solution is only weakly unstable. If physical damping is included in the problem the instability may be rendered harmless.

6.3.8 A good start

Eq. (6.36) applies for any scheme that has one computational mode, and it is not restricted to the oscillation equation. From (6.36) we see that after n time steps our two solutions will be

$$q_1^n = \lambda_1^{\color{red}n} q_1^0, \text{ and } q_2^n = \lambda_2^{\color{red}n} q_2^0 \,. \tag{6.49}$$

Here we have used red to label the superscripts n that are exponents; the black superscripts are time=level indices. The general solution of the finite-difference equation is a linear combination of these two modes, i.e.,

$$q^n = a\lambda_1^{\color{red}n} q_1^0 + b\lambda_2^{\color{red}n} q_2^0 \,, \tag{6.50}$$

where a and b are constant coefficients. *We would like to have $b = 0$, because then the computational mode has zero amplitude*.

Because we have two solutions, we need two initial conditions. These are the values of q^0 and q^1. We already know that the first time step cannot use the leapfrog scheme, because q^{-1} is not available. The first leapfrog step needs both q^0 and q^1, and is used to predict the value of q^2. We have to predict q^1 by starting from q^0, using a two-time-level scheme. For example, we could use the Euler forward scheme, or the Matsuno scheme, or a Runge-Kutta scheme.

The values of a and b, i.e., the amplitudes of the physical and computational modes, depend in part on how we compute the computational initial condition, q^1. We now show how the initial amplitudes of the physical and computational modes depend on the choice of q^1. With reference to (6.50), the two initial conditions can be written this way:

$$q^0 = aq_1^0 + bq_2^0 \text{ for } n = 0 \,, \tag{6.51}$$

and

$$q^1 = \lambda_1 \left(aq_1^0\right) + \lambda_2 \left(bq_2^0\right) \text{ for } n = 1 \,. \tag{6.52}$$

Here we have used λ_1 and λ_2 to advance q_1 and q_2 from time level 0 to time level 1. We can solve (6.51) and (6.52) for aq_1^0 and bq_2^0 in terms of q^0 and q^1, and substitute the results back into (6.50). The result is

$$q^n = \frac{\left(q^1 - \lambda_2 q^0\right)\lambda_1^n - \left(q^1 - \lambda_1 q^0\right)\lambda_2^n}{\lambda_1 - \lambda_2} \,. \tag{6.53}$$

By inspection of (6.53), we see that the initial values of the physical and computational modes are

$$q_1^0 = \frac{q^1 - \lambda_2 q^0}{\lambda_1 - \lambda_2} \,, \tag{6.54}$$

and

$$q_2^0 = -\frac{q^1 - \lambda_1 q^0}{\lambda_1 - \lambda_2} \,, \tag{6.55}$$

respectively. *If we are able to choose q^1 so that $q^1 - \lambda_1 q^0 = 0$, then the computational mode will have zero amplitude*. The condition $q^1 - \lambda_1 q^0 = 0$ simply means that q^1 is predicted from q^0 using exactly the amplification factor for the *physical mode*. That makes sense, right? Unfortunately, in realistically complicated models, this is impossible to arrange, but we can come close by using a sufficiently accurate method to predict q^1 from q^0.

6.3.9 Ad hoc damping of computational modes in time

In the idealized example discussed above, if the amplitude of the computational mode is initialized to (almost) zero through the use of a sufficiently accurate scheme to predict q^1, then it will remain small for all time. In a real numerical model, however, the computational mode can be excited, in the middle of a simulation, by various ongoing processes (e.g., nonlinear terms and parameterized physics) that have been omitted for simplicity in the present discussion. Given this reality, a model that admits computational modes needs a way to *damp* them, as a simulation progresses.

One simple approach is to "restart" the model periodically, e.g., once per simulated day. A restart means repeating the procedure used on the initial start, i.e., taking one time step with a forward-in-time scheme. One of the two solutions is abandoned or killed off, while the other lives on.

A second approach is to use a time filter, as suggested by Robert (1966) and Asselin (1972). We write

$$\overline{q}^n = q^n + \alpha \left(\overline{q}^{n-1} - 2q^n + q^{n+1} \right) . \tag{6.56}$$

Here the overbar denotes a filtered quantity, and α is a non-negative parameter that controls the strength of the filter. For $\alpha = 0$ the filter is "turned off." Models that employ this so-called Asselin filter often use $\alpha = 0.5$.

To use the filter, we predict q^{n+1} in the usual way, then use (6.56) to filter q^n. The filtered value of q^n, i.e., $\overline{q}^n$, is used to take the next leapfrog step from n to $n+2$. Note that (6.56) also uses $\overline{q}^{n-1}$. As an example, suppose that $\overline{q}^{n-1} = q^{n+1} = 1$, and $q^n = -1$. This is a zig-zag in time. Eq. (6.31) will give $\overline{q}^n = -1 + \alpha(1+2+1) = -1 + 4\alpha$. If we choose $\alpha = 0.5$, we get $\overline{q^n} = 1$, which means that the zig-zag has been eliminated.

The filter damps the computational mode, but it also damps the physical mode, and it reduces the overall order of accuracy of the time-differencing scheme from second-order (leapfrog without filter) to first-order (leapfrog with filter). Many authors have suggested alternative filtering approaches (e.g., Williams, 2013).

Two-level time-differencing schemes do not have computational modes. *The existence of computational modes is a major disadvantage of all schemes that involve more than two time levels*. There is a school of thought in the numerical modeling community that advocates using only two-time-level schemes. In the framework of (5.4) this means schemes with $l = m = 0$, i.e., the Euler forward or backward implicit schemes, which have only first-order accuracy. The iterative schemes, such as Matsuno, Heun, and Runge-Kutta, also use only two time levels, and can have higher-than-first-order accuracy. We have already seen that forward-in-time schemes can be quite accurate, e.g., the fourth-order Runge-Kutta scheme. The desirable properties of forward-in-time schemes can motivate the use of iterative schemes.

The current discussion is about computational modes in time. As mentioned earlier, there can also be computational modes in space. They will be discussed later.

6.3.10 A survey of time differencing schemes for the oscillation equation

Baer and Simons (1970) summarized the properties of various explicit and implicit schemes for the oscillation equation. These are listed in Table 6.1. Some properties of these schemes are shown in Figs. 6.9 and 6.8. All of the schemes are non-iterative.

There are many other schemes of higher-order accuracy. Since our meteorological interest mainly leads us to partial differential equations, the solutions to which will also suffer from discretization error due to space differencing, we cannot hope to gain much by increasing the accuracy of the time-differencing scheme alone.

6.4 Motivation 2

Now consider the diffusion equation, which is

$$\frac{\partial A}{\partial t} = D\frac{\partial^2 A}{\partial x^2} \, . \tag{6.57}$$

Here D is a positive diffusion coefficient, assumed constant for simplicity. Diffusion describes a process in which microscopic motions smooth out a spatially variable field. Chapter 16 is entirely devoted to equations similar to (6.57). Using (6.3), we find that

$$\frac{\partial \hat{A}}{\partial t} = -k^2 D\hat{A} \, . \tag{6.58}$$

Table 6.1: List of the time differencing schemes for the oscillation equation, as surveyed by Baer and Simons (1970). Schemes whose names begin with "E" are explicit, while those whose names begin with "I" are implicit. The numerical indices in the names are m, which is the number of "time intervals" over which the scheme steps, as defined in Eq. (5.4) and Fig. 5.1; and l, which controls the number of values of f used, again as defined in (5.4). For the coefficients, a blank entry means zero. Iterative schemes are not included in the survey.

Scheme identifier (m, l)	**Name**	β	α_n	α_{n-1}	α_{n-2}	α_{n-3}	α_{n-4}	**Order of Accuracy**
E01	Adams-Bashforth		3/2	-1/2				$(\Delta t)^2$
E02	Adams-Bashforth		23/12	-4/3	5/12			$(\Delta t)^3$
E03			55/24	-59/24	37/24	-9/24		$(\Delta t)^4$
E04			1901/720	-2774/720	2616/720	-1274/720	251/720	$(\Delta t)^5$
E11	Leapfrog		1					$(\Delta t)^2$
E12			7/6	-2/6	1/6			$(\Delta t)^3$
E33	Milne Predictor		2/3	-1/3	2/3			$(\Delta t)^4$
I01	Trapezoidal Implicit	1/2	1/2					$(\Delta t)^2$
I02		5/12	8/12	-1/12				$(\Delta t)^3$
I03	Moulton Corrector	9/24	19/24	-5/24	1/24			$(\Delta t)^4$
I04		251/720	646/720	-264/720	106/720	-19/720		$(\Delta t)^5$
I13	Milne Corrector	1/6	4/6	1/6				$(\Delta t)^4$
I14		29/180	124/180	24/180	4/180	-1/180		$(\Delta t)^5$
I35	Milne II Corrector	14/180	64/180	24/180	64/180	14/180		$(\Delta t)^6$

This is a form of the "decay equation," which we will analyze in this chapter using the notation

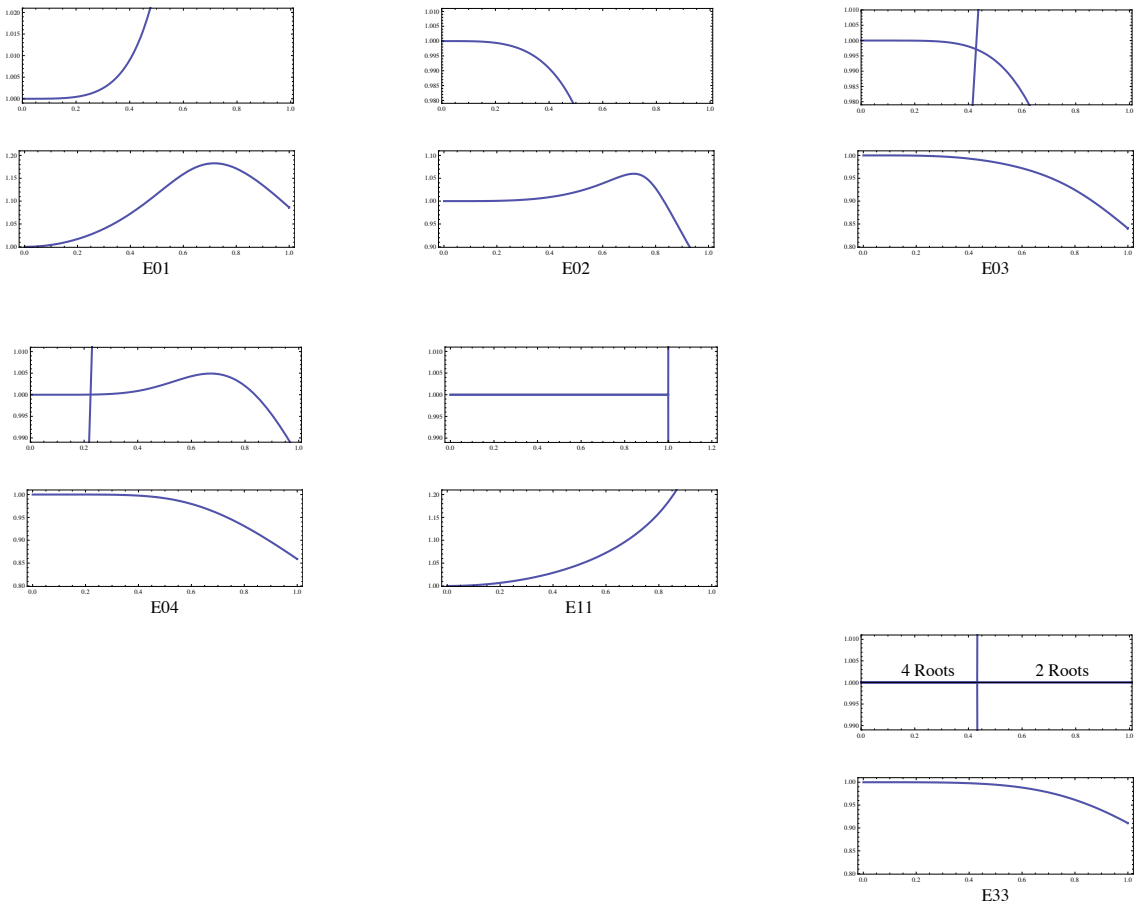


Figure 6.8: Magnitude of the amplification factor (upper panels) and θ/Ω (lower panels) for various explicit schemes as applied to the oscillation equation, following Baer and Simons (1970). The abscissa in each panel is Ω. See Table 6.1.

$$\frac{dq}{dt} = -\kappa q \,, \tag{6.59}$$

where q and κ are both real, and κ is positive. Note the similarity between (6.5) and (6.59).

6.5 Schemes for the decay equation

The solution of the continuous decay equation is

$$q(t) = q(0)\, e^{-\kappa t} \,. \tag{6.60}$$

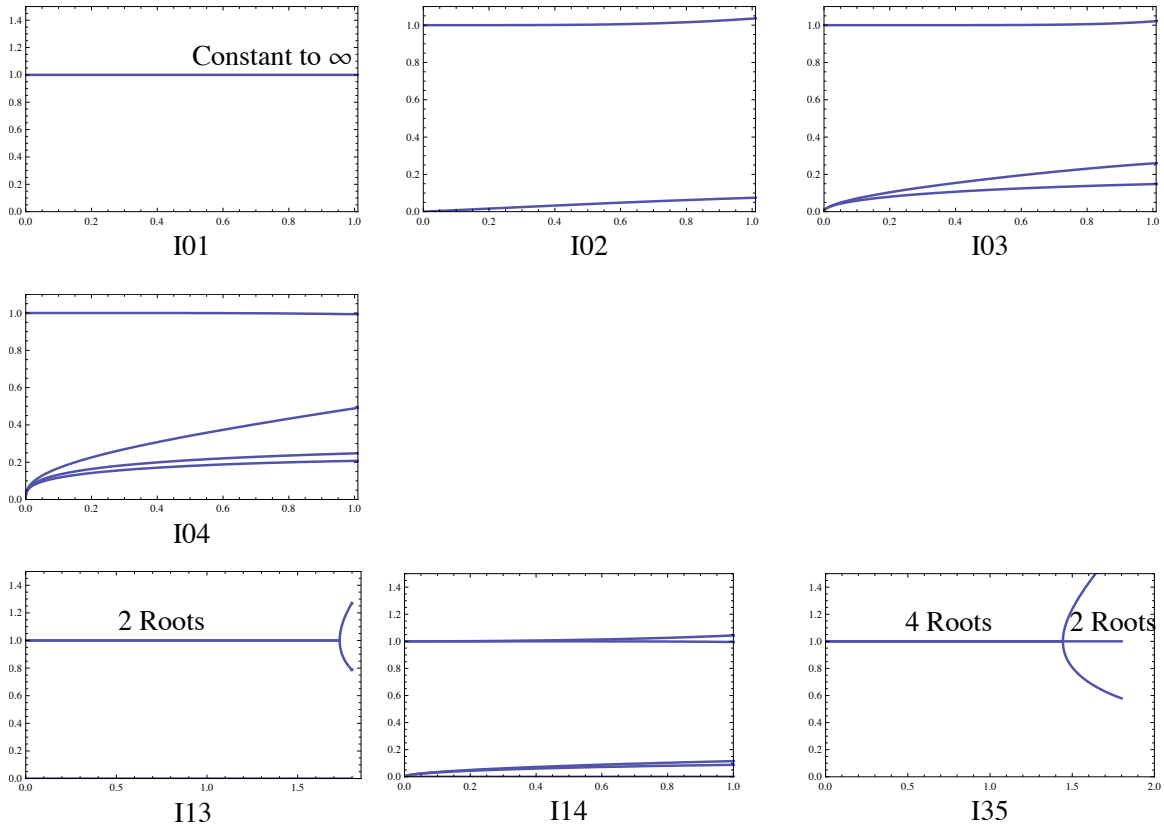


Figure 6.9: Amplification factors of various implicit schemes as applied to the oscillation equation, following Baer and Simons (1970). The abscissa in each panel is Ω. See Table 6.1.

This describes a simple exponential decay with time. For large t, $q \to 0$. Note that for the decay equation

$$q(t+\Delta t) = q(t)\, e^{-\kappa \Delta t} \ . \tag{6.61}$$

Compare with (6.7). A good scheme for the decay equation should give $q^{n+1} \to 0$ as $\kappa \Delta t \to \infty$, because that's what the solution to the differential equation does. The "true" value of λ is given by

$$\lambda_T = e^{-\kappa \Delta t} < 1 \tag{6.62}$$

This differs from the oscillation equation, for which $\mid \lambda_T \mid = 1$.

For the Euler (forward) scheme, the finite-difference analogue of (6.59) is

$$q^{n+1} - q^n = -Kq^n, \tag{6.63}$$

where $K \equiv \kappa \Delta t$. The solution is

$$q^{n+1} = (1-K)\, q^n \ . \tag{6.64}$$

Note that $\lambda = 1 - K$ is real. For $|1-K| < 1$, which is satisfied for κ or Δt small enough to give $K \leq 2$, the scheme is stable. This is, therefore, a *conditionally* stable scheme. It gives an unphysical damped oscillation for $1 < K < 2$. The oscillatory instability that occurs with the Euler forward scheme for $K \geq 2$ is an example of what is called *"overstability,"* in which the restoring force that is supposed to damp the perturbation goes too far and spuriously causes the perturbation to grow in the form of an amplifying oscillation.

The backward implicit scheme for the decay equation is

$$q^{n+1} - q^n = -Kq^{n+1}, \tag{6.65}$$

with solution

$$q^{n+1} = \frac{q^n}{1+K}. \tag{6.66}$$

Here $\lambda = 1/(1+K) < 1$, so the solution is unconditionally stable. As a bonus, for $K \to \infty$ we get $q^{n+1} \to 0$, which is consistent with the solution to the differential equation. For these reasons, the backward implicit scheme is a good choice for the decay equation, although of course it has only first-order accuracy.

It can be shown that, when applied to the decay equation,

- the trapezoidal implicit scheme is *unconditionally stable*, with better (second-order) accuracy than the backward implicit scheme;
- the Matsuno (Euler-Backward) scheme is *conditionally stable*;

- the Heun scheme is *conditionally stable*; and
- the second-order Adams-Bashforth scheme is *conditionally stable*.

There are many other possibilities, but in general implicit schemes are best for the decay equation.

Finally, the leapfrog scheme for the decay equation is

$$q^{n+1} - q^{n-1} = -2Kq^n \,, \tag{6.67}$$

and so λ satisfies

$$\lambda^2 + 2K\lambda - 1 = 0 \,. \tag{6.68}$$

The two roots are

$$\lambda_1 = -K + \sqrt{K^2+1}, \lambda_2 = -K - \sqrt{K^2+1} \tag{6.69}$$

Since $0 \leq \lambda_1 \leq 1$, and $\lambda_1 \to 1$ as $K \to 0$, we see that λ_1 corresponds to the physical mode. On the other hand, $|\lambda_2|$ is always greater than one, so *the leapfrog scheme is unconditionally unstable when applied to the decay equation*. Actually $\lambda_2 \leq -1 \quad (\lambda_2 \to -1 \text{ as } \mathit{varDeltat} \to 0)$, so the computational mode is "overstable;" it oscillates in sign from one time level to the next, and amplifies. It's just awful.

A simple interpretation is as follows. Suppose that we have $q = 0$ at $n = 0$ and $q > 0$ at $n = 1$, as shown in Fig. 6.10. From (6.68) we see that the restoring effect computed at $n = 1$ is added to q^0, resulting in a negative deviation at $n = 2$. The positive "restoring effect" computed at $n = 2$ is added to q^1, which is already positive, resulting in a *more* positive value at $n = 3$, as illustrated in Fig. 6.10. And so on. This is overstability again. Overstability is why the leapfrog scheme is a disastrous choice for the decay equation. In fact, *the leapfrog scheme is a bad choice for any "damping" component of a model*, e.g., diffusion. You should remember this fact forever.

Note that the first-order backward implicit scheme gives a good solution for the decay equation, while the second-order (i.e., "more accurate") leapfrog scheme gives a bad solution. This illustrates that a scheme can be accurate in the Taylor-series sense but hopelessly inaccurate in other ways.

$q^0 = 0$

$q^1 > 0$

$q^2 = q^0 - 2Kq^1 = 0 - 2Kq^1 < 0$

$q^3 = q^1 - 2Kq^2 = q^1 - 2K(q^0 - 2Kq^1) = q^1(1 + 4K^2) > q^1$

$q^4 = q^2 - 2Kq^3 < q^2$

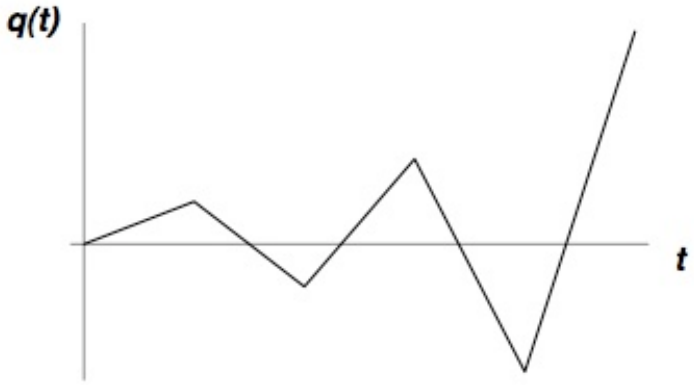


Figure 6.10: An illustration of how the leapfrog scheme leads to instability with the decay equation. The solution plotted represents the computational mode only and would be superimposed on the physical mode.

6.6 Damped oscillations

What should we do if we have an equation of the form

$$\frac{dq}{dt} = (i\omega - \kappa)\, q \,? \tag{6.70}$$

Here we are essentially combining the oscillation and decay equations. The exact solution of (6.70) is a damped oscillation. One possible scheme is based on a "mix" of the leapfrog and forward or backward schemes in the following manner. We write the finite-difference analogue of (6.70) as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n-1}\,, \tag{6.71}$$

(decay term forward differenced), or as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1} \tag{6.72}$$

(decay term backward differenced). The oscillation terms on the right-hand sides of (6.71) and (6.72) are in "centered" form, whereas the damping terms have an uncentered form. These schemes are conditionally stable.

6.7 Nonlinear damping

In real applications, it is quite typical that κ depends on q, so that the decay equation becomes nonlinear. Kalnay and Kanamitsu (1988) studied the behavior of ten time-differencing schemes for a nonlinear version of (6.59), given by

$$\frac{dq}{dt} = \left(-\kappa q^P\right) q + S \,, \tag{6.73}$$

where P is a non-negative exponent, and S is a source or sink whose form is unspecified. The reason for introducing S is simply to allow non-zero equilibrium values of q. In real applications, there is usually a term corresponding to S. In case $P = 0$ and $S = 0$, (6.73) reduces to (6.59).

An example of a real application that gives rise to an equation of the form (6.73) is boundary-layer parameterization. The soil temperature, T_g, satisfies an equation roughly of the form

$$C\frac{dT_g}{dt} = -\rho_a c_p c_T V \left(T_g - T_a\right) + S_g \,, \tag{6.74}$$

where C is the heat capacity of the soil layer, ρ_a is the density of the air, T_a is the temperature of the air at some level near the ground (often taken to be 2 m above the soil surface), c_T is a "transfer coefficient" that depends on $(T_g - T_a)$, V is the wind speed at a level near the ground (often taken to be 10 m above the soil surface), and S_g represents all other processes that affect the soil temperature, e.g., solar and infrared radiation, the latent heat flux, and the conduction of heat through the soil.

The air temperature is governed by a similar equation:

$$\rho_a D c_p \frac{dT_a}{dt} = \rho_a c_p c_T V (T_g - T_a) + S_a \,. \tag{6.75}$$

Here c_p is the specific heat of air at constant pressure, and D is the depth of the layer of air whose temperature is represented by T_a. Virtually all atmospheric models involve equations something like (6.74) and (6.75).

Subtracting (6.75) from (6.74), we find that

$$\frac{d(T_g - T_a)}{dt} = -\rho_a c_p c_T V (T_g - T_a) \left(\frac{1}{C} + \frac{1}{\rho_a D c_p} \right) + \left(\frac{S_g}{C} - \frac{S_a}{\rho_a D c_p} \right) \,. \tag{6.76}$$

The analogy between (6.73) and (6.76) should be clear. The two equations are essentially the same if the transfer coefficient c_T has a power-law dependence on $T_g - T_a$, which is a realistic possibility.

From what we have already discussed, it should seem plausible that an implicit scheme would be a good choice for (6.73), i.e.,

$$\frac{q^{n+1} - q^n}{\Delta t} = -\kappa \left(q^{n+1} \right)^{P+1} + S \,. \tag{6.77}$$

Such a scheme is in fact unconditionally stable, but for arbitrary P it must be solved iteratively, which can be expensive. For this practical reason, (6.77) may not be considered a viable choice, except where P is a small integer, in which case (6.77) can be solved analytically.

As mentioned earlier, linearization about an equilibrium solution is a necessary preliminary step before von Neumann's method can be applied to a nonlinear equation. Let $\overline{q}$ denote an equilibrium solution of (6.73), so that

$$\kappa \overline{q}^{P+1} = S \,. \tag{6.78}$$

We are assuming for simplicity that S is independent of q and time. Let q' denote a departure from the equilibrium, so that $q = \overline{q} + q'$. Then (6.74) can be linearized as follows:

$$\frac{d}{dt}\left(\overline{q}+q'\right) = -\kappa\overline{q}^{P+1} - \kappa\left(P+1\right)\overline{q}^{P}q' + S\,, \tag{6.79}$$

which reduces to

$$\frac{dq'}{dt} = -\kappa\left(P+1\right)\overline{q}^{P}q'\,. \tag{6.80}$$

This linearized equation with constant coefficients can be analyzed using von Neumann's method.

As an example, the forward time-differencing scheme, applied to (6.80), gives

$$q^{n+1} - q^{n} = -\alpha\left(P+1\right)q^{n}\,, \tag{6.81}$$

where we use the shorthand notation

$$\alpha \equiv \kappa\left(\overline{q}\right)^{P}\Delta t\,, \tag{6.82}$$

and we have dropped the "prime" notation for simplicity. We can rearrange (6.81) to

$$q^{n+1} = \left[1-\alpha\left(P+1\right)\right]q^{n}\,, \tag{6.83}$$

from which we see that

$$\lambda = 1-\alpha\left(P+1\right)\,. \tag{6.84}$$

From (6.84), we see that the forward time-differencing scheme is conditionally stable, and that the criterion for stability is more difficult to satisfy when P is large.

Table 6.2 summarizes the ten schemes that Kalnay and Kanamitsu analyzed, and gives the amplification factors and stability criteria for each. For the nonlinear forward explicit,

Table 6.2: Schemes for the nonlinear decay equation, as studied by Kalnay and Kanamitsu (1988). The source/sink term is omitted here for simplicity.

Name of Scheme	Form of Scheme	Amplification Factor	Linear stability criterion
Forward Explicit	$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(q^n\right)^{P+1}$	$1-\alpha(P+1)$	$\alpha(P+1)<2$
Backward Implicit	$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(q^{n+1}\right)^{P+1}$	$\frac{1}{1+\alpha(P+1)}$	Unconditionally stable
Centered Implicit	$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(\frac{q^n+q^{n+1}}{2}\right)^{P+1}$	$1-\alpha(P-1)+\frac{(\alpha P)^2}{(1+\alpha)^2}$	Unconditionally stable
Explicit coefficient, implicit q	$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P q^{n+1}$	$\frac{1-\alpha P}{(1+\alpha)^2}$	$\alpha(P-1)<2$
Predictor-Corrector coefficient, implicit q	$\frac{\hat{q}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$ $\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(\hat{q}\right)^P q^{n+1}$	$\frac{1-\alpha(P-1)+(\alpha P)^2}{(1+\alpha)^2}$	$\alpha(P-1)<1$
Average coefficient, implicit q	$\frac{\hat{q}-q^n}{\Delta t} = \kappa\left(q^n\right)^P \hat{q}$ $\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left[\frac{\kappa\left(q^n\right)^P+\left(\hat{q}\right)^P}{2}\right]q^{n+1}$	$\frac{1-\alpha(P-1)-\frac{\alpha^2 P}{2}+\frac{(\alpha P)^2}{2}}{(1+\alpha)^2}$	$\alpha(P-2)<2$
Explicit coefficient, extrapolated q	$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P\left[(1-\gamma)q^n\right]$	$\frac{1-\alpha(P+1-\gamma)}{1+\alpha\gamma}$	$\alpha(P+1-2\gamma)<2$
Explicit coefficient, implicit q, with time filter	$\frac{\hat{q}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$ $q^{n+1} = (1-A)\hat{q}+Aq^n$	$\frac{(1-A)(1-\alpha P)}{1+\alpha}+A$	$\alpha\left[P(1-A)-1-A\right]<2$
Double time step, explicit coefficient, implicit q with time average filter	$\frac{\hat{q}-q^n}{2\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$ $q^{n+1} = \frac{\hat{q}+q^n}{2}$	$\frac{1-\alpha(P-1)}{1+2\alpha}$	$\alpha(P-3)<2$
Linearization of backward implicit scheme	$\frac{q^{n+1}-q^n}{\Delta t} =$ $-\kappa\left(q^n\right)^P\left[(P+1)q^{n+1}-Pq^n\right]$	$\frac{1+\alpha P}{1+\alpha(P+1)}$	Unconditionally stable

backward implicit and centered implicit schemes, the amplification factor has been obtained by linearization, but the results obtained are not misleading. In the table, A is a

parameter used to adjust the properties of a time filter; and $\gamma > 1$ is an "extrapolation" parameter. For a detailed discussion, see the paper by Kalnay and Kanamitsu (1988), which you should find quite understandable at this stage.

6.8 Summary

Schemes with smaller discretization errors are not always better. For example, the second-order leapfrog scheme is unstable when applied to the decay equation, while the first-order backward implicit scheme is unconditionally stable and well behaved for the same equation. A stable but less accurate scheme is obviously preferable to an unstable but "more accurate" scheme. Accuracy in the Taylor-series sense doesn't tell the whole story.

For the advection and oscillation equations, discretization errors can be separated into amplitude errors and phase errors. Neutral schemes, like the leapfrog scheme and the trapezoidal implicit scheme, have phase errors, but no amplitude errors.

Implicit schemes are well suited to the decay equation, but can be difficult to implement when the decay term is nonlinear.

Computational modes in time are permitted by differencing schemes that involve three or more time levels. To control these modes, there are four possible approaches:

- Choose a scheme that involves only two time levels.
- Choose the computational initial condition well, and periodically "re-start" the model by taking a two-level time step.
- Choose the computational initial condition well, and use a time filter (e.g., Asselin (1972)) to suppress the computational mode.
- Choose the computational initial condition well, and choose a scheme that damps the computational mode more than the physical mode, e.g., an Adams- Bashforth scheme.

Finally, we list some properties of "good" schemes:

- High accuracy.
- Stability.
- Simplicity.
- Computational economy.
- No computational modes in time, or else damped computational modes in time.
- Graceful behavior in the limit of large time steps, as with the backward implicit scheme applied to the decay equation.

6.9 Problems

1. Consider the following pair of equations, which describe inertial oscillations:

$$\frac{Du}{Dt} = fv\,, \tag{6.85}$$

$$\frac{Dv}{Dt} = -fu\,. \tag{6.86}$$

   (a) Determine the stability of the Euler forward time-differencing scheme as applied to these two equations.

   (b) Determine the stability of the trapezoidal implicit time-differencing scheme as applied to these two equations.

   (c) Determine the stability of third-order Adams-Bashforth time-differencing scheme as applied to these two equations.

2. (a) Find the exact solution of

$$\frac{dq}{dt} = i\omega q - \kappa q. \tag{6.87}$$

   Let $q(t=0) = 100$, $\dfrac{\omega}{2\pi} = 0.1$, $\kappa = 0.1$. Plot the real part of the solution for $0 \le t \le 100$.

   (b) Find the stability criterion for the scheme given by

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1}\,. \tag{6.88}$$

$$\lambda^2 - 1 = 2i\Omega\lambda - 2K\lambda^2 \tag{6.89}$$

$$\lambda^2(1+2K) - 2i\Omega\lambda - 1 = 0 \tag{6.90}$$

The solutions are

$$\begin{aligned}\lambda =& \frac{2i\Omega \pm \sqrt{-4\Omega^2+4\left(1+2K\right)}}{2\left(1+2K\right)} \\ =& \frac{i\Omega \pm \sqrt{-\Omega^2+\left(1+2K\right)}}{1+2K}\end{aligned} \tag{6.91}$$

For $-\Omega^2+(1+2K)\geq 0$, we can write

$$\begin{aligned}|\lambda|^2 = \frac{\Omega^2+\left[-\Omega^2+(1+2K)\right]}{(1+2K)^2} & \\ =& \frac{(1+2K)}{(1+2K)^2} \\ =& \frac{1}{1+2K} < 1\end{aligned} \tag{6.92}$$

so the scheme is unconditionally stable.

For $-\Omega^2+(1+2K)<0$, we get

$$\lambda = \left(\frac{i}{1+2K}\right)\left[\Omega \pm \sqrt{\Omega^2-(1+2K)}\right] \tag{6.93}$$

Note that λ is pure imaginary, so that stability boundary occurs for $\lambda = i$ or $\lambda = -i$. For $\lambda = i$, we get

$$1+2K-\Omega = \pm\sqrt{\Omega^2-(1+2K)} \tag{6.94}$$

Squaring both sides, we obtain

$$(1+2K-\Omega)^2 = \Omega^2-(1+2K) \tag{6.95}$$

which leads to

$$\Omega = 1 + K \tag{6.96}$$

This defines a line that moves up towards the right, starting at $\Omega = 1$, $K = 0$. We have instability for $\Omega > 1 + K$, *provided that* $-\Omega^2 + (1 + 2K) < 0$. This second condition is equivalent to $\Omega > \sqrt{1 + 2K}$. It is easy to show that $1 + \kappa > \sqrt{1 + 2\kappa}$ for all κ. So the conclusion is that the stability boundary is $\Omega = 1 + \kappa$. Instability occurs for $\Omega > 1 + \kappa$.

For $\lambda = -i$, we find that

$$-(1 + 2K) - \Omega = \pm\sqrt{\Omega^2 - (1 + 2K)} \tag{6.97}$$

Squaring both sides leads to

$$\Omega = -(1 + K) \tag{6.98}$$

(c) Repeat for the scheme given by

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n-1} \, . \tag{6.99}$$

$$\lambda^2 - 1 = 2i\Omega\lambda - 2K \tag{6.100}$$

or

$$\lambda^2 - 2i\Omega\lambda + (2K - 1) = 0 \tag{6.101}$$

The solutions are

$$\begin{aligned}\lambda =& \frac{2i\Omega \pm \sqrt{-4\Omega^2 - 4(2K-1)}}{2} \\ =& i\Omega \pm \sqrt{-\Omega^2 - (2K-1)}\end{aligned} \tag{6.102}$$

For $-\Omega^2 - (2K-1) \geq 0$, we get

$$|\lambda|^2 = 1 - 2K \tag{6.103}$$

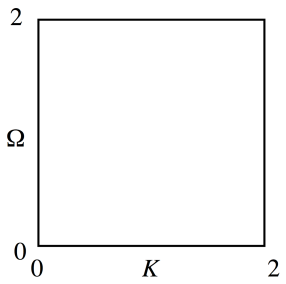
This says that the amplification factor is independent of Ω. Note, however, that we have assumed that $-\Omega^2 - (2K-1) \geq 0$, which implies that $1 - 2K > \Omega^2 > 0$. So $|\lambda|^2$ cannot be negative. In other words, everything is OK.

For $-\Omega^2 - (2K-1) < 0$, we get

$$\lambda = i\left[\Omega \mp \sqrt{\Omega^2 + (2K-1)}\right] \tag{6.104}$$

Here again, λ is pure imaginary. For $\lambda = i$, we get $\Omega = 1 - K$, and for $\lambda = -1$, we get $\Omega = -(1-K)$.

(d) Plot the neutral stability boundaries for both schemes (where $|\lambda| = 1$) as curves in the $(K,\ \Omega)$ plane, for K and Ω in the range 0 to 2, as in the sketch below. Here $\Omega \equiv \omega\Delta t$, $K \equiv \kappa\Delta t$. Indicate which part(s) of the (K, Ω) plot correspond to instability.



(e) Code equations (6.88) and (6.99). Use an Euler-forward time step for the first step only. Use $q(t=0)=100$, and $\Delta t = 1$. For each scheme, run the following cases:

$$\frac{\omega}{2\pi} = 0.1,\ \kappa = 0; \quad \frac{\omega}{2\pi} = 0,\ \kappa = 0.1; \quad \frac{\omega}{2\pi} = 0.1,\ \kappa = 0.1\ . \tag{6.105}$$

For each case, plot $\mathrm{Re}\{q\}$ for $0 \le t \le 100$ and compare with the exact solution. Discuss the numerical results as they relate to the stability analyses of parts (b), (c), and (d) above.

3. Find the stability criterion for the fourth-order Runge-Kutta scheme applied to the oscillation equation.

4. Work out the stability criteria for the Matsuno scheme and the Heun scheme as applied to the decay equation, and compare with the corresponding criteria for the backward implicit and trapezoidal implicit schemes, respectively.

5. Plot θ as a function of Ω for the exact solution of the oscillation equation, and for the Euler, trapezoidal, Matsuno, and Heun schemes, and also the leapfrog scheme's physical mode, q_1. Consider $-\pi \le \Omega \le \pi$.

Chapter 7

Riding along with the air

Up to now we have considered only ordinary differential equations in which the independent variable is time. Starting in this chapter, we will consider partial differential equations, involving both time and space derivatives. The subject of this chapter is advection, which is an extremely important process in the atmosphere. The purpose of this chapter is to review the physical nature of advection, as a prelude to a discussion, in the following chapters, of numerical schemes for the advection equation.

7.1 The Lagrangian form

Consider an arbitrary "intensive" variable A. An intensive variable is defined per unit mass. Familiar examples are the mixing ratio of water vapor, and temperature, which is proportional to the internal energy per unit mass. In Lagrangian form, the advection equation for an intensive variable A, in any number of dimensions, is

$$\frac{DA}{Dt} = 0 \,. \tag{7.1}$$

This simply means that the value of A does not change following a particle. We say that A is "conserved" following a particle. In fluid dynamics, we consider an infinite collection of fluid particles. According to (7.1), each particle maintains its value of A as it moves. If we do a survey of the values of A in our fluid system, let advection occur, and conduct a "follow-up" survey, we will find that exactly the same values of A are still in the system. The locations of the particles presumably will have changed, but the maximum value of A over the population of particles is unchanged by advection, the minimum value is unchanged, the average is unchanged, and in fact *all of the statistics of the distribution of A over the mass of the fluid are completely unchanged by the advective process*. This is an essential property of advection.

Here is another way of describing this property: If we worked out the probability density function (PDF) of A, by defining narrow "bins" and counting the mass associated with particles having values of A falling within each bin, we would find that the PDF was unchanged by advection. For instance, if the PDF of A at a certain time is Gaussian (or "bell shaped"), it will still be Gaussian at a later time (and with the same mean and standard deviation) if the only intervening process is advection and if no mass enters or leaves the system.

Pure advection is time-reversible. Ideally a numerical scheme for advection should also be time-reversible. As explained later, this can be done, although it is not usually done.

Consider a simple function of A, such as A^2. Since the A of each particle is unchanged during advection, A^2 will also be unchanged. In fact, any function of A will also be unchanged. It follows that the PDF of any function of A is unchanged by advection.

In many cases of interest, A is non-negative more or less by definition. For example, the mixing ratio of water vapor cannot be negative. Some other variables, such as the zonal component of the wind vector, can be either positive or negative.

Suppose that A is conserved following each particle. If there are no negative values of A at some initial time, then, to the extent that advection is the only process at work, there will be no negative values of A at any later time either. This is true whether the variable in question is non-negative by definition (like the mixing ratio of water vapor) or not (like the zonal component of the wind vector).

Of course, in general these various quantities are not really conserved following particles, because various sources and sinks cause the value of A to change as the particle moves. For instance, if A is temperature, one possible source is adiabatic expansion and compression, and another is radiative heating. To describe more general processes that include not only advection but also sources and sinks, we can replace (7.1) by

$$\frac{DA}{Dt} = S, \tag{7.2}$$

where S is the source of A per unit time. (A negative value of S represents a sink.) We still call (7.2) a "conservation" equation; it says that A is conserved *except* when sources or sinks come into play. We call (7.2) the *"Lagrangian form"* of the conservation equation.

7.2 The advective form

The Lagrangian (particle-following) time derivative, D/Dt, can be expanded as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \, . \tag{7.3}$$

To demonstrate this, let $\mathbf{r}$ be the position vector that corresponds to the particle's location, and write the Lagrangian time derivative of an arbitrary variable as

$$\frac{DA}{Dt} = \frac{A(\mathbf{r} + \Delta\mathbf{r},\, t + \Delta t) - A\,(\mathbf{r},\, t)}{\Delta t} \, . \tag{7.4}$$

Think of the two values of A on the right-hand side of (7.4) as "measured values" for a particular particle. The measurements are made at times t and $t + \Delta t$, when the particle's positions are $\mathbf{r}$ and $\mathbf{r} + \Delta\mathbf{r}$, respectively. Using a Taylor series expansion about the point $\mathbf{r}$, t, we can write

$$A\,(\mathbf{r} + \Delta\mathbf{r},\, t + \Delta t) - A(\mathbf{r},\, t) = \frac{\partial A}{\partial t}\Delta t + \nabla A \cdot \Delta\mathbf{r} + \text{higher order terms} \, . \tag{7.5}$$

Here the partial derivative with respect to time is taken at fixed $\mathbf{r}$. Dividing both sides of (7.5) by Δt, and using $\Delta\mathbf{r}/\Delta t \equiv \mathbf{V}$, we obtain

$$\frac{DA}{Dt} = \frac{\partial A}{\partial t} + (\mathbf{V} \cdot \nabla) A \, . \tag{7.6}$$

The left-hand side of (7.6) is the change of A experienced by a moving particle, and the right-hand side is the sum of the time-rate-of-change of as seen in a fixed "Eulerian" coordinate system, and a term representing the effects of advection as seen in the Eulerian framework.

The individual terms on the right-hand side of (7.6) depend on the Eulerian coordinate system used. The time-rate-of-change "at a fixed point in space" means one thing in an inertial frame of reference, and something very different in a frame of reference that is rotating with the Earth. Similarly, $\mathbf{V}$ takes one value in an inertial frame and a different value in the rotating frame. Nevertheless, the left-hand side of (7.6) has a meaning that is independent of the frame of reference. Therefore, the *sum of the terms on the right-hand side* must be also independent of the frame of reference.

Using (7.6), we can rewrite (7.2) as

$$\frac{\partial A}{\partial t} = -(\mathbf{V}\cdot\nabla)A + S\,. \tag{7.7}$$

With (7.7), we predict A at a particular point using information about the values of A at nearby points. We will call (7.7) the *"advective form"* of the conservation equation.

7.3 The continuity equation

In addition to conservation equations for quantities that are defined per unit mass, we need a conservation equation for mass itself. This "continuity equation" can be written as

$$\frac{\partial \rho}{\partial t} = -\nabla\cdot(\rho\mathbf{V})\ , \tag{7.8}$$

where ρ is the density (mass per unit volume) and $\mathbf{V}$ is the velocity vector. This says that the density at a point changes in time due to the convergence or divergence of the "mass flux" $\rho\mathbf{V}$. We call this the "flux form" of the continuity equation.

Using (7.6), can write (7.8) as

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\mathbf{V}\,. \tag{7.9}$$

Eq. (7.9) tells us that, in general, ρ *is not constant following a particle*. For the limiting case of an incompressible fluid, i.e., one for which ρ is an immutable property of the fluid so that $D\rho/Dt = 0$, (7.9) implies that $\nabla\cdot\mathbf{V} = 0$. This means that incompressibility implies nondivergence of the wind. In reality, there is no such thing as an incompressible fluid, but some fluids (like liquid water) are much less compressible than others (like air). Eq. (7.9) also tells us that if the wind field happens to be non-divergent, then the density is conserved following a particle.

7.4 The flux form

Multiply (7.8) by A, and (7.7) by ρ, and add the results to obtain

$$\frac{\partial}{\partial t}(\rho A) = -\nabla\cdot(\rho\mathbf{V}A) + \rho S\,. \tag{7.10}$$

This is called the *"flux form"* of the conservation equation. Like (7.7), it is Eulerian. Notice that if we put $A \equiv 1$ and $S \equiv 0$ (because there is no "source of 1"!) then (7.10) reduces to (7.8). This is an important point that can and should be used in the design of advection schemes. We can and should design the flux form of an advection scheme for A in such a way that for $A \equiv 1$ we get the scheme that we use for the continuity equation.

Obviously we could use the continuity equation to go back from the flux form to the advective form. Throughout the rest of this book we will frequently use continuity to transform back and forth between flux form and advective form.

Suppose that we integrate (7.8) over a closed or periodic domain, R. Here"closed" means that there is no flux of mass across the boundary of R, and "periodic" means that the domain has no boundaries (e.g., a spherical shell). For *either* closed or periodic boundaries, Gauss's Theorem tells us that

$$\int_R \nabla \cdot (\rho \mathbf{V} A) \; dR = 0 \, . \tag{7.11}$$

It follows that

$$\frac{d}{dt} \int_R \rho A \, dR = \int_R \rho S \, dR \, . \tag{7.12}$$

This says that the mass-weighted total of A within the domain does not change with time, except for the effects of sources and sinks. Similarly, we find that

$$\frac{d}{dt} \int_R \rho \, dR = 0 \, , \tag{7.13}$$

which simply states that the total mass within the domain does not change with time. We can describe (7.12) and (7.13) as "integral forms" of the conservation equations for mass and A, respectively.

7.5 Characteristics

Consider the one-dimensional advection equation, given by

$$\left(\frac{\partial A}{\partial t} \right)_x + u \left(\frac{\partial A}{\partial x} \right)_t = 0 \, , \tag{7.14}$$

where $A = A(x,t)$. This is the advective form, with the source term set to zero. We will assume for now that u is independent of both x and t. Eq. (7.14) is a first-order linear partial differential equation with a constant coefficient, namely u. It looks harmless, but it causes no end of trouble.

Suppose that

$$A\,(x,\,0) = F\,(x) \text{ for } -\infty < x < \infty\,. \tag{7.15}$$

This is an "initial condition," because it gives the spatial distribution of A at $t = 0$. Our goal is to determine $A(x,t)$. We first work out the analytic solution of (7.14), for later comparison with our numerical solution. Define

$$\xi \equiv x - ut, \text{ so that } \left(\frac{\partial x}{\partial t}\right)_{\xi} = u\,. \tag{7.16}$$

At $t = 0$, $x = \xi$; in other words, $x = \xi$ is the initial position of the particle. In general, $x = \xi + ut$ is the position of the particle at time t. Each particle has its own initial position, and it's own value of ξ. Think of a particle's value of ξ as the particle's "name." The value of ξ does not change as the particle moves, so we can think of it as labeling the particle that was at $x = \xi$ when $t = 0$.

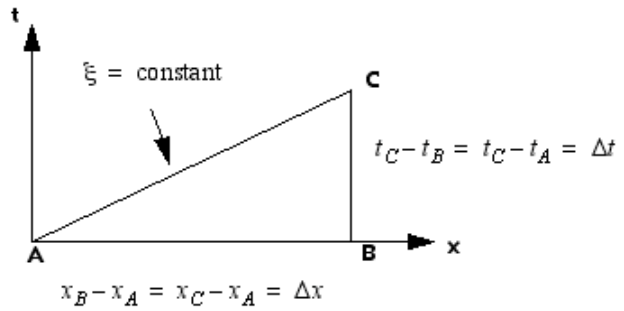


Figure 7.1: Figure used in the derivation of the first line of (7.17). XX Redraw this figure.

With reference to Fig. 7.1, we can write

$$
\begin{aligned}
\left(\frac{\partial A}{\partial x}\right)_\xi &= \frac{A_C - A_A}{x_C - x_A} \\
&= \frac{(A_B - A_A) + (A_C - A_B)}{x_C - x_A} \\
&= \frac{(A_B - A_A) + (A_C - A_B)}{x_C - x_A} \\
&= \frac{(A_B - A_A)}{x_B - x_A} + \frac{A_C - A_B}{t_C - t_B}\frac{t_C - t_B}{x_C - x_B} \\
&= \left(\frac{\partial A}{\partial x}\right)_t + \left(\frac{\partial A}{\partial t}\right)_x \left(\frac{\partial t}{\partial x}\right)_\xi \\
&= \left(\frac{\partial A}{\partial x}\right)_t + \left(\frac{\partial A}{\partial t}\right)_x \frac{1}{u} \\
&= 0 \,.
\end{aligned}
\tag{7.17}
$$

Similarly,

$$
\begin{aligned}
\left(\frac{\partial A}{\partial t}\right)_\xi &= \left(\frac{\partial A}{\partial t}\right)_x + \left(\frac{\partial A}{\partial x}\right)_t \left(\frac{\partial x}{\partial t}\right)_\xi \\
&= \left(\frac{\partial A}{\partial t}\right)_x + \left(\frac{\partial A}{\partial x}\right)_t u \\
&= \frac{DA}{Dt} \\
&= 0 \,.
\end{aligned}
\tag{7.18}
$$

We can interpret $(\partial A/\partial t)_\xi$ as the Lagrangian time rate of change of A. From (7.17) and (7.18), we conclude that

$$
A = f(\xi) \tag{7.19}
$$

is the general solution to (7.14). This means that A depends only on ξ, in the sense that if you tell me the value of ξ, that's all the information I need to tell you the value of A. (Note that "A depends only on ξ" does *not* mean that A is independent of x for fixed t, or of t for fixed x.)

The initial condition is

$$\xi \equiv x \text{ and } A(x) = f(x) \text{ at } t = 0\,, \tag{7.20}$$

i.e., the shape of $f(\xi)$ is determined by the initial condition. In order to satisfy the initial condition, we chose $f \equiv F$ [see Eq. (7.15)]. Referring to (7.19) , we see that $A(\xi) = F(\xi) \equiv F(x - ut)$ is the solution to (7.14) that satisfies the initial condition (7.20). The initial values of A simply "move along" the lines of constant ξ, which are called *characteristics*. The initial shape of $A(x)$, namely $F(x)$, is just carried along by the wind. From a physical point of view this is obvious.

The discussion above can be generalized to define "curvy" characteristics in two or three spatial dimensions, with winds that vary in both space and time.

Partial differential equations whose solutions are constant along characteristics are called *hyperbolic*. The advection equation is hyperbolic. Further discussion is given later.

7.6 Discussion

It may seem that the ideal way to simulate advection in a model would be to define a collection of particles, associate various properties of interest with each particle, and let the particles be carried about by the wind. In such a *Lagrangian model*, the properties associated with each particle would include its spatial coordinates, e.g., its longitude, latitude, and height. These would change with time in response to the predicted velocity field. The Lagrangian approach is discussed in Chapter 14.

At the present time, virtually all models in atmospheric science are based on Eulerian descriptions of horizontal advection. For the case of vertical advection, the Eulerian vertical coordinate is sometimes permitted to "move" in the sense that the height or pressure of a coordinate surface changes as the circulation evolves (e.g., Phillips, 1957; Hsu and Arakawa, 1990). These vertical coordinate systems are discussed in Chapter 22.

All three forms of the conservation equation for A imply that if A is uniform throughout the domain (e.g., if $A \equiv 1$), and if $S \equiv 0$ throughout the domain, then A will remain uniform under advection. In the case of the flux form, putting $A =$ constant (and $S_A = 0$) converts the conservation equation for A into the continuity equation.

Chapter 8

The upstream scheme for advection

8.1 From there and then to here and now

We now investigate the solution of one possible numerical scheme for (7.14). We construct a grid, as in Fig. 8.1. One of the infinitely many possible finite-difference approximations to (7.14) is

$$\boxed{\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) = 0 \quad \text{for} \quad u \geq 0\,.} \tag{8.1}$$

Here we have used the forward difference quotient in time and the backward difference quotient in space, and we have stipulated that $u \geq 0$. If we know A_j^n at some time level n for all j, then we can solve (8.1) for A_j^{n+1} at the next time level, $n+1$. Eq. (8.1) is called the "*upstream*" scheme. It is one-sided or asymmetric in both space and time. It seems naturally suited to modeling advection, in which air comes from the upstream side and goes to the downstream side, as time passes by. The upstream scheme has some serious weaknesses, but it also has some very useful properties. It is a scheme worth remembering. That's why I put (8.1) in a box.

Because

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} \to \frac{\partial A}{\partial t} \text{ as } \Delta t \to 0\,, \tag{8.2}$$

and

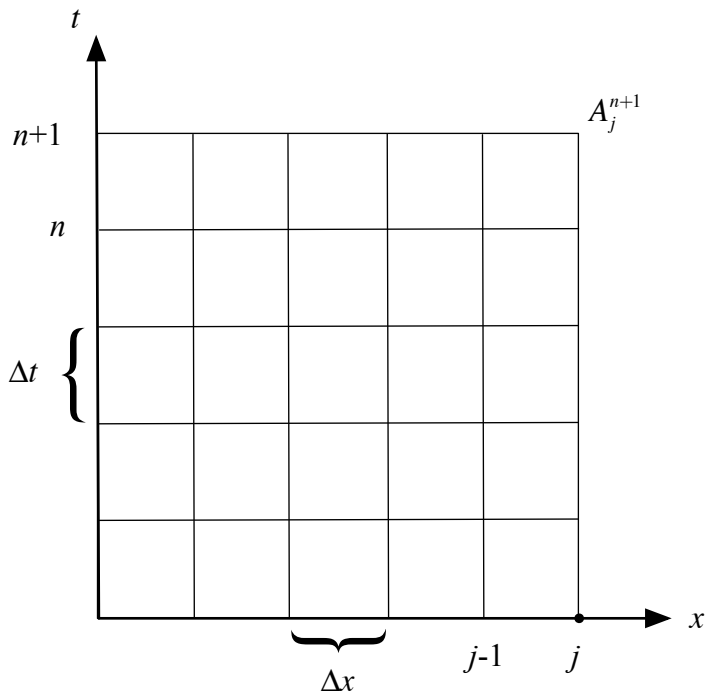


Figure 8.1: A grid for the solution of the one-dimensional advection equation.

$$\frac{A_j^n - A_{j-1}^n}{\Delta x} \to \frac{\partial A}{\partial x} \text{ as } \Delta x \to 0 \,, \tag{8.3}$$

we can say that (8.1) does approach (7.14) as Δt and Δx both approach zero. Given this fact, it may seem obvious that the *solution* of (8.1) approaches the *solution* of (7.14) as $\Delta x \to 0$ and $\Delta t \to 0$. I am now going to show you why that is not necessarily true.

8.2 The discretization error of the upstream scheme

Let $A(x,\, t)$ denote the (exact) solution of the differential equation, so that $A(j\Delta x,\, n\Delta t)$ is the value of this exact solution at the discrete point $(j\Delta x,\, n\Delta t)$ on the grid shown in Fig. 8.1. We use the notation A_j^n to denote the "exact" solution of a finite-difference equation, at the same point. In general, $A_j^n \neq A(j\Delta x,\, n\Delta t)$. To find the discretization error of the upstream scheme, we substitute the solution of the differential equation into the finite-difference equation. For the upstream scheme given by (8.1), we get

$$\left\{\frac{A\left[j\Delta x,\ (n+1)\,\Delta t\right]-A\left(j\Delta x,\ nvarDeltat\right)}{\Delta t}\right\}+u\left\{\frac{A\left(j\Delta x,\ n\Delta t\right)-A\left[(j-1)\,\Delta x,\ n\Delta t\right]}{\Delta x}\right\}=\varepsilon\,. \tag{8.4}$$

Recall that the discretization error, ε, is a measure of how accurately the solution $A(x,\ t)$ of the original differential equation (7.14), satisfies the finite-difference equation, (8.1). It is far from a perfect measure of accuracy, however, as will become evident.

If we obtain the terms of (8.4) from a Taylor Series expansion of $A(x,\ t)$ about the point $(j\Delta x,\ n\Delta t)$, *and use the fact that* $A(x,\ t)$ *satisfies* (7.14), we find that

$$\varepsilon=\left(\frac{1}{2!}\Delta t\frac{\partial^2 A}{\partial t^2}+\cdots\right)+u\left(-\frac{1}{2!}\Delta x\frac{\partial^2 A}{\partial x^2}+\cdots\right)\,. \tag{8.5}$$

We say this is a "first-order scheme" because the first powers of Δt and Δx appear in (8.5). The notations $\mathcal{O}(\Delta t,\Delta x)$ or $\mathcal{O}(\Delta t)+\mathcal{O}(\Delta x)$ can be used to express this. The upstream scheme is first-order accurate in both space and time.

A scheme is said to be "consistent" with the differential equation if the discretization error of the scheme approaches zero as Δt *and* Δx *approach zero.* We have demonstrated above that the upstream scheme is consistent. Consistency is necessary, but it is a "low bar," and nowhere near sufficient to make a good scheme.

8.3 The domain of dependence

Given acceptable levels of discretization error, we must also consider the error of the *solution* of the discrete equation, i.e., the difference between the *solution* of the discrete equation and the *solution* of the continuous differential equation, i.e., $A_j^n-A\left(j\Delta x,\ n\Delta t\right)$. How does the solution of the finite-difference scheme change as Δt and $\Delta x\rightarrow 0$? If the solution of the finite-difference scheme approaches the solution of the differential equation as the grid is refined, then we say that the solution *converges*.

Fig. 8.2 illustrates a situation in which the solution *does not converge* as the grid is refined. The thin diagonal line in the figure shows the *characteristic* along which A is "carried," i.e., A is constant along the line. This is the exact solution. To work out the numerical approximation to this solution, we first choose Δx and Δt such that the grid points are the dots in the figure. The set of grid points carrying values of A on which A_j^n depends is called the *"domain of dependence."* The shaded area in the figure shows the domain of dependence for the upstream scheme, (8.1).

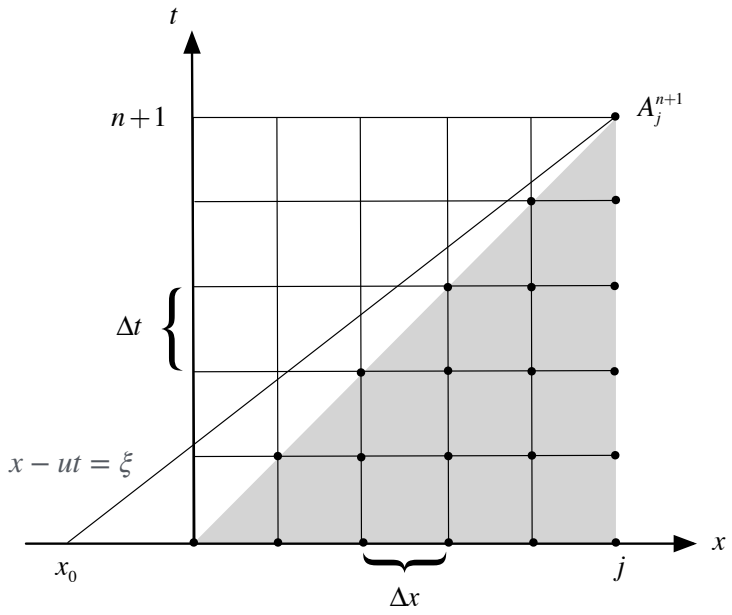


Figure 8.2: The shaded area represents the "domain of dependence" of the solution of the upstream scheme at the point $[j\Delta x, (n+1)\,\Delta t]$. The diagonal line is the characteristic of the exact solution that passes through that point.

We could increase the accuracy of the finite-difference approximation to $\partial A/\partial t$ by cutting Δt in half, and we can increase the accuracy of the finite-difference approximation to $\partial A/\partial x$ by cutting Δx in half, but the domain of dependence would not change as long as the nondimensional ratio $u\Delta t/\Delta x$ remains the same. The preceding discussion shows that $u\Delta t/\Delta x$ is an important quantity. We will give it a name:

$$\mu \equiv \frac{u\Delta t}{\Delta x}\,. \tag{8.6}$$

Consider the characteristic that passes through the point $[j\Delta x, (n+1)\,\Delta t]$, i.e., $x - ut = x_0$, where x_0 is a constant. For the case shown in Fig. 8.2, the characteristic does not lie in the domain of dependence. As a result, there is no hope of obtaining smaller discretization error, no matter how small Δx and Δt become, as long as mu is unchanged, because the true solution depends only on the initial value of A at the single point $(x_0, 0)$ which cannot influence A_j^n. You could change $A\,(x_0, 0)$ (and hence the exact solution $A\,(j\Delta x,\ n\Delta t)$), but the computed solution A_j^n would remain the same. In such a case, the error of the solution usually will not be decreased by refining the grid. We conclude that if the value of u is such that x_0 lies outside of the domain of dependence, it is not possible for the solution of the finite-difference equation to approach the solution of the differential equation, no matter how fine the mesh becomes. The finite-difference equation converges to the differential

equation, but the solution of the finite-difference equation does not converge to the solution of the differential equation. *This illustrates that it is possible to decrease the truncation errors without decreasing the discretization error.* The truncation error goes to zero, but the discretization error does not. Bummer.

The condition for x_0 to lie inside the domain of dependence is

$$0 \leq \mu \leq 1 \,. \tag{8.7}$$

This a *necessary* condition for convergence of the upstream scheme. Eq. (8.7) is a form of the famous "CFL" stability criterion associated with the names Courant, Friedrichs and Lewy (Courant et al., 1928).

Notice that if u is negative (giving what we might call a "downstream" scheme), then the characteristic lies outside the domain of dependence shown in the figure. Of course, we can use

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^n - A_j^n}{\Delta x}\right) = 0 \quad \text{for} \quad u \leq 0 \,, \tag{8.8}$$

in place of (8.1). For $u < 0$, Eq. (8.8) is the appropriate form of the upstream scheme.

A computer program can have an "if-test" that checks the sign of u, and uses (8.1) if $u \geq 0$, and (8.8) if $u < 0$. If-tests can cause slow execution on certain types of computers, however, and besides, if-tests are ugly and reduce the readability of a code. If we define

$$u^+ \equiv \frac{u + |u|}{2} \geq 0, \text{ and } u^- \equiv \frac{u - |u|}{2} \leq 0 \,, \tag{8.9}$$

then a "generalized" upstream scheme can be written as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u^-\left(\frac{A_{j+1}^n - A_j^n}{\Delta x}\right) + u^+\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) = 0 \,. \tag{8.10}$$

This form avoids the use of *if*-tests and is also convenient for use in pencil-and-paper analysis.

8.4 Interpolation and extrapolation

Referring back to (8.1), we can rewrite the upstream scheme as

$$A_j^{n+1} = A_j^n (1-\mu) + A_{j-1}^n \mu \ . \tag{8.11}$$

This scheme has the form of either an *interpolation* or an *extrapolation*, depending on the value of μ. To see this, refer to Figure 8.3. Along the line plotted in the figure

$$\begin{aligned} A &= A_{j-1}^n - \left(x - x_{j-1}\right) \left(\frac{A_j^n - A_{j-1}^n}{x_j - x_{j-1}} \right) \\ &= A_j^n \left[1 - \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \right) \right]_j + A_{j-1}^n \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \right) \ , \end{aligned} \tag{8.12}$$

which has the same form as our scheme if we identify

$$A \equiv A_j^{n+1} \text{ and } \mu \equiv \frac{x - x_{j-1}}{x_j - x_{j-1}} \ . \tag{8.13}$$

For $0 \leq \mu \leq 1$ we have *interpolation*. For $\mu < 0$ or $\mu > 1$ we have *extrapolation*.

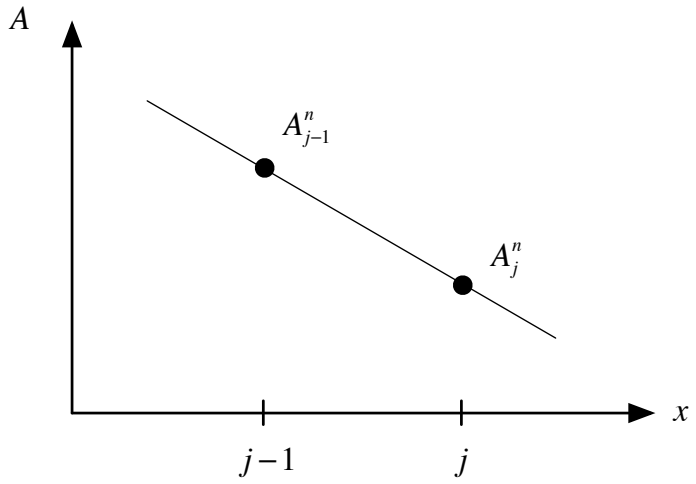


Figure 8.3: Diagram illustrating the concepts of interpolation and extrapolation. See text for details.

For the case of interpolation, the value of A_j^{n+1} will be intermediate between A_{j-1}^n and A_j^n, so it is impossible for A_j^{n+1} to "blow up," no matter how many time steps have been

taken. The repeated interpolation can lead to an unrealistic smoothing of the solution, however.

Interpolation also implies that if A_{j-1}^n and A_j^n are both positive, then A_j^{n+1} will be positive too. This is a good thing, for example, if A represents the mixing ratio of water vapor. More discussion is given later.

For the case of extrapolation, A_j^{n+1} will lie outside the range of A_{j-1}^n and A_j^n. This suggests that $|A|$ may increase with time.

Both interpolation and extrapolation are used extensively in atmospheric modeling. Much more discussion is given later.

8.5 The computational stability of the upstream scheme

We will now use the direct method, the energy method, and von Neumann's method to test the stability of the upstream scheme for advection.

8.5.1 The direct method

The direct method checks stability by asking whether the largest absolute value of A_j^{n+1} anywhere on the grid increases with time. It it does, the scheme is unstable. Recall that with the upstream scheme A_j^{n+1} is a weighted mean of A_j^n and A_{j-1}^n. Provided that (8.7) is satisfied, Eq. (8.11) implies that

$$\left|A_j^{n+1}\right| \le \left|A_j^n\right|(1-\mu) + \left|A_{j-1}^n\right|\mu \ . \tag{8.14}$$

Summing over the grid, we find that

$$\begin{aligned} \max_{\text{all j}} \left|A_j^{n+1}\right| &\le \max_{\text{all j}} \left|A_j^n\right|(1-\mu) + \max_{\text{all j}} \left|A_{j-1}^n\right|\mu \\ &= \max_{\text{all j}} \left|A_j^n\right| \quad , \end{aligned} \tag{8.15}$$

where $\max\limits_{\text{all j}}$ denotes the largest value at any point on the grid. The second line of (8.15) follows because

$$\max_{\text{all j}} \left|A_j^n\right| = \max_{\text{all j}} \left|A_{j-1}^n\right| \ . \tag{8.16}$$

Eq. (8.15) demonstrates that the scheme is stable provided that our assumption (8.7) is satisfied. This means that the solution remains bounded for all time provided that (8.7) is satisfied, and so we can conclude that (8.7) is a sufficient condition for stability. For the upstream scheme, a sufficient condition for stability has turned out to be the same as the necessary condition for convergence. In other words, if the scheme is convergent it is stable, and vice versa.

In the solution of the *exact* advection equation, the maxima and minima of $A(x,\, t)$ never change. They are just carried along to different spatial locations. So, for the exact solution, the equality in (8.15) would hold.

Eq. (8.15), the sufficient condition for stability, is actually obvious from (8.11), because when $0 \leq \mu \leq 1$, A_j^{n+1} is obtained by linear interpolation *in space* to the point $x = j\Delta x - c\Delta t$, from the neighboring values of A. This is reasonable, because for advection the time rate of change at a point is closely related to the spatial variations upstream of that point.

8.5.2 The energy method

The direct method cannot be used to check the stability of complicated schemes. The energy method is more widely applicable, even for some nonlinear equations, and it is quite important in practice. With the energy method we ask: "Is $\sum_{\text{all j}} \left(A_j^n\right)^2$ bounded after an arbitrary number of time steps?" Here the summation is over the entire domain. If the sum is bounded, then each A_j^n must also be bounded. Whereas in the direct method we checked $\max_{\text{all j}} \left|A_j^{n+1}\right|$, with the energy method we check $\sum_{\text{all j}} \left(A_j^n\right)^2$. The two approaches are somewhat similar.

Returning to (8.11), squaring both sides, and summing over the domain, we obtain

$$\begin{aligned}\sum_{\text{all j}} \left(A_j^{n+1}\right)^2 =& \sum_{\text{all j}} \left[\left(A_j^n\right)^2 (1-\mu)^2 + 2\mu\,(1-\mu) A_j^n A_{j-1}^n + \mu^2 \left(A_{j-1}^n\right)^2\right] \\ =& (1-\mu)^2 \sum_{\text{all j}} \left(A_j^n\right)^2 + 2\mu\,(1-\mu) \sum_{\text{all j}} A_j^n A_{j-1}^n + \mu^2 \sum_{\text{all j}} \left(A_{j-1}^n\right)^2 \,.\end{aligned} \tag{8.17}$$

If A is periodic in x, then

$$\sum_{\text{all j}} \left(A_{j-1}^n\right)^2 = \sum_{\text{all j}} \left(A_j^n\right)^2 \,. \tag{8.18}$$

By starting from $\sum_{\text{all j}} \left(A_j^n - A_{j-1}^n\right)^2 \geq 0$, and using (8.18), we can show that

$$\sum_{\text{all j}} A_j^n A_{j-1}^n \leq \sum_{\text{all j}} \left(A_j^n\right)^2 . \tag{8.19}$$

Another way to derive (8.19) is to use (8.18) and Schwartz's inequality (e.g., Arfken (1985), p. 257), i.e.,

$$\left(\sum_{\text{all j}} a_j b_j\right)^2 \leq \left(\sum_{\text{all j}} a_j^2\right)\left(\sum_{\text{all j}} b_j^2\right) , \tag{8.20}$$

which holds for any sets of a's and b's. An interpretation of Schwartz's inequality is that the square of the dot product of two vectors is less than or equal to the product of the squares of the magnitudes of the two vectors. Use of (8.18) and (8.19) in (8.17) gives

$$\sum_{\text{all j}} \left(A_j^{n+1}\right)^2 \leq \sum_{\text{all j}} \left(A_j^n\right)^2 , \tag{8.21}$$

provided that $\mu\left(1-\mu\right) \geq 0$, which follows from (8.7). We conclude that

$$\sum_{\text{all j}} \left(A_j^{n+1}\right)^2 \leq \sum_{\text{all j}} \left(A_j^n\right)^2 , \text{ provided that } 0 \leq \mu \leq 1 . \tag{8.22}$$

This conclusion is the same as that obtained using the direct method, i.e., $0 \leq \mu \leq 1$ is a sufficient condition for stability.

8.5.3 von Neumann's method

John von Neumann was one of the leading mathematicians of the 20th century, and worked on many important problems, including game theory, nuclear physics, computer science, and numerical analysis (e.g., Bhattacharya, 2021). He developed *von Neumann's method*, which will be used extensively in this book. Solutions to linear partial differential equations can be expressed as superpositions of waves, by means of Fourier series. Von Neumann's method simply tests the stability of each Fourier component. If all of the Fourier components are stable, then the scheme is stable. The method can only be applied to linear or linearized equations with constant coefficients, however. Because of that, it can sometimes give misleading results.

To illustrate von Neumann's method, we return to the exact advection equation, (7.14). We assume for simplicity that the domain is infinite. First, we look for a solution with the wave form

$$A(x,\,t) = \mathrm{Re}\left[\widehat{A}(t)\,e^{ikx}\right]\,, \tag{8.23}$$

where $\left|\widehat{A}(t)\right|$ is the amplitude of the wave. Here k is called the wave number. It is independent of t and x because we have assumed that u is constant in space and time. For now, we consider a single wave number, for simplicity, but we can (and soon will) generalize (8.23) by replacing the right-hand side by a sum over a range of wave numbers.

Substituting (8.23) into (7.14), we find that

$$\frac{d\widehat{A}}{dt} + iku\widehat{A} = 0\,. \tag{8.24}$$

By this substitution, we have converted the partial differential equation (7.14) into an ordinary differential equation, (8.24), whose solution is

$$\widehat{A}(t) = \widehat{A}(0)\,e^{-ikut}\,, \tag{8.25}$$

where $\widehat{A}(0)$ is the initial value of $\widehat{A}$. Substituting (8.25) back into (8.23), we find that the full solution to (8.24) is

$$A(x,\,t) = \mathrm{Re}\left[\widehat{A}(0)\,e^{ik(x-ut)}\right]\,, \tag{8.26}$$

provided that $u =$ constant. As can be seen by inspection of (8.26), the sign convention used here implies that for $u > 0$ the signal will move towards larger x as time passes. Note that the exponent in (8.26) is a constant times $\xi \equiv x - ut$, which is the line (the characteristic) along which the solution of (7.14) is expected to be constant.

For a finite-difference equation, the assumed form of the solution, given by Eq. (8.23), is replaced by

$$A_j^n = \mathrm{Re}\left[\widehat{A}^n e^{ikj\Delta x}\right]\,. \tag{8.27}$$

Here $\left|\widehat{A}^n\right|$ is the amplitude of the wave at time-level n. Recall that the wavelength is 2π divided by the wave number. It follows that the shortest resolvable wave, with wavelength $L = 2\Delta x$, has $k\Delta x = \pi$, while longer waves have $k\Delta x < \pi$. This means that *there is no need to consider* $k\Delta x > \pi$.

We now introduce the amplification factor, λ, which was defined in Eq. (6.11). We can write

$$\widehat{A}^{n+1} \equiv \lambda \widehat{A}^n \ . \tag{8.28}$$

The amplification factor can be a complex number, but we have a special interest in its magnitude, which reveals the stability of a numerical scheme. Note that

$$\left|\widehat{A}^{n+1}\right| = |\lambda| \left|\widehat{A}^n\right| \ . \tag{8.29}$$

In general, λ depends on k, so we could use the notation λ_k or $\lambda(k)$, but for now we suppress that urge for the sake of readability. Of course, the value of λ also depends on the size of the time step. As shown below, we can work out the form of λ for a particular finite-difference scheme.

Before doing that, however, we use Eq. (8.28) to identify the effective value of λ for the exact solution to the differential equation. From (8.25), we find that

$$\text{for the exact advection equation } \widehat{A}(t+\Delta t) \equiv e^{iku\Delta t}\widehat{A}(t) \ , \tag{8.30}$$

from which it follows "by inspection" that

$$\text{for the exact advection equation } \lambda = e^{iku\Delta t} \ . \tag{8.31}$$

Eq. (8.31) implies that

$$\text{for the exact advection equation } |\lambda| = 1 \ , \tag{8.32}$$

regardless of the value of Δt. To go from (8.31) to (8.32), we have used Euler's formula.

As already discussed in Chapter 6, for processes other than advection the exact value of $|\lambda|$ can differ from 1, and can depend on Δt.

From (8.28) we see that after n time steps, starting from $n = 0$, the solution will be

$$\widehat{A}^n = \widehat{A}^0 \lambda^n \,. \tag{8.33}$$

Here again red denotes a superscript. From (8.33), the requirement for stability, i.e., that the solution remains bounded after arbitrarily many time steps, implies that

$$\boxed{|\lambda| \leq 1} \,. \tag{8.34}$$

Therefore, to evaluate the stability of a finite-difference scheme using von Neumann's method, we need to work out the value of $|\lambda|$ for that scheme, and check it to see whether or not (8.34) is satisfied.

Consider the particular case of the upstream scheme, as given by (8.1). Substituting (8.27) into (8.1) leads to

$$\frac{\widehat{A}^{n+1} - \widehat{A}^n}{\Delta t} + \left(\frac{1 - e^{-ik\Delta x}}{\Delta x} \right) u \widehat{A}^n = 0 \,. \tag{8.35}$$

Make sure that you understand where (8.35) comes from. Notice that the true advection speed, u, is multiplied, in (8.35), by the factor $\left(1 - e^{-ik\Delta x}\right)/\Delta x$. Comparing (8.35) with (8.24), we see that $\left(1 - e^{-ik\Delta x}\right)/\Delta x$ is "taking the place" of ik in the exact solution. In fact, you should be able to show that

$$\lim_{\Delta x \to 0} \left(\frac{1 - e^{-ik\Delta x}}{\Delta x} \right) = ik \,. \tag{8.36}$$

This is a clue that, for a given value of k, the upstream scheme advects A at with the wrong speed. We return to this point later.

For now, we use the definition of λ, i.e., (8.28), together with (6.8) and (8.35), to infer that

$$\lambda = 1 - \mu\,(1 - \cos k\Delta x + i \sin k\Delta x)\ . \tag{8.37}$$

Note that λ is complex. This is to be expected, because the effective value of λ for the exact solution of the differential equation is also complex. Computing the square of the modulus of both sides of (8.37), we obtain

$$|\lambda|^2 = 1 + 2\mu\,(\mu - 1)\,(1 - \cos k\Delta x)\ . \tag{8.38}$$

According to (8.38), the amplification factor $|\lambda|$ depends on the wave number, k, and also on μ. For $\mu = 1/2$, (8.38) reduces to

$$|\lambda|^2 = 1 - \frac{1}{2}\,(1 - \cos k\Delta x)\ . \tag{8.39}$$

Fig. 8.4 shows how $|\lambda|^2$ varies with $k\Delta x$ and μ. For $k\Delta x = 0$, we get $|\lambda|^2 = 1$, regardless of the value of μ. For $k\Delta x > 0$, the scheme damps for $0 \le \mu \le 1$ and is unstable for both $\mu < 0$ and $\mu > 1$. For μ close to zero, the scheme is close to neutral, but many time steps are needed to complete a given simulation. For μ close to one, the scheme is again close to neutral, but it is also close to instability. If we choose intermediate values of μ, the shortest modes are strongly damped, and such strong smoothing is usually unacceptable.

Although λ depends on k, it does not depend on x (i.e., on j) or on t (i.e., on n). Why not? The reason is that out "coefficient," namely the wind speed c, has been assumed to be independent of x and t.

The fact that von Neumann's method can only be used to analyze the stability of a linearized version of the equation, with constant coefficients, is an important limitation of the method, because the equations used in numerical models are typically nonlinear and/or have spatially variable coefficients – if this were not true, we would solve them analytically! The key point is that *von Neumann's method can sometimes tell us that a scheme is stable, when in reality it is unstable*. In such cases, the instability arises from nonlinearity and/or through the effects of spatially variable coefficients. This type of instability can be detected using the energy method, and will be discussed in a later chapter.

If von Neumann's method says that a scheme is *unstable*, you can be confident that it is really is unstable.

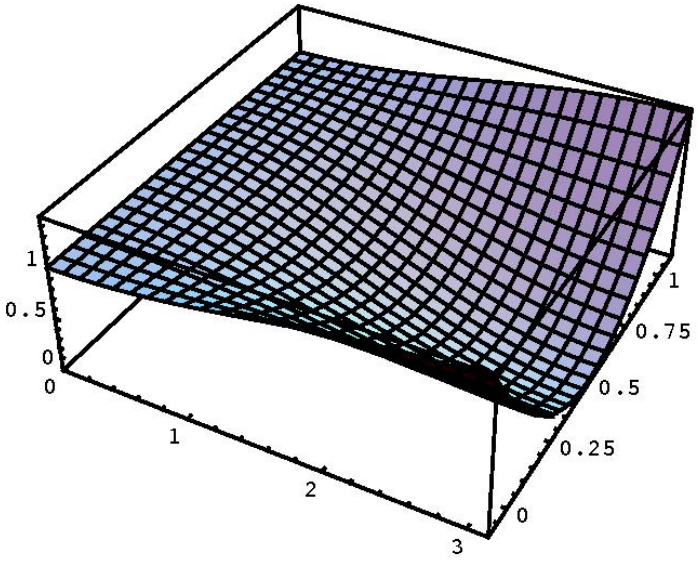


Figure 8.4: The square of the amplification factor for the upstream scheme is shown on the vertical axis. The front axis is $k\Delta x$, and the right-side axis is μ. The plot includes values of μ that are slightly less than zero, and slightly greater than one.

8.6 Including multiple wave numbers

In general, the full solution for A_j^n can be expressed as a Fourier series. For simplicity, suppose that the domain is periodic in x, with period L_0. You might want to think of L_0 as the distance around a latitude circle. Then A_j^n can be written as

$$\begin{aligned} A_j^n &= \sum_{m=-\infty}^{\infty} \widehat{A}_m^n e^{imk_0 j\Delta x} \\ &= \sum_{m=-\infty}^{\infty} \widehat{A}_m^0 e^{imk_0 j\Delta x} \lambda_m^n \,, \end{aligned} \tag{8.40}$$

where

$$k \equiv mk_0 \,, \tag{8.41}$$

$$k_0 \equiv 2\pi/L_0 \,, \tag{8.42}$$

and m is a nondimensional integer (i.e., a "pure number"), which is analogous to what is called the "zonal wave number" in large-scale dynamics. We can interpret k_0 as the *lowest* non-zero wave number in the solution, so if L_0 is the distance across the model's spatial domain then the lowest wave number k_0 corresponds to one "high" and one "low" within the domain. In (8.40), the summation has been formally taken over all integers, although of course only a finite number of ms can be used in a real application. On the second line of (8.40), $|\lambda_m|$ can be interpreted as the amplification factor for mode m, which shows explicitly that different wave numbers have different amplification factors. If *any* wave number is unstable, then the scheme is unstable.

We can write

$$\begin{aligned} \left|A_j^n\right| &\leq \left| \sum_{m=-\infty}^{\infty} \widehat{A}_m^0 e^{imk_0 j\Delta x} \lambda_m^n \right| \\ &\leq \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 e^{imk_0 j\Delta x} \lambda_m^n \right| \\ &= \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 \lambda_m^n \right| . \end{aligned} \tag{8.43}$$

If $|\lambda_m^n| \leq 1$ is satisfied for all m, then

$$\left|A_j^n\right| \leq \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 \right| . \tag{8.44}$$

Therefore, $\left|A_j^n\right|$ will be bounded provided that $\sum_{m=-\infty}^{\infty} \widehat{A}_m^0 e^{imk_0 j\Delta x}$, which gives the initial condition, is an absolutely convergent Fourier series. *The point is that* $|\lambda| \leq 1$ *for all* m *is sufficient for stability*. It is also necessary, because if $|\lambda| > 1$ for a particular m, say $m = m_1$, then the solution for the initial condition $A_{m_1} = 1$ and $A_m = 0$ for all $m \neq m_1$ will be unbounded.

From (8.11), λ_m for the upstream scheme is given by

$$\lambda_m = 1 - \mu \left(1 - \cos mk_0\Delta x + i \sin mk_0\Delta x\right) . \tag{8.45}$$

This leads to

$$|\lambda_m| = \sqrt{1 + 2\mu\,(\mu - 1)\,(1 - \cos mk_0\Delta x)}\,. \tag{8.46}$$

From (8.46) we can show that $|\lambda_m| \leq 1$ holds for all m, if and only if $\mu\,(\mu - 1) \leq 0$, which is equivalent to (8.7). This is the necessary and sufficient condition for the stability of the scheme.

8.7 Periodic boundary conditions

To explicitly allow for a finite periodic domain, the upstream scheme can be written in matrix form as

$$\begin{bmatrix} A_1^{n+1} \\ A_2^{n+1} \\ \cdots \\ A_{j-1}^{n+1} \\ A_j^{n+1} \\ A_{j+1}^{n+1} \\ \cdots \\ A_{J-1}^{n+1} \\ A_J^{n+1} \end{bmatrix} = \begin{bmatrix} 1-\mu & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \mu \\ \mu & 1-\mu & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1-\mu & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \mu & 1-\mu & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mu & 1-\mu & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \mu & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1-\mu & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \mu & 1-\mu \end{bmatrix} \begin{bmatrix} A_1^n \\ A_2^n \\ \cdots \\ A_{j-1}^{n1} \\ A_j^n \\ A_{j+1}^n \\ \cdots \\ A_{J-1}^n \\ A_J^n \end{bmatrix}, \tag{8.47}$$

or

$$\left[A_j^{n+1}\right] = [M]\left[A_j^n\right]\,, \tag{8.48}$$

where $[M]$ is the matrix written out on the right-hand side of (8.47). In writing (8.47), the cyclic boundary condition

$$A_1^{n+1} = (1-\mu)A_1^n + \mu A_J^n \tag{8.49}$$

has been assumed, and that is why μ appears in the top-right corner of the matrix. I made it red just so that you wouldn't miss it. From the definition of λ, (8.28), we can write

$$\begin{bmatrix} A_1^{n+1} \\ A_2^{n+1} \\ \cdots \\ A_{j-1}^{n+1} \\ A_j^{n+1} \\ A_{j+1}^{n+1} \\ \cdots \\ A_{J-1}^{n+1} \\ A_J^{n+1} \end{bmatrix} = \begin{bmatrix} \lambda & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \begin{bmatrix} A_1^n \\ A_2^n \\ \cdots \\ A_{j-1}^{n1} \\ A_j^n \\ A_{j+1}^n \\ \cdots \\ A_{J-1}^n \\ A_J^n \end{bmatrix}, \tag{8.50}$$

or

$$\left[A_j^{n+1}\right] = \lambda [I] \left[A_j^n\right] , \tag{8.51}$$

where $[I]$ is the identity matrix. Comparing (8.48) and (8.51), we see that

$$([M] - \lambda [I]) \left[A_j^n\right] = 0 . \tag{8.52}$$

This equation must hold regardless of the values of the A_j^n. It follows that the amplification factors, λ, are the *eigenvalues* of $[M]$, obtained by solving

$$|[M] - \lambda [I]| = 0 , \tag{8.53}$$

where the absolute value signs denote the determinant. For the current example, we can use (8.53) to show that

$$\lambda = 1 - \mu \left(1 - e^{i2m\pi/J}\right), \quad m = 0,\, 1,\, 2,\, \ldots,\, J-1\,. \tag{8.54}$$

This has essentially the same form as (8.37), and so it turns out that, again, the stability criterion is $0 \leq \mu \leq 1$.

8.8 Does the solution improve if we refine the grid?

Consider what happens when we increase the number of grid points, *while fixing the domain size, D, the wind speed, u, and the wave number k of the advected signal*. We would like to think that the solution is improved by increasing the resolution, but this must be checked because higher spatial resolution also means that we have to take a shorter time step (for stability), and a shorter time step means that more time steps are needed to simulate a given interval of time. Each time step leads to some damping, and the damping is an error. The increased spatial resolution is a good thing, but it sounds like the increased number of time steps could be a bad thing. Does the solution improve, or not?

Consider grid spacing Δx, such that

$$D = J\Delta x\,. \tag{8.55}$$

As we decrease Δx, we increase J correspondingly, so that D does not change, and

$$k\Delta x = kD/J\,. \tag{8.56}$$

Substituting this into (8.38), we find that the amplification factor satisfies

$$|\lambda|^2 = 1 + 2\mu\,(\mu - 1)\left[1 - \cos\left(\frac{kD}{J}\right)\right]\,. \tag{8.57}$$

In order to maintain computational stability, *we keep μ fixed as Δx decreases*, so that

$$\begin{aligned} \Delta t &= \frac{\mu \Delta x}{u} \\ &= \frac{\mu D}{uJ} \, . \end{aligned} \tag{8.58}$$

The time required for the air to flow through the domain is

$$T = \frac{D}{u} \tag{8.59}$$

Let N be the number of time steps needed for the air to flow through the domain, so that

$$\begin{aligned} N &= \frac{T}{\Delta t} \\ &= \frac{D}{u \; varDeltat} \\ &= \frac{D}{\mu \Delta x} \\ &= \frac{J}{\mu} \end{aligned} \tag{8.60}$$

To obtain the last line of (8.60), we have substituted from (8.55). The total amount of damping that "accumulates" as the air moves across the domain is given by

$$|\lambda|^N = \left(|\lambda|^2\right)^{N/2} = \left\{1 - 2\mu\,(1-\mu)\,[1-\cos(kD/J)]\right\}^{\frac{J}{2\mu}} \, . \tag{8.61}$$

Here we have used (8.57) and (8.60).

As we increase the resolution with a fixed domain size, J increases. In Fig. 8.5, we show the dependence of $|\lambda|^N$ on J, for two different fixed values of μ. In making the figure, the wavelength is assumed to be half the domain width, so that $kD = 4 \; / pi$. This causes the cosine factor in (8.61) to approach 1, which weakens the damping associated with $|\lambda| < 1$; but on the other hand it also causes the exponent in (8.61) to increase, which strengthens the damping. Which effect dominates? The answer can be seen in Fig. 8.5.

Increasing J leads to less total damping for a given value of μ, even though the number of time steps needed to cross the domain increases. This is good news.

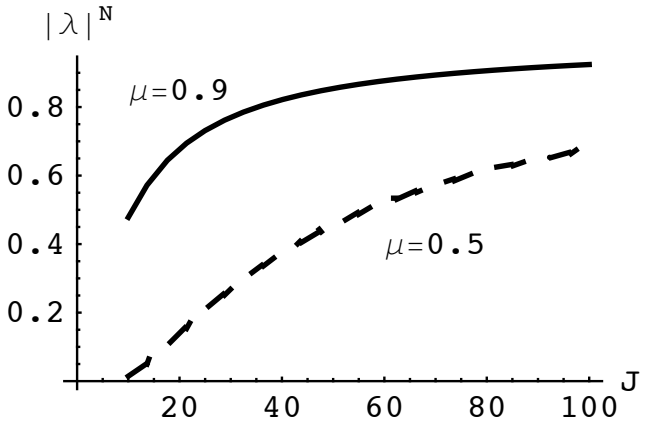


Figure 8.5: "Total" damping experienced by a disturbance crossing the domain, as a function of J, the number of grid points across the domain, for two different fixed values of μ. The point of the figure is that increasing J leads to a decrease in damping. In these examples we have assumed $D/L = 2$, i.e., the wavelength is half the width of the domain.

On the other hand, if we fix J and decrease μ (by decreasing the time step), the damping increases, so the solution becomes less accurate. This means that, *for the upstream scheme, the amplitude error can be minimized by using the largest stable value of* μ. We minimize the error by living dangerously.

8.9 Summary

This chapter gives a quick introduction to the solution of a finite-difference equation, using the upstream scheme for the advection equation as an example. We have encountered the concepts of convergence and stability, and three different ways to test the stability of a scheme.

Suppose that we are given a non-linear partial differential equation and wish to solve it by means of a finite-difference approximation. The usual procedure would be as follows:

- **Check the truncation errors.** This is done by using a Taylor series expansion to find the leading terms of the errors in approximations for the various derivatives that appear in the governing equations of the model.
- **Check linear stability** for a simplified (linearized, constant coefficients) version of the equation. The most commonly used method is that of von Neumann.
- **Check nonlinear stability,** if possible. This can be accomplished, in some cases, by using the energy method. Otherwise, empirical tests are needed. More discussion is given later.

Increased accuracy as measured by discretization error does not always imply a better scheme. For example, consider two schemes A and B, such that scheme A is first-order accurate but stable, while scheme B is second-order accurate but unstable. Given such a choice, the "less accurate" scheme is definitely better.

Almost always, the design of a finite-difference scheme is an exercise in trade-offs. For example, a more accurate scheme is usually more complicated and expensive than a less accurate scheme. We have to ask whether the additional complexity and computational expense are justified by the increased accuracy. The answer depends on the particular application.

In general, "good" schemes have the following properties, among others:

- High accuracy.
- Stability.
- Simplicity.
- Computational economy.

Later, we will extend this list.

8.10 Problems

1. Program the upstream scheme on a periodic domain with 100 grid points. Give a sinusoidal initial condition with a single mode such that exactly four wavelengths fit in the domain. Integrate for μ = -0.1, 0.1, 0.5, 0.9, 1 and 1.1. In each case, take enough time steps so that in the exact solution the signal will just cross the domain. Plot and discuss your results.

2. Work out the expressions for the phase and amplitude errors of the upstream scheme, and plot them as functions of μ and $k\Delta x$.

3. Analyze the stability of

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = 0 \tag{8.62}$$

   using von Neumann's method.

4. Consider the advection equation with centered second-order space differencing on a uniform grid. We use the trapezoidal implicit time-differencing scheme.

   (a) Using von Neumann's method, prove that scheme is unconditionally stable.

(b) Repeat, using the energy method.

(c) Prove that, with second-order centered-in-space differencing, the trapezoidal implicit scheme is *time-reversible* for advection. Also prove that the forward scheme is not time-reversible.

5. Check the stability of the Matsuno scheme with centered space differencing, for the advection equation.

Chapter 9

"Forward-in-time" advection schemes

Further examples of schemes for the advection equation can be obtained by combining centered space-differencing with two-level time-differencing schemes (see Chapter 4).

9.1 A family of advection schemes

Following Takacs (1985), we define a fairly general family of two-time-level explicit schemes of the form

$$A_j^{n+1} = \sum_{j'=-\infty}^{\infty} a_{j'} A_{j+j'}^n \,. \tag{9.1}$$

Here j' denotes various points on the grid that are used to evaluate the time-rate-of-change of A at the point j. We have assumed for simplicity that the grid spacing is uniform. Schemes that use only two time levels are often called "forward-in-time" schemes; don't confuse them with the Euler-forward time-differencing scheme.

Recall that the discretization error, which is a measure of the accuracy of the finite-difference scheme, can be evaluated by substituting the exact solution of the differential equation into the finite-difference equation. Replacing the various A's in (9.1) by the corresponding values of the true solution, represented in terms of Taylor series expansions around the point $(j\Delta x, n\Delta t)$, and replacing the right-hand side of (9.1) by the discretization error of the scheme, denoted by ε, we find that

$$\left(A + \Delta t \frac{\partial A}{\partial t} + \frac{\Delta t^2}{2!}\frac{\partial^2 A}{\partial t^2} + \ldots\right) - \sum_{j'=-\infty}^{\infty} a_{j'} \left[A + \left(j'\Delta x\right)\frac{\partial A}{\partial x} + \frac{\left(j'\Delta x\right)^2}{2!}\frac{\partial^2 A}{\partial x^2} + \ldots\right] = \varepsilon, \tag{9.2}$$

where A and all of its derivatives are evaluated at $(x,\, t) = \left(x_j,\, t^n\right)$. We now use the continuous advection equation in the forms

$$\frac{\partial A}{\partial t} = -c\frac{\partial A}{\partial x}, \quad \frac{\partial^2 A}{\partial t^2} = c^2\frac{\partial^2 A}{\partial x^2}, \quad \text{etc.}\,. \tag{9.3}$$

These relations only hold if u is constant, so the results derived below are only approximately valid when c is variable. For derivatives of order m, (9.3) generalizes to

$$\left(\frac{\partial^m}{\partial t^m}\right) A = (-u)^m \left(\frac{\partial^m}{\partial x^m}\right) A\,. \tag{9.4}$$

This allows us to write

$$\Delta t^m \left(\frac{\partial^m}{\partial t^m}\right) A = (-\mu)^m \Delta x^m \left(\frac{\partial^m}{\partial x^m}\right) A\,, \tag{9.5}$$

or

$$\Delta x^m \left(\frac{\partial^m}{\partial x^m}\right) A = \left(\frac{\Delta t}{-\mu}\right)^m \left(\frac{\partial^m}{\partial t^m}\right) A\,, \tag{9.6}$$

where μ is the usual CFL parameter. With the use of (9.6), we can rewrite (9.2) as

$$\left(A + \Delta t\frac{\partial A}{\partial t} + \frac{\Delta t^2}{2!}\frac{\partial^2 A}{\partial t^2} + \ldots\right) - \sum_{j'=-\infty}^{\infty} a_{j'}\left[A - \left(\frac{j'\Delta t}{\mu}\right)\frac{\partial A}{\partial t} + \frac{1}{2!}\left(\frac{j'\Delta t}{\mu}\right)^2\frac{\partial^2 A}{\partial t^2} + \ldots\right] = \varepsilon\,, \tag{9.7}$$

Inspection of (9.7) shows that in order to ensure first-order accuracy in both time and space, we need

$$\sum_{j'=-\infty}^{\infty} a_{j'} = 1 \tag{9.8}$$

and

$$\sum_{j'=-\infty}^{\infty} j' a_{j'} = -\mu \, . \tag{9.9}$$

If these two conditions are used to define coefficients for points j and $j-1$, the result is the upstream scheme (assuming that $\mu \geq 0$).

To have second-order accuracy in both time and space, we need

$$\sum_{j'=-\infty}^{\infty} j'^2 a_{j'} = \mu^2 \, . \tag{9.10}$$

In general, to have mth-order accuracy in both time and space, we must require (9.8), (9.9), and

$$\sum_{j'=-\infty}^{\infty} \left(j'\right)^l a_{j'} = (-\mu)^l \quad \text{for} \quad l = 0 \quad \text{to} \quad m \, . \tag{9.11}$$

Using (9.8), we can rewrite (9.9) as

$$\sum_{j'=-\infty}^{\infty} \left(j' + \mu\right) a_{j'} = 0 \, . \tag{9.12}$$

Similarly, we can use (9.8) to rewrite (9.10) as

Using the binomial theorem, it can be shown from (9.8) through (9.11) that for a scheme of mth-order accuracy

$$\sum_{j'=-\infty}^{\infty} \left(j' + \mu\right)^l a_{j'} = 0 \text{ for } 1 \leq l \leq m \, . \tag{9.13}$$

This will be used in Chapter 12.

Next, we work out the amplification factor for the family of schemes given by (9.1). As usual, we look for a solution of the form

$$A_j^n = \mathrm{Re}\left[\hat{A}^n e^{ikj\Delta x}\right] . \tag{9.14}$$

It follows from (9.1) and (9.14) that the amplification factor is given by

$$\lambda = \sum_{j'=-\infty}^{\infty} a_{j'} e^{ij'k\Delta x} . \tag{9.15}$$

In the following section, we discuss several schemes to which the preceding analysis is directly applicable.

9.2 Explicit schemes for advection

9.2.1 Matsuno time-differencing with centered space differencing

In the case of the Matsuno scheme, the first approximation to A_j^{n+1} comes from

$$\frac{A_j^{n+1*} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = 0, \tag{9.16}$$

and the final value from

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^{n+1*} - A_{j-1}^{n+1*}}{2\Delta x}\right) = 0 . \tag{9.17}$$

Eliminate the terms with $(\,)*$ from (9.17) by using (9.16) twice (first with j replaced by $j+1$, then with j replaced by $j-1$). The result can be written as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = \frac{u^2\Delta t}{(2\Delta x)^2}\left(A_{j+2}^n - 2A_j^n + A_{j-2}^n\right) . \tag{9.18}$$

It should be clear that (9.18) is a member of the family given by (9.1). The term on the right-hand side of (9.18) approaches zero as $\Delta t \to 0$, and thus (9.18) is consistent with the one-dimensional advection equation, but has only first-order accuracy. If we let $\Delta x \to 0$ (and $\Delta t \to 0$ to keep stability), this term approaches $c^2 \Delta t \partial^2 A/\partial x^2$. In effect, *it acts as a diffusion term that damps spatial variations*. The "diffusion coefficient" is $u^2 \Delta t$, which goes to zero as $\Delta t \to 0$. We say that the centered-in-space advection scheme with Matsuno time differencing is "diffusive." As you will show in one of the homework problems, it is only first-order accurate.

9.2.2 The Lax-Wendroff scheme

A similarly diffusive scheme, called the Lax-Wendroff scheme[1], has second-order accuracy. Consider an explicit two-level scheme of the form:

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + \frac{u}{\Delta x}\left(a_{j-1}A_{j-1}^n + a_j A_j^n + a_{j+1}A_{j+1}^n\right) = 0\,. \tag{9.19}$$

The scheme given by Eq. (9.23) was proposed by Lax and Wendroff (1960), and recommended by Richtmyer (1963). It is a member of the family given by (9.1). For the centered-in-space approximation to $\partial A/\partial x$, we would have $a_{j-1} = -1/2$, $a_j = 0$, and $a_{j+1} = 1/2$, but the Lax-Wendroff scheme does not use those values. To ensure at least first-order accuracy in both time and space, we must require that

$$a_{j-1} + a_j + a_{j+1} = 0, \text{ and } -a_{j-1} + a_{j+1} = 1\,. \tag{9.20}$$

To obtain second-order accuracy in both time and space, we must also enforce

$$\mu + a_{j-1} + a_{j+1} = 0\,. \tag{9.21}$$

Solving (9.20) and (9.21) for the parameters of the scheme, we find that

$$a_{j-1} = \frac{-1-\mu}{2},\ a_j = \mu,\ \text{ and } a_{j+1} = \frac{1-\mu}{2}\,. \tag{9.22}$$

[1] Peter Lax died in spring 2025.

For example, if $\mu = 1/2$, then $a_{j-1} = -3/4$, $a_j = 1/2$, and $a_{j+1} = 1/4$. On the other hand, if $\mu = -1/2$, then $a_{j-1} = -1/4$, $a_j = -1/2$, and $a_{j+1} = 3/4$. No matter which way the wind is blowing, the absolute value of the upstream coefficient is larger than the absolute value of the downstream coefficient. In other words, the scheme is automatically "upstream-weighted" regardless of which way the wind is blowing, even though the stencil is centered. That's very attractive.

Substituting from (9.22) into (9.19), we can write the scheme as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = \frac{u^2 \Delta t}{2\Delta x^2}\left(A_{j+1}^n - 2A_j^n + A_{j-1}^n\right) . \tag{9.23}$$

Compare (9.23) with (9.18), which is the corresponding result for the Matsuno scheme. The left-hand side of (9.23) looks like "forward in time, centered in space," which would be unstable. But the right-hand side looks like diffusion, and can stabilize the scheme if the time step is small enough. Note that (9.23) is second-order accurate in time, even though it involves only two time levels; this conclusion depends on our assumption that c is a constant. The scheme achieves second-order accuracy in space through the use of three grid points. This illustrates that *a non-iterative two-time-level scheme is not necessarily a first-order scheme*.

The Lax-Wendroff scheme is equivalent to and can be interpreted in terms of the following procedure: First calculate $A_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ and $A_{j1\frac{1}{2}}^{n+\frac{1}{2}}$ from

$$\frac{A_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}\left(A_{j+1}^n + A_j^n\right)}{\frac{1}{2}\Delta t} = -u\left(\frac{A_{j+1}^n + A_j^n}{\Delta x}\right) , \tag{9.24}$$

$$\frac{A_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}\left(A_j^n + A_{j-1}^n\right)}{\frac{1}{2}\Delta t} = -u\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) , \tag{9.25}$$

and then use these to obtain A_j^{n+1} from

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} = -u\left(\frac{A_{j+\frac{1}{2}}^{n+\frac{1}{2}} - A_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x}\right) . \tag{9.26}$$

Note that (9.26) is "centered in time." If (9.24) and (9.25) are substituted into (9.26), we recover (9.23). This helps to rationalize why it is possible to obtain second-order accuracy in time with this two-time-level scheme.

For the Lax-Wendroff scheme, the amplification factor is

$$\lambda = 1 - 2\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) - i\mu \sin\left(k\Delta x\right) \ , \tag{9.27}$$

To obtain (9.27), we have used the trigonometric identity $2\sin^2(\theta/2) = 1 - \cos\theta$. We find that

$$\begin{aligned} |\lambda|^2 &= \left[1 - 4\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) + 4\mu^4 \sin^4\left(\frac{k\Delta x}{2}\right)\right] + \mu^2 \sin^2\left(k\Delta x\right) \\ &= 1 - 4\mu^2\left(1-\mu^2\right)\sin^4\left(\frac{k\Delta x}{2}\right) \ . \end{aligned} \tag{9.28}$$

To obtain the second line of (9.28), we have used the trigonometric identity $\sin(2\theta) = 2\sin\theta\cos\theta$. The scheme is stable for $\mu^2 < 1$. Since (9.28) involves only μ^2, the stability criterion does not depend on the direction of the wind, which is very nice. If $\mu^2 < 1$, then $|\lambda| < 1$ and the scheme is dissipative. Fig. 9.1 shows how $|\lambda|^2$ depends on μ and L, for both the upstream and Lax-Wendroff schemes. The damping of the Lax-Wendroff scheme is more scale selective than that of the upstream scheme.

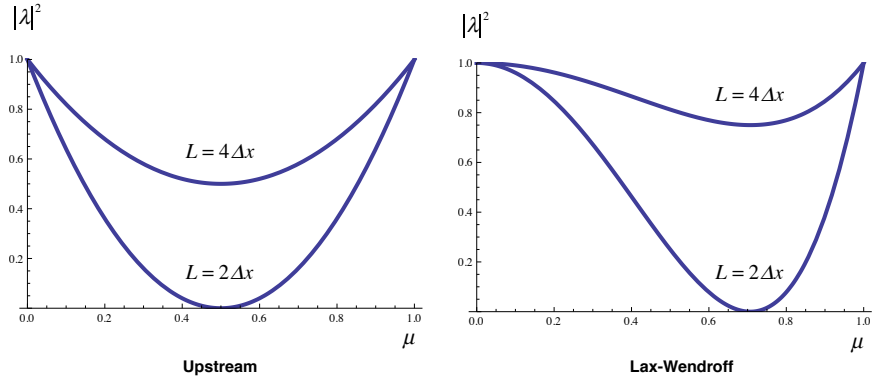


Figure 9.1: The amplification factors for the Lax-Wendroff and upstream schemes, for two different wavelengths, plotted as a function of μ^2.

9.2.3 The Takacs scheme

Takacs (1985) proposed a forward-in-time space-uncentered advection scheme that, for positive μ, uses the points $j+1$, j, $j-1$ and $j-2$, Note that, for positive μ, points $j-1$

and $j-2$ are both in the upstream direction. The four grid points bring in four coefficients, which can be chosen to achieve third-order accuracy:

$$a_{j+1} = \left(-\mu^3 + 3\mu^2 - 2\mu\right)/6\,, \tag{9.29}$$

$$a_j = 1 - \left(-\mu^3 + 6\mu^2 - 3\mu\right)/6\,, \tag{9.30}$$

$$a_{j-1} = \left(-\mu^3 + \mu^2 + 2\mu\right)/2\,, \tag{9.31}$$

$$a_{j-2} = \mu\left(\mu^2 - 1\right)/6\,. \tag{9.32}$$

9.3 Implicit schemes for advection

There are also various implicit schemes, such as the trapezoidal implicit scheme, which are neutral and unconditionally stable, so that in principle any Δt can be used if the phase error can be tolerated. Such schemes are *not* members of the family defined by (9.1). Implicit schemes have the drawback that a linear system of equations must be solved on each time step, often using an iterative method. In many cases, solving the linear system may take as much computer time as an explicit scheme with a smaller Δt.

9.4 Segue

When we solve the advection equation, space-differencing schemes can introduce diffusion-like damping. If this damping is sufficiently scale-selective, it can be beneficial.

9.5 Problems

1. Consider a periodic domain with cells numbered by $0 \le j \le 100$, with "boxcar" initial conditions:

$$\begin{aligned} q_j &= 100 \text{ for } 45 \leq j \leq 55, \\ q_j &= 0 \text{ for all other } j\,. \end{aligned} \tag{9.33}$$

Write programs to solve

$$\frac{\partial q}{\partial t} + u\frac{\partial q}{\partial x} = 0\,, \tag{9.34}$$

using the following schemes:

(a) Upstream.

(b) Lax Wendroff.

(c) Trapezoidal in time, and second-order centered in space. In order to do this, you will have to solve a linear algebra problem.

Choose $\mu = 0.7$ in each case. Run the model long enough so that *for the exact solution* the signal crosses the domain exactly once. Plot the results for the end of the run, and also the half-way point. Compare the numerical solutions with each other and with the exact solution. For each scheme, discuss the amplitude errors, phase errors, and the extent to which the scheme is sign preserving.

Chapter 10

Advection in multiple dimensions

In Chapter 3, we briefly discussed methods for approximating differential operators that involve two or more dimensions, using the Laplacian as an example. In this chapter, we discuss two-dimensional advection. The Laplacian is an "isotropic" operator; it has no built-in direction. Advection, on the other hand, carries the air in in the direction that the wind is blowing towards; the advection operator is highly directional.

The discussion in this chapter will be limited to horizontal planes. The special issues that arise with the vertical dimension are discussed in Chapters 22, 23, and 24. The special issues that arise in spherical geometry are discussed in Chapters 28 and 29.

Variable currents are usually multi-dimensional. Before we discuss variable currents, in a later chapter, it is useful to consider constant currents in two-dimensions.

Let A be an arbitrary quantity advected, in two dimensions, by a constant basic current. The advection equation is

$$\frac{\partial A}{\partial t} + u\frac{\partial A}{\partial x} + v\frac{\partial A}{\partial y} = 0, \tag{10.1}$$

where u and v are the x and y components of the current, respectively. We assume here that u and v are spatially constant, but of course that won't be the case in a real model.

Let i and j be the indices of grid points in the x and y directions, on a rectangular grid. Replacing $\partial A/\partial x$ and $\partial A/\partial y$ by the corresponding centered difference quotients, we obtain

$$\frac{dA_{i,j}}{dt} + u\left(\frac{A_{i+1,j} - A_{i-1,j}}{2\Delta x}\right) + v\left(\frac{A_{i,j+1} - A_{i,j-1}}{2\Delta y}\right) = 0\,. \tag{10.2}$$

Assume that A has the form

$$A_{i,j} = \mathrm{Re}\left\{\widehat{A}(t)\, e^{\underset{\sim}{i}[(ki\Delta x)+(lj\Delta y)]}\right\} , \tag{10.3}$$

where $\underset{\sim}{i} \equiv \sqrt{-1}$, and k and l are wave numbers in the x and y directions, respectively. Substitution gives the oscillation equation again:

$$\frac{d\widehat{A}}{dt} = i\omega\widehat{A} , \tag{10.4}$$

where this this time the frequency is given by

$$\omega \equiv -\left[u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right] . \tag{10.5}$$

If we were to use leapfrog time-differencing, the stability criterion would be

$$\left|u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right|\Delta t \le 1 . \tag{10.6}$$

Since

$$\begin{aligned}\left|u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right|\Delta t &\le \left[\left|u\frac{\sin(k\Delta x)}{\Delta x}\right| + \left|v\frac{\sin(l\Delta y)}{\Delta y}\right|\right]\Delta t \\ &\le \left(\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y}\right)\Delta t ,\end{aligned} \tag{10.7}$$

a *sufficient* condition to satisfy (10.6) is

$$\left(\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y}\right)\Delta t \le 1 . \tag{10.8}$$

If we require the scheme to be stable for all possible k and l, and for all combinations of u and v, then (10.8) is also a necessary condition.

How does the stability criterion depend on the direction of the flow and the shapes of the grid cells? To answer this, define

$$|u| \equiv S\cos\alpha \text{ and } |v| \equiv S\sin\alpha \,, \tag{10.9}$$

where the wind speed (the magnitude of the wind vector) is represented by $S \equiv \sqrt{(u^2+v^2)} \geq 0$, and $0 \leq \alpha \leq \pi/2$. Here we work with the absolute values of u and v because reversing the direction of the wind has no effect on the numerical stability. For $\alpha = 0$, the flow is purely zonal, and for $\alpha = \pi/2$ it is purely meridional. Then (10.8) becomes

$$S\left(\frac{\cos\alpha}{\Delta x} + \frac{\sin\alpha}{\Delta y}\right)\Delta t \leq 1 \,. \tag{10.10}$$

In order for the scheme to be stable *for any orientation of the current*, we must have

$$S\left(\frac{\cos\alpha_{\max}}{\Delta x} + \frac{\sin\alpha_{\max}}{\Delta y}\right)\Delta t \leq 1 \,, \tag{10.11}$$

where $\alpha_{\max}$ is the "worst-case" α, which makes the left-hand side of (10.10) as large as possible for a given value of S. We can show that $\alpha_{\max}$ satisfies

$$\begin{gathered} \tan\alpha_{\max} = \frac{\Delta x}{\Delta y}, \text{ so that} \\ \sin\alpha_{\max} = \frac{\Delta x}{\sqrt{(\Delta x)^2+(\Delta y)^2}} \text{ and } \cos\alpha_{\max} = \frac{\Delta y}{\sqrt{(\Delta x)^2+(\Delta y)^2}} \,. \end{gathered} \tag{10.12}$$

As shown in Fig. 10.1, $\alpha_{\max}$ measures the angle of the "diagonal" across a grid cell, so it is a measure of the shape of the box. For example, when $\Delta y/\Delta x \ll 1$ (a box that is much wider than it is tall), we get $\alpha_{\max} \to \pi/2$, which means that the most dangerous flow direction is meridional, because that the direction in which the grid cell is "narrowest." As a second example, for $\Delta x = \Delta y$ (a square box), we get $\alpha_{\max} = \pi/4$.

From (10.11) and (10.12) we see that the stability criterion can be written as

$$\frac{S\Delta t}{\sqrt{(\Delta x)^2+(\Delta y)^2}}\left(\frac{\Delta y}{\Delta x} + \frac{\Delta x}{\Delta y}\right) \leq 1 \,. \tag{10.13}$$

In particular, for $\Delta x = \Delta y = d$, Eq. (10.13) reduces to

$$\frac{S\Delta t}{d} \leq \frac{1}{\sqrt{2}} < 1 \, . \tag{10.14}$$

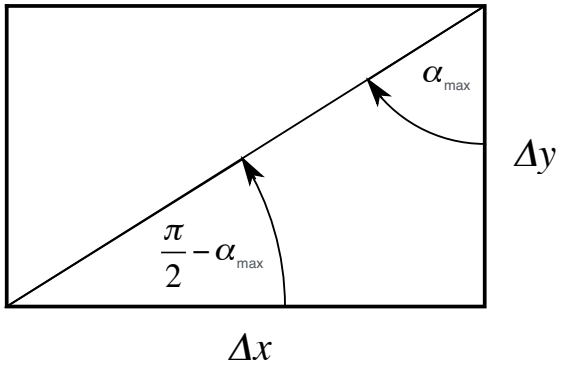


Figure 10.1: Sketch illustrating the angle $\alpha_{\max}$ on a rectangular grid.

Chapter 11

Finite-volume methods

11.1 The basic idea

The finite-difference method defines points within a volume of fluid, and predicts the properties of the fluid at those points. In contrast, the finite-volume method divides a volume into contiguous, non-overlapping finite "chunks" of fluid, contained within finite "grid cells," and predicts the average properties of the air in each cell.

Finite-volume methods were originally developed to create conservative advection schemes, but they have been generalized to create schemes for the pressure-gradient force, diffusion, etc.. They can readily be adapted to arbitrarily shaped grids.

11.2 Godunov schemes

Godunov's scheme for advection, introduced by Sergei Godunov (1959), is the first example of a finite-volume scheme. The key features of Godunov's approach are:

- Predict grid-cell averages rather than point values.
- Compute fluxes at cell walls for use in predicting the grid-cell averages. This is called "Riehmann's problem."
- Assume that the solution as piecewise constant within each grid cell.

The first items in the list above are essential to the finite-volume method for advection. The third item (an assumption) is not necessary, and can be drastically improved upon. As explained below, Godunov's approach conserves mass, and avoids non-physical oscillations.

"Godunov's order barrier theorem" shows that any linear monotone scheme that avoids creating new maxima or minima cannot be more than first-order accurate. As discussed in Chapter 13, modern schemes circumvent Godunov's theorem by using nonlinear "limiters."

11.3 Continuous advection in one dimension

In preparation for illustrating the finite-volume method for advection in a simplified framework, let A be a "conservative" variable, satisfying the following one-dimensional conservation law:

$$\frac{\partial}{\partial t}(\rho A) + \frac{\partial}{\partial x}(\rho u A) = 0 \,. \tag{11.1}$$

Here ρ is the density of the air, and ρu is a mass flux. Putting $A \equiv 1$ in (11.1) gives mass conservation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0 \,. \tag{11.2}$$

By combining (11.1) and (11.2), we can derive the "advective form" of the conservation equation for A as

$$\rho \left(\frac{\partial A}{\partial t} + u \frac{\partial A}{\partial x} \right) = 0 \,. \tag{11.3}$$

The one-dimensional continuity equation, (11.2), can be also be written as

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} = 0 \,. \tag{11.4}$$

When the wind field is non-divergent, this reduces to an "advection equation" for the density:

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} = 0 \,. \tag{11.5}$$

Note, however, that in general the density is *not* conserved following a particle, because in general the wind field is divergent.

11.4 Conserving mass

Early work on conservative schemes was reported by Godunov (1959), Godunov and Bohachevsky (1959), Lorenz (1960), Arakawa (1966), and Arakawa and Lamb (1981); Roe (1981).

Suppose that we approximate (11.2) by:

$$\frac{d\rho_j}{dt} + \frac{(\rho u)_{j+1/2} - (\rho u)_{j-1/2}}{\Delta x_j} = 0, \tag{11.6}$$

This is an example of a "differential-difference equation" (sometimes called a semi-discrete equation), because the time-rate-of-change term is in differential form, while the spatial derivative has been approximated using a finite-difference quotient. We will keep time derivatives continuous for now because the issues that we are going to discuss are mostly about space differencing.

The density ρ is defined at integer points, while u and ρu are defined at half-integer points. See Fig. 11.1. This is an example of a "staggered" grid. In order to use this approach, the wind-point densities $\rho_{j+1/2}$ and $\rho_{j-1/2}$ must be interpolated somehow from the predicted values of ρ_j. Much further discussion of staggered grids, in multiple spatial dimensions, is given in later chapters.

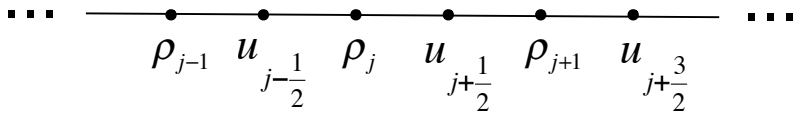


Figure 11.1: The staggered grid used in (11.10) and (11.6).

Multiply (11.6) through by Δx_j, and sum over the domain, to obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j \Delta x_j\right) + (\rho u)_{J+1/2} - (\rho u)_{-\frac{1}{2}} = 0\,. \tag{11.7}$$

If

$$(\rho u)_{J+1/2} = (\rho u)_{-\frac{1}{2}} \tag{11.8}$$

(these are periodic boundary conditions), then we obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j \Delta x_j\right) = 0\,, \tag{11.9}$$

which expresses conservation of mass (11.9). Note that (11.9) holds regardless of the form of the interpolation used for $\rho_{j+1/2}$. See Fig. 11.2.

The continuous system	*The discrete system*
$\dfrac{\partial \rho}{\partial t} = -\nabla\cdot\left(\rho\mathbf{V}\right)$	$\dfrac{\rho_j^{n+1}-\rho_j^n}{\Delta t} = \displaystyle\sum_{j'\neq j} F_{j',j}$
$\displaystyle\int_D \nabla\cdot\left(\rho\mathbf{V}\right)\, dD = 0$	$\displaystyle\sum_j \left(\sum_{j'\neq j} F_{j',j}\right) = 0$
$\dfrac{d}{dt}\displaystyle\int_D \rho\, dD = 0$	$\displaystyle\sum_j \rho_j^{n+1} - \sum_j \rho_j^n = 0$

Figure 11.2: Mass conservation in the continuous system (left) and the finite-volume system (right). The $F_{j',j}$ are mass fluxes.

11.5 Conserving an intensive scalar

In a similar way, we approximate (11.1) by

$$\frac{d}{dt}\left(\rho_j A_j\right) + \frac{(\rho u)_{j+1/2}A_{j+1/2} - (\rho u)_{j-1/2}A_{j-1/2}}{\Delta x_j} = 0\,. \tag{11.10}$$

Here A, like ρ, has an integer subscript. The half-integer values of A and ρ, i.e., $A_{j+1/2}$, $\rho_{j+1/2}$, $A_{j-1/2}$, and $\rho_{j-1/2}$, are defined on the cell walls, and must be interpolated somehow from the predicted values of A and ρ. Note that if we put $A \equiv 1$, then (11.10) reduces to the continuity equation, (11.6).

Multiply (11.10) through by Δx_j, and sum over the domain:

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j A_j \Delta x_j\right) + (\rho u)_{J+1/2} A_{J+1/2} - (\rho u)_{\frac{1}{2}} A_{\frac{1}{2}} = 0\,, \tag{11.11}$$

If

$$(\rho u)_{J+1/2} A_{J+1/2} = (\rho u)_{\frac{1}{2}} A_{\frac{1}{2}}\,, \tag{11.12}$$

(these are periodic boundary conditions), then we obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j A_j \Delta x_j\right) = 0\,, \tag{11.13}$$

which expresses conservation of the mass-weighted value of A. Note that (11.13) holds regardless of the form of the interpolation used for $A_{j+1/2}$.

Any scheme that conserves the mass-weighted value of A will also conserve any linear function of A. We may also wish to require conservation of the mass-weighted value of some function of A, such as A^2. This might correspond, for example, to conservation of kinetic energy. Energy conservation can be arranged, as we will see.

There are various additional requirements that we might like to impose. Ideally, for example, the finite-difference advection operator would not alter the PDF of A over the mass. Unfortunately this cannot be guaranteed with Eulerian methods, although we can minimize the effects of advection on the PDF, especially if the shape of the PDF is known *a priori*. This will be discussed later. In a model based on Lagrangian methods, advection does not alter the PDF of the advected quantity. That's very attractive.

11.6 An advective form

From (11.10) and (11.6), we can derive a discrete "advective form," analogous to (11.3):

$$\rho_j \frac{dA_j}{dt} + \frac{(\rho u)_{j+1/2}\left(A_{j+1/2} - A_j\right) + (\rho u)_{j-1/2}\left(A_j - A_{j-1/2}\right)}{\Delta x_j} = 0\,. \tag{11.14}$$

To understand how this approximation works, note that it can be rearranged to

$$\rho_j \frac{dA_j}{dt} + \frac{1}{2}\left[\frac{(\rho u)_{j+1/2}\left(A_{j+1/2} - A_j\right)}{\frac{1}{2}\Delta x_j} + \frac{(\rho u)_{j-1/2}\left(A_j - A_{j-1/2}\right)}{\frac{1}{2}\Delta x_j}\right] = 0\,. \tag{11.15}$$

11.7 Flux-form advection and continuity

Since (11.14) is consistent with (11.10) and (11.6), use of (11.14) and (11.6) will allow conservation of the mass-weighted value of A (and of mass itself). Also note that if A is uniform over the grid, then (11.14) gives $dA_j/dt = 0$, which is the right answer. This is ensured ***because*** (11.10) reduces to (11.6) when A is uniform over the grid. *If the flux-form advection equation did not reduce to the flux-form continuity equation when A is uniform over the grid, then a uniform tracer field would not remain uniform under advection.*

11.8 Example: A flux form of the upstream scheme

In Chapter 8, we discussed the upstream scheme in advective form. Can we write it in flux form, so that it conserves the mass-weighted value of the advected quantity? In order to do so, we must choose the interpolated values of ρ and A in the flux-form equation (11.10) so that the corresponding advective form is the upstream scheme.

We can write

$$\begin{aligned}\frac{d}{dt}\left(\rho_j A_j\right) + &\left[\frac{\rho_j u^+_{j+1/2} A_j + \rho_{j+1} u^-_{j+1/2} A_{j+1}}{\Delta x_j}\right] \\ &- \left[\frac{\rho_{j-1} u^+_{j-1/2} A_{j-1} + \rho_j u^-_{j-1/2} A_j}{\Delta x_j}\right] = 0\,.\end{aligned} \tag{11.16}$$

Here we are using the notation defined in (8.9). By setting $A \equiv 1$ in (11.16), we obtain the continuity equation in the form

$$\begin{aligned}\frac{d\rho_j}{dt} + &\left[\frac{\rho_j u^+_{j+1/2} + \rho_{j+1} u^-_{j+1/2}}{\Delta x_j}\right] \\ &- \left[\frac{\rho_{j-1} u^+_{j-1/2} + \rho_j u^-_{j-1/2}}{\Delta x_j}\right] = 0\,.\end{aligned} \tag{11.17}$$

Multiplying (11.17) by A_j, and subtracting from (11.16), we obtain the advective form

$$\rho_j \frac{dA_j}{dt} + \left[\frac{\rho_{j+1} u^-_{j+1/2} \left(A_{j+1} - A_j \right) + \rho_{j-1} u^+_{j-1/2} \left(A_j - A_{j-1} \right)}{\Delta x_j} \right] = 0 \,. \tag{11.18}$$

If we ignore local variations of the the density, then (11.18) is consistent with (8.10).

11.9 Coordinate-free definitions of operators

As discussed in Appendix A, the divergence, curl, and gradient operators can be *defined* in terms of the limits of surface integrals as the enclosed volume shrinks to zero. These definitions can be used to formulate finite-volume methods, in which the definition of the operator is used but the volume takes the form of a finite, (possibly) multi-dimensional "grid cell." See Fig. 11.3.

For example, the divergence operator can be defined using

$$\nabla \cdot \mathbf{Q} \equiv \lim_{V \to 0} \left[\frac{1}{V} \oint_S \mathbf{n} \cdot \mathbf{Q} dS \right] \,, \tag{11.19}$$

where S is the surface bounding a volume V, and $\mathbf{n}$ is the outward normal on S. In the limit shown in (11.19), both the volume and the area of its bounding surface shrink to zero. Here the terms "volume" and "bounding surface" are used in the following generalized sense: In a three-dimensional space, "volume" is literally a volume, and "bounding surface" is literally a surface. In a two-dimensional space, "volume" means an area, and "bounding surface" means the curve bounding the area. In a one-dimensional space, "volume" means a curve, and "bounding surface" means the end points of the curve.

Eq. (11.19) can be used, for example, to formulate an approximation to an advective flux divergence in terms of the normal components of the flux on the wall of the volume. The flux on each grid-cell wall adds or subtracts from the contents of the grid cell, like deposits and withdrawals from a bank account.

A definition of the gradient operator that does not make reference to any coordinate system is

$$\nabla A \equiv \lim_{V \to 0} \left[\frac{1}{V} \oint_S \mathbf{n} A \, dS \right] , \tag{11.20}$$

This can be used, for example, for formulate an approximation to the pressure-gradient force. The pressure on each grid-cell wall tries to accelerate the mass in the grid cell in the direction normal to the wall. If all of the pressures are equal, then they cancel each other out and there is no net force on the mass in the cell. But when the pressures differ from one wall to another, there can be a net force. You can connect this idea with everyday experience: If the pressure on the front of your car is higher than the pressure on the back, then the car will experience a net "drag" force that tries to slow it down.

A definition of the curl operator that does not make reference to any coordinate system is

$$\nabla \times \mathbf{Q} \equiv \lim_{V \to 0} \left[\frac{1}{V} \oint_S \mathbf{n} \times \mathbf{Q} \, dS \right] . \tag{11.21}$$

This can be used to formulate an approximation to the vorticity in terms of the tangential wind components on the bounding surface of the volume.

Finally, the Jacobian on a two-dimensional surface can be defined by

$$J(A,B) = \lim_{C \to 0} \left[\oint_C A \nabla B \cdot \mathbf{t} \, dl \right] , \tag{11.22}$$

where $\mathbf{t}$ is a unit vector that is tangent to the bounding curve C. This can be used to formulate an approximation to the Jacobian operator in terms of the grid-point values of the scalars A and B. The Jacobian operator will be used in Chapter 26.

The finite-volume method imitates the forms of (11.19) – (11.22) to construct discrete analogs of the gradient, divergence, curl, and Jacobian operators. See Fig. 11.3. The integrals in (11.19) – (11.22) are simply replaced by sums. The method can be applied to grid cells of arbitrary shapes, including those discussed in Chapters 4 and 28.

11.10 Mimetic schemes

"Mimetic" schemes *mimic* some key properties of the differential operators, such as

Divergence $$\nabla \cdot \mathbf{Q} \equiv \lim_{S\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\cdot\mathbf{Q}\,dS\right]$$ Advection

Curl $$\nabla \times \mathbf{Q} \equiv \lim_{S\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\times\mathbf{Q}\,dS\right]$$ Vorticity

Gradient $$\nabla A \equiv \lim_{S\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}A\,dS\right]$$ Pressure gradient

Figure 11.3: The coordinate-free definitions of the divergence, curl, and gradient operators. The divergence is relevant to advection, the curl is used in the definition of the vorticity, and the gradient is used to compute the pressure-gradient force.

$$\int_R \nabla\cdot\mathbf{Q}\,dR = 0 \quad \text{if the region R has either closed or periodic boundaries ,} \tag{11.23}$$

$$\int_R \nabla\times\mathbf{Q}\,dR = 0 \quad \text{if the region R has periodic boundaries ,} \tag{11.24}$$

$$\nabla\times(\nabla A) = 0\,, \tag{11.25}$$

and

$$\nabla\cdot(\nabla\times\mathbf{Q}) = 0\,. \tag{11.26}$$

In these equations, $\mathbf{Q}$ is an arbitrary vector and A is an arbitrary scalar. Although term "mimetic schemes" has been used only since the 1990s (Hyman et al., 1992), the ideas involved actually go back much further (e.g., Godunov, 1959; Godunov and Bohachevsky, 1959; Lorenz, 1960; Arakawa, 1966; Arakawa and Lamb, 1981; Roe, 1981).

11.11 Conserving a function of an advected scalar

We have already discussed the fact that, for the continuous system, conservation of A itself implies conservation of *any function* of A, e.g., A^2, A^{17}, etc. This is most easily seen from the Lagrangian form:

$$\frac{DA}{dt} = 0\,. \tag{11.27}$$

According to (11.27), A is conserved "following a particle." As discussed earlier, this implies that

$$\frac{d}{dt}\left[F\left(A\right)\right] = 0, \tag{11.28}$$

where $F(A)$ is an arbitrary function of A only. We can derive (11.28) by multiplying (11.27) by dF/dA.

In a finite-difference system, we can force conservation of at most one non-linear function of A, in addition to A itself. Here's how that works: Let F_j denote $F\left(A_j\right)$, and let F'_j denote $dF\left(A_j\right)/dA_j$. Multiplying (11.14) by F'_j gives

$$\rho_j \frac{dF_j}{dt} + \frac{(\rho u)_{j+1/2} F'_j\left(A_{j+1/2} - A_j\right) + (\rho u)_{j-1/2} F'_j\left(A_j - A_{j-1/2}\right)}{\Delta x_j} = 0\,. \tag{11.29}$$

Now use (11.6) to rewrite (11.29) in "flux form":

$$\frac{d}{dt}\left(\rho_j F_j\right) + \frac{1}{\Delta x_j}\left\{(\rho u)_{j+1/2}\left[F'_j\left(A_{j+1/2} - A_j\right) + F_j\right] - (\rho u)_{j-1/2}\left[-F'_j\left(A_j - A_{j-1/2}\right) + F_j\right]\right\} = 0\,. \tag{11.30}$$

Inspection of (11.30) shows that, to ensure conservation of $F(A)$, we must choose

$$F_{j+1/2} = F'_j\left(A_{j+1/2} - A_j\right) + F_j\,, \tag{11.31}$$

$$F_{j-1/2} = -F'_j \left(A_j - A_{j-1/2}\right) + F_j \; . \tag{11.32}$$

Let $j \to j+1$ in (11.32), giving

$$F_{j+1/2} = -F'_{j+1} \left(A_{j+1} - A_{j+1/2}\right) + F_{j+1} \; . \tag{11.33}$$

Now we have two formulae for $F_{j+1/2}$, namely (11.31) and (11.33), and they must agree if we are to to have a true flux form for prediction of F_j. This leads to

$$F'_j \left(A_{j+1/2} - A_j\right) + F_j = -F'_{j+1} \left(A_{j+1} - A_{j+1/2}\right) + F_{j+1} \; . \tag{11.34}$$

The only unknown in this equation is $A_{j+1/2}$. We find that

$$\boxed{A_{j+1/2} = \frac{\left(F'_{j+1} A_{j+1} - F_{j+1}\right) - \left(F'_j A_j - F_j\right)}{F'_{j+1} - F'_j} \; .} \tag{11.35}$$

The conclusion is that by choosing $A_{j+1/2}$ according to (11.35), we can guarantee conservation of both A and $F(A)$ (apart from time-differencing errors).

As an example, suppose that $F(A) = A^2$. Then $F'(A) = 2A$, and we find that

$$A_{j+1/2} = \frac{\left(2A_{j+1}^2 - A_{j+1}^2\right) - \left(2A_j^2 - A_j^2\right)}{2\left(A_{j+1} - A_j\right)} = \frac{1}{2}\left(A_{j+1} + A_j\right) \; . \tag{11.36}$$

We conclude that this arithmetic-mean interpolation allows conservation of the square of A. It may or may not be an *accurate* interpolation for $A_{j+1/2}$. Note that x_{j+1}, x_j, and $x_{j+1/2}$ do not appear in (11.36). This means that our spatial interpolation does not contain any information about the spatial locations of the various grid points involved – a rather awkward and somewhat strange property of the interpolation. If the grid spacing is uniform, (11.36) gives second-order accuracy in space. If the grid spacing is nonuniform, the accuracy drops to first-order, but if the grid spacing varies smoothly second-order accuracy can be maintained, as discussed in Chapter 3.

Substituting (11.36) back into (11.14) gives

$$\rho_j \frac{dA_j}{dt} + \frac{1}{2\Delta x_j}\left[(\rho u)_{j+1/2}\left(A_{j+1} - A_j\right) + (\rho u)_{j-1/2}\left(A_j - A_{j-1}\right)\right] = 0\,. \tag{11.37}$$

This is the advective form that allows conservation of A^2 (and of A).

11.12 Lots of ways to interpolate

There are infinitely many ways to interpolate a variable. A general two-point interpolation has the form

$$A_{j+1/2} = I\left(A_j, A_{j+1}\right)\,. \tag{11.38}$$

The interpolating function I may or may not include information about the spatial locations of points j and $j+1$. Any two-point interpolation should have the property that

$$\min\left\{A_j,\, A_{j+1}\right\} \le A_{j+1/2} \le \max\left\{A_j,\, A_{j+1}\right\}\,. \tag{11.39}$$

It follows from (11.39) that if $A_j = A_{j+1}$, then $A_{j+1/2} = A_j = A_{j+1}$.

It is possible and sometimes useful to create interpolations that use more than two points in more than one dimension, but two-point interpolations can help to prevent computational modes in space. Further discussion is given later.

As a simple example, we can spatially interpolate A itself in a linear fashion, e.g.,

$$A_{j+1/2} = \alpha_{j+1/2} A_j + \left(1 - \alpha_{j+1/2}\right) A_{j+1}\,, \tag{11.40}$$

where $0 \le \alpha_{j+1/2} \le 1$ is a weighting factor that might be a constant, as in (11.36), or might be a function of x_j, x_{j+1} and $x_{j+1/2}$, or a function of $\mu \equiv u\Delta t/\Delta x$. This is plotted in panel a) of Fig. 11.4, for the case $\alpha_{j+1/2} = 1/2$, which gives the "arithmetic mean." Alternatively, we can interpolate so as to conserve an arbitrary function of A, as in (11.35).

Another approach is to compute some function of $f(A)$, interpolate $f(A)$ using a form such as (11.40), and then extract an interpolated value of A by applying the inverse of $f(A)$ to the result. A practical example of this would be interpolation of the water vapor mixing ratio by computing the relative humidity from the mixing ratio, interpolating the relative humidity, and then converting back to mixing ratio. This type of interpolation does not (in general) have the property that when the two input values of A are the same the interpolated value of A is equal to the input value; instead, when the two input values of $f(A)$ are the same the interpolated value of $f(A)$ is equal to the input value.

We can also make use of "averages" that are different from the simple and familiar arithmetic mean given by (11.36). Examples are the "*geometric mean*,"

$$A_{j+1/2} = \sqrt{A_j A_{j+1}} \,. \tag{11.41}$$

and the "*harmonic mean*,"

$$A_{j+1/2} = \frac{2A_j A_{j+1}}{A_j + A_{j+1}} \,, \tag{11.42}$$

which are plotted in panels b) and c) of Fig. 11.4. The geometric mean and the harmonic mean are both nonlinear interpolations, because the geometric mean of A plus the geometric mean of B is not equal to the geometric mean of $A+B$, although it will usually be close. The geometric mean and the harmonic mean are both symmetric in the inputs; if A_j and A_{j+1} are swapped, the result is unchanged. The geometric mean and the harmonic mean both have the potentially useful property that if either A_{j+1} or A_j is equal to zero, then $A_{j+1/2}$ will also be equal to zero. More generally, both (11.41) and (11.42) tend to make the interpolated value close to the smaller of the two input values.

Define normalized values of A, such that

$$r_{j,j+\frac{1}{2}} \equiv \frac{A_j}{\max\left\{A_j,\, A_{j+1}\right\}} \quad \text{and} \quad r_{j+1,j+\frac{1}{2}} \equiv \frac{A_{j+1}}{\max\left\{A_j,\, A_{j+1}\right\}} \,. \tag{11.43}$$

Note that $0 \le r_{j,j+\frac{1}{2}} \le 1$ and $0 \le r_{j+1,j+\frac{1}{2}} \le 1$. We specify $r_{j+1/2}$ as some function of $r_{j,j+\frac{1}{2}}$ and $r_{j+1,j+\frac{1}{2}}$, such that $0 \le r_{j+1/2} \le 1$. Various interpolations can be constructed by choosing the function in particular ways, then using $A_{j+1/2} = r_{j+\frac{1}{2}} \max\left\{A_j, A_{j+1}\right\}$. For example, an interpolation that make the interpolated value close to the *larger* input value is

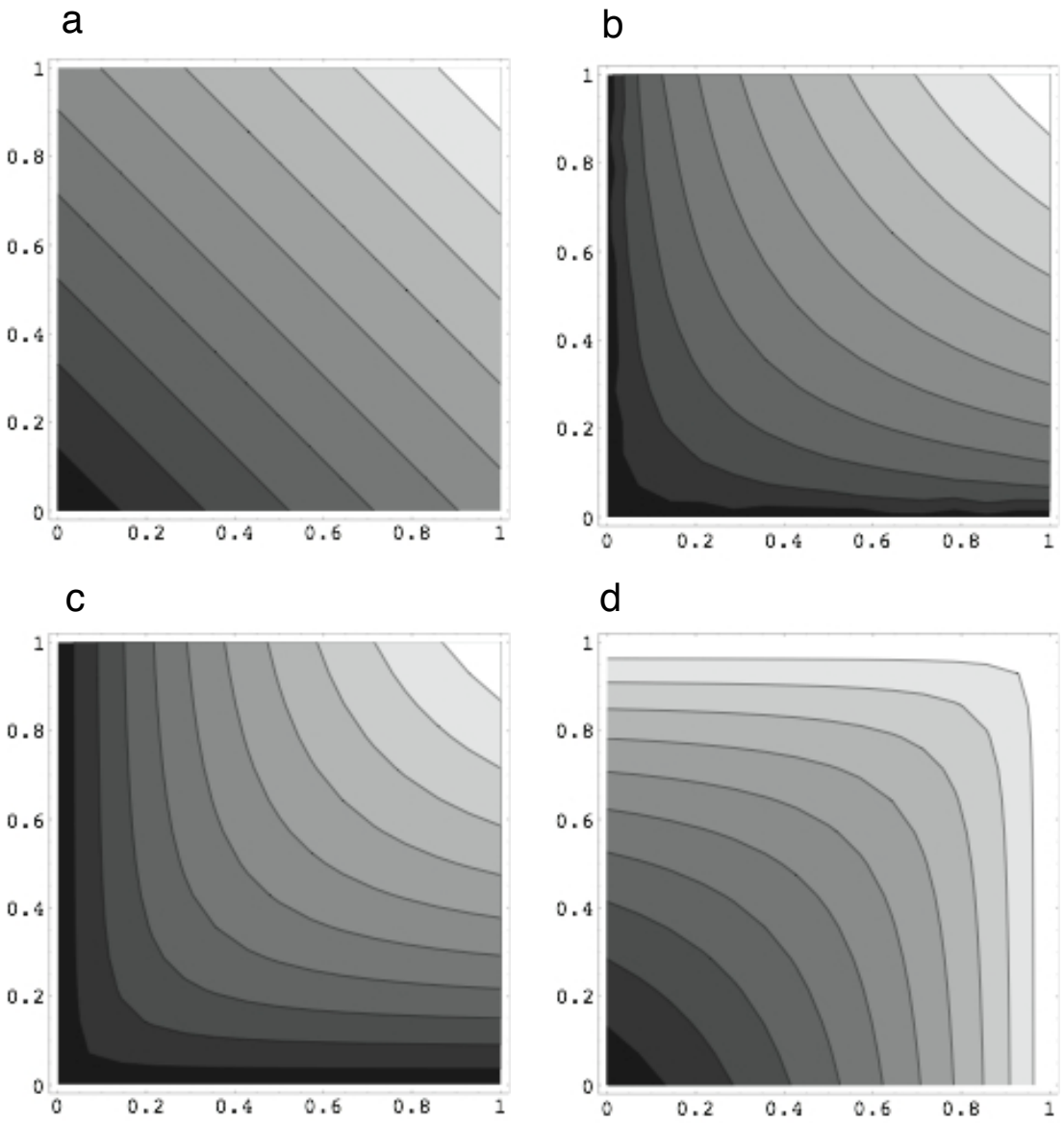


Figure 11.4: Four interpolations as functions of the input values. a) arithmetic mean, b) geometric mean, c) harmonic mean, d) Eq. (11.44), which makes the interpolated value close to the larger of the two input values. The independent variables on the abscissas and ordinates are the input values, which are assumed to run between 0 and 1. In all plots, black is close to zero, and white is close to one.

$$r_{j+1/2} = \frac{r_{j,j+\frac{1}{2}} + r_{j+1,j+\frac{1}{2}} - 2r_{j,j+\frac{1}{2}} r_{j+1,j+\frac{1}{2}}}{2 - \left(r_{j,j+\frac{1}{2}} + r_{j+1,j+\frac{1}{2}} \right)} \,. \tag{11.44}$$

Eq. (11.44) is plotted in panel d) of Fig. 11.4. When $r_{j,j+\frac{1}{2}} = r_{j+1,j+\frac{1}{2}}$, Eq. (11.44) gives $r_{j+1/2} = r_{j,j+\frac{1}{2}} = r_{j+1,j+\frac{1}{2}}$, as expected. Also, (11.44) gives $r_{j+1/2} = 1$ if either $r_{j,j+\frac{1}{2}} = 1$ or $r_{j+1,j+\frac{1}{2}} = 1$. Finally, to obtain $A_{j+\frac{1}{2}}$, we use

$$A_{j+\frac{1}{2}} = r_{j+\frac{1}{2}} \max\left\{A_j,\, A_{j+1}\right\} . \tag{11.45}$$

Where does (11.44) come from? Rewrite the harmonic mean, (11.42), as

$$r_{j+1/2} = \frac{2 r_j r_{j+1}}{r_j + r_{j+1}} , \tag{11.46}$$

where, as before, $0 \le r \le 1$. Now replace each value of r in (11.46) by $1 - r$:

$$\left(1 - r_{j+1/2}\right) = \frac{2\left(1 - r_j\right)\left(1 - r_{j+1}\right)}{\left(1 - r_j\right) + \left(1 - r_{j+1}\right)} . \tag{11.47}$$

Eq. (11.47) gives the harmonic mean of $1 - r$. Solving (11.47) for $r_{j+1/2}$, we obtain

$$r_{j+1/2} = \frac{r_j + r_{j+1} - 2 r_j r_{j+1}}{2 - \left(r_j + r_{j+1}\right)} , \tag{11.48}$$

which is similar to (11.44).

The fact that there are infinitely many ways to average and/or interpolate can be viewed as a good thing, because it means that we have the opportunity to choose the *best* interpolation for a particular application.

11.13 Fixers

For reasons that will be discussed later, some models do not use conservative forms of the continuity equation. It is possible to "fix" conservation of mass by checking at the end of the time step (or perhaps at the end of a simulated day) to see how much mass has been gained or lost globally, and then just subtracting or adding whatever it takes to restore the total mass at the beginning of the time step (or day). These *ad hoc* procedures are called "fixers" (not to be confused with shady attorneys). I don't recommend fixers, because they are unphysical and can mask underlying problems with a model. The finite-volume method can eliminate the need for fixers.

11.14 Segue

Finite-difference schemes for the advection equation can be designed to allow "exact" or "formal" conservation of mass, of the advected quantity itself (such as potential temperature), and of one arbitrary function of the advected quantity (such as the square of the potential temperature). Conservative schemes mimic the "form" of the exact equations. In addition, they are often well behaved computationally. Since coding errors often lead to failure to conserve, conservative schemes can be easier to de-bug than non-conservative schemes.

11.15 Problems

1. Find a one-dimensional advection scheme that conserves both A and $\ln(A)$. Keep the time derivative continuous.

2. Consider the continuity equation

$$\frac{d\rho_j}{dt} + \frac{(\hat{\rho}u)_{j+1/2} - (\hat{\rho}u)_{j-1/2}}{\Delta x} = 0 \,, \tag{11.49}$$

   and the advection equation

$$\frac{dA_j}{dt} + \frac{1}{2}\left(u_{j+1/2} + u_{j-1/2}\right)\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0 \,. \tag{11.50}$$

   Does this scheme conserve the mass-weighted average value of A? Give a proof to support your answer.

3. Determine the order of accuracy of

$$\left(\frac{\partial A}{\partial x}\right)_j \cong \frac{A_{j+1/2} - A_{j-1/2}}{\Delta x} \tag{11.51}$$

   when the harmonic mean,

$$A_{j+1/2} = \frac{2A_jA_{j+1}}{A_j + A_{j+1}} \,, \tag{11.52}$$

is used for interpolation to the cell walls. Assume uniform grid spacing.

4. Prove that the geometric mean cannot be larger than the arithmetic mean, assuming that the input values are non-negative.

Chapter 12

Computational dispersion

12.1 Dispersion with centered space differencing

12.1.1 The phase velocity

In the continuous world, if $A(x,\, t) = \widehat{A}_k(t)\, e^{ikx}$, where k is the wave number, then

$$\frac{\partial A}{\partial x} = ikA \;, \tag{12.1}$$

and the advection equation can be written as

$$\frac{\partial \widehat{A}_k}{\partial t} = -iku\widehat{A}_k \;. \tag{12.2}$$

We recognize this as a form of the oscillation equation. The solution is

$$\widehat{A}_k = \widehat{A}_k(0)\, e^{-ikut} \;, \tag{12.3}$$

which leads to

$$A = \widehat{A}_k(0)\, e^{ik(x-ut)} \;. \tag{12.4}$$

The signal travels at the speed u, regardless of the value of k.

With the centered-difference quotient

$$\left(\frac{\partial A}{\partial x}\right)_j \cong \frac{A_{j+1}-A_{j-1}}{2\Delta x}\,, \tag{12.5}$$

if $A_j(t)$ has the wave form $A_j(t) = \widehat{A}_k(t)\, e^{ikj\Delta x}$, then

$$\begin{aligned}
\frac{A_{j+1}-A_{j-1}}{2\Delta x} =& \frac{\widehat{A}_k}{2\Delta x}\left[e^{ik(j+1)\Delta x} - e^{ik(j-1)\Delta x}\right] \\
=& \frac{\widehat{A}_k e^{ikj\Delta x}}{2\Delta x}\left(e^{ik\Delta x} - e^{-ik\Delta x}\right) \\
=& \frac{\widehat{A}_k e^{ikj\Delta x}}{2\Delta x}\left[2i\sin(k\Delta x)\right] \\
=& ik\left[\frac{\sin(k\Delta x)}{k\Delta x}\right]\widehat{A}_k(t)\, e^{ikj\Delta x} \\
=& ik\,\mathrm{sinc}(k\Delta x)\,\widehat{A}_k(t)\, e^{ikj\Delta x}\,,
\end{aligned} \tag{12.6}$$

where for any α we define

$$\mathrm{sinc}(\alpha) \equiv \frac{\sin(\alpha)}{\alpha}\,. \tag{12.7}$$

A plot of $\mathrm{sinc}(\alpha)$ is given in Fig. 12.1. Note that $\mathrm{sinc}(\alpha) \to 1$ as $\alpha \to 0$.

Using (12.6), the discrete advection equation can be written as

$$\frac{d\widehat{A}_k}{dt} = -iku\,\mathrm{sinc}(k\Delta x)\,\widehat{A}_k\,. \tag{12.8}$$

Compare with (12.2). The signal travels at the speed $u\,\mathrm{sinc}(k\Delta x)$, which depends on the wave number; the short scales travel more slowly than the longer scales. We can define an *effective wind speed*, u^*, that depends on the wave number:

$$u^* = u\,\mathrm{sinc}(k\Delta x)\,. \tag{12.9}$$

Of course, the true wind speed, u, is independent of $k\Delta x$. A plot of u^*/u versus $k\Delta x$ is given by the upper curve in Fig. 12.2. (The second, lower curve, which illustrates the computational group velocity, will be discussed later.) If we have multiple wave components superimposed on one another, each component will move with a different effective wind speed, depending on its wave number. *The total "pattern" formed by the superimposed waves will come apart, or "disperse," as the waves separate from each other.* This is called a "*computational dispersion*."

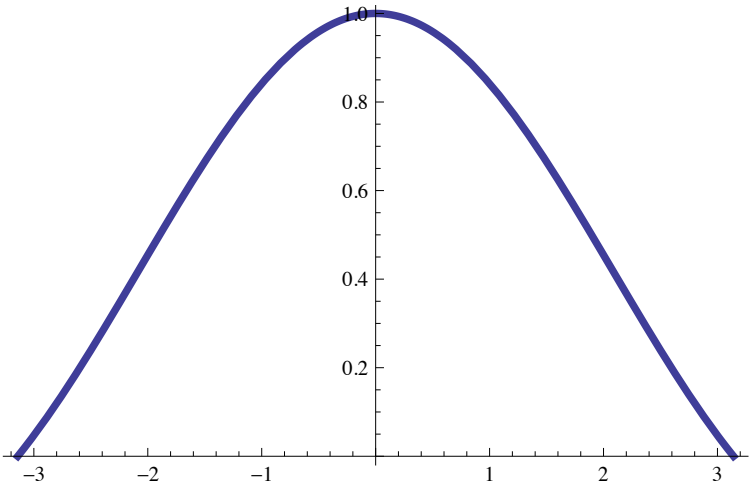


Figure 12.1: A plot of $\mathrm{sinc}\,(\alpha)$, for $-\pi \le \alpha \le \pi$.

For the case of leapfrog time differencing and centered second-order space differencing, the one-dimensional advection equation is

$$\frac{A_j^{n+1} - A_j^{n-1}}{2\Delta t} + u\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = 0\,. \tag{12.10}$$

It is already apparent from the form of (12.10) that the $2\Delta x$ wave will not be advected at all. Referring back to (12.6), we see that (12.10) leads to

$$\widehat{A}_k^{n+1} - \widehat{A}_k^{n-1} = 2i\Omega\widehat{A}_k^n\,, \tag{12.11}$$

where

$$\Omega \equiv -ku\,\mathrm{sinc}\,(k\Delta x)\,\Delta t\,. \tag{12.12}$$

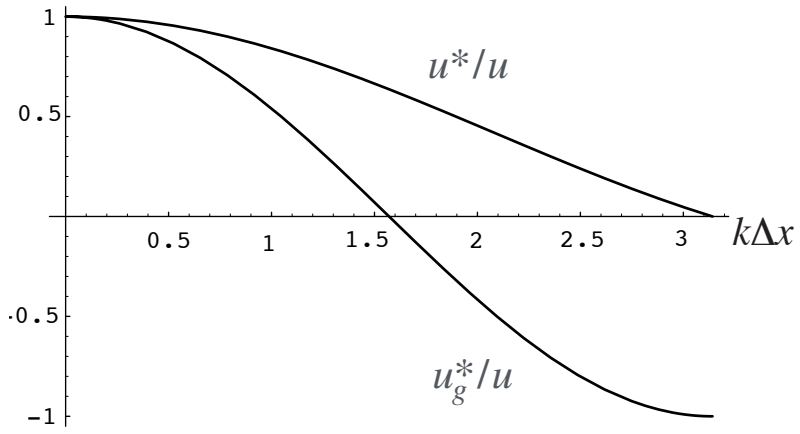


Figure 12.2: The ratio of the computational effective wind speed to the true wind speed, and also the ratio of the computational group speed to the true group speed, both plotted as functions of wave number.

We recognize Eq. (12.11) as the leapfrog scheme for the oscillation equation. Recall from Chapter 6 that $|\Omega| \leq 1$ is necessary for (12.11) to be stable. Therefore,we can simply re-use our result from Chapter 6, i.e.,

$$|ku \,\mathrm{sinc}\,(k\Delta x)\,\Delta t| = \left|\sin(k\Delta x)\,\frac{u\Delta t}{\Delta x}\right| \leq 1 \tag{12.13}$$

must hold for stability, for any and all k. The "*worst case*" is $|\sin k\Delta x| = 1$, which occurs for $k\Delta x = \pm\pi/2$, corresponding to the wavelength $L = 4\Delta x$. The $2\Delta x$ wave is not the problem here, because as mentioned above $\Omega = 0$ for that mode. It is the $4\Delta x$ wave that can most easily become unstable. We conclude that

$$\frac{|u|\,\Delta t}{\Delta x} \leq 1 \tag{12.14}$$

is the necessary condition for stability. *Here "stability" means stability for all modes*. The stability criterion (12.14) also applies to the upstream scheme, as we saw already in Chapter 8.

Recall that the leapfrog scheme gives a numerical solution with two modes - a physical mode and a computational mode. We can write these two modes as in Chapter 6:

$$\left(\widehat{A}_k\right)_1^n = \lambda_1^n \left(\widehat{A}_k\right)_1^0, \text{ and } \left(\widehat{A}_k\right)_2^n = \lambda_2^n \left(\widehat{A}_k\right)_2^0. \tag{12.15}$$

Here the subscripts denote the mode. For $|\Omega| \leq 1$, we find, as discussed in Chapter 6, that

$$\lambda_1 = e^{i\theta}, \text{ and } \lambda_2 = e^{i(\pi-\theta)} = -e^{-i\theta}, \text{ where } \theta \equiv \tan^{-1}\left(\frac{\Omega}{\sqrt{1-\Omega^2}}\right). \tag{12.16}$$

Both modes are neutral. For the physical mode,

$$\begin{aligned}\left(A_j^n\right)_1 &= \lambda_1^n \left(\widehat{A}_k\right)_1^0 e^{ikj\Delta x} \\ &= \left(\widehat{A}_k\right)_1^0 \exp\left[ik\left(j\Delta x + \frac{\theta}{k\Delta t} n\Delta t\right)\right].\end{aligned} \tag{12.17}$$

Similarly, for the computational mode,

$$\left(A_j^n\right)_2 = \left(\widehat{A}_k\right)_2^0 (-1)^n \exp\left[ik\left(j\Delta x - \frac{\theta}{k\Delta t} n\Delta t\right)\right]. \tag{12.18}$$

Note the nasty factor of $(-1)^n$ in (12.18), which comes from the leading minus sign in (12.16). Comparing (12.17) and (12.18) with the expression $A(x,t) = \widehat{A}_k(0)\, e^{ik(x-ut)}$, which is the true solution, we see that the speeds of the physical and computational modes are $-\frac{\theta}{k\Delta t}$ and $\frac{\theta}{k\Delta t}$, respectively, for even time steps. The speed of the physical mode approaches u (i.e., the right answer), while the speed of the computational mode approaches $-u$. *The computational mode goes backwards!* In other words, the computational solution advects the signal towards the upwind direction, which is obviously crazy.

For the physical mode, the effective wind speed depends on k, while the true wind speed, u, is of course independent of k. *The shorter waves move more slowly than the longer waves*, and as already mentioned the $2\Delta x$ does not move at all.

12.1.2 The group velocity

Now we briefly digress to explain the concept of group velocity, in the context of the continuous equations. Suppose that we have a superposition of two waves, with slightly different wave numbers k_1 and k_2. Define

$$k \equiv \frac{k_1+k_2}{2}, \quad u \equiv \frac{u_1^*+u_2^*}{2}, \quad \Delta k \equiv \frac{k_1-k_2}{2}, \quad \Delta u \equiv \frac{u_1^*-u_2^*}{2}, \quad \Delta(ku) \equiv \frac{k_1u_1^*-k_2u_2^*}{2} . \tag{12.19}$$

See Fig. 12.3. Note that $k_1 = k+\Delta k$ and $k_2 = k-\Delta k$. Similarly, $u_1^* = u+\Delta u$ and $u_2^* = u-\Delta u$. From these relations, it follows that

$$\Delta(ku) = k\Delta u + u\Delta k . \tag{12.20}$$

You should be able to show that

$$k_1u_1^* \cong ku+\Delta(ku) \quad \text{and} \quad k_2u_2^* \cong ku-\Delta(ku) . \tag{12.21}$$

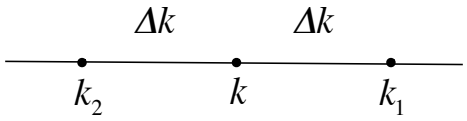


Figure 12.3: Sketch defining notation used in the discussion of the group velocity.

In these approximate equations we neglect terms involving the product $\Delta k\Delta u$, which is acceptable when $k_1 \cong k_2$ and $u_1^* \cong u_2^*$. Using (12.21), we can write the sum of two waves, each with unit amplitude, as

$$\begin{aligned}
&\exp[ik_1(x-u_1^*t)] + \exp[ik_2(x-u_2^*t)] \\
&\qquad \cong \exp(i\{(k+\Delta k)x-[ku+\Delta(ku)]t\}) + \exp(i\{(k-\Delta k)x-[ku-\Delta(ku)]t\}) \\
&\qquad = \exp[ik(x-ut)]\left(\exp\{i[(\Delta k)x-\Delta(ku)t]\} + \exp\{-i[(\Delta k)x-\Delta(ku)t]\}\right) \\
&\qquad = 2\cos[(\Delta k)x-\Delta(ku)t]\exp[ik(x-ut)] \\
&\qquad = 2\cos\left\{\Delta k\left[x-\frac{\Delta(ku)}{\Delta k}t\right]\right\}\exp[ik(x-ut)] .
\end{aligned} \tag{12.22}$$

When Δk is small, the factor $\cos\left\{\Delta k\left[x-\frac{\Delta(ku)}{\Delta k}t\right]\right\}$ behaves like the outer, slowly varying envelope in Fig. 12.4.

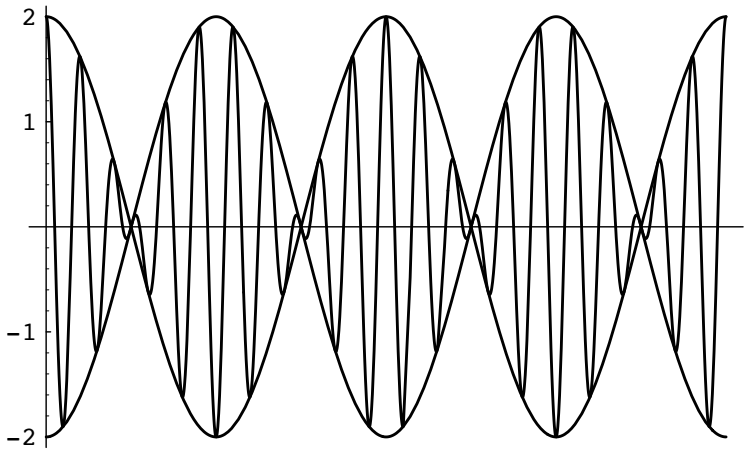


Figure 12.4: Sketch used to illustrate the concept of group velocity. The short waves are modulated by longer waves.

The envelope "modulates" wave k, which is represented by the inner, rapidly varying curve in the figure. The short waves move with effective wind speed u, but the "wave packets", i.e., the envelopes of the short waves, move with speed $\Delta(ku)/\Delta k$. The differential expression $d(ku)/dk \equiv u_g$ is called the "group velocity." Note that $u_g = u$ if u does not depend on k. *For advection by a uniform current, the "right answer" is $u_g = u$, i.e., the group velocity and wind speed should be the same*. For this reason, there is no need to discuss the group velocity for advection in the context of the continuous equations with a constant advecting current.

With our discrete scheme, however, we have

$$u_g^* = \frac{d(ku^*)}{dk} = u\frac{d}{dk}\left(\frac{\sin k\Delta x}{\Delta x}\right) = u\cos k\Delta x\,. \tag{12.23}$$

A plot of u_g^*/u versus $k\Delta x$ is given in Fig. 12.2. Note that $u_g^* = 0$ for the $4\Delta x$ wave, and $u_g^* < 0$ for the $2\Delta x$ wave. This means that wave groups with wavelengths between $L = 4\Delta x$ and $L = 2\Delta x$ have negative group velocities. Very close to $L = 2\Delta x$, u_g^* actually approaches $-u$, when in reality it should be equal to u for all wavelengths. For all waves, $u_g^* < u^* < u = u_g$. These issues arise from the space differencing; they have nothing to do with time differencing.

Fig. 12.5, which is a modified version of Fig. 12.4, makes this point in a different way, for the particular case $L = 2\Delta x$. Consider the upper solid curve and the thick red dashed curve. If we denote points on the thick red curve (corresponding to our solution with $L = 2\Delta x$) by A_j, and points on the upper solid curve (the envelope of the thick dashed curve, moving with speed u_g^*) by B_j, then

$$B_j = (-1)^j A_j \, . \tag{12.24}$$

(This is true only for the particular case $L = 2\Delta x$.) The advection equation for B_j is

$$\frac{dB_j}{dt} + (-u)\left(\frac{B_{j+1} - B_{j-1}}{2\Delta x}\right) = 0 \, . \tag{12.25}$$

Eq. (12.25) shows that the upper solid curve moves with speed $-u$; it goes backwards!

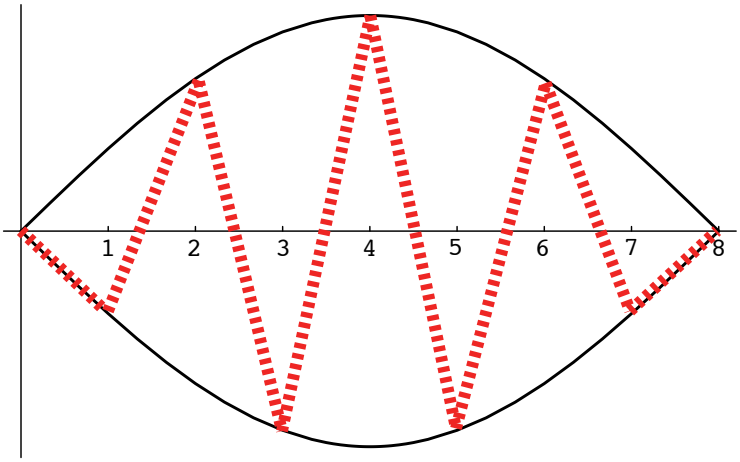


Figure 12.5: Yet another sketch used to illustrate the concept of group velocity. The short wave shown by the red dashed lin has wavelength $L = 2\Delta x$.

Recall that when we introduce time differencing, the computed phase change per time step is generally not equal to $-ku\Delta t$. This leads to changes in u^* and u_g^*, although the formulas discussed above remain valid for $\Delta t \to 0$.

12.1.3 The analyses of Matsuno and Wurtele

We now present an analysis of dispersion error with a continuous time derivative, following the work of Matsuno (1966). If we write the discrete advection equation in the form

$$\frac{dA_j}{dt} = \frac{A_{j-1} - A_{j+1}}{2\Delta x} \, , \tag{12.26}$$

then

$$2\frac{dA_j}{d\tau} = A_{j-1} - A_{j+1} \; . \tag{12.27}$$

where $\tau \equiv tu/\Delta x$ is a nondimensional time. I don't expect you to know this, but (12.27) has the same form as a recursion formula satisfied by the Bessel functions of the first kind of order j. These functions are usually denoted by $J_j(\tau)$. The $J_j(\tau)$ have the property that $J_0(0) = 1$, and $J_j(0) = 0$ for all $j \neq 0$. Because the $J_j(\tau)$ satisfy (12.27), each $J_j(\tau)$ can be interpreted as the solution of (12.27) at grid point j, as a function of the nondimensional time, τ. The "initial condition," for $\tau = 0$, is equal to one at grid point zero, and equal to zero everywhere else.

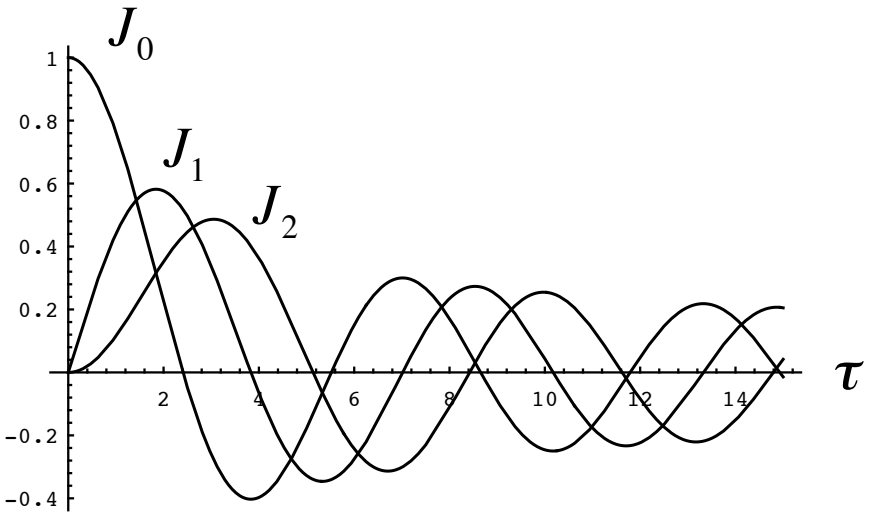


Figure 12.6: The time evolution of the solution of (12.27) at grid points $j = 0$, 1, and 2.

As an example, set $A_j = J_j(\tau)$, which is consistent with and in fact implies the initial conditions that $A_0(0) = 1$ and $A_j(0) = 0$ for all $j \neq 0$. This initial condition is an isolated "spike" at $j = 0$. The solution of (12.27) for the points $j = 0$, 1, and 2 is illustrated in Fig. 12.6. By using the identity

$$J_{-j} = (-1)^j J_j \; , \tag{12.28}$$

we can also obtain the solutions for the points $j = -1, -2, -3$, etc. This analysis is useful because it allows us to obtain the exact solution of the differential-difference equation, in which the time derivative is continuous so that there are no errors associated with time differencing.

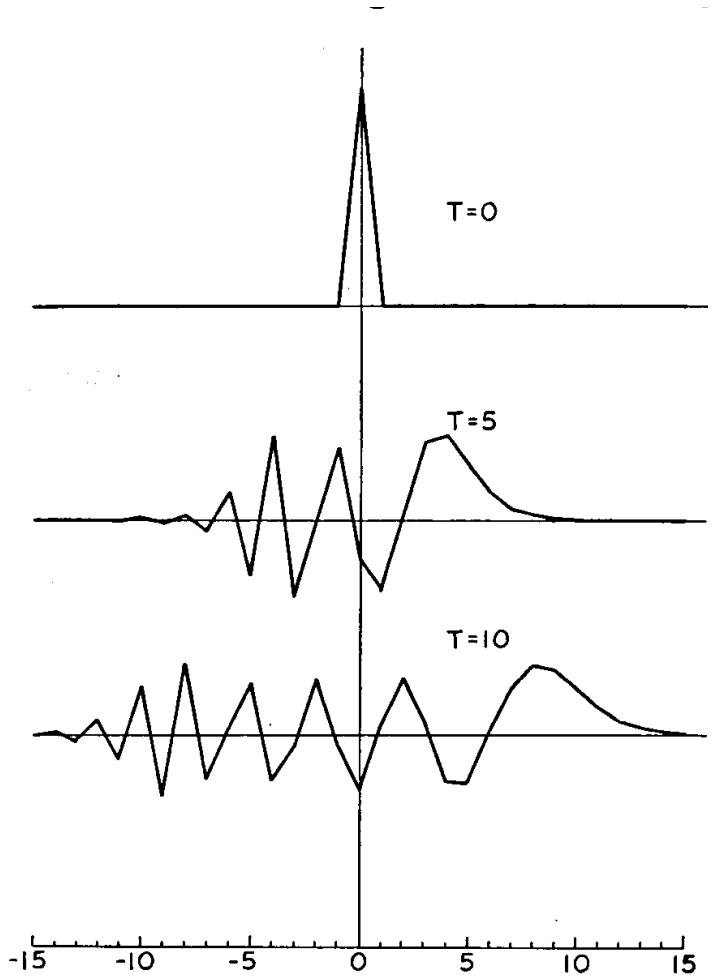


Figure 12.7: The solution of (12.27) for $\tau = 5$ and $\tau = 10$ for $-15 \leq j \leq 15$, with "spike" initial conditions. From Matsuno (1966).

Fig. 12.7 shows the solution of (12.27) for $\tau = 5$ and $\tau = 10$ for $-15 \leq j \leq 15$, with these "spike" initial conditions. The figure is taken from a paper by Matsuno (1966). Computational dispersion, schematically illustrated earlier in Fig. 12.2 and Fig. 12.5, is seen directly here. The figure also shows that u_g^* is negative for the shortest wave.

A similar type of solution is shown in Fig. 12.8, which is taken from a paper by Wurtele (1961). Here the initial conditions are slightly different, namely,

$$A_{-1} = A_0 = A_1 = 1, \text{ and } A_j = 0 \text{ for } j \leq -2, \quad j \geq 2\,. \tag{12.29}$$

This is a "top hat" or "box car" initial condition. We can construct it by combining

$$J_{j-1}(0) = 1 \text{ for } j = 1 \text{ and zero elsewhere}, \tag{12.30}$$

$$J_j(0) = 1 \text{ for } j = 0 \text{ and zero elsewhere, and} \tag{12.31}$$

and

$$J_{j+1}(0) = 1 \text{ for } j = -1 \text{ and zero elsewhere}. \tag{12.32}$$

The initial condition is $A_j(0) = J_{j-1}(0) + J_j(0) + J_{j+1}(0)$, which has the form of a "boxcar," three grid points wide. The full solution is given by

$$A_j(\tau) = J_{j-1}(\tau) + J_j(\tau) + J_{j+1}(\tau)\ . \tag{12.33}$$

The analytical solution is that the boxcar simply moves to the right, without any change in shape. This is illustrated in Fig. 12.8. Dispersion is evident in the two numerical solutions, however. The dashed curve shows the solution obtained with for centered space differencing, and the solid curve shows the solution obtained with an upstream difference like the one used in the upstream scheme. The solution for the upstream case is given in terms of the Poisson distribution rather than Bessel functions; see Wurtele (1961) for details. The figure shows that the principal disturbance moves to the right, but the short-wave components move to the left.

Don't confuse computational dispersion with instability. Both dispersion and instability can lead to "noise," but a noisy solution isn't necessarily unstable. In the case of dispersion, the waves are not growing in amplitude; instead, they separating from one another ("dispersing"), each moving at its own speed.

12.1.4 Fourth-order schemes

As discussed in Chapter 3, the fourth-order difference quotient takes the form

$$\left(\frac{\partial A}{\partial x}\right)_j = \frac{4}{3}\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) - \frac{1}{3}\left(\frac{A_{j+2} - A_{j-2}}{4\Delta x}\right) + \mathscr{O}\left[(\Delta x)^4\right]\ . \tag{12.34}$$

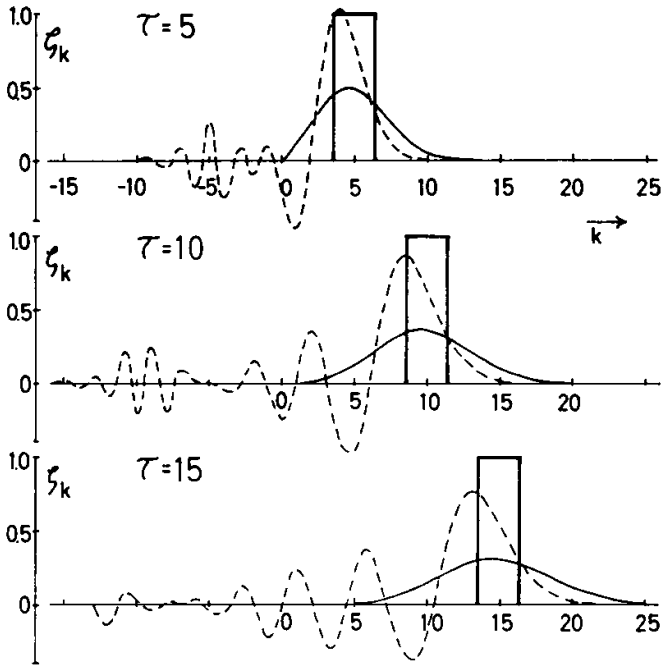


Fig. 1. Three solutions of the advection equation for (non-dimensional) times $\tau = 5, 10, 15$.
(exact) solution of continuous equation (4)
solution (8) of centered differential-difference equation
solution (11) of backward differential-difference equation
For typical meteorological values, ten units of non-dimensional time correspond to about 42 hours.

Figure 12.8: The solution of (12.24) with "box" initial conditions. From Wurtele (1961).

Recall that in our earlier discussion of the second-order scheme, we derived an expression for the effective wind speed of the numerical solution given by

$$u^* = u \,\mathrm{sinc}\,(k\Delta x) \ . \tag{12.35}$$

For this fourth-order scheme, the corresponding expression for the effective wind speed is

$$u^* = u \left[\frac{4}{3}\mathrm{sinc}\,(k\Delta x) - \frac{1}{3}\mathrm{sinc}\,(2k\Delta x) \right] \ . \tag{12.36}$$

Fig. 12.9 shows plots of u^*/u versus $k\Delta x$ for both the second-order and fourth-order schemes. For long waves, the fourth-order scheme gives a considerable improvement in the accuracy of the effective wind speed. There is no improvement for wavelengths close to $L = 2\Delta x$, however. This illustrates that increasing the order of accuracy does not help with the errors of the shortest waves.

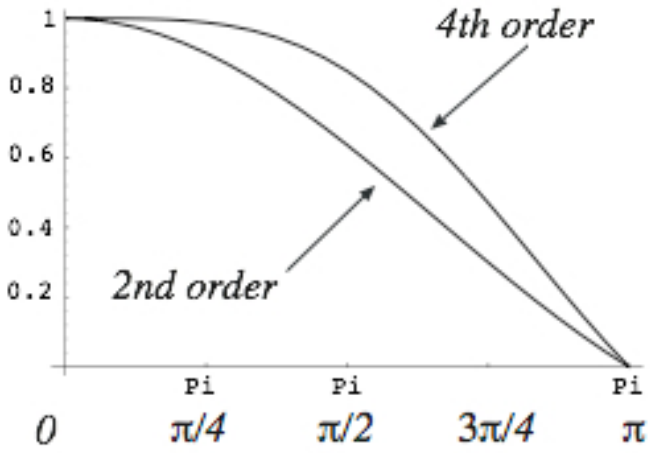


Figure 12.9: The ratio of the computational effective wind speed, u^*, to the true wind speed, u, plotted as a function of $k\Delta x$, for the second-order and fourth-order schemes.

12.2 Space-uncentered schemes

One way in which computational dispersion can be reduced in the numerical solution of the advection equation is to use uncentered space differencing, as, for example, in the upstream scheme. Recall that earlier we defined and illustrated the concept of the "domain of dependence." By reversing the idea, we can define a "domain of influence." For example, the domain of influence for explicit non-iterative space-centered schemes expands in time as is shown by the union of Regions I and II in Fig. 12.10. For $u > 0$, the domain of dependence is region II only. For $u < 0$ the domain of dependence is region I only.

The "upstream scheme," given by

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + u\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) = 0 \quad \text{for} \quad u > 0\,, \tag{12.37}$$

or

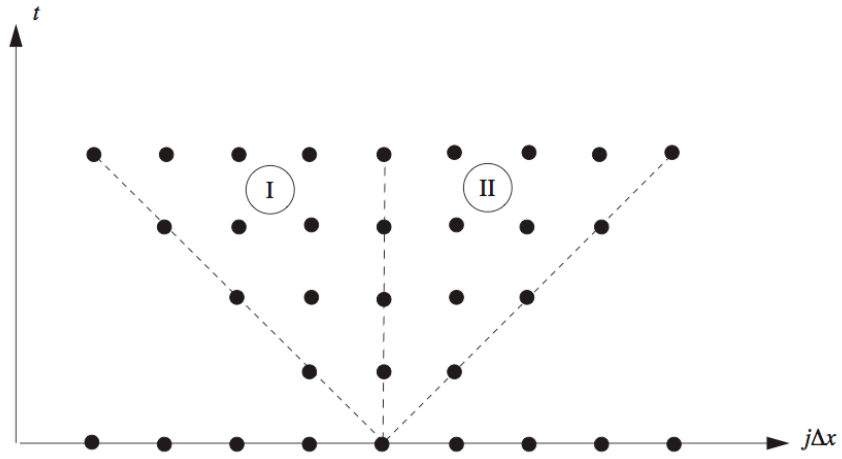


Figure 12.10: The domain of influence for explicit non-iterative space-centered schemes expands in time, as is shown by the union of Regions I and II.

$$\frac{A_j^{n+1}-A_j^n}{\Delta t}+u\left(\frac{A_{j+1}^n-A_j^n}{\Delta x}\right)=0 \quad \text{for} \quad u<0\,, \tag{12.38}$$

is the simplest example of a space-uncentered scheme. As shown earlier, we can write (12.37) as

$$A_j^{n+1}=(1-\mu)A_j^n+\mu A_{j-1}^n\,, \tag{12.39}$$

which has the form of an interpolation, and (12.38) can be written in a similar way. Obviously, for the upstream scheme, Region II alone is the domain of influence when $u>0$, and Region I alone is the domain of influence when $u<0$. This is good. The scheme produces strong damping, however, as shown in Fig. 12.8. The damping results from the interpolation. Although we can reduce the undesirable effects of computational dispersion by using the upstream scheme, usually the disadvantages of the damping outweigh the advantages of reduced dispersion. Further discussion is given in Chapter 13.

12.3 Quantifying the amplitude and phase errors

Following Takacs (1985), the *relative error*, ε, can be defined by

$$\lambda\equiv(1+\varepsilon)\lambda_T\,, \tag{12.40}$$

Where λ_T is the amplification factor for the exact solution of the differential equation, as given by (6.14). This definition of ε is applicable to any problem, not just the oscillation equation. The relative error can be separated into its real and imaginary parts:

$$\varepsilon \equiv \varepsilon_R + i\varepsilon_I \ . \tag{12.41}$$

If we know λ and λ_T, we can determine ε_R and ε_I from (12.40) and (12.41). With these definitions, we can write for the oscillation equation (which has $|\lambda_T| = 1$)

$$\begin{aligned} |\lambda|^2 &\equiv |1+\varepsilon|^2 \\ &= (1+\varepsilon_R)^2 + \varepsilon_I^2 \\ &\equiv 1 + \varepsilon_{|\lambda|} \ , \end{aligned} \tag{12.42}$$

where

$$\boxed{\varepsilon_{|\lambda|} \equiv 2\varepsilon_R + \varepsilon_R^2 + \varepsilon_I^2} \tag{12.43}$$

can be interpreted as a measure of the amplitude error. Similarly, we define ε_φ as a measure of the phase error, such that

$$\theta \equiv \Omega + \varepsilon_\varphi \ , \tag{12.44}$$

It can be shown, using a trigonometric identity, that

$$\boxed{\varepsilon_\varphi = \tan^{-1}\left(\frac{\varepsilon_I}{\varepsilon_R + 1}\right) \ .} \tag{12.45}$$

In most cases, $\varepsilon_{|\lambda|}$ is dominated by ε_R, and ε_φ is dominated by ε_I.

12.4 Even- and odd-order schemes

Takacs (1985) made an important and general point about the differences between space-centered and space-uncentered schemes, based on an analysis of how the amplitude and phase errors change as we go from an even-order scheme to the next-higher odd-order scheme, and then on to the next higher even-order scheme, and so on.

Consider the family of schemes given by (9.1), which is repeated here for convenience:

$$A_j^{n+1} = \sum_{j'=-\infty}^{\infty} a_{j'} A_{j+j'}^n \ . \tag{12.46}$$

In Chapter 9, we showed that the amplification factor for this family satisfies (9.15), which is also repeated here:

$$\lambda = \sum_{j'=-\infty}^{\infty} a_{j'} e^{ij'k\Delta x} \ . \tag{12.47}$$

Comparing (12.40) with (12.47), and using

$$\lambda_T = e^{-iku\Delta t} = e^{-i\mu k\Delta x} \ , \tag{12.48}$$

we find that

$$\varepsilon = \lambda - 1 \ . \tag{12.49}$$

The real part of the relative error is given by

$$\varepsilon_R = \sum_{j'=-\infty}^{\infty} a_{j'} \cos\left[\left(j' + \mu\right) k\Delta x\right] - 1 \ . \tag{12.50}$$

Similarly, the imaginary part of the relative error satisfies

$$\varepsilon_I = \sum_{j'=-\infty}^{\infty} a_{j'} \sin\left[\left(j'+\mu\right) k\Delta x\right] \ . \tag{12.51}$$

Now comes the key step. We want to understand how ε_R and ε_I depend on $k\Delta x$, and how this dependence changes as we increase the order of accuracy of the scheme, step-by-step. To do this, we expand (12.50) and (12.51) as Taylor series in powers of $k\Delta x$. Because ε_R involves the cosine of $(j+\mu)\,k\Delta x$, its expansion involves only even powers of $k\Delta x$:

$$\varepsilon_R = -\frac{(k\Delta x)^2}{2!} \sum_{j'=-\infty}^{\infty} \left(j'+\mu\right)^2 a_{j'} + \frac{(k\Delta x)^4}{4!} \sum_{j'=-\infty}^{\infty} \left(j'+\mu\right)^4 a_{j'} + \ldots \ . \tag{12.52}$$

Here we have used (9.8). Similarly, because ε_I involves the sine of $(j+\mu)\,k\Delta x$, its expansion involves only odd powers of $k\Delta x$:

$$\varepsilon_I = (k\Delta x) \sum_{j'=-\infty}^{\infty} \left(j'+\mu\right) a_{j'} + \frac{(k\Delta x)^3}{3!} \sum_{j'=-\infty}^{\infty} \left(j'+\mu\right)^3 a_{j'} + \ldots \ . \tag{12.53}$$

What do these two results mean? Expressions of the form $\sum_{j'=-\infty}^{\infty} (j'+\mu)^l a_{j'}$ appear repeatedly in (12.52) and (12.53). According to (9.13), for a scheme of mth-order accuracy these sums are equal to zero for l in the range 1 to m. As the order of accuracy of the scheme increases, the real and imaginary parts of the relative error both decrease, as more and more terms of (12.52) and (12.53) drop out. *But the real and imaginary parts take turns*. Moving from an odd-order scheme to an even-order scheme will reduce ε_R but not ε_I, and moving from an even-order scheme to an odd-order schemes will reduce ε_I but not ε_R.

As an example, suppose that we have a first-order scheme, such as the upstream scheme. Then the leading term of ε_R is proportional to $(k\Delta x)^2$, but the leading term of ε_I is proportional to $k\Delta x$. When we go to a second-order scheme, the leading term of ε_R is proportional to $(k\Delta x)^4$, but the leading term of ε_R does not change. If we go to a third-order scheme, the leading term of ε_I is proportional to $(k\Delta x)^3$, but the leading term of ε_R does not change.

We conclude that *the amplitude error decreases as we go from an odd order to the next even order, while the phase error decreases as we go from an even order to the next odd order*. This makes odd-order schemes attractive. For example, a third-order scheme has about the same amplitude error as a second-order scheme, but with a smaller phase error. The solution of the third-order scheme will be less noisy than the solution of the second-order scheme.

12.5 Segue

Computational dispersion arises from space differencing. It causes waves of different wavelengths to move at different speeds. In some cases, the phase speed can be zero or even negative, when it should be positive. Short waves generally move slower than longer waves. The effective wind speeds of the long waves are well simulated by the commonly used space-time differencing schemes. The group speed, which is the rate at which a wave "envelope" moves, can also be adversely affected by space truncation errors.

Higher-order schemes simulate the well resolved modes more accurately, but do not improve the solution for the short modes (e.g., the $2\Delta x$ modes), and can actually make the problems with the short modes worse, in some ways. Of course, higher-order schemes involve more arithmetic and so are computationally more expensive than lower-order schemes. An alternative is to use a lower-order scheme with more grid points. This may be preferable in many cases.

Odd-order, space-uncentered schemes are well suited to advection, which is a spatially asymmetric process, and they can minimize the effects of computational dispersion.

12.6 Problems

1. Consider the first-order upstream scheme, the second-order Lax-Wendroff scheme, and the third-order Takacs scheme. For each of these, compute the amplitude error, $\varepsilon_{|\lambda|}$, as given by (12.43), and the phase error, ε_φ, as given by (12.45). Contour plot your results as functions of μ and $k\Delta x$, for $0 \leq \mu \leq 1$ and $0 \leq k\Delta x \leq \pi$.

2. Prove that the phase error ε_φ, defined by (12.44), satisfies $\varepsilon_\varphi = \tan^{-1}\left(\dfrac{\varepsilon_I}{\varepsilon_R + 1}\right)$.

Chapter 13

Modern Eulerian advection schemes

13.1 Sign-preservation and monotonicity

We often wish to require that a non-negative variable, such as the water vapor mixing ratio, remains non-negative under advection. An advection scheme that has this property is often called "positive-definite" or, more generally, "sign-preserving." Sign-preserving schemes are obviously desirable, because negative values that arise through truncation errors the advection scheme will have to be eliminated somehow before any moist physics can be considered, and the methods used to eliminate the negative values are inevitably somewhat artificial(e.g., Williamson and Rasch, 1994). As we will see, most of the older advection schemes do not come anywhere near satisfying this requirement. Many newer schemes do satisfy it, however.

As seen in earlier examples, computational dispersion can cause new maxima and minima to develop as advection proceeds, and it can also cause the advected field to change sign, e.g., from positive to negative.

As discussed earlier, the stability condition for the upstream scheme is $|\mu| \leq 1$. When this condition is met, the "interpolation" form of (12.39) guarantees that

$$\text{Min}\left\{A_j^n,\, A_{j-1}^n\right\} \leq A_j^{n+1} \leq \text{Max}\left\{A_j^n,\, A_{j-1}^n\right\}. \tag{13.1}$$

This means that A_j^{n+1} cannot be smaller than the smallest value of A_j^n in the neighborhood of point j, or larger than the largest value of A_j^n in the neighborhood of point j. In other words, the upstream scheme does not produce any new maxima or minima, like those associated with the dispersive ripples seen in Fig. 12.8. Schemes with this property are called "*monotonic*." As discussed earlier, real advection is monotonic. The property of monotonicity is expressed by (13.1).

Real advection cannot produce negative values of A if none are present initially, and neither can the upstream scheme, provided that $0 \leq \mu \leq 1$. This means that *the upstream scheme is a sign-preserving scheme* when the stability criterion is satisfied. This is very useful if the advected quantity is intrinsically non-negative, e.g., the mixing ratio of some trace species. Even better, *the upstream scheme is a monotone scheme* when the stability criterion is satisfied.

All monotone schemes are sign-preserving schemes. The converse is not true.

sign-preservation is relative to both the continuity equation (no negative mass) and the tracer equation (no negative tracers). This is not true of monotonicity. Recall from (7.9) that the density of a compressible fluid is not constrained to be constant following a particle. *For this reason, new maxima and minima can appear in the density field, and so monotonicity is physically wrong for the continuity equation*.

13.2 Sign-preservation with fields that have both signs

In the preceding discussion, we assumed that A_j^0 is everywhere of one sign, but this assumption is not really necessary. For variable-sign fields, a similar result can be obtained by decomposing A into positive and negative parts, i.e.,

$$A = A^+ + A^- \,. \tag{13.2}$$

The idea is that A^+ is positive where A is positive, and zero elsewhere; and similarly that A^- is negative where A is negative, and zero elsewhere. The total of A is then the sum of the two parts, as stated in (13.2). Advection of A is equivalent to advection of A^+ and A^- *separately*. If we apply a sign-preserving scheme to each part, then each of these two advections is stable by the argument given above, and so the advection of A itself is also stable.

13.3 Help from the geometric and harmonic means

Although the upstream scheme is sign-preserving, it is only first-order accurate and strongly damps, as we have seen. Are there more accurate schemes that are sign-preserving or nearly so? A spurious negative value is customarily called a "hole." Second-order advection schemes that produce relatively few holes are given by (11.10) with either the geometric mean given by (11.41), or the harmonic mean given by (11.42). Both of these schemes have the property that $A_{j+\frac{1}{2}}$ also goes to zero when either A_j or A_{j+1} goes to zero. If the time step were infinitesimal, this would be enough to prevent the property denoted by A from changing sign. Because time-steps are finite in real models, however, such schemes

do not completely prevent hole production. Nevertheless, they do tend to reduce the rate of hole production.

13.4 Fixing a hole

If a non-sign-preserving advection scheme is used, and holes are produced, then a procedure is needed to fill the holes. To make the discussion concrete, we consider here a scheme to fill "water holes," in a model that advects water vapor mixing ratio.

Simply replacing negative mixing ratios by zero is unacceptable because it leads to a systematic increase in the mass-weighted total water. Hole-filling schemes therefore "borrow" mass from elsewhere on the grid. They take from the rich, and give to the poor.

There are many possible borrowing schemes. Some borrow systematically from nearby points, but of course borrowing is only possible from neighbors with positive mixing ratios, and it can happen that the nearest neighbors of a "holey" grid cell don't have enough water to fill the hole. Logic can be invented to deal with such issues, but hole-fillers of this type tend to be complicated and computationally slow.

An alternative is to borrow from *all* points on the mesh that have positive mixing ratios. The "global multiplicative hole-filler" uses that approach, and is a particularly simple and computationally fast algorithm. The first step is to add up all of the positive water on the mesh:

$$P \equiv \sum_{\text{where } A_j \geq 0} m_j A_j \geq 0 \,. \tag{13.3}$$

Here A_j is the mixing ratio in grid cell j, and m_j is the mass of dry air in that grid cell (in kg, say), so that the product $m_j A_j$ is the mass of water in the cell. Note that m_j is *not* the density of dry air in the cell; instead, it is the product of the density of dry air and the volume of the cell. The total amount of water on the mesh, including the contributions from "holes," is given by

$$T \equiv \sum_{\text{all points}} m_j A_j \,. \tag{13.4}$$

Both T and P have the dimensions of mass. Define the nondimensional ratio

$$\Phi \equiv \frac{T}{P} \leq 1 \,; \tag{13.5}$$

normally Φ is just very slightly less than one, because there are only a few holes and they are not very "deep." We replace all negative values of A_j by zero, and then set

$$A_j^{new} = \Phi A_j \, . \tag{13.6}$$

In this way, we are ensured of the following:

- No negative values of A_j remain on the mesh.
- The total mass of water in the adjusted state is equal to T, the same as the total in the "holey" state.
- Water is borrowed most heavily from grid cells with large mixing ratios, and least from cells with small mixing ratios.

This multiplicative hole fller does not have to be implemented using *global* sums for P and T; sums over sufficiently large subdomains can be used instead, and the subdomains can be corrected individually. This can be useful on parallel machines.

Hole-filling is ugly. Any hole-filling procedure is necessarily somewhat arbitrary, because in designing it we cannot mimic any natural process; nature does not fill holes, because it has no holes to fill.

In addition, hole-filling is "quasi-diffusive," because it removes water from wet cells and adds it to dry cells, thus reducing ("dissipating") the total variance of the mixing ratio. Hole filling tends to transport moisture upward, which can spuriously increase cloudiness and precipitation.

The best approach is to choose an advection scheme that does not make holes in the first place. At the very least, we should insist that an advection scheme digs holes slowly, so that the hole-filler will not have to work very hard.

13.5 Flux-corrected transport

The upstream scheme is monotone and sign-preserving, but, unfortunately, as we have seen, it is strongly damping. Damping is intrinsic to all monotone and sign-preserving schemes. Much work has been devoted to designing monotone or sign-preserving schemes that produce *as little damping as possible*. The following discussion, based on the paper of Zalesak (1979), explains how this can be done.

Monotone and sign-preserving schemes can be derived by using the approach of "flux-corrected transport," often abbreviated as FCT, which was invented by Boris and Book (1973) and extended by Zalesak (1979) and many others. Suppose that we have a "high-order" advection scheme, represented schematically by

$$A_j^{n+1} = A_j^n - \left(FH_{j+\frac{1}{2}} - FH_{j-\frac{1}{2}} \right) . \tag{13.7}$$

Here FH represents the "high-order" fluxes associated with the scheme. Note that (13.7) is in "conservation" form, and the time derivative is approximated using time levels n and $n+1$. Suppose that we have at our disposal a monotone or sign-preserving low-order scheme, whose fluxes are denoted by $FL_{j+\frac{1}{2}}$. This low-order scheme could be, for example, the upstream scheme. (From this point on we say "monotone" with the understanding that we mean "monotone or sign-preserving.") As a matter of definition, we can write

$$FH_{j+\frac{1}{2}} \equiv FL_{j+\frac{1}{2}} + FC_{j+\frac{1}{2}} . \tag{13.8}$$

Here $FC_{j+\frac{1}{2}}$ is a "corrective" flux, sometimes called an "anti-diffusive" flux. Eq. (13.8) is the definition of $FC_{j+\frac{1}{2}}$. According to (13.8), the high-order flux is the low-order flux plus a correction. We know that the low-order flux is diffusive in the sense that it damps the solution, but on the other hand by assumption the low-order flux corresponds to a monotone scheme. The high-order flux is presumably less diffusive, and more accurate, but does not have the nice monotone property that we want.

Suppose that we take a time-step using the low-order scheme. Let the result be denoted by A_j^{n+1*}, i.e.,

$$A_j^{n+1*} = A_j^n - \left(FL_{j+\frac{1}{2}} - FL_{j-\frac{1}{2}} \right) . \tag{13.9}$$

Since, by assumption, the low-order scheme is monotone, we know that

$$A_j^{\mathrm{MAX}} \geq A_j^{n+1*} \geq A_j^{\mathrm{MIN}} \text{ for all } j. \tag{13.10}$$

where A_j^{MAX} and A_j^{MIN} are, respectively, suitably chosen upper and lower bounds on the value of A within the grid-box in question. For instance, A_j^{MIN} might be zero, if A is a non-negative scalar like the mixing ratio of water vapor. Other possibilities will be discussed below.

There are two important points in connection with the inequalities in (13.10). First, the inequalities must actually be true for the low-order scheme that is being used. Second,

the inequalities should be strong enough to ensure that the solution obtained is in fact monotone.

From (13.7), (13.8), and Equation (13.9) it is easy to see that

$$A_j^{n+1} = A_j^{n+1*} - \left(FC_{j+\frac{1}{2}} - FC_{j-\frac{1}{2}} \right) . \tag{13.11}$$

This simply says that we can obtain the high-order solution from the low-order solution by adding the anti-diffusive fluxes. The anti-diffusive fluxes can be computed, given the forms of the low-order and high-order schemes.

We now define some coefficients, denoted by $C_{j+\frac{1}{2}}$, and "scaled-back" anti-diffusive fluxes, denoted by $\widehat{FC}_{j+\frac{1}{2}}$, such that

$$\widehat{FC}_{j+\frac{1}{2}} \equiv C_{j+\frac{1}{2}} FC_{j+\frac{1}{2}} . \tag{13.12}$$

In place of (13.11), we use

$$A_j^{n+1} = A_j^{n+1*} - \left(\widehat{FC}_{j+\frac{1}{2}} - \widehat{FC}_{j-\frac{1}{2}} \right) . \tag{13.13}$$

To see the idea, consider two limiting cases. If $C_{j+\frac{1}{2}} = 1$, then $\widehat{FC}_{j+\frac{1}{2}} = FC_{j+\frac{1}{2}}$, and so (13.13) will reduce to (13.11) and will simply give the high-order solution. If $C_{j+\frac{1}{2}} = 0$, then $\widehat{FC}_{j+\frac{1}{2}} = 0$, and so (13.13) will simply give the low-order solution. We enforce

$$0 \leq C_{j+\frac{1}{2}} \leq 1 \text{ for all } j , \tag{13.14}$$

and try to make $C_{j+\frac{1}{2}}$ as close to one as possible, so that we get as much of the high-order scheme as possible, and as little of the low-order scheme as possible, but we require that

$$A_j^{\text{MAX}} \geq A_j^{n+1} \geq A_j^{\text{MIN}} \text{ for all } j \tag{13.15}$$

be satisfied. Compare (13.15) with (13.10). We can always ensure that (13.15) will be satisfied by taking $C_{j+\frac{1}{2}} = 0$; this is the "worst case." Quite often it may happen that (13.15) is satisfied for $C_{j+\frac{1}{2}} = 1$; that is the "best case." The approach outlined above can be interpreted as a *nonlinear* interpolation for the value of A_j^{n+1}. This nonlinearity means that Godunov's theorem does not apply; see Chapter 11.

It remains to assign values to the $C_{j+\frac{1}{2}}$, which are called "limiters" because they limit the amount of the high-order scheme that will be added to the low-order scheme. Zalesak broke the problem down into parts. One obvious issue is that the flux into one cell is the flux out of another, so the value assigned to $C_{j+\frac{1}{2}}$ has to be sufficient to enforce (13.15) for both A_j^{n+1} and A_{j+1}^{n+1}. Set that issue aside, for now.

We can conceptually divide the anti-diffusive fluxes affecting cell j into the fluxes in and the fluxes out, simply based on sign. The anti-diffusive fluxes into cell j could cause A_j^{n+1} to exceed A_j^{MAX}, while the anti-diffusive fluxes out of cell j could cause A_j^{n+1} to fall below A_j^{MIN}. We adjust (all of) the anti-diffusive fluxes into the cell so as to avoid overshooting A_j^{MAX}, and we adjust (all of) the anti-diffusive fluxes out so as to avoid undershooting A_j^{MIN}. In this way, we obtain provisional values of $C_{i-\frac{1}{2}}$ and $C_{i+\frac{1}{2}}$, based on an analysis of cell j. Similarly, analysis of cell $j+1$ gives provisional values of $C_{j+\frac{1}{2}}$ and $C_{j+\frac{3}{2}}$, and analysis of cell $j-1$ gives provisional values of $C_{j-\frac{3}{2}}$ and $C_{j-\frac{1}{2}}$.

Note that we now have *two* provisional values for $C_{j+\frac{1}{2}}$. After looping over the whole grid, we will have two provisional Cs for each cell wall. To ensure that all constraints are simultaneously satisfied, we choose the smaller of the two provisional Cs for each cell wall. *Voila!* All constraints are satisfied.

The procedure outlined above is not unique. Many other approaches can be found in the literature.

It remains to choose the upper and lower bounds A_j^{MAX} and A_j^{MIN} that appear in (13.15) and (13.10). Zalesak (1979) proposed limiting A_j^{n+1} so that it is bounded by the largest and smallest values of its neighbors at time level n, and also by the largest and smallest values of the low-order solution at time level $n+1$. In other words, he took

$$A_j^{\text{MAX}} = \max\left\{A_{j-1}^n,\, A_j^n,\, A_{j+1}^n,\, A_{j-1}^{n+1*},\, A_j^{n+1*},\, A_{j+1}^{n+1*}\right\}, \tag{13.16}$$

and

$$A_j^{\text{MIN}} = \min\left\{A_{j-1}^n,\, A_j^n,\, A_{j+1}^n,\, A_{j-1}^{n+1*},\, A_j^{n+1*},\, A_{j+1}^{n+1*}\right\}. \tag{13.17}$$

This "limiter" is not unique. Many other possibilities are discussed in the literature.

Our analysis of FCT schemes has been given in terms of one spatial dimension, but all of the discussion given above can be extended to two or three dimensions, without time splitting. The literature on FCT schemes is very large.

FCT schemes are dissiptive, in the sense that they reduce the variance of the advected field. This numerical dissipation is the price paid for monotonicity.

The FCT method opened the door to a cottage industry of "shape-preserving" (i.e., monotone) advection schemes. The schemes have much in common, but are given different names, as discussed below.

13.6 MPDATA

Smolarkiewicz (1991) proposed an advection scheme that he called MPDATA (Multidimensional Positive Definite Advection Transport Algorithm). MPDATA is second-order accurate, sign-preserving, and conservative. It is designed for multi-dimensional applications. A monotonicity option exists. There is now a third-order accurate version (Waruszewski et al., 2018).

13.7 TVD schemes

An extremely broad class of schmemes can be described as "Total Variation Diminishing" (TVD; Harten, 1983). The total variation of a variable α can be measured as follows:

$$\mathrm{TV}^n \equiv \sum_{j=1}^{J} \left| \alpha_j^n - \alpha_{j-1}^n \right| . \tag{13.18}$$

A scheme is said to be TVD if it can be shown that the total variation does not increase:

$$\mathrm{TV}^{n+1} \leq \mathrm{TV}^n . \tag{13.19}$$

Clearly, TVD schemes are dissipative.

TVD schemes rely on flux limiters or slope limiters—mathematical functions that limit the contribution of higher-order corrections in regions where total variation might otherwise grow. In smooth regions, the limiter allows more higher-order accuracy. Near sharp changes, the limiter suppresses higher-order terms to prevent spurious oscillations,

reverting to a more diffusive but monotonic (stable) scheme. This adaptivity is key: TVD schemes automatically adjust behavior based on the local smoothness of the solution.

13.8 van Leer schemes

Bram van Leer of the University of Michigan proposed the MUSCL (Monotonic Upstream-centered Scheme for Conservation Laws) scheme (Van Leer, 1974, 1977a,b, 2008), which introduced slope limiters to achieve high-order accuracy while preserving monotonicity and avoiding oscillations near discontinuities. The MUSCL scheme replaces the piecewise constant representation in Godunov's first-order method with slope-limited linear reconstructions inside each cell to achieve higher order (second-order spatial accuracy). It reconstructs left and right states at each cell boundary using slope limiters that prevent spurious oscillations by adapting to the solution's smoothness. These reconstructed states are used to compute numerical fluxes at cell edges. MUSCL prevents the growth of new oscillations, ensuring physically realistic results near steep gradients. It can achieve third-order accuracy in special cases, though originally it was developed for second order.

13.9 The piecewise parabolic method

Phil Colella and colleagues developed the Piecewise Parabolic Method (PPM; Colella and Woodward, 1984). IT is a finite-volume method that represents the advected quantity within each grid cell by a piecewise parabolic (quadratic) function, rather than simple averages or linear functions. It interpolates data within each cell with a parabola constructed from cell averages in neighboring zones. The PPM achieves third-order accuracy in general and effectively fourth-order accuracy for small time steps on uniform grids. The parabolic interpolation allows for a sharp and accurate representation of discontinuities and steep gradients, like shocks and fronts, while minimizing numerical diffusion. It integrates the parabolic profiles over the domain to update cell averages, ensuring mass conservation. PPM uses limiters and reconstruction techniques to avoid spurious oscillations near discontinuities. With some limiters, the scheme is monotone.

13.10 Prather's scheme

Prather's scheme (Prather, 1986) is a high-order, positive-definite, and monotonic method designed to minimize numerical diffusion and prevent unphysical oscillations—goals shared by schemes like Leonard's, the Piecewise Parabolic Method (PPM), and TVD schemes. Prather's scheme employs polynomial reconstructions of tracer distributions within grid cells, ensuring mass conservation and maintaining the shape of transported scalar fields accurately over time. It was developed with an emphasis on sign-preservation and minimizing numerical diffusion, making it especially suitable for atmospheric chemical transport. Like Leonard's schemes (see below), Prather's approach uses polynomial reconstructions

and limiting procedures to achieve high-order accuracy while preserving physical realism. Compared to TVD schemes, Prather's method similarly limits spurious oscillations. With suitable limiters it can be monotone.

13.11 Leonard schemes

Leonard advection schemes (Leonard, 1993; Leonard et al., 1996) are a class of higher-order, monotone numerical advection methods developed by Brian Phillip ("Benny") Leonard of the University of Akron, and others. The schemes aim to achieve third-order accuracy while maintaining important properties such as monotonicity, positivity, and conservation, thereby limiting spurious oscillations near sharp gradients or discontinuities. Leonard schemes are based on a finite-volume framework in which higher-order polynomial reconstructions of the advected quantity are performed within each cell. Leonard introduced a universal limiter method and "ULTIMATE" schemes, which are one-step, single time-step, third-order flux-based, finite-volume advection schemes designed for multidimensional flow. Leonard schemes schemes carefully blend high-order accuracy with nonlinear limiters to ensure stability and reduce numerical diffusion without introducing unphysical oscillations. The schemes have been influential in developing monotone and positivity-preserving advection methods suitable for atmospheric and ocean modeling where tracer transport accuracy is critical. Leonard schemes often serve as the basis or inspiration for various modern high-order shape-preserving advection algorithms.

13.12 WENO schemes

WENO (Weighted Essentially Non-Oscillatory) schemes (Zhang and Shu, 2016; Shu, 2020) give high-order accuracy in smooth regions, and suppress numerical oscillations near discontinuities or sharp gradients. They can adaptively reduce order where necessary, minimizing unphysical behavior. WENO schemes can cause excessive numerical diffusion—especially when used with coarse resolution.

Chapter 14

Lagrangian and semi-Lagrangian advection schemes

14.1 Lagrangian schemes

Lagrangian schemes, in which particles are tracked through space without the use of an Eulerian grid, have been used in the atmospheric and oceanic sciences, as well as other fields including astrophysics and weapons engineering (e.g., Mesinger, 1971; Trease, 1988; Monaghan, 1992; Norris, 1996; Haertel and Randall, 2002). The Lagrangian approach has a number of attractive features:

- The PDF of the advected quantity (and all functions of the advected quantity) can be preserved "exactly" under advection. Here "exactly" has been put in quotation marks because, in practice, only a finite number of particles can be tracked.
- Lagrangian schemes are monotone and sign-preserving.
- Time steps can be arbitrarily long without triggering computational instability, although of course long time steps still lead to large discretization errors.

Alas, there are always trade-offs. Lagrangian schemes suffer from a number of practical difficulties. Some of these have to do with the possible development of "voids," i.e., regions with not enough particles to represent the physics. Others arise from the need to compute spatial derivatives (e.g., the pressure gradient force, which is needed to compute the acceleration of each particle) on the basis of a collection of particles that can be located literally anywhere within the domain, in an irregular and uncontrolled way.

14.1.1 Smoothed particle hydrodynamics

One class of Lagrangian schemes, called "smoothed particle hydrodynamics" (SPH), has been widely used by the astrophysical research community, and is reviewed by Monaghan (1992). The idea is to invent a way to compute a given field at all points throughout the domain, given the values of the field at a finite number of arbitrarily located points that are

occupied by particles. Once such a continuous field has been defined everywhere, it can be differentiated to obtain, e.g., the pressure-gradient force.

For an arbitrary field A, let $A(\mathbf{r})$ be given by a volume integral of the product of A and a differentiable weighting function W:

$$A(\mathbf{r}) = \int A(\mathbf{r}') W\left(\left|\mathbf{r}-\mathbf{r}'\right|, h\right) dV(\mathbf{r}') . \tag{14.1}$$

Here the integration is over the whole domain (e.g., the whole atmosphere), and $W(|\mathbf{r}-\mathbf{r}'|, h)$ is a differentiable interpolating "kernel" such that

$$\int W\left(\left|\mathbf{r}-\mathbf{r}'\right|, h\right) dV(\mathbf{r}') = 1 , \tag{14.2}$$

Note that W has dimensions of an inverse volume. The maximum of W occurs where $|\mathbf{r}-\mathbf{r}'| = 0$, and $W \to 0$ as $|\mathbf{r}-\mathbf{r}'| \to \infty$ In (14.1) – (14.2), h is a parameter that measures the "width" of W. With smaller h, W is more strongly peaked. With these definitions, if $A(\mathbf{r}')$ is a constant field, then $A(\mathbf{r})$ is given by the same constant.

In a discrete model, we replace (14.1) by

$$A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} W\left(\left|\mathbf{r}-\mathbf{r}'\right|, h\right) . \tag{14.3}$$

Here the index b denotes a particular particle, m_b is the mass of the particle, and ρ_b is the density of the particle. Note that m_b/ρ_b is a volume. To see what is going on in (14.3), consider the case $A \equiv \rho$. Then (14.3) reduces to

$$\rho(\mathbf{r}) = \sum_b m_b W\left(\left|\mathbf{r}-\mathbf{r}'\right|, h\right) , \tag{14.4}$$

which simply says that the density at a point $\mathbf{r}$ is a weighted sum of the masses of particles in the vicinity of $\mathbf{r}$. In case there are no particles near the point $\mathbf{r}$, the density there will be small.

We can now perform spatial differentiation simply by taking the appropriate derivatives of $W(\mathbf{r}-\mathbf{r}_b, h)$, e.g.,

$$\nabla A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} \nabla W\left(\left|\mathbf{r} - \mathbf{r}'\right|, h\right). \tag{14.5}$$

This follows because m_b, A_b and ρ_b are associated with particular particles and are, therefore, not functions of space.

Further discussion of SPH and related methods is given by Monaghan (1992) and the other references cited above.

14.1.2 Slippery sacks

A very different approach has been developed by Patrick Haertel and colleagues. They consider flexible "parcels" of fixed mass, which fill the space of the fluid, something like conforming water balloons. Haertel and Randall (2002) named these parcels "slippery sacks." The moving parcels exchange momentum and energy through the pressure force, by literally pushing on each other along their shared boundaries, like nursing kittens crawling over a mother cat. Diffusive exchanges can be parameterized in a straightforward way, in terms of the differences of the properties of neighboring parcels. The first applications of the slippery sacks method were to lakes, and then to ocean basins (Haertel and Randall, 2002; Haertel et al., 2004; Van Roekel et al., 2009; Haertel et al., 2009; Haertel and Fedorov, 2012), taking advantage of the incompressibility of water, which implies that parcels of fixed mass have fixed volumes. This work culminated in a global ocean model (Haertel, 2019). A major advantage of the method for ocean modeling is that simulated parcels can move for hundreds of years in the deep ocean circulation, with no computational diffusion whatsoever.

Haertel and colleagues later allowed the sacks to change their volumes, which made it possible to construct a global atmosphere model (Haertel and Straub, 2010; Haertel et al., 2014, 2015, 2017). The global atmosphere and ocean models have been coupled, and used to simulate climate change (Haertel and Liang, 2024). This fully Lagrangian global modeling system is truly unique, and it will be interesting to see how it evolves in the future.

14.2 Semi-Lagrangian schemes

14.2.1 Further upstream

"Semi-Lagrangian" schemes (e.g., Robert et al., 1985; Staniforth and Côté, 1991; Bates et al., 1993; Smith, 2000; Diamantakis, 2013) are of interest in part because they allow very long time steps, and also because they can easily maintain such properties as monotonicity.

The basic idea is very simple. At time step $n+1$, values of the advected field, at the various grid points, are considered to be characteristic of the particles that reside at those

points. We ask where those particles were at time step n. This question can be answered by using the (known) velocity field, averaged over the time interval $(n, n+1)$, to track the particles backward in time from their locations at the various specified grid points, at time level $n+1$, to their "departure points" at time level n. Naturally, the departure points are usually located in between grid cells. The values of the advected field at the departure points, at time level n, can be determined by spatial interpolation. If advection is the only process occurring, then the values of the advected field at the departure points at time level n will be identical to those at the grid points at time level $n+1$.

As a simple example, consider one-dimensional advection of a variable q by a constant current, u. A particle that resides at $x = x_j$ at time level $t = t^{n+1}$ has a departure point given by

$$(x_d)_j^n = x_j - u\Delta t \ . \tag{14.6}$$

Here the superscript n is used to indicate that the departure point is the location of the particle at time level n. Suppose that $u > 0$, and that the departure point is less than one Δx away from x_j, so that

$$x_{j-1} < (x_d)_j^n < x_j \ . \tag{14.7}$$

Then the simplest linear interpolation for A at the departure point is

$$\begin{aligned}
(A_d)_j^n &= A_{j-1}^n + \left[\frac{(x_d)_j^n - x_{j-1}}{\Delta x}\right]\left(A_j^n - A_{j-1}^n\right) \\
&= A_{j-1}^n + \left(\frac{\Delta x - u\Delta t}{\Delta x}\right)\left(A_j^n - A_{j-1}^n\right) \\
&= A_{j-1}^n + (1-\mu)\left(A_j^n - A_{j-1}^n\right) \\
&= \mu A_{j-1}^n + (1-\mu) A_j^n \ .
\end{aligned} \tag{14.8}$$

Here we assume for simplicity that the mesh is spatially uniform, and $\mu \equiv u\Delta t/\Delta x$, as usual. The semi-Lagrangian scheme uses

$$A_j^{n+1} = (A_d)_j^n \ , \tag{14.9}$$

so we find that

$$A_j^{n+1} = \mu A_{j-1}^n + (1-\mu) A_j^n \, . \tag{14.10}$$

This is our friend, the upstream scheme. Note that (14.7), which was used in setting up the spatial interpolation, is equivalent to

$$0 < \mu < 1 \, . \tag{14.11}$$

As shown earlier, this is the condition for stability of the upstream scheme.

What if (14.7) is not satisfied? This will be the case if the particle is moving quickly and/or the time step is long or, in other words, if $\mu > 1$. Then we might have, for example,

$$x_{j-a} < (x_d)_j^n < x_{j-a+1} \, . \tag{14.12}$$

where a is an *integer* greater than 1. For this case, we find in place of (14.8) that

$$(A_d)_j^n = \hat{\mu} A_{j-a}^n + (1-\hat{\mu}) A_{j-a+1}^n \, . \tag{14.13}$$

where

$$\hat{\mu} \equiv 1 - a + \mu \, . \tag{14.14}$$

We have assumed again here, for simplicity, that both the mesh and the advecting current are spatially uniform. It should be clear that

$$0 < \hat{\mu} < 1 \, . \tag{14.15}$$

For $a = 1$, $\mu = \hat{\mu}$. Eq. (14.9) gives

$$A_j^{n+1} = \hat{\mu} A_{j-a}^n + (1-\hat{\mu}) A_{j-a+1}^n \ . \tag{14.16}$$

This has the form of an interpolation, so we still have computational stability and monotonicity (and sign-preservation); the semi-Lagrangian scheme is computationally stable regardless of the size of the time step. This means that the only limit on the time step is that it has to be short enough to temporally resolve what we are trying to simulate.

It is clear that the semi-Lagrangian scheme outlined above is very diffusive, because it is more or less equivalent to a "generalized upstream scheme." By using higher-order interpolations (e.g., cubic interpolations), the strength of this computational diffusion can be reduced, but it cannot be eliminated completely.

14.2.2 More accurate semi-Lagrangian schemes

We now consider more accurate semi-Lagrangian schemes, still in one spatial dimension. The issues to be addressed are:

- How to find the departure point when the wind varies in space and time;
- How to interpolate the advected field to the departure point;
- How to account for sources and sinks that act along the particle's path from the departure point.

To find the departure point, we use

$$\frac{Dx}{Dt} = \int_t^{t+\Delta t} u\left[x(t),t\right]\, dt \ . \tag{14.17}$$

which we approximate by

$$x - x_d = \Delta t\ u\left(\frac{x+x_d}{2}, \frac{t+\Delta t}{2}\right) \ . \tag{14.18}$$

Here $x = x_d$ is the departure point. Next, we use

$$
\begin{aligned}
u\left(t+\frac{\Delta t}{2}\right) &\cong u(t)+\frac{\Delta t}{2}\frac{\partial u}{\partial t} \\
&\cong u(t)+\frac{\Delta t}{2}\left[\frac{u(t)-u(t-\Delta t)}{\Delta t}\right] \\
&\cong \frac{3}{2}u(t)-\frac{1}{2}u(t-\Delta t)\ .
\end{aligned}
\tag{14.19}
$$

These values are evaluated at the u points of the model's grid. As a first guess, we use $x_d^l = x - \Delta t\ u(t)$, where the superscript l is an iteration counter. A revised estimate of the departure point is then obtained from

$$
x_d^l = x - \Delta t\ u\left(\frac{x+x_d^{l-1}}{2}, \frac{t+\Delta t}{2}\right)\ , \tag{14.20}
$$

and spatial interpolation is used on the right-hand side. There is no need to run the iteration to full convergence; typically two passes suffice.

Once the departure point has been found, the next step is to spatially interpolate the advected quantity to the departure point, at time-level n. The simplest approach is to use a linear interpolation based on the two nearest grid points. This has the advantage of being sign-preserving. A more accurate (and widely used) interpolation can be based on a cubic polynomial designed to "pass through" the values of the advected quantity at the first two grid points on either side of the departure point.

14.2.3 Remapping schemes

Can semi-Langrangian schemes be conservative? To prove that a scheme is conservative, it suffices to show that it can be written in a "flux form." Note, however, that in deriving the semi-Lagrangian scheme we have used the Lagrangian form of the advection equation very directly, by considering the parcel trajectory between the mesh point at time level $n+1$ and the departure point at time level n. Because the Lagrangian form is used in their derivations, many semi-Lagrangian schemes are not conservative.

An exception is a class of conservative semi-Lagrangian schemes based on "remapping" (Dukowicz and Baumgardner, 2000; Cotter et al., 2007; Reich, 2007). One such remapping scheme is CSLAM (Lauritzen et al., 2010; Dubey et al., 2014; Lauritzen et al., 2017), which stands for "Conservative Semi-Lagrangian Multitracer."

Remapping schemes replace the departure and arrival *points* by finite volumes. Figure 14.1 illustrates the basic idea. As a result of advection, at time $n+1$ the grid cell A_k is filled

with air that at time n occupied the distorted upstream region shown with dark shading. A portion of the upstream region overlapped, at time n, with grid cell A_l. The vertices of A_k are mapped back to departure points of the upstream region, as shown. The upstream regions fill the domain. With this approach, each grid cell at time $n+1$ is filled with a mixture of air from other grid cells at time n. Because the same air fills the grid at both time levels, conservation is ensured.

Finding the upstream regions is complicated and expensive, but it only has to be done once on each time step, no matter how many species are being advected. The scheme is therefore best suited to applications in which a large number of quantities are advected, as in a model with complex chemistry.

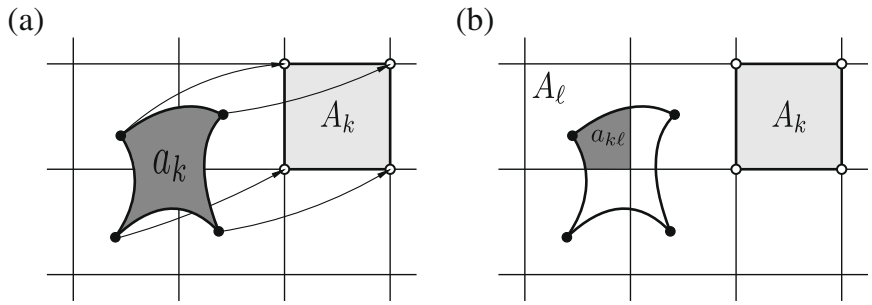


Figure 14.1: Schematic depiction of a remapping scheme. In panel (a), the deformed "departure cell" (dark shaded area) ends up, after being transported by the flow for one time step, at the arrival cell labeled A_k (light shaded area). The trajectories for the cell vertices are shown with arrows, and the departure and arrival cell vertices are marked with filled and open circles, respectively. Panel (b) illustrates the overlap region between the grid cell A_l and the departure grid cell. From Lauritzen et al. (2010).

Chapter 15

Just relax

15.1 Introduction

Systems of linear equations frequently arise in atmospheric science. They can involve simultaneous solution for thousands or even millions of unknowns. Systems of linear equations can be solved by a wide variety of methods, which are discussed in standard texts on numerical analysis. A method to solve such a system is called "solver."

There are many methods for solving linear systems. The issue is not finding a solution. It is finding a solution *fast*. We want to minimize the amount of computational work that must be done to obtain the solution, while at the same time minimizing the amount of storage required.

One source of linear systems is boundary-value problems. These involve spatial derivatives and/or integrals, but no time derivatives and/or integrals. Boundary-value problems can and do frequently arise in one, two, or three dimensions, in the atmospheric sciences. Three important examples are the Laplace equation

$$\nabla^2 f = 0 \,, \tag{15.1}$$

which is discussed in Appendix D, the Poisson equation

$$\nabla^2 f = g \,, \tag{15.2}$$

where g is a known function of the spatial coordinates, and the Helmholtz equation

$$\nabla^2 f = -k^2 f \,, \tag{15.3}$$

which arises in the study of waves. In all three of these equations, the function to be solved for is called f. All three of the equations must be supplemented by boundary conditions. In this chapter we will focus on the Poisson equation, but the methods discussed can also be used with the Laplace equation and the Helmholtz equation.

15.2 The Poisson equation

Here is an example of a physical problem that gives rise to a Poisson equation: Consider a two-dimensional flow with velocity vector $\mathbf{v}$. Let ζ and δ be the vorticity and divergence, respectively. We can define a stream function, ψ, and a velocity potential, χ, by

$$\mathbf{v}_r = \mathbf{k} \times \nabla \psi, \tag{15.4}$$

and

$$\mathbf{v}_d = \nabla \chi, \tag{15.5}$$

respectively. Here $\mathbf{k}$ is the unit vector perpendicular to the plane of the motion, and $\mathbf{v}_r$ and $\mathbf{v}_d$ are the rotational and divergent parts of the wind vector, respectively, so that

$$\mathbf{v} = \mathbf{v}_r + \mathbf{v}_d. \tag{15.6}$$

The vorticity and divergence then satisfy

$$\zeta = \nabla^2 \psi. \tag{15.7}$$

and

$$\delta = \nabla^2 \chi, \tag{15.8}$$

respectively.

Suppose that we are given the distributions of ζ and δ, and want to determine the wind vector. This can be done by first solving the two Poisson equations represented by (15.7) - (15.8), with suitable boundary conditions, then using (15.4) - (15.5) to obtain $\mathbf{v}_r$ and $\mathbf{v}_d$, and finally using Eq. (15.6) to obtain the total horizontal wind vector.

A second example is the solution of the anelastic pressure equation, in which the pressure field takes whatever shape is needed to prevent the divergence of the mass flux vector from becoming non-zero. This will be discussed further in a later chapter.

Further examples arise from implicit time-differencing combined with space-differencing, e.g., for the diffusion equation (Chapter 16) or the shallow-water equations (Chapter 19).

15.3 A continuous one-dimensional boundary-value problem

As a simple one-dimensional example, consider

$$\frac{d^2 q(x)}{dx^2} = f(x) \ , \tag{15.9}$$

on a periodic domain. Here $f(x)$ is a *given* periodic function of x. Eq. (15.9) is a Poisson equation, in which the Laplacian (in this case $\partial^2/\partial x^2$) of the unknown function is equal to a given function. Solution of (15.9) requires two boundary conditions. One of these can be periodicity. The second boundary condition can be specification of the domain-averaged value of q.

In some problems it is physically appropriate to specify the value of q on the boundary. This is called a Dirichlet boundary condition. In other cases it is physically appropriate to specify the normal derivative of q on the boundary. This is called a Neumann[1] boundary condition. Dirichlet boundary conditions constrain the solution more strongly than Neumann boundary conditions.

We are using the Poisson equation as an example here. The methods discussed in the remainder of this chapter are also applicable to other problems.

[1]Named after Carl Neumann (`https://en.wikipedia.org/wiki/Carl_Neumann`) rather than John von Neumann.

15.4 Fourier methods

Fourier methods may or may not be useful for solving the Poisson equation, depending on the geometry of the domain. For example, the exact solution of (15.9) can be obtained by expanding $q(x)$ and $f(x)$ in an infinite Fourier series. The individual Fourier modes will satisfy

$$-k^2 \widehat{q}_k = \widehat{f}_k \,, \tag{15.10}$$

which can readily be solved for the $\widehat{q}_k$, provided that the wave number k is not zero. The value of $\widehat{q}_0$, i.e., the domain average of q, must be obtained directly from the second boundary condition mentioned above. The full solution for $q(x)$ can be obtained by Fourier-summing the $\widehat{q}_k$.

This method to find the exact solution of (15.9) can be adapted to obtain an approximate numerical solution, simply by truncating the expansions of $q(x)$ and $f(x)$ after a finite number of modes. This is an example of what is called the "spectral" method. Like everything else, the spectral method has both strengths and weaknesses. It is discussed Chapter 29.

15.5 Finite-difference methods to solve the Poisson equation

Suppose, however, that the problem posed by (15.9) arises in a large numerical model, in which the functions $q(x)$ and $f(x)$ appear in many complicated equations, perhaps including time-dependent partial differential equations that are solved (approximately) through the use of spatial and temporal discretization. In that case, the need for consistency with the other equations of the model may dictate that the second derivative in (15.9) be approximated by a finite-difference method, such as

$$\frac{q_{j+1} - 2q_j + q_{j-1}}{d^2} = f_j \,. \tag{15.11}$$

Here d is the grid spacing. In this example, we have used centered second-order spatial differences.

15.6 A Fourier method for solving a finite-difference equation

Assuming that

$$q_j = \widehat{q}_k e^{ikjd} \quad \text{and} \quad f_j = \widehat{f}_k e^{ikjd} \,, \tag{15.12}$$

we can write

$$\begin{aligned}
\frac{q_{j+1} - 2q_j + q_{j-1}}{d^2} &= \frac{\widehat{q}_k}{d^2}\left[e^{ik(j+1)d} - 2e^{ikjd} + e^{ik(j-1)d}\right] \\
&= \frac{\widehat{q}_k e^{ikjd}}{d^2}\left(e^{ikd} - 2 + e^{-ikd}\right) \\
&= \frac{\widehat{q}_k e^{ikjd}}{d^2}\left[\cos(kd) + i\sin(kd) - 2 + \cos(-kd) + i\sin(-kd)\right] \\
&= \frac{\widehat{q}_k e^{ikjd}}{d^2}\left[2\cos(kd) - 2\right] \\
&= -\frac{\widehat{q}_k e^{ikjd}}{d^2} 4\sin^2(kd/2) \\
&= -k^2 \widehat{q}_k e^{ikjd}\left[\frac{\sin(kd/2)}{kd/2}\right]^2 \\
&= -k^2 \widehat{q}_k e^{ikjd} \mathrm{sinc}^2(kd/2) \,.
\end{aligned} \tag{15.13}$$

This leads to

$$-k^2 \widehat{q}_k \mathrm{sinc}^2(kd/2) = \widehat{f}_k \,. \tag{15.14}$$

Note the similarity between (15.14) and (15.10). Clearly (15.14) can be solved to obtain each of the $\widehat{q}_k$, except $\widehat{q}_0$, *and the result will be consistent with the finite-difference approximation* (15.11). For each k, the factor of $-k^2\mathrm{sinc}^2(kd/2)$ in (15.14) can be evaluated once and stored for use later in the simulation. This is advantageous if (15.14) must be solved on each of many time steps. This example illustrates that Fourier methods can be used even in combination with finite-difference approximations.

The Fourier method outlined above can produce solutions quickly, because of the existence of fast algorithms for computing Fourier transforms (not discussed here, but readily available in various scientific subroutine libraries). It is easy to see that, with suitable geometry, the method can be extended to two or three dimensions.

The Fourier method is not applicable when the problem involves spatially variable coefficients, or when the grid is nonuniform, or when the geometry of the problem is not compatible with Fourier expansion.

Further discussion is given in Chapter 29.

15.7 Solving linear systems

There are other ways to solve (15.11). It can be regarded as a system of linear equations, in which the unknowns are the q_i. The matrix of coefficients for this particular problem turns out to be "tri-diagonal." This means that the only non-zero elements of the matrix are the diagonal elements and those directly above and below the diagonal, as in the simple 6×6 example shown below:

$$\begin{bmatrix} d_1 & a_2 & 0 & 0 & 0 & b_6 \\ b_1 & d_2 & a_3 & 0 & 0 & 0 \\ 0 & b_2 & d_3 & a_4 & 0 & 0 \\ 0 & 0 & b_3 & d_4 & a_5 & 0 \\ 0 & 0 & 0 & b_4 & d_5 & a_6 \\ a_1 & 0 & 0 & 0 & b_5 & d_6 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{bmatrix}. \tag{15.15}$$

Here each element of the 6×6 matrix is labeled with a single subscript, denoting its column number. The names "d," "a," and "b" denote "diagonal," "above-diagonal," and "below-diagonal" elements, respectively. Note that a_1 and b_6 appear in the lower left and upper right corners of the matrix, respectively, because of the periodic boundary conditions.

The solution of tri-diagonal linear systems is very fast and easy. For instance, the first of the six equations represented by (15.15) can be solved for q_1 as a function of q_2 and q_6, provided that $d_1 \neq 0$. This can be used to eliminate q_1 in the five remaining equations. The (modified version of the) second equation can then be solved for q_2 as a function of q_3 and q_6, and this solution can be used to eliminate q_2 from the remaining four equations. Continuing in this way, we can ultimately obtain a single equation for the single unknown q_6. Once the value of q_6 has been determined, the other unknowns can be obtained by back-substitution. In case $d_1 = 0$ (assumed *not* to be true in the preceding discussion), we can immediately solve the first equation for q_2 in terms of q_6, provided that a_2 is not also equal to zero. And so on.

The issue of "scaling" deals with the *rate of change* in the amount of work needed to solve the system as the problem size increases. The best we can hope for is that the amount of work is simply proportional to the number of unknowns. It should be clear that this is the case for the tri-diagonal solver described above. This means that the tri-diagonal solver scales very well. Highly optimized tri-diagonal solvers are available in standard software libraries. Because tri-diagonal systems are easy to deal with, it is good news

when a problem can be expressed in terms of a tri-diagonal system. Naturally, tri-diagonal methods are not applicable when the matrix is not tri-diagonal.

We could, of course, solve the linear system by other methods that are discussed in introductory texts, such as Cramer's Rule or matrix inversion or Gaussian elimination. These "classical" methods work, but they scale badly compared to the Fourier and tri-diagonal methods discussed above. For each of the classical methods, the amount of arithmetic needed to find the solution is proportional to the *square* of the number of unknowns. If the number of unknowns is large, the methods are prohibitively expensive.

15.8 Simple relaxation methods

A better approach is solve (15.11) by a *relaxation method*. Relaxation methods are iterative, i.e., they start with an initial guess for the solution, and obtain successively better approximations to the solution by repeatedly executing a sequence of steps. Each pass through the sequence of steps is called a "*sweep*." Relaxation methods were invented during the 1940s (e.g., Southwell, 1940, 1946; Allen, 1954), and are now very widely used (e.g., Strang, 2007). They are of interest because they are very flexible and they scale better than the classical methods mentioned above. Several relaxation methods are discussed below.

15.8.1 Jacobi relaxation

Starting from this point, most of the discussion in this chapter is a condensed version of a portion of the paper by Fulton et al. (1986).

Consider the boundary-value problem

$$\begin{aligned} -\nabla^2 q &= f \text{ on a two-dimensional domain, and} \\ q &= g \text{ on the boundary of the domain ,} \end{aligned} \tag{15.16}$$

where f and g are known. Here we are using Dirichlet boundary conditions. We approximate (15.16) on a cartesian grid with uniform spacing d, and N grid points in each direction, such that $Nd = 1$, i.e., the total width of the domain in each direction is unity. Using second-order centered differences, we write:

$$\begin{aligned} d^{-2}\left(4q_{j,k} - q_{j-1,k} - q_{j+1,k} - q_{j,k-1} - q_{j,k+1}\right) &= f_{j,k} \text{ for } 0 < (j,\, k) < N, \\ q_{j,k} &= g_{j,k}, \text{ for } j = 0,\, N \quad \text{and} \quad k = 0,\, N\,. \end{aligned} \tag{15.17}$$

In order to explore relaxation methods for the solution of (15.17), we need a notation that allows us to distinguish approximate solutions from exact solutions. Here by "exact" solution I mean an exact solution to the *finite-difference problem* posed in (15.17). I will use the notation $\widehat{q}_{j,\,k}$ to denote an approximation to $q_{j,\,k}$.

The simplest relaxation method is called Jacobi relaxation or simultaneous relaxation. The Jacobi method defines the new value $\widehat{q}^{\,\text{new}}_{j,\,k}$ by applying (15.17) with the new value at the point $(j,\ k)$ and the "old" values at the neighboring points, i.e.,

$$d^{-2}\left(4\widehat{q}^{\,\text{new}}_{j,\,k} - \widehat{q}_{j-1,\,k} - \widehat{q}_{j+1,\,k} - \widehat{q}_{j,\,k-1} - \widehat{q}_{j,\,k+1}\right) = f_{j,\,k}\ , \tag{15.18}$$

so that

$$\widehat{q}^{\,\text{new}}_{j,\,k} = \frac{1}{4}\left(d^2 f_{j,\,k} + \widehat{q}_{j-1,\,k} + \widehat{q}_{j+1,\,k} + \widehat{q}_{j,\,k-1} + \widehat{q}_{j,\,k+1}\right)\ . \tag{15.19}$$

With this approach, we compute $\widehat{q}^{\,\text{new}}_{j,\,k}$ at all interior points using (15.18), and then replace the "old" approximate solution by the new one. This procedure is repeated until convergence is deemed adequate, but of course we have to ask whether or not convergence will actually occur, and if so how rapidly. Conditions for convergence and factors that influence the speed of convergence are discussed below.

Jacobi relaxation is well suited to parallelization, because we do exactly the same thing at every grid point.

Suppose that your first guess is that $q_{j,\,k}$ is uniform across the entire grid. Then, on the first sweep, the four values of $\widehat{q}$ on the right-hand side of (15.19) will all be the same number, and those four terms alone will try to make $q^{\text{new}}_{j,\,k}$ the same number again. It is true that the $d^2 f_{j,\,k}$ term prevents this, but its effect is usually small in a single sweep, because $d \ll 1$. As a result, it can take many sweeps for the iteration to converge. The finer the grid, the smaller the value of $d^2 f_{j,\,k}$, and the more sweeps are needed. Later, we will return to this simple but important point.

Let the error of a given approximation be denoted by[2]

$$\varepsilon_{j,\,k} \equiv \widehat{q}_{j,\,k} - q_{j,\,k}\ . \tag{15.20}$$

[2]Caution: In contrast to the definition used here, Fulton et al. (1986) defines the error as the exact solution minus the approximate solution; see the text above his equation (2.9). In other words, our error is minus his error.

Here again $q_{j,k}$ is the exact solution of the finite-difference system. If we know the ε associated with our current value of $\widehat{q}_{j,k}$, then we can use (15.20) to calculate q. Using (15.20) to eliminate all values of $\widehat{q}$ in (15.19), we find that

$$\varepsilon^{\text{new}}_{j,k} + q_{j,k} = \frac{1}{4}\left[d^2 f_{j,k} + \left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1}\right) + \left(q_{j-1,k} + q_{j+1,k} + q_{j,k-1} + q_{j,k+1}\right)\right] . \tag{15.21}$$

Since the exact solution satisfies (15.17), we can simplify (15.21) to

$$\varepsilon^{\text{new}}_{j,k} = \frac{1}{4}\left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1}\right) . \tag{15.22}$$

Eq. (15.22) shows that the new error (after the sweep) is the average of the current errors (before the sweep) at the four surrounding points.

One problem can be identified immediately. Suppose that the error field consists of a checkerboard pattern of 1's and -1's. Suppose further that point $(j,\ k)$ has a "current" error of +1, i.e., $\varepsilon_{j,k} = 1$. For our assumed checkerboard error pattern, it follows that the errors at the neighboring points referenced on the right-hand side of (15.22) are all equal to -1. At the end of the sweep we will have $\varepsilon_{j,k} = -1$. Then, on the next iteration, the error will flip back to $\varepsilon_{j,k} = +1$. In other words, the checkerboard error pattern "flips sign" from one iteration to the next, without any decrease in its magnitude. This means that the checkerboard error can never be removed by Jacobi iteration. That's bad.

Here is a more general way to analyze the method. First, rewrite (15.22) as

$$\varepsilon^{\text{new}}_{j,k} = \varepsilon_{j,k} + \frac{1}{4}\left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1} - 4\varepsilon_{j,k}\right) . \tag{15.23}$$

The quantity in parentheses in (15.23) is an "increment" which, when added to the "old" error, $\varepsilon_{j,k}$, gives the new error, $\varepsilon^{\text{new}}_{j,k}$. Eq. (15.23) looks like time differencing, which we have already analyzed in some detail. We can use von Neumann's method to calculate the change (hopefully a decrease) in the error from one sweep to the next. First, write

$$\varepsilon_{j,k} = \varepsilon_0 e^{i(jld+kmd)} , \tag{15.24}$$

where l and m are the wave numbers in the x and y directions, respectively. We also define an "amplification factor" by

$$\varepsilon_{j,k}^{\text{new}} \equiv \lambda \varepsilon_{j,k} \, . \tag{15.25}$$

Substituting (15.24) and (15.25) into (15.23), we find that

$$\begin{aligned} \lambda &= 1 + \frac{1}{4}\left[e^{i(j-1)ld} + e^{i(j+1)ld} + e^{i(k-1)md} + e^{i(k+1)md} - 4\right] \\ &= \frac{1}{2}\left[\cos(ld) + \cos(md)\right] \, . \end{aligned} \tag{15.26}$$

Here we have used Euler's formula. If λ is negative, the sign of the error will oscillate from one sweep to the next. To have rapid, monotonic convergence, we want

$$0 \leq \lambda \ll 1 \, . \tag{15.27}$$

For "long" modes, with $ld \ll 1$ and $md \ll 1$, λ is just slightly less than one. This means that the error in the long modes goes away slowly; it will take many sweeps for the long modes to converge. At the other extreme, for the checkerboard, which has $ld = md = \pi$, we get $\lambda = -1$. This corresponds to the oscillation already discussed above. The error goes to zero after a single sweep for $ld = md = \pi/2$, corresponding to a wavelength (in both directions) of $4d$. In short, the $2d$ error never goes away, but the $4d$ error is eliminated after a single sweep. In general, small-scale errors are killed faster than large-scale errors, but the $2d$-checkerboard is an exception.

The slow convergence of the long modes determines how many iterations are needed to reduce the overall error to an acceptable level. The reason that the long modes converge slowly is that, as you can see from the algorithm, each sweep shares information only among grid cells that are immediate neighbors. Information travels across N grid cells only after N sweeps. Many sweeps are needed for information to travel across a large grid. The long modes are the ones most accurately resolved on the grid, but ironically they limit the speed of convergence.

15.8.2 Jacobi under-relaxtion

A strategy to overcome the checkerboard problem is to "*under-relax*." To understand this approach, we first re-write (15.19) as

$$\widehat{q}^{\,\text{new}}_{j,\,k} = \widehat{q}_{j,\,k} + \left[\frac{1}{4}\left(d^2 f_{j,\,k} + \widehat{q}_{j-1,\,k} + \widehat{q}_{j+1,\,k} + \widehat{q}_{j,\,k-1} + \widehat{q}_{j,\,k+1}\right) - \widehat{q}_{j,\,k}\right] \ . \tag{15.28}$$

This simply says that $\widehat{q}^{\,\text{new}}_{j,\,k}$ is equal to $\widehat{q}_{j,\,k}$ plus an "increment." For the checkerboard error, the increment given by Jacobi relaxation tries to reduce the error, but the increment is "too large," and so "overshoots;" this is why the sign of $\widehat{q}^{\,\text{new}}_{j,\,k}$ flips from one iteration to the next. This simple observation suggests that it would be useful to reduce the increment by multiplying it by a factor less than one, which we will call ω, i.e., we replace (15.28) by

$$\widehat{q}^{\,\text{new}}_{j,\,k} = \widehat{q}_{j,\,k} + \omega\left[\frac{1}{4}\left(d^2 f_{j,\,k} + \widehat{q}_{j-1,\,k} + \widehat{q}_{j+1,\,k} + \widehat{q}_{j,\,k-1} + \widehat{q}_{j,\,k+1}\right) - \widehat{q}_{j,\,k}\right] \ . \tag{15.29}$$

where $0 < \omega < 1$. For $\omega = 0$, a sweep does nothing. For $\omega = 1$, (15.29) reverts to (15.28). We can rearrange (15.23) to

$$\widehat{q}^{\,\text{new}}_{j,\,k} = \widehat{q}_{j,\,k}\,(1-\omega) + \frac{\omega}{4}\left(d^2 f_{j,\,k} + \widehat{q}_{j-1,\,k} + \widehat{q}_{j+1,\,k} + \widehat{q}_{j,\,k-1} + \widehat{q}_{j,\,k+1}\right) \ . \tag{15.30}$$

Substitution of (15.20) into (15.30), and use of (15.17), gives

$$\varepsilon^{\text{new}}_{j,\,k} = \varepsilon_{j,\,k}\,(1-\omega) + \frac{\omega}{4}\left(\varepsilon_{j-1,\,k} + \varepsilon_{j+1,\,k} + \varepsilon_{j,\,k-1} + \varepsilon_{j,\,k+1}\right) \ . \tag{15.31}$$

Starting from (15.31), we can show that

$$\lambda = 1 - \omega + \frac{\omega}{2}\left[\cos\left(ld\right) + \cos\left(md\right)\right] \ . \tag{15.32}$$

Compare with (15.26). Inspection of (15.32) shows that the value of ω that makes $\lambda = 0$ depends on the wave numbers l and m. For $\omega = 0.5$, the checkerboard error will be

destroyed in a single pass. This demonstrates that under-relaxation can be useful with the Jacobi algorithm. On the other hand, using $\omega = 0.5$ makes the long modes converge even more slowly than with $\omega = 1$. This suggests that the optimal value of ω is in the range $0.5 < \omega < 1$. Values in the range 0.6 – 0.8 are often used.

Suppose that, on a particular sweep, the error is spatially uniform over the grid. Then, according to (15.22), the error will never change under Jacobi relaxation, and this is true even with under-relaxation, as can be seen from (15.31). When the error varies spatially, we can decompose it into its spatial average plus the departure from the average. The argument just given implies that the spatially uniform part of the error will never decrease. This sounds terrible, but it's not really a problem because, as discussed earlier, when solving a problem of this type the average over the grid has to be determined by a "boundary condition." If the appropriate boundary condition can be applied in the process of formulating the first guess, then the domain-mean error will be zero even before the relaxation begins.

Note, however, that if the error field is spatially smooth (but not uniform), it will change only a little on each sweep. This shows again that *the "large-scale" part of the error is reduced only slowly, while the smaller-scale part of the error is reduced more rapidly*. Once the small-scale or "noisy" part of the error has been removed, the remaining error will be smooth.

For a given domain size, convergence is slower (i.e., more sweeps are needed) when the grid spacing is finer. Qualitatively, this seems fair, since a finer grid can hold more information.

For errors of intermediate spatial scale, Jacobi relaxation works reasonably well.

15.8.3 Gauss-Seidel relaxation

Gauss-Seidel relaxation is similar to Jacobi relaxation, except that each value is updated immediately after it is calculated. For example, suppose that we start at the lower left-hand corner of the grid, and work our way across the bottom row, then move to the left end of the second row from the bottom, and so on. In Gauss-Seidel relaxation, as we come to each grid point we use the "new" values of all qs that have already been updated, so that (15.19) is replaced by

$$\widehat{q}^{\,\text{new}}_{j,\,k} = \frac{1}{4}\left(d^2 f_{j,\,k} + \widehat{q}^{\,\text{new}}_{j-1,\,k} + \widehat{q}_{j+1,\,k} + \widehat{q}^{\,\text{new}}_{j,\,k-1} + \widehat{q}_{j,\,k+1}\right) . \tag{15.33}$$

This "sequential updating" immediately reduces the storage requirements, because it is not necessary to save all of the old values and all of the new values simultaneously. More importantly, it also speeds up the convergence of the iteration, compared to Jacobi relaxation.

Sequential updating allows information to propagate rapidly through the system, giving the Gauss-Seidel method an intrinsic acceleration relative to the Jabobi method.

Obviously (15.33) does not apply to the very first point encountered on the very first sweep, because at that stage no "new" values are available. For the first point, we will just perform a Jacobi-style update using (15.19). It is only for the second and later *rows* of points that (15.33) actually applies. Because values are updated as they are encountered during the sweep, the results obtained with Gauss-Seidel relaxation depend on where the sweep starts. To the extent that the final result satisfies (15.17) exactly, it will be independent of where the sweep starts.

For Gauss-Seidel relaxation, the error-reduction formula corresponding to (15.22) is

$$\varepsilon^{\text{new}}_{j,\,k} = \frac{1}{4}\left(\varepsilon^{\text{new}}_{j-1,\,k} + \varepsilon_{j+1,\,k} + \varepsilon^{\text{new}}_{j,\,k-1} + \varepsilon_{j,\,k+1}\right) , \tag{15.34}$$

and the amplification factor defined by (15.25) turns out to be a complex number, which means that the error will oscillate in space as we move through a sweep.

Consider the following simple example on a 6 x 6 mesh. Suppose that f is identically zero, so that the solution (with periodic boundary conditions) is that q is spatially constant. Exercising our second "boundary condition," we choose the constant to be zero. We make the rather ill-considered first guess that the solution is a checkerboard:

$$\widehat{q}^{\,0}_{j,\,k} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} . \tag{15.35}$$

Here the superscript zero denotes the first guess. After partially completing one sweep, doing the bottom row and the left-most three elements of the second row from the bottom, we have:

$$\widehat{q}^{\,1,\text{partial}}_{j,k} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ -0.5 & 0.25 & -0.281 & -1 & 1 & -1 \\ 1 & -0.5 & 0.625 & -0.593 & 0.602 & -0.60 \end{bmatrix} . \tag{15.36}$$

Here the superscript "1, partial" means that the first sweep has been partially completed. Inspection of (15.36) shows that although the solution is flipping sign from one point to the next, the amplitude of the checkerboard is decreasing noticeably. You can finish the exercise for yourself.

15.8.4 Gauss-Seidel over-relaxation

Because Gauss-Seidel relaxation uses new values as soon as they are available, the increments tend to be smaller than with Jacobi relaxation. For this reason, convergence of Gauss-Seidel relaxation can be speeded up by multiplying the increment by a factor *greater* than one, i.e., by "*over-relaxation*." By analogy with (15.30), we replace (15.33) by

$$\widehat{q}^{\,\text{new}}_{j,k} = \widehat{q}_{j,k}(1-\omega) + \frac{\omega}{4}\left(d^2 f_{j,k} + \widehat{q}^{\,\text{new}}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}^{\,\text{new}}_{j,k-1} + \widehat{q}_{j,k+1}\right) , \tag{15.37}$$

where this time we choose $\omega > 1$. Choosing ω too large will cause the iteration to diverge. It can be shown that the convergence of (15.37) is optimized (i.e., made as rapid as possible) if we choose

$$\omega = \frac{2}{1+\sin(\pi d/L)} , \tag{15.38}$$

where L is the total width of the domain. According to (15.38), ω approaches 2 on very fine grids. In practice, some experimentation may be needed to find the best value of ω.

The algorithm represented by (15.37) and (15.38) is called "successive over-relaxation," or SOR.

15.8.5 The alternating-direction implicit method

Yet another relaxation scheme is the "alternating-direction implicit" method, often called "ADI" for short. With ADI, the spatial coordinates are treated separately and successively within each iteration sweep. We rewrite (15.17) as

$$\left(-q_{j-1,\,k}+2q_{j,\,k}-q_{j+1,\,k}\right)+\left(-q_{j,\,k-1}+2q_{j,\,k}-q_{j,\,k+1}\right)=d^2 f_{j,\,k}. \tag{15.39}$$

The first quantity in parentheses on the left-hand side of (15.39) involves variations in the x-direction only, and the second involves variations in the y-direction only. We proceed in two steps on each sweep. The first step treats the x-dependence to produce an intermediate approximation by solving

$$\left[-\widehat{q}^{\,\text{int}}_{j-1,\,k}+(2+r)\,\widehat{q}^{\,\text{int}}_{j,\,k}-\widehat{q}^{\,\text{int}}_{j+1,\,k}\right]=d^2 f_{j,\,k}-\left[-\widehat{q}_{j,\,k-1}+(2-r)\,\widehat{q}_{j,\,k}-\widehat{q}_{j,\,k+1}\right] \tag{15.40}$$

for the values with superscript "int." Here r is a parameter used to control convergence, as discussed below. Eq. (15.40) is a tri-diagonal system, which can easily be solved. The sweep is completed by solving

$$\left[-\widehat{q}^{\,\text{new}}_{j,\,k-1}+(2-r)\,\widehat{q}^{\,\text{new}}_{j,\,k}-\widehat{q}^{\,\text{new}}_{j,k+1}\right]=d^2 f_{j,\,k}-\left[-\widehat{q}^{\,\text{int}}_{j-1,\,k}+(2+r)\,\widehat{q}^{\,\text{int}}_{j,\,k}-\widehat{q}^{\,\text{int}}_{j+1,\,k}\right] \tag{15.41}$$

as a second tridiagonal system. It can be shown that the ADI method converges if r is positive and constant for all sweeps. The optimal value of r is

$$r=2\sin\left(\frac{\pi d}{L}\right)\,. \tag{15.42}$$

This gives $r\ll 1$ if $d/L\ll 1$, i.e., if many grid points are used to cover the domain. For the shortest wave, with $d/L=1/2$, we get $r=2$.

15.9 The multigrid method

15.9.1 The basic idea

Fulton et al. (1986) summarize the important multi-grid approach to solving boundary-value problems, which was developed by Achi Brandt (1973, 1977). The basic idea is very simple and elegant, although implementation can be complicated.

As we have already discussed, with Jacobi and Gauss-Seidel relaxation the small-scale errors are eliminated quickly, while the large-scale errors disappear more slowly. A key observation is that *the large-scale part of the error can be represented on a coarse grid*. Because a coarse grid has only a few grid points, the large-scale part of the error can be removed quickly. In addition, of course, less work is needed to do a sweep on a coarser grid.

Putting these ideas together, we arrive at a strategy whereby *we use a coarse grid to remove the large-scale errors, and a fine grid to remove the small-scale errors*. In practice, we use as many grids as possible; the "multi" in the multi-grid method is important.

We define M grids, each denoted by superscript l, such that $M \geq l \geq 1$. The finest grid is denoted by $l = M$, and the coarsest by $l = 1$. Typically, each grid is chosen to have half as many points in each direction as the next finer grid, and is composed of a subset of the points used in the next finer grid. With a two-dimensional domain, grid l would have $1/4$ as many points as grid $l+1$.

What we want is a solution for our unknown, q, on the finest grid, satisfying

$$\begin{aligned} -L^M q^M &= f^M, \quad \text{and} \\ q &= g \text{ on the boundary of the domain ,} \end{aligned} \tag{15.43}$$

where L is a linear operator (which could be the Laplacian). Eq. (15.43) is a generalization of (15.16).

The approximate solution on grid l is denoted by $\widehat{q}^{\,l}$. We create an f^l for each grid. The error on grid l satisfies

$$q^l = \widehat{q}^{\,l} - \widehat{\varepsilon}^{\,l} \,. \tag{15.44}$$

Here is the idea in a nutshell, postponing all discussion of the details until the next subsection.

The multigrid method starts by creating an initial guess, $\widehat{q}^M$, on the finest grid. Several sweeps are performed on that grid.

Data is then transferred to the next coarser grid, $M-1$ by "injection," which means simply copying the fine-grid values onto the corresponding points of the coarse grid[3]. Additional sweeps are performed on grid $M-1$.

This process continues until the coarsest grid is reached. We then work our way back up through the finer grids, one by one. Data is transferred from the coarser grids to the finer grids by interpolation[4]. Additional relaxations are performed on each finer grid, until we reach the finest grid with $l = M$.

This sequential fine-to-coarse and course-to-fine procedure is called a "V-cycle." It is illustrated schematically in Fig. 15.1. Usually, multiple V-cycles are needed to achieve adequate convergence.

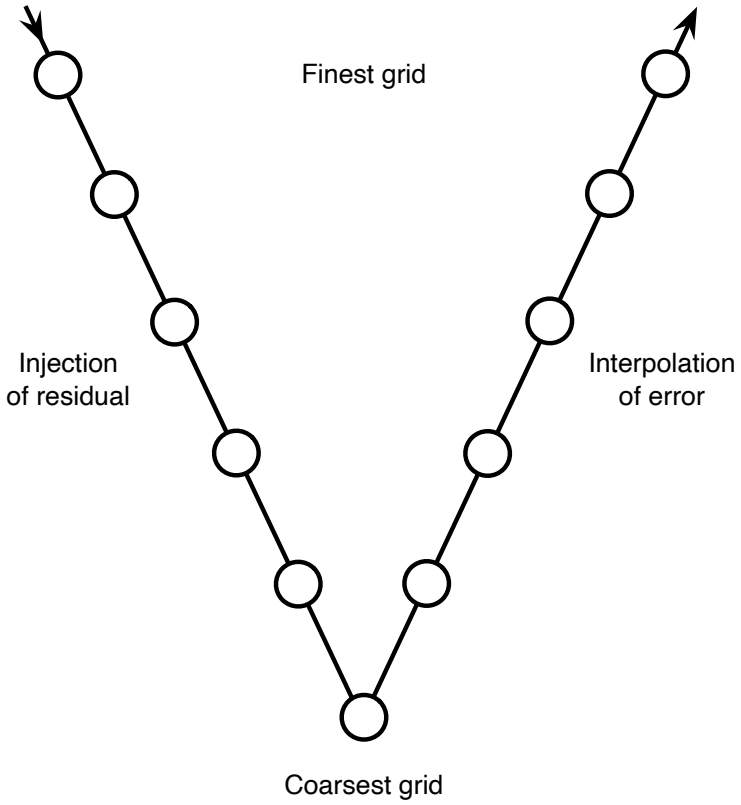


Figure 15.1: Schematic illustration of a V-cycle. See text for details. The diagram is modeled after Fig. 4 of Fulton et al. (1986).

Although the transfers between grids involve some computational work, the net effect is to speed up the solution considerably beyond what can be achieved using only the finest

[3]More generally the transfer of data from the fine grid to the coarse grid is called "restriction."

[4]More generally the transfer of data from the coarse grid to the fine grid is called "prolongation."

grid with Jacobi or Gauss-Seidel relaxation. In addition, the scaling improves, as discussed below.

15.9.2 The details

The error on grid l is denoted by

$$\widehat{\varepsilon}^{\,l} \equiv \widehat{q}^{\,l} - q^l \,. \tag{15.45}$$

Using (15.45) to eliminate q^M in (15.43), we find that

$$-L^l \left(\widehat{q}^{\,l} - \widehat{\varepsilon}^{\,l} \right) = f^l \,. \tag{15.46}$$

Since L is linear (by assumption), we know that

$$L^l \left(\widehat{q}^{\,l} - \widehat{\varepsilon}^{\,l} \right) = L^l \, \widehat{q}^{\,l} - L^l \, \widehat{\varepsilon}^{\,l} \,. \tag{15.47}$$

With the use of (15.47) and (15.43), we can rewrite (15.46) as

$$\boxed{L^l \, \widehat{\varepsilon}^{\,l} = r^l} \,, \tag{15.48}$$

where

$$r^l \equiv f^l + L^l \, \widehat{q}^{\,l} \tag{15.49}$$

is called the "*residual*," and Eq. (15.48) is called the "*residual equation*." The residual is what comes out when the linear operator is applied to the error. We can also say that the residual is the "forcing" in the residual equation. Eq. (15.48) shows that when the error is zero everywhere, the residual is also zero. In this sense, the residual is a measure of the error.

Inspection of (15.49) shows that the quantities needed to compute r^l are known. They are the forcing function, f^l, and the current partially converged solution, $\widehat{q}^{\,l}$. In contrast,

the error itself is not known. Since the residual is known, *the unknown in* (15.48) *is the error*, $\widehat{\varepsilon}^{\,l}$. Instead of solving (15.43) for $\widehat{q}^{\,l}$, we can solve the residual equation (15.48) for $\widehat{\varepsilon}^{\,l}$. We perform a few sweeps with the residual equation to obtain an estimate of $\widehat{\varepsilon}^{\,l}$. We can then correct $\widehat{q}^{\,l}$ using

$$\left(\widehat{q}^{\,l}\right)^{\text{corrected}} = \widehat{q}^{\,l} - \widehat{\varepsilon}^{\,l} \,. \tag{15.50}$$

We can then update the residual, using (15.49).

Next, we transfer r^l and $\left(\widehat{q}^{\,l}\right)^{\text{corrected}}$ to the next coarser grid, by injection. We then solve the residual equation for the error on the coarser grid.

This process can be repeated until we reach the coarsest possible grid – say a 3×3 grid as shown in Fig. 15.2. On the coarsest grids, direct solves (e.g., matrix inversion) may be preferable to relaxation sweeps.

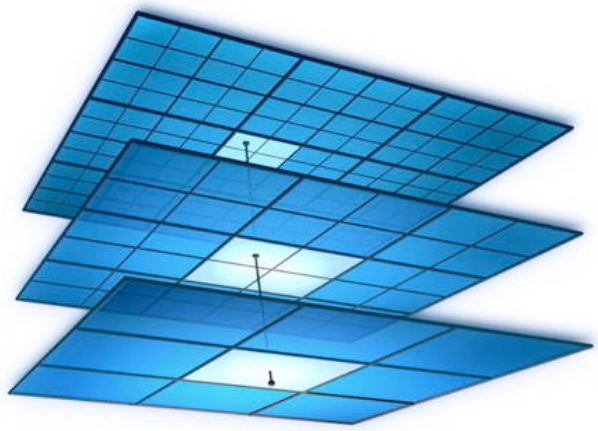


Figure 15.2: Schematic illustrating three grids used by a multigrid solver, with the finest grid on top. Source: `http://ccma.math.psu.edu/ccma-wp/?page_id=237`.

Having worked our way down to the coarsest grid, we then start back the other way, towards the finer grids, by interpolation. See Fig. 15.1. Interpolation works best if the interpolated quantity is smooth. As the iteration converges, the distribution of the error over the grid becomes smoother, because ultimately it approaches zero everywhere. In contrast, the distribution of $\widehat{q}_{j,\,k}$ may or may not become smoother, because the converged solution might, in fact, be noisy. This motivates to interpolate the smooth error, instead of the possibly noisy $\widehat{q}_{j,\,k}$.

The interpolated error on grid $l+1$ is used to correct $\widehat{q}^{\,l+1}$. Some relaxation sweeps may be performed to further refine $\widehat{q}^{\,l+1}$. The new estimate of $\widehat{q}$ can then be used to compute a

new residual from (15.49). A sweep (or two) is performed to smooth the error on the finer grid. The result is interpolated to the next finer grid, and so on, until we arrive back at the finest grid. The error on the finest grid can be used to correct $\widehat{q}^M$, using

For further discussion of multi-grid methods, see the paper by Fulton et al. (1986).

As mentioned earlier, multi-grid methods can be difficult to program. Fortunately, there are optimized multi-grid solvers available in libraries (e.g., `https://github.com/trifwn/mudpack_SP`.

15.10 Summary

Boundary-value problems occur quite frequently in atmospheric science. The main issue is the amount of work needed to find the solution. Fast solutions to one-dimensional problems are very easy to obtain, but two- and three-dimensional problems are more challenging, particularly when complex geometry is involved. Among the most useful methods available today for multi-dimensional problems are the multi-grid methods and the conjugate-gradient methods (e.g., Shewchuk, 1994). Spectral methods are also excellent, and will be discussed in a later chapter.

Table 15.1 summarizes the operations counts and storage requirements of some well known methods for solving boundary-value problems. The best possible scalings for the operation count and storage requirement are $O\left(N^2\right)$. Only the multi-grid method achieves this ideal.

Table 15.1: Well known methods for solving boundary value problems, and the operation count and storage. Here the total number of points is N^2.

Method	Operation Count	Storage Requirement
Gaussian Elimination	N^4	N^2
Jacobi	N^4	N^2
Gauss-Seidel	N^4	N^2
Successive Over-Relaxation	N^3	N^2
Alternating Direction Implicit	$N^2 \ln N$	N^2
Multigrid	N^2	N^2

15.11 Problems

1. Prove that with periodic boundary conditions the domain-average of q is not changed by a sweep using
   (a) Jacobi relaxation;
   (b) Gauss-Seidel relaxation.
2. Use von Neumann's method to analyze the convergence of

$$\widehat{q}_{j,k}^{\text{new}} = \widehat{q}_{j,k} + \omega \left[\frac{1}{4} \left(d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1} \right) - \widehat{q}_{j,k} \right] . \quad (15.51)$$

3. Consider a square domain, of width L, with periodic boundary conditions in both x and y directions. We wish to solve

$$\nabla^2 q = f(x,y) = \left(\sin \frac{4\pi x}{L} \right) \left(\cos \frac{4\pi y}{L} \right) \quad (15.52)$$

for the unknown function q, where

$$\begin{aligned} 0 \leq & x \leq L \, , \\ 0 \leq & y \leq L \, . \end{aligned} \quad (15.53)$$

Assume that the domain-average value of q is zero, and impose this condition on your numerical solution. For simplicity, use $L = 1$. Use centered second-order differences (with the + stencil) to approximate $\nabla^2 q$. Use $N = 100$ points in both directions. The periodic boundary conditions mean that $j = 1$ is the same as $j = 101$, and $k = 1$ is the same as $k = 101$.

   (a) Find and plot the exact solution.
   (b) Also find and plot the solution using each of the relaxation methods listed below.
      - Jacobi relaxation;
      - Jacobi under-relaxation, with a suitable choice of the parameter ω;
      - Gauss-Seidel relaxation;

- Gauss-Seidel over-relaxation, with a suitable choice of the parameter ω.

For each of the relaxation methods, try the following two initial guesses:

$$\begin{aligned} &1)\ q_{j,\,k} = (-1)^{j+k}, \\ &2)\ q_{j,\,k} = 0 \text{ everywhere .} \end{aligned} \tag{15.54}$$

(c) Let n be an "iteration counter," i.e., $n = 0$ for the initial guess, $n = 1$ after one sweep, etc. Define the error after n sweeps by

$$\varepsilon^n_{j,\,k} \equiv \nabla^2 \left(\widehat{q}^{\,n}_{j,\,k} \right) - f_{j,\,k} \,. \tag{15.55}$$

Here $\nabla^2 \left(\widehat{q}^{\,n}_{j,\,k} \right)$ is the finite-difference Laplacian of the the approximate solution, and $f_{j,\,k}$ is "forcing function" given in (15.25), as evaluated on the grid. Let the convergence criterion be

$$\max \forall\,(j,\,k) \left\{ \left| \varepsilon^n_{j,\,k} \right| \right\} < 10^{-2} \max \forall\,(j,\,k) \left\{ \left| f_{j,\,k} \right| \right\} \,. \tag{15.56}$$

How many iterations are needed to obtain convergence with Jacobi, Jacobi with under-relaxation, Gauss-Seidel, and Gauss-Seidel over-relaxation?

(d) Plot the RMS error $R^n \equiv \sqrt{\frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \left(\varepsilon^n_{j,\,k} \right)^2}$ as a function of n (or, if you prefer, as a function of $\ln n$) for all three methods.

4. Construct a subroutine that solves for the stream function, given the vorticity, in a periodic domain with a hexagonal grid. Start from the code that you created for the hexagonal-grid homework problem in Chapter 10. Use the Jacobi method with under-relaxation. Test your subroutine by feeding it distributions of the vorticity for which you can compute the exact solution analytically. Show the results of your test in your homework submission.

   Note that in all cases the domain-average of the specified vorticity "forcing" must be zero. Explain why that is true.

   Ensure that the domain-average of the stream function is equal to zero in your solution. Explain how you do that.

5. Construct a one-dimensional multi-grid solver. Test it by solving (15.9) on a domain with 48 equally spaced grid points and periodic boundary conditions. The coarsest grid will have 3 points. For use in the test, invent a forcing $f(x)$ that is sufficiently complicated to be interesting. Remember that $f(x)$ must be periodic. On each grid, use an under-relaxed Jacobi solver, except that on the coarsest grids, you may choose to use a non-iterative method. Provide a step-by-step written explanation of how your multi-grid solver works.

Chapter 16

It's only dissipation (But I like it)

16.1 The diffusion equation

Diffusion is a macroscopic statistical description of microscopic advection. Here "microscopic" refers to scales below the resolution of a model. In general diffusion can occur in three dimensions, but often in atmospheric science only vertical diffusion, i.e., one-dimensional diffusion, is the main issue. The process of one-dimensional diffusion can be represented in simplified form by

$$\frac{\partial q}{\partial t} = -\frac{\partial F_q}{\partial x} \,. \tag{16.1}$$

Here q is the "diffused" quantity, x is the spatial coordinate, and F_q is a flux of q due to diffusion. Although very complex parameterizations for F_q are required in many applications, a simple parameterization that is often encountered in practice is

$$F_q = -K\frac{\partial q}{\partial x} \,, \tag{16.2}$$

where K is a non-negative "diffusion coefficient," which must be determined somehow. Physically meaningful applications of are possible when

$$K \geq 0 \,. \tag{16.3}$$

Substitution of (16.2) into (16.1) gives

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial x}\left(K\frac{\partial q}{\partial x}\right) . \tag{16.4}$$

Because (16.4) involves second derivatives in space, it requires two boundary conditions, one of which might determine the value of F_q at a wall. Here, we assume for simplicity that the one-dimensional domain is periodic. It then follows immediately from (16.1) that the spatially averaged value of q does not change with time:

$$\frac{d}{dt}\left(\int\limits_{\text{spatial domain}} q dx\right) = 0 . \tag{16.5}$$

When (16.3) is satisfied, (16.4) describes "downgradient" transport, in which the flux of q is from larger values of q towards smaller values of q. Such a process tends to reduce large values of q, and to increase small values, so that the spatial variability of q decreases with time. In particular, we can show that

$$\frac{d}{dt}\left(\int\limits_{\text{spatial domain}} q^2 dx\right) \leq 0 . \tag{16.6}$$

To prove this, multiply both sides of (16.4) by q:

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{q^2}{2}\right) =& q\frac{\partial}{\partial x}\left(K\frac{\partial q}{\partial x}\right) \\ =& \frac{\partial}{\partial x}\left(qK\frac{\partial q}{\partial x}\right) - K\left(\frac{\partial q}{\partial x}\right)^2 .\end{aligned} \tag{16.7}$$

When we integrate the second line of (16.7) over a periodic domain, the first term vanishes and the second is negative (or possibly zero). The result (16.6) follows immediately.

Given the assumed periodic boundary conditions, we can expand q in a Fourier series:

$$q(x,t) = \sum_k \text{Re}\left[\widehat{q}_k(t)\, e^{ikx}\right] . \tag{16.8}$$

Substituting into (16.4), and (temporarily) assuming spatially constant K, we find that the amplitude of a particular Fourier mode satisfies

$$\frac{d\widehat{q}_k}{dt} = -k^2 K \widehat{q}_k \,, \tag{16.9}$$

which is the decay equation. This shows that there is a close connection between the diffusion equation and the decay equation, which is good to know because we have already discussed time differencing for the decay equation. The solution of (16.9) is

$$\widehat{q}_k(t) = \widehat{q}_k(0)\, e^{-k^2 K t} \,. \tag{16.10}$$

Note that higher wave numbers decay more rapidly, for a given value of K. Since

$$\widehat{q}_k(t+\Delta t) = \widehat{q}_k(0)\, e^{-k^2 K (t+\Delta t)} = \widehat{q}_k(t)\, e^{-k^2 K \Delta t} \,, \tag{16.11}$$

we see that, for the exact solution, the amplification factor is given by

$$\lambda = e^{-k^2 K \Delta t} < 1 \,. \tag{16.12}$$

16.2 A simple explicit scheme

A finite-difference analog of (16.1) is

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left(q_{j+1}^n - q_j^n \right) - \kappa_{j-\frac{1}{2}} \left(q_j^n - q_{j-1}^n \right) \,, \tag{16.13}$$

where for convenience we define the nondimensional combination

$$\kappa_{j+\frac{1}{2}} \equiv \frac{K_{j+\frac{1}{2}} \Delta t}{(\Delta x)^2} \,. \tag{16.14}$$

Here we have assumed for simplicity that Δx is a constant. With periodic boundary conditions, (16.13) guarantees conservation of q in the sense that

$$\sum_j q_j^{n+1} \Delta x = \sum_j q_j^n \Delta x \,[. \tag{16.15}$$

The scheme given by (16.13) combines forward time differencing with centered space differencing. Recall that this combination is unconditionally unstable for the advection problem, but it turns out to be conditionally stable for diffusion. To analyze the stability of (16.13) using von Neumann's method, we have to assume that κ is a constant. Then (16.13) yields

$$(\lambda - 1) = \kappa \left[\left(e^{ik\Delta x} - 1 \right) - \left(1 - e^{-ik\Delta x} \right) \right], \tag{16.16}$$

which is equivalent to

$$\lambda = 1 - 4\kappa \sin^2 \left(\frac{k\Delta x}{2} \right) \leq 1. \tag{16.17}$$

Note that λ is real and less than one. Instability occurs for $\lambda < -1$, which is equivalent to

$$\kappa \sin^2 \left(\frac{k\Delta x}{2} \right) > \frac{1}{2}. \tag{16.18}$$

The worst case is $\sin^2 (k\Delta x/2) = 1$, which occurs for $k\Delta x/2 = \frac{\pi}{2}$, or $k\Delta x = \pi$. This is the $2\Delta x$ wave. We conclude that with (16.13)

$$\kappa \leq \frac{1}{2} \text{ is required for stability.} \tag{16.19}$$

When the scheme is unstable, it blows up in an oscillatory fashion.

When the stability criterion derived above is satisfied, we can be sure that

$$\sum_j \left(q_j^{n+1} \right)^2 < \sum_j \left(q_j^n \right)^2; \tag{16.20}$$

this is the condition for stability according to the energy method discussed in Chapter 2. Eq. (16.20) is analogous to (16.6).

16.3 An implicit scheme

We can obtain unconditional stability through the use of an implicit scheme, but at the cost of some additional complexity. Replace (16.13) by

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left(q_{j+1}^{n+1} - q_j^{n+1} \right) - \kappa_{j-\frac{1}{2}} \left(q_j^{n+1} - q_{j-1}^{n+1} \right). \tag{16.21}$$

We use the *energy method* to analyze the stability of (16.21), for the case of spatially variable but non-negative K. Multiplying (16.21) by q_j^{n+1}, we obtain:

$$\left(q_j^{n+1} \right)^2 - q_j^{n+1} q_j^n = \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \kappa_{j+\frac{1}{2}} \left(q_j^{n+1} \right)^2 - \kappa_{j-\frac{1}{2}} \left(q_j^{n+1} \right)^2 + \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} \tag{16.22}$$

Summing over the domain gives

$$\begin{aligned} \sum_j \left(q_j^{n+1} \right)^2 - \sum_j q_j^{n+1} q_j^n &= \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} \left(q_j^{n+1} \right)^2 \\ &\quad - \sum_j \kappa_{j-\frac{1}{2}} \left(q_j^{n+1} \right)^2 + \sum_j \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} \\ &= \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} \left(q_j^{n+1} \right)^2 \\ &\quad - \sum_j \kappa_{j+\frac{1}{2}} \left(q_{j+1}^{n+1} \right)^2 + \sum_j \kappa_{j+\frac{1}{2}} q_j^{n+1} q_{j+1}^{n+1} \\ &= - \sum_j \kappa_{j+\frac{1}{2}} \left(q_{j+1}^{n+1} - q_j^{n+1} \right)^2, \end{aligned} \tag{16.23}$$

which can be rearranged to

$$\sum_j q_j^{n+1} q_j^n = \sum_j \left[\left(q_j^{n+1} \right)^2 + \kappa_{j+\frac{1}{2}} \left(q_{j+1}^{n+1} - q_j^{n+1} \right)^2 \right]. \tag{16.24}$$

Next, note that

$$\sum_j \left(q_j^{n+1} - q_j^n\right)^2 = \sum_j \left[\left(q_j^{n+1}\right)^2 + \left(q_j^n\right)^2 - 2q_j^{n+1}q_j^n\right] \geq 0. \tag{16.25}$$

Substitute (16.24) into (16.25), to obtain

$$\sum_j \left\{\left(q_j^{n+1}\right)^2 + \left(q_j^n\right)^2 - 2\left[\left(q_j^{n+1}\right)^2 + \kappa_{j+\frac{1}{2}}\left(q_{j+1}^{n+1} - q_j^{n+1}\right)^2\right]\right\} \geq 0, \tag{16.26}$$

which can be simplified and rearranged to

$$\sum_j \left[\left(q_j^{n+1}\right)^2 - \left(q_j^n\right)^2\right] \leq -2\sum_j \left[\kappa_{j+\frac{1}{2}}\left(q_{j+1}^{n+1} - q_j^{n+1}\right)^2\right] \leq 0. \tag{16.27}$$

Eq. (16.27) demonstrates that $\sum_j \left[\left(q_j^{n+1}\right)^2 - \left(q_j^n\right)^2\right]$ is less than or equal to a not-positive number. In short,

$$\sum_j \left[\left(q_j^{n+1}\right)^2 - \left(q_j^n\right)^2\right] \leq 0. \tag{16.28}$$

This is the desired result.

The trapezoidal implicit scheme is also unconditionally stable for the diffusion equation, and it is more accurate than the backward-implicit scheme discussed above.

Eq. (16.21) contains three unknowns, namely q_j^{n+1}, q_{j+1}^{n+1}, and q_{j-1}^{n+1}. We must therefore solve a system of such equations, for the whole domain at once. Assuming that K is independent of q (often not true in practice), the system of equations is linear and tridiagonal, so it is not hard to solve. In realistic models, however, K can depend strongly on multiple dependent variables which are themselves subject to diffusion, so that multiple coupled systems of nonlinear equations must be solved simultaneously in order to obtain a fully implicit solution to the diffusion problem. For this reason, implicit methods are not always practical.

16.4 The DuFort-Frankel scheme

The DuFort-Frankel scheme is partially implicit and unconditionally stable, but does not lead to a set of equations that must be solved simultaneously. The scheme is given by

$$\frac{q_j^{n+1}-q_j^{n-1}}{2\Delta t}=\frac{1}{(\Delta x)^2}\left[K_{j+\frac{1}{2}}\left(q_{j+1}^n-q_j^{n+1}\right)-K_{j-\frac{1}{2}}\left(q_j^{n-1}-q_{j-1}^n\right)\right]. \tag{16.29}$$

Notice that three time levels appear, which means that we will have a computational mode in time, in addition to a physical mode. *Time level $n+1$ appears only in connection with grid point j, so the solution of* (16.29) *can be obtained without solving a system of simultaneous equations:*

$$q_j^{n+1}=\frac{q_j^{n-1}+2\left[\kappa_{j+\frac{1}{2}}q_{j+1}^n-\kappa_{j-\frac{1}{2}}\left(q_j^{n-1}-q_{j-1}^n\right)\right]}{1+2\kappa_{j+\frac{1}{2}}}. \tag{16.30}$$

To apply von Neumann's method, we assume spatially constant κ, and for convenience define

$$\alpha\equiv 2\kappa\geq 0. \tag{16.31}$$

The amplification factor satisfies

$$\lambda^2-1=\alpha\left(\lambda e^{ik\Delta x}-\lambda^2-1+\lambda e^{-ik\Delta x}\right), \tag{16.32}$$

which is equivalent to

$$\lambda^2(1+\alpha)-\lambda 2\alpha\cos(k\Delta x)-(1-\alpha)=0. \tag{16.33}$$

The solutions are

$$\begin{aligned}\lambda &= \frac{\alpha\cos(k\Delta x) \pm \sqrt{\alpha^2\cos^2(k\Delta x) + (1-\alpha^2)}}{1+\alpha} \\ &= \frac{\alpha\cos(k\Delta x) \pm \sqrt{1-\alpha^2\sin^2(k\Delta x)}}{1+\alpha}.\end{aligned} \tag{16.34}$$

The plus sign corresponds to the physical mode, for which $\lambda \to 1$ as $\alpha \to 0$, and the minus sign corresponds to the computational mode. This can be seen by taking the limit $k\Delta x \to 0$.

Consider two cases. First, if $\alpha^2\sin^2(k\Delta x) \leq 1$, then λ is real, and by considering the two solutions separately it is easy to show that

$$|\lambda| \leq \frac{1+|\alpha\cos(k\Delta x)|}{1+\alpha} \leq 1. \tag{16.35}$$

Second, if $\alpha^2\sin^2(k\Delta x) > 1$, which implies that $\alpha > 1$, then λ is complex, and we find that

$$|\lambda| = \frac{\sqrt{\alpha^2\cos^2(k\Delta x) + \alpha^2\sin^2(k\Delta x) - 1}}{1+\alpha} = \frac{\sqrt{\alpha^2-1}}{1+\alpha} = \sqrt{\frac{\alpha-1}{\alpha+1}} < 1. \tag{16.36}$$

We conclude that the scheme is unconditionally stable.

It does not follow, however, that the scheme gives a good solution for large Δt. For $\alpha \to \infty$ (strong diffusion and/or a long time step), (16.36) gives

$$|\lambda| \to 1. \tag{16.37}$$

We conclude that the Dufort-Frankel scheme does not damp when the diffusion coefficient is large or the time step is large. This is very bad behavior.

16.5 Hyperdiffusion

Jiménez (1994)

16.6 Summary

Diffusion is a relatively simple process that preferentially wipes out small-scale features. The most robust schemes for the diffusion equation are fully implicit, but these give rise to systems of simultaneous equations. The DuFort-Frankel scheme is unconditionally stable and easy to implement, but behaves badly as the time step becomes large for fixed Δx.

16.7 Problems

1. Prove that the trapezoidal implicit scheme with centered second-order space differencing is unconditionally stable for the one-dimensional diffusion equation. Do not assume that K is spatially constant.

2. Program the backward-implicit version of the one-dimensional diffusion equation. Use a constant diffusion coefficient $K = 1$ and a grid-spacing of $\Delta x = 1$. Use a periodic domain consisting of 100 grid points. To solve the equations, program an under-relaxed Jabobi solver, with $\omega = 0.75$. Let the initial condition be

$$q_j = 100,\ j \in [1,\ 50]\,,\ \text{and}\ q_j = 110\ \text{for}\ j \in [51, 100]\ .$$

   Run the model twice out to $t = 1000$. In the first simulation use $\Delta t = 0.5$ and in the second use $\Delta t = 5.0$. Compare the solutions at $t = 100$, $t = 500$, and $t = 1000$.

3. Use the energy method to evaluate the stability of

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left(q_{j+1}^n - q_j^n\right) - \kappa_{j-\frac{1}{2}} \left(q_j^n - q_{j-1}^n\right).$$

   Do not assume that κ is spatially constant.

Chapter 17

The shallow-water equations

17.1 Introduction

The two-dimensional shallow water equations are very widely used to test horizontal differencing schemes for atmosphere and ocean models. They are based on three assumptions or idealizations. The first assumption is that the fluid is incompressible, and the density is uniform everywhere and for all time. As discussed earlier, water is much less compressible than air. The assumption of incompressibility has two important consequences. First, the continuity equation (2.20) simplifies to

$$\nabla \cdot \mathbf{V} = 0 \,. \tag{17.1}$$

Because the three-dimensional divergence vanishes, convergence (divergence) in the horizontal must be accompanied by divergence (convergence) in the vertical. We assume that the fluid is bounded below by an impermeable surface (like the Earth's surface) of possibly variable height h_T, and bounded above by a "free surface" (like the surface of a lake) at height $h_T + h$. By definition, no mass crosses a free surface. See Fig. 17.1. With these assumptions, (17.1) leads to a continuity equation of the form

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\mathbf{v}) = 0 \,, \tag{17.2}$$

where h is the depth of the fluid.

The second consequence of incompressibility is that the expansion-work term vanishes in the mechanical energy and thermodynamic energy equations (see Chapter 2), so that there is no conversion between internal energy and mechanical energy. The internal energy

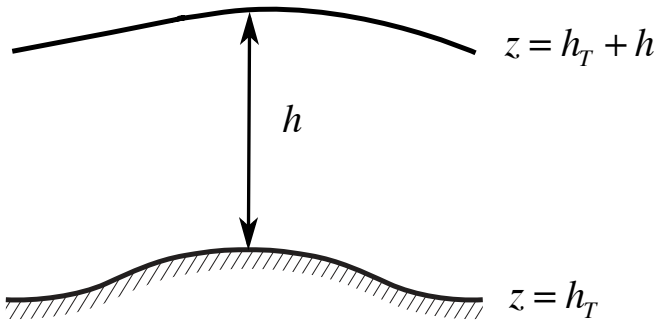


Figure 17.1: Schematic illustrating shallow water flowing over a "mountain."

can therefore be ignored, and there is no need for a thermodynamic energy equation, or a temperature. Any kinetic energy that is dissipated is just "gone," so when dissipative friction is included the shallow water equations do not conserve total energy.

The second assumption is that there is no vertical shear of the horizontal wind. This is where "shallowness" comes in.

The third assumption is that the motion is quasi-hydrostatic, so that the pressure at a give level is proportional to the depth of the water above that level, which varies only due to changes in the height of the free surface.

Using the first and third assumptions, the horizontal pressure-gradient force can be written as

$$\begin{aligned} \mathbf{HPGF} &= -\frac{1}{\rho}\nabla\left[g\rho\left(h_T+h\right)\right] \\ &= -g\nabla\left(h_T+h\right)\ , \end{aligned} \tag{17.3}$$

where $h_T + h$ is the height of the free-surface. We can then write the shallow-water form of the horizontal momentum equation as

$$\frac{D\mathbf{v}}{Dt} + f\mathbf{k}\times\mathbf{v} + g\nabla\left(h+h_T\right) = 0\ . \tag{17.4}$$

Here $\mathbf{k}$ is a unit vector pointing upward. A friction term can be added if desired. Here

$$f \equiv 2\Omega\sin\varphi \tag{17.5}$$

is the Coriolis parameter. A second, equivalent form of the shallow-water momentum equation is

$$\frac{\partial \mathbf{v}}{\partial t} + \left(\frac{\zeta + f}{h} \right) \mathbf{k} \times (h\mathbf{v}) + \nabla \left[K + g \left(h + h_T \right) \right] = 0 \, . \tag{17.6}$$

Here

$$\zeta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{v}) \tag{17.7}$$

is the relative vorticity, and

$$K \equiv \mathbf{v} \cdot \mathbf{v}/2 \tag{17.8}$$

is the kinetic energy of the horizontal wind, per unit mass. This form can be obtained using a vector identity discussed in Appendix A. In (17.6), we have multiplied and divided the vorticity term by h, for reasons that will be explained later.

We will also need the shallow-water vorticity and divergence equations. The vorticity equation can be derived by applying the operator $\mathbf{k} \cdot \nabla \times$ to each term of (17.6). The result is

$$\frac{\partial \zeta}{\partial t} + \mathbf{k} \cdot \nabla \times \left[\left(\frac{\zeta + f}{h} \right) \mathbf{k} \times (h\mathbf{v}) \right] = 0 \, . \tag{17.9}$$

Here we have used

$$(h\mathbf{v}) \cdot \left[\left(\frac{\zeta + f}{h} \right) \mathbf{k} \times (h\mathbf{v}) \right] = 0 \, , \tag{17.10}$$

which follows from the fact that the cross product of any two vectors is perpendicular to both. With the use of additional vector identities, (17.9) can be rewritten as

$$\frac{\partial}{\partial t}(\zeta + f) + \nabla \cdot [\mathbf{v}(\zeta + f)] = 0 \,, \tag{17.11}$$

where

$$\eta \equiv \zeta + f \tag{17.12}$$

is the absolute vorticity. Here we have assumed that f is independent of time. Comparison of (17.11) with (17.2) shows that the absolute vorticity $\zeta + f$ and the mass h move in the same way, i.e., *they move together*. This explains why vortices can carry tracers like smoke and dust, and why they look like little creatures moving with the fluid. The vorticity is much more interesting than the divergence.

Eq. (17.11) is equivalent to

$$\frac{\partial}{\partial t}(hq) + \nabla \cdot (h\mathbf{v}q) = 0 \,, \tag{17.13}$$

where

$$q \equiv \eta \; h \tag{17.14}$$

is the potential vorticity. As you know, conservation of potential vorticity is key to the dynamics of balanced flows. Using (17.14), we can rewrite (17.6) as

$$\frac{\partial \mathbf{v}}{\partial t} + q\mathbf{k} \times (h\mathbf{v}) + \nabla \left[K + g\,(h + h_T)\right] = 0 \,. \tag{17.15}$$

The divergence equation can be derived by applying the operator $\nabla\cdot$ to each term of (17.6). The result can be written as

$$\frac{\partial \delta}{\partial t} - q\mathbf{k} \cdot [\nabla \times (h\mathbf{v})] + \nabla^2 \left[K + g\,(h + h_T)\right] = 0 \,, \tag{17.16}$$

where

$$\delta \equiv \nabla \cdot \mathbf{v} \tag{17.17}$$

is the shallow-water divergence, and we have used identities from Appendix A. Further discussion of the divergence equation is given later in this chapter.

17.2 Energy conservation in shallow water

When we take the dot product of (17.6) with $h\mathbf{v}$, we obtain the advective form of the kinetic energy equation, i.e.,

$$h\frac{\partial K}{\partial t} + (h\mathbf{v}) \cdot \nabla \left[K + g\left(h + h_T\right)\right] = 0 \,. \tag{17.18}$$

Combining (17.18) with the continuity equation (17.2), we can obtain the flux form of the kinetic energy equation:

$$\frac{\partial}{\partial t}\left(hK\right) + \nabla \cdot \left(h\mathbf{v}K\right) + (h\mathbf{v}) \cdot \nabla \left[g\left(h + h_T\right)\right] = 0 \,. \tag{17.19}$$

The potential energy equation for shallow water can be derived by multiplying the continuity equation, (17.2), by $g\left(h_T + h\right)$:

$$g\left(h_T + h\right)\frac{\partial h}{\partial t} + g\left(h_T + h\right)\nabla \cdot \left(h\mathbf{v}\right) = 0 \,, \tag{17.20}$$

This can be rearranged to

$$\frac{\partial}{\partial t}\left[hg\left(h_T + \frac{1}{2}h\right)\right] + \nabla \cdot \left[\left(h\mathbf{v}\right)g\left(h + h_T\right)\right] - \left(h\mathbf{v}\right) \cdot \nabla \left[g\left(h + h_T\right)\right] = 0 \,, \tag{17.21}$$

which is the flux form of the potential energy equation.

Notice that h has three different jobs in the shallow water equations. It is a measure of mass, used in the continuity equation. It appears in the expression for the horizontal pressure-gradient force, used in the momentum equation. And it appears in the expression for the potential energy per unit mass.

By adding (17.19) and (17.21), we obtain conservation of total energy:

$$\frac{\partial}{\partial t}\left[hK + hg\left(h_T + \frac{1}{2}h\right)\right] + \nabla\cdot\left[h\mathbf{v}K + (h\mathbf{v})\,g\,(h+h_T)\right] = 0\,. \tag{17.22}$$

Here the energy conversion terms have cancelled.

17.3 Potential enstrophy conservation

Using the continuity equation, (17.2), we can rewrite the PV equation, (17.13), in advective form:

$$\frac{\partial q}{\partial t} + \mathbf{v}\cdot\nabla q = 0\,. \tag{17.23}$$

Multiplying each term of (17.23) by q, and using continuity to go to flux form, we obtain

$$\frac{\partial}{\partial t}\left(h\frac{q^2}{2}\right) + \nabla\cdot\left(h\mathbf{v}\frac{q^2}{2}\right) = 0\,. \tag{17.24}$$

The quantity $q^2/2$ is called the potential enstrophy. According to (17.24), the shallow-water potential enstrophy is conserved in the absence of friction.

17.4 The nondivergent barotropic vorticity equation

The nondivergent barotropic vorticity equation applies when

$$\delta = 0\,, \tag{17.25}$$

and there is no bottom topography. Here we treat (17.25) as an assumption. The continuity equation, (17.2), reduces to

$$\frac{\partial h}{\partial t} = -\mathbf{v} \cdot \nabla h \, . \tag{17.26}$$

The vorticity equation, (17.11), reduces to

$$\frac{\partial \eta}{\partial t} = -\mathbf{v} \cdot \nabla \eta \, , \tag{17.27}$$

where $\eta = \zeta = f$ is the absolute vorticity. Because the flow is nondivergent, the velocity can be computed from a stream function:

$$\mathbf{v} = \mathbf{k} \times \nabla \psi \, . \tag{17.28}$$

The stream function can be computed by solving

$$\zeta = \nabla^2 \psi \, , \tag{17.29}$$

using the methods discussed in Chapter 15.

In view of (17.25), the divergence equation, (17.16) reduces to

$$-\eta \zeta + \nabla^2 (K + gh) = 0 \, . \tag{17.30}$$

Obviously this equation is not needed (and could not be used) to determine the divergence, so what is it for? Using the methods discussed in Chapter 15, it can be solved as a diagnostic equation for the water depth, h, which appears inside the Laplacian operator. According to (17.30), in a two-dimensional nondivergent flow the spatial distribution of h is whatever is needed to maintain nondivergence over time.

The nondivergent barotropic vorticity equation will be used in Chapter 26.

17.5 Problems

1. Derive (17.2) by starting from (17.1).

Chapter 18

Conserving momentum and energy with the one-dimensional shallow-water equations

18.1 Properties of the continuous equations

Consider the one-dimensional shallow-water equations, with bottom topography, without rotation and with $v = 0$. The prognostic variables are the water depth or mass, h, and the speed, u. The equations can be written as

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \tag{18.1}$$

and

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left[K + g\left(h + h_S\right)\right] = 0. \tag{18.2}$$

Here

$$K \equiv \frac{1}{2}u^2 \tag{18.3}$$

is the kinetic energy per unit mass, g is the acceleration of gravity, and h_S is the height of the bottom topography. In Eq. (18.2), the vorticity has been assumed to vanish, which is reasonable in the absence of rotation and in one dimension. The effects of vorticity are of course absolutely critical in geophysical fluid dynamics; they will be discussed in Chapter 26.

The design of the scheme is determined by a sequence of choices. We should welcome the opportunity to make the best possible choices. The first thing that we have to choose is the particular form of the continuous equations that the space-differencing scheme is designed to mimic. Eq. (18.2) is one possible choice for the continuous form of the momentum equation. An alternative choice is

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(huu) + gh\frac{\partial}{\partial x}(h + h_S) = 0, \tag{18.4}$$

i.e., the flux form of the momentum equation, which can be derived by combining (18.1) and (18.2). The momentum per unit area is hu.

The continuous shallow-water equations have important "integral properties," which we will use as a guide in the design of our space-differencing scheme. For example, if we integrate (18.1) with respect to x, over a closed or periodic domain, we obtain

$$\frac{d}{dt}\left(\int_{\text{all } x} h dx\right) = 0, \tag{18.5}$$

which means that mass is conserved.

Using

$$h\frac{\partial h}{\partial x} = \frac{\partial}{\partial x}\left(\frac{h^2}{2}\right), \tag{18.6}$$

we can rewrite (18.4) as

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}\left(huu + g\frac{h^2}{2}\right) = -gh\frac{\partial h_S}{\partial x}. \tag{18.7}$$

If we integrate with respect to x, over a periodic domain, we obtain

$$\frac{d}{dt}\left(\int_{\text{all } x} hu\, dx\right) = -\int_{\text{all } x} gh\frac{\partial h_S}{\partial x}\, dx\,. \tag{18.8}$$

This shows that in the absence of topography, i.e., if $\partial h_S/\partial x = 0$ everywhere, the domain average of hu is invariant, i.e., momentum is conserved. When h_S is spatially variable, the atmosphere and the "solid earth" can exchange momentum through the covariance of the surface pressure, gh, with the slope of the topography, $\partial h_S/\partial x$.

The flux form of the kinetic energy equation can be derived by multiplying (18.1) by K and (18.2) by hu, and adding the results, to obtain

$$\frac{\partial}{\partial t}(hK) + \frac{\partial}{\partial x}(huK) + hu\frac{\partial}{\partial x}[g(h+h_S)] = 0\,. \tag{18.9}$$

The last term of (18.9) represents conversion between potential and kinetic energy.

The potential energy equation can be derived by multiplying (18.1) by $g(h+h_S)$ to obtain

$$\frac{\partial}{\partial t}\left[hg\left(h_S+\frac{1}{2}h\right)\right] + g(h+h_S)\frac{\partial}{\partial x}(hu) = 0\,, \tag{18.10}$$

which can be rearranged to

$$\frac{\partial}{\partial t}\left[hg\left(h_S+\frac{1}{2}h\right)\right] + \frac{\partial}{\partial x}[hug(h+h_S)] - hu\frac{\partial}{\partial x}[g(h+h_S)] = 0\,. \tag{18.11}$$

Here $g\left(h_S+\frac{1}{2}h\right)$ can be interpreted as the potential energy per unit mass of a particle that is "half-way up" in the water column. The middle term represents both advection of potential energy and the horizontal redistribution of energy by pressure-work. The last term represents conversion between kinetic and potential energy; compare with (18.9). In deriving (18.10), we have assumed that h_S is independent of time. This assumption can easily be relaxed, at the cost of an additional term in (18.11).

We pause to observe that in the shallow water system the water depth, h, plays three roles. It is proportional to the mass per unit area. It enters in the pressure-gradient force, as seen in (18.2) or (18.4). Finally, it enters ino the potential energy, as seen in (18.10).

When we add (18.9) and (18.11), the energy conversion terms cancel, and we obtain a statement of the conservation of total energy, i.e.,

$$\frac{\partial}{\partial t}\left\{h\left[K+g\left(h_S+\frac{1}{2}h\right)\right]\right\} + \frac{\partial}{\partial x}\{hu[K+g(h+h_S)]\} = 0\,. \tag{18.12}$$

Integration of (18.12) over a closed or periodic domain gives

$$\frac{d}{dt}\int_{\text{all } x} h\left[K + g\left(h_S + \frac{1}{2}h\right)\right]\, dx = 0\,, \tag{18.13}$$

which shows that the domain-integrated total energy is conserved.

18.2 The spatially discrete case

Now consider finite-difference approximations to (18.1) and (18.2). We keep the time derivatives continuous, and explore the effects of space differencing only. We use a staggered grid, with h defined at integer points (hereafter called mass points) and u at half-integer points (hereafter called wind points). This can be viewed as a one-dimensional version of the C grid. The grid spacing, Δx, is assumed to be uniform. Our selection of this particular grid is a second *choice* made in the design of the space-differencing scheme.

The finite difference version of the continuity equation is

$$\frac{dh_i}{dt} + \left[\frac{(hu)_{i+1/2} - (hu)_{i-1/2}}{\Delta x}\right] = 0\,. \tag{18.14}$$

It should be understood that

$$h_{i+1/2}u_{i+1/2} \equiv (hu)_{i+1/2}\ . \tag{18.15}$$

The "wind-point masses," e.g., $h_{i+1/2}$, are undefined at this stage, but of course we will have to settle on a way to define them before we can actually use the scheme. The finite-difference approximation used in (18.14) is consistent with second-order accuracy in space, although we cannot really determine the order of accuracy until the finite-difference form of the mass flux has been specified. We have already discussed how the "flux form" used in (18.14) makes it possible for the model to conserve mass, i.e.,

$$\frac{d}{dt}\left(\sum_{\text{all } x} h_i\right) = 0, \tag{18.16}$$

and this is true regardless of how $h_{i+1/2}$ is defined. Eq. (18.16) is analogous to (18.5).

A finite-difference momentum equation that is modeled after (18.2) is

$$\frac{du_{i+1/2}}{dt} + \left(\frac{K_{i+1} - K_i}{\Delta x}\right) + g\left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x}\right] = 0. \tag{18.17}$$

The kinetic energy per unit mass, K_i, is undefined at this stage, but its integer subscript shows that it resides at mass points. The finite-difference approximations used in (18.17) are consistent with second-order accuracy in space, although we cannot really determine the order of accuracy until the finite-difference forms of the mass flux and kinetic energy are specified. Although Eq. (18.17) is not in a "flux form," we might (or might not) be able to use the continuity equation to show that it is consistent with (i.e., can be derived from) a flux form. This possibility is discussed further below.

Multiply (18.17) by $h_{i+1/2}$ to obtain

$$h_{i+1/2}\frac{du_{i+1/2}}{dt} + h_{i+1/2}\left(\frac{K_{i+1} - K_i}{\Delta x}\right) + gh_{i+1/2}\left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x}\right] = 0. \tag{18.18}$$

In order to mimic the differential relationship (18.6), we must require that

$$h_{i+1/2}\left(\frac{h_{i+1} - h_i}{\Delta x}\right) = \left(\frac{h^2{}_{i+1} - h_i{}^2}{2\Delta x}\right), \tag{18.19}$$

which can by satisfied by choosing

$$h_{i+1/2} = \frac{h_{i+1} + h_i}{2}. \tag{18.20}$$

This choice is required to ensure that the pressure-gradient force does not prod uce any net source or sink of momentum in the absence of topography. In view of (18.20), we can write

$$(hu)_{i+1/2} = \left(\frac{h_{i+1} + h_i}{2}\right) u_{i+1/2}. \tag{18.21}$$

Combining (18.20) with the continuity equation (18.14), we see that we can write a *continuity equation for the wind points*, as follows:

$$\frac{dh_{i+1/2}}{dt} + \frac{1}{2\Delta x}\left[(hu)_{i+3/2} - (hu)_{i-1/2}\right] = 0. \tag{18.22}$$

It should be clear from the form of (18.22) that the "wind-point mass" is actually conserved by the model. Of course, we do not actually use (18.22) when we integrate the model; instead we use (18.14). Nevertheless, (18.22) will be satisfied, because it can be derived from (18.14) and (18.20). An alternative form of (18.22) is

$$\frac{dh_{i+1/2}}{dt} + \frac{1}{\Delta x}\left[(hu)_{i+1} - (hu)_i\right] = 0, \tag{18.23}$$

where

$$(hu)_{i+1} \equiv \frac{1}{2}\left[(hu)_{i+3/2} + (hu)_{i+1/2}\right] \text{ and } (hu)_i \equiv \frac{1}{2}\left[(hu)_{i+1/2} + (hu)_{i-1/2}\right]. \tag{18.24}$$

Now add (18.18) and $u_{i+1/2}$ times (18.23), and use (18.20), to obtain what "should be" the flux form of the momentum equation, analogous to (18.4):

$$\begin{aligned}\frac{d}{dt}\left(h_{i+1/2}u_{i+1/2}\right) + h_{i+1/2}\left(\frac{K_{i+1} - K_i}{\Delta x}\right) + \frac{u_{i+1/2}\left[(hu)_{i+1} - (hu)_i\right]}{\Delta x} + g\left(\frac{h_{i+1}^2 - h_i^2}{2\Delta x}\right) \\ = -gh_{i+1/2}\left[\frac{(h_S)_{i+1} - (h_S)_i}{\Delta x}\right].\end{aligned} \tag{18.25}$$

Is this really a flux form, or not? The answer is: It depends on how we define K_i. Suppose that K_i is defined by

$$K_i \equiv \frac{1}{2}u_{i+1/2}u_{i-1/2}. \tag{18.26}$$

Other possible definitions of K_i will be discussed later. Using (18.26) and (18.24), we can write

$$\begin{aligned} & h_{i+1/2}\left(\frac{K_{i+1}-K_i}{\Delta x}\right)+u_{i+1/2}\frac{1}{\Delta x}\left[(hu)_{i+1}-(hu)_i\right] \\ &= \frac{1}{2\Delta x}\left\{h_{i+1/2}\left(u_{i+3/2}u_{i+1/2}-u_{i+1/2}u_{i-1/2}\right)+u_{i+1/2}\left[(hu)_{i+3/2}-(hu)_{i-1/2}\right]\right\} \\ &= \frac{1}{\Delta x}\left[\left(\frac{h_{i+1/2}+h_{i+3/2}}{2}\right)u_{i+3/2}u_{i+1/2}-\left(\frac{h_{i+1/2}+h_{i-1/2}}{2}\right)u_{i-1/2}u_{i+1/2}\right]. \end{aligned} \tag{18.27}$$

This is, indeed, a finite-difference flux divergence. The momentum flux at the point i is

$$\frac{1}{2}\left(h_{i+1/2}+h_{i-1/2}\right)u_{i-1/2}u_{i+1/2}\,,$$

and the momentum flux at the point $i+1$ is

$$\frac{1}{2}\left(h_{i+1/2}+h_{i+3/2}\right)u_{i+3/2}u_{i+1/2}\,.$$

Because (18.27) is a finite-difference flux divergence, momentum will be conserved by the scheme *if we define the kinetic energy by* (18.26).

Note, however, that there are two big problems with (18.26). First, when u is dominated by the $2\Delta x$-mode, (18.26) will give a negative value of K_i, which is unphysical. The second big problem is that when u is dominated by the $2\Delta x$-mode, the momentum flux will always be negative, i.e., momentum will always be transferred in the $-x$ direction, assuming that the interpolated masses that appear in the momentum fluxes are positive. These problems are severe enough that the definition of kinetic energy given by (18.26) is unacceptable.

OK, that's not good, but let's see what happens with the kinetic energy equation. For this purpose, we return to the general form of K_i; Eq. (18.26) will not be used. Recall that the kinetic energy is defined at mass points. To begin the derivation, multiply (18.17) by $(hu)_{i+1/2}$ to obtain

$$(hu)_{i+1/2}\frac{du_{i+1/2}}{dt}+(hu)_{i+1/2}\left(\frac{K_{i+1}-K_i}{\Delta x}\right)+g(hu)_{i+1/2}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]=0\,. \tag{18.28}$$

Rewrite (18.28) for grid point $i-\frac{1}{2}$, simply by subtracting one from each subscript:

$$(hu)_{i-1/2}\frac{du_{i-1/2}}{dt}+(hu)_{i-1/2}\left(\frac{K_i-K_{i-1}}{\Delta x}\right)+g\,(hu)_{i-1/2}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]=0\,. \tag{18.29}$$

Now add (18.28) and (18.29), and multiply the result by $\frac{1}{2}$ to obtain the arithmetic mean:

$$\begin{aligned}&\frac{1}{2}\left[(hu)_{i+1/2}\frac{du_{i+1/2}}{dt}+(hu)_{i-1/2}\frac{du_{i-1/2}}{dt}\right]\\&+\frac{1}{2}\left[(hu)_{i+1/2}\left(\frac{K_{i+1}-K_i}{\Delta x}\right)+(hu)_{i-1/2}\left(\frac{K_i-K_{i-1}}{\Delta x}\right)\right]\\&+\frac{g}{2}\left\{(hu)_{i+1/2}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-1/2}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}=0\,.\end{aligned} \tag{18.30}$$

You should be able to see that this is an advective form of the kinetic energy equation.

Next we try to derive, from (18.30) and (18.14), a *flux form* of the kinetic energy equation. If we can do that, then kinetic energy will be conserved. Begin by multiplying (18.14) by K_i:

$$K_i\left\{\frac{dh_i}{dt}+\left[\frac{(hu)_{i+1/2}-(hu)_{i-1/2}}{\Delta x}\right]\right\}=0\,. \tag{18.31}$$

Keep in mind that we still do not know what K_i is; we have just multiplied the continuity equation by a mystery variable. Add (18.31) and (18.30) to obtain

$$\begin{aligned}&K_i\frac{dh_i}{dt}+\frac{1}{2}\left[(hu)_{i+1/2}\frac{du_{i+1/2}}{dt}+(hu)_{i-1/2}\frac{du_{i-1/2}}{dt}\right]\\&+\left\{\frac{(hu)_{i+1/2}}{\Delta x}\left[K_i+\frac{1}{2}\left(K_{i+1}-K_i\right)\right]-\frac{(hu)_{i-1/2}}{\Delta x}\left[K_i-\frac{1}{2}\left(K_i-K_{i-1}\right)\right]\right\}\\&+\frac{g}{2}\left\{(hu)_{i+1/2}\frac{\left[(h+h_S)_{i+1}-(h+h_S)_i\right]}{\Delta x}+(hu)_{i-1/2}\frac{\left[(h+h_S)_i-(h+h_S)_{i-1}\right]}{\Delta x}\right\}=0\,.\end{aligned} \tag{18.32}$$

Eq. (18.32) "should be" a flux form of the kinetic energy equation. Is it really? To answer this question, we analyze the various terms of (18.32) one by one.

The advection terms on the second line of (18.32) are very easy to deal with. They can be rearranged to

$$\frac{1}{\Delta x}\left[(hu)_{i+1/2}\frac{1}{2}(K_{i+1}+K_i)-(hu)_{i-1/2}\frac{1}{2}(K_i+K_{i-1})\right]. \tag{18.33}$$

This has the form of a "finite-difference flux divergence." The conclusion is that these terms are consistent with kinetic energy conservation under advection, simply by virtue of their form, regardless of the method chosen to determine K_i.

Next, consider the energy conversion terms on the third line of (18.32), i.e.,

$$\frac{g}{2}\left\{(hu)_{i+1/2}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-1/2}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}. \tag{18.34}$$

We want to compare these terms with the corresponding terms of the finite-difference form of the potential energy equation, which can be derived by multiplying (18.14) by $g(h+h_S)_i$:

$$\frac{d}{dt}\left[h_i g\left(h_S+\frac{1}{2}h\right)_i\right]+g(h+h_S)_i\left[\frac{(hu)_{i+1/2}-(hu)_{i-1/2}}{\Delta x}\right]=0. \tag{18.35}$$

Eq. (18.35) is analogous to (18.10). We want to recast (18.35) so that we see advection of potential energy, as well as the energy conversion term corresponding to (18.34); compare with (18.11). To accomplish this, we "put in" the energy conversion term by hand, and write the advection term symbolically, like this:

$$\begin{aligned}&\frac{d}{dt}\left[h_i g\left(h_S+\frac{1}{2}h\right)_i\right]+ADV_i\\&-\frac{g}{2}\left\{(hu)_{i+1/2}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-1/2}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}=0.\end{aligned} \tag{18.36}$$

Here "ADV_i" represents the advection of potential energy, which we hope will be in flux form. At this point, we do not have an expression for ADV_i, but we will let the equations tell us how to compute it. The second line of (18.36) is a copy of the energy conversion terms of (18.32), but with the sign reversed. We require that (18.36) be equivalent to (18.35), and ask what form of ADV_i is implied by this requirement. The answer is:

$$ADV_i = \frac{g}{2\Delta x}\left\{(hu)_{i+1/2}\left[(h+h_S)_{i+1}+(h+h_S)_i\right]-(hu)_{i-1/2}\left[(h+h_S)_i+(h+h_S)_{i-1}\right]\right\}. \tag{18.37}$$

This does indeed have the form of a finite-difference flux divergence, as desired.

What we have shown up to this point is that the conservation of potential energy and the cancellation of the energy conversion terms are pretty easy. The form of K_i is still up to us.

We are not quite finished, however, because we have not yet examined the time-rate-of-change terms of (18.32). Obviously, the first line of (18.32) must be analogous to $\partial/\partial t\,(hK)$. For convenience, we define

$$(\text{KE tendency})_i \equiv K_i\frac{dh_i}{dt}+\frac{1}{2}\left[(hu)_{i+1/2}\frac{d}{dt}u_{i+1/2}+(hu)_{i-1/2}\frac{d}{dt}u_{i-1/2}\right]. \tag{18.38}$$

Substituting for the mass fluxes from (18.21), we find that (18.38) can be rewritten as

$$(\text{KE tendency})_i \equiv K_i\frac{dh_i}{dt}+\frac{1}{8}\left[(h_{i+1}+h_i)\frac{d}{dt}\left(u_{i+1/2}^2\right)+(h_i+h_{i-1})\frac{d}{dt}\left(u_{i-1/2}^2\right)\right]. \tag{18.39}$$

The requirement for kinetic energy conservation is

$$\sum_{\text{all } x}(\text{KE tendency})_i = \sum_{\text{all } x}\frac{d}{dt}\left(h_iK_i\right). \tag{18.40}$$

Note that *only the sums over i must agree*; it is not necessary that

$$K_i \frac{dh_i}{dt} + \frac{1}{8}\left[(h_{i+1}+h_i)\frac{d}{dt}\left(u^2_{i+1/2}\right) + (h_i+h_{i-1})\frac{d}{dt}\left(u^2_{i-1/2}\right)\right]$$

be equal to $d/dt\,(h_i K_i)$ for each i. To complete our check of kinetic energy conservation, we substitute for K_i on the right-hand side of (18.40), and check to see whether the resulting equation is actually satisfied.

The bad news is that, if we use (18.26), Eq. (18.40) is not satisfied. This means that, when we start from the continuous form of (18.2), we cannot have both momentum conservation under advection and kinetic energy conservation. On the other hand, we didn't like (18.26) anyway, because for the $2\Delta x$ wave it gives negative kinetic energy and a momentum flux that is always in the negative x direction.

The good news is that there are ways to satisfy (18.40). Two alternative definitions of the kinetic energy are

$$K_i \equiv \frac{1}{4}\left(u^2{}_{i+1/2} + u^2_{i-1/2}\right), \tag{18.41}$$

$$h_i K_i \equiv \frac{1}{4}\left(h_{i+1/2}u^2{}_{i+1/2} + h_{i-1/2}u^2_{i-1/2}\right). \tag{18.42}$$

With either of these definitions, K_i cannot be negative. We can show that the sum over the domain of $h_i K_i$ given by (18.41) is equal to the sum over the domain of $h_i K_i$ given by (18.42). Either choice allows (18.40) to be satisfied, so both are consistent with kinetic energy conservation under advection. On the other hand, neither is consistent with momentum conservation under advection.

To sum up: When we start from the continuous form of (18.2), we can have either momentum conservation under advection or kinetic energy conservation under advection, but not both. Which is better depends on the application.

An alternative approach is to start from a finite-difference form of the momentum equation that mimics (18.4). With that approach, it is possible to conserve *both* momentum under advection and kinetic energy under advection. You are invited to demonstrate this in a problem at the end of this chapter.

When we generalize to the two-dimensional shallow-water equations with rotation, there are very important additional considerations having to do with vorticity, and the issues discussed here have to be revisited. This is discussed in a later chapter.

18.3 Summary

We have explored the conservation properties of spatial finite-difference approximations of the momentum and continuity equations for one-dimensional non-rotating flow, using a staggered grid. We were able to find a scheme that guarantees conservation of mass, conservation of momentum in the absence of bottom topography, conservation of kinetic energy under advection, conservation of potential energy under advection, and conservation of total energy in the presence of energy conversion terms.

This chapter has introd uced several new things. This is the first time that we have considered the full shallow water equations, including their nonlinear terms. This is the first time that we have discussed energy conversions and total energy conservation. In addition, the chapter illustrates a way of thinking about the trade-offs that must be weighed in the design of a scheme, as various alternative choices each have advantages and disadvantages.

18.4 Problems

1. Show that if we use (18.26) it is not possible to conserve kinetic energy under advection.
2. Starting from a finite-difference form that mimics (18.4), show that it is possible to conserve both momentum and total energy. Use the C grid, and keep the time derivatives continuous.

Chapter 19

Making waves

19.1 Inertia-gravity waves in shallow water

For the special case of small-amplitude gravity waves on a resting basic state, with no topography, the linearized versions of Eqs. (17.2) and (17.4) are

$$\frac{\partial h}{\partial t} + H\nabla\cdot\mathbf{v} = 0\,, \tag{19.1}$$

and

$$\frac{\partial \mathbf{v}}{\partial t} + f\mathbf{k}\times\mathbf{v} + g\nabla h = 0\,. \tag{19.2}$$

Here H is the mean depth of the shallow water. These equations have the steady, geostrophically balanced solution

$$\nabla\cdot\mathbf{v} = 0\,, \tag{19.3}$$

$$f\mathbf{k}\times\mathbf{v} + g\nabla h = 0\,. \tag{19.4}$$

From (17.9) and (17.16) we can obtain the linearized shallow-water vorticity and divergence equations:

$$\frac{\partial \zeta}{\partial t} + f\delta = 0 \,, \tag{19.5}$$

$$\frac{\partial \delta}{\partial t} - f\zeta + g\nabla^2 h = 0 \,, \tag{19.6}$$

Rewriting (19.1) as

$$\frac{\partial h}{\partial t} + H\delta = 0 \,, \tag{19.7}$$

we see that Eqs. (19.5), (19.6) and (19.7) form a closed system for the three unknowns h, ζ, and δ. Of these three equations, only (19.6) contains spatial derivatives, in the form of our friend the Laplacian. We can write

$$\nabla^2 h = -k_{\text{tot}}^2 \widehat{H} \,, \tag{19.8}$$

where k_{tot} is the total (two-dimensional) wave number and $\widehat{H}$ is the time-dependent amplitude of h. Similarly, the time derivatives can be written as

$$\frac{\partial}{\partial t}(\,) = i\sigma(\,) \,, \tag{19.9}$$

where σ is the frequency. Then Eqs. (19.5), (19.6) and (19.7) reduce to a system of algebraic equations, which can be written as:

$$i\sigma\widehat{\zeta} + f\widehat{\delta} = 0 \,, \tag{19.10}$$

$$i\sigma\widehat{\delta} - f\zeta - gk_{\text{tot}}^2\widehat{H} = 0 \,, \tag{19.11}$$

$$i\sigma\widehat{H} + H\widehat{\delta} = 0\,. \tag{19.12}$$

Setting the determinant of the coefficients to zero, we obtain the following dispersion equation:

$$\sigma\left(-\sigma^2 + gHk_{\rm tot}^2 + f^2\right) = 0\,. \tag{19.13}$$

The simple case

$$\sigma = 0 \tag{19.14}$$

corresponds to the geostrophically balanced steady solution discussed above. The two remaining solutions are given by

$$\sigma^2 = c_{\rm gw}^2 k_{\rm tot}^2 + f^2\,, \tag{19.15}$$

where

$$c_{\rm gw} \equiv \sqrt{gH} \tag{19.16}$$

is the phase speed of a pure gravity wave (i.e., a gravity wave that does not feel the effects of rotation). Note that $c_{\rm gw}$ is independent of wave number, which means that pure gravity waves are nondispersive.

The solutions given by (19.18) represent two "inertia-gravity waves," i.e., gravity waves modified by rotation. The frequencies of the two waves depend on the total wave number, which means that they are dispersive. Eq. (19.15) shows that one effect of rotation is to reduce the frequency of the waves. The phase speed of the waves, which is called c, is given by

$$\begin{aligned} c^2 &\equiv (\sigma/k_{\rm tot})^2 \\ &= c_{\rm gw}^2 + f^2/k_{\rm tot}^2\,. \end{aligned} \tag{19.17}$$

One effect of rotation is to cause the phase speed to depend on wave number, i.e., it makes the waves dispersive.

It is convenient to rewrite (19.15) as

$$(\sigma/f)^2 = 1 + r_{\text{def}}^2 k_{\text{tot}}^2 \,, \tag{19.18}$$

where

$$r_{\text{def}} \equiv c_{\text{gw}}/f \tag{19.19}$$

is the radius of deformation, which can be interpreted as the distance that a pure gravity wave propagates in one inertial period.

19.2 Red and black

For the special case of a one-dimensional, small-amplitude gravity wave on a resting basic state, with no rotation or topography, Eqs. (19.1) and (19.2) become

$$\frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} = 0 \,, \tag{19.20}$$

and

$$\frac{\partial h}{\partial t} + H\frac{\partial u}{\partial x} = 0, \tag{19.21}$$

respectively. By combining (19.20) - (19.21) we can show that

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{19.22}$$

and

$$\frac{\partial^2 h}{\partial t^2} = c^2 \frac{\partial^2 h}{\partial x^2}, \tag{19.23}$$

which are both examples of "the wave equation."

Assuming solutions of the form $e^{i(kx \pm \sigma t)}$, and substituting into either (19.22) or (19.23), we obtain the dispersion equation

$$\sigma^2 = c^2 k^2 \, . \tag{19.24}$$

There are two waves, one propagating in the positive x-direction, and the other in the negative x-direction.

Now consider the differential-difference equations

$$\frac{du_j}{dt} + g\left(\frac{h_{j+1} - h_{j-1}}{2d}\right) = 0 \, , \tag{19.25}$$

$$\frac{dh_j}{dt} + H\left(\frac{u_{j+1} - u_{j-1}}{2d}\right) = 0 \tag{19.26}$$

where d is the grid spacing. These are, of course, differential-difference analogs of the one-dimensional shallow water equations, (19.20) - (19.21). We are temporarily neglecting rotation. We keep the time derivatives continuous here because the issues that we are going discuss next have to do with space differencing only. Consider a distribution of the dependent variables on the grid as shown in Fig. 19.1. Notice that according (19.25) and (19.26) *the set of red quantities will act completely independently of the set of black quantities*, if there are no boundaries. With cyclic boundary conditions, this is still true if the number of grid points in the cyclic domain is even. What this means is that we have two families of waves on the grid: "red" waves that propagate both left and right, and "black" waves that propagate both left and right. Physically there should only be one family of waves.

A good way to think about this situation is that we have two non-interacting models living on the same grid: a red model and a black model. That's a problem. The red model may think it's winter, while the black model thinks it's summer. In such a case we will have tremendous noise at the grid scale.

The two models are noninteracting so long as they are linear. If we include nonlinear terms, then interactions can occur, but that doesn't mean that the nonlinear terms solve the problem.

Since we have two models, we have two solutions. One of the solutions is physical, and the other is computational. In Chapter 6, we encountered computational modes in time, which arise from the time-differencing scheme. Here, for the first time, we are meeting computational modes in space, which arise from the space differencing.

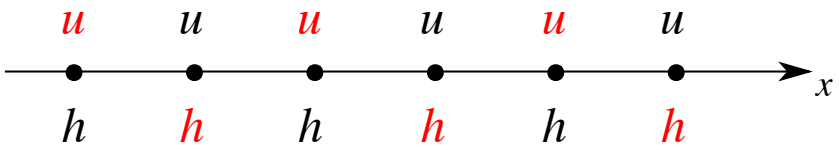


Figure 19.1: A-grid for solution of the one-dimensional shallow water equations.

Here is a mathematical way to draw the same conclusion. The wave solutions of (19.25) and (19.26) are

$$\left(u_j, h_j\right) \sim e^{i(kjd-\sigma t)}, \tag{19.27}$$

giving

$$\begin{aligned} \sigma u_j - g h_j \frac{\sin(kd)}{d} &= 0, \\ \sigma h_j - H u_j \frac{\sin(kd)}{d} &= 0\,. \end{aligned} \tag{19.28}$$

Provided that u_j and h_j are not both identically zero, we obtain the dispersion relation

$$\sigma^2 = k^2 g H \mathrm{sinc}^2(kd)\ . \tag{19.29}$$

This shows that the square of the finite-difference phase speed is

$$c^2 = g H \mathrm{sinc}^2(kd)\ . \tag{19.30}$$

In the exact solution, the phase speed $c = \pm\sqrt{gH}$ is independent of wave number, but the finite-difference phase speed depends on wave number, is generally less than the true phase speed, and is zero for the shortest wave that fits on the grid. This is computational dispersion again, but for wave propagation rather than advection.

Suppose that σ is given. We assume that $\sigma \geq 0$, so that the direction of propagation is determined by the sign of the wave number, k, and we allow $-\pi \leq kd \leq \pi$.

Eq. (19.29) is quadratic, so at first glance you may think that it describes two solutions, but it actually has *four* solutions! To see why, let's start by defining $p \equiv kd$. Note that $-\pi \leq p \leq \pi$. If $p = p_0$ satisfies (19.29), then $p = -p_0$, $p = \pi - p_0$ and $p = -(\pi - p_0)$ also satisfy it. *The two "extra" solutions are computational modes in space*. They come from the redundancy of the unstaggered grid. Earlier we encountered troublesome computational modes in time. Now we find that computational modes in space can also cause problems.

The two solutions $p = p_0$ and $p = -p_0$ are approximations to the true solution, and so could be considered as physical, while the other two, $p = \pi - p_0$ and $p = -(\pi - p_0)$, could be considered as computational. This distinction is less meaningful than in the case of the advection equation, however. For advection, the envelope of a computational mode moves toward the downstream direction. For the wave equation, there is no "downstream" direction. Although upstream-weighted schemes are best for advection, centered schemes are best for wave propagation.

For a given value of σ, the general solution for u_j is a linear combination of the four modes, and can be written as

$$u_j = \left[Ae^{ip_0 j} + Be^{-ip_0 j} + Ce^{i(\pi - p_0)j} + De^{-i(\pi - p_0)j} \right] e^{-i\sigma t} \,. \tag{19.31}$$

By substituting (19.31) into the second of (19.28), we find that h_j satisfies

$$h_j = \frac{H \sin p_0}{\sigma \Delta x} \left[Ae^{ip_0 j} - Be^{-ip_0 j} + Ce^{i(\pi - p_0)j} - De^{-i(\pi - p_0)j} \right] e^{-i\sigma t} \,. \tag{19.32}$$

Remember that we are assuming $\sigma \geq 0$, so that $\sin(p_0) = \sigma d / \sqrt{gH}$ [see (19.29)]. Then (19.32) reduces to

$$h_j = \sqrt{\frac{H}{g}} \left[Ae^{ip_0 j} - Be^{-ip_0 j} + Ce^{i(\pi - p_0)j} - De^{-i(\pi - p_0)} \right] e^{-i\sigma t} \,. \tag{19.33}$$

We now repeat the analysis for the case in which the red variables are "erased" in Fig. 19.1. The governing equations can be written as

$$\frac{du_{j+1/2}}{dt} + g\left(\frac{h_{j+1} - h_j}{d}\right) = 0\,, \tag{19.34}$$

$$\frac{dh_j}{dt} + H\left(\frac{u_{j+1/2} - u_{j-1/2}}{d}\right) = 0\,. \tag{19.35}$$

Here we use half-integer subscripts for the wind points, and integer subscripts for the mass points. The solutions of (19.34) and (19.35) are assumed to have the form

$$\begin{aligned} u_{j+1/2} &= u_0 e^{i[k(j+1/2)d - \sigma t]}, \\ h_j &= h_0 e^{i(kjd - \sigma t)}\,. \end{aligned} \tag{19.36}$$

Substitution into (19.34) and (19.35) gives

$$\begin{aligned} -\sigma d u_0 + 2g h_0 \sin(kd/2) &= 0, \\ -\sigma d h_0 + 2H u_0 \sin(kd/2) &= 0\,. \end{aligned} \tag{19.37}$$

The resulting dispersion equation is

$$\sigma^2 = k^2 g H \,\mathrm{sinc}^2(kd/2)\,. \tag{19.38}$$

Compare with (19.29), which applies when both the red and black variables are included in the model. In (19.38), there are only two solutions of (19.38) for a given value of σ, because the range of $kd/2$ is $-\pi/2$ and $\pi/2$, rather than $-\pi$ to π. *This demonstrates that omitting the red variables eliminates the spurious computational modes.*

For a given σ, the solution for $u_{j+\frac{1}{2}}$ is a linear combination of the two modes, and can be written as

$$u_{j+\frac{1}{2}} = \left[A e^{ip_0\left(j+\frac{1}{2}\right)} + B e^{-ip_0\left(j+\frac{1}{2}\right)}\right] e^{-i\sigma t}\,. \tag{19.39}$$

By substituting (19.39) into (19.35), we find that h_j satisfies

$$h_j = -\frac{H}{\sigma d}\left(Ae^{ip_0 j} - Be^{-ip_0 j}\right) 2\sin\left(\frac{p_0}{2}\right) e^{-i\sigma t} \ . \tag{19.40}$$

Since we are assuming $\sigma \geq 0$, so that $2\sin(p_0/2) = \sigma d/\sqrt{gH}$, (19.40) reduces to

$$h_j = -\sqrt{H/g}\left(Ae^{ip_0 j} - Be^{-ip_0 j}\right) e^{-i\sigma t} \ . \tag{19.41}$$

19.3 Inertia-gravity waves on two-dimensional staggered grids

19.3.1 The continuous equations

Consider the shallow water equations linearized about a resting basic state, in the following continuous form:

$$\frac{\partial u}{\partial t} - fv + g\frac{\partial h}{\partial x} = 0, \tag{19.42}$$

$$\frac{\partial v}{\partial t} + fu + g\frac{\partial h}{\partial y} = 0, \tag{19.43}$$

$$\frac{\partial h}{\partial t} + H\delta = 0. \tag{19.44}$$

Here H is the constant depth of the shallow water in the basic state,

$$\delta \equiv \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \tag{19.45}$$

is the divergence, and all other symbols have their conventional meanings.

From (19.42) - (19.44), we can derive the dispersion relation

$$\left(\frac{\sigma}{f}\right)^2 = 1 + r_{\text{def}}^2\left(k^2 + l^2\right) \ . \tag{19.46}$$

Here σ is the frequency, and k and l are the wave numbers in the x and y directions, respectively. The frequency and group speed increase monotonically with wave number and are non-zero for all wave numbers. As discussed by AL, these characteristics of (19.46) are important for the geostrophic adjustment process.

19.3.2 Staggering the wind components

Winninghoff (1968) and Arakawa and Lamb (1977) (hereafter AL) discussed the extent to which finite-difference approximations to the shallow water equations can simulate the process of geostrophic adjustment, in which the dispersion of inertia-gravity waves leads to the establishment of a geostrophic balance, as the energy density of the inertia gravity waves decreases with time due to their dispersive phase speeds and non-zero group velocity. These authors considered the momentum and mass conservation equations, and defined five different staggered grids for the velocity components and mass.

In their discussion of various numerical representations of (19.35) - (19.37), AL defined five-grids denoted by "A" through "E," as shown in Fig. 19.2. For all of the grids, the mass variable, h, is defined at the centers of the grid cells. The grids differ in the placement of the velocity components, u and v. Fig. 19.2 also shows the Z-grid, which will be discussed later. AL also gave the simplest centered finite-difference approximations to of (19.35) - (19.37), for each of the five gridsgrids A - E; these equations are fairly obvious and will not be repeated here. The two-dimensional dispersion equations for the various schemes were derived but not published by AL; they are included in Fig. 19.3, which also gives a plot of the nondimensional frequency, σ/f, as a function of kd and ld, for the special case $r_{\text{def}}/d = 2$, where d is the grid spacing, assumed to be the same in the x and y directions. The particular choice $r_{\text{def}}/d = 2$ means that the radius of deformation is twice as large as the grid spacing, so that the radius of deformation is at least crudly resolved. The significance of the choice $r_{\text{def}}/d = 2$ is discussed later.

Let's start by considering grids A - E, one by one:

- *The A-grid* is arguably the simplest, because it is unstaggered. For example, the Coriolis terms of the momentum equations are easily evaluated, since u and v are defined at the same points. The A-grid has a serious problem, however. In the sketch of the A-grid in Fig. 19.2, some of the variables are colored red, some blue, some orange, and some black. Each of these colors defines a C-grid, which will be discussed below. The red winds are used to predict the red masses, and the red masses are used to compute the red winds. Similarly, the blue winds interact with the blue masses,

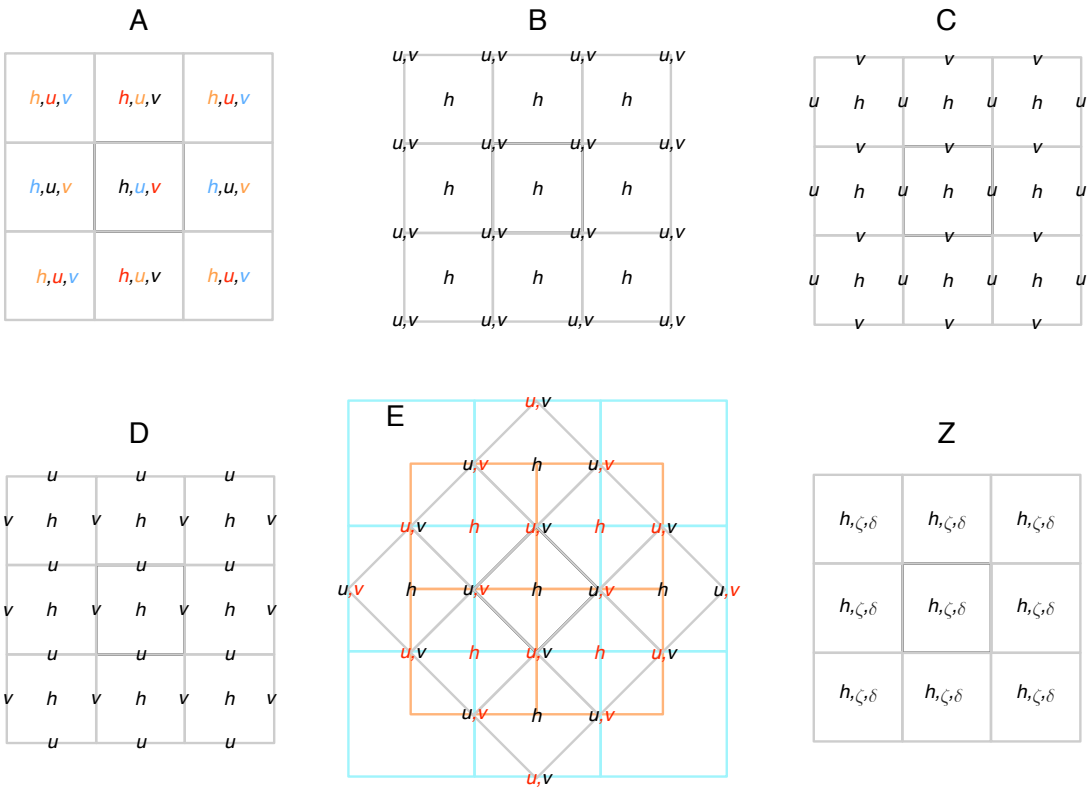


Figure 19.2: Grids A-E and Z, on a square mesh. The A-grid is equivalent to four shifted C-grids. The E-grid is equivalent to two shifted C-grids, which live in the overlapping blue and orange boxes. The E-grid can also be interpreted as a B-grid that has been rotated by 45°, but note that the directions of the wind components are *not rotated*. In the sketch of the E-grid, the mass variables can be considered to live in the rotated gray boxes.

the orange winds interact with the orange masses, and the black winds interact with the black masses. The only "mixing" of colors occurs with the Coriolis terms. As you probably know the Coriolis terms have little effect on small-scale motions. If we ignore the Coriolis terms, the the A-grid is home to "four independent models," which can do four independent things, even though they occupy the same space. In such a case, the pattern of the variables on the A-grid is characterized by strong noise at the smallest scales, i.e., a checkerboard pattern. As a result, the high-wavenumber behavior of a model based on the A-grid is poor. This problem with the A-grid is closely analogous to that of the unstaggered one-dimensional grid discussed above.

The plots in Fig. 19.3 show how the nondimensional frequency varies out to $kd = \pi$ and $ld = \pi$; recall that these wave numbers correspond to the shortest waves that can be represented on the grid. The top "row" of the figure shows the continous case. The plot of the dispersion equation for the A-grid, shown in Fig. 19.3, indicates a maximum of the frequency (group speed equal to zero) for some combinations of k and l. As a result, solutions on the A-grid are extremely noisy in practice and must

be smoothed, e.g., through artificial diffusion or filtering (Kalnay-Rivas et al., 1977). Because of this well known problem, the A-grid is rarely used today. An example of a current A-grid model is NICAM (Satoh et al., 2008). It will be discussed later.

We conclude that the existence of "multipl models"can lead to computational modes.

Thinking of the formula for the vorticity in Cartesian coordinates, (19.47), $\partial/v/\partial x$ lives on the east-west cell walls of the A-grid, while $\partial u/\partial y$ lives on the north-south walls. This means that the vorticity does not have a natural home on the A-grid. It can only be defined by averaging. Similarly, the divergence does not have a natural home, and can only be defined by averaging.

- *The B-grid* puts both velocity components on the corners of the grid cells. Fig. 19.2 shows that the velocity vectors are defined at the corners of the mass cells. The velocity components, i.e., u and v, point along the directions of the walls that intersect at the corners.

  The Coriolis terms are easily evaluated, without averaging, since u and v are defined at the same points. The averaging used to approximate the x-component of the pressure-gradient force is in the *y-direction*. Therefore, an oscillation in the x-direction, on the smallest represented scale, is not averaged out in the computation of $\partial h/\partial x$; it can, therefore, participate in the model's dynamics, and so is subject to geostrophic adjustment. Similarly, an oscillation in the y-direction, on the smallest represented scale, is not averaged out in the computation of $\partial h/\partial y$. A similar conclusion holds for the convergence / divergence terms of the continuity equation.

  Nevertheless, the averaging does do some harm, as is apparent in the plot of the B-grid dispersion equation, as shown in Fig. 19.3. The frequency does not increase monotonically with total wave number; for certain combinations of k and l, the group speed is zero. AL concluded that the B-grid gives a fairly good simulation of geostrophic adjustment, but with some tendency to small-scale noise.

  We conclude that averaging can lead to computational modes.

  On the B-grid, $\partial/v/\partial x$ lives on the north-south cell walls, while $\partial u/\partial y$ lives on the east-west walls. Again, the vorticity does not have a natural home. As with the A-grid, the divergence also does not have a natural home on the B-grid.

- *The C-grid* makes it easy to compute pressure gradient terms, without averaging, because h is defined east and west of u points, and north and south of v points. Similarly, the mass convergence / divergence terms of the continuity equation can be evaluated without averaging the winds. The divergence has a natural home, and, even better, it coincides with h points. On the other hand, averaging is needed to obtain the Coriolis terms, since u and v are defined at different points. This averaging can lead to noise in the wind field.

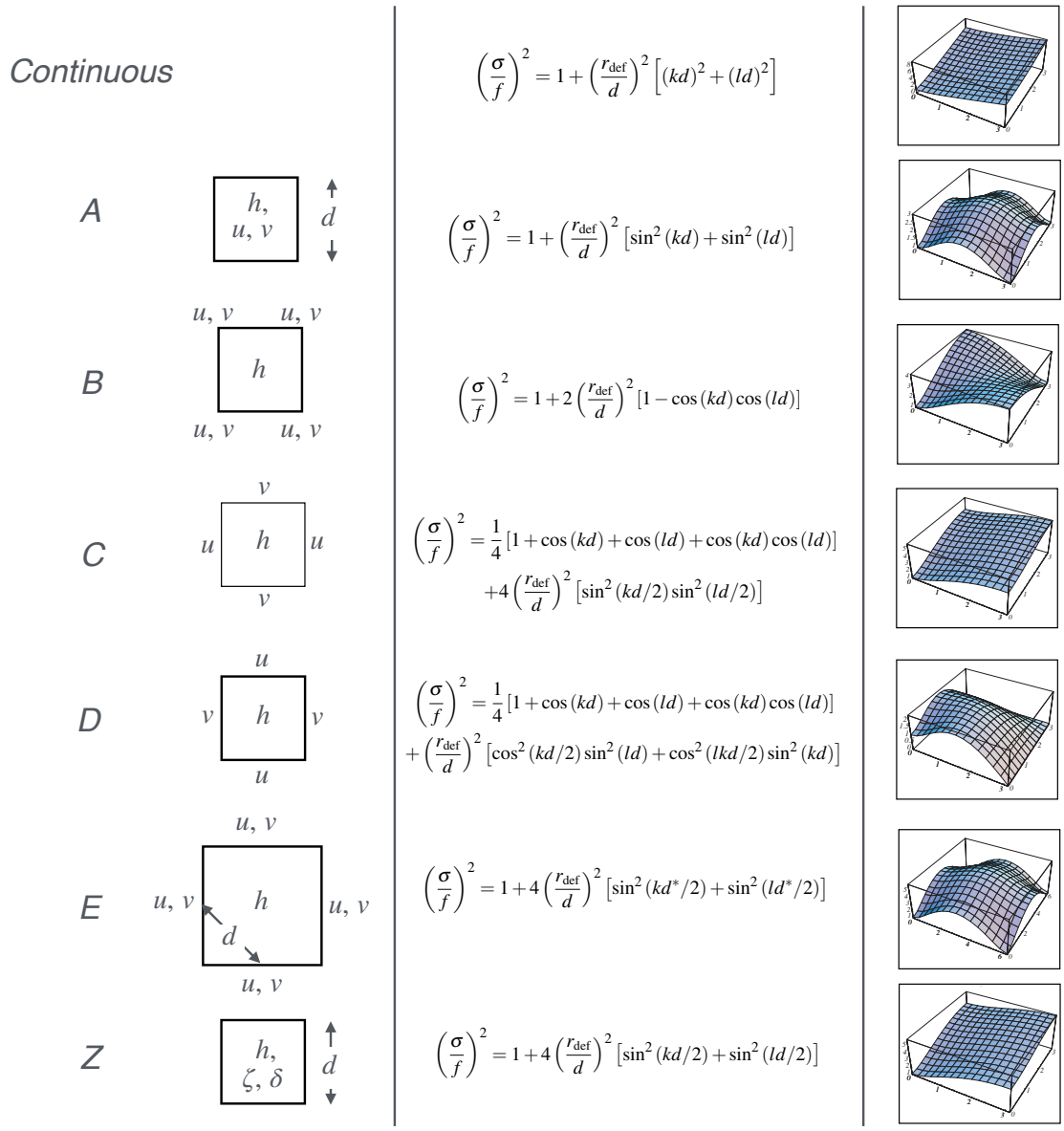


Figure 19.3: Grids, dispersion equations, and plots of dispersion equations for grids A - E and Z. The continuous dispersion equation and its plot are also shown for comparison. For plotting, it has been assumed that $r_{\text{def}}/d = 2$. This is a re-drawn version of a figure that appeared in Randall (1994).

It is natural to define the vorticity at cell corners. Since the mass is defined at cell centers this complicates the definition of the potential vorticity, $(\zeta + f)/h$.

For very small-scale inertia-gravity waves, geostrophic balance is not expected and the Coriolis terms are negligible; we essentially have pure gravity waves. This sug-

gests that the C-grid will perform well if the horizontal resolution of the model is fine enough so that the smallest waves that can be represented on the grid are insensitive to the Coriolis force. More precisely, AL argued that the C-grid does well when the grid size is small compared to r_{def}, the radius of deformation, which was defined in (19.19). A plot of the dispersion equation, given in Fig. 19.3, shows that the frequency increases monotonically with wave number, as in the exact (contiuous) dispersion equation, although not as rapidly. Recall, however, that this plot is for the particular case $r_{\text{def}}/d = 2$. Other values of r_{def}/d are discussed later.

From Fig. 19.1, you can see that the one-dimensional A-grid is equivalent to the super-position of two one-dimensional C-grids, shifted with respect to each other. The A-grid is equivalent to a super-position of *four* (shifted) two-dimensional C-grids (blue, red, orange, and black).

- *The D grid* allows a simple evaluation of the geostrophic wind, because the h points are north and south of the u points, and east and west of the v points. Since geostrophic balance is very important for large-scale motions, this sounds like a good thing. It is also apparent, however, that considerable averaging is needed in the pressure-gradient force, mass convergence / divergence, and even in the Coriolis terms. As a result, the dispersion equation for the D grid, shown in Fig. 19.3, is very badly behaved, giving zero phase speed for the shortest represented waves, and also giving a zero group speed for some modes.

  FV3, the dynamical core of the model used operationally by the U.S. National Centers for Environmental Prediction, uses what is called a "C-D" grid, but in practice it behaves like the D-grid (Skamarock, 2008; Konor and Randall, 2018a,b). It can only be used with strong damping of the horizontal divergence, and gives poor solutions when run with grid spacings of a few kilometers or finer (Carley et al., 2023; Wicker, 2023).

- *The E-grid* can be viewed as a modified B-grid, rotated by 45°. The rotation of the grid does not change the meaning of the u and v components of the wind, however; they point in the same directions as before, diagonally across the corners of the gray cells surrounding the mass points; this is different from the B-grid, on which, as mentioned above, the velocity components point along the directions of the walls that intersect at the corners of the mass cells. The gray grid cells are the same size as the B-grid cells, but rotated. The spatial "density" of h points is the same as in the other four grids. Because the E grid can be interpreted as a rotated B grid, the grid spacing for the E-grid is $d^* \equiv \sqrt{2}d$.

  At first, the E-grid seems perfect; no averaging is needed for the Coriolis terms, the pressure-gradient terms, or the mass convergence / divergence terms. Note, however, that the E-grid can be considered to live within the *overlapping* but unrotated orange and blue boxes. From this point of view, the E-grid is the superposition of two C-grids, shifted with respect to each other, so that the v points on one of the C-grids

coincide with the u points on the other, and vice versa. One of the two C-grid models is represented by the red variables in the figure, and the other by the black variables.

In 19.3, the nondimensional frequency for the E-grid is plotted as a function of kd^* and ld^*, out to a value of $\sqrt{2}\pi \cong 4.44$; this corresponds to the shortest "one-dimensional" mode. The group speed is zero for some combinations of kd^* and ld^*.

Fig. 19.3 also shows that the A-grid can also be viewed as a super-position of two E-grids, in which one of the E-grids is shifted by one-half of the grid spacing. Since each E-grid is equivalent to two superimposed but shifted C-grids, this is consistent with our earlier statement that the two-dimensional A-grid is equivalent to four shifted two-dimensional C-grids.

19.3.3 Dependence on the radius of deformation

Now recall the conclusion of AL, mentioned earlier, that the C-grid gives a good simulation of geostrophic adjustment provided that $r_{\text{def}}/d > 1$. Large-scale modelers are never happy to choose d and r_{def} so that r_{def}/d can be less than one. Nevertheless, in practice modes for which $r_{\text{def}}/d << 1$ can be unavoidable, at least for some situations. For example, Hansen et al. (1983) described a low-resolution atmospheric global circulation model, which they called Model II, designed for very long climate simulations in which low resolution was a necessity. Model II used a grid size of 10 degrees of longitude by 8 degrees of latitude; this means that the grid size was larger than the radius of deformation for many of the physically important modes that could be represented on the grid. As shown by AL, such modes cannot be well simulated using the C-grid. Having experienced these problems with the C-grid, Hansen et al. (1983) chose the B-grid for Model II.

The radius of deformation for the ocean is much smaller than that for the atmosphere, because the stratification of the ocean is weak due to the incompressibility of water. Ocean models must, therefore, contend with small radii of deformation. As a result, very fine-grids are needed to ensure that $r_{\text{def}}/d > 1$, even for external modes. For this reason, many ocean models have used the B-grid (e.g., Semtner and Chervin, 1992). This has changed in recent years, because faster computers have made it possible to use finer grids in ocean models, for which $r_{\text{def}}/d > 1$. Today's ocean models use the C-grid.

In addition, three-dimensional models of the atmosphere and ocean generate internal modes. With vertical structures typical of current atmospheric global circulation models, the highest internal modes can have radii of deformation on the order of 50 km or less. The same model may have a horizontal grid spacing on the order of 500 km, so that r_{def}/d can be on the order of 0.1. Fig. 19.4 demonstrates that the C-grid behaves very badly for $r_{\text{def}}/d = 0.1$. The phase speed actually decreases monotonically as the wave number increases, and becomes very small for the shortest waves that can be represented on the grid. Janjić and Mesinger (1989) have emphasized that, as a result, models that use the C-grid have difficulty in representing the geostrophic adjustment of high internal modes.

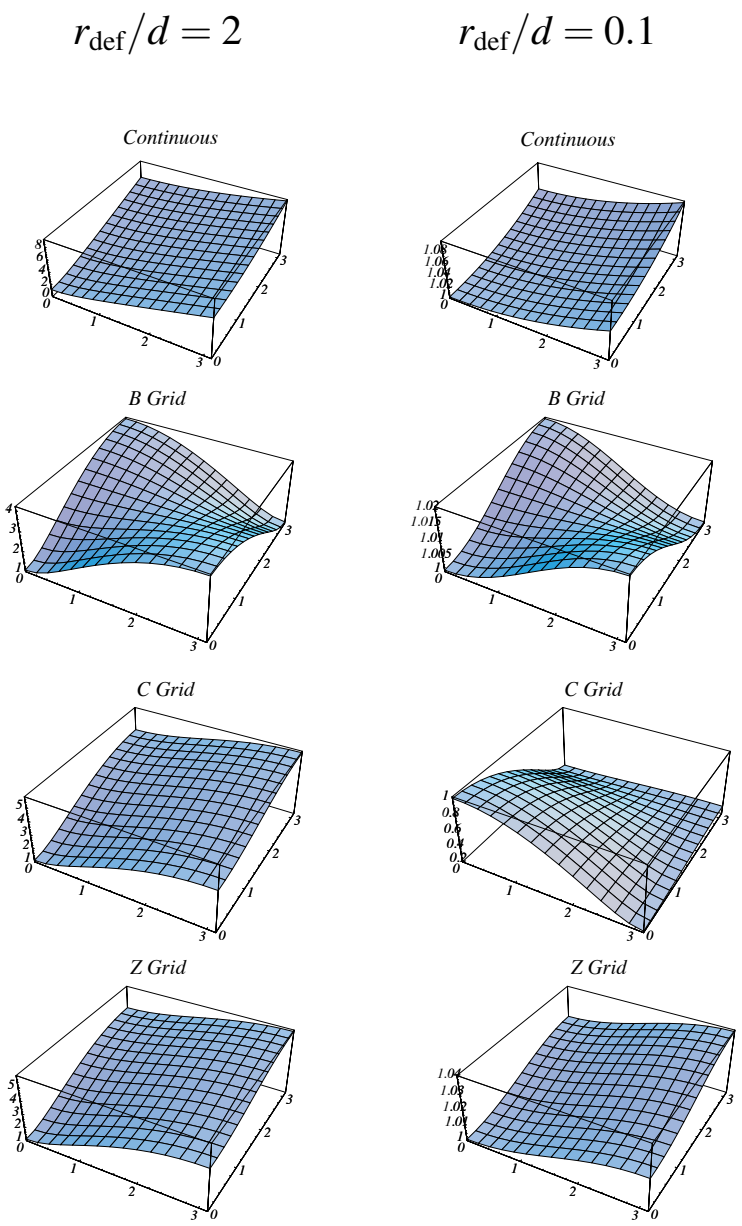


Figure 19.4: Dispersion relations for the continuous shallow water equations, and for finite-difference approximations based on the B, C, and Z-grids. The horizontal coordinates in the plots are kd and ld, respectively. The vertical coordinate is the normalized frequency, σ/f. The left column shows results for $r_{\text{def}}/d = 2$, and the right column for $r_{\text{def}}/d = 0.1$.

In contrast, the dispersion relation for the B-grid is qualitatively insensitive to the value of r_{def}/d. The B-grid has moderate problems for $r_{\text{def}}/d = 2$, but these problems do not become significantly worse for $r_{\text{def}}/d = 0.1$.

In summary, the C-grid does well with deep, external modes, but has serious problems with high internal modes, whereas the B-grid has moderate problems with all modes. *The C-grid's problem with high internal modes can be avoided by using a sufficiently fine horizontal grid spacing for a given vertical grid spacing.*

19.3.4 The Z-grid

From (19.42) - (19.44), we can derive an equivalent set in terms of vorticity,

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \,, \tag{19.47}$$

and divergence:

$$\frac{\partial \delta}{\partial t} - f\zeta + g\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right) = 0 \,, \tag{19.48}$$

$$\frac{\partial \zeta}{\partial t} + f\delta = 0 \,, \tag{19.49}$$

$$\frac{\partial h}{\partial t} + H\delta = 0 \,. \tag{19.50}$$

Of course, (19.50) is identical to (19.44).

Now consider an unstaggered grid for the integration of (19.48) - (19.50), which was called the Z-grid by Randall (1994). The Z-grid is also illustrated in Fig. 19.2. Inspection shows that with the Z-grid the components of the divergent part of the wind "want" to be staggered as in the C-grid, while the components of the rotational part of the wind "want" to be staggered as in the D-grid. This means that the Z-grid does not correspond to any of the grids A through E.

No averaging is required with the Z-grid. The only spatial differential operator appearing in (19.38) - (19.40) is the Laplacian, which is applied to h in the divergence equation. With the usual centered finite-difference stencils, the finite-difference approximation to $\nabla^2 h$ is defined in the same place as h itself. An unstaggered grid is thus a natural choice for the numerical integration of (19.48) - (19.50).

Fig. 19.4 shows that the dispersion relation for the Z-grid is very close to that of the C-grid, for $r_{\rm def}/d = 2$, but is drastically different for $r_{\rm def}/d = 0.1$. Whereas the C-grid behaves badly for $r_{\rm def}/d = 0.1$, the dispersion relation obtained with the Z-grid is qualitatively insensitive to the value of $r_{\rm def}/d$; it resembles the dispersion relation for the continuous equations, in that the phase speed increases monotonically with wave number and the group speed is non-zero for all wave numbers. Since the Z-grid is unstaggered, collapsing it to one dimension has no effect.

To implement the Z-grid, it is necessary to compute the normal velocity components on the cell walls from the vorticity and divergence. To do this, we start by solving for the stream function ψ and the velocity potential χ using

$$\nabla^2 \psi = \zeta \ , \tag{19.51}$$

and

$$\nabla^2 \chi = \delta \ . \tag{19.52}$$

These equations can be solved using the methods of Chapter 15. The stream function is and velocity potential are defined at the cell centers. The outward normal component of the divergent part of the wind on the cell walls can be obtained from

$$v_{\rm div} = \left(\chi_{\rm neighbor} - \chi_{\rm home\ base} \right) / d \ , \tag{19.53}$$

where d is the distance between cell centers. The outward normal component of the rotational part of the wind on a cell wall is given by the tangential derivative of ψ along the wall. To compute this, we first interpolate ψ to the cell corners. The we take the difference of ψ between neighboring corners, and divide by the length of the cell wall. The total outward normal component of the wind on the cell walls is the sum of the divergent and rotational parts.

Experience shows that the procedure described above, involving the solution of two two-dimensional Poisson equations, adds about 25% to the cost of a dynamical core.

19.4 Shifting shapes

In order to define the grids A-E and Z for a square mesh, we have had to specify both the locations and the orientations of the velocity components that are used to represent the horizontal wind. As discussed in Chapter 4, triangles, squares, and hexagons are the only regular polygons that tile the plane. Can we define grids A - E and Z so that the definitions work on triangular and hexagonal meshes, as well as quadrilateral meshes? In Fig. 19.5 we repeat the upper portion of Fig. 4.2.

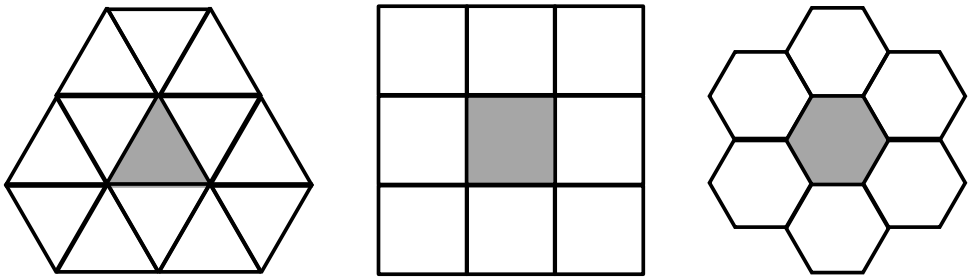


Figure 19.5: A copy of the upper portion of Fig. 4.2.

The A-Grid and Z-grid do not involve any staggering, so they can be unambiguously defined on triangular or hexagonal meshes, or for that matter meshes of any other shape. In the case of the A-grid, we need to define and predict two mutually orthogonal components of the horizontal wind vector, usually the zonal and meridional components. In the case of the Z-grid, only scalars are involved.

The B-grid can be generalized by defining it to have the horizontal velocity vectors at the corners of mass cells. *The vectors are represented using components that point along the walls that intersect at the corners.* On a triangular mesh, there are 6 intersecting walls at each corner, on a quadrilateral mesh there are two, and on a hexagonal mesh there are three. You would not wish to predict six (highly redundant) velocity components at the corners of a triangular mesh, or three (still redundant) velocity components at the corners of a hexagonal mesh. From this point of view, the B-grid is really only compatible with quadrilateral meshes.

The C-grid can be generalized by defining the normal component of the velocity on the edges of all mass-cells. This works on triangular, quadrilateral, and hexagonal meshes.

The generalized D-grid has the tangential velocity component on the edges of mass-cells. This also works on triangular, quadrilateral, and hexagonal meshes.

Like the B-grid, the generalized E-grid has wind vectors on the corners of the mass-cells. In contrast to the B-grid, however, *the E-grid's wind components point diagonally across the cells*, as shown for the gray cells in the illustration of the E-grid in Fig. 19.2.

There would be six such components on a triangular mesh, two on a quadrilateral mesh, and three on a hexagonal mesh. With this definition, the E-grid cannot be defined for the triangular mesh or hexagonal meshes.

Alternatively, we could try to define the E-grid as the superposition of multiple triangular or hexagonal C-grids, such that two-dimensional velocity vectors, represented by the tangential and normal components, are defined on each cell wall, as with the orange-grid cells shown for the E-grid in Fig. 19.2. It is not possible to create triangular or hexagonal E-grids in this way, so again we conclude that the E-grid cannot be defined for triangular or hexagonal meshes. It is possible, however, to create an E-grid by combining a hexagonal C-grid with a triangular C-grid. Naturally, the resulting grid suffers from computational modes.

From this point of view, the E-grid is really only compatible with quadrilateral meshes.

Table 19.1: The numbers of corners and edges per face, on the triangular, square, and hexagonal meshes.

	Triangles	**Squares**	**Hexagons**
Corners per face	1/2	1	2
Edges per face	3/2	2	3

19.5 The “degrees-of-freedom” problem

Table 19.1 lists the numbers of corners and edges per face, on the triangular, square, and hexagonal meshes. Table 19.2 lists the number of prognostic degrees of freedom in the wind field per mass point, for the generalized A-E and Z-grids, on triangular, square, and hexagonal meshes. From a physical point of view, *there should be two prognostic degrees of freedom in the wind field per mass point.* The A-grid and Z-grid achieve this ideal on all three meshes. None of the other grids has two degrees of freedom in the horizontal wind, per mass point, for all three mesh shapes.

Table 19.2 suggests that, if C-staggering is desired, then a square (or quadrilateral) mesh should be used. If squares are not used, then Z-staggering is best, but Z-staggering works fine for squares (and triangles), too.

Computational modes exist when the number of wind components per mass point is greater than or less than 2. We call this the “degrees-of-freedom” problem. With hexagonal or triangular meshes, the degrees-of-freedom problem can be avoided by using the A-grid (which has its own very serious issues) or the Z-grid.

Table 19.2: The number of prognostic degrees of freedom in the horizontal wind field, per mass point, on grids A-E and Z, and for triangular, square, and hexagonal meshes. For the Z-grid, the vorticity and divergence carry the information about the wind field.

Grid	Triangles	Squares	Hexagons
A	2	2	2
B	1	2	4
C	3/2	2	3
D	3/2	2	3
E	Does not exist	2	Does not exist
Z	2	2	2

19.6 Time-differencing schemes for the shallow-water equations

19.6.1 Explicit schemes

Consider the one-dimensional (1D) case, with spatial coordinate x and velocity component u. We neglect rotation and ssume that $v \equiv 0$. This means that we have divergence (i.e., $\partial u/\partial x$), but no vorticity. Linearizing about a state of rest, the continuous equations are (19.20) and (19.21).

We use a staggered 1D grid, which for this simple problem can be interpreted as the 1D C-grid, or the 1D B-grid, or the 1D Z-grid.

We can anticipate from our earlier analysis of the oscillation equation that forward time-differencing for both the momentum equation and the continuity equation is unstable, and that turns out to be true. We can also anticipate that a scheme that is centered in both space and time will be conditionally stable and neutral when stable. Such a scheme is given by:

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n-1}_{j+\frac{1}{2}}}{2\Delta t} + g\left(\frac{h^n_{j+1} - h^n_j}{d}\right) = 0\,, \tag{19.54}$$

$$\frac{h^{n+1}_j - h^{n-1}_j}{2\Delta t} + H\left(\frac{u^n_{j+\frac{1}{2}} - u^n_{j-\frac{1}{2}}}{d}\right) = 0\,. \tag{19.55}$$

Here we have used the C-grid. Compare with (19.25) - (19.26). With assumed solutions of the form $u^n_{j+1/2} = \widehat{u}^n e^{[ik(j+1/2)d]}$ and $h^n_j = \widehat{H}^n\ e^{(ikjd)}$, and the usual definition of the amplification factor, we find that

$$\left(\lambda^2 - 1\right)\widehat{u}^n + \lambda\frac{g\Delta t}{d}4i\sin\left(kd/2\right)\widehat{H}^n = 0\,, \tag{19.56}$$

$$\lambda\frac{H\Delta t}{d}4i\sin\left(kd/2\right)\widehat{u}^n + \left(\lambda^2 - 1\right)\widehat{H}^n = 0\,. \tag{19.57}$$

Non-trivial solutions exist for

$$\left(\lambda^2 - 1\right)^2 + \lambda^2\left(\frac{4c_{\rm gw}\Delta t}{d}\right)^2\sin^2\left(kd/2\right) = 0\,, \tag{19.58}$$

As should be expected with the leapfrog scheme, there are four modes altogether. Two of these are physical and two are computational.

We can solve (19.58) as a quadratic equation for λ^2. As a first step, rewrite it as

$$\left(\lambda^2\right)^2 + \lambda^2\left(-2 + b\right) + 1 = 0\,, \tag{19.59}$$

where, for convenience, we define

$$b \equiv \left(\frac{4c_{\rm gw}\Delta t}{d}\right)^2\sin^2\left(kd/2\right) \geq 0\,. \tag{19.60}$$

Obviously, for $\Delta t \to 0$ with fixed d we get $b \to 0$. The solution of (19.59) is

$$\begin{aligned}\lambda^2 &= \frac{-\left(b-2\right) \pm \sqrt{\left(b-2\right)^2 - 4}}{2} \\ &= \frac{-\left(b-2\right) \pm \sqrt{b\left(b-4\right)}}{2}\,.\end{aligned} \tag{19.61}$$

Inspection of (19.61) shows that for $b \to 0$, we get $|\lambda| \to 1$, as expected. For $\lambda = |\lambda| e^{i\theta}$ we see that

$$|\lambda|^2 [\cos(2\theta) + i \sin(2\theta)] = \frac{-(b-2) \pm \sqrt{b(b-4)}}{2} . \tag{19.62}$$

First consider the case $b \leq 4$. It follows from (19.62) that

$$|\lambda|^2 \cos(2\theta) = -\left(\frac{b-2}{2}\right), \text{ and } |\lambda|^2 \sin(2\theta) = \frac{\pm\sqrt{b(4-b)}}{2}, \text{ for } b \leq 4 , \tag{19.63}$$

from which we obtain

$$\tan(2\theta) = \frac{\sqrt{b(4-b)}}{2-b} \text{ for } b \leq 4 , \tag{19.64}$$

and

$$|\lambda|^4 = \left(\frac{b-2}{2}\right)^2 + \frac{b(4-b)}{4} = 1 \text{ for } b \leq 4 . \tag{19.65}$$

The scheme is thus neutral for $b \leq 4$, as was anticipated based on our earlier analysis of the oscillation equation.

Next, consider the case $b > 4$. Returning to (19.62), we find that

$$\sin(2\theta) = 0, \cos(2\theta) = \pm 1 \text{ and } |\lambda|^2 = \frac{-(b-2) \pm \sqrt{b(b-4)}}{2} \text{ for } b > 4 . \tag{19.66}$$

You should be able to see that for $b > 4$ there are always unstable modes.

We conclude that the scheme is stable and neutral for $b \leq 4$. This condition can also be written as

$$\left(\frac{c_{\rm gw}\Delta t}{d}\right)|\sin(kd/2)| \leq \frac{1}{2} \tag{19.67}$$

The worst case occurs for $|\sin(kd/2)| = 1$, which corresponds to $kd = \pi$, i.e., the shortest wave that can be represented on the grid. It follows that

$$\frac{c_{\rm gw}\Delta t}{d} < \frac{1}{2} \text{ is required for stability,} \tag{19.68}$$

and that the shortest wave will be the first to become unstable.

19.6.2 Implicit schemes

In atmospheric models, the fastest gravity waves, i.e., the external-gravity or "Lamb" waves, have speeds on the order of 300 m s^{-1}, which is the speed of sound in the Earth's atmosphere. The stability criterion for the leapfrog scheme as applied to the wave problem, i.e., (19.60), can therefore be painful. In models that do not permit vertically propagating sound waves (i.e., quasi-static models, or anelastic models, or shallow-water models), the external gravity wave is almost always the primary factor limiting the size of the time step. This is unfortunate, because the external gravity modes are believed to play only a minor role in weather and climate dynamics.

With this in mind, the gravity-wave terms of the governing equations are often approximated using implicit differencing. For the simple case of first-order backward-implicit differencing, we replace (19.42) - (19.43) by

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n}_{j+\frac{1}{2}}}{\Delta t} + \frac{g}{d}\left(h^{n+1}_{j+1} - h^{n+1}_{j}\right) = 0\,, \tag{19.69}$$

$$\frac{h^{n+1}_{j} - h^{n}_{j}}{\Delta t} + \frac{H}{d}\left(u^{n+1}_{j+\frac{1}{2}} - u^{n+1}_{j-\frac{1}{2}}\right) = 0\,. \tag{19.70}$$

This leads to

$$(\lambda-1)\,\widehat{u}^n+\lambda\frac{g\Delta t}{d}2i\sin(kd/2)\,\widehat{H}^n=0\,, \tag{19.71}$$

$$\lambda\frac{H\Delta t}{d}2i\sin(kd/2)\,\widehat{u}^n+(\lambda-1)\,\widehat{H}^n=0\,. \tag{19.72}$$

The condition for non-trivial solutions is

$$(\lambda-1)^2+\lambda^2 4\left(\frac{c_{\rm gw}\Delta t}{d}\right)^2\sin^2(kd/2)=0\,, \tag{19.73}$$

which, using the definition (19.60), is equivalent to

$$\lambda^2(1+b/4)-2\lambda+1=0\,. \tag{19.74}$$

This time there are no computational modes; the two physical modes satisfy

$$\lambda^2=\frac{2\pm\sqrt{4-4(1+b/4)}}{2(1+b/4)}=\frac{1\pm i\sqrt{b/4}}{1+b/4}\,. \tag{19.75}$$

The solutions are always oscillatory, and

$$|\lambda|^2=\frac{1+b/4}{(1+b/4)^2}=\frac{1}{1+b/4}\leq 1, \tag{19.76}$$

i.e., the scheme is unconditionally stable, and in fact it damps all modes.

The trapezoidal implicit scheme gives superior results; it is more accurate, and unconditionally neutral. We replace (19.61) - (19.62) by

$$\frac{u^{n+1}_{j+\frac12}-u^n_{j+\frac12}}{\Delta t}+\frac{g}{d}\left[\left(\frac{h^n_{j+1}+h^{n+1}_{j+1}}{2}\right)-\left(\frac{h^n_j+h^{n+1}_j}{2}\right)\right]=0\,, \tag{19.77}$$

$$\frac{h_j^{n+1}-h_j^n}{\Delta t}+\frac{H}{d}\left[\left(\frac{u^n_{j+\frac{1}{2}}+u^{n+1}_{j+\frac{1}{2}}}{2}\right)-\left(\frac{u^n_{j-\frac{1}{2}}+u^{n+1}_{j-\frac{1}{2}}}{2}\right)\right]=0\,. \tag{19.78}$$

This leads to

$$(\lambda-1)\,\widehat{u}^n+\left(\frac{1+\lambda}{2}\right)\frac{g\Delta t}{d}2i\sin(kd/2)\,\widehat{H}^n=0\,, \tag{19.79}$$

$$\left(\frac{1+\lambda}{2}\right)\frac{H\Delta t}{d}2i\sin(kd/2)\,\widehat{u}^n+(\lambda-1)\,\widehat{H}^n=0\,. \tag{19.80}$$

For non-trivial solutions, we need

$$(\lambda-1)^2+(1+\lambda)^2\left(\frac{c_{\rm gw}\Delta t}{d}\right)^2\sin^2(kd/2)=0\,. \tag{19.81}$$

Using (19.60) we can show that this is equivalent to

$$\lambda^2-2\lambda\left(\frac{16-b}{16+b}\right)+1=0\,. \tag{19.82}$$

The solutions are

$$\lambda=\left(\frac{16-b}{16+b}\right)\pm i\sqrt{1-\left(\frac{16-b}{16+b}\right)^2}\,. \tag{19.83}$$

It follows that $|\lambda|^2=1$ for all modes, i.e., the trapezoidal scheme is unconditionally neutral.

19.6.3 The forward-backward scheme

The disadvantage of implicit schemes is that they give rise to matrix problems, i.e., the various unknowns must be solved for simultaneously at all grid points. A simpler alternative, which is conditionally stable but allows a longer time step, is the *"forward-backward"* scheme, given by

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n}_{j+\frac{1}{2}}}{\Delta t} + \frac{g}{d}\left(h^{n+1}_{j+1} - h^{n+1}_{j}\right) = 0\,, \tag{19.84}$$

$$\frac{h^{n+1}_{j} - h^{n}_{j}}{\Delta t} + \frac{H}{d}\left(u^{n}_{j+\frac{1}{2}} - u^{n}_{j-\frac{1}{2}}\right) = 0\,. \tag{19.85}$$

This scheme can be called "partially implicit," because the end-of-time-step mass field predicted using (19.85) is used to compute the pressure-gradient force in (19.84). The continuity equation uses a forward time step. There is no need to solve a matrix problem.

We know that the forward scheme for both equations is unconditionally unstable, and that the backward scheme for both equations is unconditionally stable and damping. When we "combine" the two approaches, in the forward-backward scheme, the result turns out to be conditionally stable with a fairly long allowed time step, and neutral when stable. From (19.77) and (19.78), we get

$$(\lambda - 1)\,\widehat{u}^{n} + \lambda\,\frac{g\Delta t}{d} 2i \sin\left(kd/2\right)\widehat{H}^{n} = 0\,, \tag{19.86}$$

$$\frac{H\Delta t}{d} 2i \sin\left(kd/2\right)\widehat{u}^{n} + (\lambda - 1)\,\widehat{H}^{n} = 0\,. \tag{19.87}$$

This leads to

$$(\lambda - 1)^2 + 4\lambda\left(\frac{c_{\rm gw}\Delta t}{d}\right)^2 \sin^2\left(kd/2\right) = 0\,, \tag{19.88}$$

which is equivalent to

$$\lambda^2 + (b/4 - 2)\,\lambda + 1 = 0\,. \tag{19.89}$$

The solutions are

$$\lambda = \frac{(2 - b/4) \pm \sqrt{(2 - b/4)^2 - 4}}{2} = (1 - b/8) \pm i\sqrt{b/4 - (b/8)^2}\,. \tag{19.90}$$

The discriminant is non-negative for

$$b \leq 16\,, \tag{19.91}$$

which corresponds to

$$\left(\frac{c_{\mathrm{gw}}\Delta t}{d}\right)^2 \sin^2(kd/2) \leq 1\,. \tag{19.92}$$

It follows that

$$\frac{c_{\mathrm{gw}}\Delta t}{d} \leq 1 \text{ is required for stability}\,. \tag{19.93}$$

The time step can thus be twice as large as with the leapfrog scheme. When (19.84) is satisfied, we have $|\lambda|^2 = 1$ for all modes, i.e., the scheme is neutral when stable (like the leapfrog scheme). The forward-backward scheme is thus very attractive: It allows a long time step, it is neutral when stable, it is non-iterative, and it has no computational modes.

Going to two dimensions and adding rotation does not change much. The Coriolis terms can easily be made implicit if desired, since they are linear in the dependent variables and do not involve spatial derivatives.

19.7 Summary and conclusions

Horizontally staggered grids are important because they make it possible to avoid or minimize computational modes in space, and to realistically simulate geostrophic adjustment. The Z-grid gives the best overall simulation of geostrophic adjustment, for a range of grid sizes relative to the radius of deformation. In order to use the Z-grid, it is necessary to solve a pair of Poisson equations on each time step.

Computational modes in space can be caused in several ways:

1. Multiple models on the same grid, as with the A- and E-grids.
2. Averaging, as with the B-, C-, and D-grids.
3. The degrees-of-freedom problem, as with C-staggering on a hexagonal grid.

The rapid phase speeds of external gravity waves limit the time step that can be used with explicit schemes. Implicit schemes can be unconditionally stable, but in order to use them it is necessary to solve the equations simultaneously for all grid points.

19.8 Problems

1. Derive the dispersion equation for the C-grid, as given in Fig. 19.3.

2. Consider the linearized (about a resting basic state) shallow-water equations without rotation on the *one-dimensional* versions of the A-grid and the C-grid. Let the distance between neighboring mass points be d on both grids. Use the forward-backward time differencing discussed in Section 19.6.3. Derive the stability criteria for both cases, and compare the two results.

3. Program the two-dimensional linearized shallow water equations for the square A-grid and the square C-grid, using a mesh of 101 x 101 mass points, with periodic boundary conditions in both directions. Use the forward-backward time differencing discussed in Section 19.6.3. Set $f = 10^{-4}$ s^{-1}, $g = 0.1$ m s^{-1}, $H = 10^3$ m, and $d = 10^5$ m. Use implicit time differencing for the Coriolis terms.

   In the square region

$$\begin{aligned} 45 \leq i \leq 55, \\ 45 \leq j \leq 55, \end{aligned} \tag{19.94}$$

   apply a noisy forcing in the continuity equation, of the form

$$\left(\frac{\partial h}{\partial t}\right)_{\text{noisy}} = (-1)^{i+j} N \sin(\omega_N t) \ , \tag{19.95}$$

   and set $(\partial h/\partial t)_{\text{noisy}} = 0$ at all other grid points. Adopt the values $\omega_N = 2\pi \times 10^{-3}$ s^{-1}; and $N = 10^{-4}$ m s^{-1}. In addition, apply a smooth forcing to the entire domain of the form

$$\left(\frac{\partial h}{\partial t}\right)_{\text{smooth}} = S \sin\left(\frac{2\pi x}{L}\right) \sin\left(\frac{2\pi y}{L}\right) \sin(\omega_S t) \tag{19.96}$$

   with $\omega_S = 2\pi\sqrt{gH}/L$ s^{-1} and $S = 10^{-4}$ m s^{-1}. Here $L = 101 \times d$ is the width of the domain.

   Finally, include friction in the momentum equations, of the form

$$\left(\frac{\partial u}{\partial t}\right)_{\text{fric}} = -\kappa u, \qquad \left(\frac{\partial v}{\partial t}\right)_{\text{fric}} = -\kappa v, \tag{19.97}$$

where $K = 2 \times 10^{-5}$ s^{-1}. Use implicit time differencing for these friction terms. Because the model has both forcing and damping, it is possible to obtain a statistically steady solution.

(a) Using the results of Problem 2 above, choose a suitable time step for each model.

(b) As initial conditions, put $u = 0$, $v = 0$, and $h = 0$. Run both versions of the model for at least 10^5 simulated seconds, and analyze the results.

(c) Repeat your runs using $f = 3 \times 10^{-3}$ s^{-1}. Discuss the changes in your results.

4. Show that there is no pressure-gradient term in the vorticity equation for the shallow-water system on the C-grid.

5. Derive the form of the pressure-gradient term in the divergence equation for the shallow-water system on the C-grid, and compare with the continuous case.

Chapter 20

Up against the wall

20.1 Introduction

Boundary conditions can be real or fictitious, because boundaries can be real or fictitious. With some vertical coordinate systems (such as height), topography imposes a lateral boundary condition in an atmospheric model. Ocean models describe flows in basins, and so (depending again on the vertical coordinate system used) have "real" lateral boundary conditions. Limited area models have artificial lateral boundaries. Models in which the grid spacing changes rapidly (e.g., nested-grid models) effectively apply boundary conditions where the dissimilar grids meet.

The Earth's surface is a "real wall" at the lower boundary of the atmosphere. Most numerical models have "fictitious walls" or "lids" at their tops. Upper boundary conditions will be discussed in a later chapter.

20.2 Real walls

No mass passes through a real wall, so the normal component of the velocity vanishes at the wall. We should position and orient the walls so that they are located where the wind component normal to the wall is predicted. If we use the C-grid, for example, a wall that is oriented in the north-south direction should be located at a u point. See Fig. 20.1 for a one-dimensional example.

20.3 Advection at inflow boundaries

Consider a fictitious wall in a limited-area model. We consider advection with the wind blowing into our domain at the boundary.

Suppose that the initial condition is given only in a certain limited domain. Fig. 20.2 shows lines of constant $x - ut$, for $u > 0$. These are characteristics. If the initial condition is

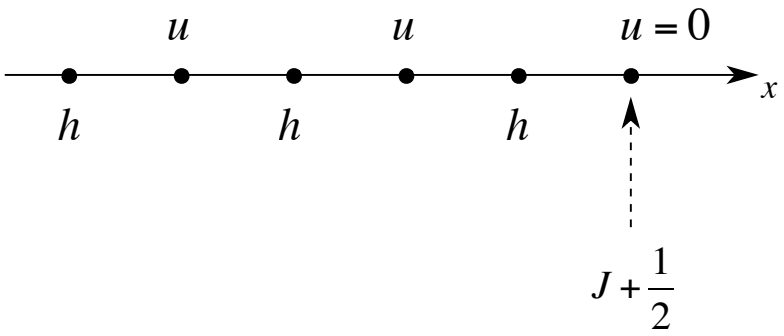


Figure 20.1: A one-dimensional staggered grid for solution of the shallow water equations, near a wall where $j = J + \frac{1}{2}$.

specified between the points $(x = x_0\ , t = 0)$, and $(x = x_1,\ t = 0)$, then $A\,(x,t)$ is determined in the triangular domain ABC. To determine $A\,(x,\ t)$ above the line $x - ut = x_0$, we need an "inflow boundary condition" at $x = x_0$ for $t > 0$. When this boundary condition and the initial condition at $t = 0$ between the points A and B are specified, we can obtain the solution within the entire domain $x_0 \leq x \leq x_1$. We say that the problem is well-posed. Note that a boundary condition at $x = x_1$ is of no use.

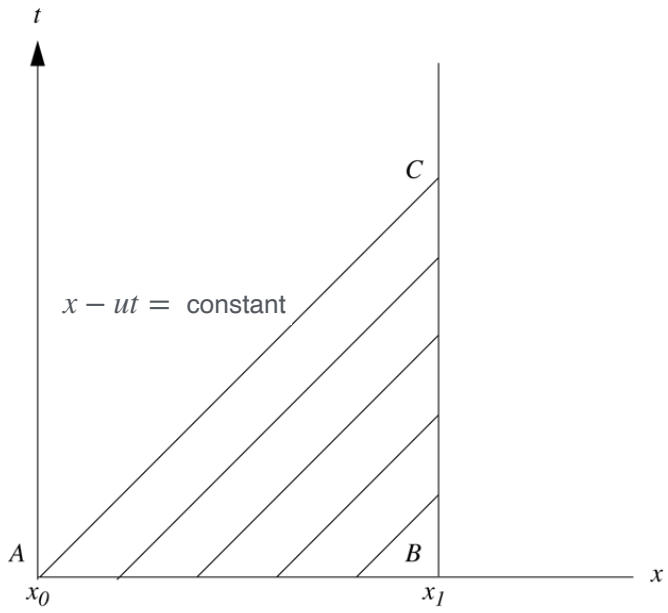


Figure 20.2: An initial condition that is specified between $x = x_0$ and $x = x_1$ determines the solution only along the characteristics shown.

At an inflow boundary, say at $x = 0$, we have to prescribe $A\,(0,\ t)$ for all time. As an example, suppose that $A\,(0,\ t)$ is a simple harmonic function, with frequency σ. The value of σ is determined by the upstream boundary condition. We can choose σ any way we

please.

Referring again to the continuous advection equation, i.e.,

$$\frac{\partial A}{\partial t} + u\frac{\partial A}{\partial x} = 0 \,, \tag{20.1}$$

we assume $u > 0$ and write

$$A(x,t) = \text{ Re}\left[\widehat{A}(x)\, e^{-i\sigma t}\right] \text{ for } \sigma \neq 0 \,, \tag{20.2}$$

where $\widehat{A}(0)$ is a real constant. Then A at the inflow boundary satisfies

$$A(0,t) = \widehat{A}(0)\ \text{Re}\left(e^{-i\sigma t}\right) = \widehat{A}(0)\cos(\sigma t)\ . \tag{20.3}$$

Use of (20.2) in (20.1) gives

$$-i\sigma\widehat{A} + u\frac{d\widehat{A}}{dx} = 0 \,, \tag{20.4}$$

which has the solution

$$\widehat{A}(x) = \widehat{A}(0)\, e^{ikx} \,. \tag{20.5}$$

The dispersion equation is obtained by substituting (20.5) into (20.4):

$$\sigma = uk \tag{20.6}$$

The full solution is thus

$$\begin{aligned} A(x,\, t) &= \widehat{A}_0\, \text{Re}\left\{\exp\left[i\left(kx - \sigma t\right)\right]\right\} \\ &= \widehat{A}_0\, \text{Re}\left\{\exp\left[ik\left(x - ut\right)\right]\right\} \end{aligned} \tag{20.7}$$

20.3.1 A space-centered scheme

Now consider the same problem again, this time as represented through the differential-difference equation

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0\,. \tag{20.8}$$

We assume a solution of the form

$$A_j(t) = \mathrm{Re}\left\{\widehat{A}_j e^{-i\sigma t}\right\}\,. \tag{20.9}$$

Note that the frequency is given. For convenience, we now define the nondimensional ratio

$$\theta \equiv \frac{\sigma \Delta x}{u}\,. \tag{20.10}$$

The frequency and wind speed imply a wave number. We obtain the dispersion relation

$$\theta = \sin(k\Delta x)\,. \tag{20.11}$$

For comparison, the exact dispersion equation, (20.6), can be written as

$$\theta = k\Delta x\,. \tag{20.12}$$

Fig. 20.3 gives a schematic plot, with θ and $k\Delta x$ as coordinates, for the true dispersion equation (20.6) and the approximate dispersion equation (20.11). *For a given* θ there is only one k in the exact solution. In the numerical solution, however, there are two ks, which we will call k_1 and k_2. As discussed below, for $\theta > 0$, k_1 corresponds to the exact solution. The figure makes it clear that

$$k_2 \Delta x = \pi - k_1 \Delta x \, . \tag{20.13}$$

The group velocity is positive, as it should be, for $k\Delta x < \pi/2$, but it is negative for $k\Delta x > \pi/2$.

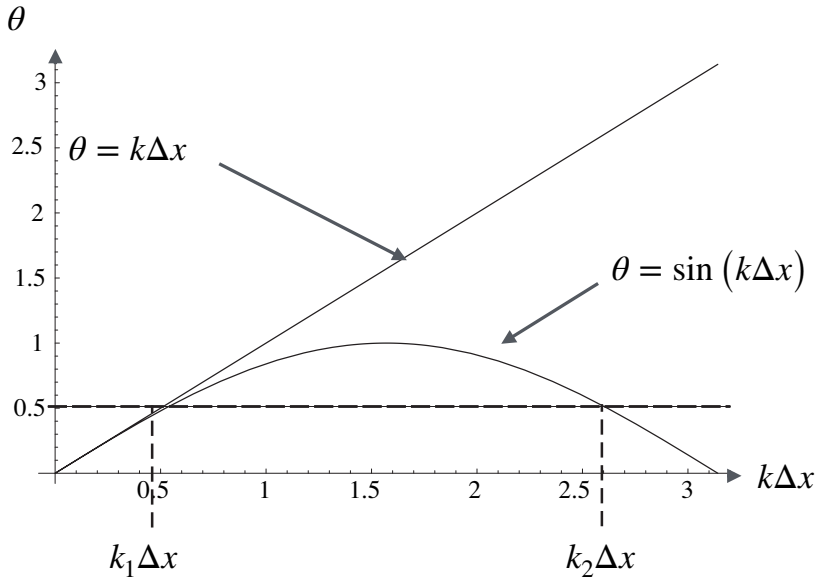


Figure 20.3: A schematic plot, with $k\Delta x$ and θ as coordinates, for the true solution (20.6) and the approximate solution (20.11). The dashed line illustrates that, for a given σ, the approximate solution is consistent with two different wave numbers.

When we studied computational modes in time with the leapfrog time-differencing scheme in Chapter 6, we found two solutions with different frequencies, for a given wave number. Here we have two solutions with different wave numbers, for a given frequency. The solution corresponding to k_1 is a "physical mode" in space, and the solution corresponding to k_2 is a "computational mode" in space. The wavelength that corresponds to k_2, i.e., the wavelength of the computational mode, will always be between $2\Delta x$ (which corresponds to $k\Delta x = \pi$) and $4\Delta x$ (which corresponds to $k\Delta x = \pi/2$). For $k\Delta x > \pi/2$, there really is no physical mode. In other words, *the physical mode exists only for wavelengths longer than* $4\Delta x$. The figure makes it clear that when the physical mode has a very long wavelength, the computational mode has a very short wavelength.

In view of (20.11), the condition $k_1 \Delta x < \pi/2$, which is required for a physical mode to exist, corresponds to

$$\sin^{-1}(\sigma \Delta x / c) < \pi/2 \, . \tag{20.14}$$

This condition can be satisfied by choosing Δx small enough, for given values of σ and c. In other words, *for a given frequency and wind speed the grid spacing must be small enough to represent the implied physical wavelength.* This seems like common sense.

Referring back to (20.7) and (20.13), we see that the two modes can be written as

$$\text{Physical mode:} \quad A_j = \widehat{A}_0 \,\mathrm{Re}\left\{\exp\left[ik_1\left(j\Delta x - \frac{\sigma}{k_1}t\right)\right]\right\}, \tag{20.15}$$

$$\begin{aligned}\text{Computational mode:} \quad A_j &= \widehat{A}_0 \,\mathrm{Re}\left\{\exp\left[ik_2\left(j\Delta x - \frac{\sigma}{k_2}t\right)\right]\right\} \\ &= \widehat{A}_0 \,\mathrm{Re}\left\{\exp\left[i\left(j\pi - k_1 j\Delta x - \sigma t\right)\right]\right\} \\ &= (-1)^j \widehat{A}_0 \,\mathrm{Re}\left\{\exp\left[-ik_1\left(j\Delta x + \frac{\sigma}{k_1}t\right)\right]\right\}.\end{aligned} \tag{20.16}$$

Here we have used (20.13) and $e^{ij\pi} = (-1)^j$. The phase velocity of the computational mode is equal and opposite to that of the physical mode, and the computational mode oscillates in space with wave length $2\Delta x$, due to the factor of $(-1)^j$. That's just terrible.

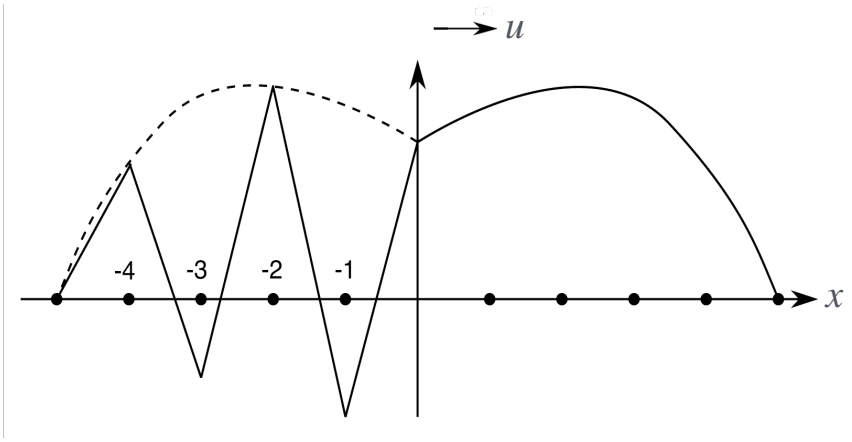


Figure 20.4: Schematic illustration of the solution of the advection equation in the neighborhood of a "smokestack," which acts as a time-dependent source of the advected quantity. In the numerical solution, a computational mode appears in the domain $j < 0$, and a physical mode appears in the domain $j > 0$.

In general, the solution is a superposition of the physical and computational modes. For the case $c > 0$, and if the point $j = 0$ is the "source of influence," like a smoke stack,

only a physical mode appears for $j > 0$ and only a computational mode appears for $j < 0$. Fig. 20.4 shows this schematically for some arbitrary time. The dashed line for $j < 0$ represents (20.16), without the factor $(-1)^j$; the solid line represents the entire expression. The influence of the computational mode propagates to the left. If the wave length of the physical mode is very large (compared to Δx), the computational mode will appear as an oscillation from point to point, i.e., a wave of length $2\Delta x$.

Since we control σ, Δx, and u, we also control the value of θ. Suppose that we give $\theta > 1$, e.g., by choosing a large value of Δx. In that case, k has to be complex:

$$k = k_r + ik_i \ . \tag{20.17}$$

To see what this means, we use the mathematical identity

$$\sin(k\Delta x) = \sin(k_r\Delta x)\cosh(k_i\Delta x) + i\cos(k_r\Delta x)\sinh(k_i\Delta x) \ , \tag{20.18}$$

which can be derived from Euler's formula. Here cosh and sinh are the hyperbolic cosine and sine functions. Substituting from (20.18) into the right-hand side of (20.10), and equating real and imaginary parts, we find that

$$\begin{aligned} \theta &= \sin(k_r\Delta x)\cosh(k_i\Delta x) \ , \\ 0 &= \cos(k_r\Delta x)\sinh(k_i\Delta x) \ . \end{aligned} \tag{20.19}$$

We cannot accept a solution with $\sinh(k_i\Delta x) = 0$, because this would imply $k_i\Delta x = 0$, and we already know that $k_i \neq 0$. Therefore we must take

$$\cos(k_r\Delta x) = 0, \text{ which implies that } k_r\Delta x = \pi/2 \ . \tag{20.20}$$

This is the $4\Delta x$ wave, for which

$$\sin(k_r\Delta x) = 1 \ , \tag{20.21}$$

and so from (20.19) we find that

$$k_i \Delta x = \cosh^{-1}(\theta) > 0\,. \tag{20.22}$$

The inequality follows because we have assumed that $\theta > 1$. We can now write (20.9) as

$$A_j = \widehat{A}_0\, e^{-k_i j \Delta x}\, \mathrm{Re}\left[e^{-i\sigma t} e^{i k_r j \Delta x}\right]\,. \tag{20.23}$$

Since $k_i > 0$, the signal dies out downstream, as shown schematically in Fig. 20.5. This is unrealistic but will not make the model crash.

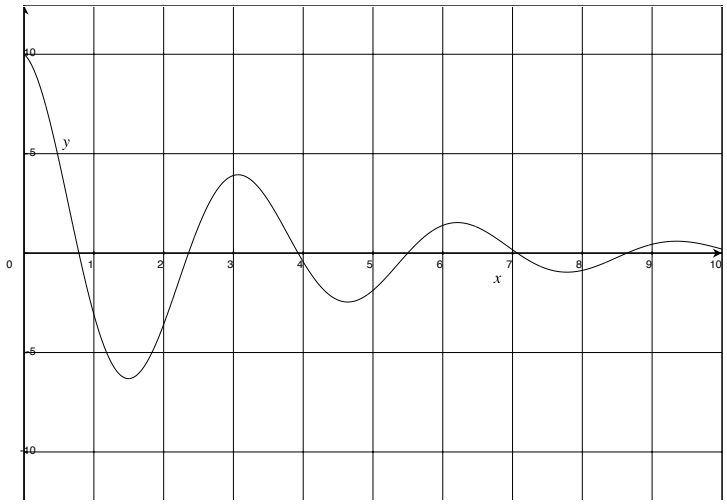


Figure 20.5: Sketch of a damped oscillation.

20.3.2 A space-uncentered scheme

Suppose that we use an uncentered scheme instead of the centered scheme (20.8), e.g.,

$$\frac{dA_j}{dt} + \frac{u}{\Delta x}\left(A_j - A_{j-1}\right) = 0, \tag{20.24}$$

with $c > 0$. This scheme has no computational mode in space, because it involves only two grid points. We will show that the uncentered scheme damps the solution regardless of the values of σ, Δx, and u. using

$$A_j = \widehat{A} e^{-i\sigma t} e^{ikj\Delta x}\,, \tag{20.25}$$

we obtain the dispersion equation

$$-i\theta + \left(1 - e^{-ik\Delta x}\right) = 0 \,. \tag{20.26}$$

First, assume that k is real. Setting the real and imaginary parts of (20.26) to zero gives

$$\cos(k\Delta x) = 1, \text{ and } -\theta + \sin(k\Delta x) = 0 \,. \tag{20.27}$$

Note that $\cos(k\Delta x) = 1$ implies that $k\Delta x = 0$. This constant-with-x solution is not interesting. We conclude that *k must be complex when A varies with x*. We therefore use (20.17) in (20.26) to obtain

$$-i\theta + \left(1 - e^{-ik_r\Delta x} e^{k_i\Delta x}\right) = 0 \,. \tag{20.28}$$

Setting the real part to zero gives

$$1 - e^{k_i\Delta x}\cos(k_r\Delta x) = 0 \,, \tag{20.29}$$

and setting the imaginary part to zero gives

$$\theta + e^{k_i\Delta x}\sin(k_r\Delta x) = 0 \,. \tag{20.30}$$

These two equations can be solved for the two unknowns k_r and k_i. From (20.29) we see that

$$e^{k_i\Delta x} = \sec(k_r\Delta x) \geq 1 \,. \tag{20.31}$$

Using this in (20.30), we obtain

$$k_r \Delta x = -\tan^{-1}(\theta) \ . \tag{20.32}$$

Then (20.31) gives

$$e^{k_i \Delta x} = \sec\left[\tan^{-1}(\theta)\right] > 1 \ . \tag{20.33}$$

From (20.33), we conclude that $k_i > 0$. Substituting back into (20.25), we obtain

$$A_j = \widehat{A}\, e^{-k_i j \Delta x} \mathrm{Re}\left[e^{i(k_r j \Delta x - \sigma t)}\right] \ . \tag{20.34}$$

This shows that the signal weakens as $j \to \infty$.

20.4 Advection at outflow boundaries

20.4.1 Nitta's 8 methods

Suppose that we are carrying out our numerical solution of the one-dimensional advection equation over the region between $j = 1$ and $j = J$, as shown in Fig. 20.6, using centered space-differencing with a continuous time derivative, i.e.,

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0 \ . \tag{20.35}$$

Figure 20.6: A finite domain with boundaries on both sides.

We assume "without loss of generality" that $u > 0$, so that the inflow boundary is on the left in the figure, and the outflow boundary is on the right. We need inflow boundary conditions for both the continuous and finite-difference advection equations. With the finite-difference

scheme, the inflow boundary condition essentially determines A_0 as a function of time. At $j = 1$ we can write, using centered space differencing,

$$\frac{dA_1}{dt} + u\left(\frac{A_2 - A_0}{2\Delta x}\right) = 0\,. \tag{20.36}$$

We also need a "*computational boundary condition*" at the fictitious outflow boundary. At $j = J-1$ our scheme gives

$$\frac{dA_{J-1}}{dt} + u\left(\frac{A_J - A_{J-2}}{2\Delta x}\right) = 0\,, \tag{20.37}$$

and at $j = J$ we have

$$\frac{dA_J}{dt} + u\left(\frac{A_{J+1} - A_{J-1}}{2\Delta x}\right) = 0\,. \tag{20.38}$$

Eq. (20.38) shows that in order to predict A_J, we need to know A_{J+1}, which is not available because it lies outside our domain. We need to give a condition that can be used to determine A_J as a function of time. Ideally, this outflow boundary condition should not affect the solution in the interior. Its only purpose is to limit the size of the domain. If the computational outflow boundary condition is not chosen well, there is a possibility of exciting a strong computational mode. Tsuyoshi Nitta (1964) considered eight methods for dealing with the outflow boundary condition. They are listed in Table 20.1:

Nitta (1964) integrated the advection equation with each of the eight methods, using leapfrog time differencing. His paper deals mainly with space differencing, but as discussed later his conclusions are influenced by his choice of leapfrog time differencing.

With Nitta's Method 1, A_J is constant in time. This is obviously a bad assumption, and it leads to bad results.

With Method 2, $A_J^n = A_{J-1}^n$, i.e., the first derivative of A vanishes at the wall. Method 3 is similar, but in terms of the time derivatives.

With Method 4, $A_J^n = A_{J-2}^n$, so it looks similar to Method 2. Method 4 is just asking for trouble, though, because $A_J^n = A_{J-2}^n$ is characteristic of the $2\Delta x$ mode.

Name of scheme	Form of scheme
Method 1	$A_J =$ constant in time
Method 2	$A_J^n = A_{J-1}^n$
Method 3	$\partial A_J/\partial t = \partial A_{J-1}/\partial t$
Method 4	$A_J^n = A_{J-2}^n$
Method 5	$A_J^n = 2A_{J-1}^n - A_{J-2}^n$
Method 6	$\partial A_J/\partial t = \partial\left(2A_{J-1} - A_{J-2}\right)/\partial t$
Method 7	$\partial A_J/\partial t = -u\left(A_J^n - A_{J-1}^n\right)/\Delta x$
Method 8	$\partial A_J/\partial t = -u\left(3A_J^n - 4A_{J-1}^n + A_{J-2}^n\right)/2\Delta x$

Table 20.1: A summary of the computational boundary conditions studied by Nitta (1964).

Method 5, on the other hand, sets $A_J^n = 2A_{J-1}^n - A_{J-2}^n$. This is a linear extrapolation from the two interior points to A_J^n. It is equivalent to setting the second derivative to zero at $J-1$.

Method 6 is similar to Method 5, but uses a linear extrapolation of the time derivative.

Method 7 predicts A_j by means of

$$\frac{dA_J}{dt} + u\left(\frac{A_J^n - A_{J-1}^n}{\Delta x}\right) = 0\,, \tag{20.39}$$

which uses uncentered differencing in space. It works well.

Method 8 is similar to Method 7, but has higher-order accuracy.

20.4.2 An analysis of Nitta's methods

We now perform an analysis to understand Nitta's results. We assume that the domain extends far upstream towards decreasing j. We can use the results of our earlier analysis. In general, the solution can be expressed as a linear combination of the physical and computational modes. Recall that the physical mode is given by $A_j \sim e^{\left[ik_1\left(j\Delta x - \frac{\sigma}{k_1}t\right)\right]}$ and the computational mode is given by $A_j \sim (-1)^j e^{\left[-ik_1\left(j\Delta x + \frac{\sigma}{k_1}t\right)\right]}$. Since the computational mode propagates "upstream," the outflow boundary for the physical mode is effectively the inflow boundary for the computational mode. We examine the solution at the outflow

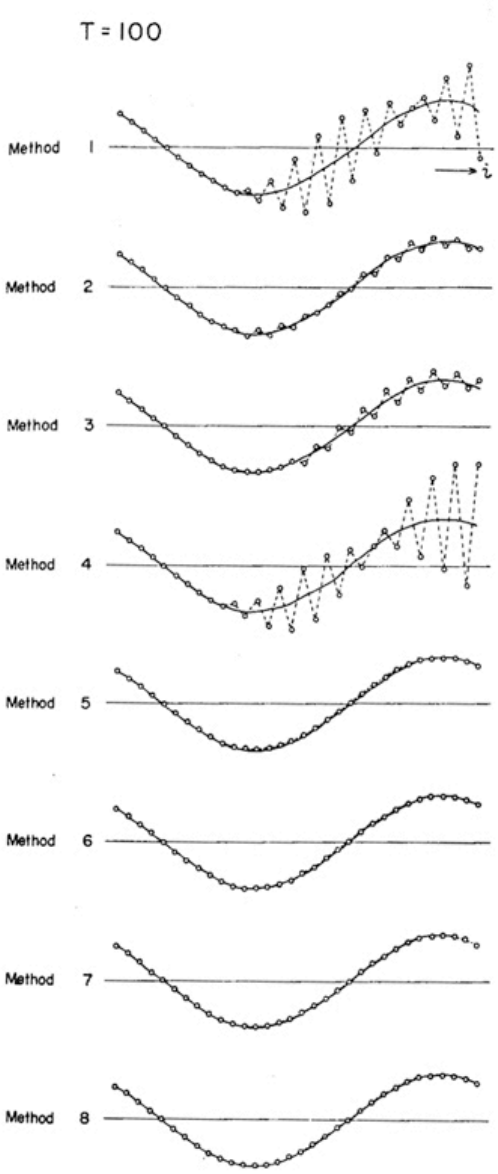


Figure 20.7: A summary of Nitta's numerical results, with various computational boundary conditions. Here leapfrog time differencing was used.

boundary in order to determine the "initial" amplitude of the computational mode that is potentially excited there. Obviously we want that amplitude to be as small as possible.

Referring to (20.15) and (20.16), we can write

$$A_j = \widehat{A}\,\mathrm{Re}\left\{\exp\left[ik\left(j\Delta x - \frac{\sigma}{k}t\right)\right] + r(-1)^j \exp\left[-ik\left(j\Delta x + \frac{\sigma}{k}t\right)\right]\right\}\,, \tag{20.40}$$

where k (now without a subscript) is the wave number of the physical mode and r is the "virtual reflection rate" at the boundary for the computational mode, so that $|r|$ is the ratio of the amplitude of the computational mode to that of the physical mode. We would like to make $r = 0$.

In Method 1, A_J is kept constant. Assume $A_J = 0$, for simplicity, and let J be even ("without loss of generality"), so that $(-1)^J = 1$. We then can write, from (20.40),

$$\begin{aligned} A_J &= \widehat{A}\,\mathrm{Re}\left\{\exp\left[ik\left(j\Delta x - \frac{\sigma}{k}t\right)\right] + r\exp\left[-ik\left(j\Delta x + \frac{\sigma}{k}t\right)\right]\right\} \\ &= \widehat{A}\left[\exp\left(ikj\Delta x\right) + r\exp\left(-ikj\Delta x\right)\right]\exp\left(-i\sigma t\right) \\ &= 0. \end{aligned} \tag{20.41}$$

Since $e^{-i\sigma t} \neq 0$, we conclude that

$$r = \frac{-\exp\left(ikj\Delta x\right)}{\exp\left(-ikj\Delta x\right)} = -\exp\left(2ikj\Delta x\right)\ , \tag{20.42}$$

which implies that $|r| = 1$. *This means that the incident wave is totally reflected*. The computational mode's amplitude is equal to that of the physical mode - a very unsatisfactory situation, as can be seen from Fig. 20.7.

With Method 2, and still assuming that J is even, we put $u_J = u_{J-1}$. This leads to

$$\exp\left(ikJ\Delta x\right) + r\exp\left(-ikJ\Delta x\right) = \exp\left[ik\left(J-1\right)\Delta x\right] - r\exp\left[-ik\left(J-1\right)\Delta x\right]\ , \tag{20.43}$$

which implies that

$$|r| = \tan\left(k\Delta x/2\right)\ . \tag{20.44}$$

For $L = 4\Delta x$, we get $|r| = 1$. Recall that $L < 4\Delta x$ need not be considered. For large L, we get $|r| \to 0$, i.e., long waves are reflected only weakly. In Fig. 20.7, the incident mode is relatively long.

With Method 5, it turns out that $|r| = \tan^2(k\Delta x/2)$. Fig. 20.8 is a graph of $|r|$ versus $k\Delta x$ for Methods 2 and 5. Because k is the wave number of the physical mode, the plot shows only the region $0 \leq k\Delta x \leq \pi/2$. Higher-order extrapolations give even better results for the lower wave numbers, but there is little motivation for doing this.

In actual computations there will also be an inflow boundary, and *this will then act as an outflow boundary for the computational mode, which has propagated back upstream.* A secondary mode will then be reflected from the inflow boundary and will propagate downstream, and so on. There exists the possibility of multiple reflections back and forth between the boundaries. Can this process amplify in time, as in a laser? It can if there is a source of energy for the computational mode, and such a source exists if the computational mode is damped.

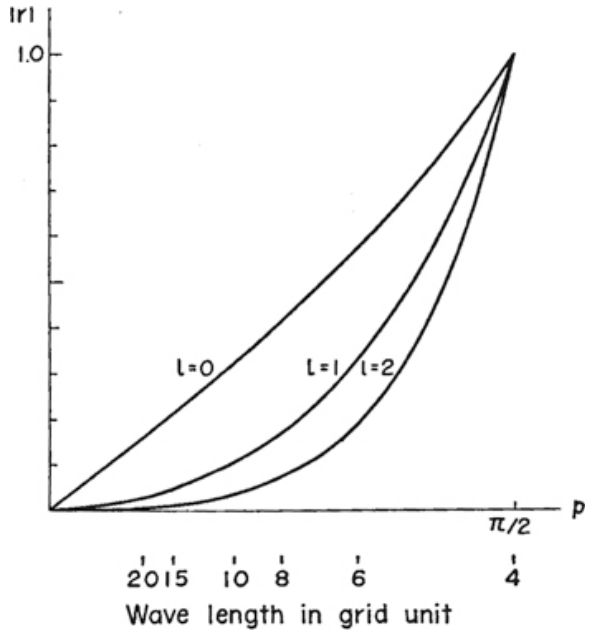


Figure 20.8: A graph of $|r|$ versus $k\Delta x$ for Methods 2 (labeled $l = 0$) and 5 (labeled $l = 1$). The curve labeled $l = 2$ corresponds to a scheme that is more accurate than any of Nitta's schemes. From Matsuno (1966).

With Methods 2 - 8, the computational mode is damped. Recall from Chapter 6 that any damping process is unstable with the leapfrog scheme. This explains why Platzman (1954) concluded *in an analysis based on the leapfrog scheme* that Method 1 is necessary for stability. But we don't have to (and should not) use the leapfrog scheme.

If we use Method 1 with the leapfrog scheme, the domain is quickly filled with small scale noise, but the noise remains stable. If we use Methods 5 or 7 with the leapfrog scheme, the domain will be littered with noise after a considerable length of time (depending on the width of the domain and u), but once the noise becomes noticeable, it will amplify and the model will blow up. This problem can be avoided by avoiding the leapfrog scheme.

With Method 1, all of the energy stays in the domain. With Methods 5 and 7 some of the energy is lost due to incomplete reflection at the outflow boundary. This energy loss paradoxically causes the leapfrog scheme to blow up. The situation is analogous to using the leapfrog scheme with a damping term, which was discussed in Chapter 6. The bottom line? Don't use the leapfrog scheme.

A more complete model with a larger domain would in fact permit energy to pass out through the artificial boundaries of the smaller domain considered here. Schemes that permit such loss, such as Methods 5 and 7, are therefore more realistic, if used with a suitable(i.e., non-leapfrog) time-differencing scheme.

20.4.3 Energy fluxes at outflow boundaries

The discussion above suggests that Nitta's schemes can also be analyzed in terms of the energy flux in the neighborhood of the outflow boundary. Multiplying the continuous one-dimensional advection equation by $2A$, we obtain

$$\frac{\partial A^2}{\partial t} + \frac{\partial}{\partial x}\left(uA^2\right) = 0\,. \tag{20.45}$$

This shows that A^2 is advected. Defining A^2 as the "energy," we see that uA^2 is the energy flux and $\partial\left(uA^2\right)/\partial x$ is the energy flux divergence. Now suppose that the advection equation is approximated by the differential difference equation:

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0\,. \tag{20.46}$$

Multiplying (20.46) by $2A_j$, we obtain

$$\frac{d}{dt}A_j^2 + \left(\frac{uA_jA_{j+1} - uA_jA_{j-1}}{\Delta x}\right) = 0\,. \tag{20.47}$$

Comparing (20.47) with (20.45), we see that uA_jA_{j+1} and uA_JA_{J-1} are the energy fluxes from grid point j to grid point $j+1$, and from grid point $j-1$ to grid point j, respectively. Applying (20.47) to the grid point $j+1$ (by adding one to each subscript) gives

$$\frac{d}{dt}A_{j+1}^2 + \left(\frac{uA_{j+1}A_{j+2} - uA_jA_{j+1}}{\Delta x}\right) = 0\,. \tag{20.48}$$

Inspection shows that the energy flux between j and $j+1$ is given by uA_jA_{j+1}. In the differential case, the sign of the energy flux is the same as the sign of u. This is not necessarily true for the differential-difference equation, however, because A_jA_{j+1} is not necessarily positive. When A_jA_{j+1} is negative, as when A oscillates from one grid point to the next, the direction of energy flow is opposite to the direction of u. This implies a negative group velocity u_g^* for $\frac{\pi}{2} < k\Delta x < \pi$, meaning that for short waves, for which $A_JA_{J-1} < 0$, energy flows in the $-x$ direction, i.e., "backward." This is consistent with our earlier analysis of the group velocity.

When we put an *artificial* boundary at $j = J$, and if we let $A_j = 0$ as in Nitta's Method 1, the energy flux from the point $J-1$ to the point J is zero. This is possible only when a computational mode, which transfers energy in the upstream direction, is superposed. This is a tip-off that Nitta's Method 1 is bad, but that is obvious anyway.

For Nitta's Method 2, $A_J = A_{J-1}$. This gives

$$uA_JA_{J-1} = uA_J^2 = uA_{J-1}^2 > 0\,. \tag{20.49}$$

Since energy can leave the domain, there is less reflection. Of course, using the present approach, the actual energy flux cannot be determined, because we do not know the value of A_J.

For Nitta's Method 4, $A_J = A_{J-2}$. Then for short waves

$$uA_JA_{J-1} = uA_{J-1}A_{J-2} < 0\,. \tag{20.50}$$

Short-wave energy moves back upstream, and the computational mode is strongly excited.

For Nitta's Method 5,

$$A_J = 2A_{J-1} - A_{J-2}\,, \tag{20.51}$$

so

$$\begin{aligned} uA_JA_{J-1} &= uA_{J-1}\left(2A_{J-1} - A_{J-2}\right) \\ &= u\left(2A_{J-1}^2 - A_{J-1}A_{J-2}\right)\,. \end{aligned} \tag{20.52}$$

As discussed above, for very short waves,

$$A_{J-1}A_{J-2} < 0\,, \tag{20.53}$$

so that the flux given by (20.52) is positive, as it should be. For very long waves,

$$A_{J-1}A_{J-2} \cong A_{J-1}A_{J-1}\,, \tag{20.54}$$

so the flux is approximately

$$uA_JA_{J-1} \cong uA_{J-1}^2 > 0\,. \tag{20.55}$$

For Nitta's Method 7,

$$\frac{dA_J}{dt} + u\left(\frac{A_J - A_{J-1}}{\Delta x}\right) = 0\,, \tag{20.56}$$

so we find that

$$\frac{dA_J^2}{dt} + 2u\left(\frac{A_J^2 - A_JA_{J-1}}{\Delta x}\right) = 0\,. \tag{20.57}$$

The energy flux "into J" is A_JA_{J-1}, while that "out of J" is $A_J^2 > 0$. Applying (20.47) to $J-1$, we find that

$$\frac{d}{dt}\left(A_{J-1}^2\right) + u\left(\frac{A_JA_{J-1} - A_{J-2}A_{J-1}}{\Delta x}\right) = 0\,. \tag{20.58}$$

Comparison of (20.57) and (20.58) shows that the energy flux out of $J-1$ is the same as the flux into J, which means that energy is conserved. This is good.

20.5 Nested grids

If we use a nested grid in which the grid spacing changes discontinuously at the boundary between a coarse grid and a fine grid, we will encounter a problem similar to the one met at the boundaries; a reflection occurs because the fine portion of the grid permits modes that are too short to be represented on the coarse portion of the grid. The problem can be minimized by various techniques, but it cannot be eliminated.

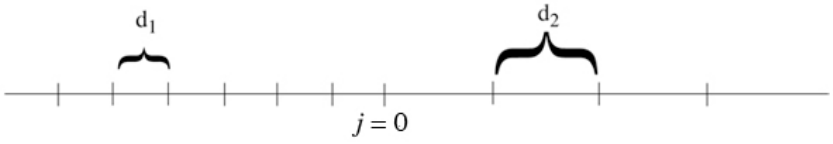


Figure 20.9: Schematic showing a change in the grid size at $j = 0$.

In this section, we consider the simple case in which the grid spacing changes discontinuously at $j = 0$, from d_1 to d_2, as shown in Fig. 20.9. The grid spacing is assumed to be uniform elsewhere. This example corresponds to what are called "nested" grids. The figure shows $d_2 > d_1$, but we will also consider the opposite case. We analyze the solution of the one-dimensional advection equation with a positive and spatially constant advecting current u. We use the following differential-difference equations:

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2d_1}\right) = 0 \text{ for } j < 0\,, \tag{20.59}$$

$$\frac{dA_0}{dt} + u\left[\alpha\left(\frac{A_0 - A_{-1}}{d_1}\right) + \beta\left(\frac{A_1 - A_0}{d_2}\right)\right] = 0 \text{ for } j = 0\,, \tag{20.60}$$

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2d_2}\right) = 0 \text{ for } j > 0\,. \tag{20.61}$$

In (20.60), α and β are weights, and we assume that

$$\alpha + \beta = 1\,. \tag{20.62}$$

If we choose $d_1 = d_2$ and $\alpha = \beta = 1/2$, then the scheme used at $j = 0$ reduces to the scheme used at all of the other grid points. The method used in (20.60) is not completely general, but it allows us to consider various choices for the weights α and β.

As a shorthand notation, define

$$k_1 d_1 = k_1 d_1, \text{ and } k_2 d_2 = k_2 d_2, \tag{20.63}$$

where k_1 is the *known* wave number of the incoming signal, moving in from the left, and k_2 is the wave number for $j \geq 0$. Since the incident wave must have $u_g^* > 0$ (this is what is meant by "incident"), we see from Fig. 20.3 that

$$0 < k_1 d_1 < \pi/2, \tag{20.64}$$

i.e., the wavelength of the incident wave has to be at least as long as $4d_1$.

Referring back to (20.40), we see that the solution for $j \leq 0$ is given by:

$$A_J = e^{i(jk_1 d_1 - \sigma t)} + r(-1)^j e^{-i(jk_1 d_1 + \sigma t)} , \tag{20.65}$$

where the frequency satisfies

$$\sigma = u \frac{\sin k_1 d_1}{d_1} . \tag{20.66}$$

In writing (20.65), we have assumed for simplicity that the incident wave has unit amplitude, and we define r as the amplitude of the reflected wave (relative to unity).

Similarly, the solution for $j \geq 0$ is

$$A_J = R e^{i(jk_2 d_2 - \sigma t)} , \tag{20.67}$$

where R is the amplitude of the transmitted wave and the frequency must satisfy

$$\sigma = u\frac{\sin k_2 d_2}{d_2} \ . \tag{20.68}$$

Note that *the frequency has to be the same throughout the domain*. This allows us to eliminate σ between (20.66) and (20.68), giving

$$\boxed{\frac{\sin k_1 d_1}{d_1} = \frac{\sin k_2 d_2}{d_2}} \ , \tag{20.69}$$

which relates $k_2 d_2$ to $k_1 d_1$, or k_2 to k_1.

At $j = 0$, the solutions given by (20.65) and (20.67) have to agree. This implies that

$$\boxed{1 + r = R} \ . \tag{20.70}$$

20.5.1 What does the downstream signal look like?

To determine the form of the solution for $j > 0$, we have to consider several cases:

1. $d_2/d_1 > 1$.

   Suppose that advecting current carries the signal from a finer grid onto a coarser grid. This can be expected to cause problems. Define

   $$\sin k_2 d_2 = (d_2/d_1)\sin k_1 d_1 \equiv a \ , \tag{20.71}$$

   so that $\cos k_2 d_2 = \pm\sqrt{1-a^2}$. From Euler's formula, this implies that

   $$e^{ik_2 d_2} = ia \pm \sqrt{1-a^2} \ . \tag{20.72}$$

   Since we can choose d_1, d_2, and k_1 any way we want, it is possible to make $|a|$ either greater than one or less than one. We consider these two possibilities separately.

(a) $a > 1$.

In this case $k_2 d_2$ has to be complex. From (20.72), we find that

$$\begin{aligned} e^{ik_2 d_2} &= i\left(a \pm \sqrt{a^2-1}\right) \\ &= e^{i\frac{\pi}{2}}\left(a \pm \sqrt{a^2-1}\right) . \end{aligned} \tag{20.73}$$

Using (20.73), the solution for $j \geq 0$ can be written as

$$A_j = R\left(a \pm \sqrt{a^2-1}\right)^j e^{i\left(\frac{\pi}{2} j - \sigma t\right)} \text{ for } j \geq 0 \tag{20.74}$$

Note the exponent, j, on the expression in parentheses. Since $a > 1$ by assumption, it is clear that $a + \sqrt{a^2-1} > 1$ and $a - \sqrt{a^2-1} < 1$. To ensure that A_j remains bounded as $j \to \infty$, we have to choose the minus sign. Then (20.74) reduces to

$$A_J = R\left(a - \sqrt{a^2-1}\right)^j e^{i\left(\frac{\pi}{2} j - \sigma t\right)} \text{ for} j \geq 0. \tag{20.75}$$

Eq. (20.75) describes a damped oscillation, as shown in Fig. 20.5. The wavelength of the transmitted is $4d_2$, regardless of the wavelength of the incident mode. The amplitude of the transmitted mode decreases as j increases. This is similar to our earlier result for what happens at an inflow boundary when the downstream grid is too coarse to be consistent with the imposed frequency and wind speed.

(b) $a \leq 1$.

In this case $k_2 d_2$ is real, and

$$|k_2 d_2| < \pi/2 , \tag{20.76}$$

which means that the transmitted wave has a wavelength longer than four times the grid spacing. The solution is

$$A_J = R\, e^{i(jk_2d_2 - \sigma t)} \,. \tag{20.77}$$

Since we are currently considering the case $d_2/d_1 > 1$, (20.69) implies that $k_2d_2 > k_1d_1$. We also have from (20.69) that

$$\frac{k_2}{k_1} = \frac{\operatorname{sinc}(k_1d_1)}{\operatorname{sinc}(k_2d_2)} \,. \tag{20.78}$$

Recall that $\operatorname{sinc}(x)$ is a decreasing function of x for $0 < x < \pi/2$. We conclude, then, that $k_2/k_1 > 1$. This means that *the wavelength of the transmitted wave is shorter than that of the incident wave, even though the transmitted wave is being advected on a coarser grid.*

2. $d_2/d_1 < 1$.

   Next, suppose that the advecting current blows from a coarser grid to a finer grid, which is a relatively benign situation. In this case, k_2d_2 is always real. The analysis is similar to (1b) above. It turns out that the wavelength of the transmitted wave is longer than that of the incident wave. Using the fact that $k_2d_2 \le \sin^{-1}(d_2/d_1)$, it can be shown that the minimum wavelength of the transmitted wave, which occurs for $k_1d_1 = \pi/2$, is

$$L_{\min} = \frac{2\pi d_2}{\sin^{-1}(d_2/d_1)} \,. \tag{20.79}$$

   As an example, for $d_2/d_1 = 1/2$, the minimum wavelength is $12d_2$.

20.5.2 Reflection and transmission

Recall that the amplitudes of the transmitted and reflected signals are denoted by R and r, respectively. We can use (20.65) and (20.67) to substitute for A_{-1}, A_0, and A_1 in (20.60). The result is

$$-i\sigma R + c\left\{\frac{\alpha}{d_1}\left[1 - e^{-ik_1d_1} + r\left(1 + e^{ik_1d_1}\right)\right] + \frac{\beta}{d_2}R\left(e^{ik_2d_2} - 1\right)\right\} = 0 \,. \tag{20.80}$$

Next, use (20.70) to eliminate r in (20.80), and solve for R:

$$R = \frac{2u\frac{\alpha}{d_1}\cos k_1 d_1}{-i\sigma + u\left[\frac{\alpha}{d_1}\left(1+e^{ik_1d_1}\right)+\frac{\beta}{d_2}\left(e^{ik_2d_2}-1\right)\right]} \,. \tag{20.81}$$

Now use (20.66) to eliminate σ. Also use (20.62) and (20.69). The result is

$$R = \frac{2\cos k_1 d_1}{1+\cos k_1 d_1 - \gamma(1-\cos k_2 d_2)} \,, \tag{20.82}$$

where we have defined

$$\gamma \equiv \left(\frac{\beta}{\alpha}\right)\left(\frac{d_1}{d_2}\right) \,. \tag{20.83}$$

Substituting (20.82) back into (20.70) gives the reflection coefficient as

$$r = -\left[\frac{1-\cos k_1 d_1 - \gamma(1-\cos k_2 d_2)}{1+\cos k_1 d_1 - \gamma(1-\cos k_2 d_2)}\right]. \tag{20.84}$$

Eq.s (20.82) and (20.84) are basic results. Ideally, we want to have $R = 1$ and $r = 0$.

As a check, suppose that $d_1 = d_2$ and $\alpha = \beta = 1/2$. Then $j = 0$ is "just another point," and so there should not be any computational reflection, and the transmitted wave should be identical to the incident wave. From (20.69), we see that for this case $k_2 = k_1$ and $k_1 d_1 = k_2 d_2$. Then (20.82) and (20.84) give $R = 1$, $r = 0$, i.e., complete transmission and no reflection, as expected. So it works.

20.5.3 Choosing the weights at a seam

For $\alpha \to 0$ with finite d_1/d_2, we get $\gamma \to \infty$, $R \to 0$, and $|r| \to 1$, unless $\cos k_2 d_2 = 1$, which is the case of an infinitely long wave, i.e., $k_2 d_2 = 0$. This is like Nitta's Method 1.

For $\beta \to 0$ with finite d_1/d_2, $\gamma \to \infty$, so that

$$\begin{aligned} R &\to \frac{2\cos k_1 d_1}{1+\cos k_1 d_1} = 1 - \tan^2\left(\frac{k_1 d_1}{2}\right), \\ r &\to -\left(\frac{1-\cos k_1 d_1}{1+\cos k_1 d_1}\right) = -\tan^2\left(\frac{k_1 d_1}{2}\right). \end{aligned} \tag{20.85}$$

This is like Nitta's Method 5.

If both $k_1 d_1 \to 0$ and $k_2 d_2 \to 0$, we get $R \to 1$ and $r \to 0$, regardless of the value of γ. When $k_1 d_1$ and $k_2 d_2$ are small but not zero, we get

$$\cos k_1 d_1 \cong 1 - k_1 d_1^2/2,\ \cos k_2 d_2 \cong 1 - k_2 d_2^2/2, \text{ and } k_2 d_2 \cong (d_2/d_1)\, k_1 d_1 . \tag{20.86}$$

Then we find that

$$\begin{aligned} R \cong \frac{2\left(1 - \frac{k_1 d_1^2}{2}\right)}{2 - \frac{k_1 d_1^2}{4} + \gamma \frac{k_2 d_2^2}{4}} &\cong \left(1 - \frac{k_1 d_1^2}{2}\right)\left(1 + \frac{k_1 d_1^2}{4} + \gamma \frac{k_2 d_2^2}{4}\right) \\ &\cong 1 - \frac{k_1 d_1^2}{4} + \gamma \frac{p^2{}_2}{4} \\ &\cong 1 - \frac{k_1 d_1^2}{4}\left[1 - \gamma (d_2/d_1)^2\right]. \end{aligned} \tag{20.87}$$

The choice

$$\gamma = (d_1/d_2)^2 \tag{20.88}$$

gives $R = 1 + \mathscr{O}\left(k_2 d_2^4\right)$. Referring back to (20.69), we see that this choice of γ corresponds to

$$\frac{\beta}{\alpha} = \frac{d_1}{d_2}\,. \tag{20.89}$$

This gives R close to one and $|r|$ close to zero. It can be shown that (20.89) is the requirement for second-order accuracy at the "seam" that connects the two grids. Since the given equations have second-order accuracy elsewhere, (20.89) essentially expresses the requirement that the order of accuracy be the same everywhere on the grid.

20.6 Boundary conditions for waves

What happens when a wave encounters a real wall? Consider an incident wave traveling toward the right with a certain wave number k_0, such that $0 < k_0 d < \pi$. For $\sigma \geq 0$, $e^{ik_0 dj} e^{-i\sigma t}$ represents such a wave. An additional wave, with $p = -k_0 d$, can be produced by reflection at a boundary. We assume that the amplitude of the incident wave with wavenumber k_0 is 1, and let r denote the amplitude of the reflected wave. In other words, we take $A = 1$ and $B = r$. Then (19.39) and (19.41) can be written as

$$u_{j+\frac{1}{2}} = \left[e^{ik_0 d\left(j+\frac{1}{2}\right)} + re^{-ik_0 d\left(j+\frac{1}{2}\right)} \right] e^{-i\sigma t} \ , \tag{20.90}$$

$$h_j = -\sqrt{\frac{H}{g}} \left(e^{ik_0 dj} - re^{-ik_0 dj} \right) e^{-i\sigma t} \ . \tag{20.91}$$

Suppose that at $j = J + \frac{1}{2}$ we have a real, physical wall, as shown in Fig. 20.1. Since there is no flow through the wall, we know that $u_{J+\frac{1}{2}} = 0$, for all time. This is a physical boundary condition. Then (20.90) reduces to

$$0 = e^{ik_0 d\left(J+\frac{1}{2}\right)} + re^{-ik_0 d\left(J+\frac{1}{2}\right)} \ , \tag{20.92}$$

which implies that

$$|r| = 1 \ . \tag{20.93}$$

This means that the reflected wave has the same amplitude as the incident wave. The reflection is "complete."

Recall from (19.22) that shallow-water gravity waves in one dimension (and without rotation) are governed by

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \ , \tag{20.94}$$

where c is the phase speed. Eq. (20.94) actually applies to a wide variety of waves, including but not restricted to gravity waves in shallow water. In Chapter 21, we will apply it to the study of sound waves. For any type of wave, c is the phase speed; the expression for c will of course depend on the wave type.

We now demonstrate this second-order wave equation is equivalent to a pair of first-order equations, *each of which describes the flow of information in just one direction*.

First, note that the solutions of (20.94) are constant along characteristics. The characteristics can, and generally do, intersect boundaries. As with the advection equation, $f(x-ct)$ is a solution of the wave equation (19.22), but $g(x+ct)$ is a second solution. Note that we are not assuming a single Fourier mode; the functions f and g can include many Fourier components. The general solution of (20.94) is given by a super-position of the left- and right-propagating solutions:

$$u(x,t) = f(x-ut) + g(x+ct) \ . \tag{20.95}$$

The forms of $f(x-ut)$ and $g(x+ct)$ are completely determined by the initial conditions, which can be written as

$$u(x,\,0) = f(x) \ , \tag{20.96}$$

$$\frac{\partial u}{\partial t}(x,\,0) = g(x) \ . \tag{20.97}$$

For the case of the shallow-water waves, $g(x)$ contains information about $h(x,0)$, so together these two initial conditions contain information about both the initial mass field and the initial wind field.

Substituting (20.95) into (20.96), we find that

$$\begin{aligned} f(x,\,0) + g(x,\,0) &= f(x)\,, \\ -cf'(x,\,0) + cg'(x,\,0) &= g(x) \ . \end{aligned} \tag{20.98}$$

Here a prime denotes differentiation. Differentiating the first of (20.98), and then using the second, we can solve for $f'(x,\,0)$ and $g'(x,\,0)$:

$$
\begin{aligned}
f'(x,\,0) &= \frac{1}{2}\left[F'(x) - \frac{g(x)}{c}\right], \\
g'(x,\,0) &= \frac{1}{2}\left[F'(x) + \frac{g(x)}{c}\right].
\end{aligned}
\tag{20.99}
$$

These can be integrated to obtain $f(x)$ and $g(x)$:

$$
\begin{aligned}
f(x,\,0) &= \frac{1}{2}\left[f(x) - \frac{1}{c}\int_0^x g(\xi)\,d\xi\right] + C_1, \\
g(x,\,0) &= \frac{1}{2}\left[f(x) + \frac{1}{c}\int_0^x g(\xi)\,d\xi\right] + C_2.
\end{aligned}
\tag{20.100}
$$

Here ξ is a dummy variable of integration, and C_1 and C_2 are constants of integration. Finally, we obtain $u(x,t)$ by replacing $(x,\,0)$ by $x-ct$ and $x+ct$, respectively, in the expressions for $f(x,\,0)$ and $g(x,\,0)$ in (20.100), and then substituting back into (20.95). This gives

$$
u(x,t) = \frac{1}{2}\left[f(x-ut) + f(x+ct) + \frac{1}{c}\int_{x-ct}^{x+ct} g(\xi)\,d\xi\right]. \tag{20.101}
$$

Here we have set $C_1 + C_2 = 0$ in order to satisfy $u(x,0) = f(x)$. As mentioned above, $g(x)$ contains information about $h(x,0)$. Obviously, that information is needed to predict $u(x,t)$, and it is in fact used on the right-hand side of (20.101).

Once $u(x,t)$ is known from (20.101), we can obtain $\partial h/\partial x$ from the momentum equation, (19.20). The constant "background" value of h cannot be determined without additional information.

In order to make connections between the wave equation, (20.94), and the advection equation that we have already analyzed, we define

$$
p \equiv \frac{\partial u}{\partial t}\,, \tag{20.102}
$$

and

$$q \equiv -c\frac{\partial u}{\partial x} \,. \tag{20.103}$$

Substitution of (20.102) into the wave equation (20.94) gives

$$\frac{\partial p}{\partial t} + c\frac{\partial q}{\partial x} = 0, \tag{20.104}$$

and differentiation of (20.103) with respect to t, with the use of (20.102), gives

$$\frac{\partial q}{\partial t} + c\frac{\partial p}{\partial x} = 0. \tag{20.105}$$

If we alternately add (20.104) to (20.105), and subtract (20.105) from (20.104), we obtain

$$\frac{\partial P}{\partial t} + c\frac{\partial P}{\partial x} = 0 \,, \tag{20.106}$$

and

$$\frac{\partial Q}{\partial t} - c\frac{\partial Q}{\partial x} = 0 \,, \tag{20.107}$$

respectively, where

$$P \equiv p + q \,, \tag{20.108}$$

where

$$Q \equiv p - q \,. \tag{20.109}$$

Now we have a system of two first-order equations, namely (20.106) and (20.107), each of which "looks like" an advection equation with an advecting current of magnitude $|c|$.

Note, however, that the "advections" are in opposite directions! Assuming that $c > 0$, P is "advected" towards increasing x, while Q is "advected" towards decreasing x. From (20.106) and (20.107), it is clear that P is constant along the line $x - ct =$ constant, and Q is constant along the line $x + ct =$ constant. Eqs. (20.106) and (20.107) are called the *normal forms* of (20.104) and (20.105).

To apply these ideas to the shallow water equations, set $p \equiv u$ and $q \equiv \sqrt{g/H}\, h$. With these substitutions, Eqs. (20.104) and (20.105) reduce to the shallow water equations (19.20) and (19.20), which are repeated here for your convenience:

$$\frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} = 0\,, \tag{20.110}$$

and

$$\frac{\partial h}{\partial t} + H\frac{\partial u}{\partial x} = 0\,. \tag{20.111}$$

Note the similarity of (20.110) - (20.111) to (20.104) and (20.105). We can make the similarity more complete by rewriting (20.110) - (20.111) as

$$\frac{\partial u}{\partial t} + \sqrt{gH}\frac{\partial}{\partial x}\left(\sqrt{g/H}\, h\right) = 0\,, \tag{20.112}$$

and

$$\frac{\partial}{\partial t}\left(\sqrt{g/H}\, h\right) + \sqrt{gH}\frac{\partial u}{\partial x} = 0\,. \tag{20.113}$$

Setting $P = u + \sqrt{g/H}\, h$ and $Q \equiv u - \sqrt{g/H}\, h$, we find that the normal forms of the one-dimensional shallow water equations without rotation are

$$\left(\frac{\partial}{\partial t} + \sqrt{gH}\frac{\partial}{\partial x}\right)\left(u + \sqrt{g/H}\, h\right) = 0\,, \tag{20.114}$$

and

$$\left(\frac{\partial}{\partial t} - \sqrt{gH}\frac{\partial}{\partial x}\right)\left(u - \sqrt{g/H}\, h\right) = 0\,. \tag{20.115}$$

Inspection of (20.114) shows that $u + \sqrt{g/H}\, h$ is carried towards positive x, and inspection of (20.115) shows that $u - \sqrt{g/H}\, h$ is carried towards negative x. Therefore, we must not specify $u + \sqrt{g/H}\, h$ on a "right" wall, and we must not specify $u - \sqrt{g/H}\, h$ on a "left" wall. To prevent waves from propagating into the domain, the best option is to specify $u + \sqrt{g/H}\, h = 0$ on the left wall, and $u - \sqrt{g/H}\, h = 0$ on the right wall. Then, if h is given on the wall, we can solve for u, or vice versa.

20.7 The effects of a mean flow

We now generalize our system of equations to include advection by a mean flow U, in the following manner:

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)u + g\frac{\partial h}{\partial x} = 0, \tag{20.116}$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)A = 0\,, \tag{20.117}$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)h + H\frac{\partial u}{\partial x} = 0\,. \tag{20.118}$$

We have also added a tracer, A. The normal forms of these equations are

$$\left[\frac{\partial}{\partial t} + \left(U + \sqrt{gH}\right)\frac{\partial}{\partial x}\right]\left(u + \sqrt{g/H}\, h\right) = 0\,, \tag{20.119}$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)A = 0\,, \tag{20.120}$$

$$\left[\frac{\partial}{\partial t} + \left(U - \sqrt{gH}\right)\frac{\partial}{\partial x}\right]\left(u - \sqrt{g/H}\, h\right) = 0. \tag{20.121}$$

We assume that $\sqrt{gH} > |U|$, which is often true in the atmosphere. For (20.119) and (20.121), the lines $x - \left(U + \sqrt{gH}\right)t =$ constant and $x - \left(U - \sqrt{gH}\right)t =$ constant are the characteristics, and are shown as the solid lines in Fig. 20.10. Everything is similar to the case without advection, except that now the slopes of the two characteristics differ not only in sign but also in magnitude, because one wave is propagating with the wind and the other is propagating against the wind.

We also have an additional equation, namely (20.120). This is an advection equation, and so A is a constant along the lines $x - Ut =$ constant, which are shown schematically by the broken lines in Fig. 20.10. We should specify A only on the inflow boundary.

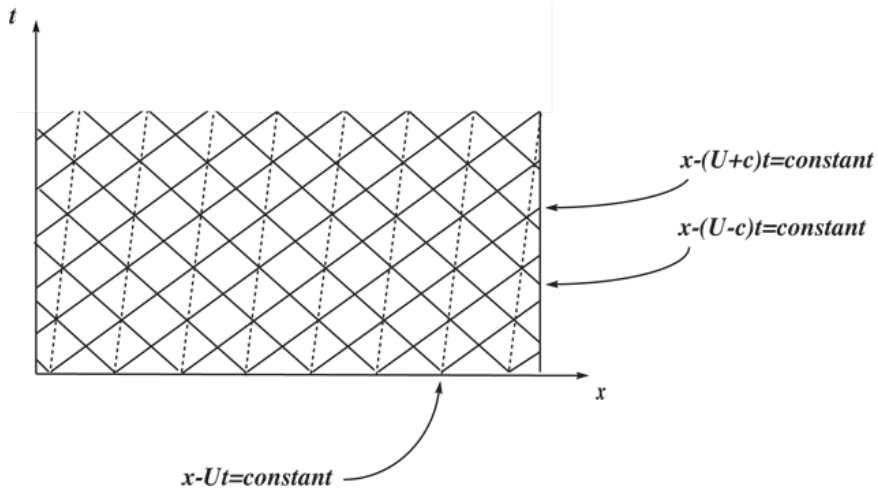


Figure 20.10: Characteristics for the case of shallow water wave propagation with an advecting current U.

20.8 Summary

20.9 Problems

1. Program the discontinuous-grid model given discussed in the text, i.e.,

$$\begin{aligned}
\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2d_1}\right) &= 0, \text{ for } j < 0\,, \\
\frac{dA_0}{dt} + u\left[\alpha\left(\frac{A_0 - A_{-1}}{d_1}\right) + \beta\left(\frac{A_1 - A_0}{d_2}\right)\right] &= 0, \quad \alpha + \beta = 1, \text{ for } j = 0\,, \\
\frac{dA_j}{dt} + u\left(\frac{A_{j+1} - A_{j-1}}{2d_2}\right) &= 0, \text{ for } j > 0\,,
\end{aligned} \tag{20.122}$$

where

$$-\alpha d_1 + \beta d_2 = 0 \text{ and } \alpha + \beta = 1\,. \tag{20.123}$$

Replace the time derivatives by leapfrog time differencing. Consider the case $d_1 = 2d_2$. Let $\mu = 1/2$ on the finer of the two grids. Use a periodic domain whose width is 100 times the larger of the two grid spacings. The periodicity means that there will be two discontinuities. Use a "square bump" initial condition that is ten points wide on the finer grid. Run the model long enough for the signal to circle the domain twice. Discuss the evolution of the results.

Chapter 21

The sound of silence

21.1 How sound waves work

The simplest set of linearized equations that can support sound waves is this:

$$\overline{\rho}\frac{\partial \mathbf{V}'}{\partial t} = -\nabla p' \,, \tag{21.1}$$

$$\frac{\partial \rho'}{\partial t} = -\overline{\rho}\,\nabla\cdot\mathbf{V}' \,, \tag{21.2}$$

$$\theta' = 0 \,. \tag{21.3}$$

These equations describe linear, elastic, isentropic wave motions in a basic state in which the potential temperature and density are the same everywhere. From (21.3) and the equation of state, we can show that

$$\frac{\rho'}{\overline{\rho}} = \frac{1}{\gamma}\frac{p'}{\overline{p}} \,, \tag{21.4}$$

where

$$\gamma = \frac{c_p}{c_v} \cong 1.4 \, . \tag{21.5}$$

Using (21.4) in (21.2), we can rewrite the continuity equation in terms of the pressure perturbation:

$$\frac{\partial p'}{\partial t} = -\gamma \overline{p} \, \nabla \cdot \mathbf{V}' \, . \tag{21.6}$$

We can now combine (21.1) and (21.6) to obtain a wave equation:

$$\frac{\partial^2 p'}{\partial t^2} = \gamma R \overline{T} \, \nabla^2 p' \, . \tag{21.7}$$

We recognize $\gamma R \overline{T}$ as the square of the phase speed, c, and conclude that the speed of sound is given by

$$|c| = \sqrt{\gamma R \overline{T}} \, . \tag{21.8}$$

For $\gamma = 1.4$, $R = 287$ J kg^{-1} K^{-1}, and $T = 280$ K, we find that $|c| \cong 334$ m s^{-1}, which is much faster than typical wind speeds, and also much faster than the phase speeds of other waves.

If a model includes vertically propagating sound waves then, with explicit time differencing, the largest time step that is compatible with linear computational stability can be quite small. For example, with a vertical grid spacing on the order of 300 m, the allowed time step will be on the order of one second. This may be acceptable if the horizontal grid spacing is comparably small. On the other hand, with a horizontal grid spacing of 30 km and a vertical grid spacing of 300 m, vertically propagating sound waves will limit the time step to about one percent of the value that would be compatible with the horizontal grid spacing. That's hard to take.

21.2 Coping with acoustic waves

There are four ways to reduce or eliminate the time-step constraints associated with sound waves. Two of them involve modifying the continuous equations to eliminate or "filter"

sound wave solutions. The other two methods involve numerical approximations in the discrete equations.

Nonhydrostatic system sof equations that filter sound waves are generically referred to as "anelastic" systems, although one particular version is called "*the* anelastic system." Anelastic models of various types are very widely used, especially for high-resolution modeling. There are several types of anelastic systems, developed over a period of fifty years or so (e.g., Ogura and Phillips, 1962; Lipps and Hemler, 1982; Durran, 1989; Bannon, 1996; Arakawa and Konor, 2009). Some of the systems filter both vertically and horizontally propagating sound waves, while other "partially" anelastic systems filter only the vertically propagating sound waves without affecting the horizontally propagating sound waves. Filtering sound waves introduces errors and can also interfere with conservation properties. Fortunately, the newest anelastic systems are both accurate and conservative.

A second approach is to adopt the quasi-static system of equations, in which the equation of vertical motion is replaced by the hydrostatic equation. The quasi-static system filters vertically propagating sound waves, while permitting Lamb waves, which are sound waves that propagate only in the horizontal. The quasi-static approximation has been widely used in global models for both weather prediction and climate, but its errors become unacceptably large for some small-scale weather systems, so its use is limited to models with horizontal grid spacings on the order of about 10 km or larger, depending on the particular application.

The third approach is to use implicit or partially implicit time differencing, which can permit a long time step even when vertically propagating sound waves occur. There is a family of schemes called "HEVI", which stands for "horizontally explicit, vertically implicit."

The fourth approach is to "sub-cycle" Klemp and Wilhelmson (1978). This means that small time steps are used to integrate the terms of the equations that govern sound waves, i.e., the terms that qppear in Eqs. (21.1) - (21.3), while longer time steps are used for the remaining terms. This is also called "time-splitting."

21.3 Acoustic filters

21.3.1 The anelastic and Boussinesq systems

In order to obtain the wave equation, (21.7) from the two first-order equations (21.1) and (21.2), the time-derivative terms of (21.1) and (21.2) are essential. If either one of the time-derivative terms is neglected, the sound waves are eliminated.

The anelastic system neglects the time derivative term in the continuity equation, so that (21.2) reduces to

$$\nabla \cdot \mathbf{V}' = 0 \,, \tag{21.9}$$

which says that the wind is nondivergent. Combining (21.9) with (21.1), we find that

$$\nabla^2 p' = 0 \,. \tag{21.10}$$

This means that in order to ensure that the wind field remains nondivergent, as required by (21.9), the Laplacian of the pressure field must vanish. It should be clear that the use of (21.9) prevents the derivation of a wave equation; i.e., it filters sound waves. Versions of the anelastic system were derived by Ogura and Phillips (1962) and Lipps and Hemler (1982).

Up to now we have used a highly simplified linearized system. Now consider the more complete versions of the momentum and continuity equations, given by

$$\frac{\partial}{\partial t}(\rho \mathbf{V}) = -\nabla p + \text{other terms} \,, \tag{21.11}$$

and

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{V}) \,. \tag{21.12}$$

In the full anelastic system, (21.11) and (21.12) are replaced by

$$\frac{\partial}{\partial t}(\rho_0 \mathbf{V}) \cong -\nabla p + \text{other terms} \,, \tag{21.13}$$

and

$$0 \cong \nabla \cdot (\rho_0 \mathbf{V}) \,, \tag{21.14}$$

where ρ_0 is a "reference state" density *that depends only on height*. The use of such a reference state introduces errors and is one of the weaknesses of the anelastic system.

Taking the divergence of (21.13), and using (21.14) we find that

$$\nabla^2 p = \nabla \cdot (\text{other terms}) \ . \tag{21.15}$$

The physical interpretation is that the pressure field takes whatever shape is needed to ensure that the three-dimensional mass flux remains non-divergent over time. The unknown in (21.15) is the pressure. It can be solved for the pressure using the methods discussed in Chapter 15.

The anelastic continity equation can also be written as

$$\nabla_z \cdot (\rho_0 \mathbf{v}) + \frac{\partial}{\partial z} (\rho_0 w) \cong 0 \ . \tag{21.16}$$

Here ∇_z acts on a constant-height surface, $\mathbf{v}$ is the horizontal wind vector, and w is the vertical component of the velocity. An anelastic model predicts $\mathbf{v}$, and uses (21.16) to determine w.

Bannon (1995, 1996) discussed other versions of the anelastic system. Durran (1989, 2008) and Durran and Arakawa (2007) proposed an improved anelastic system, called the "pseudo-incompressible system."

For the case of shallow motion, such as eddies in the boundary layer, (21.14) can be further simplified to

$$0 \cong \nabla \cdot \mathbf{V} \ . \tag{21.17}$$

This is called the Boussinesq approximation (Spiegel and Veronis, 1960). It is essentially a special case of the anelastic approximation. It is often used in boundary-layer models, and in oceanography.

21.3.2 The quasi-static system

The quasi-static system drops the time-derivative term in the equation of motion, (21.1), *but only for the vertical component of the velocity*. The time-derivative term is retained for

the horizontal momentum equation. The result is that vertically propagating sound waves are eliminated, but horizontally propagating sound waves are retained.

With an appropriate boundary condition, the hydrostatic approximation,

$$\partial p/\partial z = -\rho g \,, \tag{21.18}$$

allows us to compute $p(z)$ from $\rho(z)$. Even when the air is moving, the hydrostatic approximation gives a good approximation to $p(z)$, simply because Dw/Dt and the other terms of the equation of vertical motion are almost always very small compared to g. The hydrostatic approximation is applicable to virtually all meteorological phenomena, including even violent thunderstorms.

For large-scale circulations, the approximate pressure determined through the use of the hydrostatic approximation can be used to compute the pressure gradient force in the equation of horizontal motion. This is called the *quasi-static* approximation. It is accurate for large-scale motions, but it distorts small-scale motions, such as thunderstorms and gravity waves. When the quasi-static approximation is made, the effective kinetic energy is due to the horizontal wind only; the contribution of the vertical component, w, is neglected. For large-scale motions, the vertical motion is very weak compared to the horizontal wind speed, so that this quasi-static kinetic energy is very close to the true kinetic energy. Further discussion is given by Holton (1973).

When the equation of vertical motion is replaced by the hydrostatic approximation, there are two fundamental changes. First, we can no longer use the equation of vertical motion to determine the vertical velocity. As discussed in a later chapter, the method that is actually used to obtain the vertical motion depends on the vertical coordinate system used. This is expected because the actual *meaning* of the mass flux across iso-surfaces of the vertical coordinate also depends on the vertical coordinate used. For example, in the case of pressure coordinates, the vertical motion is $\omega \equiv Dp/Dt$, which can be determined using the pressure-coordinate version of the continuity equation, with the upper boundary condition $\omega = 0$ at $p = 0$. Further discussion is given in Chapter 23.

The Lamb wave is a horizontally propagating sound wave, and it is a solution of the quasi-static system. Lamb waves are constantly present in the atmosphere, but usually with small amplitudes. Large-amplitude Lamb waves are sometimes generated by volcanic explosions, such as the eruption of Hunga Tonga-Hunga Ha'apai (Abbrescia et al., 2022; Dalal et al., 2023). Quasi-static models can simulate Lamb waves, but anelastic models can't.

The quasi-static system does not give accurate solutions for high frequencies and small spatial scales, but it is very useful in atmospheric models with horizontal grid spacings larger than about 20 km.

21.4 The Unified System

Arakawa and Konor (2009) derived a nonhydrostatic system of equations that filters vertically propagating sound waves, while allowing the Lamb wave. They called it the "Unified System," although Dubos and Voitus (2014) call it the "semi-hydrostatic system." The key to the Unified System is to replace the density by the "quasi-static density."

21.4.1 The quasi-static sounding

We can *define* a quasi-static density and a quasi-static pressure without actually *using* the hydrostatic approximation. The use of the quasi-static pressure was also advocated by Miller (1974) and Laprise (1992a). Here's how it works:

Suppose that the potential temperature is predicted by a model that uses height as its vertical coordinate. A convenient form of the hydrostatic system is

$$\frac{\partial \Pi_{\rm qs}}{\partial z} \equiv -\frac{g}{\theta} \,, \tag{21.19}$$

where

$$\Pi_{\rm qs} \equiv c_p \left(\frac{p_{\rm qs}}{p_0} \right)^{\kappa} \,. \tag{21.20}$$

Eq. (21.19) is the definition of $\Pi_{\rm qs}$. Eq. (21.19) can be used to compute $p_{\rm qs}$ given a boundary condition such as the surface value of $\Pi_{\rm qs}$. Then $p_{\rm qs}$ can be computed from (21.20). We use

$$\frac{\partial p_{\rm qs}}{\partial z} = -\rho_{\rm qs} g \tag{21.21}$$

to determine $\rho_{\rm qs}$. These quasi-static values of Π, p, and ρ do not comprise a reference state, because they are determined from the predicted value of θ, which can vary in any way. By combining (21.19), (21.20), and (21.21), we find that

$$\rho_{\rm qs} = \frac{p_{\rm qs}}{\kappa \Pi_{\rm qs} \theta} \,. \tag{21.22}$$

As mentioned above, we need the value of $(\Pi_{\rm qs})_S$, which is the surface value of $\Pi_{\rm qs}$, in order to determine $\Pi_{\rm qs}(z)$ by vertical integration of (21.19). Arakawa and Konor (2009) show how to predict $(\Pi_{\rm qs})_S$ by using the vertically integrated mass budget.

21.4.2 The continuity equation for the quasi-static density

In height coordinates, the continuity equation for the Unified System is

$$\frac{\partial \rho_{\rm qs}}{\partial t} + \nabla_z \cdot (\rho_{\rm qs} \mathbf{v}) + \frac{\partial}{\partial z}(\rho_{\rm qs} w) = 0 \,. \tag{21.23}$$

Compare with (21.16). Eq. (21.23) looks like it would be used to predict $\rho_{\rm qs}$, but that's not right! As discussed above, $\rho_{\rm qs}$ can be computed diagnostically from θ and $\Pi_{\rm qs}$. If $\rho_{\rm qs}$ is known at two successive time levels, as it would be in the middle of a simulation, then its finite-difference time-tendency can also be diagnosed, by subtraction. What, then, is the unknown in (21.23)? It's w. In the Unified System, just like the anelastic system, we vertically integrate the continuity equation to diagnose the vertical velocity. This is the *only* approximation of the Unified System. It filters vertically propagating sound waves, but not horizontally propagating sound waves.

For further discussion of the Unified System see see Dukowicz (2013), Konor (2014) Dubos and Voitus (2014), Voitus et al. (2019), and Qaddouri et al. (2021).

21.5 Dispersion curves for the various systems of equations

Figure 21.1 shows the dispersion curves for the various systems of equations discussed in this chapter. In each case, the equations have been linearized about a resting basic state in which the temperature is uniform with height. The index n is a nondimensional vertical wave number, such that $n = 1$ means one node in the vertical, etc. The Lamb wave corresponds to $n = 0$.

The figure shows that all four approximate systems omit the vertically propagating sound waves. The quasi-static approximation gives unrealistic high-frequency inertia-gravity waves for large values of k. The anelastic and pseudo-incompressible approximations omit the Lamb wave, and distort the frequencies of the inertia-gravity waves. The Unified System retains the Lamb wave, and gives more accurate frequencies for the inertia-gravity waves.

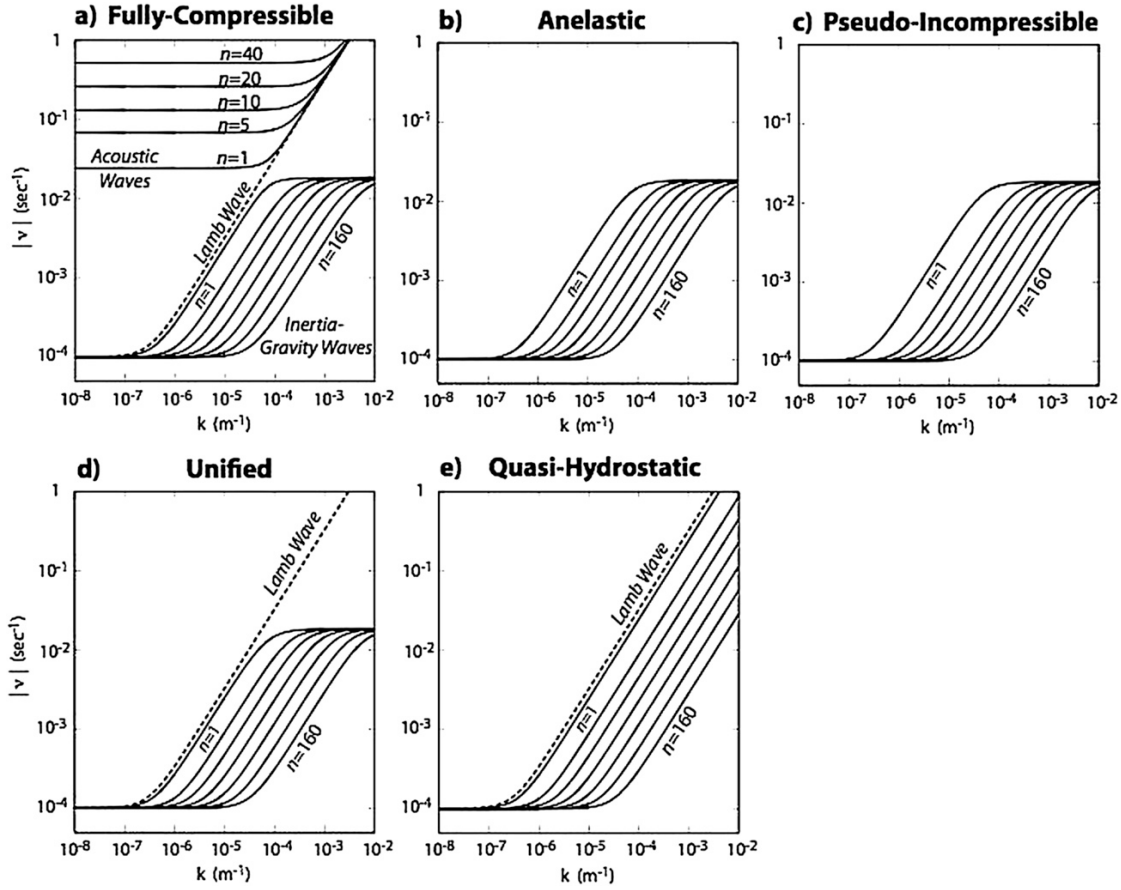


Figure 21.1: The absolute frequencies $|\nu|$ of normal modes on an f plane, as functions of horizontal wavenumber k, for (a) the fully compressible, (b) anelastic, (c) pseudo-incompressible, (d) unified, and (e) quasi-hydrostatic systems. The index n is a nondimensional vertical wave number. The equations have been linearized about an isothermal resting basic state. From Arakawa and Konor (2009).

Chapter 22

Stairways to heaven

To construct a numerical model, we have to make a lot of choices, including these:

- The governing equations: Quasi-static or not? Shallow atmosphere or not? Anelastic or not?
- The vertical coordinate system;
- The vertical staggering of the model's dependent variables;
- The properties of the continuous equations that we want the discrete equations to mimic.

The first two choices can be discussed in the context of the continuous system. Chapter 21 discussed the first choice. This chapter deals with second choice. Chapters 23 and 24 deal with the remaining two choices, which involve the discrete system. As usual, every choice will have strengths and weaknesses. We must also pay attention to possible interactions between the vertical differencing and the horizontal and temporal differencing.

22.1 The third dimension is special

Up to now we have ignored the vertical structure of the atmosphere, because it brings several issues that don't come up in modeling the horizontal structure:

- Gravitational effects strongly control vertical motions, and gravitational potential energy is the source of atmospheric kinetic energy.
- The Earth's atmosphere is very shallow compared to its horizontal extent, so that *on large horizontal scales* vertical gradients are much stronger than horizontal gradients, and horizontal motions are much faster than vertical motions. (On small horizontal scales vertical and horizontal gradients can be comparable, and the horizontal and vertical wind components can be comparable in magnitude.) Strong vertical gradi-

ents can only be simulated accurately with a small vertical grid spacing. The small vertical grid spacing can require short time steps to maintain computational stability.

- The atmosphere has a complex lower boundary that can strongly influence the circulation through both mechanical blocking and thermal forcing. Surface fluxes of momentum, energy, and moisture act at the lower boundary. Models have to enforce appropriate lower boundary conditions.
- The density of the air decreases exponentially upward, gradually giving way to the vacuum of space. Models have to deal with this by using upper boundary conditions.

For these reasons, it is important to consider the vertical structure of the atmosphere separately from the horizontal structure.

22.2 Choosing a vertical coordinate system

The most obvious choice of vertical coordinate system is height. As you probably already know, the equations of motion are frequently expressed using vertical coordinates other than height, such as pressure, normalized pressure (usually called σ), and potential temperature.

The most basic requirement for a variable to be used as a vertical coordinate is that it must change monotonically with height. Even this requirement can be relaxed, slightly. For example, a vertical coordinate can be independent of height over some layer of the atmosphere, provided that the layer is not too deep.

Factors to be weighed in choosing a vertical coordinate system for a particular application include the following:

- the form of the lower boundary conditions (simpler is better);
- the form of the continuity equation (simpler is better);
- the form of the horizontal pressure gradient force (simpler is better, and a pure gradient is particularly good);
- the form of the hydrostatic equation, which matters even for non-hydrostatic models (simpler is better);
- the method used to compute the vertical motion (simpler is better);
- the intensity of the vertical motion as seen in the coordinate system (weaker is better).

In Chapter 23, each of these factors will be discussed for particular vertical coordinates.

Kasahara (1974) published a detailed discussion of general vertical coordinates for *quasi-static* models. A more modern discussion of the same subject is given by Konor

and Arakawa (1997). In this chapter, we discuss general vertical coordinates for *nonhydrostatic* models. Naturally, the quasi-static limit can be recovered as a special case.

22.3 The basic equations in height coordinates

In Chapter 2, we presented the basic equations that govern the atmosphere using the three-dimensional velocity vector, $\mathbf{V}$, and without using any vertical (or horizontal) coordinate system. Given the unique character of the vertical coordinate, as discussed in Section 22.1 above, it is advantageous (I'm tempted to say "necessary") to treat the horizontal and vertical motion separately. Therefore, in this section, we rewrite the basic equations using height as the vertical coordinate, while still avoiding a horizontal coordinate system. We adopt the usual symbol z for height relative to some reference level at or near the Earth's surface.

First, we separate the three-dimensional wind vector into a horizontal wind vector, $\mathbf{v}$, and the vertical velocity, $w\mathbf{k}$:

$$\mathbf{V} = \mathbf{v} + w\mathbf{k} \,. \tag{22.1}$$

With the use of (22.1), the Lagrangian time derivative can be written as

$$\frac{D}{Dt} \equiv \left(\frac{\partial}{\partial t}\right)_z + \mathbf{v}\cdot\nabla_z + w\frac{\partial}{\partial z} \,. \tag{22.2}$$

Here the partial derivative with respect to time means the time rate of change at a fixed point in space, because z is a true spatial coordinate. With a different vertical coordinate, the partial derivative with respect to time means the time rate of change at a fixed horizontal location and on a particular isosurface of the vertical coordinate. As mentioned earlier, the Lagrangian time rate of change operator, D/Dt, has a meaning that is independent of coordinate system. Height coordinates have been used in Eq. (22.2), but D/Dt can be expressed in other coordinate systems too, as will be shown later in this chapter.

With this preparation, the equations of horizontal and vertical motion can be expressed in z coordinates as

$$\rho\left(\frac{D\mathbf{v}}{Dt} + 2\boldsymbol{\Omega}\times\mathbf{v} + 2\Omega w\cos\varphi\,\mathbf{e}_\lambda\right) = -\nabla_z p - \nabla_z\cdot\mathbf{S} \,, \tag{22.3}$$

and

$$\rho\left(\frac{Dw}{Dt}-2\Omega\cos\varphi u\right)=-\frac{\partial p}{\partial z}-\rho g-\mathbf{e}_r\frac{\partial}{\partial z}\left[\mathbf{e}_r\cdot(\nabla\cdot\mathbf{S})\right]\ , \tag{22.4}$$

where u is the zonal component of the velocity. We will use these forms of the continuity equation, the thermodynamic equation, and the latent heat equation:

$$\left(\frac{\partial\rho}{\partial t}\right)_z+\nabla_z\cdot(\rho\mathbf{v})+\frac{\partial}{\partial z}(\rho w)=0 \tag{22.5}$$

$$\rho\frac{D}{Dt}(c_pT)=\rho\omega\alpha+\rho LC-\nabla_z\cdot\mathbf{R}-\frac{\partial}{\partial z}(\mathbf{R}\cdot\mathbf{e}_r)+\rho\delta\ , \tag{22.6}$$

$$\rho\frac{D}{Dt}(Lq_v)=-\rho LC\ . \tag{22.7}$$

22.4 A general vertical coordinate

22.4.1 Up is up

Consider an arbitrary vertical coordinate denoted by $\hat{z}$. We assume that $\hat{z}$ is a *monotonic* differentiable function of z. We also assume that $\hat{z}$ can be differentiated with respect to time and the horizontal coordinates. The dimensions of $\hat{z}$ can be different from those of z. For example, $\hat{z}$ could be pressure or potential temperature. Of course, we include the possibility that $\hat{z}\equiv z$.

In general, isosurfaces of $\hat{z}$ are not parallel to isosurfaces of z. The direction of $\mathbf{k}$ is perpendicular to isosurfaces of z, and therefore not necessarily perpendicular to the isosurfaces of $\hat{z}$. *The meaning and direction of* $\mathbf{k}$ *are not affected by the choice of vertical coordinate system*; regardless of the choice of $\hat{z}$, "*up is up*." Similarly, the meanings of $\mathbf{v}$ and the horizontal and vertical pressure-gradient forces are independent of the choice of vertical coordinate system.

22.4.2 The pseudodensity

We now define the "pseudodensity," $\rho_{\hat{z}}$, which is given by

$$\rho_{\hat{z}} \equiv \rho \partial z / \partial \hat{z} \,. \tag{22.8}$$

The pseuodensity is the amount of mass (per unit horizontal area) between two $\hat{z}$-surfaces. Note that

$$\rho \partial z = \rho_{\hat{z}} \partial \hat{z} \,, \tag{22.9}$$

and

$$\frac{1}{\rho}\frac{\partial}{\partial z} = \frac{1}{\rho_{\hat{z}}}\frac{\partial}{\partial \hat{z}} \,. \tag{22.10}$$

The actual form of the pseudodensity depends on the vertical coordinate system used. For example, with height coordinates $\rho_z = \rho$. If $\hat{z}$ is the quasi-static pressure defined in Section 21.4, then we replace ρ by $\rho_{\rm qs}$ in (22.8), and write

$$\begin{aligned} \rho_{p_{\rm qs}} =& \rho_{\rm qs} \partial z / \partial p_{\rm qs} \\ =& -1/g \,. \end{aligned} \tag{22.11}$$

In this case, the pseudodensity is independent of height.

22.5 Transforming to a general vertical coordinate

The first step in transforming (22.3) - (22.7) to generalized vertical coordinates is to multiply each of them by $\partial z/\partial \hat{z}$, and use (22.8). The results are:

$$\rho_{\hat{z}} \left(\frac{D\mathbf{v}}{Dt} + 2\boldsymbol{\Omega} \times \mathbf{v} + 2\Omega w \cos\varphi \mathbf{e}_\lambda \right) = -\frac{\partial z}{\partial \hat{z}} \nabla_z p - \frac{\partial z}{\partial \hat{z}} \nabla_z \cdot \mathbf{S} \,, \tag{22.12}$$

$$\rho_{\hat{z}}\left(\frac{Dw}{Dt}-2\Omega\cos\varphi u\right)=-\frac{\partial p}{\partial\hat{z}}-\rho_{\hat{z}}g-\mathbf{e}_r\frac{\partial}{\partial\hat{z}}\left[\mathbf{e}_r\cdot(\nabla\cdot\mathbf{S})\right]\,, \tag{22.13}$$

$$\frac{\partial z}{\partial\hat{z}}\left(\frac{\partial\rho}{\partial t}\right)_z+\frac{\partial z}{\partial\hat{z}}\nabla_z\cdot(\rho\mathbf{v})+\frac{\partial}{\partial\hat{z}}(\rho w)=0\,, \tag{22.14}$$

$$\rho_{\hat{z}}\frac{D}{Dt}(c_pT)=\rho_{\hat{z}}\omega\alpha+\rho_{\hat{z}}LC-\frac{\partial z}{\partial\hat{z}}\nabla_z\cdot\mathbf{R}-\frac{\partial}{\partial\hat{z}}(\mathbf{R}\cdot\mathbf{e}_r)+\rho_{\hat{z}}\delta\,, \tag{22.15}$$

$$\rho_{\hat{z}}\frac{D}{Dt}(Lq_v)=-\rho_{\hat{z}}LC\,. \tag{22.16}$$

Then, using (A.62), (A.65) and (A.70), we can rewrite the equations as

$$\begin{aligned}\rho_{\hat{z}}\left(\frac{D\mathbf{v}}{Dt}+2\boldsymbol{\Omega}\times\mathbf{v}+2\Omega w\cos\varphi\,\mathbf{e}_\lambda\right)=&-\left[\nabla_{\hat{z}}\left(p\frac{\partial z}{\partial\hat{z}}\right)-\frac{\partial}{\partial\hat{z}}(p\nabla_{\hat{z}}z)\right]\\&-\left[\nabla_{\hat{z}}\cdot\left(\mathbf{S}\frac{\partial z}{\partial\hat{z}}\right)+\frac{\partial}{\partial\hat{z}}(\mathbf{S}\cdot\nabla_{\hat{z}}z)\right]\,,\end{aligned} \tag{22.17}$$

$$\rho_{\hat{z}}\left(\frac{Dw}{Dt}-2\Omega\cos\varphi u\right)=-\frac{\partial p}{\partial\hat{z}}-\rho_{\hat{z}}g-\frac{\partial}{\partial\hat{z}}\left[\mathbf{e}_r\cdot(\nabla\cdot\mathbf{S})\right]\,, \tag{22.18}$$

$$\left(\frac{\partial\rho_{\hat{z}}}{\partial t}\right)_{\hat{z}}+\nabla_{\hat{z}}\cdot(\rho_{\hat{z}}\mathbf{v})+\frac{\partial}{\partial\hat{z}}(\rho_{\hat{z}}\hat{w})=0\,, \tag{22.19}$$

$$\rho_{\hat{z}}\frac{D}{Dt}(c_pT)=\rho_{\hat{z}}\omega\alpha+\rho_{\hat{z}}LC-\nabla_{\hat{z}}\cdot\left(\mathbf{R}\frac{\partial z}{\partial\hat{z}}\right)+\frac{\partial}{\partial\hat{z}}\left[\mathbf{R}\cdot\nabla_{\hat{z}}z-(\mathbf{R}\cdot\mathbf{e}_r)\right]+\rho_{\hat{z}}\delta\,, \tag{22.20}$$

$$\rho_{\hat{z}} \frac{D}{Dt} (Lq_v) = -\rho_{\hat{z}} LC \, . \tag{22.21}$$

In (22.19),

$$\boxed{\hat{w} \equiv -\frac{\partial \hat{z}}{\partial z} \left[\left(\frac{\partial z}{\partial t} \right)_{\hat{z}} + \mathbf{v} \cdot \nabla_{\hat{z}} z - w \right] \, .} \tag{22.22}$$

The product $\rho_{\hat{z}} \hat{w}$ is the rate at which mass is moving upward across the surface $z = \hat{z}$. For $\hat{z} \equiv z$ Eq. (22.22) reduces to $\hat{w} = w$. Notice that (22.22) can be rearranged to

$$w = \left(\frac{\partial z}{\partial t} \right)_{\hat{z}} + \mathbf{v} \cdot \nabla_{\hat{z}} z + \hat{w} \frac{\partial z}{\partial \hat{z}} \, . \tag{22.23}$$

This says that w is the Lagrangian change of z, expressed in $\hat{z}$ coordinates.

22.6 ALE

Eq. (22.22) is used in is called the "Arbitrary Lagrangian-Eulerian" (ALE) method (Hirt et al., 1974). ALE is arbitrary[1] in the sense that any physically useful rule can be used to determine $(\partial z/\partial t)_{\hat{z}}$. ALE can be used in many ways:

- In the limiting "Lagrangian" case, we specify $\hat{w} = 0$, This means that no mass crosses the *hatz* surface, which is simply advected by the wind. We use (22.22) in the form

$$\left(\frac{\partial z}{\partial t} \right)_{\hat{z}} = -\mathbf{v} \cdot \nabla_{\hat{z}} z + w \, , \tag{22.24}$$

  to predict the height of the advected $\hat{z}$ surface.

- We could creatively modify (22.24) so that the advected surfaces don't intersect the lower boundary or the model top, and don't get too close together or too far apart (e.g., Toy and Randall, 2009). Here the motivation would be to let the *hatz* surfaces be advected by the wind *except* when it causes problems.

[1] A better word (with a worse acronym) would be "flexible."

- If $\hat{z} = \theta$, then $\hat{w}$ is $\dot{\theta}$, which is proportional to the *parameterized* heating rate per unit mass. Given $\dot{\theta}$, we use (22.24) to predict $z(\theta)$.

For practical reasons, we might choose to impose upper and/or lower limits on the height of the $\hat{z}$ surfaces. We might also want to enforce a minimum amount of mass between neighboring $\hat{z}$ surfaces. We can use (22.22) to determine the value of $\hat{w}$ required to avoid violating such rules. This is an application of the first approach. Toy and Randall (2009) present another example of the two approaches in combination.

ALE does not define a vertical coordinate system. Instead, it defines a vertical grid, which can evolve with time. We have already discussed triangular and hexagonal horizontal grids that are not associated with a coordinate system. Further discussion is given in Chapter 28.

22.7 Boundary conditions on the continuity equation

Let $\hat{z}_S$ and $\hat{z}_T$ be the values of $\hat{z}$ at the Earth's surface and the top of the model, respectively. We allow the possibility that $\hat{z}_T$ is at a finite height above the Earth's surface. For example, if $\hat{z}$ is height, we might place the top of the model at 80 km above sea level. If $\hat{z}$ is pressure, the corresponding isobaric surface would be close to 1 Pa. We also consider the possibility that both $\hat{z}_S$ and $\hat{z}_T$ vary with horizontal location and time.

Integration of (22.19) through the entire depth of the model, and use of Leibniz's rule, gives

$$\begin{aligned} &\frac{\partial}{\partial t}\int_{\hat{z}_S}^{\hat{z}_T} \rho_{\hat{z}}\, d\hat{z} - (\rho_{\hat{z}})_T \frac{\partial \hat{z}_T}{\partial t} + (\rho_{\hat{z}})_S \frac{\partial \hat{z}_S}{\partial t} \\ +\nabla\cdot\int_{\hat{z}_S}^{\hat{z}_T} (\rho_{\hat{z}}\mathbf{v})\, d\hat{z} &- (\rho_{\hat{z}}\mathbf{v})_T\cdot\nabla\hat{z}_T + (\rho_{\hat{z}}\mathbf{v})_S\cdot\nabla\hat{z}_S \\ &+ (\rho_{\hat{z}}\hat{w})_T - (\rho_{\hat{z}}\hat{w})_S = 0\,. \end{aligned} \tag{22.25}$$

The condition that no mass crosses the top of the model can be written as

$$(\rho_{\hat{z}})_T \frac{\partial \hat{z}_T}{\partial t} + (\rho_{\hat{z}}\mathbf{v})_T\cdot\nabla\hat{z}_T - (\rho_{\hat{z}}\hat{w})_T = 0\,. \tag{22.26}$$

If the top of the model is a surface of constant $\hat{z}$, which is usually the case, then (22.26) reduces to

$$\hat{w}_T = 0\,. \tag{22.27}$$

Similarly, the condition that no mass crosses the Earth's surface is expressed by

$$(\rho_{\hat{z}})_S \frac{\partial \hat{z}_S}{\partial t} + (\rho_{\hat{z}}\mathbf{v})_S \cdot \nabla \hat{z}_S - (\rho_{\hat{z}}\hat{w})_S = 0. \tag{22.28}$$

With the use of (22.26) and (22.28), Eq. (22.25) simplifies to

$$\frac{\partial}{\partial t}\int_{\hat{z}_S}^{\hat{z}_T} \rho_{\hat{z}}\, d\hat{z} + \nabla \cdot \int_{\hat{z}_S}^{\hat{z}_T} (\rho_{\hat{z}}\mathbf{v})\, d\hat{z} = 0\,, \tag{22.29}$$

which expresses conservation of mass for the entire column of air.

22.8 The vertically integrated pressure-gradient force

22.8.1 The coordinate-free case

Let **VIPGF** be the vertical integral of the pressure-gradient term of (2.16). This integral represents the effects of the pressure gradient on the whole column of air. The lower limit of integration is the Earth's surface, where $z = z_S$, and the upper limit is $z = z_T$, which is the (possibly infinite) height of a model's top. We can write

$$\begin{aligned} \mathbf{VIPGF} &\equiv -\int_{z_S}^{z_T} \nabla p\, dz \\ &= -\nabla\left(\int_{z_S}^{z_T} p,\, dz\right) - p_S \nabla z_S + p_T \nabla z_T\,. \end{aligned} \tag{22.30}$$

Because of the vertical integration, each term of (22.30) is independent of height. On the second line of (22.30) we take the integral inside the gradient operator, and allow for the possible spatial variations of z_S and z_T.

We now integrate each term of (22.30) around *an arbitrary closed path*, which could be a circle of constant latitude. The first term on the right-hand side of the second line of (22.30) is a gradient, and therefore does not contribute to the line integral. There are only two non-zero terms:

$$\oint_C \mathbf{VIPGF} \cdot d\mathbf{l} = -\oint_C p_S \nabla z_S \cdot d\mathbf{l} + \oint_C p_T \nabla z_T \cdot d\mathbf{l}\,. \tag{22.31}$$

Here C is the closed path, and $d\mathbf{l}$ is a (vector) increment of distance along C. Eq. (22.31) shows under what conditions the vertically integrated pressure-gradient force and spin up or spin down a circulation contained within C. The first or "surface" term on the right-hand side of (22.31) vanishes if *either* p_S *or* z_S is constant along C. This tells us that *the surface term vanishes in the absence of topography*, i.e., if z_S is constant along C. We conclude that in the absence of topography the surface term cannot spin up or spin down a circulation contained within C.

For $z_T \to \infty$ we have $p_T \to 0$, so the second or "top" term on the right-hand side of (22.31) vanishes. In a model with a finite z_T, the top term can potentially spin up or spin down a circulation contained within C. It is important to prevent such a thing because the real atmosphere has no top. We see that the top term vanishes if *either* p_T *or* z_T is constant along C. *We should, therefore, choose either* $z_T =$ *constant or* $p_T =$ *constant.* Hydrostatically, the assumption p_T = constant means that the amount of mass above the model top is the same everywhere and at all times.

Further discussion is given later.

22.8.2 With the general vertical coordinate

From (22.17), we see that the with the general vertical coordinate the vertically integrated horizontal pressure-gradient force is given by

$$\begin{aligned} \mathbf{VIHPGF} &= -\int_{\hat{z}_S}^{\hat{z}_T} \left[\frac{\partial}{\partial \hat{z}} \left(z \nabla_{\hat{z}} p \right) - \nabla_{\hat{z}} \left(z \frac{\partial p}{\partial \hat{z}} \right) \right] d\hat{z} \\ &= \left[(z\nabla p)_S - (z\nabla p)_T \right] + \nabla \left(\int_{\hat{z}_S}^{\hat{z}_T} z \frac{\partial p}{\partial \hat{z}} \, d\hat{z} \right) - \left(z \frac{\partial p}{\partial \hat{z}} \right)_T \nabla \hat{z}_T + \left(z \frac{\partial p}{\partial \hat{z}} \right)_S \nabla \hat{z}_S \, . \end{aligned} \tag{22.32}$$

In order to obtain agreement with (22.30), the last two terms on the bottom line of (22.32) must vanish, leaving

$$\mathbf{VIHPGF} = \left[(z\nabla p)_S - (z\nabla p)_T \right] + \nabla \left(\int_{\hat{z}_S}^{\hat{z}_T} z \frac{\partial p}{\partial \hat{z}} \, d\hat{z} \right) \, . \tag{22.33}$$

This will be the case if $\hat{z}_T$ and $\hat{z}_S$ are both constant. As discussed in Chapter 23, $\hat{z}_T$ and $\hat{z}_S$ are, in fact, both constant for σ-coordinates, sigma-pressure coordinates, and the Klemp vertical coordinate. A vertical coordinate for which $\hat{z}_T$ and $\hat{z}_S$ are *not* both constant can lead to a spurious spin-up or spin-down of the circulation enclosed by an arbitrary closed curve.

22.9 Energy conservation

22.9.1 The coordinate-free case

Energy conservation is a very important issue in numerical models, especially for long simulations. Energy conservation also helps to maintain numerical stability. The two types of energy that we will consider are mechanical energy and thermodynamic energy. The sum of these is the total energy.

There are two distinct but closely related issues. The first is whether the total energy is actually conserved by the model. The second is whether the change in mechanical energy is exactly compensated by the change in thermodynamic energy.

To form the mechanical equation, we take the dot product of $\mathbf{V}$ with (2.16), giving

$$\rho \frac{DK}{Dt} = -\rho \mathbf{V} \cdot \nabla \phi - \mathbf{V} \cdot \nabla p - \mathbf{V} \cdot (\nabla \cdot \mathbf{S}) \ , \tag{22.34}$$

where

$$K \equiv \frac{1}{2} (\mathbf{V} \cdot \mathbf{V}) \tag{22.35}$$

is the kinetic energy per unit mass. We now assume that the geopotential ϕ depends only on r, the distance from the Earth's center, and is independent of latitude and longitude. This is not precisely true, but it is a good approximation that is used in virtually all atmospheric models. If ϕ depends only on r, then

$$\begin{aligned} \mathbf{V} \cdot \nabla \phi =& \frac{D\mathbf{r}}{Dt} \cdot \left(\frac{d\phi}{dr} \mathbf{e}_r \right) \\ =& \frac{D\phi}{Dt} \ . \end{aligned} \tag{22.36}$$

Using (22.36), we can rewrite (22.34) as

$$\rho \frac{D}{Dt} (K + \phi) = -\mathbf{V} \cdot \nabla p - \mathbf{V} \cdot (\nabla \cdot \mathbf{S}) \ , \tag{22.37}$$

This is a form of the "mechanical energy equation."

The pressure-gradient term of (22.37) can be rewritten as

$$\begin{aligned} -\mathbf{V}\cdot\nabla p &= -\nabla\cdot(p\mathbf{V}) + p\nabla\cdot\mathbf{V} \\ &= -\nabla\cdot(p\mathbf{V}) + \frac{p}{\alpha}\frac{D\alpha}{Dt} \\ &= -\nabla\cdot(p\mathbf{V}) + \frac{1}{\alpha}\left[\frac{D}{Dt}(p\alpha) - \alpha\frac{Dp}{Dt}\right] \\ &= -\nabla\cdot(p\mathbf{V}) + \rho\left[\frac{D}{Dt}(RT) - \alpha\omega\right] . \end{aligned} \tag{22.38}$$

To obtain the second line of (22.38), we have used the continuity equation in the form (2.21). The term $pD\alpha/Dt$ represents the work done by expansion ($D\alpha/Dt > 0$) or compression ($D\alpha/Dt < 0$). Note that a similar term appears in (2.22).

In a similar way, the friction term of (22.37) can be expanded to reveal two physically distinct parts:

$$-\mathbf{V}\cdot(\nabla\cdot\mathbf{S}) = -\nabla\cdot\mathbf{W} - \rho\delta\ . \tag{22.39}$$

Here

$$\mathbf{W} \equiv \mathbf{V}\cdot\mathbf{S} \tag{22.40}$$

is the (three-dimensional vector) flux of energy due to frictional work, and

$$\rho\delta \equiv -(\mathbf{S}\cdot\nabla)\cdot\mathbf{V} \geq 0 \tag{22.41}$$

is the (scalar) rate of kinetic energy dissipation. Because $\nabla\cdot\mathbf{W}$ is a divergence, it represents a spatial redistribution of mechanical energy as friction causes neighboring air parcels to do work on each other. It does not change the total amount of mechanical energy in the atmosphere, except where friction does work on the lower boundary. In contrast, the dissipation rate, δ, is a true sink of mechanical energy. As discussed in Chapter 2, the dissipation rate appears as a source of thermodynamic energy. It is, therefore, an energy *conversion* term.

Substituting (22.38) and (22.39) into (22.37), we obtain an alternative form of the mechanical energy equation:

$$\rho \frac{D}{Dt}(K+\phi-RT)+\nabla\cdot(p\mathbf{V}+\mathbf{W})=\rho(\omega\alpha)-\rho\delta\,. \tag{22.42}$$

Here we have placed the flux divergence terms on the left-hand side of the equation, along with the Lagrangian time rate of change. The true sources and sinks of mechanical energy appear on the right-hand side. We could cancel the ρ and α in the $\omega\alpha$ term, but choose not to.

To complete our derivation of the total energy equation, we need the enthalpy form of the thermodynamic equation, (2.23), and the latent energy equation, (2.32). These are repeated here for convenience:

$$\frac{D}{Dt}(c_pT)-\omega\alpha=LC-\nabla\cdot\mathbf{R}+\delta\,, \tag{22.43}$$

$$\frac{D}{Dt}(Lq_v)=-LC\,. \tag{22.44}$$

We now add (22.42), (22.43), and (22.44) to obtain the total energy equation in the form

$$\rho\frac{De_{\rm tot}}{Dt}+\nabla\cdot(p\mathbf{V}+\mathbf{W}+\mathbf{R})=0\,, \tag{22.45}$$

where

$$e_{\rm tot}\equiv K+\phi+c_vT+Lq_v \tag{22.46}$$

is the total energy per unit mass. The $\omega\alpha$ and dissipation terms of (22.43) and (22.42) have cancelled, because those terms represent rates of conversion between mechanical energy and thermodynamic energy. The condensation terms have also cancelled, because they represent conversion between latent energy and thermodynamic energy.

We can rewrite (22.45) in flux form, separating the horizontal and vertical velocity terms, as

$$\left[\frac{\partial}{\partial t}(\rho e_{\text{tot}})\right]_z + \nabla_z \cdot [\rho \mathbf{v} e_{\text{tot}}] + \frac{\partial}{\partial z}(\rho w e_{\text{tot}}) + \nabla \cdot (p\mathbf{V} + \mathbf{S}\cdot\mathbf{V} + \mathbf{R}) = 0\,. \tag{22.47}$$

22.9.2 Energy conservation with the generalized vertical coordinate

As shown in Appendix C, for the non-hydrostatic case the total energy equation with the general vertical coordinate is

$$\begin{aligned}\left[\frac{\partial}{\partial t}(\rho_{\hat{z}} e_{\text{tot}})\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot [\rho_{\hat{z}} \mathbf{v}(e_{\text{tot}} + p\alpha)] + \frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}(e_{\text{tot}} + p\alpha) + p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] \\ = -\nabla \cdot (\mathbf{W} + \mathbf{R})\,.\end{aligned} \tag{22.48}$$

When the quasi-static approximation is used, the result is rather different:

$$\begin{aligned}\left[\frac{\partial}{\partial t}\rho_{\hat{z}}(K + c_p T + L q_v)\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot [\rho_{\hat{z}} \mathbf{v}(K + h)] + \frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}(K + c_p T + \phi + L q_v) - z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] \\ = -\nabla \cdot (\mathbf{W} + \mathbf{R})\,.\end{aligned} \tag{22.49}$$

Compare with (22.48).

22.10 The vorticity equation

Let $q_{\hat{z}} \equiv (\mathbf{k} \cdot \nabla_{\hat{z}} \times \mathbf{v}) + f$ be the vertical component of the absolute vorticity. Note that *the meaning of $q_{\hat{z}}$ depends on the choice of $\hat{z}$*, because the curl of the velocity is taken along a $\hat{z}$-surface. Starting from the momentum equation, we can derive the vorticity equation in the form

$$\begin{aligned}\left(\frac{\partial q_{\hat{z}}}{\partial t}\right)_{\hat{z}} + (\mathbf{v} \cdot \nabla_{\hat{z}})\, q_{\hat{z}} + \hat{w}\frac{\partial q_{\hat{z}}}{\partial \hat{z}} = \\ -q_{\hat{z}}(\nabla_{\hat{z}} \cdot \mathbf{v}) + \frac{\partial \mathbf{v}}{\partial \hat{z}} \times (\nabla_{\hat{z}} \hat{w}) - \mathbf{k} \cdot \left[\nabla_{\hat{z}} \times \left(\nabla_{\hat{z}} p - \frac{\partial p}{\partial z}\nabla_{\hat{z}} z\right)\right] + \mathbf{k} \cdot (\nabla_{\hat{z}} \times \mathbf{F})\,.\end{aligned} \tag{22.50}$$

The first term on the right-hand side of (22.50) represents the effects of stretching, and the second represents the effects of twisting. When the HPGF can be written as a gradient, it has no effect in the vorticity equation, because the curl of a gradient is always zero, *provided that the curl and gradient are taken along the same isosurfaces*. In general, however, the HPGF is not simply a gradient along a $\hat{z}$-surface. When $\hat{z}$ is such that the HPGF is not a gradient, it can spin up or spin down a circulation on a $\hat{z}$-surface.

22.11 Segue

The next chapter surveys the vertical coordinates used in practice.

Chapter 23

A survey of vertical coordinates

In this chapter, we discuss ten particular vertical coordinates:

- height, z
- pressure, p, and also log-pressure, z^*, which is used in many theoretical studies and some numerical models (e.g., Girard et al., 2014)
- σ, defined by

$$\sigma = \frac{p - p_T}{p_S - p_T} \,, \tag{23.1}$$

  which is designed to simplify the lower boundary condition;
- a "hybrid," or "mix," of σ and p coordinates, used in many global circulation models, including the model of the European Centre for Medium Range Weather Forecasts;
- a terrain-following height coordinate proposed by Klemp (2011);
- η, which is a modified σ coordinate, defined by

$$\eta \equiv \left(\frac{p - p_T}{p_S - p_T} \right) \eta_S \,, \tag{23.2}$$

  where η_S is a time-independent function of the horizontal coordinates;
- potential temperature, θ, which has many attractive properties;
- entropy, $\varepsilon = c_p \ln(\theta/\theta_0)$, where θ_0 is a constant reference value;

- a hybrid $\sigma - \theta$ coordinate, which behaves like σ near the Earth's surface, and like θ away from the Earth's surface.

23.1 Height coordinates

The nonhydrostatic system in height coordinates is very straightforward. Here we discuss the quasi-static system in height coordinates.

23.1.1 The continuity equation

In height coordinates, the hydrostatic equation is

$$\frac{\partial p}{\partial z} = -\rho g \,. \tag{23.3}$$

As discussed in Section 21.3.2, when we use the quasi-static system, we can only predict (time-step) one thermodynamic variable. Choices include ρ, T, and θ.

The continuity equation in height coordinates is

$$\left(\frac{\partial \rho}{\partial t}\right)_z + \nabla_z \cdot (\rho \mathbf{v}) + \frac{\partial}{\partial z}(\rho w) = 0 \,. \tag{23.4}$$

This equation has a clear physical interpretation, but it is nonlinear and involves the time derivative of the density, which decreases exponentially with height.

The lower boundary condition in height coordinates is

$$\frac{\partial z_S}{\partial t} + \mathbf{v}_S \cdot \nabla z_S - w_S = 0 \,. \tag{23.5}$$

Normally we can assume that z_S is independent of time, but (23.5) can accommodate the effects of a specified time-dependent value of z_S (e.g., to represent the effects of an earthquake, or a wave on the sea surface). Because height surfaces intersect the Earth's surface, height-coordinates are relatively difficult to implement in numerical models. This complexity is mitigated somewhat by the fact that the horizontal spatial coordinates where the height surfaces meet the Earth's surface are normally independent of time.

23.1.2 The thermodynamic equation

The thermodynamic energy equation in height coordinates can be written as

$$c_p\rho\left(\frac{\partial T}{\partial t}\right)_z = -c_p\rho\left(\mathbf{v}\cdot\nabla_z T + w\frac{\partial T}{\partial z}\right) + \omega + \rho Q\,. \tag{23.6}$$

Here

$$\begin{aligned}\omega &= \left(\frac{\partial p}{\partial t}\right)_z + \mathbf{v}\cdot\nabla_z p + w\frac{\partial p}{\partial z}\\ &= \left(\frac{\partial p}{\partial t}\right)_z + \mathbf{v}\cdot\nabla_z p - \rho g w\,.\end{aligned} \tag{23.7}$$

By using (23.7) in (23.6), we find that

$$c_p\rho\left(\frac{\partial T}{\partial t}\right)_z = -c_p\rho\mathbf{v}\cdot\nabla_z t - \rho w c_p\left(\Gamma_d - \Gamma\right) + \left[\left(\frac{\partial p}{\partial t}\right)_z + \mathbf{v}\cdot\nabla_z p\right] + \rho Q\,, \tag{23.8}$$

where the actual lapse rate and the dry-adiabatic lapse rate are given by

$$\Gamma \equiv -\frac{\partial T}{\partial z}\,, \tag{23.9}$$

and

$$\Gamma_d \equiv \frac{g}{c_p}\,, \tag{23.10}$$

respectively. Eq. (23.8) is awkward because it involves the time derivatives of both T and p. The time derivative of the pressure can be eliminated by using the height-coordinate version of (23.94), which is

$$\left(\frac{\partial p}{\partial t}\right)_z = -g\nabla_z\cdot\int_z^{z_T}(\rho\mathbf{v})\,dz + g\rho\left(z\right)w\left(z\right) + \frac{\partial p_T}{\partial t}. \tag{23.11}$$

Substitution into (23.8) gives

$$\begin{aligned} c_p\rho\left(\frac{\partial T}{\partial t}\right)_z &= -c_p\rho\mathbf{v}\cdot\nabla_z t - \rho w c_p\left(\Gamma_d-\Gamma\right) \\ &\quad + \left[-g\nabla_z\cdot\int_z^{z_T}\left(\rho\mathbf{v}\right)dz + g\rho\left(z\right)w\left(z\right)+\frac{\partial p_T}{\partial t}\right]+\mathbf{v}\cdot\nabla_z p+\rho Q\,. \end{aligned} \tag{23.12}$$

According to (23.12), the time rate of change of the temperature at a given height is influenced by the convergence of the horizontal wind field through the layer above. The reason is that convergence above causes a pressure increase, which leads to compression, which warms.

An alternative, considerably simpler form of the thermodynamic energy equation in height coordinates is

$$\left(\frac{\partial\theta}{\partial t}\right)_z = -\left(\mathbf{v}\cdot\nabla_z\theta + w\frac{\partial\theta}{\partial z}\right)+\frac{Q}{\Pi}\,. \tag{23.13}$$

Given these choices, prediction of θ looks like the best option.

23.1.3 Richardson's equation

We need the vertical velocity, w, for vertical advection, among other things. In quasi-static models based on height coordinates, the equation of vertical motion is replaced by the hydrostatic equation, in which w does not even appear. How then can we determine w? It has to be computed using "Richardson's equation," which is an expression of the physical fact that hydrostatic balance applies not just at a particular instant, but continuously through time. The derivation of Richardson's equation is complicated. Here it comes:

As the state of the atmosphere evolves, the temperature, pressure, and density all change, at a location in the three-dimensional space. Many complicated and somewhat independent processes contribute to these changes, and it is easy to imagine that a hydrostatically balanced initial state would quickly be pushed out of balance. Balance is actually maintained over time through a process called hydrostatic adjustment (e.g., Bannon, 1995). The statement that balance is maintained leads to Richardson's equation. It can be derived by starting from the equation of state, in the form

$$p=\rho RT\,. \tag{23.14}$$

"Logarithmic differentiation" of (23.14) with respect to time gives

$$\frac{1}{p}\left(\frac{\partial p}{\partial t}\right)_z = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial t}\right)_z + \frac{1}{T}\left(\frac{\partial T}{\partial t}\right)_z . \tag{23.15}$$

The time derivatives can be eliminated by using continuity (23.4), the thermodynamic energy equation (23.8) and the pressure tendency equation (23.11). Note that the derivation of (23.11) involves use of the hydrostatic equation. After some manipulation, we find that

$$\begin{aligned} c_p T \frac{\partial}{\partial z}(\rho w) + \rho w \left[g\frac{c_v}{R} + c_p(\Gamma_d - \Gamma)\right] = (&-c_p \rho \mathbf{v}\cdot\nabla_z T + \mathbf{v}\cdot\nabla_z p) \\ &- c_p T\, \nabla_z\cdot(\rho\mathbf{v}) + g\frac{c_v}{R}\nabla_z \cdot \int_z^\infty (\rho\mathbf{v})\, dz' + \rho Q . \end{aligned} \tag{23.16}$$

where

$$c_v \equiv c_p - R \tag{23.17}$$

is the specific heat of air at constant volume.

Eq. (23.16) has been arranged so that the vertical velocity appears in both terms on the left-hand side, but not at all on the right-hand side. Expand the first term on the left-hand side using the product rule:

$$c_p T \frac{\partial(\rho w)}{\partial z} = \rho c_p T \frac{\partial w}{\partial z} + w c_p T \frac{\partial \rho}{\partial z} , \tag{23.18}$$

A second logarithmic differentiation of (23.14), this time with respect to height, gives

$$\frac{1}{p}\frac{\partial p}{\partial z} = \frac{1}{\rho}\frac{\partial \rho}{\partial z} + \frac{1}{T}\frac{\partial T}{\partial z} . \tag{23.19}$$

Using the hydrostatic equation again, we can rewrite (23.19) as

$$\begin{aligned}\frac{1}{\rho}\frac{\partial \rho}{\partial z} &= -\frac{\rho g}{p} + \frac{\Gamma}{T} \\ &= \frac{1}{T}\left(-\frac{g}{R} + \Gamma\right) .\end{aligned} \tag{23.20}$$

Substitution of (23.20) into (23.18) gives

$$c_p T \frac{\partial}{\partial z}(\rho w) = \rho c_p T \frac{\partial w}{\partial z} + \rho w c_p \left(-\frac{g}{R} + \Gamma\right) . \tag{23.21}$$

Finally, using (23.21) in (23.16), and combining terms, we obtain

$$\boxed{\begin{aligned}\frac{\partial w}{\partial z} = &\left(\frac{-c_p \rho \mathbf{v}\cdot\nabla_z T + \mathbf{v}\cdot\nabla_z p}{\rho c_p T}\right) - \frac{1}{\rho}\nabla_z\cdot(\rho\mathbf{v}) \\ &+ \frac{c_v}{c_p p}\left[g\nabla_z \cdot \int_z^{z_T} (\rho\mathbf{v})\, dz' - \frac{\partial p_T}{\partial t}\right] + \frac{Q}{c_p T} .\end{aligned}} \tag{23.22}$$

This beast is Richardson's equation. It can be integrated to obtain $w(z)$, given a lower boundary condition and the information needed to compute the various terms on the right-hand side, which involve both the mean horizontal motion and the heating rate, as well as various horizontal derivatives. A physical interpretation of (23.22) is that *the vertical motion is whatever it takes to maintain hydrostatic balance through time*, despite the fact that the various processes represented on the right-hand side may (individually) tend to upset that balance.

As a very simple illustration of the use of (23.22), suppose that we have no horizontal motion. Then (23.22) drastically simplifies to

$$\frac{\partial w}{\partial z} = \frac{1}{c_p}\left(\frac{Q}{T} - \frac{c_v}{p}\frac{\partial p_T}{\partial t}\right) . \tag{23.23}$$

If $w = 0$ at both the surface z_S and the finite top height z_T, then the pressure at the model top changes according to

$$\frac{\partial p_T}{\partial t} = \frac{\int_{z_S}^{z_T} \frac{Q}{c_v T}\, dz'}{\int_{z_S}^{z_T} \frac{1}{p}\, dz'} \ , \tag{23.24}$$

This shows that the addition of heat causes the pressure at the model top to increase with time, like the internal pressure acting on the lid of a pressure-cooker. The vertical velocity satisfies

$$w(z) = \int_0^z \left(\frac{Q}{c_p t} - \frac{c_v}{c_p p} \frac{\partial p_T}{\partial t} \right) dz' \ , \tag{23.25}$$

which says that heating (cooling) below a given level induces rising (sinking) motion at that level, as the air expands (or contracts) above the rigid lower boundary.

The complexity of Richardson's equation has discouraged the use of height coordinates in quasi-static models; one of the very few exceptions was the early NCAR GCM (Kasahara and Washington, 1967). We are now entering an era of non-hydrostatic global models, in which use of height coordinates is becoming more common, but of course Richardson's equation is not needed (and cannot be used) in non-hydrostatic models.

23.2 Pressure and log pressure coordinates

23.2.1 The hydrostatic pressure

In a non-hydrostatic model, the pressure can be literally "noisy," because it plays an essential role in acoustic waves. For this reason, it is useful to define a "hydrostatic pressure" that can be used as a vertical coordinate, even though the actual dynamics of the model are non-hydrostatic. A method to do this was presented already in Chapter 21. We define the non-hydrostatic "perturbation" pressure and density, δp and $\delta \rho$, by

$$p \equiv p_{\rm qs} + \delta p \tag{23.26}$$

and

$$\rho \equiv \rho_{\rm qs} + \delta \rho \ , \tag{23.27}$$

respectively. Notice that $\delta\theta = 0$ because, as discussed in Chapter 21, we have assumed that the model predicts (i.e., time-steps) the potential temperature.

23.2.2 The hydrostatic pressure as a vertical coordinate

In the following discussion we will drop the subscript "qs" on the hydrosatic pressure.

The hydrostatic equation in pressure coordinates is

$$\frac{\partial \phi}{\partial p} = -\alpha \, . \tag{23.28}$$

Eq. (22.8) tells us that the pseudodensity in pressure coordinates is

$$\rho_p = \rho \frac{\partial z}{\partial p} \, , \tag{23.29}$$

which reduces to

$$\rho_p \cong -1/g \tag{23.30}$$

when the hydrostatic equation is used. With the hydrostatic pressure coordinate, the pseudodensity is a negative constant! The continuity equation in hydrostatic pressure coordinates is relatively simple; it is linear and does not involve a time derivative:

$$\nabla_p \cdot \mathbf{v} + \frac{\partial \omega}{\partial p} = 0 \, . \tag{23.31}$$

On the other hand, the lower boundary condition is complicated in hydrostatic pressure coordinates:

$$\frac{\partial p_S}{\partial t} + \mathbf{v}_S \cdot \nabla p_S - \omega_S = 0 \, . \tag{23.32}$$

Recall that p_S can be predicted using the surface pressure-tendency equation, (23.93). Substitution from (23.93) into (23.32) gives

$$\omega_S = \frac{\partial p_T}{\partial t} - \nabla \cdot \left(\int_{p_T}^{p_S} \mathbf{v}\, dp \right) + \mathbf{v}_S \cdot \nabla p_S \,, \tag{23.33}$$

which can be used to diagnose ω_S. The fact that pressure surfaces intersect the ground at locations that change with time (unlike height surfaces), means that models that use pressure coordinates are complicated. Largely for this reason, pressure coordinates are hardly ever used in numerical models. One of the few exceptions was the early and short-lived general circulation model developed by Cecil "Chuck" Leith (1965b,a) at the Lawrence Radiation Laboratory (now the Lawrence Livermore National Laboratory).

With the pressure coordinate, we can write

$$\left[\frac{\partial}{\partial t} \left(\frac{\partial \phi}{\partial p} \right) \right]_p = -\frac{R}{p} \left(\frac{\partial T}{\partial t} \right)_p \,. \tag{23.34}$$

This allows us to eliminate the temperature in favor of the geopotential, which is often done in theoretical studies.

Let T_0 be a *constant* reference temperature, and $H \equiv RT_0/g$ the corresponding (constant) scale height. *Define* the "log-pressure coordinate," denoted by z^*, in terms of the differential relationship

$$dz^* \equiv -H \frac{dp}{p} \,. \tag{23.35}$$

The minus sign is used to make z^* increase upwards. Note that z^* has the units of length, so in that way it is "like" height. It is easy to show that

$$dz^* = \frac{T_0}{T} dz \,, \tag{23.36}$$

so that

$$dz^* = dz \text{ when } T(p) = T_0 \,. \tag{23.37}$$

Integration of (23.35) gives

$$z^* = -H \ln(p/p_0) \,. \tag{23.38}$$

where p_0 is a reference pressure, at which $z^* = 0$.

Obviously a surface of constant p is also a surface of constant z^*. Nevertheless, the equations take different forms in the p and z^* coordinate systems. From (23.35), we see that

$$\frac{\partial \phi^*}{\partial p} = -\frac{RT_0}{p} \,, \tag{23.39}$$

where

$$\phi^* \equiv g z^* \,. \tag{23.40}$$

This looks like the hydrostatic equation, but it is really just a form of (23.38), which is the definitio of z^*. Since z^* is a constant along a pressure surface, ϕ^* is also constant. We do of course have the true hydrostatic equation, which can be written as

$$\frac{\partial \phi}{\partial p} = -\frac{RT}{p} \,. \tag{23.41}$$

Here the true (non-constant) temperature appears. Subtracting (23.39) from (23.41), we obtain a useful form of the hydrostatic equation:

$$\frac{\partial (\phi - \phi^*)}{\partial p} = -\frac{R(T - T_0)}{p} \,. \tag{23.42}$$

Since ϕ^*, T_0, and p are all independent of time on z^* surfaces, we see that

$$\left[\frac{\partial}{\partial t}\left(\frac{\partial \phi}{\partial p}\right)\right]_{z^*} = -\frac{R}{p}\left(\frac{\partial T}{\partial t}\right)_{z^*} \,. \tag{23.43}$$

This is a form of the thermodynamic equation that is used in quasi-geostrophic models.

The pseudo-density in log-pressure coordinates is given by

$$\rho_{z^*} = \frac{p}{RT_0} \,. \tag{23.44}$$

The vertical velocity in log-pressure coordinates is

$$\begin{aligned} w_{z^*} &\equiv \frac{Dz^*}{Dt} \\ &= -\frac{H}{p}\frac{Dp}{Dt} \\ &= -\frac{H}{p}\omega \,. \end{aligned} \tag{23.45}$$

Finally, the continuity equation in log-pressure coordinates is given by

$$\left(\frac{\partial \rho_{z^*}}{\partial t}\right)_{z^*} + \nabla_{z^*} \cdot (\rho_{z^*}\mathbf{v}) + \frac{\partial}{\partial z^*}(\rho_{z^*} w_{z^*}) = 0 \,. \tag{23.46}$$

23.3 The sigma coordinate

23.3.1 Definition

The σ-coordinate of Norman Phillips (1957) is defined in such a way that the Earth's surface and the model top are both surfaces of constant σ:

$$\sigma \equiv \frac{p - p_T}{\pi} \,, \tag{23.47}$$

where

$$\pi \equiv p_S - p_T \,, \tag{23.48}$$

which is independent of height. From (23.47) and (23.48), it is clear that

$$\sigma_S = 1 \text{ and } \sigma_T = 0 \, . \tag{23.49}$$

This is by design, of course. Notice that if $p_T =$ constant, which is always assumed, than the top of the model is an isobaric surface. Phillips (1957) discussed the case $p_T = 0$.

Rearranging (23.47), we can write

$$p = p_T + \sigma\pi \, . \tag{23.50}$$

For a fixed value of σ, i.e., along a surface of constant σ, this implies that

$$dp = \sigma d\pi \, , \tag{23.51}$$

where the differential can represent a fluctuation in either time or horizontal position. Also,

$$\frac{\partial}{\partial p} = \frac{1}{\pi}\frac{\partial}{\partial \sigma} \, . \tag{23.52}$$

Here the partial derivatives are evaluated at fixed horizontal position and time. In view of (23.52), the hydrostatic equation in σ-coordinates can immediately be written down as

$$\frac{1}{\pi}\frac{\partial \phi}{\partial \sigma} = -\alpha \, , \tag{23.53}$$

which is closely related to the hydrostatic equation in pressure coordinates.

23.3.2 The continuity equation in sigma coordinates

The pseudodensity in σ-coordinates is

$$\rho_\sigma = -\pi / g \, , \tag{23.54}$$

which is independent of height. The continuity equation in σ-coordinates can therefore be written as

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{v}) + \frac{\partial (\pi \dot{\sigma})}{\partial \sigma} = 0 \,. \tag{23.55}$$

Although this equation does contain a time derivative, the differentiated quantity, π, is independent of height, which makes (23.55) considerably simpler than than the continuity equation in height coordinates. The lower boundary condition in σ-coordinates is very simple:

$$\dot{\sigma} = 0 \text{ at } \sigma = 1 \,. \tag{23.56}$$

This simplicity was in fact Phillips' motivation for the invention of σ-coordinates. The upper boundary condition is similar:

$$\dot{\sigma} = 0 \text{ at } \sigma = 0. \,. \tag{23.57}$$

The continuity equation in σ-coordinates plays a dual role. First, it is used to predict π. This is done by integrating (23.55) through the depth of the vertical column and using the boundary conditions (23.56) and (23.57), to obtain the surface pressure-tendency equation in the form

$$\frac{\partial \pi}{\partial t} = -\nabla \cdot \left(\int_0^1 \pi \mathbf{v} d\sigma \right) . \tag{23.58}$$

The continuity equation is also used to determine $\pi\dot{\sigma}$. Once $\partial\pi/\partial t$ has been evaluated using (23.58), which does not involve $\pi\dot{\sigma}$, we can substitute back into (23.55) to obtain

$$\frac{\partial}{\partial \sigma} (\pi \dot{\sigma}) = \nabla \cdot \left(\int_0^1 \pi \mathbf{v} d\sigma \right) - \nabla_\sigma \cdot (\pi \mathbf{v}) \,. \tag{23.59}$$

This can be integrated vertically to obtain $\pi\dot{\sigma}$ as a function of σ, starting from either the Earth's surface or the top of the atmosphere, and using the appropriate boundary condition at the bottom or top. The same result is obtained regardless of the direction of integration.

23.3.3 The horizontal pressure-gradient force

Starting from (22.17), we find that in σ-coordinates, the horizontal pressure-gradient force takes the relatively complicated form:

$$\begin{aligned} -\frac{1}{\rho}\nabla_z p &= -\frac{1}{\rho}\nabla_\sigma p + \frac{1}{\rho}\frac{\partial p}{\partial z}\nabla_\sigma z \\ &= -\sigma\alpha\nabla\pi - \nabla_\sigma\phi \ . \end{aligned} \tag{23.60}$$

To obtain the second line of (23.60), we have used the hydrostatic equation. Consider the behavior of the two terms on the right-hand side of (23.60) near a steep mountain, as illustrated in Fig. 23.1. In such a situation, the spatial variations of p_S and the near-surface value of ϕ along a σ-surface are strong and of opposite sign. For example, when moving uphill p_S decreases while ϕ_S increases. As a result, the two terms on the right-hand side of (23.60) are individually large and opposing, and the horizontal pressure-gradient force is the relatively small difference between them – a dangerous situation. Near steep mountains the relatively small discretization errors in the individual terms of the right-hand side of (23.60) can be as large as the horizontal pressure-gradient force.

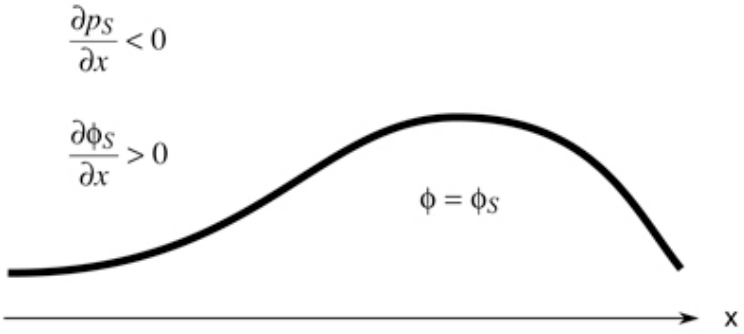


Figure 23.1: Sketch illustrating the opposing terms of the horizontal pressure gradient force as measured in σ-coordinates.

Using the hydrostatic equation, (23.53), we can rewrite (23.60) as

$$-\frac{\pi}{\rho}\nabla_z p = \sigma\left(\frac{\partial\phi}{\partial\sigma}\right)\nabla\pi - \pi\nabla_\sigma\phi \ . \tag{23.61}$$

The problem with the horizontal pressure-gradient force in σ-coordinates may appear to be an issue mainly with horizontal differencing, because Eq. (23.60) involves only horizontal derivatives, but Eq. (23.61) shows that vertical differencing also comes in. To see how,

consider Fig. 23.2. At the point O, the $\sigma = \sigma^*$ and $p = p^*$ surfaces intersect. As we move away from point O, the two surfaces separate. Eq. (23.61) shows that the horizontal pressure-gradient force depends on both the horizontal change in ϕ along a σ-surface, say between two neighboring horizontal grid points (mass points), and the vertical change in ϕ between neighboring model layers, which depends, hydrostatically, on the temperature. Compare with (23.53). If the σ-surfaces are very steeply tilted relative to constant height surfaces, which is expected near steep mountains, the thickness needed on the right-hand side of (23.61) will depend on the temperatures of two or more σ-layers, rather than a single layer. If the temperature is changing rapidly with height, this can lead to large errors. It can be shown that the problem is minimized if the model has sufficiently high horizontal resolution relative to its vertical resolution (Janjic, 1977; Mesinger, 1982; Mellor et al., 1994), i.e., it is good to have

$$\frac{\delta\sigma}{\delta x} \geq \frac{\left|\left(\frac{\delta\phi}{\delta x}\right)_\sigma\right|}{\left|\left(\frac{\delta\phi}{\delta\sigma}\right)_x\right|} \,. \tag{23.62}$$

This is a condition on the aspect ratio of the grid cells. It means that the vertical grid spacing must be large enough for a given horizontal grid spacing, or that the horizontal grid spacing must be fine enough for a given vertical grid spacing. This implies that an increase in the vertical resolution without a corresponding increase in the horizontal resolution can cause problems, and it suggests that changes in the vertical grid spacing must be accompanied by changes in the horizontal grid spacing, and vice versa. The numerator of the right-hand side of (23.62) increases when the terrain is steep, especially in the lower troposphere. The denominator increases when T is warm, i.e., near the surface, which means that it is easier to satisfy (23.62) near the surface.

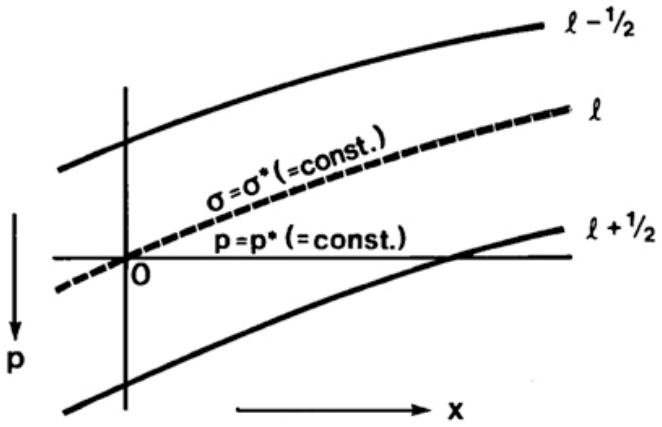


Figure 23.2: Sketch illustrating the pressure-gradient force as seen in σ-coordinates and pressure coordinates.

23.3.4 The vertically integrated horizontal pressure-gradient force

Eq. (23.61) can be rewritten as

$$\begin{aligned} -\frac{\pi}{\rho}\nabla_z p &= \left[\frac{\partial}{\partial\sigma}(\sigma\phi) - \phi\right]\nabla\pi - \pi\nabla_\sigma\phi \\ &= \frac{\partial}{\partial\sigma}(\sigma\phi)\,\nabla\pi - \nabla_\sigma(\pi\phi)\ . \end{aligned} \tag{23.63}$$

Vertically integrating (23.63) through the entire vertical column, and using (23.49), we find that

$$\int_0^1 \left(-\frac{\pi}{\rho}\nabla_z p\right)\,d\sigma = \phi_S\nabla\pi - \nabla\left(\int_0^1 \pi\phi\right)\,d\sigma\ . \tag{23.64}$$

When integrated around any closed horizontal path, the first term gives the mountain torque, and the second term vanishes.

23.3.5 The pressure vertical velocity

Finally, the Lagrangian time derivative of pressure can be expressed in σ-coordinates as

$$\begin{aligned} \omega \equiv \frac{Dp}{Dt} &= \left(\frac{\partial p}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma p + \dot\sigma\frac{\partial p}{\partial\sigma} \\ &= \sigma\left(\frac{\partial\pi}{\partial t} + \mathbf{v}\cdot\nabla\pi\right) + \pi\dot\sigma\ . \end{aligned} \tag{23.65}$$

23.4 Hybrid sigma-pressure coordinates

The advantage of the σ-coordinate is realized in the lower boundary condition. The disadvantage is the complicated and poorly behaved pressure-gradient force, which is realized at all levels. This has motivated the use of hybrid coordinates that reduce to σ at the lower boundary, and become pure pressure-coordinates at higher levels. In principle there are many ways of doing this. The most widely cited paper on this topic is by Simmons and Burridge (1981). They recommended the coordinate

$$\sigma_p(p,\,p_S) \equiv \left(\frac{p-p_T}{p_0-p_T}\right)\left(\frac{p_S-p}{p_S-p_T}\right) + \left(\frac{p-p_T}{p_S-p_T}\right)^2\ , \tag{23.66}$$

where p_0 is a positive constant. It can be shown that σ_p is monotonic with pressure provided that $p_0 > (p_S + p_T)/2$. Inspection of (23.66) shows that

$$(\sigma_p)_T = 0 \quad \text{and} \quad (\sigma_p)_S = 1 \, . \tag{23.67}$$

The boundary conditions on the vertical mass flux are just like with σ coordinates:

$$(\dot{\sigma}_p)_T = 0 \quad \text{and} \quad (\dot{\sigma}_p)_S = 0 \, . \tag{23.68}$$

It can be demonstrated that σ_p-surfaces are nearly parallel to isobaric surfaces in the upper troposphere and stratosphere, despite possible variations of the surface pressure in the range $\sim$1000 hPa to $\sim$500 hPa. When we evaluate the horizontal pressure-gradient force with the σ_p-coordinate, there are still two terms, as with the σ-coordinate, but above the lower troposphere one of the terms is strongly dominant, regardless of the topography.

23.5 Terrain-following vertical coordinates based on height

A σ-like vertical coordinate based on height (instead of pressure) is used in NICAM (Satoh et al., 2008). Klemp (2011) invented a different approach for use in the MPAS atmosphere dynamical core (Skamarock et al., 2012). Let z be the height of a coordinate surface, let $h(\lambda, \phi)$ be the terrain height, and let $h_s(\lambda, \phi, \zeta)$ be a smoothed version of the terrain height, such that $h_s(\lambda, \phi, 0) = h(\lambda, \phi)$. MPAS uses

$$z \equiv \zeta + A(\zeta) h_s(\lambda, \phi, \zeta) \, . \tag{23.69}$$

This is the definition of what I call the "Klemp coordinate." ζ. Note that $\zeta = z$ where $h_s = 0$.

There are two smoothers at work in (23.69):

- The smoothed terrain height $h_s(\lambda, \phi, \zeta)$. The values of h_s are smoothed for $\zeta > 0$ using an iterative diffusion operator, *before the model is started up*. The number of iterations is chosen to suit the planned simulation. The smoothing is done in such a way that the fractional vertical grid spacing does not fall below a specified minimum value.

- The "hybrid attenuation profile," $A(\zeta)$, controls the rate at which the coordinate transitions from terrain-following at the surface to constant height surfaces aloft. For $A = 0$ we get $z = \zeta$. MPAS uses $A(0) = 1$, which ensures that $\zeta = 0$ at the surface. It also uses $A(\zeta) = 0$ for $\zeta \geq z_H$, where z_H is specified, which ensures that the top of the model is a surface of constant $\zeta = z = z_T$. Therefore $\zeta =$ constant at both the surface and the model top. See Klemp (2011) for further explanation.

Some issues with making the top of the model a surface of constant z are discussed by Klemp and Skamarock (2022).

23.6 The eta-coordinate

As a solution to the problem with the horizontal pressure-gradient force in σ-coordinates, Mesinger (1984) and Mesinger and Janjic (1985) proposed the η-coordinate, which was used operationally at NCEP (the U.S. National Centers for Environmental Prediction) from 1993 to 2006. The coordinate is defined by

$$\eta \equiv \sigma \eta_S \,, \tag{23.70}$$

where

$$\eta_S = \frac{p_{\text{ref}}(z_S) - p_T}{p_{\text{ref}}(0) - p_T} \,. \tag{23.71}$$

Whereas $\sigma = 1$ at the Earth's surface, Eq. (23.70) shows that $\eta = \eta_S$ at the Earth's surface. According to (23.71), $\eta_S = 1$ (just as $\sigma_S = 1$) if $z_S = 0$. Here z_S is set to zero at or near "sea level." Only a finite number of discrete values of z_S are permitted, which means that *mountains must come in a few discrete sizes*, like off-the-rack clothing. The number of mountain sizes increases as the vertical resolution of the model increases. This is sometimes called the "step-mountain" approach.

The function $p_{\text{ref}}(z_S)$ is pre-specified so that it gives typical surface pressures for each value of z_S. Because z_S depends on the horizontal coordinates, η_S does too. Only a finite number of discrete (and constant) values of η_S are permitted. After choosing the function $p_{\text{ref}}(z_S)$ and the map $z_S(x, y)$, it is possible to make a map of $\eta_S(x, y)$, and of course this map is independent of time.

When we build a σ-coordinate model, we must choose values of σ to serve as layer-edges and/or layer centers. These values are constant in the horizontal and time. Similarly, when we build an η-coordinate model, we must specify fixed values of η to serve as layer

edges and/or layer centers. The layer-edge values of η for the lower troposphere include all of the values of η_S. Fig. 23.3 shows how this works. Note that, unlike σ-surfaces, η-surfaces are nearly flat, in the sense that they are close to being isobaric surfaces. The circled u-points have $u = 0$, which is the appropriate boundary condition on the cliff-like sides of the step mountains.

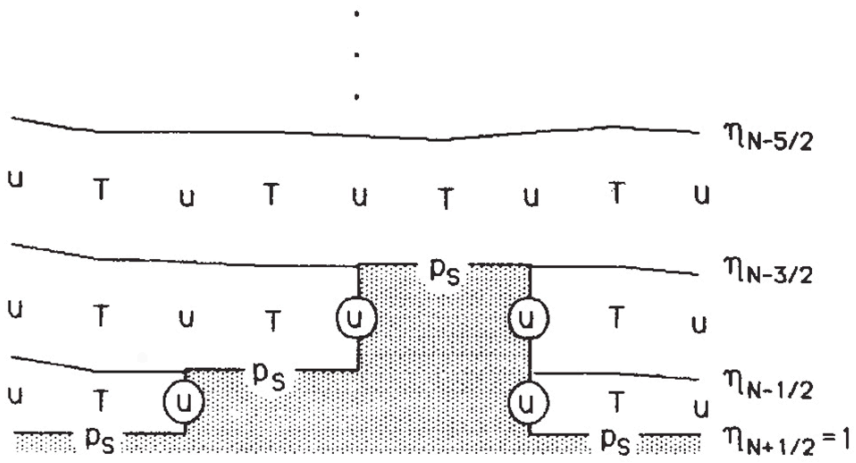


Figure 23.3: Sketch illustrating the η-coordinate.

In η-coordinates, the horizontal pressure-gradient force still consists of two terms:

$$-\nabla_p \phi = -\nabla_\eta \phi - \alpha \nabla_\eta p \, . \tag{23.72}$$

Because the η-surfaces are nearly flat, however, these two terms are each comparable in magnitude to the horizontal pressure-gradient force itself, even near mountains, so the problem of near-cancellation is greatly mitigated.

23.7 Potential temperature and entropy coordinates

23.7.1 Definition and attractions

The potential temperature is defined by

$$\theta \equiv T \left(\frac{p_0}{p} \right)^{\kappa} \, . \tag{23.73}$$

The potential temperature increases upwards in a statically stable atmosphere, so that there is a monotonic relationship between θ and z. Note, however, that potential temperature

cannot be used as a vertical coordinate when static instability occurs, and that the vertical resolution of a θ-coordinate model becomes very poor when the atmosphere is close to neutrally stable.

Potential temperature coordinates have highly useful properties that have been recognized for many years:

- In the absence of heating, θ is conserved following a particle. This means that the vertical motion in θ-coordinates is proportional to the heating rate:

$$\dot{\theta} = \frac{\theta}{c_p t} Q \,; \tag{23.74}$$

  in the absence of heating, there is "no vertical motion," from the point of view of θ-coordinates. It is also true that, in the absence of heating, a particle that is on a given θ-surface stays on that surface. This minimizes errors associated with vertical advection. Any quasi-Lagrangian system has this property. Eq. (23.74) is equivalent to (23.96), and is an expression of the thermodynamic energy equation in θ-coordinates.

- The pressure-gradient force is a gradient. This minimizes pressure-gradient errors near topography, and spurious generation of vorticity.
- The Ertel potential vorticity is easily computed from the wind vector and the isentropic pseudo-density.
- Wave momentum transport occurs via isentropic form drag.
- It is easy to implement diffusion along θ-surfaces.
- The available potential energy can be calculated exactly.
- Both energy and entropy can be conserved.
- The sharp inversion at the top of the boundary layer (or at a front) can be represented by layers that have small pseudo-densities.

In fact, θ-coordinates provide an especially simple pathway for the derivation of many important results, including the conservation equation for the Ertel potential vorticity. In addition, θ-coordinates have some important advantages for both observational and numerical studies. Their utility has been endorsed by many scientists over the past century (e.g., Rossby and Collaborators, 1937; Namias and Stone, 1940; Starr, 1945; Lorenz, 1955; Danielsen, 1961; Bleck, 1973; Uccellini and Johnson, 1979; Johnson and Uccellini, 1983; Hoskins et al., 1985; Hsu and Arakawa, 1990; Benjamin et al., 2004).

The continuity equation in θ-coordinates is

$$\left(\frac{\partial \rho_\theta}{\partial t}\right)_\theta + \nabla_\theta \cdot (\rho_\theta \mathbf{v}) + \frac{\partial}{\partial \theta}\left(\rho_\theta \dot{\theta}\right) = 0 \,. \tag{23.75}$$

For $\dot{\theta} = 0$, (23.75) reduces to

$$\left(\frac{\partial \rho_\theta}{\partial t}\right)_\theta + \nabla_\theta \cdot (\rho_\theta \mathbf{v}) = 0 \,, \tag{23.76}$$

which is closely analogous to the continuity equation of a shallow-water model. In the absence of heating, a model that uses θ-coordinates behaves like "a stack of shallow-water models."

23.7.2 Massless layers

Although θ-surfaces can intersect the lower boundary, we can consider, following Lorenz (1955), that they actually follow along the boundary, like coats of paint. This leads to the concept of "massless layers," as shown in the middle panel of Fig. 23.4. The massless layers exist in between coats of paint. Obviously, a model that follows the massless-layer approach is susceptible to producing negative mass. This can be avoided, for example, through the use of flux-corrected transport. Nevertheless, this practical difficulty has led most modelers to avoid θ-coordinates.

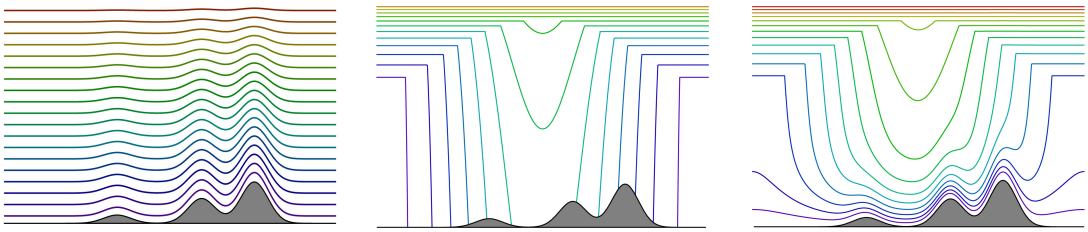


Figure 23.4: Coordinate surfaces with topography: Left, the σ-coordinate. Center, the θ-coordinate. Right, a hybrid σ-θ coordinate.

The massless layer approach allows us to use values of θ that are colder than any actually present in an atmospheric column, particularly in the tropics of a global model. The coldest possible value of θ is zero Kelvin. Consider the lower boundary condition on the hydrostatic equation, (23.85). We can write

$$s(\theta) - s(0) = \int_0^\theta \Pi\left(\theta'\right)\, d\theta' \,, \tag{23.77}$$

where θ' is a dummy variable of integration. From the definition of s, we have $s(0) = \phi_S$. For "massless" portion of the integral, the integrand, $\Pi(\theta')$, is just a constant, namely Π_S, i.e., the surface value of Π. We can therefore write

$$\begin{aligned} s(\theta) &= \phi_S + \int_0^{\theta_S} \Pi\left(\theta'\right) d\theta' + \int_{\theta_S}^{\theta} \Pi\left(\theta'\right) d\theta' \\ &= \phi_S + \Pi_S \theta_S + \int_{\theta_S}^{\theta} \Pi\left(\theta'\right) d\theta' \\ &= \phi_S + c_p T_S + \int_{\theta_S}^{\theta} \Pi\left(\theta'\right) d\theta' \,. \end{aligned} \tag{23.78}$$

It follows that

$$s(\theta) = s_S + \int_{\theta_S}^{\theta} \Pi\left(\theta'\right) d\theta' \,, \tag{23.79}$$

as expected.

We can take the lower boundary to be located at $\theta = \theta_{\min} =$ constant, where $\theta_{\min}$ is smaller than any value of theta that we expect to encounter in our simulation. For example, could choose $\theta_{\min} = 0$ K. Note that $(\rho_\theta)_{\theta_{\min}} = 0$. It follows that

$$\left(\rho_\theta \dot{\theta}\right)_{\theta_{\min}} = 0 \,, \tag{23.80}$$

This means that no mass crosses the $\theta = \theta_{\min} =$ surface.

23.7.3 The hydrostatic equation

Recall that, with a general vertical coordinate, $\hat{z}$, the hydrostatic equation can be expressed as

$$\frac{\partial p}{\partial \hat{z}} = -\rho_{\hat{z}} g \tag{23.81}$$

For the case of θ-coordinates, the hydrostatic equation, (23.81), reduces to

$$\frac{\partial \phi}{\partial \theta} = \alpha \frac{\partial p}{\partial \theta} \ . \tag{23.82}$$

Logarithmic differentiation of (23.71) gives

$$\frac{d\theta}{\theta} = \frac{dT}{T} - \kappa \frac{dp}{p} \ . \tag{23.83}$$

It follows that

$$\alpha \frac{\partial p}{\partial \theta} = c_p \frac{\partial T}{\partial \theta} - c_p \frac{T}{\theta} \ . \tag{23.84}$$

Substitution of (23.84) into (23.82) gives

$$\frac{\partial s}{\partial \theta} = \Pi, \tag{23.85}$$

where s is the dry static energy.

With the quasi-static approximation, the horizontal pressure-gradient force in θ-coordinates can be written as

$$-\alpha \nabla_z p = -\alpha \nabla_\theta p - \nabla_\theta \phi \ . \tag{23.86}$$

From (23.83) it follows that

$$\nabla_\theta p = c_p \left(\frac{p}{RT} \right) \nabla_\theta T. \tag{23.87}$$

Substitution of (23.87) into (23.86) gives

$$-\alpha\nabla_z p = -\nabla_\theta s \ . \tag{23.88}$$

23.7.4 The isentropic potential vorticity

The dynamically important isentropic potential vorticity, q, is easily constructed in θ-coordinates, since it involves the curl of $\mathbf{v}$ on a θ-surface:

$$q \equiv (\mathbf{k}\cdot\nabla_\theta \times \mathbf{v} + f)\,\frac{\partial\theta}{\partial p} \ . \tag{23.89}$$

The available potential energy is also easily obtained, since it involves the distribution of pressure on θ-surfaces.

The entropy coordinate is very similar to the θ-coordinate. We define the entropy by

$$\varepsilon \equiv c_p \ln(\theta/\theta_0) \ , \tag{23.90}$$

where θ_0 is a constant reference value of θ. It follows that

$$d\varepsilon = c_p d\theta/\theta \ . \tag{23.91}$$

The hydrostatic equation can then be written as

$$\frac{\partial s}{\partial \varepsilon} = T \ . \tag{23.92}$$

This is a particularly attractive form because the "thickness" (in terms of s) between two entropy surfaces is simply the temperature.

23.8 Vertical mass flux for a family of vertical coordinates with the quasi-static approximation

23.8.1 Preliminaries

In view of (23.81), Eq. (22.29) is equivalent to

$$\frac{\partial p_S}{\partial t} = -\nabla \cdot \left(\int_{\hat{z}_t}^{\hat{z}_S} \rho_{\hat{z}} \mathbf{v} \, d\hat{z} \right) + \frac{\partial p_T}{\partial t} \, , \tag{23.93}$$

which is the surface pressure tendency equation. Depending on the definitions of $\hat{z}$ and $\hat{z}_T$, it may or may not be appropriate to set $p_T =$ constant as an upper boundary condition, which would cause the $\partial p_T / \partial t$ term of (23.93) to drop out. This was discussed already in Chapter 22. Corresponding to (23.93), we can show that the pressure tendency on an arbitrary $\hat{z}$-surface satisfies

$$\left(\frac{\partial p}{\partial t} \right)_{\hat{z}} = -\nabla \cdot \left(\int_{\hat{z}_t}^{\hat{z}} \rho_{\hat{z}} \mathbf{v} d \, \hat{z} \right) + \frac{\partial p_T}{\partial t} + \rho_{\hat{z}} \hat{w} \, (\hat{z}) \, . \tag{23.94}$$

The thermodynamic equation can be written as

$$c_p \left[\left(\frac{\partial T}{\partial t} \right)_{\hat{z}} + (\mathbf{v} \cdot \nabla_{\hat{z}}) \, T + \hat{w} \frac{\partial T}{\partial \hat{z}} \right] = \omega \alpha + Q \, . \tag{23.95}$$

An alternative form is

$$\left(\frac{\partial \theta}{\partial t} \right)_{\hat{z}} + (\mathbf{v} \cdot \nabla_{\hat{z}}) \, \theta + \hat{w} \frac{\partial \theta}{\partial \hat{z}} = \frac{Q}{\Pi} \, . \tag{23.96}$$

23.8.2 A family of vertical coordinates

Konor and Arakawa (1997) derived a diagnostic equation that can be used to compute the vertical velocity, $\hat{w}$, for a large family of vertical coordinates that can be expressed as functions of the potential temperature, the pressure, and the surface pressure, i.e.,

$$\hat{z} \equiv F \, (\theta, \, p, \, p_S) \, . \tag{23.97}$$

We are free to choose the form of $F(\theta,\ p,\ p_S)$, subject to the condition that $\hat{z}$ is monotonic with height. While not completely general, Eq. (23.97) does include a variety of interesting cases, namely:

- Pressure coordinates
- Sigma coordinates
- The hybrid sigma-pressure coordinate of Simmons and Burridge (1981)
- Theta coordinates
- The hybrid sigma-theta coordinate of Konor and Arakawa (1997).

The height coordinate and the normalized height coordinate are *not* included in (23.97).

23.8.3 The vertical velocity

By taking the partial derivative (23.97) with respect to time, on a surface of constant $\hat{z}$, we have

$$0 = \left[\frac{\partial}{\partial t} F(\theta,\ p,\ p_S)\right]_{\hat{z}} . \tag{23.98}$$

The chain rule tells us that this is equivalent to

$$\frac{\partial F}{\partial \theta}\left(\frac{\partial \theta}{\partial t}\right)_{\hat{z}} + \frac{\partial F}{\partial p}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} + \frac{\partial F}{\partial p_S}\left(\frac{\partial p_S}{\partial t}\right) = 0 . \tag{23.99}$$

Substituting from (23.96), (23.94), and (23.93), we obtain

$$\begin{aligned} &\frac{\partial F}{\partial \theta}\left[-\left((\mathbf{v}\cdot\nabla_{\hat{z}})\,\theta + \hat{w}\frac{\partial \theta}{\partial \hat{z}}\right) + \frac{Q}{\Pi}\right] \\ &+\frac{\partial F}{\partial p}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_t}^{\hat{z}} \rho_{\hat{z}}\mathbf{v}\,d\hat{z}\right) + (\rho_{\hat{z}}\hat{w})_{\hat{z}}\right] \\ &+\frac{\partial F}{\partial p_S}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_t}^{\hat{z}_S} \rho_{\hat{z}}\mathbf{v}\,d\hat{z}\right)\right] = 0 . \end{aligned} \tag{23.100}$$

Eq. (23.100) can be solved for the vertical velocity, $\hat{w}$:

$$\boxed{\hat{w} = \frac{\frac{\partial F}{\partial \theta}\left(-\left(\mathbf{v}\cdot\nabla_{\hat{z}}\right)\theta + \frac{Q}{\Pi}\right) + \frac{\partial F}{\partial p}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}} \rho_{\hat{z}}\mathbf{v}d\hat{z}\right)\right] + \frac{\partial F}{\partial p_S}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}_S} \rho_{\hat{z}}\mathbf{v}\, d\hat{z}\right)\right]}{\frac{\partial \theta}{\partial \hat{z}}\frac{\partial F}{\partial \theta} - \rho_{\hat{z}}\frac{\partial F}{\partial p}}} \,. \tag{23.101}$$

As a check of (23.101), consider the special case $F \equiv p$, so that $\rho_{\hat{z}} = -1/g$, and assume that $\partial p_T/\partial t = 0$, as would be natural for the case of pressure coordinates. Then (23.101) reduces to

$$\omega = -\nabla\cdot\left(\int_{p_T}^{p} \mathbf{v}\, dp\right) \,. \tag{23.102}$$

As a second special case, suppose that $F \equiv \theta$. Then (23.101) becomes

$$\dot{\theta} = Q/\Pi \,. \tag{23.103}$$

Both of these results are as expected.

23.8.4 The upper boundary

We assume that the model top is a surface of constant $\hat{z}$, i.e., $\hat{z}_T = \text{constant}$. Then (23.99) must apply at the model top, so that we can write

$$\left(\frac{\partial F}{\partial \theta}\right)_{\theta_T,\, p_T} \frac{\partial \theta_T}{\partial t} + \left(\frac{\partial F}{\partial p}\right)_{\theta_T,\, p_T} \frac{\partial p_T}{\partial t} + \left(\frac{\partial F}{\partial p_S}\right)_{\theta_T,\, p_T} \frac{\partial p_S}{\partial t} = 0 \,. \tag{23.104}$$

Suppose that $F(\theta,\, p,\, p_S)$ is chosen in such a way that $(\partial F/\partial p_S)_{\theta_T,\, p_T} = 0$. This is a natural choice, because the model top is far away from the surface, so that p_S is expected to be irrelevant. Then Eq. (23.104) simplifies to

$$\left(\frac{\partial F}{\partial \theta}\right)_{\theta_T,\, p_T} \frac{\partial \theta_T}{\partial t} + \left(\frac{\partial F}{\partial p}\right)_{\theta_T,\, p_T} \frac{\partial p_T}{\partial t} = 0 \,. \tag{23.105}$$

Now consider two possibilities:

1. When the model top is both a surface of constant $\hat{z}$ and an isobaric surface, so that $\partial p_T/\partial t = 0$, then the second term of (23.105) goes away, and we have the following situation: By assumption, $[F(\theta,\ p,\ p_S)]_T$ is a constant (because the top of the model is also a surface of constant $\hat{z}$). Also by assumption, $[F(\theta,\ p,\ p_S)]_T$ does not depend on p_S. Finally we have assumed that the top of the model is an isobaric surface. It follows that, *when the model top is an isobaric surface, the form of* $F(\theta,\ p,\ p_S)$ *must be chosen so that* $(\partial F/\partial p_S)_{\theta_T,\ p_T} = 0$.

2. When the model top is an isentropic surface, $\partial\theta_T/\partial t = 0$, so the last term of (22.104) goes away. The form of $F(\theta,\ p,\ p_S)$ must be chosen so that $(\partial F/\partial p_S)_{\theta_T,\ p_T} = 0$. Recall, however, from Section 22.8, that to avoid unphysical momentum fluxes at the model top, we should make the model top either a surface of constant height or a surface of constant pressure. Making the model top a surface of constant potential temperature is not a good idea.

23.9 Hybrid sigma-theta coordinates

Konor and Arakawa (1997) discuss a hybrid σ - θ vertical coordinate, which we will call $\hat{z}_{\rm KA}$, that reduces to θ away from the surface, and to σ near the surface. This hybrid coordinate is a member of the family of schemes given by (23.97). It is designed to combine the strengths of θ and σ coordinates, while avoiding their weaknesses. Hybrid σ - θ coordinates have also been considered by other authors, e.g., Johnson and Uccellini (1983) and Zhu et al. (1992).

To specify the scheme, we must choose the function $F(\theta,\ p,\ p_S)$ that appears in (23.97). Following Konor and Arakawa (1997), define

$$\hat{z}_{\rm KA} = F(\theta,\ p,\ p_S) \equiv f(\sigma) + g(\sigma)\,\theta\ , \tag{23.106}$$

where $\sigma \equiv \sigma(p, p_S)$ is a modified sigma coordinate, defined so that it is (as usual) a constant at the Earth's surface, and (not as usual) increases upwards, e.g.,

$$\sigma \equiv \frac{p_S - p}{p_S}\ . \tag{23.107}$$

Note that with this definition $\sigma_S = 0$. If we specify $f(\sigma)$ and $g(\sigma)$, then the hybrid coordinate is fully determined.

We require, of course, that $\hat{z}_{\rm KA}$ itself increases upwards, so that

$$\frac{\partial \hat{z}_{\mathrm{KA}}}{\partial \sigma} > 0 \, . \tag{23.108}$$

We also require that

$$\hat{z}_{\mathrm{KA}} = \text{constant for } \sigma = 0, \tag{23.109}$$

which means that $\hat{z}_{\mathrm{KA}}$ is σ-like at the Earth's surface, and that

$$\hat{z}_{\mathrm{KA}} = \theta \text{ for } \sigma = \sigma_T \, , \tag{23.110}$$

which means that $\hat{z}_{\mathrm{KA}}$ becomes θ at the model top (or lower). These conditions imply, from (23.106), that

$$g(\sigma) \to 0 \text{ as } \sigma \to 0 \, , \tag{23.111}$$

$$f(\sigma) \to 0 \text{ and } g(\sigma) \to 1 \text{ as } \sigma \to \sigma_T \, . \tag{23.112}$$

Now substitute (23.106) into (23.108), to obtain

$$\frac{df}{d\sigma} + \frac{dg}{d\sigma}\theta + g\frac{\partial \theta}{\partial \sigma} > 0 \, . \tag{23.113}$$

This is the requirement that $\hat{z}_{\mathrm{KA}}$ increases monotonically upward. Any choices for f and g that satisfy (23.111) - (23.113) can be used to define the hybrid coordinate.

Here is a way to satisfy those requirements: First, we agree to choose $g(\sigma)$ so that it is a monotonically increasing function of σ, i.e.,

$$\frac{dg}{d\sigma} > 0 \text{ for all } \sigma \, . \tag{23.114}$$

Obviously, there are many ways to do this. Since θ also increases upward, the condition (23.114) simply ensures that $g(\sigma)$ and θ change in the same sense, and the middle term on the left-hand side of (23.113) is guaranteed to be positive. We also choose $g(\sigma)$ so that the conditions (23.111) - (23.112) are satisfied. There are many possible choices for $g(\sigma)$ that meet these requirements.

Next, define $\theta_{\min}$ and $(\partial\theta/\partial\sigma)_{\min}$ as lower bounds on θ and $\partial\theta/\partial\sigma$, respectively, so that

$$\theta > \theta_{\min} \text{ and } \frac{\partial\theta}{\partial\sigma} > \left(\frac{\partial\theta}{\partial\sigma}\right)_{\min} . \tag{23.115}$$

When we choose the value of $\theta_{\min}$, *we are saying that we have no interest in simulating situations in which* θ *is actually colder than* $\theta_{\min}$. For example, we could choose $\theta_{\min} = 10$ K. This is not necessarily an ideal choice, for reasons to be discussed below, but we can be sure that θ in our simulations will exceed 10 K everywhere at all times, unless the model is in the final throes of blowing up. *Similarly, when we choose the value of* $(\partial\theta/\partial\sigma)_{\min}$, *we are saying that we have no interest in simulating situations in which* $\partial\theta/\partial\sigma$ *is actually less stable (or more unstable) than* $(\partial\theta/\partial\sigma)_{\min}$.

Now, with reference to the *inequality* (23.113), we write the following *equation*:

$$\frac{df}{d\sigma} + \frac{dg}{d\sigma}\theta_{\min} + g\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min} = 0 . \tag{23.116}$$

Remember that we have agreed to specify $g(\sigma)$ in such a way that (23.110) is satisfied. You should be able to see that if the *equality* (23.116) is satisfied, then the *inequality* (23.113) will also be satisfied, i.e., $\hat{z}_{\mathrm{KA}}$ will increase monotonically upward. *This will be true even if the sounding is statically unstable in some regions, provided that* (23.111) *is satisfied.* Eq. (23.116) is a first-order ordinary differential equation for $f(\sigma)$, which can be solved subject to the boundary condition (23.112).

The vertical velocity is obtained using (23.101).

That's all there is to it. Amazingly, the scheme does not involve any "if-tests." It is simple and fairly flexible.

23.10 Summary

There are many possible vertical coordinate systems, some old and some new. The hybrid sigma-pressure coordinate and the hybrid sigma-theta coordinate are particularly attractive.

As mentioned in Chapter 22, the ALE method is not really a vertical coordinatem but it is very useful and may be more widely used in the future.

23.11 Problems

1. Derive the form of the horizontal pressure-gradient force in σ_p coordinates, and compare with the corresponding formula in σ coordinates.
2. Starting from $\partial p/\partial z = -\rho g$, show that $\partial \phi/\partial \Pi = -\theta$.
3. Prove that the method to determine $\pi\dot{\sigma}$ with the σ coordinate, i.e.,

$$\frac{\partial}{\partial \sigma}(\pi\dot{\sigma}) = \nabla \cdot \left(\int_0^1 \pi \mathbf{v}\, d\sigma \right) - \nabla_\sigma \cdot (\pi \mathbf{v}) \tag{23.117}$$

   is consistent with the method to determine the vertical velocity for a general family of schemes (that includes the σ coordinate), as given by (23.101).
4. For the hybrid sigma-pressure coordinate of Simmons and Burridge (1981), work out
   (a) The form of the pseudo-density, expressed as a function of the vertical coordinate.
   (b) The condition for σ_p to be a monotonic function of height.
   (c) A method to determine the vertical velocity, $\dot{\sigma}_p$.
   (d) The form of the horizontal pressure-gradient force.

   You may use the hydrostatic equation, and you may assume that the model predicts both p_S and $\theta(\sigma_p)$.

   Suggestion: Start by writing σ_p as a function of σ, rather than as a function of p.
5. Assume that the surface elevation is given by $\partial z_S/\partial x = A[1+\sin(2\pi x/L)]$. Also assume that the temperature is independent of height, and equal to 250 K, and that the horizontal temperature gradient is given by $\partial T/\partial x = 10^{-5}\,\mathrm{K\,m^{-1}}\,\sin(2\pi x/M)$ K per meter at all levels throughout the atmospheric column. Set $M = 10^6$ m. Calculate the x-component of the horizontal pressure gradient force using both the sigma coordinate and the hybrid sigma-pressure coordinate of Simmons and Burridge (1981).

Chapter 24

Vertical differencing

24.1 Vertical staggering

After the choice of vertical coordinate system, the next issue is the choice of vertical staggering. Two possibilities are discussed here, and are illustrated in Fig. 24.1. These are the "Lorenz" or "L" staggering (Lorenz, 1960), and the "Charney-Phillips" or "CP" staggering (Charney and Phillips, 1953). On both grids the mass and the horizontal wind are defined at the same levels, which are called "layer centers," and are represented by the horizontal dashed lines in the figure. The vertical velocities are located at layer edges, which are represented by the horizontal solid lines in the figure The key difference between the two grids is in the location of the potential temperature, θ. On the L grid, θ is defined at the layer centers, along with the mass and horizontal wind. On the CP grid, θ is defined at the layer edges. Note that the upper and lower boundaries are both located at layer edges.

Suppose that both grids have N wind-levels. The L-grid also has N θ-levels, while the CP-grid has $N+1$ θ-levels. On both grids, ϕ is hydrostatically determined on the wind-levels, and

$$\phi_l - \phi_{l+1} \sim \theta_{l+\frac{1}{2}} \ . \tag{24.1}$$

On the CP-grid, θ is located between ϕ-levels, so (24.1) is convenient. With the L-grid, $\theta_{l+\frac{1}{2}}$ must be determined by interpolation. For example, we might choose

$$\theta_{l+\frac{1}{2}} = \frac{1}{2}\left(\theta_l + \theta_{l+1}\right) \ . \tag{24.2}$$

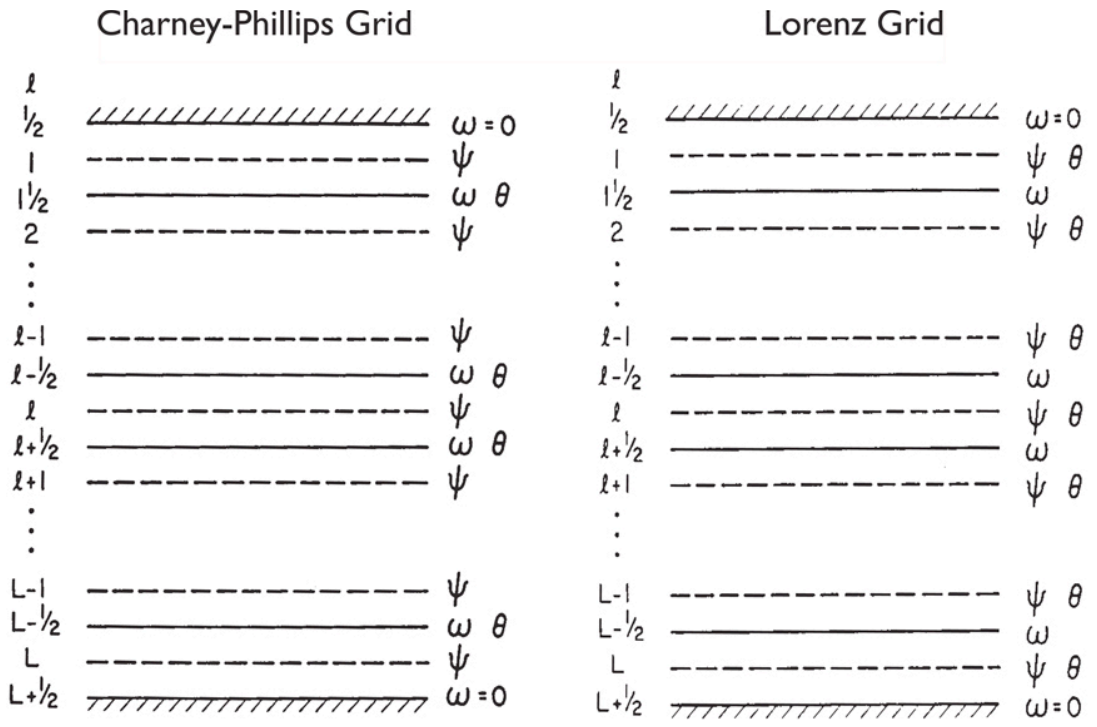


Figure 24.1: A comparison of the Lorenz and Charney-Phillips staggering methods.

Because (24.2) involves averaging, a vertical oscillation in θ is not "seen" by ϕ, and so has no effect on the winds; it is dynamically inert. This is a problem because the "invisible" oscillation can't be removed by the model's dynamics. No such problem occurs with the CP-grid.

There is a second, less obvious problem with the L-grid. The vertically discrete potential vorticity corresponding to (23.89) is

$$q_l \equiv (\mathbf{k} \cdot \nabla_\theta \times \mathbf{v}_l + f) \left(\frac{\partial \theta}{\partial p} \right)_l . \tag{24.3}$$

Inspection shows that (24.3) "wants" the potential temperature to be defined at levels in between the wind levels, as they are on the CP-grid. Suppose that we have N wind levels. Then with the CP-grid we will have $N+1$ potential temperature levels and N potential vorticities. This is nice. With the L-grid, on the other hand, it can be shown that we effectively have $N+1$ potential vorticities. The "extra" degree of freedom in the potential vorticity is spurious, and allows a spurious "computational baroclinic instability" (Arakawa and Moorthi, 1988). This is a further drawback of the L-grid.

With the Charney-Phillips staggering, we need continuity equations at the layer edges that are consistent with and actually *implied by* the continuity equations at layer centers. It

is possible to construct a set of layer-edge continuity equations that are *implied by* the layer-center continuity equations. In other words, given that we time-step the layer-center continuity equations, the layer-edge continuity equations are satisfied "automatically." There is no need to time-step them separately. Further discussion is given by Arakawa and Konor (1996).

As Lorenz (1960) pointed out, however, the L-grid is convenient for maintaining total energy conservation, because the kinetic and thermodynamic energies are defined at the same levels. Today, most models use the L-grid. Exceptions are the UK's Unified Model (Davies et al., 2005) and the Canadian Environmental Multiscale model (Girard et al., 2014), both of which use the CP-grid.

24.2 Conservation of total energy with continuous sigma coordinates

In Chapter 22, we discussed conservation of total energy with the basic equations using height coordinates and general coordintes. We now present the corresponding derivation using σ coordinates, with the quasi-static approximation. The starting equations are

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{v}) + \frac{\partial (\pi \dot{\sigma})}{\partial \sigma} = 0 \,, \tag{24.4}$$

$$\begin{aligned} \omega &\equiv \frac{Dp}{Dt} \\ &= \left(\frac{\partial p}{\partial t} \right)_\sigma + \mathbf{v} \cdot \nabla_\sigma p + \dot{\sigma} \frac{\partial p}{\partial \sigma} \\ &= \sigma \left(\frac{\partial \pi}{\partial t} + \mathbf{v} \cdot \nabla \pi \right) + \pi \dot{\sigma} \,, \end{aligned} \tag{24.5}$$

$$\left(\frac{\partial \mathbf{v}}{\partial t} \right)_\sigma + [f + \mathbf{k} \cdot (\nabla_\sigma \times \mathbf{v})] \, \mathbf{k} \times \mathbf{v} + \dot{\sigma} \frac{\partial \mathbf{v}}{\partial \sigma} + \nabla_\sigma K = -\sigma \alpha \nabla \pi - \nabla_\sigma \phi \,, \tag{24.6}$$

$$\left[\frac{\partial}{\partial t} (\pi \theta) \right]_\sigma + \nabla_\sigma \cdot (\pi \mathbf{v} \theta) + \frac{\partial}{\partial \sigma} (\pi \dot{\sigma} \theta) = 0 \,, \tag{24.7}$$

$$\frac{\partial \phi}{\partial \sigma} = -\pi\alpha \; . \tag{24.8}$$

Using continuity, (24.7) can be expressed in advective form:

$$\left(\frac{\partial \theta}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma\theta + \dot{\sigma}\frac{\partial \theta}{\partial \sigma} = 0 \; . \tag{24.9}$$

We can write (24.9) in terms of temperature, as follows:

$$\begin{aligned} c_p\left[\left(\frac{\partial T}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma T + \dot{\sigma}\frac{\partial T}{\partial \sigma}\right] &= \frac{c_p T}{\Pi}\left[\left(\frac{\partial \pi}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma \Pi + \dot{\sigma}\frac{\partial \pi}{\partial \sigma}\right] \\ &= \frac{c_p T\kappa}{p}\left[\left(\frac{\partial p}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma p + \dot{\sigma}\frac{\partial p}{\partial \sigma}\right] \\ &= \sigma\alpha\left(\frac{\partial \pi}{\partial t} + \mathbf{v}\cdot\nabla\pi + \pi\dot{\sigma}\right) \\ &= \omega\alpha \; . \end{aligned} \tag{24.10}$$

Continuity allows us to rewrite (24.10) in flux form:

$$\boxed{\left[\frac{\partial}{\partial t}(\pi c_p T)\right]_\sigma + \nabla_\sigma\cdot(\pi\mathbf{v}c_p T) + \frac{\partial}{\partial \sigma}(\pi\dot{\sigma}c_p T) = \sigma\pi\alpha\left(\frac{\partial \pi}{\partial t} + \mathbf{v}\cdot\nabla\pi + \pi\dot{\sigma}\right) \; .} \tag{24.11}$$

Here we have used

$$\pi\omega\alpha = \sigma\pi\alpha\left(\frac{\partial \pi}{\partial t} + \mathbf{v}\cdot\nabla\pi + \pi\dot{\sigma}\right) \; . \tag{24.12}$$

To derive the kinetic energy equation in σ coordinates, we dot (24.6) with $\mathbf{v}$ to obtain

$$\left(\frac{\partial K}{\partial t}\right)_\sigma + \mathbf{v}\cdot\nabla_\sigma K + \dot{\sigma}\frac{\partial K}{\partial \sigma} = -\mathbf{v}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi) \; . \tag{24.13}$$

The corresponding flux form is

$$\left[\frac{\partial(\pi K)}{\partial t}\right]_\sigma + \nabla_\sigma\cdot(\pi\mathbf{v}K) + \frac{\partial(\pi\dot{\sigma}K)}{\partial\sigma} = -\pi\mathbf{v}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi)\ . \tag{24.14}$$

The pressure-work term on the right-hand side of (24.14) has to be manipulated to facilitate comparison with (24.11). Begin as follows:

$$\begin{aligned}
-\pi\mathbf{v}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi) &= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi) + \phi\nabla_\sigma\cdot(\pi\mathbf{v}) - \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi) - \phi\left[\frac{\partial\pi}{\partial t} + \frac{\partial(\pi\dot{\sigma})}{\partial\sigma}\right] - \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi) - \frac{\partial(\pi\dot{\sigma}\phi)}{\partial\sigma} + \pi\dot{\sigma}\frac{\partial\phi}{\partial\sigma} - \phi\frac{\partial\pi}{\partial t} - \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi) - \frac{\partial(\pi\dot{\sigma}\phi)}{\partial\sigma} - \left(\pi\dot{\sigma}\alpha\pi + \phi\frac{\partial\pi}{\partial t} + \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi\right)\ .
\end{aligned} \tag{24.15}$$

To get the second line of (24.15) we have used continuity, and to get the final line we have used hydrostatics. One more step is needed, and it is not at all obvious. We know that we need $\pi\omega\alpha$, where ω is given by (24.5). With this in mind, we rewrite the last three terms (in parentheses) on the bottom line of (24.15) as follows:

$$\begin{aligned}
\pi\dot{\sigma}\alpha\pi + \phi\frac{\partial\pi}{\partial t} + \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi &= \pi\omega\alpha - \pi\alpha\left[\sigma\left(\frac{\partial\pi}{\partial t} + \mathbf{v}\cdot\nabla\pi\right) + \pi\dot{\sigma}\right] \\
&\quad + \pi\dot{\sigma}\alpha\pi + \phi\frac{\partial\pi}{\partial t} + \pi\sigma\alpha\mathbf{v}\cdot\nabla\pi \\
&= \pi\omega\alpha - \pi\alpha\sigma\left(\frac{\partial\pi}{\partial t}\right) + \phi\frac{\partial\pi}{\partial t} \\
&= \pi\omega\alpha + \left(\frac{\partial\phi}{\partial\sigma}\sigma + \phi\right)\frac{\partial\pi}{\partial t} \\
&= \pi\omega\alpha + \frac{\partial}{\partial\sigma}\left(\phi\sigma\frac{\partial\pi}{\partial t}\right)\ .
\end{aligned} \tag{24.16}$$

What is the $\dfrac{\partial}{\partial\sigma}\left(\phi\sigma\dfrac{\partial\pi}{\partial t}\right)$ term doing on the last line of (24.16)? It is a contribution to the vertical pressure-work term. Substituting (24.16) back into (24.15), we conclude that

$$\begin{aligned}-\pi\mathbf{v}\cdot(\nabla_\sigma\phi+\sigma\alpha\nabla\pi) &= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi)-\frac{\partial(\pi\dot\sigma)}{\partial\sigma}-\left[\pi\omega\alpha+\frac{\partial}{\partial\sigma}\left(\phi\sigma\frac{\partial\pi}{\partial t}\right)\right]\\ &= -\nabla_\sigma\cdot(\pi\mathbf{v}\phi)-\frac{\partial}{\partial\sigma}\left[\phi\left(\pi\dot\sigma+\sigma\frac{\partial\pi}{\partial t}\right)\right]-\pi\omega\alpha.\end{aligned} \tag{24.17}$$

Using (24.17) in (24.14), we obtain the kinetic energy equation in the form

$$\boxed{\begin{aligned}&\left[\frac{\partial(\pi K)}{\partial t}\right]_\sigma+\nabla_\sigma\cdot[\pi\mathbf{v}(K+\phi)]+\frac{\partial}{\partial\sigma}\left[\pi\dot\sigma K+\phi\left(\pi\dot\sigma+\sigma\frac{\partial\pi}{\partial t}\right)\right]\\ &\qquad=-\sigma\pi\alpha\left(\frac{\partial\pi}{\partial t}+\mathbf{v}\cdot\nabla\pi+\pi\dot\sigma\right),\end{aligned}} \tag{24.18}$$

where (24.12) has been used.

We can now add (24.18) and (24.11) to obtain the total energy equation in σ coordinates:

$$\boxed{\begin{aligned}&\left\{\frac{\partial}{\partial t}[\pi(K+c_pT)]\right\}_\sigma+\nabla_\sigma\cdot[\pi\mathbf{v}(K+c_pT+\phi)]\\ &+\frac{\partial}{\partial\sigma}\left[\pi\dot\sigma(K+c_pT)+\phi\left(\pi\dot\sigma+\sigma\frac{\partial\pi}{\partial t}\right)\right]=0\,.\end{aligned}} \tag{24.19}$$

24.3 Conservation of total energy in vertically discrete sigma-coordinate models

We now investigate conservation properties of the vertically discrete equations, using σ-coordinates, and *using the L-grid*. The discussion follows Arakawa and Lamb (1977), although some of the ideas originated with Lorenz (1960). For simplicity, we keep both the temporal and horizontal derivatives in continuous form.

We begin by writing down the vertically discrete prognostic equations of the model. Conservation of mass is expressed, in the vertically discrete system, by

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{v}_l) + \left[\frac{\delta (\pi \dot{\sigma})}{\delta \sigma} \right]_l = 0, \tag{24.20}$$

where

$$[\delta ( )]_l \equiv ( )_{l+\frac{1}{2}} - ( )_{l-\frac{1}{2}}. \tag{24.21}$$

Similarly, conservation of potential temperature is expressed, in flux form, by

$$\frac{\partial (\pi \theta_l)}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{v}_l \theta_l) + \left[\frac{\delta (\pi \dot{\sigma} \theta)}{\delta \sigma} \right]_l = 0. \tag{24.22}$$

Here we omit the heating term, for simplicity. In order to use (24.22) it is necessary to define values of θ at the layer edges, via an interpolation. In Chapter 11 we analyzed the interpolation issue in the context of horizontal advection, and that same discussion applies to vertical advection as well. As one possibility, the interpolation methods that allow conservation of an arbitrary function of the advected quantity can be used for vertical advection. As discussed later, a different choice may be preferable.

The hydrostatic equation is

$$\left(\frac{\delta \phi}{\delta \sigma} \right)_l = \pi \alpha_l \, . \tag{24.23}$$

This equation involves the geopotentials at the layer edges, and also the specific volume in the layer center. These must be determined somehow, by starting from the prognostic variables of the model.

Finally, the momentum equation is

$$\frac{\partial \mathbf{v}_l}{\partial t} + [f + \mathbf{k} \cdot (\nabla_\sigma \times \mathbf{v}_l)] \, \mathbf{k} \times \mathbf{v}_l + \left(\dot{\sigma} \frac{\partial \mathbf{v}}{\partial \sigma} \right)_l + \nabla K_l = -\nabla \phi_l - (\sigma \alpha)_l \nabla \pi \, . \tag{24.24}$$

Here we omit the friction term, for simplicity. The momentum equation involves the geopotentials at the layer centers, which will have to be determined somehow, presumably using

the hydrostatic equation. Note, however, that the hydrostatic equation listed above involves the geopotentials at the layer edges, rather than the layer centers.

To complete the system, we need the upper and lower boundary conditions

$$\dot{\sigma}_{\frac{1}{2}} = \dot{\sigma}_{L+\frac{1}{2}} = 0 \, . \tag{24.25}$$

We define the vertical coordinate, σ, at layer edges, which are denoted by half-integer subscripts. The change in σ across layer l is written as $\delta\sigma_l$. Note that

$$\sum_{l=1}^{L} \delta\sigma_l = 1 \, , \tag{24.26}$$

and

$$p_{l+\frac{1}{2}} = \pi\sigma_{l+\frac{1}{2}} + p_T \, , \tag{24.27}$$

where p_T is a constant, and the constant values of $\sigma_{l+\frac{1}{2}}$ are assumed to be prescribed for each layer edge. Eq. (24.27) tells how to compute layer-edge pressures. A method to determine layer-center pressures is also needed, and will be discussed later.

By summing (24.20) over all layers, and using (24.25), we obtain

$$\frac{\partial \pi}{\partial t} + \nabla \cdot \left\{ \sum_{l=1}^{L} \left[\pi \mathbf{v}_l \left(\delta\sigma \right)_l \right] \right\} = 0 \, , \tag{24.28}$$

which is the vertically discrete form of the surface pressure tendency equation. From (24.28), we see that mass is, in fact, conserved, i.e., the vertical mass fluxes do not produce any net source or sink of mass. We can use (24.28) with (24.20) to determine $\pi\dot{\sigma}$ at the layer edges, exactly paralleling the method used to determine $\pi\dot{\sigma}$ with the vertically continuous system of equations.

24.3.1 The horizontal pressure-gradient force

A finite-difference analog of (23.63) is

$$-\left(\frac{\pi}{\rho}\nabla_z p\right)_l = \left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l \nabla\pi - \nabla(\pi\phi_l) \ . \tag{24.29}$$

Multiplying (24.29) by $\delta\sigma_l$, and summing over all layers, we obtain

$$\begin{aligned} -\sum_{l=1}^{L}\left(\frac{\pi}{\rho}\nabla_z p\right)_l (\delta\sigma)_l &= \sum_{l=1}^{L}\left[\delta(\sigma\phi)\right]_l \nabla\pi - \sum_{l=1}^{L}\left[\nabla(\pi\phi_l)(\delta\sigma)_l\right] \\ &= \phi_S\nabla\pi - \nabla\left\{\sum_{l=1}^{L}\left[(\pi\phi_l)(\delta\sigma)_l\right]\right\} \ . \end{aligned} \tag{24.30}$$

This is analogous to our earlier result for the continuous system. Inspection of (24.25) shows that, if we use the form of the horizontal pressure-gradient force given by (24.29), the vertically summed horizontal pressure-gradient force cannot spin up or spin down a circulation inside a closed path, in the absence of topography (Arakawa and Lamb, 1977).

The idea outlined above provides a rational way to choose which of the many possible forms of the horizontal pressure-gradient force should be used in a model. At this point the form is not fully determined, however, because we do not yet have a method to compute either ϕ_l or the layer-edge values of ϕ that appear in (24.29).

Eq. (24.29) is equivalent to

$$-\left(\frac{\pi}{\rho}\nabla_z p\right)_l = \left\{\left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l - \phi_l\right\}\nabla\pi - \pi\nabla\phi_l \ . \tag{24.31}$$

By comparison with (23.60), we identify

$$\pi(\sigma\alpha)_l = \phi_l - \left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l \ . \tag{24.32}$$

An analogous equation is true in the continuous case. This allows us to write (24.31) as

$$-\left(\frac{\pi}{\rho}\nabla_z p\right)_l = -\pi(\sigma\alpha)_l\nabla\pi - \pi\nabla\phi_l \ . \tag{24.33}$$

Eq. (24.33) will be used later.

24.3.2 The thermodynamic energy equation

Suppose that we choose to predict θ_l by using (24.22), because we want to conserve the globally mass-integrated value of θ in the absence of heating. We relate the temperature to the potential temperature using

$$c_p T_l = \Pi_l \theta_l \, . \tag{24.34}$$

In order to use (24.34), we need a way to determine

$$\Pi_l \equiv c_p \left(\frac{p_l}{p_0} \right)^{\kappa} \, . \tag{24.35}$$

Norman Phillips (1974) suggested

$$\Pi_l = \left(\frac{1}{1+\kappa} \right) \left[\frac{\delta \left(\Pi p \right)}{\delta p} \right]_l \, , \tag{24.36}$$

on the grounds that this form leads to a good simulation of vertical wave propagation. Eq. (24.36) gives us away to compute the layer-center value of the Exner function, and the layer-center value of the pressure, from the neighboring layer-edge values. Tokioka (1978) showed that with (24.36), the finite-difference hydrostatic equation (discussed later) is exact for atmospheres in which the potential temperature is uniform with height.

The advective form of the potential temperature equation can be obtained by combining (24.22) with (24.20):

$$\pi \left(\frac{\partial \theta_l}{\partial t} + \mathbf{v}_l \cdot \nabla \theta_l \right) + \left[\frac{(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\theta_{l+\frac{1}{2}} - \theta_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\theta_l - \theta_{l-\frac{1}{2}} \right)}{(\delta \sigma)_l} \right] = 0 \, . \tag{24.37}$$

A similar manipulation was shown way back in Chapter 7. Substitute (24.34) into (24.37), to obtain the corresponding prediction equation for T_l:

$$c_p\pi\left(\frac{\partial T_l}{\partial t}+\mathbf{v}_l\cdot\nabla T_l\right)-\pi\theta_l\frac{\partial \Pi_l}{\partial \pi}\left(\frac{\partial \pi}{\partial t}+\mathbf{v}_l\cdot\nabla\pi\right)$$
$$+\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(\Pi_l\theta_{l+\frac{1}{2}}-c_pT_l\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(c_pT_l-\Pi_l\theta_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]=0\,. \tag{24.38}$$

The derivative $\partial\Pi_l/\partial\pi$ can be evaluated using (24.36). We now introduce the terms that represent the vertical advection of temperature, modeled after the corresponding terms of (24.37). These involve the layer-edge temperatures, i.e., $T_{l+\frac{1}{2}}$ and $T_{l-\frac{1}{2}}$, but keep in mind that a method to determine the layer-edge temperatures has not yet been specified. By simply "adding and subtracting," we rewrite (24.38) as

$$c_p\pi\left(\frac{\partial T_l}{\partial t}+\mathbf{v}_l\cdot\nabla T_l\right)+c_p\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(T_{l+\frac{1}{2}}-T_l\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(T_l-T_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]$$
$$=\pi\theta_l\frac{\partial \Pi_l}{\partial \pi}\left(\frac{\partial \pi}{\partial t}+\mathbf{v}_l\cdot\nabla\pi\right) \tag{24.39}$$
$$+\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(c_pT_{l+\frac{1}{2}}-\Pi_l\theta_{l+\frac{1}{2}}\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(\Pi_l\theta_{l-\frac{1}{2}}-c_pT_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]\,.$$

The layer-edge temperatures can simply be cancelled out in (24.33) to recover (24.32). Obviously, the left-hand side of (24.39) can be rewritten in flux form through the use of the vertically discrete continuity equation:

$$c_p\left\{\frac{\partial}{\partial t}(\pi T_l)+\nabla\cdot(\pi\mathbf{v}_lT_l)+\left[\frac{\delta(\pi\sigma T)}{\delta\sigma}\right]_l\right\}=\pi\theta_l\frac{\partial \Pi_l}{\partial \pi}\left(\frac{\partial \pi}{\partial t}+\mathbf{v}_l\cdot\nabla\pi\right)$$
$$+\frac{1}{(\delta\sigma)_l}\left[(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(c_pT_{l+\frac{1}{2}}-\Pi_l\theta_{l+\frac{1}{2}}\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(\Pi_l\theta_{l-\frac{1}{2}}-c_pT_{l-\frac{1}{2}}\right)\right]\,. \tag{24.40}$$

We now observe, by comparison of (24.40) with the continuous form (24.10), that the expression on the right-hand side of (24.40) must be a form of $\pi\omega\alpha$, i.e.,

$$\boxed{\begin{aligned}\pi\left(\omega\alpha\right)_l &= \pi\theta_l\frac{\partial\Pi_l}{\partial\pi}\left(\frac{\partial\pi}{\partial t}+\mathbf{v}_l\cdot\nabla\pi\right)\\ &+\left[\frac{\left(\pi\dot{\sigma}\right)_{l+\frac{1}{2}}\left(c_pT_{l+\frac{1}{2}}-\Pi_l\theta_{l+\frac{1}{2}}\right)+\left(\pi\dot{\sigma}\right)_{l-\frac{1}{2}}\left(\Pi_l\theta_{l-\frac{1}{2}}-c_pT_{l-\frac{1}{2}}\right)}{\left(\delta\sigma\right)_l}\right].\end{aligned}} \tag{24.41}$$

Eq. (24.41) is a finite-difference analog of the not-so-obvious continuous equation

$$\pi\omega\alpha=\pi\theta\frac{\partial\pi}{\partial\pi}\left(\frac{\partial\pi}{\partial t}+\mathbf{v}\cdot\nabla\pi\right)+\frac{\partial\left(\pi\dot{\sigma}c_pT\right)}{\partial\sigma}-\Pi\frac{\partial\left(\pi\dot{\sigma}\theta\right)}{\partial\sigma}\,, \tag{24.42}$$

which you should be able to prove is correct. We will return to (24.41) below, after deriving the corresponding expression from the mechanical energy side of the problem.

24.3.3 The mechanical energy equation

We now derive the mechanical energy equation using the vertically discrete system. Taking the dot product of $\pi\mathbf{v}_l$ with the horizontal pressure-gradient force for layer l, we write, closely following the continuous case,

$$\begin{aligned}-\pi\mathbf{v}_l\cdot\left[\nabla\phi_l+\left(\sigma\alpha\right)_l\nabla\pi\right] &= -\nabla\cdot\left(\pi\mathbf{v}_l\phi_l\right)+\phi_l\nabla\cdot\left(\pi\mathbf{v}_l\right)-\pi\left(\sigma\alpha\right)_l\mathbf{v}_l\cdot\nabla\pi\\ &= -\nabla\cdot\left(\pi\mathbf{v}_l\phi_l\right)-\phi_l\left\{\frac{\partial\pi}{\partial t}+\left[\frac{\delta\left(\pi\dot{\sigma}\phi\right)}{\delta\sigma}\right]_l\right\}-\pi\left(\sigma\alpha\right)_l\mathbf{v}_l\cdot\nabla\pi\\ =-\nabla\cdot\left(\pi\mathbf{v}_l\phi_l\right)-\left[\frac{\delta\left(\pi\dot{\sigma}\phi\right)}{\delta\sigma}\right]_l &+\left[\frac{\left(\pi\sigma\right)_{l+\frac{1}{2}}\left(\phi_{l+\frac{1}{2}}-\phi_l\right)+\left(\pi\dot{\sigma}\right)_{l-\frac{1}{2}}\left(\phi_l-\phi_{l-\frac{1}{2}}\right)}{\left(\delta\sigma\right)_l}\right]\\ &-\phi_l\frac{\partial\pi}{\partial t}-\pi\left(\sigma\alpha\right)_l\mathbf{v}_l\cdot\nabla\pi\,.\end{aligned} \tag{24.43}$$

Continuing down this path, we construct the terms that we need by adding and subtracting

$$\begin{aligned}-\pi\mathbf{v}_l\left[\nabla\phi_l+\left(\sigma\alpha\right)_l\nabla\pi\right] &= -\nabla\cdot\left(\pi\mathbf{v}_l\phi_l\right)-\left[\frac{\delta\left(\pi\dot{\sigma}\phi\right)}{\delta\sigma}\right]_l+\left[\pi(\sigma\alpha)_l-\phi_l\right]\frac{\partial\pi}{\partial t}\\ -\pi\left\{\left(\sigma\alpha\right)_l\left(\frac{\partial\pi}{\partial t}+\mathbf{v}_l\cdot\nabla\pi\right)\right. &\left.-\left[\frac{\left(\pi\dot{\sigma}\right)_{l+\frac{1}{2}}\left(\phi_{l+\frac{1}{2}}-\phi_l\right)+\left(\pi\dot{\sigma}\right)_{l-\frac{1}{2}}\left(\phi_l-\phi_{l-\frac{1}{2}}\right)}{\pi(\delta\sigma)_l}\right]\right\}.\end{aligned} \tag{24.44}$$

Using the continuity equation (24.20), we can rewrite (24.44) as

$$-\pi \mathbf{v}_l \cdot [\nabla \phi_l + (\sigma\alpha)_l \nabla \pi] = -\nabla \cdot (\pi \mathbf{v}_l \phi_l) - \left\{ \frac{\delta \left[\left(\pi\dot{\sigma} + \sigma \frac{\partial \pi}{\partial t} \right) \phi \right]}{\delta \sigma} \right\}_l$$
$$-\pi \left\{ (\sigma\alpha)_l \left(\frac{\partial \pi}{\partial t} + \mathbf{v}_l \cdot \nabla \pi \right) - \left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right)}{\pi (\delta\sigma)_l} \right] \right\}. \tag{24.45}$$

By comparing with the continuous form, (24.17), we infer that

$$\boxed{\begin{array}{c} \pi(\omega\alpha)_l = \pi(\sigma\alpha)_l \left(\dfrac{\partial \pi}{\partial t} + \mathbf{v}_l \cdot \nabla \pi \right) \\ -\dfrac{1}{(\delta\sigma)_l} \left[(\pi\dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right) \right] . \end{array}} \tag{24.46}$$

24.3.4 Total energy conservation

We have now reached the crux of the problem. *To ensure total energy conservation, the form of* $\pi(\omega\alpha)_l$ *given by* (24.46) *must match the form given by* (24.41). Comparison of the two equations shows that this can be accomplished by setting:

$$(\sigma\alpha)_l = \theta_l \frac{\partial \Pi_l}{\partial \pi} , \tag{24.47}$$

$$\phi_l - \phi_{l+\frac{1}{2}} = c_p T_{l+\frac{1}{2}} - \Pi_l \theta_{l+\frac{1}{2}} , \tag{24.48}$$

and

$$\phi_{l-\frac{1}{2}} - \phi_l = \Pi_l \theta_{l-\frac{1}{2}} - c_p T_{l-\frac{1}{2}} . \tag{24.49}$$

As discussed below, all three of these equations are vertically discrete forms of the hydrostatic equation.

Eq. (24.47) gives us an expression for $(\sigma\alpha)_l$. We already had one, though, in Eq. (24.32). By requiring that these two expressions to agree, we obtain

$$\boxed{\phi_l - \left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l = \pi\theta_l \frac{\partial\Pi_l}{\partial\pi}}. \tag{24.50}$$

This is yet another a finite-difference form of the hydrostatic equation. It involves geopotentials at both layer centers and layer edges. You should be able to derive the continuous form of the hydrostatic equation that corresponds to (24.50).

By adding $\Pi_l\theta_l$ to both sides of both (24.48) and (24.49), and using (24.34), we find that

$$\left(c_p T_{l+\frac{1}{2}} + \phi_{l+\frac{1}{2}}\right) - (c_p T_l + \phi_l) = \Pi_l \left(\theta_{l+\frac{1}{2}} - \theta_l\right) , \tag{24.51}$$

and

$$(c_p T_l + \phi_l) - \left(c_p T_{l-\frac{1}{2}} + \phi_{l-\frac{1}{2}}\right) = \Pi_l \left(\theta_l - \theta_{l-\frac{1}{2}}\right) , \tag{24.52}$$

respectively. These finite-difference analogs of the hydrostatic equation have the familiar form $\partial s/\partial\theta = \Pi$. Add one to each subscript in (24.52), and add the result to (24.51). The layer-edge geopotential cancels out, and we get

$$\boxed{\phi_{l+1} - \phi_l = -\theta_{l+\frac{1}{2}}(\Pi_{l+1} - \Pi_l)}. \tag{24.53}$$

This is a finite-difference version of yet another form of the hydrostatic equation, namely $\partial\phi/\partial\Pi = -\theta$. What have we gained by the manipulation just performed? If the forms of Π_l and $\theta_{l+1/2}$ are specified, we can use (24.53) to integrate the hydrostatic equation upward from level $l+1$ to level l.

24.3.5 The problem with the L grid

In (24.53), the problem with the L grid becomes apparent. We must determine $\theta_{l+\frac{1}{2}}$ by some form of interpolation, e.g., the arithmetic mean of the neighboring layer-center values of θ. The interpolation will "hide" a vertical zig-zag in the solution for θ. A hidden zig-zag cannot influence the pressure-gradient force, so it cannot propagate, as a physical solution would. The dynamically inert zig-zag can become a permanent, unwelcome feature of the simulated temperature sounding. This problem does not arise with the CP grid.

The problem is actually both more complicated and more serious than it may appear at this point. Although we can use (24.53) to integrate the hydrostatic equation upward, it is still necessary to provide a boundary condition to determine the starting value, ϕ_l, i.e., the layer-center geopotential for the lowest layer. This can be done by first summing $(\delta\sigma)_l$ times (24.50) over all layers:

$$\sum_{l=1}^{L} \phi_l \, (\delta\sigma)_l - \phi_S = \sum_{l=1}^{L} \pi\theta_l \frac{\partial \Pi_l}{\partial \pi} \, (\delta\sigma)_l \ . \tag{24.54}$$

Now we use the mathematical identity

$$\begin{aligned} \sum_{l=1}^{L} \phi_l (\delta\sigma)_l &= \sum_{l=1}^{L} \phi_l \left(\sigma_{l+\frac{1}{2}} - \sigma_{l-\frac{1}{2}} \right) \\ &= \phi_L + \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} \left(\phi_l - \phi_{l+1} \right) \ , \end{aligned} \tag{24.55}$$

which applies for any quantity defined at layer centers. Substitution of (24.55) into the left-hand side of (24.54), and use of (24.53), gives

$$\phi_L = \phi_S + \sum_{l=1}^{L} \pi\theta_l \frac{\partial \Pi_l}{\partial \pi} \, (\delta\sigma)_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} \left(\Pi_{l+1} - \Pi_l \right) \theta_{l+\frac{1}{2}} \ , \tag{24.56}$$

which can be used to determine the geopotential height at the lowest layer center. We can then use (24.53) to determine the geopotential for the remaining layers above.

Eq. (24.56) is a bit odd, howver, because it says that *the thickness between the Earth's surface and the middle of the lowest model layer depends on all of the values of θ_l, throughout the entire column*. An interpretation is that all values of θ_l are being used to estimate the effective value of θ between the surface and level L. Since we integrate up from ϕ_l

to determine ϕ_l for $l < L$, *all values of* θ_l *are being used to determine each value of* ϕ_l *throughout the entire column*. This means that the hydrostatic equation is very non-local, i.e., the thickness between each pair of layers is influenced by the potential temperature at all model levels.

To avoid this problem, Arakawa and Suarez (1983) proposed an interpolation for $\theta_{l+\frac{1}{2}}$ in which only θ_l influences the thickness between the surface and the middle of the bottom layer. To see how this works, the starting point is to write local hydrostatic equation in the form

$$\phi_l - \phi_{l+1} = c_p \left(A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1} \right) , \tag{24.57}$$

where $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ are non-dimensional parameters to be determined. Comparing with (24.53), we see that

$$\boxed{(\Pi_{l+1} - \Pi_l)\, \theta_{l+\frac{1}{2}} = A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1}} . \tag{24.58}$$

In order that (24.58) have the form of an interpolation, we must choose $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ so that

$$\frac{A_{l+\frac{1}{2}} + B_{l+\frac{1}{2}}}{\Pi_{l+1} - \Pi_l} = 1 . \tag{24.59}$$

Eq. (24.58) essentially determines the form of $\theta_{l+\frac{1}{2}}$, if the forms of $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ are specified.

After substitution from (24.57), Eq. (24.56) becomes

$$\phi_L - \phi_S = \sum_{l=1}^{L} \pi \theta_l \frac{\partial \Pi_l}{\partial \pi} (\delta\sigma)_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} \left(A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1} \right) . \tag{24.60}$$

Every term on the right-hand-side of (24.60) involves a layer-center value of θ. *To eliminate any dependence of* ϕ_l *on the values of* θ *above the lowest layer, we "collect terms" around individual values of* θ_l*, and force the coefficients to vanish for* $l < L$. This leads to

$$\pi \frac{\partial \Pi_l}{\partial \pi} (\delta\sigma)_l = \sigma_{l+\frac{1}{2}} A_{l+\frac{1}{2}} + \sigma_{l-\frac{1}{2}} B_{l-\frac{1}{2}} \text{ for } l < L \tag{24.61}$$

With the use of (24.61), (24.60) simplifies to

$$\boxed{\phi_L = \phi_S + \left[\pi \frac{\partial \Pi_l}{\partial \pi} (\delta\sigma)_L - \sigma_{L-\frac{1}{2}} B_{L-\frac{1}{2}} \right] c_p \theta_l} . \tag{24.62}$$

Because the coefficient of each θ_l has been forced to vanish for all $l < L$, only θ_L influences ϕ_L. We have succeeded in making the thickness between the surface and the middle of the lowest layer depend only on the lowest-layer temperature. Note, however, that the thicknesses between the layer centers still depend on interpolated or averaged potential temperatures, so we still have the L grid's problem of the dynamically inert zig-zag in temperature, although it is not as serious as before.

Once the lowest-layer geopotential has been determined from (24.62), we can use either (24.53) or (24.57) to determine the geopotentials for the remaining layers; the result is the same with either method.

Methods to choose $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ are discussed by Arakawa and Suarez (1983). They recommended

$$A_{l+\frac{1}{2}} = \Pi_{l+\frac{1}{2}} - \Pi_l \text{ and } B_{l+\frac{1}{2}} = \Pi_{l+1} - \Pi_{l+\frac{1}{2}} , \tag{24.63}$$

which satisfy (24.59).

If we use (24.58)*, we are not free to use the methods of Chapter 7 to choose* $\theta_{l+\frac{1}{2}}$ *in such a way that some* $F(\theta)$ *is conserved.* A choice has to be made between these two alternatives. It seems preferable to use (24.58).

24.4 Summary and conclusions

The problem of representing the vertical structure of the atmosphere in numerical models is receiving a lot of attention at present. Among the most promising of the current approaches are those based on isentropic or hybrid-isentropic coordinate systems. Similar methods are being used in ocean models.

At the same time, models are more commonly being extended through the stratosphere and beyond, and vertical resolutions are increasing; the era of hundred-layer models is upon us.

Chapter 25

Aliasing instability

25.1 Scale interactions and nonlinearity

"Nonlinear" is a mathematical term. A more physical perspective is that the processes that are described by nonlinear mathematical terms bring about interactions among scales in a fluid system. Scale interactions arise when we try to solve either nonlinear equations or linear equations with variable coefficients. For example, suppose that we have two modes on a one-dimensional grid, given by

$$A\left(x_j\right) = \hat{A}e^{ikj\Delta x} \text{ and } B\left(x_j\right) = \hat{B}e^{ilj\Delta x}\,, \tag{25.1}$$

respectively. Here the wave numbers of A and B are denoted by k and l, respectively. We assume that k and l both "fit" on the grid in question. If we combine A and B linearly, e.g., form

$$\alpha A + \beta B\,, \tag{25.2}$$

where α and β are *spatially constant* coefficients, then no "new" waves are generated; k and l continue to be the only wave numbers present. In contrast, if we multiply A and B together, then we generate the new wave number, $k+l$:

$$AB = \hat{A}\hat{B}e^{i(k+l)j\Delta x}\,, \tag{25.3}$$

Other nonlinear operations such as division, exponentiation, etc., will also generate new wave numbers. It can easily happen that $(k+l)\Delta x > \pi$, in which case the new mode created by multiplying A and B together does not fit on the grid. *What actually happens in such a case is that the new mode is "aliased" onto a mode that* ***does*** *fit on the grid.*

25.1.1 Aliasing error

Suppose that we have a wave given by the continuous solid line in Fig. 25.1. There are discrete, evenly spaced grid points along the x-axis, as shown by the black dots in the figure. The wave has been drawn with a wave length of $(4/3)\Delta x$, corresponding to a wave number of $\frac{3\pi}{2\Delta x}$. Because $(4/3)\Delta x < 2\Delta x$, *the wave is too short to be represented on the grid.* What the grid points "see" instead is not the wave represented by the solid line, but rather the wave of wavelength $4\Delta x$, as indicated by the dashed line (again drawn as a continuous function of x). At the grid points, the wave of length $4\Delta x$ takes exactly the values that the wave of $(4/3)\Delta x$ would take at those same grid points, if it could be represented on the grid at all. This misrepresentation of a wavelength too short to be represented on the grid is called "aliasing error." *Aliasing is a high wave number (or frequency) masquerading as a low wave number (or frequency).* In the example of Fig. 25.1, aliasing occurs because the grid is too coarse to resolve the wave of length $(4/3)\Delta x$. Another way of saying this is that the wave is not adequately "sampled" by the grid. *Aliasing error is always due to inadequate sampling.*

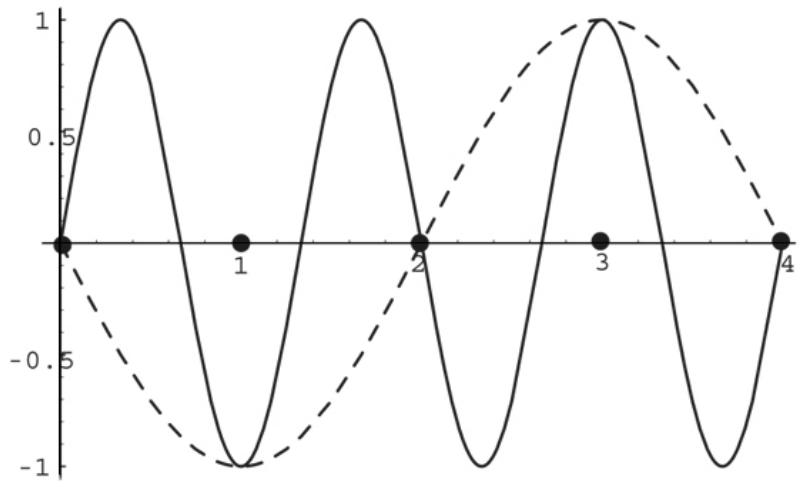


Figure 25.1: An example of aliasing error. Distance along the horizontal axis is measured in units of Δx. The wave given by the solid line has a wave length of $(4/3)\Delta x$. This is shorter than $2\Delta x$, and so the wave cannot be represented on the grid. Instead, the grid "sees" a wave of wavelength $4\Delta x$, as indicated by the dashed line. Note that the $4\Delta x$-wave is "upside-down."

25.1.2 Almost famous

Aliasing error can be important in observational studies, because observations taken "too far apart" in space (or time) can make a short wave (or high frequency) appear to be a longer wave (or lower frequency). Fig. 25.2 is an example, from real life. The blue curve in the figure makes it appear that the precipitation rate averaged over the global tropics fluctuates with a period of 23 days and an amplitude approaching 1 mm day^{-1}. If this tropical precipitation oscillation (TPO) were real, it would be one of the most amazing phenomena in atmospheric science, and its discoverer would no doubt appear on the cover of *Rolling Stone*. But alas, the TPO is bogus, even though you can see it with your own eyes in Fig. 25.2, and even though the figure is based on real data.

How is that possible? The satellite from which the data was collected has an orbit that takes it over the same point on Earth *at the same time of day* once every 23 days. Large regions of the global tropics have a strong diurnal (i.e., day-night) oscillation of the precipitation rate. This high-frequency diurnal signal is aliased onto a much lower frequency, i.e., 23 days, because *the sampling by the satellite is not frequent enough* to resolve the diurnal cycle.

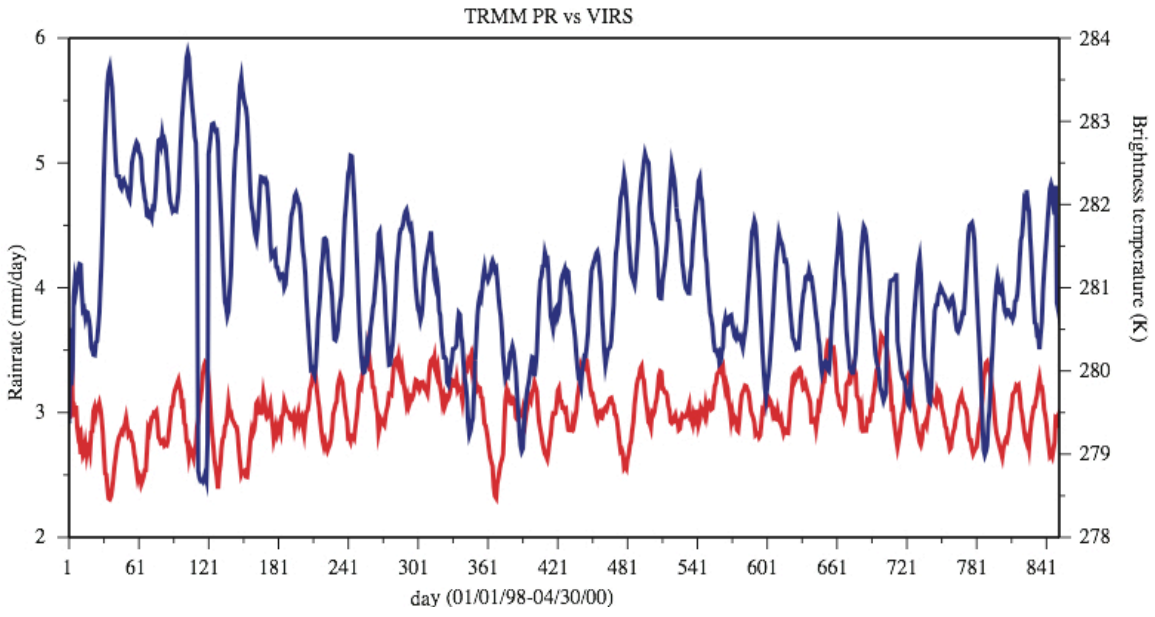


Figure 25.2: An example of aliasing in the analysis of observations. The blue curve shows the precipitation rate, averaged over the global tropics (20° S to 20° N), and the red curve shows a the thermal radiation in the 11.8 μm band, averaged over the same region. The horizontal axis is time, and the period covered is slightly more than two years. The data were obtained from the TRMM (Tropical Rain Measuring Mission) satellite. The obvious oscillation in both curves, with a period close to 23 days, is an artifact due to aliasing. See the text for further explanation.

25.1.3 A mathematical view of aliasing

In an earlier chapter, we saw that the shortest wavelength that a grid can represent is $L = 2\Delta x$, the maximum representable wave number is

$$k_{\max} \equiv \pi/\Delta x \,. \tag{25.4}$$

What happens when a wave with $k > k_{\max}$ is produced through nonlinear interactions? Since $2k_{\max}\Delta x = 2\pi$, a wave with $k > 2k_{\max} = 2\pi/\Delta x$ "fold s back." We can therefore assume that

$$2k_{\max} > k > k_{\max} \,. \tag{25.5}$$

The expression $\sin(kj\Delta x)$ can be written as as

$$\begin{aligned} \sin[kj\Delta x] &= \sin[(2k_{\max} - 2k_{\max} + k)\, j\Delta x] \\ &= \sin[2\pi j - (2k_{\max} - k)\, j\Delta x] \\ &= \sin[-(2k_{\max} - k)\, j\Delta x] \\ &= \sin[k^*(j\Delta x)] \,, \end{aligned} \tag{25.6}$$

where

$$\boxed{k^* = k \quad \text{for} \quad |k| \le k_{\max} \quad \text{and} \quad k^* \equiv -(2k_{\max} - k) \ \text{for} \ |k| > k_{\max} \,.} \tag{25.7}$$

Note that $0 < |k^*| < k_{\max}$ because of (25.5). Corresponding to (25.6), we have

$$\cos(kj\Delta x) = \cos(k^* j\Delta x) \,. \tag{25.8}$$

Eqs. (25.6) and (25.8) show that *a wave of wave number $k > k_{\max}$ is interpreted (or misinterpreted) by the grid as a wave of wave number k^**. The minus sign means that the phase change per Δx is reversed, or "backward s ."

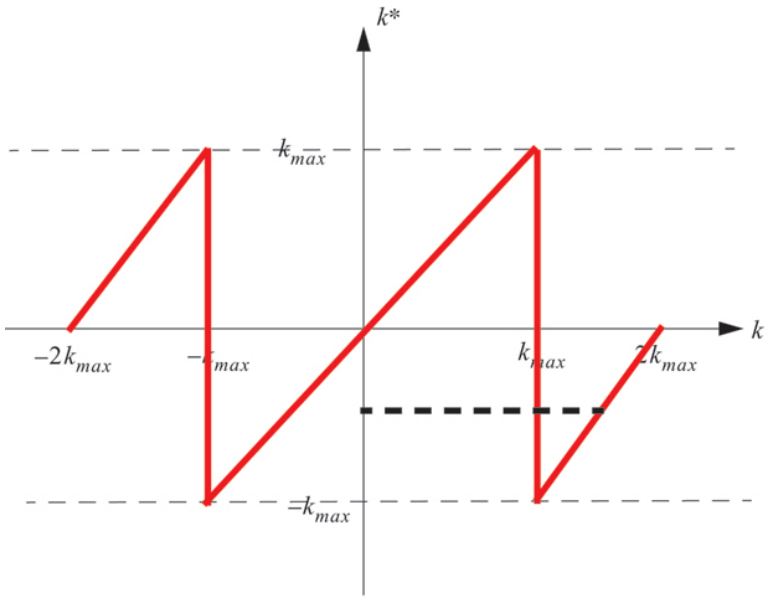


Figure 25.3: The red line is a plot of k^* on the vertical axis, versus k on the horizontal axis. The dashed black line connects $k\Delta x = 3\pi/2$ with $k^*\Delta x = -\pi/2$ (for a wavelength of $4\Delta x$) , corresponding to the example of Fig. 25.1.

Fig. 25.3 illustrates how k^* varies with k. For $-k_{\max} \le k \le k_{\max}$, we simply have $k^* = k$. For $k > k_{\max}$, we get $0 > k^* > -k_{\max}$, and so on.

In the example shown in Fig. 25.1, $L = (4/3)\,\Delta x$ so $k = \frac{2\pi}{L} = \frac{3\pi}{2\Delta x}$. Therefore $k^* \equiv -(2k_{\max} - k) = \dfrac{2\pi}{\Delta x} - \dfrac{3\pi}{2\Delta x} = \dfrac{\pi}{2\Delta x}$, which implies that $L^* = 4\Delta x$, as we have already surmised by inspection of the figure.

For $k < k_{\max}$, the phase change, as j increases by one, is less than π. This is shown in Fig. 25.4 a. For $k > k_{\max}$, the phase change, as j increases by one, is greater than π. This is shown in Fig. 25.4 b. For $k > k_{\max}$, the dot in the figure appears to move clockwise, i.e., "backward s ." This is a manifestation of aliasing that is familiar from the movies, in which wheels appear to spin backward s when the frame rate is too slow to capture the true motion. It also helps in understanding the minus sign that appears in Eq. (25.7).

25.2 Advection by a variable, non-divergent current

We now prepare for a discussion of aliasing errors in numerical models of the atmosphere. Some background is needed on two-dimensional nondivergent flow.

Suppose that an arbitrary variable q is advected in two dimensions, so that

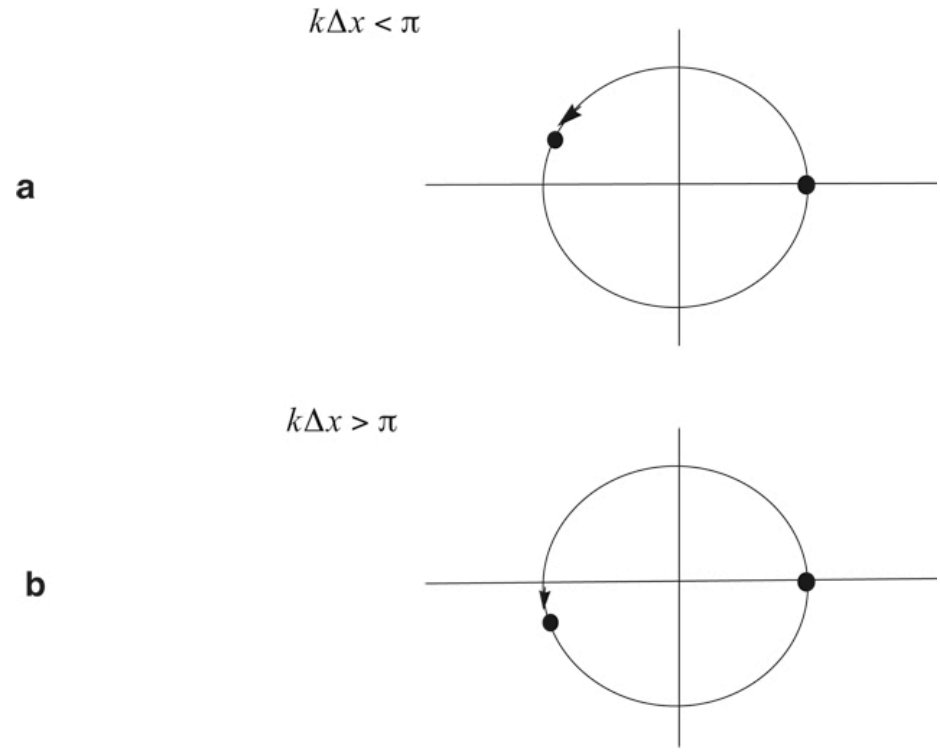


Figure 25.4: The phase change per grid point for: a) $k\Delta x < \pi$, and b) $k\Delta x > \pi$.

$$\frac{\partial q}{\partial t} + \mathbf{v} \cdot \nabla q = 0 \,, \tag{25.9}$$

where the flow is assumed to be non-divergent, i.e.,

$$\nabla \cdot \mathbf{v} = 0 \,. \tag{25.10}$$

Two-dimensional non-divergent flow is a not-too-drastic idealization of the large-scale circulation of the atmosphere. In view of (25.10), we can describe $\mathbf{v}$ in terms of a stream function ψ, such that

$$\mathbf{v} = \mathbf{k} \times \nabla \psi \tag{25.11}$$

In Cartesian coordinates, $u = -\partial \psi / \partial y$, and $v = \partial \psi / \partial x$. (These sign conventions are

arbitrary, but is essential that one of the derivatives has a plus sign and the other a minus sign.)

25.3 The Jacobian

Substituting (25.11) into (25.9), we get

$$\frac{\partial q}{\partial t} + (\mathbf{k} \times \nabla \psi) \cdot \nabla q = 0 \,. \tag{25.12}$$

Using the identity

$$(\mathbf{v}_1 \times \mathbf{v}_2) \cdot \mathbf{v}_3 = \mathbf{v}_2 \cdot (\mathbf{v}_3 \times \mathbf{v}_1) \ , \tag{25.13}$$

which holds for any three vectors, we set $\mathbf{v}_1 \equiv \mathbf{k}$, $\mathbf{v}_2 \equiv \nabla \psi$, and $\mathbf{v}_3 \equiv \nabla q$, to obtain

$$(\mathbf{k} \times \nabla \psi) \cdot \nabla q = \mathbf{k} \cdot (\nabla \psi \times \nabla q) \ . \tag{25.14}$$

This allows us to re-write (25.12) as

$$\frac{\partial q}{\partial t} + J(\psi,\, q) = 0 \,, \tag{25.15}$$

or alternatively as

$$\frac{\partial q}{\partial t} = J(q,\, \psi) \ . \tag{25.16}$$

Here J is the *Jacobian* operator, which is defined by

$$\begin{aligned} J(A,\, B) \equiv & \mathbf{k} \cdot (\nabla A \times \nabla B) \\ = & - \mathbf{k} \cdot \nabla \times (A \nabla B) \\ = & \mathbf{k} \cdot \nabla \times (B \nabla A) \ , \end{aligned} \tag{25.17}$$

for arbitrary A and B. See Appendix A. Note that

$$J(A,\,B) = -J(B,\,A)\ , \tag{25.18}$$

which can be deduced from (25.17). Eq. (25.18) has been used to go from (25.15) to (25.16). From the definition of the Jacobian, it follows that $J(p,q) = 0$ if either A or B is constant.

For the case of Cartesian coordinates, we can write $J(A,\,B)$, in the following three alternative forms, which are suggested by (25.17):

$$J(A,\,B) = \frac{\partial A}{\partial x}\frac{\partial B}{\partial y} - \frac{\partial A}{\partial y}\frac{\partial B}{\partial x} \tag{25.19}$$

$$J(A,\,B) = \frac{\partial}{\partial y}\left(B\frac{\partial A}{\partial x}\right) - \frac{\partial}{\partial x}\left(B\frac{\partial A}{\partial y}\right) \tag{25.20}$$

$$J(A,\,B) = \frac{\partial}{\partial x}\left(A\frac{\partial B}{\partial y}\right) - \frac{\partial}{\partial y}\left(A\frac{\partial B}{\partial x}\right)\ . \tag{25.21}$$

These will be used later.

Let an overbar denote an average over a two-dimensional domain that has no boundaries (e.g., a sphere or a torus), or on the boundary of which either A or B is constant. You should be able to prove the following:

$$\overline{J(A,\,B)} = 0\ , \tag{25.22}$$

$$\overline{AJ(A,\,B)} = 0\ , \tag{25.23}$$

$$\overline{BJ(A,\, B)} = 0\,. \tag{25.24}$$

Multiplying both sides of the advection equation (25.16) by q, we obtain

$$\begin{aligned} \frac{1}{2}\frac{\partial q^2}{\partial t} &= qJ(q,\, \psi) \\ &= J\left(\frac{1}{2}q^2,\, \psi\right)\,. \end{aligned} \tag{25.25}$$

Averaging over the entire domain, we find that

$$\begin{aligned} \overline{\frac{1}{2}\frac{\partial q^2}{\partial t}} &= \overline{J\left(\frac{1}{2}q^2,\, \psi\right)} \\ &= -\overline{\mathbf{v}\cdot\nabla\frac{1}{2}q^2} \\ &= -\overline{\nabla\cdot\left(\mathbf{v}\frac{1}{2}q^2\right)} = 0\,. \end{aligned} \tag{25.26}$$

25.4 Aliasing instability

Scale interactions are an essential aspect of aliasing errors. Aliasing is particularly likely when the "input" scales have high wave numbers, close to the truncation scale. It is possible for aliasing error to lead to a form of instability, which it seems reasonable to call "aliasing instability." It is more often called "non-linear" instability, but this is somewhat misleading because aliasing instability can also occur in the numerical integration of a linear equation with spatially variable coefficients. Aliasing instability is caused by a spurious growth of small-scale features of the flow, due in part to the aliasing error arising from the nonlinear interaction of the finite-difference analogs of *any* two spatially varying quantities. Aliasing instability has nothing to do with time-differencing, and can happen even in the time-continuous case. It is quite different from the linear computational instability discussed in earlier chapters.

25.4.1 An example of aliasing instability

We now work through a simple example that illustrates the possibility of aliasing instability. The example was invented by Phillips (1959a); see also (Lilly, 1965). We keep the time derivatives continuous.

We begin by writing down a differential-difference version of (25.16), on a plane, using a simple finite-difference approximation for the Jacobian. For simplicity, we take $\Delta x = \Delta y = d$. We start by investigating

$$\frac{dq_{i,\,j}}{dt} = \left[J_1\left(q,\ \psi\right)\right]_{i,\,j} \tag{25.27}$$

where J_1 is a particular finite-difference form of the Jacobian, given by

$$\left[J_1\left(q,\ \psi\right)\right]_{i,j} \equiv \frac{\left(q_{i+1,\,j} - q_{i-1,\,j}\right)\left(\psi_{i,\,j+1} - \psi_{i,\,j-1}\right) - \left(q_{i,\,j+1} - q_{i,\,j-1}\right)\left(\psi_{i+1,\,j} - \psi_{i-1,\,j}\right)}{4d^2} \,. \tag{25.28}$$

Note that J_1 is based on (25.19). Later we are going to discuss several other finite-difference approximations to the Jacobian. Combining (25.27) and (25.28), we obtain

$$\frac{dq_{i,\,j}}{dt} \equiv \frac{\left(q_{i+1,\,j} - q_{i-1,\,j}\right)\left(\psi_{i,\,j+1} - \psi_{i,\,j-1}\right) - \left(q_{i,\,j+1} - q_{i,\,j-1}\right)\left(\psi_{i+1,\,j} - \psi_{i-1,\,j}\right)}{4d^2} \,. \tag{25.29}$$

We *assume* that the solution, $q_{i,\,j}\left(t\right)$, has the form

$$q_{i,\,j}\left(t\right) = \left[C\left(t\right)\cos\left(\frac{\pi i}{2}\right) + S\left(t\right)\sin\left(\frac{\pi i}{2}\right)\right]\sin\left(\frac{2\pi j}{3}\right) \,. \tag{25.30}$$

This strange-looking assumption will be justified later. You should be able to show that the x-wavenumber, k, of the sine and cosine functions in the square brackets satisfies $kd = \pi/2 < \pi$, while the y-wavenumber, l, satisfies $ld = 2\pi/3 < \pi$. Both k and l correspond to wavelengths slightly longer than $2d$, which means that they are resolvable, but (25.30) describes a very noisy wind field. For all t, we *prescribe* the stream function $\psi_{i,\,j}$ as

$$\psi_{i,\,j} = \Psi\,\cos\left(\pi i\right)\sin\left(\frac{2\pi j}{3}\right) \,. \tag{25.31}$$

This is a time-independent but *spatially variable* advecting current. The advecting current was almost always prescribed in earlier chapters too, but until now it has been spatially uniform. Because $\psi_{i,\,j}$ is prescribed, the model under discussion here is linear, but with spatially variable coefficients. The forms of $q_{i,\,j}$ and $\psi_{i,\,j}$ given by (25.30) and (25.31), are plotted in Fig. 25.5. They are nasty functions.

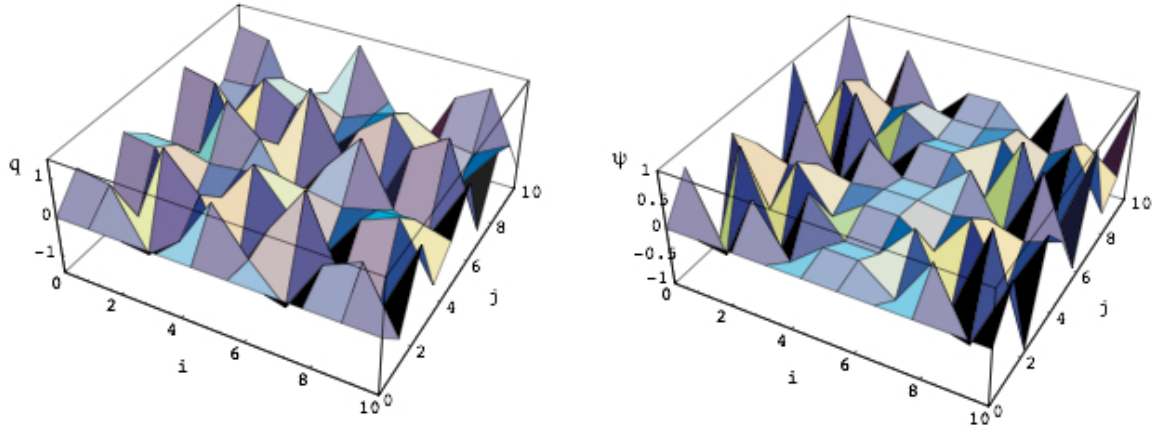


Figure 25.5: Plots of the functions $q_{i,\,j}(t=0)$ and $\psi_{i,\,j}$ given by (26.1) and (26.2), respectively. For plotting purposes, we have used $C = S = \Psi = 1$. The functions have been evaluated only for integer values of i and j, which gives them a jagged appearance. Nevertheless it is fair to say that they are rather ugly. This is the sort of thing that can appear in your simulations as a result of aliasing instability.

Because (25.31) says that $\psi_{i,\,j}$ has a wavelength of $2d$ in the x-direction, we can simplify (25.29) to

$$\frac{\partial q_{i,\,j}}{\partial t} = \frac{1}{4d^2}\left(q_{i+1,\,j} - q_{i-1,\,j}\right)\left(\psi_{i,\,j+1} - \psi_{i,\,j-1}\right) \ . \tag{25.32}$$

From (25.30), we can show that

$$q_{i+1,\,j} - q_{i-1,\,j} = 2\left(-C\sin\frac{\pi i}{2} + S\cos\frac{\pi i}{2}\right)\sin\left(\frac{2\pi j}{3}\right) \ . \tag{25.33}$$

Here we have used some trigonometric identities. Similarly, we can use some trig identities with (25.31) to show that

$$\begin{aligned}\psi_{i,\,j+1} - \psi_{i,\,j-1} &= \Psi \cos(\pi i)\, 2\cos\left(\frac{2\pi j}{3}\right) \sin\left(\frac{2\pi}{3}\right) \\ &= \sqrt{3}\, U \cos(\pi i) \cos\left(\frac{2\pi j}{3}\right) \,.\end{aligned} \tag{25.34}$$

As already mentioned, (25.34) holds for all t. The product of (25.33) and (25.34) gives the right-hand side of (25.32), which can be written, again using trigonometric identities, as

$$\frac{dq_{i,\,j}}{dt} = \frac{\sqrt{3}}{4d^2}\Psi \left[C\sin\left(\frac{\pi i}{2}\right) + S\cos\left(\frac{\pi i}{2}\right)\right] \sin\left(\frac{4\pi j}{3}\right) \,. \tag{25.35}$$

Now we observe that in (25.35) the wave number in the y-direction satisfies

$$ld = \frac{4\pi}{3} > \pi \,. \tag{25.36}$$

This means that the product on the right-hand side of (25.32) *has produced a wave number in the y-direction that is too short to be represented on the grid*. According to our earlier analysis, this wave will be interpreted by the grid as having the smaller wave number $l^* = -(2l_{\max} - l) = -\dfrac{2\pi}{3d}$. Therefore (25.35) can be re-written as

$$\frac{dq_{i,\,j}}{dt} = -\frac{\sqrt{3}\,\Psi}{4d^2}\left(C\sin\frac{\pi i}{2} + S\cos\frac{\pi i}{2}\right)\sin\frac{2\pi j}{3} \,. \tag{25.37}$$

Rewriting (25.35) as (25.37) is a key step in the analysis, because this is where aliasing enters. Because we are doing the problem algebraically, we have put in the aliasing "by hand."

Next, we observe that the spatial form of $dq_{i,\,j}/dt$, as given by (25.37), agrees with the assumed form of $q_{i,\,j}$, given by (25.30). This means that the shapes of the sine and cosine parts of $q_{i,\,j}$ do not change with time, *thus justifying our the assumed form of* (25.30)*, in which the only time-dependence is in* $C(t)$ *and* $S(t)$. In order to recognize this, we had to take into account that aliasing occurs.

If we now simply differentiate (25.30) with respect to time, and substitute the result into the left-hand side of (25.37), we find that

$$\frac{dC}{dt}\cos\frac{\pi i}{2}+\frac{dS}{dt}\sin\frac{\pi i}{2}=-\frac{\sqrt{3}}{4d^2}\Psi\left(C\sin\frac{\pi i}{2}+S\cos\pi\frac{i}{2}\right)\,. \tag{25.38}$$

Note that time derivatives of $C(t)$ and $S(t)$ appear on the left-hand side of (25.38). Using the linear independence of the sine and cosine functions, we find by inspection of (25.38) that

$$\frac{dC}{dt}=-\frac{\sqrt{3}}{4d^2}\Psi\,S,\quad\text{and}\quad\frac{dS}{dt}=-\frac{\sqrt{3}}{4d^2}\Psi\,C\,. \tag{25.39}$$

From (25.39), it follows that

$$\frac{d^2C}{dt^2}=\sigma^2C\quad\text{and}\quad\frac{d^2S}{dt^2}=\sigma^2S\,, \tag{25.40}$$

where

$$\sigma\equiv\frac{\sqrt{3}\,\Psi}{4d^2}\,. \tag{25.41}$$

According to (25.40), C and S will grow exponentially. This demonstrates that the finite-difference scheme is unstable. We note with horror that the growth rate actually increases as the grid spacing becomes finer, so the problem gets worse with higher resolution. The unstable modes will have the ugly form given by (25.30) and plotted in Fig. 25.5.

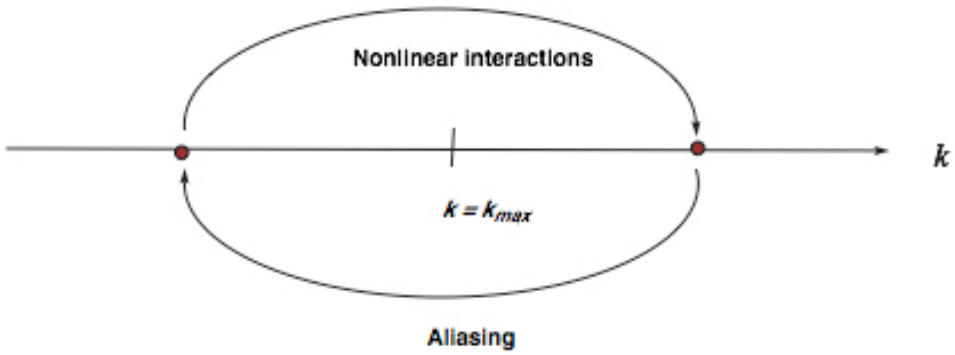


Figure 25.6: Sketch illustrating the mechanism of aliasing instability.

Fig. 25.6 summarizes the mechanism of aliasing instability. Nonlinear interactions feed energy into waves that cannot be represented on the grid. Aliasing causes this energy to "fold back" onto scales that do fit on the grid, but typically these are rather small scales that are not well resolved and suffer from large truncation errors. In the example given, the truncation errors lead to further production of energy on scales too small to be represented.

Note, however, that *if the numerical scheme conserved energy, the total amount of energy could not increase, and the instability would be prevented, even though aliasing would still occur, and even though the truncation errors for the smallest scales would still be large*. In the example discussed above, we used J_1. Later we demonstrate that J_1 does not conserve energy. As we will discuss, some other finite-difference Jacobians do conserve energy. Instability does not occur with those Jacobians.

25.4.2 Analysis in terms of discretization error

Further general insight into this type of instability can be obtained by investigating the discretization error of (25.29). This can be expressed as

$$\begin{aligned}\left(\frac{dq}{dt}\right)_{i,\,j} &= \left[J_1\left(q,\,\psi\right)\right]_{i,\,j} \\ &= \left[J\left(q,\,\psi\right)\right]_{i,\,j} + \frac{d^2}{6}\left[\frac{\partial q}{\partial x}\frac{\partial^3\psi}{\partial y^3} - \frac{\partial q}{\partial y}\frac{\partial^3\psi}{\partial x^3} + \frac{\partial^3 q}{\partial x^3}\frac{\partial\psi}{\partial y} - \frac{\partial^3 q}{\partial y^3}\frac{\partial\psi}{\partial x}\right]_{i,\,j} + \mathscr{O}\left(d^4\right)\,.\end{aligned} \tag{25.42}$$

Here $\left[J\left(q,\,\psi\right)\right]_{i,\,j}$ is the exact Jacobian at the point $(i,\,j)$. The second line of (25.42) is obtained by Taylor-series expansion, and we see that the smallest term in the error is proportional to d^2, so the leading term in the error is proportional to d^2. Multiplying (25.42) by q, integrating over the whole domain, and using (25.26), we find, after repeated integration by parts and a page or so of algebra, that the second-order part of the discretization error (i.e., the leading term in the discretization error) causes the square of q to change at the rate

$$\frac{1}{2}\frac{d}{dt}\int q^2\,ds = \frac{d^2}{4}\int \frac{\partial^2\psi}{\partial x\partial y}\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right]\,ds + \int \mathscr{O}\left(d^4\right)\,ds\,. \tag{25.43}$$

Note that $\dfrac{\partial^2\psi}{\partial x\partial y} = -\dfrac{\partial u}{\partial x} = \dfrac{\partial v}{\partial y}$. Eq. (25.43) means that, for $\frac{\partial^2\psi}{\partial x\partial y} > 0$, q^2 will falsely grow with time if $\left(\dfrac{\partial q}{\partial x}\right)^2$ is bigger than $\left(\dfrac{\partial q}{\partial y}\right)^2$, in an overall sense. In such a situation, instability will occur. The scheme will blow up *locally*, in the particular portions of the domain

where $\frac{\partial^2\psi}{\partial x\partial y} > 0$, and the growing modes will have $\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right] > 0$. Similarly, where $\frac{\partial^2\psi}{\partial x\partial y} < 0$, there will be growing modes with $\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right] < 0$.

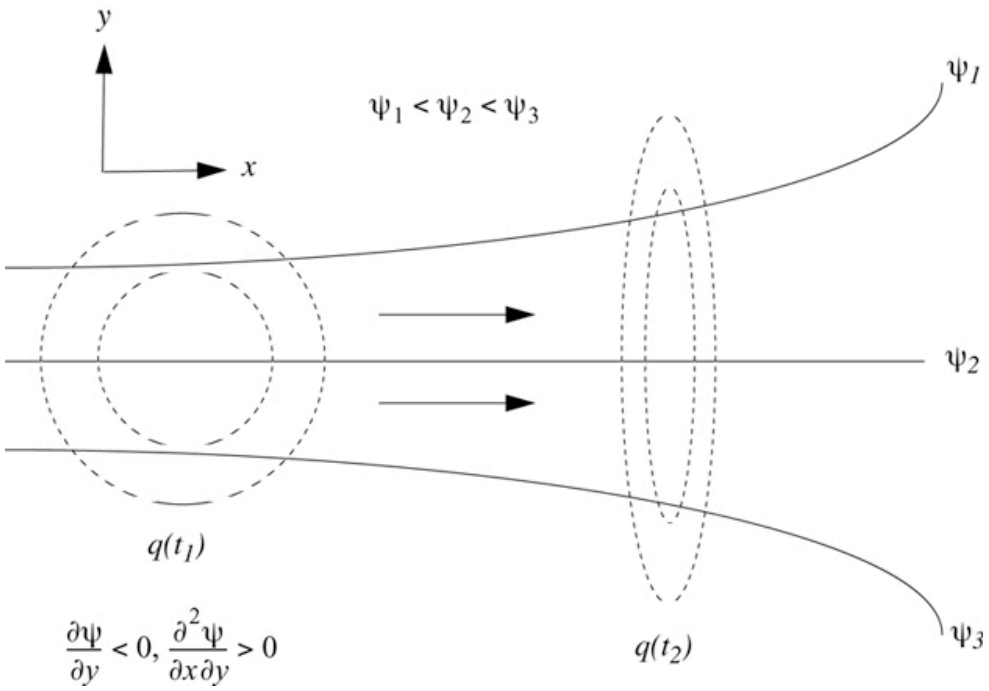


Figure 25.7: Schematic illustration of the mechanism of aliasing instability. The solid lines denote the stream function, which is independent of time. The implied mostly zonal flow is from left to right in the figure, and becomes weaker towards the right. The dashed lines show contours of the advected quantity q, at two different times, denoted by t_1 and $t_2 > t_1$. As q is carried towards the right, its contours are stretched in the y direction.

Now look at Fig. 25.7. In the figure, the streamlines are given such that $\psi_1 < \psi_2 < \psi_3$, so that $(\partial\psi/\partial y) < 0$, which implies westerly flow. The figure shows that the westerly flow is decreasing towards the east, as in the "exit" region of the jet stream, so that $\frac{\partial^2\psi}{\partial x\partial y} < 0$. In fact, the solution of the differential-difference equation tends to prefer a positive value of the integrand of the right-hand side of (25.43), as illustrated schematically in Fig. 25.7. Notice that at t_2, $\frac{\partial q}{\partial x}$ becomes greater than it was at t_1, and the reverse is true for $\frac{\partial q}{\partial y}$. The spatial distribution of q is being stretched out in the meridional direction. This is called "noodling." Although the expression $\int \frac{\partial^2\psi}{\partial x\partial y}\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right] ds$ vanishes at t_1, it has become positive at t_2. From (25.43), it can be seen that the area-integrated q^2 will then tend to increase with time, due to the discretization error.

In contrast to the linear computational instabilities discussed earlier in this course, *alias-*

ing instability has nothing to do with time truncation error. Making the time step shorter cannot prevent the instability, which can occur, in fact, even in the time-continuous case. The example we have just considered illustrates this fact, because we have left the time derivatives in continuous form.

25.5 Discussion

A number of methods have been proposed to prevent or control aliasing instability. One approach is to eliminate aliasing. As will be discussed in Chapter 13, aliasing error can actually be eliminated in a spectral model, at least for terms that involve only "quadratic" aliasing, i.e., aliasing that arises from the multiplication of two spatially varying fields. Aliasing instability can also be prevented without eliminating aliasing, however.

Phillips (1959a) suggested that aliasing instability can be avoided if a Fourier analysis of the predicted fields is made after each time step, and all waves of wave number $k > k_{\max}/2$ are simply discarded. With this "filtering" method, Phillips could guarantee absolutely no aliasing error due to quadratic nonlinearities, because the shortest possible wave would have wave number $k_{\max}/2$ (his maximum wave number), and thus any wave generated by quadratic nonlinearities would have a wave number of at most $k_{\max}$. The filter is strongly dissipative, however, because it removes variance.

Others have suggested that use of a dissipative advection scheme can suppress aliasing instability. Dissipative schemes have undesirable side effects, however. They damp or otherwise distort the solution.

A fourth way to eliminate aliasing instability is to use advection schemes that conserve the square of the advected quantity. This has the advantage that stability is ensured simply by mimicking a property of the exact equations. In particular, to prevent aliasing instability with the momentum equations, we can use finite-difference schemes that conserve either kinetic energy, or enstrophy (squared vorticity), or both. This approach was proposed by Arakawa (1966). It is discussed in Chapter 26.

25.6 Problems

1. A wagon wheel rotates at R revolutions per second. It is featureless except for a single dot painted near its outer edge. The wheel is filmed at F frames per second.

   (a) What inequality must F satisfy to avoid aliasing?

   (b) How does the *apparent* rotation rate, R^*, vary as a function of F and R? Assume $R > 0$ and $F > 0$.

Chapter 26

When the advector is the advectee

26.1 introduction

If the wind field is specified, as for example in the discussion of Chapter 7, then the advection of a tracer can be considered as a linear problem; it is, at least, linear in the tracer. With *momentum advection*, however, the wind field is both the "advector" and the "advectee." Momentum advection is thus unavoidably nonlinear. Up to now, we have mostly avoided the subject of momentum advection, except for a brief discussion in Chapter 18, which was limited to the one-dimensional case, without rotation. We now consider the advection terms of the momentum equation for the multi-dimensional case, including both Earth-rotation, f, and the relative vorticity, ζ, that is associated with the wind field. Vorticity is key to almost all atmospheric dynamics, on both large and small scales.

26.2 Fjortoft's Theorem

Ragnar Fjørtoft (1953), a Norwegian meteorologist, arrived at some very fundamental and famous insights regarding momentum advection in two-dimensional non-divergent flows. When the flow is non-divergent, so that (25.2) is satisfied, the vorticity equation, (25.3), reduces to

$$\frac{\partial}{\partial t}(\zeta + f) = -\mathbf{v} \cdot \nabla (\zeta + f) \ . \tag{26.1}$$

This says that the absolute vorticity is simply advected by the mean flow. We also see that only the sum $(\zeta + f)$ matters for the vorticity equation; henceforth we just replace $(\zeta + f)$ by ζ, for simplicity. The vorticity and the stream function are related by

$$\zeta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{v}) = \nabla^2 \psi \,. \tag{26.2}$$

(This relationship was used as an example of a Poisson equation, back in Chapter 15.) Eq. (26.1) can be rewritten as

$$\frac{\partial \zeta}{\partial t} = -\nabla \cdot (\mathbf{v}\zeta) \,, \tag{26.3}$$

or, alternatively, as

$$\frac{\partial \zeta}{\partial t} = J(\zeta,\, \psi) \,. \tag{26.4}$$

From (26.4) and (25.22) we see that the domain-averaged vorticity is conserved:

$$\frac{d\overline{\zeta}}{dt} = \overline{\frac{\partial \zeta}{\partial t}} = 0 \,. \tag{26.5}$$

By combining (26.4) and (25.23), we can show that

$$\overline{\zeta \frac{\partial \zeta}{\partial t}} = 0 \,, \tag{26.6}$$

from which it follows that the domain-average of the enstrophy is also conserved:

$$\frac{d}{dt}\left(\frac{1}{2}\overline{\zeta^2}\right) = 0 \,. \tag{26.7}$$

Similarly, from (26.4) and (25.24) we find that

$$\overline{\psi \frac{\partial \zeta}{\partial t}} = 0 \,. \tag{26.8}$$

To see what Eq. (26.8) implies, substitute (26.2) into (26.8), to obtain

$$\overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} = 0 \,. \tag{26.9}$$

Note, however, that

$$\begin{aligned} \overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} &= \overline{\psi \frac{\partial}{\partial t} \left[\nabla \cdot (\nabla \psi) \right]} \\ &= \overline{\psi \nabla \cdot \frac{\partial}{\partial t} \nabla \psi} \\ &= \overline{\nabla \cdot \left(\psi \frac{\partial}{\partial t} \nabla \psi \right)} - \overline{\nabla \psi \cdot \frac{\partial}{\partial t} \nabla \psi} \\ &= -\overline{\frac{\partial}{\partial t} \left(\frac{1}{2} |\nabla \psi|^2 \right)} \,. \end{aligned} \tag{26.10}$$

Taken together, Eqs. (26.10) and (26.9) demonstrate that *for two-dimensional nondivergent flow* the vorticity equation (26.4) implies kinetic energy conservation. The derivation in (26.10) goes through even if the $\partial / \partial t$ operators are erased on each line. This means that, for two-dimensional nondivergent flow,

$$\overline{K} = -\overline{\psi \zeta} \,. \tag{26.11}$$

This interesting result says that the domain average of the kinetic energy is equal to minus the domain average of the product of the stream function and the vorticity.

Since both kinetic energy and enstrophy are conserved in frictionless two-dimensional flows, their ratio is also conserved. It has the dimensions of a length squared:

$$\frac{\text{kinetic energy}}{\text{enstrophy}} \sim \frac{L^2 t^{-2}}{t^{-2}} = L^2 \,. \tag{26.12}$$

The length L can be interpreted as the diameter of the most energetic vortices, and (26.12) states that it is invariant. An implication is that energy does not cascade in frictionless two-dimensional flows; it "stays where it is" in wave-number space.

This argument is a little too simple. We now show that kinetic energy actually tends to move "upscale" in two-dimensional nondivergent flow. The exchanges of energy and enstrophy among different scales in two-dimensional turbulence were studied Fjørtoft (1953). Fjortoft's results can be summarized in a simplified way as follows. Consider three equally spaced wave numbers, as shown in Fig. (26.1). By "equally spaced," we mean that

$$\lambda_2 - \lambda_1 = \lambda_3 - \lambda_2 = \Delta\lambda \ . \tag{26.13}$$

The enstrophy, E, is

$$E = E_1 + E_2 + E_3 \ , \tag{26.14}$$

and the kinetic energy is

$$K = K_1 + K_2 + K_3 \ . \tag{26.15}$$

It can be shown that

$$E_n = \lambda_n^2 K_n, \tag{26.16}$$

where λ_n is a wave number, and the subscript n denotes a particular Fourier component.

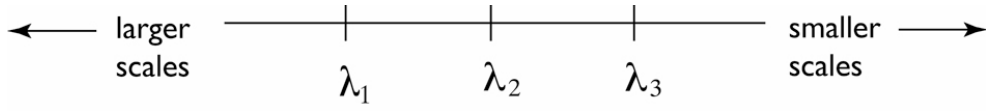


Figure 26.1: Diagram used in the explanation of Fjørtoft (1953) analysis of the exchanges of energy and enstrophy among differing scales in two-dimensional motion

Suppose that nonlinear processes redistribute kinetic energy among the three wave numbers, i.e.,

$$K_n \to K_n + \delta K_n \ , \tag{26.17}$$

while conserving both kinetic energy and enstrophy so that

$$\sum \delta K_n = 0\,, \tag{26.18}$$

and

$$\sum \delta E_n = 0\,. \tag{26.19}$$

From (26.16), we see that

$$\delta E_n = \lambda_n^2 \delta K_n. \tag{26.20}$$

It follows from (26.18) that

$$\delta K_1 + \delta K_3 = -\delta K_2, \tag{26.21}$$

and from (26.19) and (26.20) that

$$\begin{aligned} \lambda_1^2 \delta K_1 + \lambda_3^2 \delta K_3 &= -\lambda_2^2 \delta K_2 \\ &= \lambda_2^2 \left(\delta K_1 + \delta K_3\right). \end{aligned} \tag{26.22}$$

Collecting terms in (26.22), we find that

$$\frac{\delta K_3}{\delta K_1} = \frac{\lambda_2^2 - \lambda_1^2}{\lambda_3^2 - \lambda_2^2} > 0. \tag{26.23}$$

The fact that $\delta K_3/\delta K_1$ is positive means that either both ΔK_3 and ΔK_1 are positive, or both are negative. Using (26.13), Eq. (26.23) can be simplified to

$$0 < \frac{\delta K_3}{\delta K_1} = \frac{\lambda_2 + \lambda_1}{\lambda_3 + \lambda_2} < 1. \tag{26.24}$$

Eq. (26.24) shows that the energy transferred to or from higher wave numbers (ΔK_3) is less than the energy transferred to or from lower wave numbers (ΔK_1). This conclusion is based on both (26.18) and (26.19), i.e., on both energy conservation and enstrophy conservation. The implication is that kinetic energy actually "migrates" from higher wave numbers to lower wave numbers, i.e., from smaller scales to larger scales.

We now perform a similar analysis for the enstrophy. As a first step, we use (26.20) and (26.24) to write

$$\begin{aligned} \frac{\delta E_3}{\delta E_1} &= \frac{\lambda_3^2}{\lambda_1^2}\left(\frac{\lambda_2+\lambda_1}{\lambda_3+\lambda_2}\right) \\ &= \frac{(\lambda_2+\delta\lambda)^2}{(\lambda_2-\delta\lambda)^2}\frac{\left(\lambda_2-\frac{1}{2}\delta\lambda\right)}{\left(\lambda_2+\frac{1}{2}\delta\lambda\right)} > 1. \end{aligned} \tag{26.25}$$

To show that this ratio is actually greater than one, as indicated above, we demonstrate that $\delta E_3/\delta E_1 = a\cdot b\cdot c$, where a, b, and c are each greater than one. We can choose:

$$a = \frac{\lambda_2+\Delta\lambda}{\lambda_2-\Delta\lambda} > 1, \quad b = \frac{\lambda_2-\frac{1}{2}\Delta\lambda}{\lambda_2-\Delta\lambda} > 1, \quad \text{and} \quad c = \frac{\lambda_2+\Delta\lambda}{\lambda_2+\frac{1}{2}\Delta\lambda} > 1\,. \tag{26.26}$$

The conclusion is that enstrophy cascades to higher wave numbers in two-dimensional turbulence. Of course, such a cascade ultimately leads to enstrophy dissipation by viscosity.

When viscosity acts on two-dimensional turbulence, enstrophy is dissipated but kinetic energy is (almost) unaffected. Then the denominator of (26.12) decreases with time, while the numerator remains nearly constant. It follows that the length scale, L, will tend to increase with time. This means that the most energetic vortices will become larger. This is an "anti-cascade" of kinetic energy. The implication is that two-dimensional turbulence tends to remain smooth, so that the kinetic energy of the atmosphere tends to remain on large, quasi-two-dimensional scales, instead of cascading down to small scales where it can be dissipated.

In three dimensions, vorticity is not conserved because of stretching and twisting, and enstrophy is not conserved because of stretching (although it is unaffected by twisting). Vortex stretching causes small scales to gain energy at the expense of larger scales. As a result, kinetic energy cascades in three-dimensional turbulence. Ultimately the energy is converted (dissipated) from kinetic to internal by viscosity. This is relevant to mesoscale and small-scale atmospheric circulations, such as turbulent eddies and convective cells.

In summary, advection and rotation have no effect on the domain-averaged vorticity, enstrophy, or kinetic energy in two-dimensional flows. Because two-dimensional flows conserve both energy and enstrophy, they "have fewer options" than three-dimensional flows. In particular, a kinetic energy cascade cannot happen in two dimensions. What happens instead is an enstrophy cascade. Enstrophy is dissipated but kinetic energy is (almost) not.

Because kinetic energy does not cascade in two-dimensional flows, the motion remains smooth and is dominated by "large" eddies. This is true with the continuous equations, and we want it to be true in our models as well.

26.3 Kinetic energy and enstrophy conservation in two-dimensional non-divergent flow

We have shown that both kinetic energy and enstrophy are conserved in two-dimensional non-divergent flows, and that this has profound consequences for the spectrum of kinetic energy. Arakawa (1966) developed a method for numerical simulation of two-dimensional, purely rotational motion, that conserves both kinetic energy and enstrophy, as well as vorticity. His method has been and still is very widely used, and is explained below.

26.3.1 The vorticity equation

We begin by writing down a spatially discrete version of (26.18), keeping the time derivative in continuous form:

$$\begin{aligned} \sigma_{\boldsymbol{i}} \frac{d\zeta_{\boldsymbol{i}}}{dt} &= \sigma_{\boldsymbol{i}} J_{\boldsymbol{i}}(\zeta,\ \psi) \\ &= \sum_{\boldsymbol{i}'} \sum_{\boldsymbol{i}''} c_{\boldsymbol{i},\,\boldsymbol{i}',\,\boldsymbol{i}''} \zeta_{\boldsymbol{i}'} \psi_{\boldsymbol{i}''} \ . \end{aligned} \tag{26.27}$$

The bold subscripts are two-dimensional counters that can be used to specify a grid cell on a two-dimensional grid by giving a single number, as was done already in Chapter 3. The area of grid cell $\boldsymbol{i}$ is denoted by $\sigma_{\boldsymbol{i}}$. The $c_{\boldsymbol{i},\,\boldsymbol{i}',\,\boldsymbol{i}''}$ are coefficients that must be specified to define a finite-difference scheme, following the approach that we first used in Chapter 3. For later reference, with double subscripts, (26.27) would become

$$\sigma_{i,\,j} \frac{d\zeta_{i,\,j}}{dt} = \sum_{j'} \sum_{i'} \sum_{j''} \sum_{i''} \left(c_{i,\,j;i',\,j';i'',j''} \right) \zeta_{i',j'} \psi_{i'',j''}. \tag{26.28}$$

The second line of (26.27) looks a little mysterious and requires some explanation. As can be seen by inspection of (25.17), the Jacobian operator, $J(\zeta,\ \psi)$, involves derivatives of the vorticity, multiplied by derivatives of the stream function. We can anticipate, therefore, that the finite-difference form of the Jacobian at the point $\boldsymbol{i}$ will involve products of the vorticity at some nearby grid points with the stream function at other nearby grid points. We have already seen an example of this in (25.28). Such products appear in (26.27). The neighboring grid points can be specified in (26.27) by assigning appropriate values to $\boldsymbol{i'}$ and $\boldsymbol{i''}$. As you can see from the subscripts, $\boldsymbol{i'}$ picks up vorticity points, and $\boldsymbol{i''}$ picks up stream-function points. The $c_{\boldsymbol{i},\,\boldsymbol{i'},\,\boldsymbol{i''}}$ are "interaction coefficients;" their form will be chosen later. By making appropriate choices of the $c_{\boldsymbol{i},\,\boldsymbol{i'},\,\boldsymbol{i''}}$ we can construct an approximation to the Jacobian. The double sum in (26.27) essentially picks out the combinations of ζ and ψ, at neighboring grid points, that we wish to bring into our finite-difference operator. This is similar to the notation that we used in Chapter 3, but a bit more complicated.

Of course, there is nothing about the form of (26.27) that shows that it is actually a consistent finite-difference approximation to the Jacobian operator; all we can say at this point is that (26.27) has the *potential* to be a consistent finite-difference approximation to the Jacobian, if we choose the interaction coefficients properly. The coefficients can be chosen to give any desired order of accuracy (in the Taylor series sense), using the methods discussed in Chapter 3.

The form of (26.27) is sufficiently general that it is impossible to tell what kind of grid is being used. It could be a rectangular grid on a plane, or a latitude-longitude grid on the sphere, or something more exotic like a geodesic grid on the sphere (to be discussed in Chapter 28).

As an example, consider the finite-difference Jacobian J_1, introduced in Eq. (25.28). Applying J_1 to vorticity advection on a square grid with grid spacing d, we can write, corresponding to (26.28),

$$\begin{aligned} d^2\frac{d\zeta_{i,\,j}}{dt} &= d^2\left\{\frac{1}{4d^2}\left[(\zeta_{i+1,\,j}-\zeta_{i-1,\,j})(\psi_{i,\,j+1}-\psi_{i,\,j-1})-(\zeta_{i,\,j+1}-\zeta_{i,\,j-1})(\psi_{i+1,\,j}-\psi_{i-1,\,j})\right]\right\} \\ &= \frac{1}{4}\left(\zeta_{i+1,\,j}\psi_{i,\,j+1}-\zeta_{i+1,\,j}\psi_{i,\,j-1}-\zeta_{i-1,\,j}\psi_{i,\,j+1}+\zeta_{i-1,\,j}\psi_{i,\,j-1}-\zeta_{i,\,j+1}\psi_{i+1,\,j}+\zeta_{i,\,j+1}\psi_{i-1,\,j}\right. \\ &\quad \left.+\zeta_{i,\,j-1}\psi_{i+1,\,j}-\zeta_{i,\,j-1}\psi_{i-1,\,j}\right). \end{aligned} \tag{26.29}$$

By inspection of the second line of (26.29), and comparing with (26.28), we see that for J_1 each value of $c_{i,\,j;i',j';i'',j''}$ is either $+1/4$ or $-1/4$. The eight non-zero values of c, for J_1, can be listed as follows:

$$
\begin{aligned}
c_{i,\,j;\,i+1,\,j;\,i,\,j+1} &= +\frac{1}{4},\\
c_{i,\,j;\,i+1,\,j;\,i,\,j-1} &= -\frac{1}{4},\\
c_{i,\,j;\,i-1,\,j;\,i,\,j+1} &= -\frac{1}{4},\\
c_{i,\,j;\,i-1,\,j;\,i,\,j-1} &= +\frac{1}{4},\\
c_{i,\,j;\,i,\,j+1;\,i+1,\,j} &= -\frac{1}{4},\\
c_{i,\,j;\,i,\,j+1;\,i-1,\,j} &= +\frac{1}{4},\\
c_{i,\,j;\,i,\,j-1;\,i+1,\,j} &= +\frac{1}{4},\\
c_{i,\,j;\,i,\,j-1;\,i-1,\,j} &= -\frac{1}{4}\,.
\end{aligned}
\tag{26.30}
$$

Look carefully at the subscripts. As an example, you should be able to see that $c_{i,\,j;\,i+1,\,j;\,i,\,j+1}$ specifies the contribution of the vorticity east of the point $(i,\ j)$ combined with the stream function north of the point $(i,\ j)$ to the time-rate-of-change of the vorticity at the point $(i,\ j)$. See Fig. 26.2. Each coefficient involves three (not necessarily distinct) points. With the uniform square grid on a plane, the forms of the coefficients are very simple, as seen in (26.30). The same methods can be applied to very different cases, however, such as nonuniform grids on a sphere.

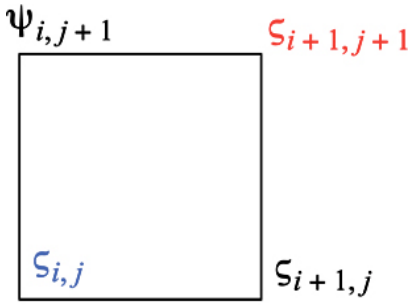


Figure 26.2: Stencil used in the discussion of vorticity conservation for J_1. See text for details.

26.3.2 Three basic requirements

Any finite-difference Jacobian should give zero if both of the input fields are spatially constant. From (26.27), this implies that

$$\boxed{0 = \sum_{i'} \sum_{i''} c_{i,\,i',\,i''} \quad \text{for all} \quad i} \,, \tag{26.31}$$

i.e., the sum of the coefficients is zero for all i. We *require* that our scheme satisfies (26.31). This requirement would emerge automatically if we enforced, for example, second-order accuracy of the Jacobian. You can easily confirm that J_1 satisfies (26.31).

Similarly, any finite-difference Jacobian should have the property that if the vorticity is spatially uniform, it will remain so for all time. From (26.27), this requirement takes the form

$$\boxed{0 = \sum_{i'} \sum_{i''} c_{i,\,i',\,i''} \psi_{i''} \quad \text{for all} \quad i} \,. \tag{26.32}$$

Eq. (26.32) can be interpreted as the condition that the motion is non-divergent, i.e., $\nabla \cdot (\mathbf{k} \times \nabla \psi) = -\mathbf{k} \cdot (\nabla \times \nabla \psi) = 0$. *Note that* (26.32) *must be true regardless of how the stream function varies in space*. This is only possible if each grid-point value of $\psi_{i''}$ appears more than once (in other words , at least twice) in the sum. Then we can arrange that the "total coefficient" multiplying $\psi_{i''}$, i.e., the sum of the two or more $c_{i,\,i',\,i''}$'s that multiply $\psi_{i''}$, is zero. In that case, the actual values of $\psi_{i''}$ have no effect on the sum in (26.32). You should confirm that J_1 satisfies (26.32).

In order to ensure conservation of the domain-averaged vorticity under advection, we must require that

$$\boxed{0 = \sum_{i} \sum_{i'} \sum_{i''} c_{i,\,i',\,i''} \zeta_{i'} \psi_{i''}} \,. \tag{26.33}$$

Here we have a *triple* sum, because we are taking a spatial average. *In this way, Eq.* (26.33) *is different in kind from* (26.31) *and* (26.32). Eq. (26.33) has to be true *regardless of how the vorticity and stream function vary in space*. This is only possible if each product $\zeta_{i'} \psi_{i''}$ appears more than once in the sum, such that the sum of the two or more $c_{i,\,i',\,i''}$s, that multiply each $\zeta_{i'} \psi_{i''}$ pair, is zero. In that case, the actual values of $\zeta_{i'} \psi_{i''}$ will have no effect on the triple sum in (26.33), because in the end they are multiplied by zero.

We now demonstrate that J_1 satisfies (26.33), i.e., it conserves vorticity. As pointed out above, $c_{i,\,j;\,i+1,\,j;\,i,\,j+1}$ specifies the contributions of the vorticity at $(i+1,\ j)$ and the stream function at $(i,\ j+1)$ to the time-rate-of-change of the vorticity at the point $(i,\ j)$. See Fig. 26.2. Similarly, $c_{i+1,\,j+1;\,i+1,\,j;\,i,\,j+1}$ specifies the contributions of the vorticity at $(i+1,\ j)$

and the stream function at $(i,\ j+1)$ to the time-rate-of-change of the vorticity at the point $(i+1,\ j+1)$. *The point here is that the vorticity at* $(i+1,\ j)$ *and the stream function at* $(i,\ j+1)$ *are "paired up" twice, once to predict the vorticity at* $(i,\ j)$ *and a second time to predict the vorticity at the point* $(i+1,\ j+1)$. Therefore, when we sum over the domain, this pair will appear twice, each time with a coefficient, c. Cancellation will occur if the two coefficients sum to zero, or in other words if they satisfy

$$c_{i,\,j;\,i+1,\,j;\,i,\,j+1} = -c_{i+1,\,j+1;\,i+1,\,j;\,i,\,j+1}\ . \tag{26.34}$$

How can we use (26.34) to choose the forms of the coefficients? *The key is that we use the same scheme for all points on the grid.* We can "shift" the stencil for the scheme from one grid cell to another by adding any (positive or negative) integer to all i subscripts for each coefficient, and adding a (generally different, positive or negative) integer to all j subscripts, without changing the numerical values of the coefficients. For example, the value of $c_{i+1,\,j+1;\,i+1,\,j;\,i,\,j+1}$, which appears on the right-hand side of (26.34), has to remain unchanged if we subtract one from each i subscript and one from each j subscript, thus shifting left and down. In other words, it must be true that

$$c_{i+1,\,j+1;\,i+1,\,j;\,i,\,j+1} = c_{i,\,j;\,i,\,j-1;\,i-1,\,j}\ . \tag{26.35}$$

Eq. (26.35) allows us to rewrite (26.34) as

$$\boxed{c_{i,\,j;\,i+1,\,j;\,i,\,j+1} = -c_{i,\,j;\,i,\,j-1;\,i-1,\,j}}\ . \tag{26.36}$$

What has been accomplished here is that, in (26.36), *both of the coefficients are associated with the time-rate of change of the vorticity at the same point, i.e.,* $(i,\ j)$, and so both of them are listed in (26.30). The meaning of (26.36) is that the "right, up" coefficient is equal to minus the "down, left" coefficient. Inspection of (26.30) shows that (26.36) is indeed satisfied. Similar results apply for the remaining terms. In this way, it can be demonstrated that J_1 conserves vorticity.

26.3.3 Conservation of enstrophy and kinetic energy

What we are going to do now is find a way to enforce finite-difference analogs of (26.6) and (26.8), namely:

$$\boxed{\begin{aligned} 0 &= \sum_i \sigma_i \zeta_i J_i(\zeta,\ \psi) \\ &= \sum_i \zeta_i \left(\sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i'} \psi_{i''} \right) \\ &= \sum_i \left(\sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_i \zeta_{i'} \psi_{i''} \right) , \end{aligned}} \tag{26.37}$$

$$\boxed{\begin{aligned} 0 &= \sum_i \sigma_i \psi_i J_i(\zeta,\ \psi) \\ &= \sum_i \psi_i \left(\sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i'} \psi_{i''} \right) \\ &= \sum_i \left(\sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i'} \psi_i \psi_{i''} \right) . \end{aligned}} \tag{26.38}$$

By enforcing these two requirements, we can ensure conservation of enstrophy and kinetic energy in the finite-difference model (although to fully ensure kinetic energy conservation we also need to attend to one additional issue, discussed later). We demonstrate below that the requirements (26.37) and (26.38) can be satisfied by suitable choices of the interaction coefficients. The requirements look daunting, though, because they involve triple sums. How in the world are we ever going to figure this out?

Inspection of (26.37) shows that the individual terms of the triple sum are going to involve products of vorticities at pairs of grid points. With this in mind, we go back to (26.27) and rewrite the scheme as

$$\begin{aligned} \sigma_i J_i(\zeta,\ \psi) &= \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i'} \psi_{i''} \\ &= \sum_{i'} a_{i,i'} \zeta_{i'} , \end{aligned} \tag{26.39}$$

where, by definition,

$$a_{i,i'} \equiv \sum_{i''} c_{i,i',i''} \psi_{i''} . \tag{26.40}$$

Here $a_{i,\,i'}$ is a weighted sum of stream functions. It does not have a $\boldsymbol{i''}$ subscript. This simplifies things because, in going from the first line of (26.39) to the second line, we replace a double sum involving three subscripts with a single sum involving two subscripts. The c coefficients don't appear on the second line of (26.39), because they are buried inside $a_{i,\,i'}$.

We can now write (26.39) times ζ_i as

$$\sigma_i \zeta_i J_i\left(\zeta,\,\psi\right) = \sum_{i'} a_{i,\,i'} \zeta_i \zeta_{i'} \,. \tag{26.41}$$

Here we have simply taken ζ_i inside the sum, which we can do because the sum is over $\boldsymbol{i'}$, not $\boldsymbol{i}$. From this point it is straightforward to enforce (26.37), which can be rewritten using our new notation as

$$0 = \sum_{i} \left(\sum_{i'} a_{i,\,i'} \zeta_i \zeta_{i'} \right) \,. \tag{26.42}$$

The value of $a_{i,\,i'}$ measures the influence of $\zeta_{i'}$ on the time-rate-of-change of the enstrophy at the point $\boldsymbol{i}$. Similarly, the value of $a_{i',\,i}$ measures the influence of ζ_i on the time-rate-of-change of the enstrophy at the point $\boldsymbol{i'}$. If these effects are equal and opposite, there will be no effect on the total enstrophy. In other words, we can achieve enstrophy conservation by enforcing

$$\boxed{a_{i,\,i'} = -a_{i',\,i} \quad \text{for all} \quad \boldsymbol{i} \quad \text{and} \quad \boldsymbol{i'} \,.} \tag{26.43}$$

This is a symmetry condition; it means that if we exchange the order of the subscript pairs, and also flip the sign, there is no net effect on the value of the coefficient a. As a special case, Eq. (26.43) implies that

$$a_{i,i} = 0 \quad \text{for all} \quad \boldsymbol{i} \,. \tag{26.44}$$

This means that the stream function at the point $\boldsymbol{i}$ has no effect on the time-rate-of-change of the enstrophy at point $\boldsymbol{i}$.

With the definition (26.40), we can rewrite the non-divergence condition (26.32) as

$$0 = \sum_{i'} a_{i,\,i'} \quad \text{for all} \quad \boldsymbol{i}\,. \tag{26.45}$$

Any scheme that satisfies (26.43) and (26.44) will also satisfy (26.45). In other words , any enstrophy-conserving scheme satisfies the non-divergence condition "automatically."

Kinetic energy conservation can be achieved using a very similar approach. We rewrite (26.27) as

$$\begin{aligned} \sigma_i J_i\left(\zeta,\ \psi\right) &= \sum_{i'} \sum_{i''} c_{i,\,i',\,i''}\, \zeta_{i'}\, \psi_{i''} \\ &= \sum_{i''} b_{i,\,i''}\, \psi_{i''}\ , \end{aligned} \tag{26.46}$$

where

$$b_{i,\,i''} \equiv \sum_{i'} c_{i,\,i',\,i''}\, \zeta_{i'} \tag{26.47}$$

is a weighted sum of vorticities that does not have a subscript $\boldsymbol{i'}$. The requirement for kinetic energy conservation, (26.38), can then be written as

$$0 = \sum_{i} \left(\sum_{i''} b_{i,\,i''}\, \psi_i\, \psi_{i''} \right)\ , \tag{26.48}$$

which is analogous to (26.42). Kinetic energy conservation can be achieved by requiring that

$$\boxed{b_{i,\,i''} = -b_{i'',i} \quad \text{all} \quad \boldsymbol{i} \quad \text{and} \quad \boldsymbol{i''}\,,} \tag{26.49}$$

which is analogous to (26.43).

The results obtained above are very general; they apply on an arbitrary grid, and on a two-dimensional domain of arbitrary shape. For instance, the domain could be a sphere.

26.3.4 One more thing

As mentioned above, there is one more thing to check, with respect to kinetic energy conservation. We have to make sure that the finite-difference analog of (26.10) holds, i.e.,

$$\sum_i \left(\sigma_i \psi_i \frac{d\zeta_i}{dt} \right) = -\sum_i \left[\sigma_i \frac{d}{dt} \left(\frac{1}{2} |\nabla \psi|_i^2 \right) \right] , \tag{26.50}$$

so that we can mimic, with the finite-difference equations, the demonstration of kinetic energy conservation that we performed with the continuous equations. In order to pursue this objective, we have to define a finite-difference Laplacian. We will consider a square grid with the "+" stencil for the Laplacian, discussed in Chapter 3, and grid spacing d:

$$\begin{aligned} \zeta_{i,\,j} &= \left(\nabla^2 \psi \right)_{i,\,j} \\ &\equiv \frac{1}{d^2} \left(\psi_{i+1,\,j} + \psi_{i-1,\,j} + \psi_{i,\,j+1} + \psi_{i,\,j-1} - 4\psi_{i,\,j} \right) . \end{aligned} \tag{26.51}$$

Here we have reverted to a conventional double-subscripting scheme, for clarity. We also define a finite-difference kinetic energy by

$$\begin{aligned} K_{i,\,j} &\equiv \frac{1}{2} |\nabla \psi|_{i,\,j}^2 \\ &\equiv \frac{1}{4d^2} \left[\left(\psi_{i+1,\,j} - \psi_{i,\,j} \right)^2 + \left(\psi_{i,\,j+1} - \psi_{i,\,j} \right) 2 + \left(\psi_{i,\,j} - \psi_{i-1,\,j} \right)^2 + \left(\psi_{i,\,j} - \psi_{i,\,j-1} \right)^2 \right] . \end{aligned} \tag{26.52}$$

Here we have used the definition of the gradient operator given by (A.6). Because the right-hand side of (26.52) is a sum of squares, we are guaranteed that the kinetic energy is non-negative. It should be clear that $K_{i,\,j}$ is defined in the same place as the vorticity and stream function. (This is compatible with the C-grid staggering, although we are not working with a staggered grid at the moment.) With the use of (26.51) and (26.52), it can be demonstrated, after a little algebra, that (26.50) is actually satisfied.

26.3.5 Do we have a Jacobian?

This is all fine, as far as it goes, but we still have some very basic and important business to attend to: We have not yet ensured that the sum in (26.27) is actually a consistent finite-difference approximation to the Jacobian operator. The approach that we will follow is to write down three *independent* finite-difference Jacobians and then identify, by inspection,

the c's in (26.27). When we say that the Jacobians are "independent," we mean that it is not possible to write any one of the three as a linear combination of the other two. The three finite-difference Jacobians are:

$$(J_1)_{i,\,j} = \frac{1}{4d^2}\left[\left(\zeta_{i+1,\,j} - \zeta_{i-1,\,j}\right)\left(\psi_{i,\,j+1} - \psi_{i,\,j-1}\right) - \left(\zeta_{i,\,j+1} - \zeta_{i,\,j-1}\right)\left(\psi_{i+1,\,j} - \psi_{i-1,\,j}\right)\right] , \tag{26.53}$$

$$\begin{aligned}(J_2)_{i,\,j} = \frac{1}{4d^2}\big[&-\left(\zeta_{i+1,\,j+1} - \zeta_{i+1,\,j-1}\right)\psi_{i+1,\,j} + \left(\zeta_{i-1,\,j+1} - \zeta_{i-1,\,j-1}\right)\psi_{i-1,\,j} \\ &+ \left(\zeta_{i+1,\,j+1} - \zeta_{i-1,\,j+1}\right)\psi_{i,\,j+1} - \left(\zeta_{i+1,\,j-1} - \zeta_{i-1,\,j-1}\right)\psi_{i,\,j-1}\big] ,\end{aligned} \tag{26.54}$$

$$\begin{aligned}(J_3)_{i,\,j} = \frac{1}{4d^2}\big[&\zeta_{i+1,\,j}\left(\psi_{i+1,\,j+1} - \psi_{i+1,\,j-1}\right) - \zeta_{i-1,\,j}\left(\psi_{i-1,\,j+1} - \psi_{i-1,\,j-1}\right) \\ &-\zeta_{i,\,j+1}\left(\psi_{i+1,\,j+1} - \psi_{i-1,\,j+1}\right) + \zeta_{i,\,j-1}\left(\psi_{i+1,\,j-1} - \psi_{i-1,\,j-1}\right)\big] .\end{aligned} \tag{26.55}$$

These can be interpreted as finite-difference analogs to the right-hand sides of (25.19) - (25.21), respectively. We can show that all three of these finite-difference Jacobians vanish if either of the input fields is spatially constant, and that all three conserve vorticity, i.e., they all satisfy (26.33).

What we need to do next is identify ("by inspection") the coefficients a and b for each of (26.53) - (26.55), and then check each scheme to see whether the requirements (26.43) and (26.49) are satisfied. In order to understand more clearly what these requirements actually mean, look at Fig. 26.3. The Jacobians J_1, J_2, and J_3 are represented in the top row of the figure. The colored lines show how each Jacobian at the point $(i,\,j)$ is influenced (or not) by the stream function and vorticity at the various neighboring points. We can interpret that $a_{\boldsymbol{i},\,\boldsymbol{i'}}$ denotes ζ-interactions of point $\boldsymbol{i}$ with point $\boldsymbol{i'}$, while $a_{\boldsymbol{i'},\,\boldsymbol{i}}$ denotes ζ-interactions of point $\boldsymbol{i'}$ with point $\boldsymbol{i}$. When we compare $a_{\boldsymbol{i},\,\boldsymbol{i'}}$ with $a_{\boldsymbol{i'},\,\boldsymbol{i}}$, it is like peering along one of the red (or purple) lines in Fig. 26.3, first outward from the point $(i,\,j)$, to one of the other points, and then back toward the point $(i,\,j)$. The condition (26.43) on the as essentially means that all such interactions are "equal and opposite," thus allowing algebraic cancellations to occur when we sum over all points. The condition (26.49) on the bs has a similar interpretation.

To check whether $(J_1)_i$ conserves enstrophy, we begin by rewriting (26.53) as

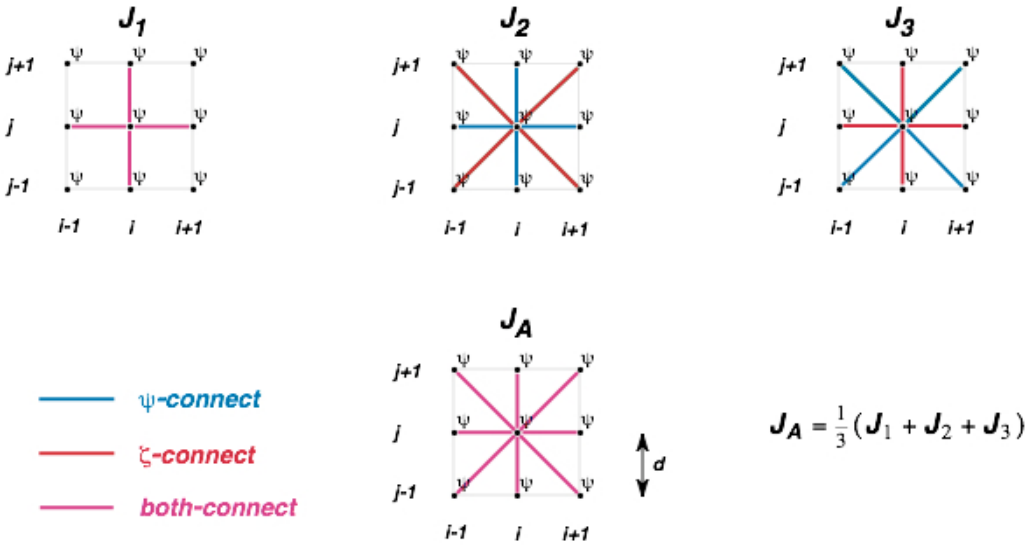


Figure 26.3: The central point in each figure is $(i,\ j)$. Stream function and vorticity are both defined at each of the mesh points indicated by the black dots. The colored lines represent contributions to $J_{i,\,j}$ from ψ, ζ, or both, from the various neighboring points. For J_1 and J_A, the red and blue lines overlap, so you only see the pink lines.

$$\begin{aligned}&\sigma_{i,\,j}\,(J_1)_{i,\,j}\,(\zeta,\ \psi) = \\ &\frac{1}{4}\left[\zeta_{i+1,\,j}\left(\psi_{i,\,j+1}-\psi_{i,\,j-1}\right)-\zeta_{i-1,\,j}\left(\psi_{i,\,j+1}-\psi_{i,\,j-1}\right)\right. \\ &\left.-\zeta_{i,\,j+1}\left(\psi_{i+1,\,j}-\psi_{i-1,\,j}\right)+\zeta_{i,\,j-1}\left(\psi_{i+1,\,j}-\psi_{i-1,\,j}\right)\right]\ .\end{aligned} \tag{26.56}$$

Here we have collected the coefficients of the four distinct values of the vorticity. By inspection of (26.56) and comparison with (26.40), we can read off the expressions for the $a_{i,\,j;\,i',\,j'}$:

$$a_{i,\,j;\,i+1,\,j} = \frac{1}{4}\left(\psi_{i,\,j+1}-\psi_{i,\,j-1}\right)\ , \tag{26.57}$$

$$a_{i,\,j;\,i-1,\,j} = -\frac{1}{4}\left(\psi_{i,\,j+1}-\psi_{i,\,j-1}\right)\ , \tag{26.58}$$

$$a_{i,\,j;\,i,\,j+1} = -\frac{1}{4}\left(\psi_{i+1,\,j}-\psi_{i-1,\,j}\right)\ , \tag{26.59}$$

$$a_{i,\,j;\,i,\,j-1} = \frac{1}{4}\left(\psi_{i+1,\,j} - \psi_{i-1,\,j}\right) \ . \tag{26.60}$$

Are these consistent with (26.43)? As a first step towards checking this, add one to each i subscript in (26.58); this gives:

$$a_{i+1,\,j;\,i,\,j} = -\frac{1}{4}\left(\psi_{i+1,\,j+1} - \psi_{i+1,\,j-1}\right) \ . \tag{26.61}$$

Now simply compare (26.61) with (26.57), to see that the requirement (26.43) is *not* satisfied by J_1. We conclude that J_1 does not conserve enstrophy.

In this way, we can reach the following conclusions:

- J_1 conserves neither enstrophy nor kinetic energy;
- J_2 conserves enstrophy but not kinetic energy; and
- J_3 conserves kinetic energy but not enstrophy.

It looks like we are out of luck.

We can form a new Jacobian, however, by combining J_1, J_2, and J_3 with weights, as follows:

$$J_A = \alpha J_1 + \beta J_2 + \gamma J_3 \ . \tag{26.62}$$

Here

$$\alpha + \beta + \gamma = 1 \ . \tag{26.63}$$

With three unknown coefficients, and only one constraint, namely (26.63), we are free to satisfy two additional constraints; and we take these to be (26.43) and (26.49). In this way, we can show that J_A will conserve both enstrophy and kinetic energy if we choose

$$\alpha = \beta = \gamma = 1/3 \ . \tag{26.64}$$

The composite Jacobian, J_A, is often called the "Arakawa Jacobian." It is also called J_7.

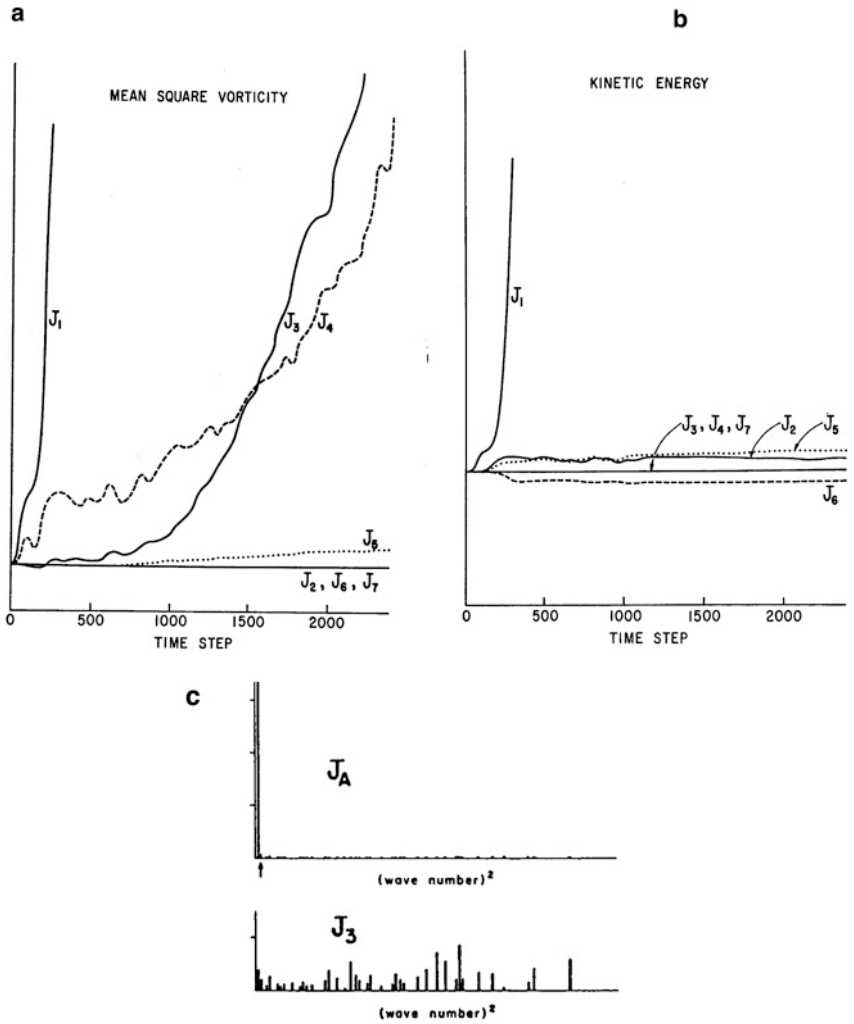


Figure 26.4: Results of tests with the various finite-difference Jacobians. Panel c shows that the initial kinetic energy is at a low wave number.

Fig. 26.4 shows the results of tests with J_1, J_2, and J_3, and also with three other Jacobians called J_4, J_5, and J_6, as well as with J_A. The leapfrog time-differencing scheme was used in these tests. The influence of time differencing on the conservation properties of the schemes is discussed in the next subsection; it is minor, as long as the criterion for linear computational instability is not violated. The various space-differencing schemes do indeed display the conservation properties expected on the basis of the preceding analysis.

The approach outlined above yields a second-order accurate (in space) finite-difference approximation to the Jacobian that conserves vorticity, kinetic energy, and enstrophy. Arakawa

(1966) also showed how to obtain a fourth-order Jacobian with the same conservation properties.

In Chapter 7, we concluded that, by suitable choice of the interpolated "cell-wall" values of an arbitrary advected quantity, A, it is possible to conserve exactly one non-trivial function of A, i.e., $F(A)$, in addition to A itself. Conserving more than A and one $F(A)$ was not possible because the only freedom that we had to work with was the form of the interpolated "cell-wall" value, which will be denoted here by $\hat{A}$. Once we chose $\hat{A}$ so as to conserve, say, A^2, we had no room left to maneuver, so we could not conserve anything else.

We have just shown, however, that the vorticity equation for two-dimensional nondivergent flow can be discretized so as to conserve *two* quantities, namely the kinetic energy and the enstrophy, in addition to the vorticity itself. How is that possible?

The key difference with the vorticity equation is that we can choose not only how to interpolate the vorticity (so as to conserve the enstrophy), but also *the actual finite-difference expression for the advecting wind itself*, in terms of the stream function, because that expression is implicit in the form of the Jacobian that we use. In choosing the form of the advecting current, we have a second "freedom," which allows us to conserve a second quantity, namely the kinetic energy.

As discussed earlier, the constraint of enstrophy conservation is needed to ensure that kinetic energy does not cascade in two-dimensional non-divergent flow. If kinetic energy does not cascade, the flow remains smooth. When the flow is smooth, kinetic energy conservation is approximately satisfied, even if it is not exactly guaranteed by the scheme. This means that a scheme that exactly conserves enstrophy and approximately conserves kinetic energy will behave well.

These considerations suggest that formal enstrophy conservation is "more important" than formal kinetic energy conservation.

26.4 The effects of time differencing on the conservation of squares

When time differencing is included, a family of finite-difference schemes for (25.16) can be written in the generic form

$$\frac{q_{i,\,j}^{n+1} - q_{i,\,j}^{n}}{\Delta t} = J_{i,\,j}\left(q^{*},\ \psi\right) \ , \tag{26.65}$$

where q is a generic scalar, and $J_{i,\,j}$ is a finite difference analog to the Jacobian at the point $(i,\ j)$. *Different choices of q^{*} give different time-differencing schemes*. Examples are given in Table 26.1.

Multiplying (26.65) by q^*, we get

$$q^* \left(q^{n+1} - q^n\right) = \Delta t q^* J\left(q^*,\ \psi\right) \ . \tag{26.66}$$

Here we drop the subscripts for simplicity. Eq. (26.66) can be rearranged to

$$\left(q^{n+1}\right)^2 - \left(q^n\right)^2 = 2\left(\frac{q^{n+1}+q^n}{2} - q^*\right)\left(q^{n+1} - q^n\right) + 2\Delta t q^* J\left(q^*,\ \psi\right) \ . \tag{26.67}$$

The left-hand side of (26.67) gives the change of q^2 in one time step. Let an overbar denote a sum over all grid points, divided by the number of grid points. Applying this averaging operator to (26.67), we find that

$$\overline{\left(q^{n+1}\right)^2} - \overline{\left(q^n\right)^2} = 2\overline{\left(\frac{q^{n+1}+q^n}{2} - q^*\right)\left(q^{n+1} - q^n\right)} + 2\Delta t \overline{q^* J\left(q^*,\ \psi\right)} \ . \tag{26.68}$$

We have already shown that we can choose our space differencing scheme in such a way that $\overline{q^* J\left(q^*,\ \psi\right)} = 0$. Time differencing will not enter in that choice, because only one "time level" of q, namely q^*, appears in $\overline{q^* J\left(q^*,\ \psi\right)}$.

The first term on the right-hand side of (26.68) is where the time-differencing comes in. For $q^* = q^n$, the contribution of this term is positive and so tends to increase $\overline{q^2}$. For $q^* = q^{n+1}$, the contribution of the first term is negative and so tends to decrease $\overline{q^2}$. With the trapezoidal implicit scheme, i.e., $q^* = \left(q^{n+1} + q^n\right)/2$, which is absolutely stable and neutral (in the linear case with constant coefficients), there is no contribution from the first term. This means that the trapezoidal implicit scheme is consistent with (allows) exact energy conservation. This could be anticipated given the time-reversibility of the trapezoidal implicit scheme, which was discussed earlier. Of course, the form of the finite-difference Jacobian must also be consistent with conservation of q^2.

In most cases, time truncation errors that interfere with exact energy conservation do not cause serious problems, provided that the scheme is stable in the linear sense, e.g., as indicated by von Neumann's method.

Table 26.1: Examples of time differencing schemes corresponding to various choices of q^*.

Name of Scheme	Form of Scheme
Euler forward	$q^* = q^n$
Backward implicit	$q^* = q^{n+1}$
Trapezoidal implicit	$q^* = \frac{1}{2}\left(q^n + q^{n+1}\right)$
Leapfrog, with time interval Δt / 2	$q^* = q^{n+\frac{1}{2}}$
Second-order Adams Bashforth	$q^* = \frac{3}{2}q^n - \frac{1}{2}q^{n-1}$
Heun	$q^* = q^n + \frac{\Delta t}{2}J\left(q^n, \psi\right)$
Lax-Wendorff (here *S* is a smoothing operator)	$q^* = Sq^n + \frac{\Delta t}{2}J\left(q^n, \psi\right)$
Matsuno	$q^* = q^n + \Delta t J\left(q^n, \psi\right)$

26.5 Summary

We began this chapter by discussing two-dimensional advection. We showed in Chapter 25 that when the advecting current is variable, a new type of instability can occur, which can be called "aliasing instability." In practice, it is often called "non-linear instability." This type of instability occurs regardless of the time step or time-differencing scheme, and cannot be detected by von Neumann's method. It can be detected by the energy method, and it can be controlled by enforcing conservation of appropriate quadratic variables, such as energy or enstrophy. It is particularly likely to cause trouble with the momentum equations, which describe how the wind is "advected by itself."

26.6 Problems

1. (a) Prove that J_2 conserves vorticity.

(b) Prove that J_3 conserves kinetic energy.

2. Work out the *continuous* form of the Jacobian for the case of spherical coordinates (longitude, λ, and latitude, φ).

3. For the case of two-dimensional non-divergent flow on a periodic domain, prove that if the vorticity is an eigensolution of the Laplacian, then the time-rate-of-change of the vorticity is zero.

4. Using the form of the Laplacian for the hexagonal grid that you worked out earlier in the semester, show that

$$\sum_i \left(\sigma_i \psi_i \frac{d\zeta_i}{dt} \right) = -\sum_i \left[\sigma_i \frac{d}{dt} \left(\frac{1}{2} |\nabla \psi|_i^2 \right) \right] \tag{26.69}$$

can be satisfied. Note that this condition only has to hold *for the sum over all grid points*, as shown.

5. For a hexagonal grid on a plane, show that a finite-difference Jacobian of the form

$$\frac{d\zeta_0}{dt} = \frac{1}{A} \sum_{i=1}^{6} \left(\frac{\psi_{i+1} - \psi_{i-1}}{\Delta s} \right) \left(\frac{\zeta_0 + \zeta_i}{2} \right) \Delta s \tag{26.70}$$

conserves vorticity, enstrophy, and kinetic energy. Here subscript 0 denotes the central point, the sum is over the six surrounding points (assumed to be numbered consecutively in a counter-clockwise fashion), A is the area of a hexagon, and Δs is the length of a side.

6. (a) Make a finite-difference model that solves the non-divergent vorticity equation on a doubly periodic plane, using an approximately square hexagonal grid with about 8000 grid cells, like the one used in the Chapter 3 homework. Use the Jacobian given in Problem 7 above, with Matsuno time differencing. You should check your Jacobian code by using a test function for which you can compute the Jacobian analytically.

   (b) Create diagnostics for the domain-averaged enstrophy and kinetic energy.

   (c) Invent an analytical function that you can use to specify an initial condition such that the periodic domain contains two pairs of (nearly) circular large-scale vortices of equal strength but opposite sign – two "highs" and two "lows." Because of the periodic boundary conditions, the solution will actually represent infinitely many vortices. Run the model with this smooth initial condition and discuss the results.

(d) Run the model again using initial conditions that approximate "white noise," and examine the time evolution of the solution. Does it follow the behavior expected for two-dimensional turbulence? You may have to run a thousand time steps or more to see the expected evolution.

Chapter 27

Conservative schemes for the two-dimensional shallow water equations with rotation

27.1 Kinetic energy conservation

In Chapter 26, we have shown how vorticity, kinetic energy and enstrophy can be conserved under advection in numerical simulations of two-dimensional non-divergent flow. In practice, however, we have to consider the presence of divergence. When the flow is divergent, vorticity and enstrophy are not conserved, but potential vorticity and potential enstrophy are conserved. Conservation of potential vorticity is an extremely important dynamical principle, as discussed in courses on atmospheric dynamics. Conservation of potential enstrophy is key to determining the distribution of kinetic energy with scale. Schemes that permit conservation of potential vorticity and potential enstrophy under advection therefore provide major benefits in the simulation of geophysical circulations.

The approach outlined below follows Arakawa and Lamb (1981). We adopt the C-grid, as shown in Fig. 27.1. Recall that on the C-grid, the zonal winds are east and west of the mass points, and the meridional winds are north and south of the mass points. The divergence "wants" to be defined at mass points, e.g., at point $\left(i+\frac{1}{2},\ j+\frac{1}{2}\right)$, and the vorticity "wants" to be defined at the corners of the mass boxes that lie along the diagonal lines connecting mass points, e.g., at the point $(i,\ j)$. The kinetic energy has to be defined by averaging u^2 and v^2 to a central location. It can be defined at the mass points, or the vorticity points, or both.

Note that in Fig. 27.1 and in the equations below, the vorticity is at the integer points, which is a departure from the indexing system used earlier.

The finite-difference form of the continuity equation is

$$\frac{dh_{i+\frac{1}{2},\,j+\frac{1}{2}}}{dt} = \frac{(hu)_{i,\,j+\frac{1}{2}} - (hu)_{i+1,\,j+\frac{1}{2}}}{\Delta x} + \frac{(hv)_{i+\frac{1}{2},\,j} - (hv)_{i+\frac{1}{2},\,j+1}}{\Delta y}\,. \tag{27.1}$$

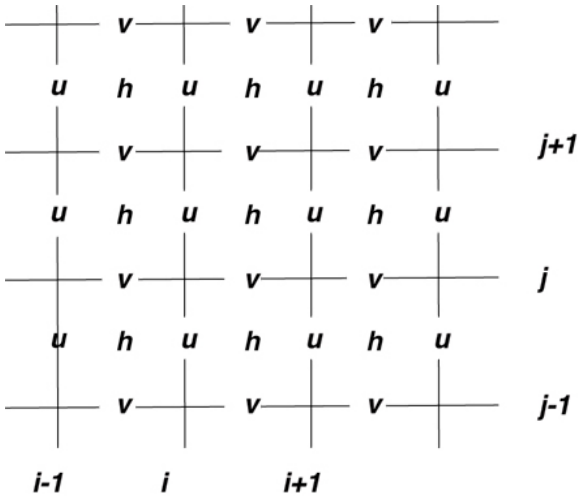


Figure 27.1: The arrangement of the mass, zonal wind, and meridional wind on the C grid. In this setup, the vorticity is at the integer points.

Here h is defined at half-integer points. The various mass fluxes that appear in (27.1) have not yet been defined, but mass will be conserved regardless of how we define them.

Simple finite-difference analogs of the two components of the momentum equation are

$$\frac{du_{i,\,j+\frac{1}{2}}}{dt} - \left[\left(\frac{\zeta+f}{h}\right)(hv)\right]_{i,\,j+\frac{1}{2}} + \left(\frac{K_{i+\frac{1}{2},\,j+\frac{1}{2}} - K_{i-\frac{1}{2},\,j+\frac{1}{2}}}{\Delta x}\right) + g\left[\frac{(h+h_S)_{i+\frac{1}{2},\,j+\frac{1}{2}} - (h+h_S)_{i-\frac{1}{2},\,j+\frac{1}{2}}}{\Delta x}\right] = 0\,, \tag{27.2}$$

and

$$\frac{dv_{i+\frac{1}{2},\,j}}{dt} + \left[\left(\frac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},\,j} + \left(\frac{K_{i+\frac{1}{2},\,j+\frac{1}{2}} - K_{i+\frac{1}{2},\,j-\frac{1}{2}}}{\Delta y}\right) + g\left[\frac{(h+h_S)_{i+\frac{1}{2},\,j+\frac{1}{2}} - (h+h_S)_{i+\frac{1}{2},\,j-\frac{1}{2}}}{\Delta y}\right] = 0\,, \tag{27.3}$$

respectively. As in the one-dimensional case discussed in Chapter 18, the kinetic energy per unit mass, $K_{i+\frac{1}{2},\,j+\frac{1}{2}}$, is undefined at this stage, but the values used in (27.2) and (27.3)

are defined at mass points. The potential vorticities $\left(\frac{\zeta+f}{h}\right)_{i,\,j+\frac{1}{2}}$ and $\left(\frac{\zeta+f}{h}\right)_{i+\frac{1}{2},\,j}$, and the mass fluxes $(hv)_{i,\,j+\frac{1}{2}}$ and $(hu)_{i+\frac{1}{2},\,j}$ are also undefined.

Note that the mass fluxes that appear in (27.2) and (27.3) are in the "wrong" places; the meridional mass flux $(hv)_{i,\,j+\frac{1}{2}}$ that appears in the equation for the u-wind is evidently at a u-wind point, and the zonal mass flux $(hu)_{i+\frac{1}{2},\,j}$ that appears in the equation for the v-wind is evidently at a v-wind point. The vorticities that appear in (27.2) and (27.3) are also in the "wrong" places. Obviously, what we have to do is interpolate somehow to obtain mass fluxes and vorticities suitable for use in the vorticity terms of (27.2) and (27.3). Note, however, that it is actually *products* of mass fluxes and vorticities that are needed.

There are two problems with the vorticity terms on the C grid. Both arise from the averaging needed, and both can occur on Cartesian or longitude-latitude C grids, as well as geodesic C grids. The first problem is that the averaging can lead to computational modes. This was discussed already in Chapter 19, in which the linearized equations included only the Coriolis parameter, rather than the absolute vorticity. The second problem is that although the vorticity terms do not affect the kinetic energy in the continuous system, we have to work to make this true in the finite-difference system.

One important question is: *Is there a finite-difference scheme that allows us to "mimic" the relationship*

$$(h\mathbf{v}) \cdot \left[\left(\frac{\zeta+f}{h} \right) \mathbf{k} \times (h\mathbf{v}) \right] = 0 \, ? \tag{27.4}$$

Here $\mathbf{k}$ is a vertical unit vector that is perpendicular to $\mathbf{v}$. Eq. (27.4) is what makes the vorticity term of the momentum equation drop out when we form the kinetic energy equation. Its essence is

$$\mathbf{v} \cdot (\mathbf{k} \times \mathbf{v}) = 0 \, , \tag{27.5}$$

which is a purely mathematical identity. Our goal is to mimic the identity itself. Arakawa and Lamb (1981) constructed the finite-difference vorticity terms of (27.2) and (27.3) in such a way that a finite-difference analog of (27.4) is satisfied, regardless of the specific forms of the mass fluxes and potential vorticities that are chosen. Their approach is to write:

$$\left[\left(\frac{\zeta+f}{h}\right)(hv)\right]_{i,\,j+\frac{1}{2}} = \\ \alpha_{i,\,j+\frac{1}{2};i+\frac{1}{2},\,j+1}(hv)_{i+\frac{1}{2},\,j+1} + \beta_{i,\,j+\frac{1}{2};i-\frac{1}{2},\,j+1}(hv)_{i-\frac{1}{2},\,j+1} + \gamma_{i,\,j+\frac{1}{2};i-\frac{1}{2},\,j}(hv)_{i-\frac{1}{2},\,j} \\ +\delta_{i,\,j+\frac{1}{2};i+\frac{1}{2},\,j}(hv)_{i+\frac{1}{2},\,j} + \varepsilon_{i+\frac{1}{2},\,j+\frac{1}{2}}(hu)_{i+1,\,j+\frac{1}{2}} - \varepsilon_{i-\frac{1}{2},\,j+\frac{1}{2}}(hu)_{i-1,\,j+\frac{1}{2}} \tag{27.6}$$

and

$$\left[\left(\frac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},\,j} = \\ \gamma_{i+\frac{1}{2},\,j;\,i+1,\,j+\frac{1}{2}}(hu)_{i+1,\,j+\frac{1}{2}} + \delta_{i+\frac{1}{2},\,j;\,i,\,j+\frac{1}{2}}(hu)_{i,\,j+\frac{1}{2}} + \alpha_{i+\frac{1}{2},\,j;\,i,\,j-\frac{1}{2}}(hu)_{i,\,j-\frac{1}{2}} \\ +\beta_{i+\frac{1}{2},\,j;\,i+1,\,j-\frac{1}{2}}(hu)_{i+1,\,j-\frac{1}{2}} + \phi_{i+\frac{1}{2},\,j+\frac{1}{2}}(hv)_{i+\frac{1}{2},\,j+1} - \phi_{i+\frac{1}{2},\,j-\frac{1}{2}}(hv)_{i+\frac{1}{2},\,j-1}\ . \tag{27.7}$$

In (27.6) and (27.7), the αs, βs, γs, δs, εs, and ϕs are interpolated potential vorticities whose forms are not yet specified. Each of these quantities has four subscripts, to indicate that it links a specific u-wind point with a specific v-wind point. The αs, βs, γs, δs, εs, and ϕs are somewhat analogous to the as and bs that were defined in the discussion of two-dimensional non-divergent flow, in that the as and bs also linked pairs of points. In (27.6), the interpolated potential vorticities multiply the mass fluxes hu and hv.

When we form the kinetic energy equation, we have to take the dot product of the vector momentum equation with the mass flux $h\mathbf{v}$. This means that we have to multiply (27.6) by $(hu)_{i+\frac{1}{2},\,j}$ and (27.7) by $(hv)_{i,\,j+\frac{1}{2}}$, and add the results. With the forms given by (27.6) and (27.7), the vorticity terms will combine to give

$$-(hu)_{i,\,j+\frac{1}{2}}\left[\left(\frac{\zeta+f}{h}\right)(hv)\right]_{i,\,j+\frac{1}{2}} + (hv)_{i+\frac{1}{2},\,j}\left[\left(\frac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},\,j} \\ = -(hu)_{i,\,j+\frac{1}{2}}\left[\alpha_{i,\,j+\frac{1}{2};i+\frac{1}{2},\,j+1}(hv)_{i+\frac{1}{2},\,j+1} + \beta_{i,\,j+\frac{1}{2};i-\frac{1}{2},\,j+1}(hv)_{i-\frac{1}{2},\,j+1} + \gamma_{i,\,j+\frac{1}{2};i-\frac{1}{2},\,j}(hv)_{i-\frac{1}{2},\,j}\right. \\ \left. +\delta_{i,\,j+\frac{1}{2};i+\frac{1}{2},\,j}(hv)_{i+\frac{1}{2},\,j} + \varepsilon_{i+\frac{1}{2},\,j+\frac{1}{2}}(hu)_{i+1,\,j+\frac{1}{2}} - \varepsilon_{i-\frac{1}{2},\,j+\frac{1}{2}}(hu)_{i-1,\,j+\frac{1}{2}}\right] \\ +(hv)_{i+\frac{1}{2},\,j}\left[\gamma_{i+\frac{1}{2},\,j;\,i+1,\,j+\frac{1}{2}}(hu)_{i+1,\,j+\frac{1}{2}} + \delta_{i+\frac{1}{2},\,j;\,i,\,j+\frac{1}{2}}(hu)_{i,\,j+\frac{1}{2}} + \alpha_{i+\frac{1}{2},\,j;\,i,\,j-\frac{1}{2}}(hu)_{i,\,j-\frac{1}{2}}\right. \\ \left. +\beta_{i+\frac{1}{2},\,j;\,i+1,\,j-\frac{1}{2}}(hu)_{i+1,\,j-\frac{1}{2}} + \phi_{i+\frac{1}{2},\,j+\frac{1}{2}}(hv)_{i+\frac{1}{2},\,j+1} - \phi_{i+\frac{1}{2},\,j-\frac{1}{2}}(hv)_{i+\frac{1}{2},\,j-1}\right]. \tag{27.8}$$

We want this long expression to sum to zero, over the whole grid, regardless of the numerical values assigned to the interpolated potential vorticies and the mass fluxes. We can make (27.8) much easier to look at by replacing the potential vorticities by 1s, and replacing hu and hv by u amd v, respectively. The result is

$$\begin{aligned} &-u_{i,\,j+\frac{1}{2}}\left(v_{i+\frac{1}{2},\,j+1}+v_{i-\frac{1}{2},\,j+1}+v_{i-\frac{1}{2},\,j}+v_{i+\frac{1}{2},\,j}+u_{i+1,\,j+\frac{1}{2}}-u_{i-1,\,j+\frac{1}{2}}\right) \\ &+v_{i+\frac{1}{2},\,j}\left(u_{i+1,\,j+\frac{1}{2}}+u_{i,\,j+\frac{1}{2}}+u_{i,\,j-\frac{1}{2}}+u_{i+1,\,j-\frac{1}{2}}+v_{i+\frac{1}{2},\,j+1}-v_{i+\frac{1}{2},\,j-1}\right) . \end{aligned} \tag{27.9}$$

An analysis of (27.9) shows that all the terms cancel out *when summed over the whole grid*. As a result, the vorticity terms will drop out of the finite-difference kinetic energy equation, just as they drop out of the continuous kinetic energy equation. This cancellation occurs regardless of the expressions that we choose of the mass fluxes, and regardless of the expressions that we choose for the αs, βs, γs, δs, εs, and ϕs. The cancellation arises purely from the forms of (27.6) and (27.7).

The discussion above shows that the finite-difference momentum equations represented by (27.2) and (27.3) with the use of (27.6) and (27.7), will guarantee kinetic energy conservation under advection, regardless of the forms chosen for the mass fluxes and the interpolated potential vorticities αs, βs, γs, δs, εs, and ϕs. From this point, the methods used in Chapter 18 will carry through essentially without change to give us conservation of mass, potential energy, and total energy.

Arakawa and Lamb (1981) went much further, showing how the finite-difference momentum equations presented above allow conservation of both potential vorticity and potential enstrophy. The details are rather complicated and will not be presented here.

27.2 TRiSK

Thuburn (2008), Thuburn et al. (2009), and Ringler et al. (2010) generalized the approach of Arakawa and Lamb (1981) for use an arbitrary C-grids, including spherical geodesic C-grids. Their approach is called "TRiSK," which stands for Thuburn, Ringler, Skamarock, and Klemp. TRiSK is used in MPAS. It does not solve the degrees-of-freedom problem with the geodesic C-grid.

27.3 Problems

1. As discussed in Chapter 19, the *linearized* shallow water equations can be written in vorticity-divergence form as

$$\frac{\partial h}{\partial t} + H\delta = 0 \,, \tag{27.10}$$

$$\frac{\partial \zeta}{\partial t} + f\delta = 0 \,, \tag{27.11}$$

$$\frac{\partial \delta}{\partial t} - f\zeta + g\nabla^2 h = 0 \,, \tag{27.12}$$

(a) Program these equations on the same doubly periodic hexagonal grid that you have used in the earlier homeworks. Use the unstaggered Z-grid, so that h, η, and δ are all defined at the cell centers.

(b) Use third-order Adams-Bashforth time differencing, which is discussed in Chapter 5 and in the paper by Durran (1991).

(c) Parameter settings:

i. Use a cell-center spacing of $d = 10,000$ m.

ii. Use $H = 5000$ m.

iii. Use $g = 0.1$ m s^{-2}.

iv. Use $f = 10^{-4}$ s^{-1}.

(d) Initial conditions:

(e) Use $\zeta = \delta = 0$. Start the simulation with a bump in h near the center of the domain. Make the bump smooth, with a maximum value of $h = 100$ m. Set $h = 0$ everywhere else. Include a map showing the shape of the initial bump.

(f) Run the model long enough for the area-average of h^2 to approach a constant value. Plot the area-average of h^2 as a function of time.

(g) Make an animation of h that shows the wave propagation that occurs as the model runs.

2. This problem includes nonlinearity, waves, diffusion, and the effects of bottom topography. As discussed by Heikes and Randall (1995a), the *nonlinear* shallow water equations can be written in vorticity-divergence form as

$$\frac{\partial h}{\partial t} + \nabla \cdot (h \nabla \chi) - J(h,\ \psi) = S_h \ , \tag{27.13}$$

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (\eta \nabla \chi) - J(\eta,\ \psi) = D \nabla^2 \eta \ , \tag{27.14}$$

and

$$\frac{\partial \delta}{\partial t} - \nabla \cdot (\eta \nabla \psi) - J(\eta,\ \chi) + \nabla^2 \left[K + g\,(h + h_T) \right] = 0 \ . \tag{27.15}$$

Here S_h is a (positive or negative) source of mass, η is the absolute vorticity, and $D \geq 0$ is a constant diffusion coefficient. Note that if the flow is nondivergent, so that $\nabla \chi = 0$, then the vorticity equation given above reduces to the form used in Chapter 26, except for the additional diffusion term.

(a) Program these equations on the same doubly periodic hexagonal grid that you have used in the earlier homeworks. Use the unstaggered Z-grid, so that h, η, and δ are all defined at the cell centers.

(b) You will need to solve a pair of Poisson equations with the relative vorticity and divergence as inputs. The solutions of the Poisson equations are the stream function, ψ, and the velocity potential, χ. Present the results of a test demonstrating that your Poisson solver is working properly.

(c) Use the Jacobian shown in (26.70), which is repeated here for your convenience:

$$J(\eta,\ \psi) = \frac{1}{A} \sum_{i=1}^{6} \left(\frac{\psi_{i+1} - \psi_{i-1}}{\Delta s} \right) \left(\frac{\eta_0 + \eta_i}{2} \right) \Delta s \tag{27.16}$$

Here Δs is the length of one side of a hexagon, and A is the area of a hexagon. Present the results of a test demonstrating that your Jacobian is working properly.

(d) The kinetic energy, K, is computed at the cell centers using

$$K = \frac{1}{2}\left[\nabla\cdot(\psi\nabla\psi) - \psi\nabla^2\psi + \nabla\cdot(\chi\nabla\chi) - \chi\nabla^2\chi\right] + J(\psi,\,\chi)\ . \qquad (27.17)$$

Present the results of a test demonstrating that your calculation of the kinetic energy is working properly.

(e) Choose the grid-cell edge values of h as the arithmetic means of the neighboring cell-center values. Choose the grid-cell edge values of q so as to allow conservation of q^2 under advection.

(f) Use third-order Adams-Bashforth time differencing, which is discussed in Chapter 5 and in the paper by Durran (1991).

(g) Parameter settings:

i. Use a cell-center spacing of $d = 10,000$ m.

ii. Use $g = 0.1$ m s^{-2}.

iii. Use $f = 10^{-4}$ s^{-1}.

iv. Use $D = 10^5$ m^2 s^{-1}.

v. Give a ridge bottom topography halfway out from the center of the domain, on the north side only, such that the largest value of h_T is 100 m. Use $h_T = 0$ elsewhere. Make h_T vary smoothly with a cross-ridge width of several grid cells. Make the ridge five times as long as it is wide. Include a map showing the distribution of h_T that you have specified.

vi. Give negative values of S_h near the center of the domain, with a largest negative value of $-h \times 10^{-4}$ m s^{-1}. Give positive values of S_h away from the center. Make S_h vary smoothly across the domain, and remember that it has to be doubly periodic. Make the area-averaged value of S_h zero, so that there is no net source or sink of mass. This set-up will make mass flow from the domain edges towards the domain center.

(h) Intial conditions:

Start the model with $h + h_T = 5000$ m everywhere. This is a flat free surface, with no pressure-gradient. Start with $q = f/h$ and $\delta = 0$ (no motion).

(i) Run the model until it reaches a statistically steady state.

(j) Make an animation that shows how the circulation spins up, and what the wave pattern looks like.

(k) Plot the area averages of q^2 and K as functions of time.

Chapter 28

Finite differences on the sphere

28.1 Introduction

Before we can use grid-point methods to discretize a model on the sphere, we must first define a grid that covers the sphere, i.e., we must discretize the sphere itself. There are many ways to do this, as discussed in the excellent review by Staniforth and Thuburn (2012).

Perhaps the most obvious possibility is to generate a grid using lines of constant latitude and longitude. With a grid based on such a "spherical coordinate system," indexing can be defined along coordinate lines, so that the neighbors of a particular cell can be referenced by defining an index for each coordinate direction, and then simply incrementing the indices to specify neighbors, as we have done many times when using Cartesian grids.

It is also possible to define grids without starting from a coordinate system. Examples are planar hexagonal and triangular grids, and spherical grids derived from the icosahedron, the octahedron, and the cube. These are often called "unstructured grids," although the term is not very descriptive, and seems almost pejorative. With an unstructured grid, the locations of each cell and its neighbors are listed in a table, which can be generated once and saved.

The governing equations can be written either with or without a coordinate system. When a coordinate system is used, the components of the wind vector are defined along the coordinate directions. On an unstructured grid, the orientations of the cell walls can be used to define local normal and tangent components of the wind vector on the cell walls. For example, a model that uses an unstructured C-grid will predict the normal component of the wind on each cell wall.

28.2 Spherical coordinates

28.2.1 Vector calculus in spherical coordinates

As discussed in Appendix A, in three-dimensional spherical coordinates $(\lambda,\ \varphi, r)$, i.e., longitude, latitude, and radius, the gradient, divergence, and curl operators take the following forms:

$$\nabla A = \left(\frac{1}{r\cos\varphi}\frac{\partial A}{\partial \lambda}, \frac{1}{r}\frac{\partial A}{\partial \varphi}, \frac{\partial A}{\partial r} \right) , \tag{28.1}$$

$$\nabla \cdot \mathbf{V} = \frac{1}{r\cos\varphi}\frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{r\cos\varphi}\frac{\partial}{\partial \varphi}\left(V_\varphi \cos\varphi\right) + \frac{1}{r^2}\frac{\partial}{\partial r}\left(V_r r^2\right) , \tag{28.2}$$

$$\nabla \times \mathbf{V} = \left\{ \frac{1}{r}\left[\frac{\partial V_r}{\partial \varphi} - \frac{\partial}{\partial r}\left(rV_\varphi\right)\right], \frac{1}{r}\frac{\partial}{\partial r}\left(rV_\lambda\right) - \frac{1}{r\cos\varphi}\frac{\partial V_r}{\partial \lambda}, \frac{1}{r\cos\varphi}\left[\frac{\partial V_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi}\left(V_\lambda \cos\varphi\right)\right] \right\} . \tag{28.3}$$

For use with the two-dimensional shallow-water equations, we can simplify these to

$$\nabla A = \left(\frac{1}{a\cos\varphi}\frac{\partial A}{\partial \lambda}, \frac{1}{a}\frac{\partial A}{\partial \varphi} \right) , \tag{28.4}$$

$$\nabla \cdot \mathbf{v} = \frac{1}{a\cos\varphi}\frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{a\cos\varphi}\frac{\partial}{\partial \varphi}\left(V_\varphi \cos\varphi\right) , \tag{28.5}$$

$$\mathbf{k} \cdot (\nabla \times \mathbf{v}) = \frac{1}{a\cos\varphi}\left[\frac{\partial V_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi}\left(V_\lambda \cos\varphi\right)\right] . \tag{28.6}$$

Here $\mathbf{v}$ is the horizontal velocity vector, and a is the radius of the spherical planet. Much further discussion is given in Appendix A.

28.2.2 The "pole problem"

In a spherical coordinate system, the lines of constant longitude converge at the poles, so longitude is multivalued at the poles. Fig. 28.1 shows one eighth of a uniform latitude-longitude grid. The zonal rows of grid points nearest the two poles consist of "pizza slices" which come together at a point at each pole. The other zonal rows consist of grid points which are roughly trapezoidal (i.e., quadrilateral) in shape.

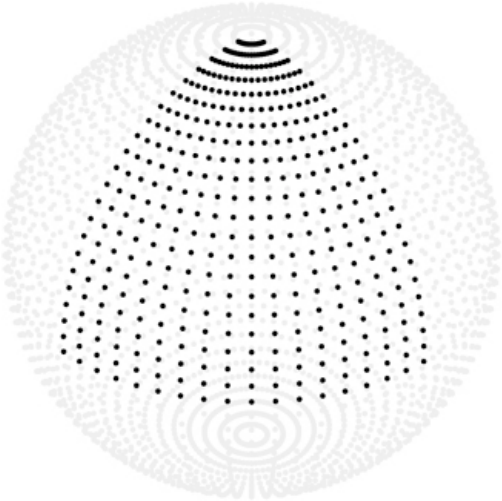


Figure 28.1: One octant of the latitude-longitude grid. In this example, there are 72 grid points around a latitude circle and 44 latitude bands from pole to pole. The longitudinal grid spacing is globally uniform, and in this example is 5°. The latitudinal grid spacing is 4° except that the pizza slices are 6° tall.

The components of the wind vector (and all other vectors) are discontinuous at the poles, although of course the (real-world) wind vector itself doesn't even know that there is a pole. For example, consider a jet directed over the North Pole, represented by the shaded arrow in Fig. 28.2. Measured at points along the prime meridian, the wind consists entirely of a positive v component. Measured along the international date line, however, the wind consists entirely of a negative v component. A discontinuity occurs at the pole, where "north" and "south" have no meaning. Similarly, the u component of the wind is positive measured near the pole along longitude 90°, and is negative measured along longitude 270°.

Such ambiguity does not occur in a Cartesian coordinate system centered on the pole. At each point along a great circle that includes the pole, the components measured in such a Cartesian coordinate system are well defined and vary continuously. But a Cartesian coordinate system centered on the pole is not very useful far away from the pole.

The scales of meteorological action do not vary dramatically from place to place. This suggests that average distance between neighboring grid points should not depend on location, and also that the distances between grid points in the zonal direction should not

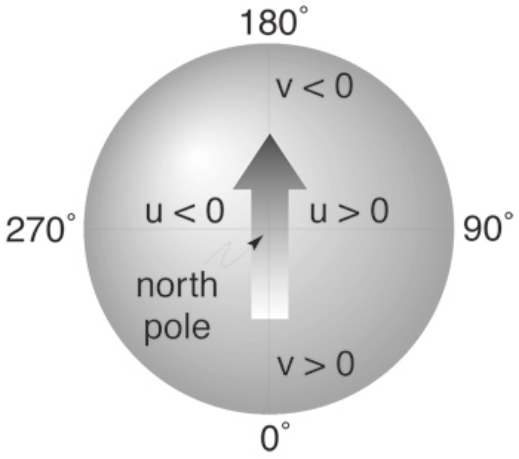


Figure 28.2: For the wind vector shown in the sketch, points along the prime meridian have a strong northward component. There is a discontinuity at the pole, and points along international date line have a strong southward component. Points near 90° longitude have a positive zonal component, while points near 270° longitude have a negative zonal component.

be substantially different from the distances in the meridional direction. Unfortunately, latitude-longitude grids lack these two desirable properties.

In addition, the convergence of the meridians at the poles demands a short time step in order to satisfy the Courant-Friedrichs-Lewy (CFL) requirement for computational stability, as discussed in Chapters 7 (for advection) and 19 (for wave propagation). The short time step is a practical problem, so we often talk about "the pole *problem*." There are actually two pole problems: one for advection, and another for wave propagation. Semi-Lagrangian advection schemes can eliminate the pole problem for advection. Semi-implicit time-differencing schemes can eliminate the pole problem for wave propagation. Spectral models make it easy to implement semi-implicit time differencing for wave propagation. Further discussion spectral methods is given in Chapter 29.

28.3 The shallow water equations in spherical coordinates

To show how filters can be used to address the pole problem for wave propagation, we use the shallow water equations in spherical coordinates, which can be written as

$$\frac{\partial u}{\partial t} + \frac{u}{a\cos\varphi}\frac{\partial u}{\partial \lambda} + \frac{v}{a}\frac{\partial u}{\partial \varphi} - \left(f + \frac{u}{a}\tan\varphi\right)v + \frac{g}{a\cos\varphi}\frac{\partial}{\partial \lambda}(h+h_S) = 0\,, \tag{28.7}$$

$$\frac{\partial v}{\partial t} + \frac{u}{a\cos\varphi}\frac{\partial v}{\partial \lambda} + \frac{v}{a}\frac{\partial v}{\partial \varphi} + \left(f + \frac{u}{a}\tan\varphi\right)u + \frac{g}{a}\frac{\partial}{\partial \varphi}(h+h_S) = 0\,, \tag{28.8}$$

$$\frac{\partial h}{\partial t}+\frac{u}{a\cos\varphi}\frac{\partial h}{\partial \lambda}+\frac{v}{a}\frac{\partial h}{\partial \varphi}+\frac{h}{a\cos\varphi}\left[\frac{\partial u}{\partial \lambda}+\frac{\partial}{\partial \varphi}\left(v\cos\varphi\right)\right]=0\,. \tag{28.9}$$

The peculiar $u\tan\varphi/a$ terms of (28.7) and (28.8) are sometimes called the "metric terms." They arise from the use of spherical coordinates and do not appear in the vector form of the equations. The metric terms vanish at the equator but can become large near the poles.

We linearize (28.7) - (28.9) about a state of rest, neglecting rotation and bottom topography:

$$\frac{\partial u}{\partial t}+\frac{g}{a\cos\varphi}\frac{\partial h}{\partial \lambda}=0\,, \tag{28.10}$$

$$\frac{\partial v}{\partial t}+\frac{g}{a}\frac{\partial h}{\partial \varphi}=0\,, \tag{28.11}$$

$$\frac{\partial h}{\partial t}+\frac{h}{a\cos\varphi}\left[\frac{\partial u}{\partial \lambda}+\frac{\partial}{\partial \varphi}\left(v\cos\varphi\right)\right]=0\,. \tag{28.12}$$

Here H denotes the mean depth of the fluid. The linearization about a resting basic state has eliminated the nonlinear metric terms. We spatially discretize (28.10) - (28.12) using the C-grid, as follows:

$$\frac{du_{j+\frac{1}{2},\,k}}{dt}+\frac{g\left(h_{j+1,k}-h_{j,\,k}\right)}{a\cos\varphi\Delta\lambda}=0 \tag{28.13}$$

$$\frac{dv_{j,\,k+\frac{1}{2}}}{dt}+\frac{g\left(h_{j,\,k+1}-h_{j,\,k}\right)}{a\Delta\varphi}=0\,, \tag{28.14}$$

$$\frac{dh_{j,k}}{dt} + H\left\{\left(\frac{u_{j+\frac{1}{2},k} - u_{j-\frac{1}{2},k}}{a\cos\varphi_j \Delta\lambda}\right) + \left[\frac{(v\cos\varphi)_{j,k+\frac{1}{2}} - (v\cos\varphi)_{j,k-\frac{1}{2}}}{a\cos\varphi_j \Delta\varphi}\right]\right\} = 0\,. \quad (28.15)$$

Here j is the zonal index, k is the meridional index, and the time derivatives have been left in continuous form. We look for solutions of the form

$$u_{j+\frac{1}{2},k} = \mathrm{Re}\left\{\hat{u}_k \exp\left[im\left(j+\frac{1}{2}\right)\Delta\lambda + i\sigma t\right]\right\}\,, \quad (28.16)$$

$$v_{j,k+\frac{1}{2}} = \mathrm{Re}\left\{\hat{v}_{k+\frac{1}{2}} \exp\left[i\left(mj\Delta\lambda + \sigma t\right)\right]\right\}\,, \quad (28.17)$$

$$h_{j,k} = \mathrm{Re}\left\{\hat{h}_k \exp\left[i\left(mj\Delta\lambda + \sigma t\right)\right]\right\}\,. \quad (28.18)$$

Note that the zonal wave number, m, is defined with respect to longitude rather than distance, and that the "hat" variables depend on latitude. By substitution of (28.16) - (28.18) into (28.13) - (28.15), we obtain

$$\sigma\hat{u}_k + \frac{m}{a\cos\varphi_j}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2} S_k(m) g\hat{h}_k = 0\,, \quad (28.19)$$

$$i\sigma\hat{v}_{k+\frac{1}{2}} + g\left(\frac{\hat{h}_{k+1} - \hat{h}_k}{a\Delta\varphi}\right) = 0\,, \quad (28.20)$$

$$i\sigma\hat{h}_k + H\left\{\frac{im}{a\cos\varphi_k}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2} S_k(m)\hat{u}_k + \left[\frac{(\hat{v}\cos\varphi)_{k+\frac{1}{2}} - (\hat{v}\cos\varphi)_{k-\frac{1}{2}}}{a\cos\varphi_k\Delta\varphi}\right]\right\} = 0\,, \quad (28.21)$$

where $S_k(m)$ is an artificially inserted "smoothing parameter" that depends on zonal wave number and latitude. The smoothing parameter has been inserted into the term of (28.19) corresponding to the *zonal* pressure gradient force, and also into the term of (28.21) corresponding to the *zonal* mass flux divergence. These are the key terms for zonally propagating gravity waves. For now, just consider it $S_k(m)$ be equal to one.

By eliminating $\hat{u}_k$ and $\hat{v}_{k+\frac{1}{2}}$ in (28.19) - (28.21), we can obtain a "meridional structure equation" for $\hat{h}_k$:

$$\begin{aligned} &c^2\left[\frac{|c|\ m\sin(m\Delta\lambda/2)}{a\cos\varphi_k\, m\Delta\lambda/2}S_k(m)\right]^2\hat{h}_k \\ &+\frac{c^2}{(a\Delta\varphi)^2}\left[\left(\hat{h}_k-\hat{h}_{k-1}\right)\frac{\cos\varphi_{k-\frac{1}{2}}}{\cos\varphi_k}-\left(\hat{h}_{k+1}-\hat{h}_k\right)\frac{\cos\varphi_{k+\frac{1}{2}}}{\cos\varphi_k}\right]=\sigma^2\hat{h}_k\ . \end{aligned} \tag{28.22}$$

Here $c^2\equiv gH$ is the square of the phase speed of a pure gravity wave. For the shortest zonal wavelengths, with $m\Delta\lambda\cong\pi$, the first term on the left-hand side of (28.22) dominates the second, so we can simplify to

$$\sigma\cong\frac{|c|\ \sin(m\Delta\lambda/2)}{a\cos\varphi_k\,\Delta\lambda/2}S_k(m)\ . \tag{28.23}$$

Although we have not used a time-differencing scheme here, we know that for a conditionally stable scheme the condition for linear computational stability takes the form

$$\sigma\Delta t<\varepsilon\ , \tag{28.24}$$

where ε is a constant of order one. Using (28.23), this criterion can be written as

$$\frac{|c|\ \Delta t\ \sin(m\Delta\lambda/2)}{a\cos\varphi_k\,\Delta\lambda/2}S_k(m)<\varepsilon\ , \tag{28.25}$$

In view of (28.23) and (28.24), the CFL criterion will place more stringent conditions on Δt as $a\cos\varphi_k\Delta\lambda$ decreases, i.e., near the poles. In addition, the criterion becomes more stringent as m increases, for a given latitude. The worst case is $\sin(m\Delta\lambda/2)=1$, which happens for $m\Delta\lambda=\pi$, i.e., the shortest possible wave length. For this worst case, (28.25) reduces to

$$\frac{|c|\,\Delta t}{a\cos\varphi_k \Delta\lambda}\, S_k\,(m) < \varepsilon\ . \tag{28.26}$$

This shows that, as expected, the time step required for stability depends on latitude. For the grid shown in Fig. 28.1, with a longitudinal grid spacing of $\Delta\lambda = 5^\circ$ and a latitudinal grid spacing of $\Delta\varphi = 4^\circ$ (the values used to draw the figure), the northernmost row of grid points where the zonal component of velocity is defined is at latitude 86°N. The zonal distance between grid points there is $\Delta x \cong 39$ km, which is less than one-tenth the zonal grid spacing at the Equator. Recall that the fast, external gravity wave has a phase speed of approximately 300 m s^{-1}. Substituting into (28.26), we find that with that resolution and $S_k\,(m) = 1$, the largest permissible time step near the pole is about 70 seconds. This is about one tenth of the largest permissible time step at the Equator.

28.4 Polar filters

It would be nice if the CFL criterion were the same at all latitudes, permitting time steps near the pole as large as those near the Equator. In order to make this possible, models that use latitude-longitude grids typically employ "polar filters" that prevent computational instability, so that a longer time step can be used.

The simplest method is to use a Fourier filter to remove the high-wave number components of the prognostic fields themselves, near the poles. This can prevent a model from blowing up, but it leads to drastic violations of mass conservation (and many other conservation principles). The cure is as bad as the disease.

A better approach is to longitudinally smooth the longitudinal pressure gradient in the zonal momentum equation and the longitudinal contribution to the mass flux divergence in the continuity equation. This has the effect of reducing the zonal phase speeds of the gravity waves sufficiently so that the CFL criterion is not violated. The smoothing parameter $S_k\,(m)$ serves this function. This is how the smoothing paramter has been used in (28.19) and (28.21).

To implement the smoothing parameter, we compute the Fourier coefficients of the zonal pressure gradient and the zonal mass flux divergence, multiply the coefficients by numbers less than or equal to one, and then do the inverse transform to construct the smoothed tendencies on the grid.

It remains to choose the form of $S_k\,(m)$. Our goal is to make the CFL criterion independent of latitude. Inspection of (28.23) shows that this can be accomplished by choosing the smoothing parameter $S_k\,(m)$ so that

$$\frac{S_k(m)\sin(m\Delta\lambda/2)}{a\cos\varphi_k\Delta\lambda} = \frac{1}{d^*}\,, \tag{28.27}$$

where d^* is a suitably chosen constant length scale, *comparable to the zonal grid spacing at the Equator*. When $S_k(m)$ satisfies (28.27), the stability criterion (28.23) reduces to

$$\sigma = \frac{2|c|}{d^*}, \tag{28.28}$$

and the CFL condition reduces to

$$\frac{|c|\,\Delta t}{d^*} < \frac{\varepsilon}{2}\,, \tag{28.29}$$

so that *the time step required is independent of latitude*, as desired. If we choose

$$d^* \equiv a\Delta\lambda\,, \tag{28.30}$$

i.e., the zonal distance between grid points at the equator, then, referring back to (28.27), we see that $S_k(m)$ must satisfy

$$S_k(m) = \frac{\cos\varphi_k}{\sin(m\Delta\lambda/2)}\,. \tag{28.31}$$

Of course, at low latitudes (28.31) can give values of $S_k(m)$ which are greater than one; these should be replaced by one, so that we actually use

$$S_k(s) = \min\left\{\frac{\cos\varphi_k}{\sin(m\Delta\lambda/2)},\ 1\right\}\,. \tag{28.32}$$

For modes with $\sin(m\Delta\lambda/2) < 1$, smoothing occurs only near the poles. For the shortest zonal wavelength, with $\sin(m\Delta\lambda/2)\,11$, smoothing is needed all the way down to the Equator.

We are free to choose d^*. Eq. (28.30) is not the only possibility. Smaller values of d^* can eliminate filtering in low latitudes.

How is the filter implemented? First, the fields to be filtered are expanded in Fourier series, to obtain an amplitude for each wave number. Then the amplitudes are multiplied by $S_k(m)$. Finally, an inverse Fourier transform is performed.

The Fourier transforms involve all grid cells at a given latitude. This can be a slow operation with domain-decomposition on parallel machines.

Polar filters are a solution to the pole problem for wave propagation. They work, but they are not ideal. They can have undesirable side effects on the simulation. They do not solve the pole problem for advection.

28.5 The Kurihara grid

Many authors have sought alternatives to the latitude-longitude grid, hoping to make the grid spacing more uniform, still within the "latitude-longitude" framework.

For example, Kurihara (1965) proposed a grid in which the number of grid points along a latitude circle varies with latitude. By placing fewer points at higher latitudes, he was able to more homogeneously cover the sphere. The grid is constructed by evenly placing $N+1$ grid points along the longitude meridian, from the North Pole to the Equator. The point at the North Pole is given the label $j=1$, the next latitude circle south is given the label $j=2$, and so on until the Equator is labeled $j=N+1$. Along latitude circle j there are $4(j-1)$ equally spaced grid points, except at each pole, where there is a single point. One octant of the sphere is shown in Fig. 28.3; compare with Fig. 28.1. For a given N, the total number of grid points on the sphere is $4N^2+2$. The Southern Hemisphere grid is a mirror image of the Northern Hemisphere grid.

Kurihara built a model using this grid, based on the shallow water equations. He tested it in a simulation of the Rossby-Haurwitz wave, with zonal wave number 4 as the initial condition. [1] Kurihara's model was run with a variety of time-stepping schemes and with varying amounts of viscosity. Each simulation covered 16 simulated days, with $N=20$. The Rossby-Haurwitz wave should move from west to east, without distortion. In several of Kurihara's tests, the wave spuriously degenerated to higher wave numbers.

Conceptually similar "skipped" grids were discussed by James Purser (1988), and by Halem and Russell (1973) as described by Herman and Johnson (1978) and Shukla and Sud (1981).

[1] This set of initial conditions was also used by Phillips (1959a), and later in the suite of seven test cases for shallow water models proposed by Williamson et al. (1992).

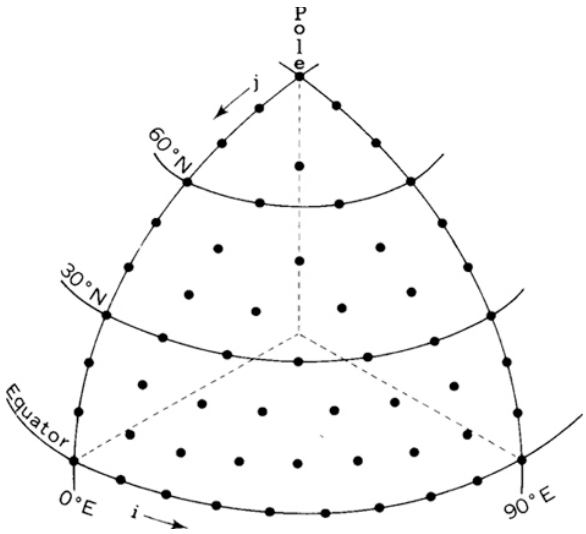


Figure 28.3: Kurihara grid on one octant of the sphere. Compare with Fig. 28.1.

28.6 Displaced poles

The important eddies of the ocean circulation are an order of magnitude smaller than those of the atmosphere. Consequentially, the horizontal grids of ocean models have, historically, been finer than those of their coupled atmosphere models. Ocean models have also used grids with shapes different from those of their coupled atmosphere models. The land model almost always runs on the same grid as the atmosphere.

In order to avoid the pole problem, and also to allow more uniform resolution in the Arctic Ocean, some ocean models have modified the longitude-latitude grid near the North Pole. In particular, MOM6 uses a tri-polar grid (Murray, 1996, Fig 28.4) in which the North Pole of the longitude-latitude system is "split in two," and the two resulting poles are displaced onto the continents, away from the pole of the Earth's rotation. The third pole is on the Antarctic continent and so lies outside the domain of an ocean model.

Today, the grid spacings used in some global atmosphere and ocean models are on the order of 25 km or finer, and the trend to increasing resolution seems likely to continue. Grids with spacings of a few kilometers can resolve mesoscale convective systems and gravity waves in the atmosphere, and mesoscale eddies in the ocean, as well as large mountains, lakes, some rivers, many cities, and various other features on the land surface. With such fine grid spacing there is no motivation to increase complexity by using different grids for the different model components. The use of identical grids for all components brings architectural simplicity, eliminates the need for interpolation and averaging as the atmosphere and ocean models exchange information, and increases computational efficiency.

Figure 28.4: A tri-polar grid used in an ocean model. One pole can be seen in northern Europe, and a second occurs in northern North America. The third pole is the conventional South Pole, in Antarctica.

28.7 Grids based on map projections

An early approach to numerically solving the shallow water equations on the sphere was to project the sphere onto a plane, and solve the equations on a regular grid using a coordinate system defined in the plane. The surface of a sphere and that of a plane are not topologically equivalent, however. Distances and areas can be badly distorted near the singular points of the projections. Nevertheless, we can use a projection to map the piece of the sphere away from the singular points. An approach to map the entire sphere using projections is the composite mesh method, discussed later.

We can derive the equations of motion in various map projections if we first express them in a general orthogonal coordinate system $(x,\ y)$. Here x and y *do not necessarily have the dimensions of length;* for example, they could be angles. Define the metric coefficients α_x and α_y so that the distance increment satisfies

$$dl^2 = \alpha_x^2 dx^2 + \alpha_y^2 dy^2 \ . \tag{28.33}$$

The metric coordinates convert coordinate increments (whatever their dimensions might be) into true distances. In the $(x,\ y)$ coordinate system, the horizontal velocity components are given by

$$u = \alpha_x \frac{dx}{dt} \ , \tag{28.34}$$

$$v = \alpha_y \frac{dy}{dt} \ . \tag{28.35}$$

Williamson (1969) gives the equations of motion in terms of the general velocity components:

$$\frac{Du}{Dt} - \left[f + \frac{1}{\alpha_x \alpha_y} \left(v \frac{\partial \alpha_y}{\partial x} - u \frac{\partial \alpha_x}{\partial y} \right) \right] v + \frac{g}{\alpha_x} \frac{\partial}{\partial x} (h + h_S) = 0, \tag{28.36}$$

$$\frac{Dv}{Dt} + \left[f + \frac{1}{\alpha_x \alpha_y} \left(v \frac{\partial \alpha_y}{\partial x} - u \frac{\partial \alpha_x}{\partial y} \right) \right] u + \frac{g}{\alpha_y} \frac{\partial}{\partial y} (h + h_S) = 0 \ . \tag{28.37}$$

The Lagrangian time derivative is given by

$$\frac{D}{Dt}(\ ) = \frac{\partial}{\partial t}(\ ) + \frac{u}{\alpha_x} \frac{\partial}{\partial x}(\ ) + \frac{v}{\alpha_y} \frac{\partial}{\partial y}(\ ) \ . \tag{28.38}$$

The continuity equation can be written as

$$\frac{\partial h}{\partial t} + \frac{1}{\alpha_x \alpha_y} \left[\frac{\partial}{\partial x} (\alpha_y h u) + \frac{\partial}{\partial y} (\alpha_x h v) \right] = 0 \ . \tag{28.39}$$

As an example, with spherical coordinates we have

$$x = \lambda \quad \text{and} \quad y = \varphi \,, \tag{28.40}$$

and the corresponding metric coefficients are

$$\alpha_x = a\cos\varphi \quad \text{and} \quad \alpha_y = a \,. \tag{28.41}$$

Then, by (28.34) and (28.35), we have

$$u = a\cos\varphi \frac{D\lambda}{Dt} \quad \text{and} \quad v = a\frac{D\varphi}{Dt} \,. \tag{28.42}$$

Substituting (28.41) and (28.42) into (28.36), (28.37) and (28.39) gives (28.7), (28.8) and (28.9), which are the shallow water equations in spherical coordinates.

The Polar Stereographic and Mercator projections are sometimes used in modeling the atmospheric circulation. Both are examples of conformal projections, that is, they preserve angles, but not distances. Also, in both of these projections the metric coefficients are independent of direction at a given point, i.e., $\alpha_x = \alpha_y$. The effects of these projections on the outlines of the continents are shown in Fig. 28.5.

The polar stereographic projection can be visualized in terms of a plane tangent to the Earth at the North Pole. A line drawn from the South Pole that intersects the Earth will also intersect the plane. This line establishes a one-to-one correspondence between all points on the plane and all points on the sphere except for the South Pole itself. In the plane, we can define a Cartesian coordinate system (X,Y), where the positive X axis is in the direction of the image of $\lambda = 0$ (the Greenwich meridian), and the positive Y axis is in the direction of the image of $\lambda = \pi/2$. Obviously, similar mappings can be obtained by placing the plane tangent to the sphere at points other than the North Pole. Haltiner and Williams (1980) give the equations relating the projection coordinates (X,Y) and the spherical coordinates $(\lambda,\ \varphi)$:

$$X = \frac{2a\cos\varphi\cos\lambda}{1+\sin\varphi}, \tag{28.43}$$

$$Y = \frac{2a\cos\varphi\sin\lambda}{1+\sin\varphi}. \tag{28.44}$$

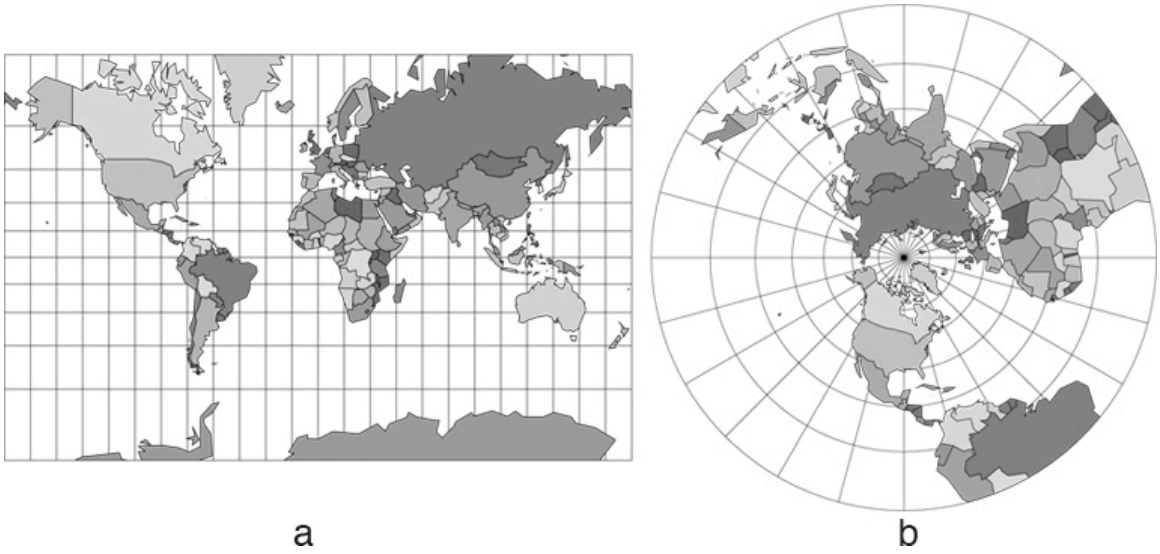


Figure 28.5: Map projections of the continents: a.) Mercator projection, in which the surface of the sphere is projected onto a cylinder that has its axis aligned with the poles, and the poles are stetched out into lines. b.) North polar stereographic projection, in which a hemisphere is projected onto a plane centered at the pole.

Note that there is a problem at the South Pole, where the denominators of (28.43) and (28.44) go to zero. From (28.43) and (28.44) we find that

$$\begin{bmatrix} dX \\ dY \end{bmatrix} = \left(\frac{2a}{1+\sin\varphi}\right) \begin{bmatrix} -\cos\varphi\sin\lambda & -\cos\lambda \\ \cos\varphi\cos\lambda & -\sin\lambda \end{bmatrix} \begin{bmatrix} d\lambda \\ d\varphi \end{bmatrix}. \tag{28.45}$$

The metrics of the polar stereographic map projection can be determined as follows: Substituting $x = \lambda$, $y = \varphi$, and the metrics for spherical coordinates into (28.33) gives

$$dl^2 = (a\cos\varphi)^2 d\lambda^2 + a^2 d\varphi^2. \tag{28.46}$$

Solving the linear system (28.45) for $d\varphi$, and $d\lambda$, and substituting the results into (28.46), we obtain

$$dl^2 = \left(\frac{1+\sin\varphi}{2}\right)^2 dX^2 + \left(\frac{1+\sin\varphi}{2}\right)^2 dY^2. \tag{28.47}$$

Comparing (28.47) with (28.33), we see that metric coefficients for the polar stereographic projection are given by

$$\alpha_x = \alpha_y = \frac{1+\sin\varphi}{2} \,. \tag{28.48}$$

We define the map factor, $m(\varphi)$, as the inverse of the metric coefficient, so that, for example, $m(\varphi) = 2/(1+\sin\varphi)$. Using (28.36), (28.37), and (28.39), we can write the shallow water equations in north polar stereographic coordinates:

$$\frac{DU}{Dt} - \left(f + \frac{uY - vX}{2a^2} \right) v + gm(\varphi) \frac{\partial}{\partial X} (h + h_S) = 0 \,, \tag{28.49}$$

$$\frac{Dv}{Dt} + \left(f + \frac{uY - vX}{2a^2} \right) u + gm(\varphi) \frac{\partial}{\partial Y} (h + h_S) = 0 \,, \tag{28.50}$$

$$\frac{\partial h}{\partial t} + m^2(\varphi) H \left\{ \frac{\partial}{\partial X} \left[\frac{u}{m(\varphi)} \right] + \frac{\partial}{\partial Y} \left[\frac{v}{m(\varphi)} \right] \right\} = 0 \,. \tag{28.51}$$

The total derivative is given by (28.38).

28.8 Composite grids

As discussed above, a finite region of the plane will only map onto a piece of the sphere, and vice versa. One technique to map the entire sphere is to partition it, for example, into hemispheres, and project the pieces separately. Each set of projected equations then gets its boundary conditions from the solutions of the other projected equations.

For example, Phillips (1957) divided the sphere into three regions: a tropical belt, and extratropical caps to the north and south of the tropical belt. On each region, the shallow water equations are mapped to a new coordinate system. He used a Mercator coordinate system in the tropics, a polar stereographic coordinate system fixed to the sphere at the North Pole for the northern extratropical cap, and similarly, a polar stereographic coordinate system fixed to the sphere at the South Pole for the southern extratropical cap. When

a computational stencil required data from outside the region covered by its coordinate system, that piece of information was obtained by interpolation within the neighboring coordinate system. The model proved to be unstable at the boundaries between the coordinate systems.

Browning et al. (1989) proposed a composite-mesh model in which the Northern and Southern Hemispheres are mapped to the plane with a polar stereographic projection. The equations used for the northern projection are just (28.49), (28.50), and (28.51). The equations for the southern projection are the same as those for the northern, except for a few sign differences. This model is different from Phillips' in that the regions interior to the coordinate systems overlap a little bit as shown in Fig. 28.6. Values for dependent variables at grid points not covered by the current coordinate system are obtained by interpolation in the other coordinate system. The overlapping of the coordinate systems made Browning's model more stable than Phillips' model, in which the coordinate systems were simply "bolted together" at a certain latitude.

Composite grids are rarely used today, although the idea does occasionally resurface, as in the Yin-Yang grid, which is briefly described later in this chapter.

28.9 Unstructured spherical grids

Fig. (28.7) shows seven alternative discretizations of the sphere. The left-most panel shows the latitude-longitude grid. Some ocean models use a modified latitude-longitude grid in which the north pole is replaced by two "displaced" poles – one in northern North America, and a second in northern Europe (Murray, 1996). The grid within the Arctic ocean is then free of singularities and relatively uniform. Grids of this type are called "tri-polar."

The second and third panels of Fig. (28.7) show triangular and hexagonal-pentagonal grids, respectively; both of these are generated by starting from the icosahedron. They will be discussed further below.

The fourth panel shows a "cubed sphere" grid, generated from the sphere (e.g., Ronchi et al. (1996); Nair et al. (2005); Putman and Lin (2007);Lauritzen and Nair (2008); Ullrich et al. (2009)). The cells of the cubed sphere grid are quadrilaterals.

The fifth panel shows the "Ying-Yang" grid proposed by Kageyama and Sato (2004), and Kageyama (2005). The grid is composed of two "sleeves" that overlap like the two leather patches that are stitched together to cover the outside of a baseball. The sleeves overlap slightly, and an interpolation is used to patch them together, giving a composite grid.

The sixth panel shows a spiraling "Fibonacci grid" (Swinbank and Purser, 2006).

The last panel shows the HEALpix grid (Calabretta and Roukema, 2007). This recnet idea was developed for use in astrophysics but has attracted interest by geophysical mod-

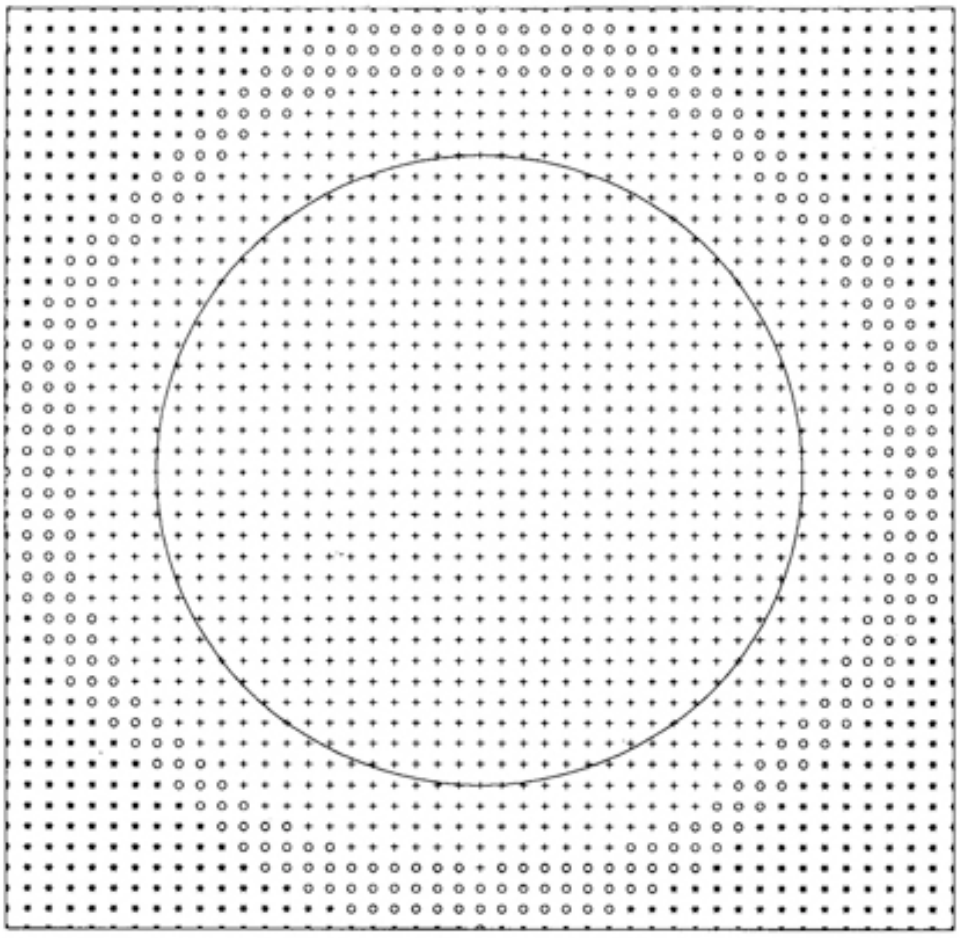


Figure 28.6: Composite grid method grid. Two such grids are used to cover the sphere. Points labeled with ○ are the boundary conditions for the points labeled with +. Values at the ○ points are obtained by interpolation from the other grid. The big circle is the image of the Equator. Points labeled * are not used. From Browning et al. (1989).

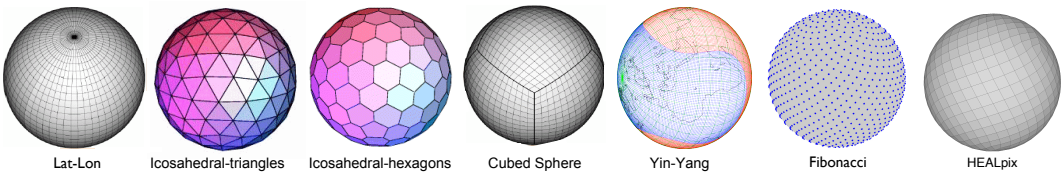


Figure 28.7: Seven grids on the sphere.

elers (Chang et al., 2023). All cells of the Healpix grid cover exactly the same area on the sphere. The cells are arranged along lines of constant latitude, facilitating fast and efficient zonal averaging, Fourier analysis, and spherical harmonic analysis. Healpix supports a multi-resolution, hierarchical tree structure that enables local operations and fast range

queries on the sphere.

28.9.1 Wandering electrons

This is just for your amusement: One idea for constructing a mesh that homogeneously covers a sphere is to calculate the equilibrium distribution of a set of electrons confined to the surface of the sphere. Because each electron is repelled by every other electron, we can hypothesize that the electrons will position themselves so as to maximize the distance between closest neighbors, and thus distribute themselves as evenly as possible over the sphere. We can then associate a grid point with each electron. It seems advantageous to constrain the grid so that it is symmetric across the Equator. An Equator can be defined by restricting the movement of a subset of the electrons to the equatorial great circle. We can also fix an electron at each of the poles. The remaining electrons can be paired so that each has a mirror image in the opposite hemisphere. Experience shows that unless the positions of some of the electrons are fixed, their positions will continue to wander indefinitely. Fig. 28.8 shows a grid constructed using this "wandering electron" algorithm. Most cells have six walls, but some have five or seven walls. While this approach does more or less homogeneously cover the sphere, it is not very satisfactory.

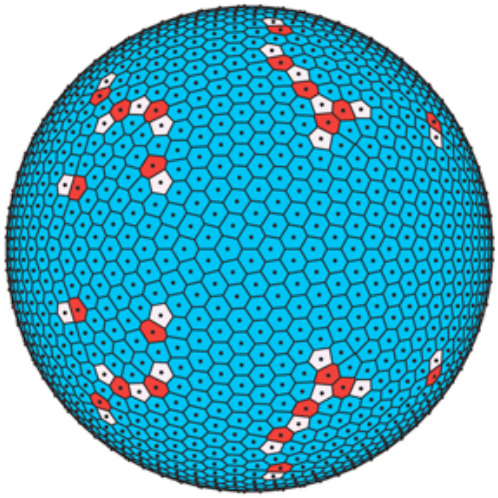


Figure 28.8: Wandering electron grid. White cells have five edges, blue cells have six edges, and red cells have seven edges.

28.9.2 Spherical grids based on the Platonic solids

28.9.2.1 The cube

Cubed sphere grids are generated by projecting the faces of a cube onto a circumscribed sphere, creating six adjoining grid faces that cover the sphere seamlessly. Grid points from each of the six cube faces are projected onto the spherical surface. This is often done using,

gnomonic projection, which maps straight lines on each cube face to great circles on the sphere. A cubed-sphere grid consists of six curved quadrilateral faces mapped from the cube sides. Coordinate lines do not extend continuously across cube face boundaries, so special numerical techniques are needed to handle connections between faces, especially where three faces of the cube come together. Some variants improve uniformity or orthogonality by slightly adjusting the initially projected grid points. Nesting can be done on the cubed sphere by subdividing faces into smaller cells for higher resolution.

28.9.2.2 The octahedron

There have been attempts to use grids based on octahedrons (e.g., McGregor (1996); Purser and Rančcić (1998)).

28.9.2.3 The icoasahedron

Grids based on the icosahedron (20 faces and 12 vertices) offer an attractive framework for simulation of the global circulations of the atmosphere and ocean. Their advantages include almost uniform and quasi-isotropic resolution over the sphere. Such grids are termed "geodesic," because they resemble the geodesic domes designed by Buckminster Fuller. Williamson (1968) and Sadourny et al. (1968) simultaneously proposed using geodesic grids based on equilateral spherical triangles that are nearly equal in area. Because the grid points are not regularly spaced and do not lie in orthogonal rows and columns, alternative finite-difference schemes are used to discretize the equations. Initial tests using the grid proved encouraging, and further studies were carried out. These were reported by Sadourny et al. (1968), Sadourny and Morel (1969), Sadourny (1969), Williamson (1970), and Masuda and Ohnishi (1986).

A conceptually simple scheme for constructing a spherical geodesic grid is to divide the edges of the icosahedral faces into equal lengths, create new smaller equilateral triangles in the plane, and then project onto the sphere. See Fig. 28.9. One can construct a more homogeneous grid by partitioning the spherical equilateral triangles instead. Williamson (1968) and Sadourny et al. (1968) use slightly different techniques to construct their grids. However, both begin by partitioning the spherical icosahedral triangle. On these geodesic grids, all but twelve of the cells are hexagons. The remaining twelve are pentagons. They are associated with the twelve vertices of the original icosahedron.

Arbitrarily fine grids can be constructed using the recursive procedure described above. Figure 28.10 lists some of the possible grid spacings.

Geodesic grids are Voronoi grids. A Voronoi grid is a way of dividing a space into regions based on distance to a specific set of points, called seeds or generator points. Each region, called a Voronoi cell, consists of all points that are closer to the region's seed than to any other seed. This leads to a tessellation of the domain into polygons (or polyhedra in 3D). The concept is named after the Russian mathematician Georgy Voronoi, who defined

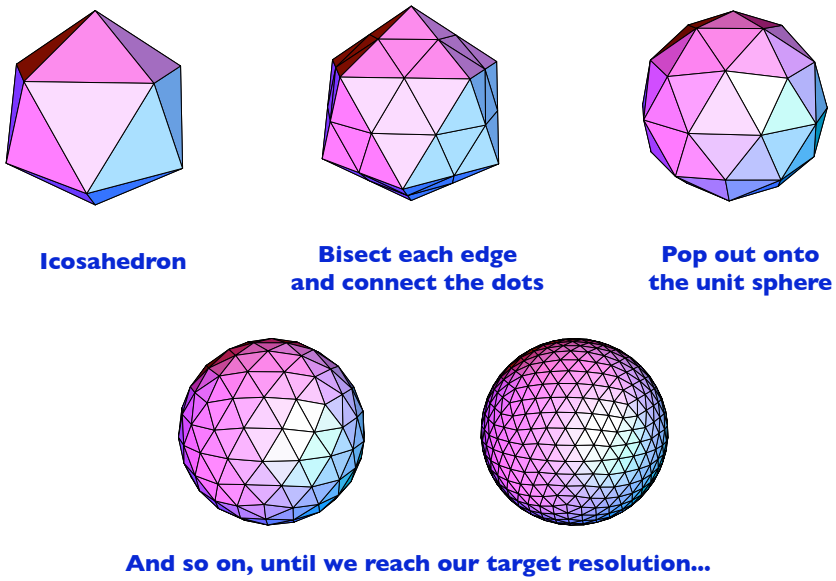


Figure 28.9: A spherical geodesic grid is generated recursively by starting from an icosahedron

Grid	No. of grid points N	Avg grid distance ℓ (km)
G0	12	6699.1
G1	42	3709.8
G2	162	1908.8
G3	642	961.4
G4	2562	481.6
G5	10 242	240.9
G6	40 962	120.4
G7	163 842	60.2
G8	655 362	30.1
G9	2 621 442	15.0
G10	10 485 762	7.53
G11	41 943 042	3.76
G12	167 772 162	1.88
G13	671 088 642	0.94

Figure 28.10: Some of the geodesic grids that can be generated through recursive subdivision of the faces of an icosahedron.

and studied the general n-dimensional case in 1908.

Williamson (1968) chose the non-divergent shallow water equations to test his geodesic

grid. He solved the non-divergent barotropic vorticity equation discussed in Chapter 26. It is repeated here for your convenience:

$$\frac{\partial \zeta}{\partial t} = J(\eta, \psi), \tag{28.52}$$

where ζ is relative vorticity, $\eta = \zeta + f$ is absolute vorticity and ψ is the stream function, such that

$$\zeta = \nabla^2 \psi. \tag{28.53}$$

For arbitrary functions α and β, it follows from the form $J(\alpha, \beta) = \mathbf{k} \cdot \nabla \times (\alpha \nabla \beta)$ that the Jacobian satisfies

$$J(\alpha,\, \beta) = \lim_{A \to 0} \left(\frac{1}{A} \oint_S \alpha \frac{\partial \beta}{\partial s}\, ds \right), \tag{28.54}$$

where A is the area of a grid cell, and s measures distance along the curve bounding A, i.e., the perimeter of the grid cell. Integrating (28.52) over the area A, and using (28.54), we get

$$\frac{d}{dt} \int_A \zeta \, dA = \oint_S \eta \frac{\partial \psi}{\partial s}\, ds\,. \tag{28.55}$$

This can be discretized with reference to Fig. 28.11. We approximate the line integral along the polygon defined by the path $P_1,\ P_2, \ldots P_K$. Let ζ_0 be the relative vorticity defined at the point P_0, and let η_i be the absolute vorticity defined at the point P_i. We can approximate (28.55) by

$$\frac{d\zeta_0}{dt} = \frac{1}{A} \sum_{i=1}^{K} \left(\frac{\psi_{i+1} - \psi_{i-1}}{\Delta s} \right) \left(\frac{\eta_0 + \eta_i}{2} \right) \Delta s\,. \tag{28.56}$$

We must also discretize the Laplacian. Consider the smaller, inner polygon in Fig. 28.11. Its walls are formed from the perpendicular bisectors of the line segments $\overline{P_0 P_i}$. We can use Gauss's Theorem to write

$$\int_a \zeta \, dA = - \oint_{s'} (\nabla \psi) \cdot \mathbf{n} \, ds' \, , \tag{28.57}$$

where A is the area of the small polygon, s' is its boundary, and $\mathbf{n}$ is the outward-normal unit vector on the boundary. Eq. (28.57) is approximated by

$$a\zeta_0 = \sum_{i=1}^{K} \frac{l_i}{|P_0 P_i|} (\psi_i - \psi_0) \, , \tag{28.58}$$

where $|P_0 P_i|$ is the distance from P_0 to P_i, and l_i is the length of wall i. Eq. (28.58) can be solved for ψ_i by relaxation, using the methods discussed in Chapter 15.

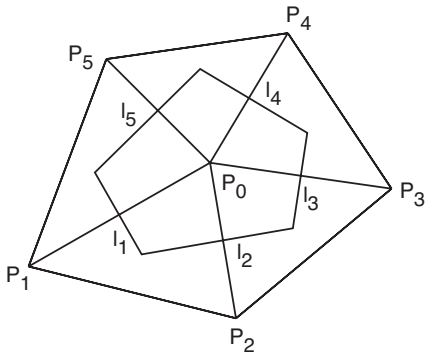


Figure 28.11: Configuration of grid triangles for the case of a pentagon. The pentagon is bounded by the line segments labeled I_1 to I_5. Each vertex of the pentagon is surrounded by a triangle.

Williamson showed that his scheme conserves kinetic energy and enstrophy, as the exact equations do. When applied to regular grid on a plane, the scheme is second-order accurate. Williamson performed a numerical experiment, using a Rossby-Haurwitz wave as the initial condition. A run of 12 simulated days produced good results. Sadourny et al. (1968) discussed a nondivergent model very similar to Williamson's. Also, Sadourny and Morel (1969) developed a geodesic-grid model based on the free-surface shallow water equations.

Masuda and Ohnishi (1986) developed an elegant spherical shallow water model, based on the Z-grid (see Chapter 19). Like Williamson, Masuda chose the Rossby-Haurwitz wave with wave number 4 as his initial condition. Heikes and Randall (1995a,b) and Heikes et al. (2013) extended Masuda's work by introducing a multi-grid method to compute the stream function and velocity potential from the vorticity and divergence, respectively. Heikes

and Randall (1995b) also showed that the grid can be "optimized," to permit consistent finite-difference approximations to the divergence, Jacobian, and Laplacian operators that are used in the construction of the model. They tested their model using certain standard test cases for shallow water on the sphere Williamson et al. (1992), and obtained good results. Ringler et al. (2000) constructed a full-physics global atmospheric model using this approach.

28.10 Summary

In order to construct a numerical model on the sphere, it is necessary to map the sphere onto a computational domain. There are various ways of doing this. The most straightforward is to use latitude-longitude coordinates, but this leads to the pole problem for both advection and wave propagation. The pole problem can be dealt with by using filters, which come with their own set of problems. Semi-implicit differencing can be used to avoid the need for filtering.

Another approach is to use a regular grid on the sphere. A perfectly regular grid is mathematically impossible, but geodesic grids come close.

A third approach, discussed in the next chapter, is to use the spectral method, with spherical harmonics as the basis functions.

28.11 Problems

1. Suppose that you are running a model in which the phase speed of the fastest gravity wave is 300 m s^{-1}. Design a filter that makes it possible run the model with a 10-km grid spacing and a time step of 1000 s. Plot the coefficients of the filter as a function of wave number.

Chapter 29

Spectral methods

29.1 Introduction

Spectral models represent the horizontal structure of a field by using "functional expansions," rather than finite differences. An elementary example of a functional expansion is a Fourier series. In the first part of this chapter we will use Fourier series to explain some of the basic concepts of spectral models. In practice, most spectral models use spherical-harmonic expansions on the global domain. For reasons that will be explained below, most spectral models also make use of grid-point representations, and go back and forth between "wave-number space" (the functional expansion) and "physical space" (the grid-point representation) as the model runs.

29.2 Transform pairs

Consider a function $q(x, t)$ with one spatial dimension, x, and also time dependence. We assume that $q(x, t)$ is real and integrable. If the domain is periodic, with period L, we can express the spatial structure of $q(x, t)$ *exactly* by a Fourier series expansion:

$$q(x, t) = \sum_{k=-\infty}^{\infty} \hat{q}_k(t)\, e^{ikx} \,. \tag{29.1}$$

The complex coefficients $\hat{q}_k(t)$ can be computed using

$$\hat{q}_k(t) = \frac{1}{L} \int_{x-L/2}^{x+L/2} q\left(x', t\right) e^{-ikx'}\, dx' \,. \tag{29.2}$$

Recall that the proof of (29.1) and (29.2) involves use of the orthogonality condition

$$\frac{1}{L}\int_{x-L/2}^{x+L/2} e^{-ikx'}\, e^{ilx'}\, dx' = \delta_{k,\,l}\,, \tag{29.3}$$

where

$$\delta_{k,\,l} \equiv \begin{cases} 1,\ k = l \\ 0,\ k \neq l \end{cases} \tag{29.4}$$

is the Kronecker delta. Eqs. (29.1) and (29.2) are a "transform pair." They can be used to go back and forth between physical space and wave-number space.

29.3 Differentiation

From (29.1), we see that the x-derivative of q satisfies

$$\frac{\partial q}{\partial x}(x,\,t) = \sum_{k=-\infty}^{\infty} ik\hat{q}_k(t)\, e^{ikx}\,. \tag{29.5}$$

Inspection of (29.5) shows that $\partial q/\partial x$ does not receive a contribution from $\hat{q}_0$; the reason for this should be clear.

29.4 Truncation

A spectral model uses equations similar to (29.1), (29.2), and (29.5), but with a finite set of wave numbers, and with x defined on a finite mesh:

$$q\left(x_j,\,t\right) \cong \sum_{k=-n}^{n} \hat{q}_k(t)\, e^{ikx_j}\,, \tag{29.6}$$

$$\hat{q}_k(t) = \frac{1}{M}\sum_{j=1}^{M} q\left(x_j,\,t\right) e^{-ikx_j}, -n \leq k \leq n\,, \tag{29.7}$$

$$\frac{\partial q}{\partial x}\left(x_j,\, t\right) \cong \sum_{k=-n}^{n} ik\hat{q}_k\left(t\right) e^{ikx_j} \,. \tag{29.8}$$

The sums in (29.6) and (29.8) are *truncated*, in that they do not include wave numbers outside the range $\pm n$. The value of n is chosen by the modeler. The sum that appears in (29.7) is over a grid with M points. Note that we have used "approximately equal signs" in (29.6) and (29.8), but not in (29.7). Why?

A key point is that *for each wave number the derivative in* (29.8) *is exact*. The approximation in (29.8) comes from the truncation.

There should be some relationship between M and n, because M measures the amount of information available on the grid, and n measures the amount of information available in the spectral coefficients. Apart from the effects of round-off error, the transform (29.6) is *exactly* reversible by the inverse transform (29.7), provided that M is large enough, i.e., provided that there are enough points on the grid.

So how many points do we need? To see the answer, substitute (29.6) into (29.7) to obtain

$$\hat{q}_k\left(t\right) = \frac{1}{M}\sum_{j=1}^{M}\left\{\left[\sum_{l=-n}^{n}\hat{q}_l\left(t\right)e^{ilx_j}\right]e^{-ikx_j}\right\} \quad \text{for} \quad -n \le k \le n \,. \tag{29.9}$$

This is, of course, a rather circular substitution, but the result serves to clarify some basic ideas. If expanded, each term on the right-hand side of (29.9) involves the product of two wave numbers, l and k, each of which lies in the range $-n$ to n. The range for wave number l is explicitly spelled out in the inner sum on the right-hand side of (29.9); the range for wave number k is understood because, as indicated, we will evaluate the left-hand side of (29.9) for k in the range $-n$ to n. Because each term on the right-hand side of (29.9) involves the product of two Fourier modes with wave numbers in the range $-n$ to n, each term includes wave numbers up to $\pm 2n$. We therefore need $2n+1$ complex coefficients, i.e., $2n+1$ values of the $\widehat{q}_k\left(t\right)$. In general, this is the equivalent of $4n+2$ real numbers, which suggests that we need $M \ge 4n+2$ in order to represent the real-valued function $q\left(x_j,\, t\right)$ on a grid.

The required value of M is actually much smaller, however, for the following reason. Because q is assumed to be real, it turns out that

$$\hat{q}_{-k} = \hat{q}_k^* \,, \tag{29.10}$$

where the superscript $*$ denotes the conjugate. This helps because $\hat{q}_{-k}$ and $\hat{q}_k^*$ together involve only two real numbers, rather than four. To see why (29.10) is true, consider the *combined* contributions of the $+k$ and $-k$ coefficients to the sum in (29.6). Define $T_k\left(x_j,\, t\right)$ by

$$\begin{aligned} T_k\left(x_j,\, t\right) &\equiv \hat{q}_k\left(t\right) e^{ikx_j} + \hat{q}_{-k}\left(t\right) e^{-ikx_j} \\ &\equiv R_k e^{i\theta} e^{ikx_j} + R_{-k} e^{i\mu} e^{-ikx_j} \\ &= R_k e^{i\left(\theta + kx_j\right)} + R_{-k} e^{i\left(\mu - kx_j\right)} \,. \end{aligned} \tag{29.11}$$

Here $R_k e^{i\theta} \equiv \hat{q}_k\left(t\right)$ and $R_{-k} e^{i\mu} \equiv \hat{q}_{-k}\left(t\right)$, where R_k and R_{-k} are real and non-negative. With this definition, we can rewrite (29.6) as

$$q\left(x_j,\, t\right) \cong \sum_{k=0}^{n} T_k\left(x_j,\, t\right) \,. \tag{29.12}$$

Our assumption that $q\left(x_j,\, t\right)$ is real, combined with the linear independence of distinct Fourier modes, implies that the imaginary part of $T_k\left(x_j\right)$ must be zero, for all x_j. With the use of Euler's formula, it follows that

$$R_k \sin\left(\theta + kx_j\right) + R_{-k} \sin\left(\mu - kx_j\right) = 0 \text{ for all } x_j \,. \tag{29.13}$$

The only way to satisfy (29.13) for all x_j is to set

$$R_k = R_{-k} \tag{29.14}$$

and

$$\theta + kx_j = -\left(\mu - kx_j\right) = -\mu + kx_j \,. \tag{29.15}$$

From (29.15), we see that $\theta = -\mu$. Eq. (29.10) follows from (29.14) and (29.15).

As mentioned above, because $\hat{q}_k$ and $\hat{q}_{-k}$ are complex conjugates of each other, they involve only two distinct real numbers. If you know $\hat{q}_k$ you can immediately write down $\hat{q}_{-k}$. In addition, it follows from (29.10) that $\hat{q}_0$ is real. Therefore, the $2n+1$ complex values of $\hat{q}_k$ actually embody the equivalent of only $2n+1$ distinct real numbers, rather than $4n+2$ real numbers. The Fourier representation up to wave number n is thus equivalent to a representation of the real function $q(x,\, t)$ using $2n+1$ equally spaced grid points, in the sense that the information content is the same. We conclude that, *in order to use a grid of M points to represent the amplitudes and phases of all waves up to $k = \pm n$, we need $M \geq 2n+1$*; we can use a grid with more than $2n+1$ points, but not fewer. If $M \geq 2n+1$, the transform pair (29.6) - (29.7) is perfectly reversible.

As a simple example, a Fourier representation of q, including just wave numbers zero and one, is equivalent to a grid-point representation of q using 3 grid points. The real values of q assigned at the three grid points suffice to compute the coefficient of wave number zero (i.e., the mean value of q) and the phase and amplitude (or "sine and cosine coefficients") of wave number one.

29.5 Spectral differentiation in terms of finite-differences

Substituting (29.7) into (29.8) gives

$$\frac{\partial q}{\partial x}\left(x_j,\, t\right) \cong \sum_{k=-n}^{n} \left[\frac{ik}{M} \sum_{l=1}^{M} q\left(x_l,\, t\right) e^{-kx_l} \right] e^{ikx_j} \tag{29.16}$$

Reversing the order of summation leads to

$$\frac{\partial q}{\partial x}\left(x_j,\, t\right) \cong \sum_{l=1}^{M} a_j^l q\left(x_l,\, t\right) \ , \tag{29.17}$$

where we define

$$a_j^l \equiv \frac{1}{M} \sum_{k=-n}^{n} ike^{ik\left(x_j - x_l\right)} \ . \tag{29.18}$$

A problem at the end of this chapter invites you to prove that the a_j^l are real numbers. The point of this exercise is that (29.17) can be interpreted as a member of the family of

finite-difference approximations discussed many times in this course, starting with (3.23). The scheme given by (29.17) is somewhat special in that it involves *all* grid points in the domain. From this point of view, spectral models can be regarded as a class of finite-difference models.

29.6 Solving linear equations with the spectral method

Now consider the one-dimensional advection equation with a constant current, u:

$$\frac{\partial q}{\partial t} = -u\frac{\partial q}{\partial x} \,. \tag{29.19}$$

Substituting (29.6) and (29.8) into (29.19) gives

$$\sum_{k=-n}^{n} \frac{d\hat{q}_k}{dt} e^{ikx} = -u \sum_{k=-n}^{n} ik\hat{q}_k e^{ikx} \,. \tag{29.20}$$

By linear independence, we obtain

$$\frac{d\hat{q}_k}{dt} = -iku\hat{q}_k \quad \text{for} \quad -n \le k \le n \,. \tag{29.21}$$

Note that $d\hat{q}_0/dt$ will be equal to zero; the reason for this should be clear. We can use (29.21) to predict $\hat{q}_k(t)$. When we need to know $q\left(x_j,\, t\right)$, we can get it from (29.6).

Compare (29.21) with

$$\frac{d\hat{q}_k}{dt} = -iku\left[\frac{\sin(k\Delta x)}{k\Delta x}\right]\hat{q}_k \,, \tag{29.22}$$

which, as discussed in earlier chapters, is obtained by using centered second-order space differencing. For a spatially uniform advecting current, the spectral method gives the *exact* speed for each Fourier mode, while the finite-difference method gives a slower value, especially for high wave numbers. Similarly, spectral methods give the *exact* phase speeds for linear waves propagating through a uniform medium, while finite-difference methods generally underestimate the phase speeds.

Keep in mind, however, that the spectral solution is not really exact, because only a finite number of modes are kept. In addition, the spectral method does not give the exact answer, even for individual Fourier modes, when the advection speed (or a wave's phase speed) is spatially variable.

Another strength of spectral methods is that they make it very easy to solve boundary value problems. As an example, consider

$$\nabla^2 q = f\,(x,\, y)\ , \tag{29.23}$$

as a problem to determine q for given $f(x,y)$. In one dimension, (29.23) becomes

$$\frac{d^2 q}{dx^2} = f\,(x)\ . \tag{29.24}$$

We assume periodic boundary conditions and compute the Fourier coefficients of both q and f, using (29.7). Then (29.24) can be written as

$$\sum_{k=-n}^{n} \left(-k^2\right) \hat{q}_k e^{ikx} = \sum_{k=-n}^{n} \hat{f}_k e^{ikx}\ , \tag{29.25}$$

Invoking linear independence to equate coefficients of e^{ikx}, we find that

$$\hat{q}_k = -\frac{\hat{f}_k}{k^2} \quad \text{for} \quad -n \le k \le n \quad (\text{unless } k = 0)\ . \tag{29.26}$$

Eq. (29.26) can be used to obtain $\hat{q}_k$, for $k = 1,\ n$. Then $q\,(x)$ can be constructed using (29.6). This completes the solution of (29.24), apart from the application of an additional "boundary condition" to determine $\hat{q}_0$. The solution is exact *for the modes that are included*; it is approximate because not all modes are included.

29.7 Solving nonlinear equations with the spectral method

Now consider a *nonlinear* problem, such as momentum advection, i.e.,

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x}\ , \tag{29.27}$$

again with a periodic domain. Fourier expansion gives

$$\sum_{k=-n}^{n} \frac{d\hat{u}_k}{dt} e^{ikx} = -\left(\sum_{l=-n}^{n} \hat{u}_l e^{ilx} \right) \left(\sum_{m=-n}^{n} im\hat{u}_m e^{imx} \right) \tag{29.28}$$

Our goal is to predict $\hat{u}_k(t)$ for k in the range $-n$ to n. Wave numbers outside that range are excluded by our choice of truncation. The right-hand-side of (29.28) involves products of the form $e^{ilx}e^{imx}$, where l and m are each in the range $-n$ to n. These products can generate "new" wave numbers, some of which lie outside the range $-n$ to n. Those that lie outside the range are simply neglected, i.e., they are not included when we evaluate the left-hand side of (29.28).

For a given Fourier mode, (29.28) implies that

$$\frac{d\hat{u}_k}{dt} = -\left\{ \sum_{l=-\alpha}^{\alpha} \sum_{m=-\alpha}^{\alpha} im \left[\hat{u}_l \hat{u}_m e^{i(l+m)x} \right] \right\} e^{-ikx}, \quad \text{for} \quad -n \le k \le n\,. \tag{29.29}$$

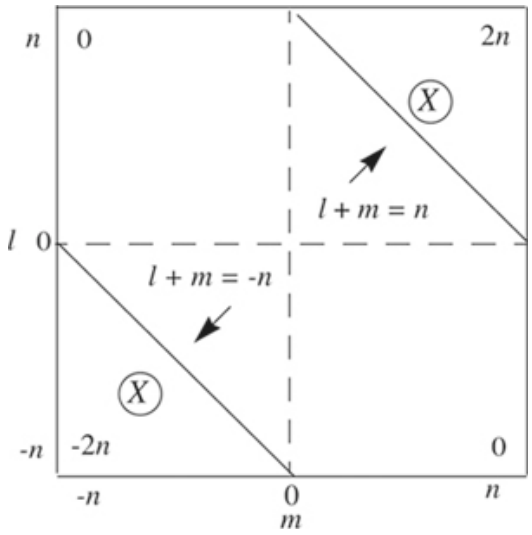


Figure 29.1: Cartoon "table" of $l+m$, showing which $(l,\, m)$ pairs can contribute to wave numbers k in the range $-n$ to n in Eq. (29.29). The pairs in the triangular regions marked by X's do not contribute.

In (29.29)*, the quantity in curly braces involves sums over wave numbers and is therefore defined at grid points*. The summations on the right-hand side of (29.29) are over the range $\pm\alpha$.

In order to get the exact value of $d\hat{u}_k/dt$ for all k in the range $-n$ to n, we must choose α large enough so that we pick up all possible combinations of l and m that lie in the range $-n$ to n. See Fig. 29.1. The circled Xs in the figure denote excluded triangular regions. The number of points in each triangular region is

$$1+2+3\cdots+(n-1)=\frac{n(n-1)}{2}\ . \tag{29.30}$$

The number of points *retained* is therefore given by

$$\begin{aligned}(2n+1)^2-2\left[\frac{n(n-1)}{2}\right] &= \left(4n^2+4n+1\right)-\left(n^2-n\right)\\ &= 3n^2+5n+1\ .\end{aligned} \tag{29.31}$$

This is the number of terms that must be evaluated inside the square brackets in (29.29). The key point is that the number of terms is of order n^2, i.e., it grows quadratically as n increases. As a result, the amount of computation also grows rapidly as n increases, and of course the problem is "twice as hard" in two dimensions. At first, this poor scaling with problem size appeared to make spectral methods prohibitively expensive for nonlinear (i.e., realistic) problems. Note that the same issue would arise in a linear problem with spatially variable coefficients.

29.8 The transform method

A way around this practical difficulty was proposed independently by Orszag (1970) and Eliasen et al. (1970). They suggested a "transform method" in which (29.6) and (29.8) are used to evaluate both u and $\partial u/\partial x$ on a grid. The product $u\partial u/\partial x$ is computed on the grid, and the result is transformed back into wave-number space.

We can obtain an exact result exact up to wave number n by making the number of grid points used large enough to allow the exact representation, for wave numbers in the range $-n$ to n, of *quadratic* nonlinearities like $u\partial u/\partial x$. Here "exact" means "exact up to wave number n." Because the solution is exact for wave numbers up to n, there is no error for those wave numbers, and in particular, there *is no aliasing error*. Therefore, a model of this type is not subject to aliasing instability arising from quadratic terms like $u\partial u/\partial x$. Aliasing can still arise, however, from "cubic" or higher-order nonlinearities.

To analyze the transform method, we proceed as follows. By analogy with (29.7), we can write

$$\widehat{\left(u\frac{\partial u}{\partial x}\right)}_k = \frac{1}{M}\sum_{j=1}^{M}\left\{\left[u\left(x_j\right)\frac{\partial u}{\partial x}\left(x_j\right)\right]e^{-ikx_j}\right\},\ -n\le k\le n\,. \tag{29.32}$$

Here we are computing the spectral coefficients of the *product* $u\partial u/\partial x$, by starting from $u\left(x_j\right)\partial u/\partial x\left(x_j\right)$, which is the grid-point representation of the same quantity. Using (29.6) and (29.8), we can express $u\left(x_j\right)$ and $\partial u/\partial x\left(x_j\right)$ in terms of Fourier series:

$$\widehat{\left(u\frac{\partial u}{\partial x}\right)}_k = \frac{1}{M}\sum_{j=1}^{M}\left[\left(\sum_{l=-n}^{n}\hat{u}_l e^{ilx_j}\right)\left(\sum_{m=-n}^{n}im\hat{u}_m e^{imx_j}\right)e^{-ikx_j}\right], -n\le k\le n\,. \tag{29.33}$$

It is important to note that, in (29.33), *the derivative, i.e.,* $\partial u/\partial x$*, has been computed using the spectral method*, rather than a grid-point method. The grid is being used to allow efficient implementation of the spectral method. Eq. (29.33) is analogous to (29.9).

When expanded, each term on the right-hand side of (29.33) involves the product of three Fourier modes (k, l, and m), and therefore includes zonal wave numbers up to $\pm 3n$. We need $3n+1$ complex coefficients to encompass wave numbers up to $\pm 3n$. Because $u\partial u/\partial x$ is real, those $3n+1$ complex coefficients actually correspond to $3n+1$ independent real numbers. Therefore, we need

$$M \ge 3n+1 \tag{29.34}$$

grid points to represent $u\partial u/\partial x$ exactly, up to wave number n. This is about 50% more than the $2n+1$ grid points needed to represent u itself exactly up to wave number n. Obviously, the number of grid points needed, i.e., $3n+1$, scales linearly (not quadratically) with n.

Because wave numbers higher that $\pm n$ are not included in (29.33), aliasing does not occur. This eliminates the possibility of aliasing instability. The grid with $3n+1$ points is sometimes called a "non-aliasing" grid.

In practice, the transform method to solve (29.27) works as follows:

1. Initialize the spectral coefficients $\hat{u}_k$, for $-n\le k\le n$. Of course, this would normally be done by using measurements of u to initialize u on a grid, and then using a transform to obtain the spectral coefficients.
2. Evaluate both u and $\partial u/\partial x$ *on a grid* with M points, where $M\ge 3n+1$. Here $\frac{\partial u}{\partial x}$ is computed *using the spectral method*, i.e., Eq. (29.8), and the result obtained is used to compute grid-point values of $\partial u/\partial x$.

3. Form the product $u\partial u/\partial x$ *on the grid.*
4. Using (29.33), transform $u\partial u/\partial x$ back into wave-number space, for $-n \le k \le n$. This gives the coefficients $(\widehat{u\partial u/\partial x})_k$.
5. Predict new values of the $\hat{u}_k$, using $d\hat{u}_k/dt = -(\widehat{u\partial u/\partial x})_k$.
6. Return to Step 2, and repeat this cycle as many times as desired.

Note that the grid-point representation of u contains more information ($3n+1$ real values) than the spectral representation ($2n+1$ real values). For this one-dimensional example the ratio is approximately 3/2. The additional information embodied in the grid-point representation is *thrown away* in Step 4 above, when we transform from the grid back into wave-number space. Therefore, the additional information is not "remembered" from one time step to the next. In effect, we throw away about 1/3 of the information that is represented on the grid. This is the price that we pay to avoid errors (for wave numbers up to $\pm n$) in the evaluation of quadratic nonlinearities.

As described above, the transform method uses a grid to evaluate quadratic nonlinearities. The same grid is used to implement complicated and often highly nonlinear physical parameterizations.

The transform method was revolutionary; it made spectral models a practical possibility.

29.9 Spectral methods on the sphere

29.9.1 Spherical harmonics

Spectral methods on the sphere were first advocated by Silberman (1954). A function F that is defined on the sphere can be represented by

$$F(\lambda, \varphi) = \sum_{m=-\infty}^{\infty} \sum_{n=|m|}^{\infty} F_n^m Y_n^m(\lambda, \mu), \tag{29.35}$$

where the

$$Y_n^m(\lambda, \mu) = e^{im\lambda} P_n^m(\mu) \tag{29.36}$$

are spherical harmonics (see Appendix D),

$$\mu \equiv \sin\varphi \,, \tag{29.37}$$

and the $P_n^m(\mu)$ are the associated Legendre functions of the first kind, which are given by

$$\begin{aligned} P_n^m(\sin\varphi) = \frac{(2n)!}{2^n n!\,(n-m)!}\left(1-x^2\right)^{m/2}\Bigg[\mu^{n-m} - \frac{(n-m)(n-m-1)}{2(2n-1)}\mu^{n-m-2} \\ + \frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2\cdot 4(2n-1)(2n-3)}\mu^{n-m-4} - \cdots\Bigg] \,. \end{aligned} \tag{29.38}$$

Here m is the zonal wave number and $n-m$ is the "meridional nodal number." As discussed in Appendix D, we require

$$n \geq m \,. \tag{29.39}$$

The spherical harmonics ${Y_n}^m$ are the eigenfunctions of the Laplacian on the sphere:

$$\boxed{\nabla^2 Y_n^m = \frac{-n(n+1)}{a^2} Y_n^m} \,. \tag{29.40}$$

Here a is the radius of the sphere. See Appendix D for further explanation.

We can approximate F by a truncated sum:

$$\overline{F} = \sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} F_n^m Y_n^m . \tag{29.41}$$

Here the overbar indicates that $\overline{F}$ is an approximation to F. In (29.41), the sum over m from $-M$ to M ensures that $\overline{F}$ is real. The choice of $N(m)$ is discussed below. For smooth F, $\overline{F}$ converges to F very quickly, in the sense that the root-mean-square error decreases quickly towards zero.

Why should we expand our variables in terms of the eigenfunctions of the Laplacian on the sphere? The Fourier representation discussed earlier is also based on the eigenfunctions

of the Laplacian, in just one dimension, i.e., sines and cosines. There are infinitely many differential operators. What is so special about the Laplacian? A justification is that:

- The Laplacian can be defined without reference to any coordinate system;
- The Laplacian consumes scalars and returns scalars, unlike, for example, the gradient, the curl, or the divergence;
- The Laplacian is isotropic, i.e., it does not favor any particular direction on the sphere;
- The Laplacian is simple.

29.9.2 Truncation

How should we choose $N(m)$? This is called the problem of truncation. The two best-known possibilities are *rhomboidal truncation* and *triangular truncation*. In both cases, we can choose the value of M, i.e., the highest zonal wave number to be included in the model, and the value of N follows.

$$\text{Rhomboidal truncation: } N = M + |m| \text{, and} \tag{29.42}$$

$$\text{Triangular truncation: } N = M \text{ .} \tag{29.43}$$

Rhomboidal and triangular truncation are illustrated in Fig. 29.2. With rhomboidal truncation, the value of N (the maximum value of n to be included in the sum) increases with the value of m, in such a way that the highest meridional nodal number is the same for all values of m. In the case of triangular truncation, the "two-dimensional wave number" N is a the same for all of the spherical harmonics that are included in the model, so the highest meridional nodal number is smaller for the larger values of m.

Fig. 29.2 also shows that, for a given value of M, rhomboidal truncation includes more spectral coefficients than triangular truncation. The numbers of complex coefficients needed are

$$(M+1)^2 + M^2 + M \text{ for rhomboidal truncation ,} \tag{29.44}$$

and

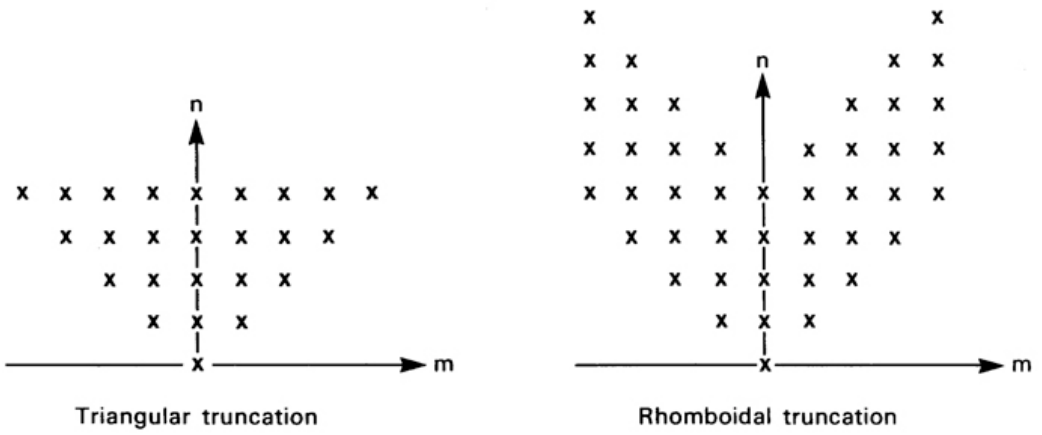


Figure 29.2: Rhomboidal and triangular truncation, both illustrated for the case $M = 4$. From Jarraud and Simmons (1983). In both cases, $n \geq m$. For a given value of M, rhomboidal truncation includes more spherical harmonics than triangular truncation.

$$(M+1)^2 \text{ for triangular truncation} . \tag{29.45}$$

You can count the Xs in Fig. 29.2 to persuade yourself that these formulas are correct.

Finally, the figure shows that, with both types of truncation, choosing the value of M is enough to determine which spectral coefficients are included. It is conventional to designate the resolution of a spectral model by designating the type of truncation by "R" or "T," and appending the value of M, as in "T106." The equivalent grid-point resolution of a spectral model is discussed later in this chapter.

Triangular truncation has the beautiful property that it is not tied to a coordinate system, in the following sense: In order to actually perform a spherical harmonic transform, it is necessary to adopt a spherical coordinate system (λ, φ). There are, of course, infinitely many such systems, which differ in the orientations of their poles. There is no reason, in principle, that the coordinates have to be chosen in the conventional way, in which the poles of the coordinate system coincide with the Earth's poles of rotation. The choice of a particular spherical coordinate system is, therefore, somewhat arbitrary. Suppose that we are given an analytical function on the sphere. We choose two different spherical coordinate systems (tilted with respect to one another in an arbitrary way), perform a *triangularly truncated* expansion in both, and then transform the results back to physical space. It can be shown that the two results will be identical, i.e.,

$$\overline{F}(\lambda_1, \varphi_1) = \overline{F}(\lambda_2, \varphi_2) , \tag{29.46}$$

where the subscripts indicate alternative spherical coordinate systems. This means that the arbitrary orientations of the spherical coordinate systems used have no effect whatsoever on the results obtained. The coordinate system used "disappears" at the end. Triangular truncation is very widely used today, in part because of this beautiful property, which is not shared by rhomboidal truncation.

As shown in Fig. 29.3, triangular truncation represents the *observed* kinetic energy spectrum more efficiently than does rhomboidal truncation (Baer, 1972). The thick lines in the figure show the observed kinetic energy percentage that comes from each component. The thin, straight, diagonal lines show the modes kept with triangular truncation. It looks like nature uses triangular truncation! With rhomboidal truncation the thin lines would be horizontal, and more modes would be kept, but they would not add much useful information.

29.9.3 Spherical harmonic transforms

In order to use (29.41) we need a "spherical harmonic transform," analogous to a Fourier transform. From (29.36), we see that a spherical harmonic transform is equivalent to the combination of a Fourier transform and a Legendre transform. The Legendre transform is formulated using a classical method called "Gaussian quadrature." The idea is as follows. Suppose that we are given a function $f(x)$ defined on the interval $-1 \leq x \leq 1$, and we wish to evaluate

$$I = \int_{-1}^{1} F(x)\, dx, \tag{29.47}$$

by a numerical method. If $f(x)$ is defined at a finite number of points, denoted by x_j, then

$$I \cong \sum_{i=1}^{N} f(x_i)\, w_i\ , \tag{29.48}$$

where the w_i are "weights."

Consider the special case in which $f(x)$ is itself a weighted sum of Legendre polynomials, as in a Legendre transform. We want the transform to be "exact," within the round-off error of the machine, so that we can recover $f(x)$ without error. Gauss showed that for such a case (29.46) gives the *exact* value of I, provided that the x_i are chosen to be the

roots of the highest Legendre polynomial used, i.e., the latitudes where the highest Legendre polynomial passes through zero. In other words, we can use (29.46) to evaluate the integral (29.45) *exactly*, provided that we choose the latitudes so that they are the roots of the highest Legendre polynomial used. These latitudes can be found by a variety of iterative methods, and of course this only has to be done once, before the model is run. The Gaussian quadrature algorithm is used to perform the Legendre transform. A grid that uses these "Gaussian latitudes" is called a "Gaussian grid."

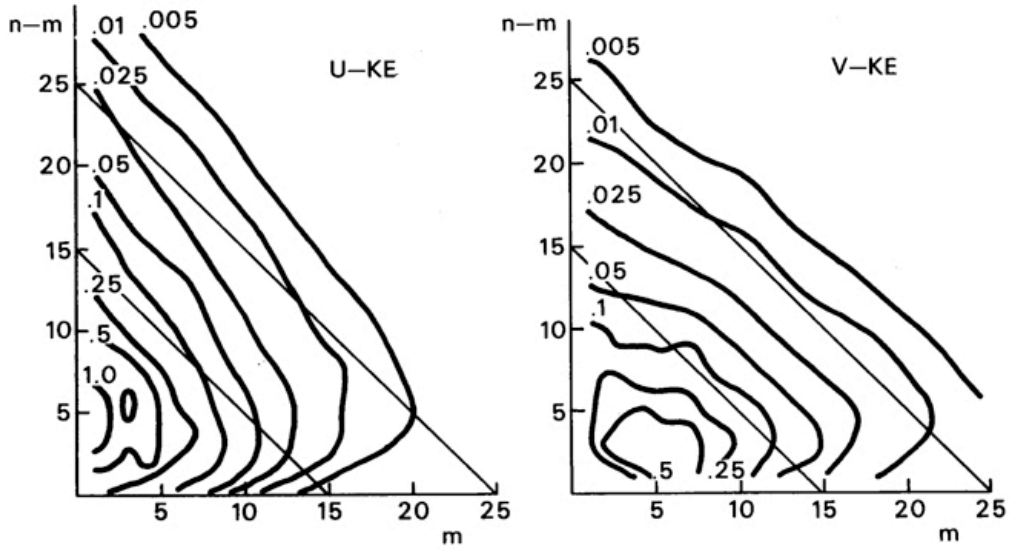


Figure 29.3: Percentage of total kinetic energy in each spectral component. From Jarraud and Simmons (1983), based on Baer (1972).

With the transform method described earlier, the number of grid points needed to avoid errors in the evalution of quadratic nonlinearities exceeds the number of degrees of freedom in the spectral representation. The number of grid points around a latitude circle must be $\geq 3M+1$. The number of latitude circles must be $\geq \frac{(3M+1)}{2}$ for triangular truncation, and so the total number of grid points needed is $\geq \frac{(3M+1)^2}{2}$. Referring back to (29.48), we see that, for large M, the grid representation uses about 2.25 times as many equivalent real numbers as the triangularly truncated spectral representation. A similar conclusion holds for rhomboidal truncation. The physics is often computed on a "nonaliasing grid," but doing so is wasteful.

29.9.4 How it works

In summary, the spectral transform method as applied to global models works as follows:

First, we choose a spectral truncation, e.g., T42. Then we identify the number of grid points needed in the longitudinal and latitudinal directions, perhaps with a view to avoiding aliasing due to quadratic nonlinearities. Next, we identify the highest degree Legendre

polynomial needed with the chosen spectral truncation, and find the latitudes where the roots of that polynomial occur. At this point, we can set up our Gaussian grid.

The horizontal derivatives are evaluated in the spectral domain, essentially through "multiplication by wave number." When we transform from the spectral domain to the grid, we combine an inverse fast Fourier transform (in the zonal direction) with an inverse Legendre transform (in the meridional direction). The nonlinear terms and the model physics are computed on the grid. Then we use the Legendre and Fourier transforms to return to the spectral domain. The basic logic of this procedure is very similar to that described earlier for the simple one-dimensional case.

We have a fast Fourier transform, but no one has yet discovered a "fast Legendre transform," although some recent work points towards one. Lacking a fast Legendre transform, the operation count for a spectral model is of $\mathcal{O}\left(N^3\right)$, where N is the number of spherical harmonics used. Finite-difference methods are, in effect, of $\mathcal{O}\left(N^2\right)$. This means that as the resolution increaess, spectral models become increasingly expensive, relative to grid-point models. Further comments are given later in this chapter.

29.10 Semi-implicit time differencing

As we have already discussed in Chapters 19 and28, gravity waves limit the time step that can be used in a primitive-equation (or shallow water) model. A way to get around this is to use semi-implicit time differencing, in which the gravity-wave terms of the equations are treated implicitly, while the other terms are treated explicitly. This can be accomplished much more easily in a spectral model than in a finite-difference model.

A detailed discussion of this approach will not be given here, but the basic ideas are as follows. The relevant terms are the pressure-gradient terms of the horizontal equations of motion, and the mass convergence term of the continuity equation. These are the same terms that we focused on in the discussion of the pole problem, in Chapter 28. The terms involve horizontal derivatives of the "height field" and the winds, respectively. Typically the Coriolis terms are also included, so that the waves in question are inertia-gravity waves.

Consider a finite-difference model. If we implicitly difference the gravity-wave terms, the resulting equations will involve the "$n+1$" time-level values of the heights and the winds at multiple grid points in the horizontal. This means that we must solve simultaneously for the "new" values of the heights and winds. This can be done, of course, but it can be computationally expensive, especially on very large grids. For this reason, most finite-difference models do not use semi-implicit time differencing.

In spectral models, on the other hand, we prognose the spectral coefficients of the heights and winds, and so we can apply the gradient and divergence operators simply by multiplying by wave number (roughly speaking). This is a "local" operation in wave-number space, so it is not necessary to solve a system of simultaneous equations.

The use of semi-implicit time differencing allows spectral models to take time steps several times longer than those of (explicit) grid-point models. This is a major advantage in terms of computational speed, which compensates, to some extent, for the expense of the spectral transform.

29.11 Conservation properties and computational stability

Because the spectral transform method prevents aliasing for quadratic nonlinearities, but not cubic nonlinearities, spectral models are formulated so that the highest nonlinearities that appear in the equations (other than in the physical parameterizations) are quadratic. This means that the equations must be written in advective form, rather than flux form. As a result, spectral models do not exactly conserve anything – not even mass – for a general, divergent flow.

It can be shown, however, that in the limit of two-dimensional non-divergent flow, spectral models do conserve both kinetic energy and enstrophy. Because of this property, they are well behaved computationally. Nevertheless, all spectral models need some artificial diffusive damping to avoid computational instability. In contrast, it is possible to formulate finite-difference models that are very highly conservative and can run indefinitely with no damping at all.

29.12 The accuracy of spectral models

With either triangular or rhomboidal truncation, the resolution of a spectral model increases as n increases. How does the accuracy of the spectral representation of a field increase as n increases? This is analogous to asking how the accuracy a grid-point representation of a field increases as the number of grid points increases. As discussed by Orszag (1974) and Jarraud and Simmons (1983), the root-mean-square error of the spectral representation of an infinitely differentiable field decreases to zero faster than any finite power of $1/n$. At first, this impressive rate of spectral convergence suggested that for a given level of accuracy spectral models could use fewer prognostic degrees of freedom than finite-difference models. Experience has not supported this idea; the flaws in the argument are that the atmosphere is subjected to discontinuous and "spikey" boundary conditions, such as topography, and the atmosphere itself produces nearly discontinuous fields associated with fronts and precipitating cloud systems. Spectral truncation of these not-so-differentiable fields produces oscillations that are often called "spectral ringing." In fact, in the vicinity of discontinuities such as sharp mountains (most famously the Andes) the spectral method produces errors that do not go to zero as $n \to \infty$. This is called Gibbs' phenomenon (Gottlieb and Shu, 1997).

Recall that in many models, e.g., those that use the σ coordinate, it is necessary to take horizontal derivatives of the terrain height in order to evaluate the horizontal pressure gradient force. The terrain heights have to be expanded to compute spectral coefficients, and

of course the expansion is truncated at some finite wave number. Given spikey topography and flat oceans as input, spectral truncation can lead to "bumpy" oceans (Fig. 29.4). Various approaches have been suggested to alleviate this problem (Hoskins, 1980; Navarra et al., 1994; Bouteloup, 1995; Holzer, 1996; Lindberg and Broccoli, 1996).

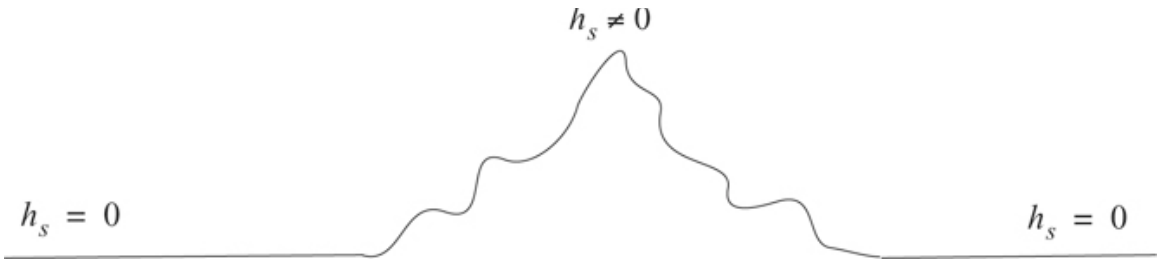


Figure 29.4: The Earth is bumpy.

The question remains, "What is the equivalent grid-spacing of a spectral model?" Laprise (1992b) offered four possible answers:

1. One might argue that the effective grid spacing of a spectral model is *the average distance between latitudes on the Gaussian grid*. With triangular truncation, this is the same as the spacing between longitudes at the Equator, which is $L_1 = \frac{2\pi a}{3M+1}$. Given the radius of the Earth, and using units of thousands of kilometers, this is equivalent to $13.5/M$. For a T31 model (with $M = 31$), we get $L_1 \cong 425$ km. An objection to this measure is that, as discussed above, much of the information on the Gaussian grid is thrown away when we transform back into spectral space.

2. A second possible measure of resolution is *half the wavelength of the shortest resolved zonal wave at the Equator*, which is $L_2 = \frac{\pi a}{M}$, or about $20/M$ in units of thousands of kilometers. For a T31 model, $L_2 \cong 650$ km.

3. A third method is based on the idea that the spectral coefficients, which are the prognostic variables of the spectral model, can be thought of *as a certain number of real variables per unit area*, distributed over the Earth. A triangularly truncated model has the equivalent of $(M+1)^2$ real coefficients. The corresponding resolution is then $L_3 = \sqrt{\frac{4\pi a^2}{(M+1)^2}} = \frac{2\sqrt{\pi}a}{M+1}$, which works out to about 725 km for a T31 model.

4. A fourth measure of resolution is based on the *equivalent total wave number associated with the Laplacian operator*, for the highest mode. The square of this total wave number is $K^2 = \frac{M(M+1)}{2a^2}$. Suppose that we equate this to the square of the equivalent total wave number on a square grid, i.e. $K^2 = {k_x}^2 + {k_y}^2$, and let $k_x = k_y = k$ for simplicity. One half of the corresponding wavelength is $L_4 = \frac{\pi}{k} = \frac{\sqrt{2}\pi a}{M}$, which is equivalent to $28.3/M$ in units of thousands of kilometers. For a T31 model this gives about 900 km.

These four measures of spectral resolution range over more than a factor of two. The measure that makes a spectral model "look good" is L_1, and so it is not surprising that spectral modelers almost always use it when specifying the equivalent grid spacing of their models.

29.13 Physical parameterizations

Because most physical parameterizations are highly nonlinear, spectral models evaluate such things as convective heating rates, turbulent exchanges with the Earth's surface, and radiative transfer on their Gaussian grids. The tendencies due to these parameterizations are then applied to the prognostic variables, which are promptly transformed back into wave-number space.

Recall that when this transform is done, the spectral representation contains less information than is present on the grid, due to the spectral truncation used to avoid aliasing due to quadratic nonlinearities. This means that if the fields were immediately transformed back onto the grid (without any changes due, e.g., to advection), the physics would not "see" the fields that it had just finished with. Instead, it would see spectrally truncated versions of these fields.

For example, suppose that the physics package includes a convective adjustment that is supposed to modify the soundings of convectively unstable columns so as to remove the convective instability. Suppose further that on a certain time step this parameterization has done its work, removing all instability as seen on the Gaussian grid. After spectral truncation, some convective instability may re-appear, even though "physically" nothing has happened!

In effect, the spectral truncation that is inserted between the grid domain and the spectral domain prevents the physical parameterizations from doing their work properly. This is a problem for all spectral models. It is not an issue when the physics is evaluated on a "linear" grid that has the same number of degrees of freedom as the spectral representation.

29.14 Moisture advection

The mixing ratio of water vapor is non-negative. In Chapter 12, we discussed the possibility of spurious negative mixing ratios caused by dispersion errors in finite-difference schemes, and we also discussed the families of finite-difference advection schemes that are "sign-preserving" and do not suffer from this problem.

Spectral models have a very strong tendency to produce negative water vapor mixing ratios (e.g., Williamson and Rasch (1994)). In the global mean, the rate at which "negative water" is produced can be a significant fraction of the globally averaged precipitation rate. Negative water vapor mixing ratios can occur not only locally on individual time steps, but even in zonal averages that have been time-averaged over a month.

Because of this very serious problem, many spectral models are now using monotone semi-Lagrangian methods for advection (e.g. Williamson and Olson (1994)). This means that they are only "partly spectral."

29.15 Linear grids

When non-spectral methods are used to evaluate the nonlinear advection terms, the motivation for using the high-resolution, non-aliasing grid disappears. Such models can then use a coarser "linear grid," with the same number of grid points as the number of independent real coefficients in the spectral representation. The physics is of course evaluated on the same linear grid. Linear grids lead greatly reduce the computational cost of a model.

29.16 Reduced linear grids

The cost of the spherical-harmonic transforms can be reduced by decreasing the numbers of grid points around latitude circles, near the poles (Hortal and Simmons (1991)). Recall from Chapter 28 that a similar idea was tried with grid point methods. The approach works with spectral methods because with high resolution the spectral coefficients corresponding to the largest values of m are very close to zero near on the poles.

29.17 Summary

In summary, the spectral method has both strengths and weaknesses:

Strengths:

- Especially with triangular truncation, it eliminates the "pole problem" associated with wave propagation, although it does not elimiinate the pole problem for zonal advection.
- It gives the exact phase speeds for linear waves and advection by a constant current such as solid-body rotation.
- It converges very rapidly, in the sense that it can give good results with just a few modes.
- Semi-implicit time-differencing schemes are easily implemented in spectral models.

Weaknesses:

- Spectral models do not exactly conserve anything - not even mass.
- Partly because of failure to conserve the mass-weighted total energy, artificial damping is needed to maintain computational stability.
- Spectral models have bumpy oceans.

- Because of truncation in the transform method, physical parameterizations do not always have the intended effect.
- Moisture advection does not work well in the spectral domain.
- At high resolution, spectral methods are computationally expensive compared to grid point models.

29.18 Problems

1. Write subroutines to compute Fourier transforms and inverse transforms, for arbitrary complex $q\left(x_j\right)$. The number of waves to be included in the transform, and the number of grid points to be used in the inverse transform, should be set through the argument lists of subroutines. Let

$$q\left(x_j\right) = 14\cos\left(k_0 x_j\right) + 6i\cos\left(k_1 x_j\right) + 5, \tag{29.49}$$

where

$$\begin{aligned} k_0 &= \frac{2\pi}{L_0} \text{ and } L_0 = \frac{X}{4}, \\ k_1 &= \frac{2\pi}{L_1} \text{ and } L_1 = \frac{X}{8}. \end{aligned} \tag{29.50}$$

Here X is the size of the periodic domain. Compute the Fourier coefficients starting from values of x_j on a grid of M points, for $M = 3$, $M = 9$, $M = 17$, and $M = 101$. Discuss your results.

2. Consider a *periodic* step function $H(x)$. The step function is defined by

$$\begin{aligned} H(x) &= -1 \quad \text{when the integer part of } x \text{ is odd,} \\ H(x) &= +1 \quad \text{when the integer part of } x \text{ is even.} \end{aligned}$$

With this definition, $H(x)$ is discontinuous at integer points, and infinitely differentiable elsewhere. Let x take values in the range $0 \leq x \leq 4$. Compute approximate versions of $H(x)$ based on Fourier expansions up to wave numbers $n = 10$, 100, 1000, and 10000. Plot the approximate versions of $H(x)$, for each value of n, with special attention to the discontinuities. Make a table that shows both the root-mean-square error and the maximum absolute error as functions of n.

3. Prove that the a_j^l that appear in (29.18) are real numbers.

4. For the transform method with a one-dimensional problem, many grid points would be needed to give the exact answer up to wave number n for the case of *cubic* non-linearities?

5. You are given the values of u at 21 points, as listed below:

$$1,2,4,5,4,3,4,7,8,5,3,1,-1,-3,-4,-3,-4,-2,-1,-1,0$$

The points are 1000 km apart, and the domain is periodic. In the following sub-problems, use $n = 10$.

(a) List the spectral coefficients of u and $\partial u/\partial x$.

(b) Work out the numerical values of the finite-difference coefficients that appear in Eq. (29.17).

(c) Show a plot that compares $\partial u/\partial x$ obtained using the spectral method with the corresponding result based on second-order centered finite-differences.

(d) Compute $-u\partial u/\partial x$ with both the direct method (29.28) and the transform method. Do the two methods agree?

Chapter 30

Finite-Element Methods

McRae (2015); Maynard et al. (2020)

Like spectral methods, finite-element methods are used to approximate derivatives as weighted sums the derivatives of a finite set of basis functions. With spectral methods the basis functions are global (e.g., spherical harmonics), while with finite-element methods the basis functions are local. The domain of a model is partitioned into a finite set of non-overlapping subdomains, called "elements." The basis functions are defined on the elements, and can be either continous or discontinuous at points shared by neighboring elements.

Giraldo (2020)

Gassner and Winters (2021)

Souza et al. (2022)

$$a + b = c. \tag{30.1}$$

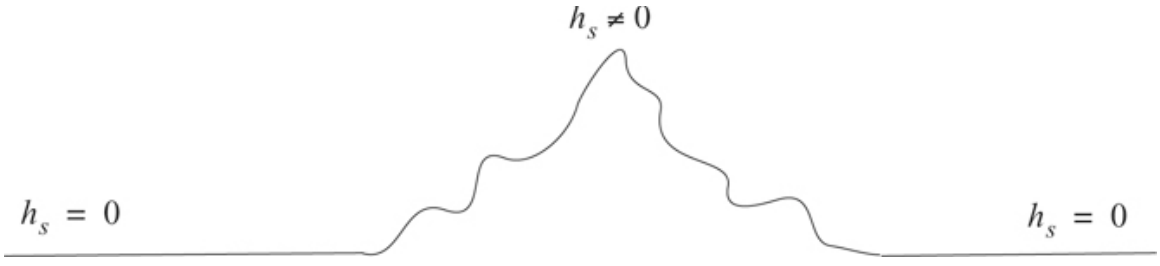


Figure 30.1: A figure about finite-element methods.

Our goal is

$$\begin{aligned}(2n+1)^2 - 2\left[\frac{n(n-1)}{2}\right] &= \left(4n^2+4n+1\right) - \left(n^2-n\right) \\ &= 3n^2+5n+1.\end{aligned} \tag{30.2}$$

Last bit of text.

30.1 Problems

1. Write subroutines to compute Fourier Let

$$a = b + c. \tag{30.3}$$

2. Consider a periodic step function, defined by

Chapter 31

Concluding discussion

The end

Appendix A

Vectors, Coordinates, and Coordinate Transformations

A.1 Physical laws and coordinate systems

For the present discussion, we define a "coordinate system" as a tool for describing positions in space. Coordinate systems are human inventions, and therefore are not part of physics, although they can be used to describe physics. For obvious reasons, spherical coordinates are particularly useful in geophysics.

Any physical law should be expressible in a form that makes no reference to any particular coordinate system; we certainly do not expect that the laws of physics change when we switch from spherical coordinates to cartesian coordinates! Nevertheless, it is useful to understand how physical laws can be expressed in different coordinate systems, and in particular how various quantities "transform" as we change from one coordinate system to another.

A.2 Scalars, vectors, and tensors

Tensors can be defined without reference to any particular coordinate system. A tensor is simply "out there," and has a meaning that is the same whether we happen to be working in spherical coordinates, or Cartesian coordinates, or whatever. Tensors are, therefore, just what we need to formulate physical laws.

The simplest kind of tensor, called a "tensor of rank 0," is a scalar, which is represented by a single number – essentially a magnitude with no direction. An example of a scalar is temperature. Not all quantities that are represented by a single number are scalars, because not all of them are defined without reference to any particular coordinate system. An example of a (single) number that is not a scalar is the longitudinal component of the wind, which is defined with respect to a particular coordinate system, i.e., spherical coordinates.

A scalar is expressed in exactly the same way regardless of what coordinate system may be in use to describe non-scalars in a problem. For example, if someone tells you

the temperature in Fort Collins, you don't have to ask whether they are using spherical coordinates or some other coordinate system, because it makes no difference at all.

Vectors are "tensors of rank 1;" a vector can be represented by a magnitude and one direction. An example is the wind vector. In atmospheric science, vectors are normally either three-dimensional or two-dimensional, but in principle they have any number of dimensions. A scalar can be considered to be a vector in a one-dimensional space.

A vector can be expressed in a particular coordinate system by an ordered list of numbers, which are called the "components" of the vector. The components have meaning only with respect to the particular coordinate system. More or less by definition, the number of components needed to describe a vector is equal to the number of dimensions in which the vector is "embedded."

We can define "unit vectors" that point in each of the coordinate directions. A vector can then be written as the vector sum of each of the unit vectors times the "component" associated with the unit vector. In general, the directions in which the unit vectors point depend on position.

Unit vectors are always non-dimensional; here we are using the word "dimension" to refer to physical quantities, such as length, time, and mass. Because the unit vectors are non-dimensional, all components of a vector must have the same dimensions as the vector itself.

Spatial coordinates may or may not have the dimensions of length. In the familiar Cartesian coordinate system, the three coordinates, (x, y, z), each have dimensions of length. In spherical coordinates, (λ, φ, r), where λ is longitude, φ is latitude, and r is distance from the origin, the first two coordinates are non-dimensional angles, while the third has the dimension of length.

When we change from one coordinate system to another, an arbitrary vector $\mathbf{V}$ transforms according to

$$\mathbf{V}' = \mathbf{M}\mathbf{V}. \tag{A.1}$$

Here $\mathbf{V}$ is the representation of the vector in the first coordinate system (i.e., $\mathbf{V}$ is the list of the components of the vector in the first coordinate system), $\mathbf{V}'$ is the representation the vector in the second coordinate system, and $\mathbf{M}$ is a "rotation matrix" that maps $\mathbf{V}$ onto $\mathbf{V}'$. The rotation matrix used to transform a vector from one coordinate system to another is a property of the two coordinate systems in question; it is the same for all vectors, but it *does* depend on the particular coordinate systems involved, so it is not a tensor.

The transformation rule (A.1) is actually part of the definition of a vector, i.e., a vector must, by definition, transform from one coordinate system to another via a rule of the form

(A.1) . It follows that not all ordered lists of numbers are vectors. For example, the list

(mass of the moon, distance from Fort Collins to Denver)

is not a vector.

Now let $\mathbf{V}$ be the a vector representing the three-dimensional velocity of a particle in the atmosphere. The Cartesian and spherical representations of are

$$\mathbf{V} = \dot{x}\,\mathbf{i} + \dot{y}\,\mathbf{j} + \dot{z}\,\mathbf{k} \tag{A.2}$$

$$\mathbf{V} = \dot{\lambda}\; r\cos\varphi\;\mathbf{e}_\lambda + r\;\dot{\varphi}\;\mathbf{e}_\varphi + \dot{r}\;\mathbf{e}_r \tag{A.3}$$

Here a "dot" denotes a Lagrangian time derivative, i.e., a time derivative following a moving particle, $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the cartesian coordinate system, and $\mathbf{e}_\lambda$, $\mathbf{e}_\varphi$, and $\mathbf{e}_r$ are unit vectors in the spherical coordinate system. Eqs. (A.2) and (A.3) both describe the same vector, $\mathbf{V}$, i.e., the meaning of $\mathbf{V}$ is independent of the coordinate system that is chosen to represent it.

Vectors are considered to be tensors of rank one, and scalars are tensors of rank zero. The number of directions associated with a tensor is called the "rank" of the tensor. In principle, the rank can be arbitrarily large, but in atmospheric science we rarely meet tensors with ranks higher than two.

A tensor of rank 2 that is important in atmospheric science is the advective flux of momentum, which has a magnitude and "two directions." One of the directions is associated with the advected momentum itewslf, and the other is associated with the direction in which the momentum is being transported. The advective flux of momentum can be written as $\rho\mathbf{V}\otimes\mathbf{V}$, where ρ is the density of the air, $\mathbf{V}$ is the wind vector, and $\otimes$ is the dyadic or outer product.

Like a vector, a tensor of rank 2 can be expressed in a particular coordinate system, i.e., we can define the "components" of the tensor with respect to a particular coordinate system. The components of a tensor of rank 2 can be arranged in the form of a two-dimensional matrix, in contrast to the components of a (column or row) vector, which form an ordered one-dimensional list. When we change from one coordinate system to another, a tensor of rank 2 transforms according to

$$\mathbf{T}' = \mathbf{M}\mathbf{T}\mathbf{M}^{-1} \tag{A.4}$$

where $\mathbf{T}$ is the representation of a rank-2 tensor in the first coordinate system, $\mathbf{T}'$ is the representation of the same tensor in the second coordinate system, $\mathbf{M}$ is the matrix introduced in Eq. (1) above, and $\mathbf{M}^{-1}$ is its inverse.

A.3 Differential operators

Several familiar differential operators can be defined without reference to any coordinate system. This makes them more fundamental than, for example, $\partial/\partial x$, where x is a particular spatial coordinate. The coordinate-independent operators that we need most often for atmospheric science (and for most other branches of physics) are:

- the gradient, denoted by $\nabla\alpha$, where α is an arbitrary scalar;
- the divergence, denoted by $\nabla\cdot\mathbf{V}$, where $\mathbf{V}$ is an arbitrary vector;
- the curl, denoted by $\nabla\times\mathbf{V}$, and
- the Laplacian, given by $\nabla^2 A = \nabla\cdot(\nabla A)$.

Note that the gradient and curl are vectors, while the divergence is a scalar. The gradient operator accepts scalars as "input," while the divergence and curl operators consume vectors.

In discussions of two-dimensional motion, it is often convenient to introduce an additional operator called the Jacobian, denoted by

$$\begin{aligned} J(\alpha,\beta) \equiv &\quad \mathbf{k}\cdot(\nabla\alpha\times\nabla\beta) \\ = &\quad \mathbf{k}\cdot\nabla\times(\alpha\nabla\beta) \\ = &-\mathbf{k}\cdot\nabla\times(\beta\nabla\alpha)\,. \end{aligned} \tag{A.5}$$

Here the gradient operators are understood to produce vectors in the two-dimensional space, α and β are arbitrary scalars, and $\mathbf{k}$ is a unit vector perpendicular to the two-dimensional surface. The second and third lines of (A.5) can be derived with the use of vector identities found in a table later in this appendix.

A definition of the gradient operator that does not make reference to any coordinate system is:

$$\nabla\alpha \equiv \lim_{V\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\alpha\, dS\right]\,, \tag{A.6}$$

where S is the surface bounding a volume V, and $\mathbf{n}$ is the outward normal on S. Here the terms "volume" and "bounding surface" are used in the following generalized sense: In a

three-dimensional space, "volume" is literally a volume, and "bounding surface" is literally a surface. In a two-dimensional space, "volume" means an area, and "bounding surface" means the curve bounding the area. In a one-dimensional space, "volume" means a curve, and "bounding surface" means the end points of the curve. The limit in (A.6) is one in which the volume and the area of its bounding surface shrink to zero.

As an example, consider a Cartesian coordinate system on a plane, with unit vectors **i** and **j** in the x and y directions, respectively. Consider a "box" of width Δx and height Δy, as shown in Figure A.1. We can write

$$\begin{aligned}\nabla A \equiv \lim_{(\Delta x,\Delta y)\to 0} \Bigg\{ \frac{1}{\Delta x \Delta y} \Bigg[& A\left(x_0+\frac{\Delta x}{2}, y_0\right)\Delta y\mathbf{i} + A\left(x_0, y_0+\frac{\Delta y}{2}\right)\Delta x\mathbf{j} \\ & - A\left(x_0-\frac{\Delta x}{2}, y_0\right)\Delta y\mathbf{i} - A\left(x_0, y_0-\frac{\Delta y}{2}\right)\Delta x\mathbf{j}\Bigg]\Bigg\} \\ & = \frac{\partial A}{\partial x}\mathbf{i} + \frac{\partial A}{\partial y}\mathbf{j}\,.\end{aligned} \tag{A.7}$$

This is the answer that we expect.

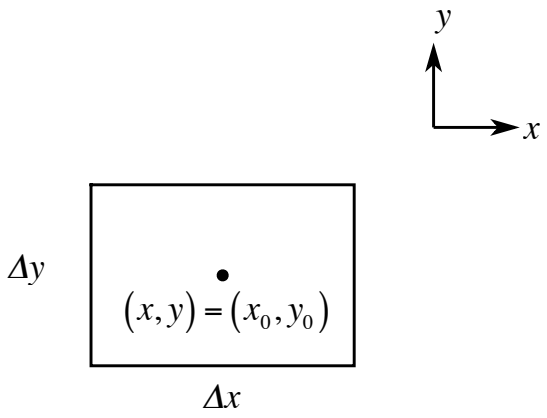


Figure A.1: A rectangular box in a planar two-dimensional space, with center at (x_0, y_0), width Δx, and height Δy.

Definitions of the divergence and curl operators that do not make reference to any coordinate system are:

$$\nabla \cdot \mathbf{Q} \equiv \lim_{V\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\cdot\mathbf{Q}\,dS\right] \tag{A.8}$$

$$\nabla \times \mathbf{Q} \equiv \lim_{V \to 0} \left[\frac{1}{V} \oint_S \mathbf{n} \times \mathbf{Q} dS \right] \tag{A.9}$$

It is possible to work through exercises similar to (A.7) for these operators too. You might want to try it yourself, to see if you understand.

Finally, the Jacobian on a two-dimensional surface can be defined by

$$J(\alpha, \beta) = \lim_{C \to 0} \left[\oint_C \alpha \nabla \beta \cdot \mathbf{t} dl \right], \tag{A.10}$$

where $\mathbf{t}$ is a unit vector that is tangent to the bounding curve C.

A.4 Vector identities

Many useful identities relate the divergence, curl, and gradient operators. Most of the following identities can be found in any mathematics reference manual, e.g., Beyer (1984). As before, let α and β be arbitrary scalars, let $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ be arbitrary vectors, and let $\mathbf{T}$ be an arbitrary tensor of rank 2. Then:

$$\nabla \times (\nabla \alpha) = 0 \tag{A.11}$$

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0 \tag{A.12}$$

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \tag{A.13}$$

$$\nabla \cdot (\alpha \mathbf{A}) = \alpha (\nabla \cdot \mathbf{A}) + \mathbf{A} \cdot \nabla \alpha \tag{A.14}$$

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = (\nabla \times \mathbf{A}) \cdot \mathbf{B} - (\nabla \times \mathbf{B}) \cdot \mathbf{A} \tag{A.15}$$

$$\nabla \times (\alpha \mathbf{A}) = \nabla \alpha \times \mathbf{A} + \alpha (\nabla \times \mathbf{A}) \tag{A.16}$$

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) \tag{A.17}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} (\mathbf{C} \cdot \mathbf{A}) - \mathbf{C} (\mathbf{A} \cdot \mathbf{B}) \tag{A.18}$$

$$\nabla \times (\mathbf{A} \times \mathbf{B}) = \mathbf{A} (\nabla \cdot \mathbf{B}) - \mathbf{B} (\nabla \cdot \mathbf{A}) - (\mathbf{A} \cdot \nabla) \mathbf{B} + (\mathbf{B} \cdot \nabla) \mathbf{A} \tag{A.19}$$

$$\nabla (\mathbf{A} \cdot \mathbf{B}) = (\mathbf{A} \cdot \nabla) \mathbf{B} + (\mathbf{B} \cdot \nabla) \mathbf{A} + \mathbf{A} \times (\nabla \times \mathbf{B}) + \mathbf{B} \times (\nabla \times \mathbf{A}) \tag{A.20}$$

$$\begin{aligned} J(\alpha, \beta) \equiv \mathbf{k} \cdot (\nabla \alpha \times \nabla \beta) = &\quad \mathbf{k} \cdot \nabla \times (\alpha \nabla \beta) \\ &= -\mathbf{k} \cdot \nabla \times (\beta \nabla \alpha) \\ &= -\mathbf{k} \cdot (\nabla \beta \times \nabla \alpha) \end{aligned} \tag{A.21}$$

$$\nabla^2 \mathbf{A} \equiv (\nabla \cdot \nabla) \mathbf{A} = \nabla (\nabla \cdot \mathbf{A}) - \nabla \times (\nabla \times \mathbf{A}) \tag{A.22}$$

$$\nabla \cdot (\mathbf{A} \otimes \mathbf{B}) = (\mathbf{A} \cdot \nabla) \mathbf{B} + (\mathbf{B} \cdot \nabla) \mathbf{A} \tag{A.23}$$

$$\nabla \cdot (\alpha \mathbf{T}) = (\nabla \alpha) \cdot \mathbf{T} + \alpha (\nabla \cdot \mathbf{T}) \tag{A.24}$$

In (A.23), $\otimes$ denotes the outer or dyadic product of two vectors, which yields a tensor of rank 2.

A special case of (A.20) is

$$\frac{1}{2} \nabla (\mathbf{A} \cdot \mathbf{A}) = (\mathbf{A} \cdot \nabla) \mathbf{A} + \mathbf{A} \times (\nabla \times \mathbf{A}) \tag{A.25}$$

This identity is used to write the advection terms of the momentum equation in alternative forms.

Identity (A.22) says that the Laplacian of a vector is the gradient of the divergence of the vector, minus the curl of the curl of the vector. The first term involves only the divergent part of the wind field, and the second term involves only the rotational part. Eq. (A.22) can be used, for example, in a parameterization of momentum diffusion.

A.5 Spherical coordinates

A.5.1 Vector operators in spherical coordinates

The gradient, divergence, curl, Laplacian, and Jacobian operators can be expressed in spherical coordinates as follows:

$$\nabla \alpha = \left(\frac{1}{r \cos \varphi} \frac{\partial \alpha}{\partial \lambda}, \frac{1}{r} \frac{\partial \alpha}{\partial \varphi}, \frac{\partial \alpha}{\partial r} \right) , \tag{A.26}$$

$$\nabla \cdot \mathbf{A} = \frac{1}{r \cos \varphi} \frac{\partial A_\lambda}{\partial \lambda} + \frac{1}{r \cos \varphi} \frac{\partial}{\partial \varphi} \left(A_\varphi \cos \varphi \right) + \frac{1}{r^2} \frac{\partial}{\partial r} \left(A_r r^2 \right) , \tag{A.27}$$

$$\nabla \times \mathbf{A} = \left\{ \frac{1}{r} \left[\frac{\partial A_r}{\partial \varphi} - \frac{\partial}{\partial r} \left(r A_\varphi \right) \right], \frac{1}{r} \frac{\partial}{\partial r} \left(r A_\lambda \right) - \frac{1}{r \cos \varphi} \frac{\partial A_r}{\partial \lambda}, \frac{1}{r \cos \varphi} \left[\frac{\partial A_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi} \left(A_\lambda \cos \varphi \right) \right] \right\} . \tag{A.28}$$

$$\nabla^2\alpha = \frac{1}{r^2\cos^2\varphi}\frac{\partial^2\alpha}{\partial\lambda^2} + \frac{1}{r^2\cos\varphi}\frac{\partial}{\partial\varphi}\left(\frac{\partial\alpha}{\partial\varphi}\cos\varphi\right) + \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\alpha}{\partial r}\right) , \tag{A.29}$$

$$J(\alpha,\beta) = \frac{1}{r^2\cos\varphi}\left(\frac{\partial\alpha}{\partial\lambda}\frac{\partial\beta}{\partial\varphi} - \frac{\partial\beta}{\partial\lambda}\frac{\partial\alpha}{\partial\varphi}\right) , \tag{A.30}$$

The divergence operator can be expanded as

$$\nabla\cdot\mathbf{A} = \frac{1}{r\cos\varphi}\frac{\partial A_\lambda}{\partial\lambda} + \frac{1}{r\cos\varphi}\frac{\partial}{\partial\varphi}\left(A_\varphi\cos\varphi\right) + \frac{\partial A_r}{\partial r} + \frac{2A_r}{r} . \tag{A.31}$$

Because the Earth's atmosphere is very thin compared to the radius of the Earth, the last term of (A.31) is negligible, and the divergence operator can be approximated by

$$\nabla\cdot\mathbf{A} \cong \frac{1}{a\cos\varphi}\frac{\partial A_\lambda}{\partial\lambda} + \frac{1}{a\cos\varphi}\frac{\partial}{\partial\varphi}\left(A_\varphi\cos\varphi\right) + \frac{\partial A_r}{\partial r} . \tag{A.32}$$

Note that r has been replaced by a in the first two terms. In this book, we normally use (A.32) rather than (A.31), largely because it is conventional to do so, although it is not clear that the approximation (A.32) actually makes our work simpler. The approximation would not be applicable to a deep atmosphere, such as that of a star, or of Jupiter.

A.5.2 Horizontal and vertical vectors in spherical coordinates

The unit vectors in spherical coordinates are denoted by $\mathbf{e}_\lambda$ pointing towards the east, $\mathbf{e}_\varphi$ pointing towards the north, and $\mathbf{e}_r$ pointing outward from the origin (in geophysics, outward from the center of the Earth).

A useful result that is a special case of (A.19) is

$$\mathbf{e}_r\cdot[\nabla\times(\alpha\mathbf{e}_r\times\mathbf{H})] = \alpha\nabla\cdot\mathbf{H}, \tag{A.33}$$

where $\mathbf{H}$ is an arbitrary horizontal vector. Similarly, a useful special case of (A.15) is

$$\nabla \cdot (\alpha \mathbf{e}_r \times \mathbf{H}) = -\alpha \mathbf{e}_r \cdot (\nabla \times \mathbf{H}) \tag{A.34}$$

If a three-dimensional vector $\mathbf{V}$ is separated into a horizontal vector and a vertical vector, as in

$$\mathbf{V} = \mathbf{v} + w\mathbf{e}_r, \tag{A.35}$$

then (A.28)) can be written as

$$\boxed{\nabla \times (\mathbf{v} + w\mathbf{e}_r) = \nabla_r \times \mathbf{v} + \mathbf{e}_r \times \left[\frac{1}{r}\frac{\partial}{\partial r}(r\mathbf{v}) - \nabla_r w\right].} \tag{A.36}$$

In case $\mathbf{V}$ is the velocity, the first term on the right-hand side of (A.36) is the vertical component of the vorticity, and the second term is the horizontal vorticity vector. Eq. (A.36) shows that the curl of a purely vertical vector is minus $\mathbf{e}_r$ crossed with the horizontal gradient of the magnitude of that vector. The three-dimensional curl of a purely horizontal vector has both a vertical part, given by $\nabla_r \times \mathbf{v}$, and a horizontal part, given by $\mathbf{e}_r \times \left[\frac{1}{r}\frac{\partial}{\partial r}(r\mathbf{v}) - \nabla_r w\right]$. The *two-dimensional* curl of a horizontal vector has only a vertical component, namely $\nabla_r \times \mathbf{v}$.

Finally, the 3D curl of the 3D curl of a 3D vector is given by

$$\nabla \times [\nabla \times (\mathbf{v} + w\mathbf{e}_r)] = \nabla_r \times \boldsymbol{\eta} + \mathbf{e}_r \times \left[\frac{1}{r}\frac{\partial}{\partial r}(r\boldsymbol{\eta}) - \nabla_r \zeta\right], \tag{A.37}$$

where

$$\boldsymbol{\eta} \equiv \mathbf{e}_r \times \left[\frac{1}{r}\frac{\partial}{\partial r}(r\mathbf{v}) - \nabla_r w\right], \tag{A.38}$$

and

$$\zeta \equiv \mathbf{e}_r \cdot (\nabla_r \times \mathbf{v}). \tag{A.39}$$

A.5.3 Derivation of the gradient operator in spherical coordinates

Consider how the two-dimensional version of (A.26) can be derived from (A.6). Figure A.2 illustrates the problem. Here we have replaced r by a, the radius of the Earth. The angle θ depicted in the figure arises from the gradual rotation of $\mathbf{e}_\lambda$ and $\mathbf{e}_\varphi$, the unit vectors associated with the spherical coordinates, as the longitude changes; the directions of $\mathbf{e}_\lambda$ and $\mathbf{e}_\varphi$ in the center of the area element, where ∇A is defined, are different from their respective directions on either east-west wall of the area element. Inspection of Figure A.2 shows that θ satisfies

$$\begin{aligned}\sin\theta &= \frac{-\frac{1}{2}\left[a\cos\left(\varphi+d\varphi\right)-a\cos\varphi\right]d\lambda}{ad\varphi}\\ &\rightarrow -\frac{1}{2}\left(\frac{\partial}{\partial\varphi}\cos\varphi\right)d\lambda\\ &= \frac{1}{2}\sin\varphi d\lambda\,.\end{aligned} \tag{A.40}$$

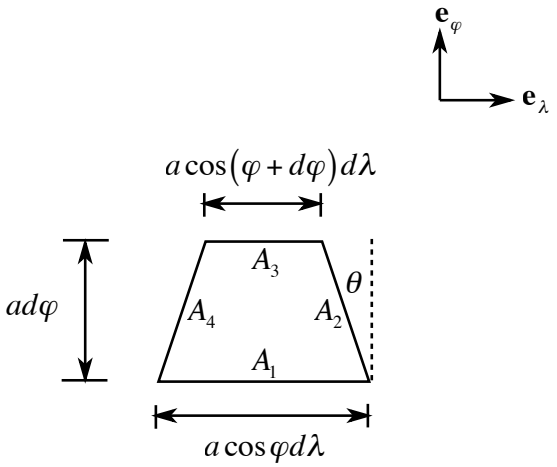


Figure A.2: A patch of the sphere, with longitudinal width $a\cos\varphi d\lambda$, and latitudinal height $ad\varphi$.

The angle θ is of "differential" or infinitesimal size. Nevertheless, it is needed in the derivation of (A.26). The line integral in (A.6) can be expressed as

$$\begin{aligned}\frac{1}{\text{Area}}\oint A\mathbf{n}dl =&\frac{1}{a^2\cos\varphi d\lambda d\varphi}\left[-\mathbf{e}_\varphi A_1 a\cos\varphi d\lambda+\mathbf{e}_\lambda A_2\cos\theta a d\varphi+\mathbf{e}_\varphi A_2\sin\theta a d\varphi\right.\\&\left.+\mathbf{e}_\varphi A_3 a\cos(\varphi+d\varphi)-\mathbf{e}_\lambda A_4\cos\theta a d\varphi+\mathbf{e}_\varphi A_4\sin\theta a d\varphi\right]\\=&\mathbf{e}_\lambda\frac{(A_2-A_4)\cos\theta}{a\cos\varphi d\lambda}\\&+\mathbf{e}_\varphi\left\{\frac{[A_3\cos(\varphi+d\varphi)-A_1\cos\varphi]d\lambda+(A_2+A_4)\sin\theta d\varphi}{a\cos\varphi d\lambda d\varphi}\right\}\end{aligned} \tag{A.41}$$

Note how the angle θ has entered here. Put $\cos\theta\to 1$ and $\sin\theta\to\frac{1}{2}\sin\varphi d\lambda$ to obtain

$$\begin{aligned}\frac{1}{\text{Area}}\oint\alpha\mathbf{n}dl=&\mathbf{e}_\lambda\frac{(A_2-A_4)}{a\cos\varphi d\lambda}+\mathbf{e}_\varphi\left\{\left[\frac{A_3\cos(\varphi+d\varphi)-A_1\cos\varphi}{a\cos\varphi d\varphi}\right]+\left(\frac{A_2+A_4}{2}\right)\frac{\sin\varphi}{a\cos\varphi}\right\}\\\to&\mathbf{e}_\lambda\frac{1}{a\cos\varphi}\frac{\partial A}{\partial\lambda}+\mathbf{e}_\varphi\left[\frac{1}{a\cos\varphi}\frac{\partial}{\partial\varphi}(A\cos\varphi)+\frac{A\sin\varphi}{a\cos\varphi}\right]\\=&\mathbf{e}_\lambda\frac{1}{a\cos\varphi}\frac{\partial A}{\partial\lambda}+\mathbf{e}_\varphi\frac{1}{a}\frac{\partial A}{\partial\varphi},\end{aligned} \tag{A.42}$$

which agrees with the two-dimensional version of (A.26).

Similar (but more straightforward) derivations can be given for (A.27) - (A.30).

A.5.4 Applying vector operators to the unit vectors in spherical coordinates

Using (A.11) - (A.13), we can prove the following about the unit vectors in spherical coordinates:

$$\nabla\cdot\mathbf{e}_\lambda=0, \tag{A.43}$$

$$\nabla\cdot\mathbf{e}_\varphi=-\frac{\tan\varphi}{r}, \tag{A.44}$$

$$\nabla \cdot \mathbf{e}_r = \frac{2}{r}, \tag{A.45}$$

$$\nabla \times \mathbf{e}_\lambda = \frac{\mathbf{e}_\varphi}{r} + \frac{\tan\varphi}{r}\mathbf{e}_r, \tag{A.46}$$

$$\nabla \times \mathbf{e}_\varphi = -\frac{\mathbf{e}_\lambda}{r}, \tag{A.47}$$

$$\nabla \times \mathbf{e}_r = 0. \tag{A.48}$$

The following relations are useful when working with the momentum equation in spherical coordinates:

$$(\mathbf{v}\cdot\nabla)\,\mathbf{e}_\lambda = \frac{u\sin\varphi}{r}\mathbf{e}_\varphi - \frac{u\cos\varphi}{r}\mathbf{e}_r\,, \tag{A.49}$$

$$(\mathbf{v}\cdot\nabla)\,\mathbf{e}_\varphi = -\frac{u\sin\varphi}{r}\mathbf{e}_\lambda - \frac{v\sin\varphi}{r}\mathbf{e}_r, \tag{A.50}$$

$$(\mathbf{v}\cdot\nabla)\,\mathbf{e}_r = \frac{\mathbf{v}}{r}\,. \tag{A.51}$$

Here $\mathbf{v}$ is the horizontal wind vector.

A.6 Solid body rotation

As an example of the application of (A.28), the vertical component of the vorticity is

$$\zeta = \frac{1}{r\cos\varphi}\left[\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \varphi}(u\cos\varphi)\right] \tag{A.52}$$

For the case of pure solid body rotation of the atmosphere about the Earth's axis of rotation, we have

$$u = \dot{\lambda} r\cos\varphi \text{ and } v = 0, \tag{A.53}$$

where $\dot{\lambda}$ is independent of φ (because that is what "solid body rotation" means). Substitution of (A.53) into (A.52) gives

$$\begin{aligned}\zeta =& \frac{-1}{r\cos\varphi}\frac{\partial}{\partial\varphi}\left(\dot{\lambda} r\cos^2\varphi\right)\\ =& 2\dot{\lambda}\sin\varphi\ ,\end{aligned} \tag{A.54}$$

which looks just like the familiar Coriolis parameter.

A.7 Formulas that are useful for two-dimensional flow

Consider the special case of two-dimensional flow. Two useful identities are

$$\nabla_r \times (\mathbf{e}_r \times \nabla_r \alpha) = \mathbf{e}_r \nabla_r^2 \alpha, \tag{A.55}$$

and

$$\nabla_r \cdot (\mathbf{e}_r \times \nabla_r \alpha) = 0. \tag{A.56}$$

Also for two-dimensional flow, the Laplacian of a vector can be written in a very simple way. Let $\zeta \mathbf{e}_r \equiv \nabla_r \times \mathbf{v}$ and $\delta \equiv \nabla_r \cdot \mathbf{v}$. Then (A.22) reduces to

$$\nabla_r^2 \mathbf{v} = \nabla_r \delta - \nabla_r \times (\zeta \mathbf{e}_r) \tag{A.57}$$

Using (A.28), we can write

$$\begin{aligned}\nabla_r \times (\zeta \mathbf{e}_r) &= \left\{ \frac{1}{r}\frac{\partial \zeta}{\partial \varphi},\ -\frac{1}{r\cos\varphi}\frac{\partial \zeta}{\partial \lambda},\ 0 \right\} \\ &= -\,\mathbf{e}_r \times \nabla_r \zeta\ .\end{aligned} \tag{A.58}$$

Then (A.57) becomes

$$\nabla_r^2 \mathbf{v} = \nabla_r \delta + \mathbf{e}_r \times \nabla_r \zeta. \tag{A.59}$$

A.8 Vertical coordinate transformations

Consider two vertical coordinates, denoted by "z" and "$\hat{z}$," respectively. Although the symbol z suggests height, no such implication is intended here; z and $\hat{z}$ can be any variables at all, so long as they vary monotonically with height (and with each other). For example, z could be pressure and $\hat{z}$ could be potential temperature. We assume that we have a rule telling how to compute $\hat{z}$ for a given value of z, and vice versa. For example, we could define $\hat{z} \equiv z - z_S(x,y)$, where $z_S(x,y)$ is the distribution of z along the Earth's surface.

Now consider the variation of an arbitrary dependent variable, f, with the independent variables x and z, as sketched in Figure A.3. Our goal is to relate $(\partial f/\partial x)_{\hat{z}}$ to $(\partial f/\partial x)_z$. Here x is not necessarily a spatial coordinate; it could be time. With reference to Fig. A.3, we can write

$$\begin{aligned}\frac{f_B - f_A}{x_2 - x_1} &= \left(\frac{f_C - f_A}{x_2 - x_1}\right) - \left(\frac{f_C - f_B}{x_2 - x_1}\right) \\ &= \left(\frac{f_C - f_A}{x_2 - x_1}\right) - \left(\frac{f_C - f_B}{z_2 - z_1}\right)\left(\frac{z_2 - z_1}{x_2 - x_1}\right)\ .\end{aligned} \tag{A.60}$$

Taking the limit as the increments become small, we find that

$$\left(\frac{\partial f}{\partial x}\right)_z = \left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}}\ . \tag{A.61}$$

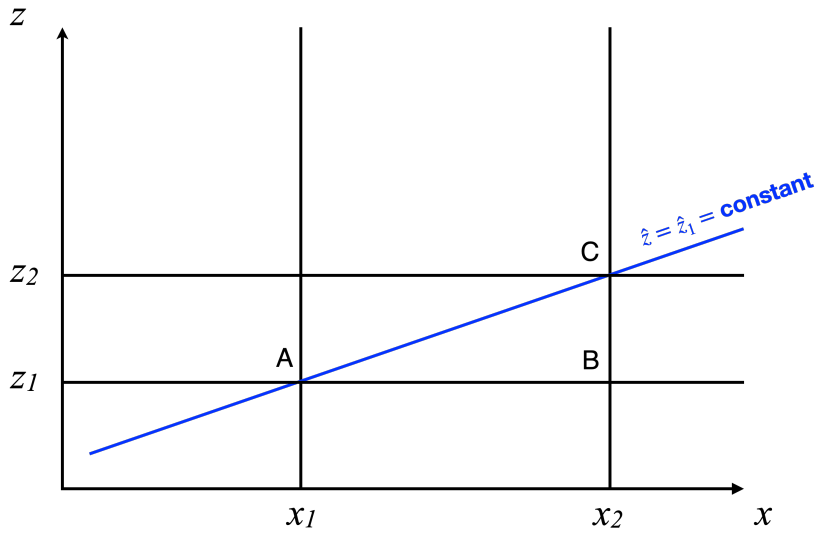


Figure A.3: A sketch used in the derivation of (A.61). The black horizontal lines have $z =$ constant. The blue line has $\hat{z} =$ constant.

It is useful to rewrite this as

$$\begin{aligned}\frac{\partial z}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_z = \frac{\partial z}{\partial \hat{z}}\left[\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}}\right] &= \frac{\partial z}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial \hat{z}}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}} \\ &= \frac{\partial z}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left(f\frac{\partial z}{\partial x}\right)_{\hat{z}} + f\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial \hat{z}}\right) \\ &= \frac{\partial}{\partial x}\left(f\frac{\partial z}{\partial \hat{z}}\right)_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left(f\frac{\partial z}{\partial x}\right)_{\hat{z}}\,.\end{aligned} \tag{A.62}$$

Alternatively, we can write

$$\begin{aligned}\frac{\partial z}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_z = \frac{\partial z}{\partial \hat{z}}\left[\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}}\right] &= \frac{\partial z}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial \hat{z}}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}} \\ &= \frac{\partial}{\partial \hat{z}}\left(z\frac{\partial f}{\partial x}\right)_{\hat{z}} - z\frac{\partial}{\partial \hat{z}}\left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial \hat{z}}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}} \\ &= \frac{\partial}{\partial \hat{z}}\left(z\frac{\partial f}{\partial x}\right)_{\hat{z}} - z\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial \hat{z}}\right)_{\hat{z}} - \left(\frac{\partial f}{\partial \hat{z}}\right)_x \left(\frac{\partial z}{\partial x}\right)_{\hat{z}} \\ &= \frac{\partial}{\partial \hat{z}}\left(z\frac{\partial f}{\partial x}\right)_{\hat{z}} - \frac{\partial}{\partial x}\left[z\left(\frac{\partial f}{\partial \hat{z}}\right)_x\right]_{\hat{z}} .\end{aligned} \tag{A.63}$$

From (A.61), we see that

$$\nabla_z f = \nabla_{\hat{z}} f - \frac{\partial f}{\partial z}\nabla_{\hat{z}} z \,. \tag{A.64}$$

Note that $\partial f/\partial z$ term is still a horizontal vector. By analogy with (A.62) and (A.63), we can rewrite (A.64) as

$$\frac{\partial z}{\partial \hat{z}}(\nabla_z f) = \nabla_{\hat{z}}\left(f\frac{\partial z}{\partial \hat{z}}\right) - \frac{\partial}{\partial \hat{z}}(f\nabla_{\hat{z}} z) \,, \tag{A.65}$$

or

$$\frac{\partial z}{\partial \hat{z}}\nabla_z f = \frac{\partial}{\partial \hat{z}}(z\nabla_{\hat{z}} f) - \nabla_{\hat{z}}\left(z\frac{\partial f}{\partial \hat{z}}\right) \,. \tag{A.66}$$

The gradient terms on the right-hand sides of these equations vanish when integrated around any closed path, including a latitude circle. The other terms can be interpreted as vertical flux divergences.

Similarly, the curl of a horizontal vector can be written as

$$\nabla_z \times \mathbf{v} = \nabla_{\hat{z}} \times \mathbf{v} + \frac{\partial \mathbf{v}}{\partial z} \times \nabla_{\hat{z}} z \,. \tag{A.67}$$

All terms are vertical vectors. Multiplying by $\partial z/\partial\hat{z}$, we find that

$$\begin{aligned}\frac{\partial z}{\partial\hat{z}}(\nabla_z\times\mathbf{v})=\frac{\partial z}{\partial\hat{z}}\left(\nabla_{\hat{z}}\times\mathbf{v}+\frac{\partial\mathbf{v}}{\partial z}\times\nabla_{\hat{z}}z\right)=&\frac{\partial z}{\partial\hat{z}}(\nabla_{\hat{z}}\times\mathbf{v})+\frac{\partial\mathbf{v}}{\partial\hat{z}}\times\nabla_{\hat{z}}z\\=&\frac{\partial z}{\partial\hat{z}}(\nabla_{\hat{z}}\times\mathbf{v})+\frac{\partial}{\partial\hat{z}}(\mathbf{v}\times\nabla_{\hat{z}}z)-\mathbf{v}\times\nabla_{\hat{z}}\left(\frac{\partial z}{\partial\hat{z}}\right)\\=&\nabla_{\hat{z}}\times\left(\mathbf{v}\frac{\partial z}{\partial\hat{z}}\right)+\frac{\partial}{\partial\hat{z}}(\mathbf{v}\times\nabla_{\hat{z}}z)\ .\end{aligned}\tag{A.68}$$

Finally, the divergence of a horizontal vector can be written as

$$\nabla_z\cdot\mathbf{v}=\nabla_{\hat{z}}\cdot\mathbf{v}-\frac{\partial\mathbf{v}}{\partial z}\cdot\nabla_{\hat{z}}z\ .\tag{A.69}$$

Multiplying (A.69) by $\partial z/\partial\hat{z}$, we find that

$$\frac{\partial z}{\partial\hat{z}}(\nabla_z\cdot\mathbf{v})=\nabla_{\hat{z}}\cdot\left(\mathbf{v}\frac{\partial z}{\partial\hat{z}}\right)-\frac{\partial}{\partial\hat{z}}(\mathbf{v}\cdot\nabla_{\hat{z}}z)\ .\tag{A.70}$$

A.9 Concluding summary

Scalars, vectors, and tensors have meanings independent of any particular coordinate system, although they can be expressed using coordinate systems.

This brief overview is intended mainly as a refresher for students who learned these concepts once upon a time, but may have not thought about them for awhile. We have also included many useful equations that are not readily available elsewhere, even on the Web.

Appendix B

A Demonstration that the Fourth-Order Runge-Kutta Scheme Really Does Have Fourth-Order Accuracy

We wish to obtain an approximate numerical solution of the ordinary differential equation

$$\frac{dq}{dt} = f(q,t). \tag{B.1}$$

Here, as indicated, the function f depends on both q and t, but q itself depends only on t.

As discussed earlier, the fourth-order Runge-Kutta scheme is given by

$$\frac{q^{n+1} - q^n}{\Delta t} = \frac{1}{6}(k_1, +2k_2 + 2k_3 + k_4), \tag{B.2}$$

where

$$\begin{aligned}
k_1 &= f(q^n, n\Delta t), \\
k_2 &= f\left[q^n + \frac{k_1 \Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right], \\
k_3 &= f\left[q^n + \frac{k_2 \Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right], \\
k_4 &= f\left[q^n + k_3 \Delta t, (n+1)\Delta t\right].
\end{aligned} \tag{B.3}$$

To demonstrate that B.2 has fourth-order accuracy, we substitute the exact solution into B.2 - B.3, and rearrange the result, to obtain

$$\frac{q^{n+1}-q^n}{\Delta t}-\frac{1}{6}\left(k_1,+2k_2+2k_3+k_4\right)=\varepsilon \tag{B.4}$$

where ε is the discretization error of the scheme.

Taylor-series expansion allows us to write

$$\frac{q^{n+1}-q^n}{\Delta t}=\frac{dq}{dt}+\frac{1}{2!}\left(\Delta t\right)\frac{d^2q}{dt^2}+\frac{1}{3!}(\Delta t)^2\frac{d^3q}{dt^3}+\frac{1}{4!}(\Delta t)^3\frac{d^4q}{dt^4}+\mathrm{O}\left[(\Delta t)^4\right]. \tag{B.5}$$

Each term on the right-hand side of of B.5 can be expressed in terms of $f(q,t)$ and its derivatives, as follows. The *total* time rate of change of an arbitrary function $A(q,t)$ that depends on both q and t is given by

$$\begin{aligned}\frac{dA}{dt}&\equiv\frac{\partial A}{\partial t}\frac{dt}{dt}+\frac{\partial A}{\partial q}\frac{dq}{dt}\\&=\frac{\partial A}{\partial t}+f\frac{\partial A}{\partial q}\\&\equiv\delta\left(A\right).\end{aligned} \tag{B.6}$$

Here a partial derivative with respect to t is taken while holding q constant, and vice versa. As a special case of B.6,

$$\delta f\equiv\frac{\partial f}{\partial t}+f\frac{\partial f}{\partial q}. \tag{B.7}$$

We can now write

$$\boxed{\frac{dq}{dt}=f}, \tag{B.8}$$

$$\boxed{\frac{d^2q}{dt^2} = \delta f}\,, \tag{B.9}$$

and

$$\begin{aligned}
\frac{d^3q}{dt^3} &= \delta\left(\delta f\right) \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right)\frac{dq}{dt} + \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial q}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)\frac{dq}{dt} + \frac{\partial^2 f}{\partial t^2} \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f + \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial q}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) f + \frac{\partial^2 f}{\partial t^2} \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) f + \frac{\partial^2 f}{\partial t^2} \\
&= \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f^2 + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial^2 f}{\partial t^2} \\
&= f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2} + \left(\frac{\partial f}{\partial q}\right)(\delta f) \\
&= \left(f\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)^2 f + \left(\frac{\partial f}{\partial q}\right)(\delta f)\,,
\end{aligned} \tag{B.10}$$

so that

$$\boxed{\frac{d^3q}{dt^3} = \left(\delta^2 f\right) + f_q\left(\delta f\right)}\,. \tag{B.11}$$

Here we have used the notation $f_q \equiv \frac{\partial f}{\partial q}$. Finally, $\frac{d^4q}{dt^4}$ is given by

$$\begin{aligned}\frac{d^4q}{dt^4} &= \delta\left[\left(\delta^2 f\right) + f_q\left(\delta f\right)\right] \\ &= \left[f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right)\right] + \left[f\frac{\partial}{\partial q}\left(f_q\right) + \frac{\partial}{\partial t}\left(f_q\right)\right]\left(\delta f\right) + f_q\left[f\frac{\partial}{\partial q}\left(\delta f\right) + \frac{\partial}{\partial t}\left(\delta f\right)\right].\end{aligned} \tag{B.12}$$

This is a bit messy. To break the analysis of B.12 into steps, we first manipulate the first term in square brackets, separately. Expanding, we find that

$$\begin{aligned}&f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right) \\ &= f\frac{\partial}{\partial q}\left[f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2}\right] + \frac{\partial}{\partial t}\left[f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2}\right] \\ &= f\left[2f\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q^2}\right) + f^2\left(\frac{\partial^3 f}{\partial q^3}\right) + 2\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q\partial t}\right) + 2f\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + \left(\frac{\partial^3 f}{\partial t^2\partial q}\right)\right] \\ &+ 2f\frac{\partial f}{\partial t}\left(\frac{\partial^2 f}{\partial q^2}\right) + f^2\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + 2\left(\frac{\partial f}{\partial t}\right)\left(\frac{\partial^2 f}{\partial q\partial t}\right) + 2f\left(\frac{\partial^3 f}{\partial t^2\partial q}\right) + \frac{\partial^3 f}{\partial t^3}\end{aligned} \tag{B.13}$$

The terms can be collected and grouped as follows:

$$\begin{aligned}f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right) &= \left[f^3\left(\frac{\partial^3 f}{\partial q^3}\right) + 3f^2\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + 3f\left(\frac{\partial^3 f}{\partial t^2\partial q}\right) + \frac{\partial^3 f}{\partial t^3}\right] \\ &+ f\left[2f\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q^2}\right) + 2\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q\partial t}\right)\right] + 2f\frac{\partial f}{\partial t}\left(\frac{\partial^2 f}{\partial q^2}\right) + 2\left(\frac{\partial f}{\partial t}\right)\left(\frac{\partial^2 f}{\partial q\partial t}\right) \\ &= \delta^3 f + 2\left(\delta f\right)\left(\delta f_q\right).\end{aligned} \tag{B.14}$$

Substituting into B.2, we find that

$$\begin{aligned}
\frac{d^4q}{dt^4} &= \left[f\frac{\partial}{\partial q}\left(\delta^2 f\right)+\frac{\partial}{\partial t}\left(\delta^2 f\right)\right]+\left[f\frac{\partial}{\partial q}\left(f_q\right)+\frac{\partial}{\partial t}\left(f_q\right)\right](\delta f)+f_q\left[f\frac{\partial}{\partial q}(\delta f)+\frac{\partial}{\partial t}(\delta f)\right] \\
&= \left[\left(\delta^3 f\right)+2(\delta f)\left(\delta f_q\right)\right]+\left(\delta f_q\right)(\delta f)+f_q\left\{f\frac{\partial}{\partial q}\left[f\left(f_q\right)+\frac{\partial f}{\partial t}\right]+\frac{\partial}{\partial t}\left[f\left(f_q\right)+\frac{\partial f}{\partial t}\right]\right\} \\
&= \left(\delta^3 f\right)+3(\delta f)\left(\delta f_q\right)+f_q\left[\left(f\frac{\partial f}{\partial q}\right)\left(f_q\right)+f\left(f\frac{\partial f_q}{\partial q}\right)+f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)+\left(\frac{\partial f}{\partial t}\right)\left(f_q\right)+f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right] \\
&= \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2\left(f\frac{\partial f}{\partial q}+\frac{\partial f}{\partial t}\right)+f_q\left[f\left(f\frac{\partial f_q}{\partial q}\right)+f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)+f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right] \\
&= \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2(\delta f)+f_q\left[f\left(f\frac{\partial f_q}{\partial q}\right)+2f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right],
\end{aligned} \tag{B.15}$$

which can be written as

$$\boxed{\frac{d^4q}{dt^4} = \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2(\delta f)+f_q\left(\delta^2 f\right)}\,. \tag{B.16}$$

Next, we express k_1 - k_4 in terms of $f(q,t)$ and its derivatives. We write

$$\boxed{k_1 = f}\,, \tag{B.17}$$

and

$$\begin{aligned}
k_2 &= f\left[q^n+k_1\frac{\Delta t}{2},\left(n+\frac{1}{2}\right)\Delta t\right] \\
&= f+\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]f \\
&+\frac{1}{2!}\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]^2 f+\frac{1}{3!}\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]^3 f+\mathrm{O}\left[(\Delta t)^4\right],
\end{aligned} \tag{B.18}$$

which is equivalent to

$$\boxed{k_2 = f + \left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 \delta^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 \delta^3 f + \mathrm{O}\left[(\Delta t)^4\right]} \tag{B.19}$$

Here we have used a two-dimensional Taylor's series expansion, because we have two independent variables, namely q and t. Similarly,

$$\begin{aligned} k_3 &= f\left[q^n + k_2\frac{\Delta t}{2}, \left(n+\frac{1}{2}\right)\Delta t\right] \\ &= f + \left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right] f + \frac{1}{2!}\left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^2 f \\ &+ \frac{1}{3!}\left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^3 f + \mathrm{O}\left[(\Delta t)^4\right]. \end{aligned} \tag{B.20}$$

From B.6 and B.20, we see that

$$k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t} = \delta + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 \delta^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 \delta^3 f\right]\frac{\partial}{\partial q} + \mathrm{O}\left[(\Delta t)^4\right]. \tag{B.21}$$

Substituting B.9 into B.10, we obtain

$$\begin{aligned} k_3 = f &+ \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 (\delta^2 f)\right]\frac{\partial f}{\partial q}\right\} \\ &+ \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left[\delta f + \left(\frac{\Delta t}{2}\right)\delta f\frac{\partial}{\partial q}\right]^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 (\delta^3 f) + \mathrm{O}\left[(\Delta t)^4\right]. \end{aligned} \tag{B.22}$$

Expand, and combine terms:

$$
\begin{aligned}
k_3 &= f + \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left(\delta^2 f\right)\right] f_q\right\} \\
&\quad + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left\{\Delta t\left(\delta f\right)\left(\delta f_q\right) + \left(\delta^2 f\right) + \mathrm{O}\left[(\Delta t)^2\right]\right\} + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3\left(\delta^3 f\right) + \mathrm{O}\left[(\Delta t)^4\right] \\
&= f + \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left(\delta^2 f\right)\right] f_q\right\} \\
&\quad + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left[\Delta t\left(\delta f\right)\left(\delta f_q\right) + \left(\delta^2 f\right)\right] + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3\left(\delta^3 f\right) + \mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}
\tag{B.23}
$$

Here we have included only the terms of k_2 that contribute up to $O\left[(\Delta t)^3\right]$; the remaining terms have been tossed onto the $O\left[(\Delta t)^4\right]$ pile at the end of B.23. Collecting powers of Δt, we conclude that

$$
\boxed{
\begin{aligned}
k_3 &= f + \Delta t\left(\frac{\delta f}{2}\right) + \frac{(\Delta t)^2}{2!}\left[\frac{\left(\delta f\right) f_q}{2} + \frac{\delta^2 f}{4}\right] \\
&\quad + \frac{(\Delta t)^3}{3!}\left[\frac{3\left(\delta^2 f\right) f_q}{8} + \frac{3\left(\delta f\right)\left(\delta f_q\right)}{4} + \frac{\left(\delta^3 f\right)}{8}\right] + \mathrm{O}\left[(\Delta t)^4\right]
\end{aligned}
}.
\tag{B.24}
$$

It remains to assemble k_4. We start with the two-dimensional Taylor series expansion:

$$
\begin{aligned}
k_4 &= f\left[q^n + k_3\Delta t, (n+1)\Delta t\right] \\
&= f + \Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right) f + \frac{1}{2!}\left[\Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^2 f + \frac{1}{3!}\left[\Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^3 f + \mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}
\tag{B.25}
$$

Next, use B.6 and B.14 to write

$$k_3\frac{\partial}{\partial q}+\frac{\partial}{\partial t}=$$
$$\delta+\left\{\Delta t\left(\frac{\delta f}{2}\right)+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]+\frac{(\Delta t)^3}{3!}\left[\frac{3\left(\delta^2 f\right)f_q}{8}+\frac{3(\delta f)\left(\delta f_q\right)}{4}+\frac{\left(\delta^3 f\right)}{8}\right]\right\}$$
$$\frac{\partial}{\partial q}+\mathrm{O}\left[(\Delta t)^4\right]. \tag{B.26}$$

Substituting B.26 into B.25, we obtain

$$\begin{aligned}
k_4 &= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left[\delta+\Delta t\left(\frac{\delta f}{2}\right)\frac{\partial}{\partial q}\right]^2 f+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right]\\
&= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left\{\left[\left(\delta^2 f\right)+\Delta t(\delta f)\left(\delta f_q\right)\right]+\mathrm{O}\left[(\Delta t)^2\right]\right\}+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right]\\
&= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left[\left(\delta^2 f\right)+\Delta t(\delta f)\left(\delta f_q\right)\right]+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right].
\end{aligned} \tag{B.27}$$

As before, we have written only terms that contribute up to $O\left[(\Delta t)^3\right]$. Now collect powers of Δt:

$$\boxed{\begin{aligned}
k_4 &= f+(\Delta t)\,\delta f+\frac{(\Delta t)^2}{2!}\left[(\delta f)f_q+\left(\delta^2 f\right)\right]\\
&+\frac{(\Delta t)^3}{3!}\left[\frac{3(\delta f)\left(f_q\right)^2}{2}+\frac{3\left(\delta^2 f\right)f_q}{4}+3(\delta f)\left(\delta f_q\right)+\left(\delta^3 f\right)\right]+\mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}} \tag{B.28}$$

The final step is to substitute the various boxed relations above into B.4. Everything cancels out, and we are left with $\varepsilon = O\left[(\Delta t)^4\right]$. This completes the proof.

Appendix C

Total energy conservation with the genralized vertical coordinate

C.1 The general case

We now derive a statement of total energy conservation with the generalized vertical coordinate, without using the quasi-static approximation. The first step is to form the kinetic energy equation, by taking $\mathbf{v}\cdot$ each term of (22.17), and adding the result to w times each term of (22.18). The result can be written as

$$\begin{aligned}\rho_{\hat{z}}\frac{DK}{Dt} = &-\mathbf{v}\cdot\left[\frac{\partial}{\partial\hat{z}}\left(z\nabla_{\hat{z}}p\right)-\nabla_{\hat{z}}\left(z\frac{\partial p}{\partial\hat{z}}\right)\right]+w\left(-\frac{\partial p}{\partial\hat{z}}-\rho_{\hat{z}}g\right)\\ &-\cdot\left[\nabla_{\hat{z}}\cdot\left(\mathbf{S}\frac{\partial z}{\partial\hat{z}}\right)+\frac{\partial}{\partial\hat{z}}\left(\mathbf{S}\cdot\nabla_{\hat{z}}z\right)\right]-w\frac{\partial}{\partial\hat{z}}\left[\mathbf{e}_r\cdot\left(\nabla\cdot\mathbf{S}\right)\right]\ ,\end{aligned} \tag{C.1}$$

where

$$K\equiv\frac{1}{2}\left(\mathbf{v}\cdot\mathbf{v}+w^2\right)=\frac{1}{2}\left(\mathbf{V}\cdot\mathbf{V}\right) \tag{C.2}$$

is the kinetic energy per unit mass. We can show that

$$\mathbf{v}\cdot\left[\nabla_{\hat{z}}\cdot\left(\mathbf{S}\frac{\partial z}{\partial\hat{z}}\right)+\frac{\partial}{\partial\hat{z}}\left(\mathbf{S}\cdot\nabla_{\hat{z}}z\right)\right]+w\frac{\partial}{\partial\hat{z}}\left[\mathbf{e}_r\cdot\left(\nabla\cdot\mathbf{S}\right)\right]=\nabla\cdot\mathbf{W}+\delta\ . \tag{C.3}$$

Using (22.22) to eliminate w in (C.1), we find that

$$\rho_{\hat{z}}\frac{DK}{Dt} = -\mathbf{v}\cdot\left[\frac{\partial}{\partial\hat{z}}\left(z\nabla_{\hat{z}}p\right) - \nabla_{\hat{z}}\left(z\frac{\partial p}{\partial\hat{z}}\right)\right] + \left[\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} + \mathbf{v}\cdot\nabla_{\hat{z}}z + \hat{w}\frac{\partial z}{\partial\hat{z}}\right]\left(-\frac{\partial p}{\partial\hat{z}} - \rho_{\hat{z}}g\right) - \nabla\cdot\mathbf{W} - \delta\,, \tag{C.4}$$

which can be simplified and rearranged to

$$\rho_{\hat{z}}\frac{D}{Dt}\left(K+\phi\right) = -\mathbf{v}\cdot\left[\frac{\partial}{\partial\hat{z}}\left(z\nabla_{\hat{z}}p\right) - \nabla_{\hat{z}}\left(z\frac{\partial p}{\partial\hat{z}}\right)\right] - \left[\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} + \mathbf{v}\cdot\nabla_{\hat{z}}z + \hat{w}\frac{\partial z}{\partial\hat{z}}\right]\left(\frac{\partial p}{\partial\hat{z}}\right) - \nabla\cdot\mathbf{W} - \delta\,. \tag{C.5}$$

The terms involving $\mathbf{v}$ mostly cancel:

$$\begin{aligned} &-\frac{\partial}{\partial\hat{z}}\left(z\nabla_{\hat{z}}p\right) + \nabla_{\hat{z}}\left(z\frac{\partial p}{\partial\hat{z}}\right) - \nabla_{\hat{z}}z\left(\frac{\partial p}{\partial\hat{z}}\right) \\ &= -\frac{\partial z}{\partial\hat{z}}\nabla_{\hat{z}}p - z\frac{\partial}{\partial\hat{z}}\left(\nabla_{\hat{z}}p\right) + \nabla_{\hat{z}}z\left(\frac{\partial p}{\partial\hat{z}}\right) + z\nabla_{\hat{z}}\left(\frac{\partial p}{\partial\hat{z}}\right) - \nabla_{\hat{z}}z\left(\frac{\partial p}{\partial\hat{z}}\right) \\ &= -\frac{\partial z}{\partial\hat{z}}\nabla_{\hat{z}}p \end{aligned} \tag{C.6}$$

With the use of (C.6), (C.5) simplifies to

$$\rho_{\hat{z}}\frac{D}{Dt}\left(K+\phi\right) = -\frac{\partial z}{\partial\hat{z}}\left(\mathbf{v}\cdot\nabla_{\hat{z}}p + \hat{w}\frac{\partial p}{\partial\hat{z}}\right) - \left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\left(\frac{\partial p}{\partial\hat{z}}\right) - \nabla\cdot\mathbf{W} - \delta\,. \tag{C.7}$$

Using

$$\omega = \left(\frac{\partial p}{\partial t}\right)_{\hat{z}} + \mathbf{v}\cdot\nabla_{\hat{z}}p + \hat{w}\frac{\partial p}{\partial\hat{z}}\,, \tag{C.8}$$

we can rewrite (C.7) as

$$\rho_{\hat{z}}\frac{D}{Dt}(K+\phi) = -\rho_{\hat{z}}\omega\alpha + \frac{\partial z}{\partial \hat{z}}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} - \left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\left(\frac{\partial p}{\partial \hat{z}}\right) - \nabla\cdot\mathbf{W} - \delta\,, \tag{C.9}$$

where, as a reminder, $\alpha = 1/\rho$. In general, we can write

$$\begin{aligned}\frac{\partial z}{\partial \hat{z}}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} &= \left[\frac{\partial}{\partial t}\left(p\frac{\partial z}{\partial \hat{z}}\right)\right]_{\hat{z}} - p\left[\frac{\partial}{\partial t}\left(\frac{\partial z}{\partial \hat{z}}\right)\right] \\ &= \left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}}RT\right)\right]_{\hat{z}} - p\frac{\partial}{\partial \hat{z}}\left[\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right]\,,\end{aligned} \tag{C.10}$$

and

$$\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\frac{\partial p}{\partial \hat{z}} = \frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] - p\frac{\partial}{\partial \hat{z}}\left[\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right]\,. \tag{C.11}$$

Combining these, cancellation occurs, and we find that

$$\frac{\partial z}{\partial \hat{z}}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} - \left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\frac{\partial p}{\partial \hat{z}} = \left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}}RT\right)\right]_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right]\,. \tag{C.12}$$

Now we substitute (C.12) into (C.9), and rearrange, to obtain

$$\rho_{\hat{z}}\frac{D}{Dt}(K+\phi) = -\rho_{\hat{z}}\omega\alpha + \left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}}RT\right)\right]_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] - \nabla\cdot\mathbf{W} - \delta\,. \tag{C.13}$$

In flux form, this can be written as

$$\begin{aligned}\frac{\partial}{\partial t}\left[\rho_{\hat{z}}\left(K+\phi-RT\right)\right]_{\hat{z}} + \nabla_{\hat{z}}\cdot\left[\rho_{\hat{z}}\mathbf{v}\left(K+\phi\right)\right] + \frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}}\hat{w}\left(K+\phi\right) + p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] \\ = -\rho_{\hat{z}}\omega\alpha - \nabla\cdot\mathbf{W} - \delta\,.\end{aligned} \tag{C.14}$$

Now we turn to the thermodynamic energy equation in terms of enthalpy, Eq. (2.23). Using the general vertical cordinate, it can be written in flux form as

$$\frac{\partial}{\partial t}\left(\rho_{\hat{z}} c_p T\right)_{\hat{z}} + \nabla_{\hat{z}} \cdot \left(\rho_{\hat{z}} \mathbf{v} c_p T\right) + \frac{\partial}{\partial \hat{z}}\left(\rho_{\hat{z}} \hat{w} c_p T\right) = \rho_{\hat{z}} \omega \alpha + \rho_{\hat{z}}\left(LC - \nabla \cdot \mathbf{R} + \delta\right) . \tag{C.15}$$

The latent energy equation can be written in flux form as

$$\left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} L q_v\right)\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot \left(\rho_{\hat{z}} \mathbf{v} L q_v\right) + \frac{\partial}{\partial \hat{z}}\left(\rho_{\hat{z}} \hat{w} L q_v\right) = -\rho_{\hat{z}} LC. \tag{C.16}$$

Adding (C.14), (C.15), and (C.16) gives a statement of total energy conservation:

$$\begin{gathered} \frac{\partial}{\partial t}\left[\rho_{\hat{z}} e_{\text{tot}}\right]_{\hat{z}} \\ + \nabla_{\hat{z}} \cdot \left[\rho_{\hat{z}} \mathbf{v}\left(K + h\right)\right] + \frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}\left(K + h\right) + p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] \\ = -\nabla \cdot \mathbf{W} + \rho_{\hat{z}} - \nabla \cdot \mathbf{R} , \end{gathered} \tag{C.17}$$

where

$$h \equiv c_p T + \phi + L q_v \tag{C.18}$$

is the moist static energy. Eq. (C.17) can also be written as

$$\boxed{\begin{gathered} \left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} e_{\text{tot}}\right)\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot \left[\rho_{\hat{z}} \mathbf{v}\left(e_{\text{tot}} + p\alpha\right)\right] + \frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}\left(e_{\text{tot}} + p\alpha\right) + p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] \\ = -\nabla \cdot \left(\mathbf{W} + \mathbf{R}\right) . \end{gathered}} \tag{C.19}$$

C.2 The energy equation with the quasi-static approximation

To derive the energy equation using the quasi-static approximation, we start from

$$\rho_{\hat{z}}\frac{DK}{Dt} = \mathbf{v}\cdot\left(-\frac{\partial z}{\partial \hat{z}}\nabla_{\hat{z}}p + \frac{\partial p}{\partial \hat{z}}\nabla_{\hat{z}}z\right) - \nabla\cdot\mathbf{W} - \delta\,, \tag{C.20}$$

where the kinetic energy is approximated by neglecting the contribution from the vertical velocity:

$$K \cong \frac{1}{2}\mathbf{v}\cdot\mathbf{v}\,. \tag{C.21}$$

Notice that (C.20) does not involve the vertical pressure-gradient force; compare with (C.1). Despite this rather different starting point, a derivation leads to (C.13) again:

$$\rho_{\hat{z}}\frac{D}{Dt}(K+\phi) = -\rho_{\hat{z}}\omega\alpha + \left[\frac{\partial}{\partial t}(\rho_{\hat{z}}RT)\right]_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] - \nabla\cdot\mathbf{W} - \delta\,. \tag{C.22}$$

There is no change from the fully compressible case, except that K has been approximated as shown in (C.21). However, with the use of the hydrostatic equation we can write (C.13) in a different way that has been used in the design of quasi-static models. Start with

$$\begin{aligned}
\rho_{\hat{z}}g\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} &= -\frac{\partial p}{\partial \hat{z}}\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} \\
&= -\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] + p\frac{\partial}{\partial \hat{z}}\left[\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] \\
&= -\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] + p\left[\frac{\partial}{\partial t}\left(\frac{\partial z}{\partial \hat{z}}\right)\right]_{\hat{z}} \\
&= -\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] + \left[\frac{\partial}{\partial t}\left(p\frac{\partial z}{\partial \hat{z}}\right)\right]_{\hat{z}} - \frac{\partial z}{\partial \hat{z}}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} \\
&= -\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] + \left[\frac{\partial}{\partial t}(\rho_{\hat{z}}RT)\right]_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left[z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] + z\frac{\partial}{\partial \hat{z}}\left[\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] \\
&= -\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right] + \left[\frac{\partial}{\partial t}(\rho_{\hat{z}}RT)\right]_{\hat{z}} - \frac{\partial}{\partial \hat{z}}\left[z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] - z\left[\frac{\partial}{\partial t}(\rho_{\hat{z}}g)\right]_{\hat{z}}\,.
\end{aligned} \tag{C.23}$$

Rearranging gives

$$-\frac{\partial}{\partial \hat{z}}\left[p\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}\right]=\left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} g z\right)\right]_{\hat{z}}-\left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} R T\right)\right]_{\hat{z}}+\frac{\partial}{\partial \hat{z}}\left[z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] \tag{C.24}$$

Substitution into (C.22) gives

$$\rho_{\hat{z}} \frac{D}{Dt}\left(K+\phi\right)=-\rho_{\hat{z}} \omega \alpha+\left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} g z\right)\right]_{\hat{z}}+\frac{\partial}{\partial \hat{z}}\left[z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right]-\nabla \cdot \mathbf{W}-\delta \ . \tag{C.25}$$

Compare with (C.22). In (C.25) there is a "new" $\rho_{\hat{z}} g z$ term, the $\rho_{\hat{z}} R T$ term is gone, and the pressure work term is in terms of $+z\partial p/\partial t$ instead of $-p\partial z/\partial t$. Converting (C.25) to flux form, we obtain

$$\begin{aligned}\left[\frac{\partial}{\partial t}\left(\rho_{\hat{z}} K\right)\right]_{\hat{z}}+\nabla_{\hat{z}} \cdot\left[\rho_{\hat{z}} \mathbf{v}\left(K+\phi\right)\right]+\frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}\left(K+\phi\right)-z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] \\ =-\rho_{\hat{z}} \omega \alpha-\nabla \cdot \mathbf{W}-\delta \ .\end{aligned} \tag{C.26}$$

Compare with (C.14). The time-tendency terms involving ϕ and RT are both gone. Adding (C.15) and (C.16) to (C.26), we obtain total energy conservation in the form

$$\boxed{\begin{aligned}\left[\frac{\partial}{\partial t} \rho_{\hat{z}}\left(K+c_p T+L q_v\right)\right]_{\hat{z}}+\nabla_{\hat{z}} \cdot\left[\rho_{\hat{z}} \mathbf{v}\left(K+h\right)\right]+\frac{\partial}{\partial \hat{z}}\left[\rho_{\hat{z}} \hat{w}\left(K+c_p T+\phi+L q_v\right)-z\left(\frac{\partial p}{\partial t}\right)_{\hat{z}}\right] \\ =-\nabla \cdot\left(\mathbf{W}+\mathbf{R}\right) \ .\end{aligned}} \tag{C.27}$$

Compare with (C.19).

Appendix D

Spherical Harmonics

The spherical surface harmonics are convenient functions for representing the distribution of geophysical quantities over the surface of the spherical Earth.

We consider a three-dimensional space and look for solutions of Laplace's differential equation, which is

$$\nabla^2 S = 0 \,. \tag{D.1}$$

Using spherical coordinates, the operator ∇^2 can be expanded as:

$$\nabla^2 S = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial S}{\partial r}\right) + \nabla_h^2 S = 0 \tag{D.2}$$

Here r is the distance from the origin, and ∇_h^2 is the Laplacian on a two-dimensional surface of constant r, i.e., on a spherical surface. The form of ∇_h^2 is discussed below. Inspection of (D.2) suggests that S should be proportional to a power of r, as in

$$S = r^n Y_n \,. \tag{D.3}$$

The Y_n are called spherical surface harmonics of order n. We assume that they are independent of radius. This is a "separability" assumption. The subscript n is attached to Y_n to remind us that it corresponds a particular exponent in the radial dependence of S.

In order for S to remain finite as $r \to 0$, we need $n \geq 0$. Since $n = 0$ would mean that S is independent of radius, we conclude that n must be a positive integer. Using

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial S}{\partial r}\right)=\frac{n\left(n+1\right)}{r^2}S\,, \tag{D.4}$$

which follows immediately from (D.3), we can rewrite (D.2) as

$$\boxed{\nabla_h^2 Y_n+\frac{n\left(n+1\right)}{r^2}Y_n=0\,.} \tag{D.5}$$

Before continuing with the separation of variables, it is important to point out that all of the quantities appearing in (D.5) have meaning without the need for any particular coordinate system on the two-dimensional spherical surface. We are going to use spherical (i.e., longitude and latitude) coordinates below, but we do not need them to write down (D.5).

At this point we make an analogy with a trigonometric functions. Suppose that we have a "doubly periodic" function $W\left(x,y\right)$ defined on a plane, with the usual Cartesian coordinates x and y. As an example, let

$$W\left(x,\,y\right)=A\sin\left(kx\right)\cos\left(ly\right)\,. \tag{D.6}$$

where A is an arbitrary constant. In Cartesian coordinates the two-dimensional Laplacian of W is

$$\nabla_h^2 W+\left(k^2+l^2\right)W=0\,. \tag{D.7}$$

Comparing (D.5) and (D.7), we see that they are closely analogous. In particular, $n\left(n+1\right)/r^2$ in (D.5) corresponds to k^2+l^2 in (D.7). This shows that $n\left(n+1\right)$ is proportional to a "total horizontal wave number" on the sphere. For a given value of r, larger n means larger total horizontal wave number, which implies a smaller horizontal scale.

We now expand the "horizontal Laplacian" in spherical coordinates as

$$\nabla_h^2 S=\frac{1}{r^2\cos\varphi}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial S}{\partial\varphi}\right)+\frac{1}{r^2\cos^2\varphi}\frac{\partial^2 S}{\partial\lambda^2}\,, \tag{D.8}$$

where λ is longitude and φ is latitude. Then (D.5) can be rewritten as

$$\frac{1}{\cos\varphi}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial Y_n}{\partial\varphi}\right)+\frac{1}{\cos^2\varphi}\frac{\partial^2 Y_n}{\partial\lambda^2}+n\,(n+1)\,Y_n=0\,. \tag{D.9}$$

Factors of $1/r^2$ have cancelled out in (D.9), and as a result r has disappeared. Nevertheless, n, the exponent of r, is still visible, like the smile of the Cheshire cat. In the following discussion, remember where n came from.

Next, we separate the longitude and latitude dependence in the $Y_n(\varphi,\,\lambda)$, i.e., we assume that

$$Y_n(\varphi,\,\lambda)=\Phi(\varphi)\Lambda(\lambda)\;, \tag{D.10}$$

where $\Lambda(\lambda)$ and $\Phi(\varphi)$ are to be determined. By substitution of (D.9) into (D.9), we find that

$$\frac{\cos^2\varphi}{\Phi}\left[\frac{1}{\cos\varphi}\frac{d}{d\varphi}\left(\cos\varphi\frac{d\Phi}{d\varphi}\right)+n\,(n+1)\,\Phi\right]=-\frac{1}{\Lambda}\frac{d^2\Lambda}{d\lambda^2}\;. \tag{D.11}$$

The left-hand side of (D.11) does not depend on λ, and the right-hand side does not depend on φ, so both sides must be equal to a constant, c. Then the longitudinal structure of the solution is governed by

$$\frac{d^2\Lambda}{d\lambda^2}+c\Lambda=0\,. \tag{D.12}$$

It follows that Λ must be a trigonometric function of longitude, i.e.,

$$\Lambda=A_\Lambda\exp(im\lambda)\;\text{ where }m=\pm\sqrt{c}\,, \tag{D.13}$$

and A_Λ is an arbitrary complex constant. The cyclic condition $\Lambda(\lambda+2\pi)=\Lambda(\lambda)$ implies that m is an integer. We refer to m as the zonal wave number. It is non-dimensional, and can be either positive or negative.

Notice that m has meaning only with respect to a particular spherical coordinate system. In this way m is less fundamental than n, which comes from the radial dependence of the three dimensional function S, and has a meaning that is independent of any particular spherical coordinate system.

The equation for $\Phi(\varphi)$, which determines the meridional structure of the solution, is

$$\frac{1}{\cos\varphi}\frac{d}{d\varphi}\left(\cos\varphi\frac{d\Phi}{d\varphi}\right)+\left[n(n+1)-\frac{m^2}{\cos^2\varphi}\right]\Phi=0\,, \tag{D.14}$$

Note that the zonal wave number, m, appears in this "meridional structure equation," as does the radial exponent, n. The longitude and radius dependencies have disappeared, but the zonal wave number and the radial exponent are still visible.

For convenience, we define a new independent variable to measure latitude,

$$\mu\equiv\sin\varphi\,, \tag{D.15}$$

so that $d\mu=\cos\varphi\,d\varphi$. Then (D.14) can be written as

$$\frac{d}{d\mu}\left[\left(1-\mu^2\right)\frac{d\Phi}{d\mu}\right]+\left[n(n+1)-\frac{m^2}{1-\mu^2}\right]=0\,. \tag{D.16}$$

Eq. (D.16) is simpler than (D.14), in that (D.16) does not involve trigonometric functions of the independent variable. This added simplicity is the motivation for using (D.15). The solutions of (D.16) are called the associated Legendre functions, are denoted by $P_n^m(\mu)$, and are given by

$$\begin{aligned}P_n^m(\mu)=\frac{(2n)!}{2^n n!\,(n-m)!}\left(1-\mu^2\right)^{m/2}\Bigg[\mu^{n-m}-\frac{(n-m)(n-m-1)}{2(2n-1)}\mu^{n-m-2}\\+\frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2\cdot4(2n-1)(2n-3)}\mu^{n-m-4}-\ldots\Bigg]\,.\end{aligned} \tag{D.17}$$

The superscript m and subscript n on $P_n^m(\mu)$ are just "markers" to remind us that m and n appear as parameters in (D.16), denoting the radial exponent and zonal wave number, respectively, of $S(r,\lambda,\varphi)$. The sum in (D.17) is continued out as far as necessary to include

all non-negative powers of μ. The factor in square brackets is, therefore, a polynomial of degree $n-m$, and that is why we must require that

$$n \geq m \,, \tag{D.18}$$

Substitution can be used to demonstrate that, when (D.18) is satisfied, the associated Legendre functions are indeed solutions of (D.16).

In view of the leading factor of $\left(1-\mu^2\right)^{m/2}$ in (D.17), the complete function $P_n^m(\mu)$ is a polynomial in μ for even values of m, but not for odd values of m. The functions $P_n^m(\mu)$ are said to be of "order n" and "rank m." Figure A7.1 gives some examples of associated Legendre functions, which you might want to check for their consistency with (D.17). It can be shown that the associated Legendre functions are mutually orthogonal, i.e.,

$$\begin{gathered}\int_{-1}^{1} P_n^m(\mu)\cdot P_l^m(\mu)\; d\mu = 0, \quad \text{for} \quad n \neq l \text{ and} \\ \int_{-1}^{1} \left[P_n^m(\mu)\right]^2 d\mu = \left(\frac{2}{2n+1}\right)\frac{(n+m)!}{(n-m)!}\,.\end{gathered} \tag{D.19}$$

It follows that the functions

$$\sqrt{\left(\frac{2n+1}{2}\right)\frac{(n-m)!}{(n+m)!}}P_n^m(\mu), \quad n=m, \quad m+1, \quad m+2,\ldots \tag{D.20}$$

are mutually orthonormal for $-1 \leq \mu \leq 1$.

Referring back to (D.7), we see that a particular spherical surface harmonic can be written as

$$\boxed{Y_n^m(\mu,\,\lambda) = P_n^m(\mu)\exp(im\lambda)\;.} \tag{D.21}$$

It is the product of an associated Legendre function of μ with a trigonometric function of λ. Note that the arbitrary constant has been set to unity. For a given value of n, there are $2n+1$ spherical harmonics, correspoinding to $m = 0, 1, \ldots n$.

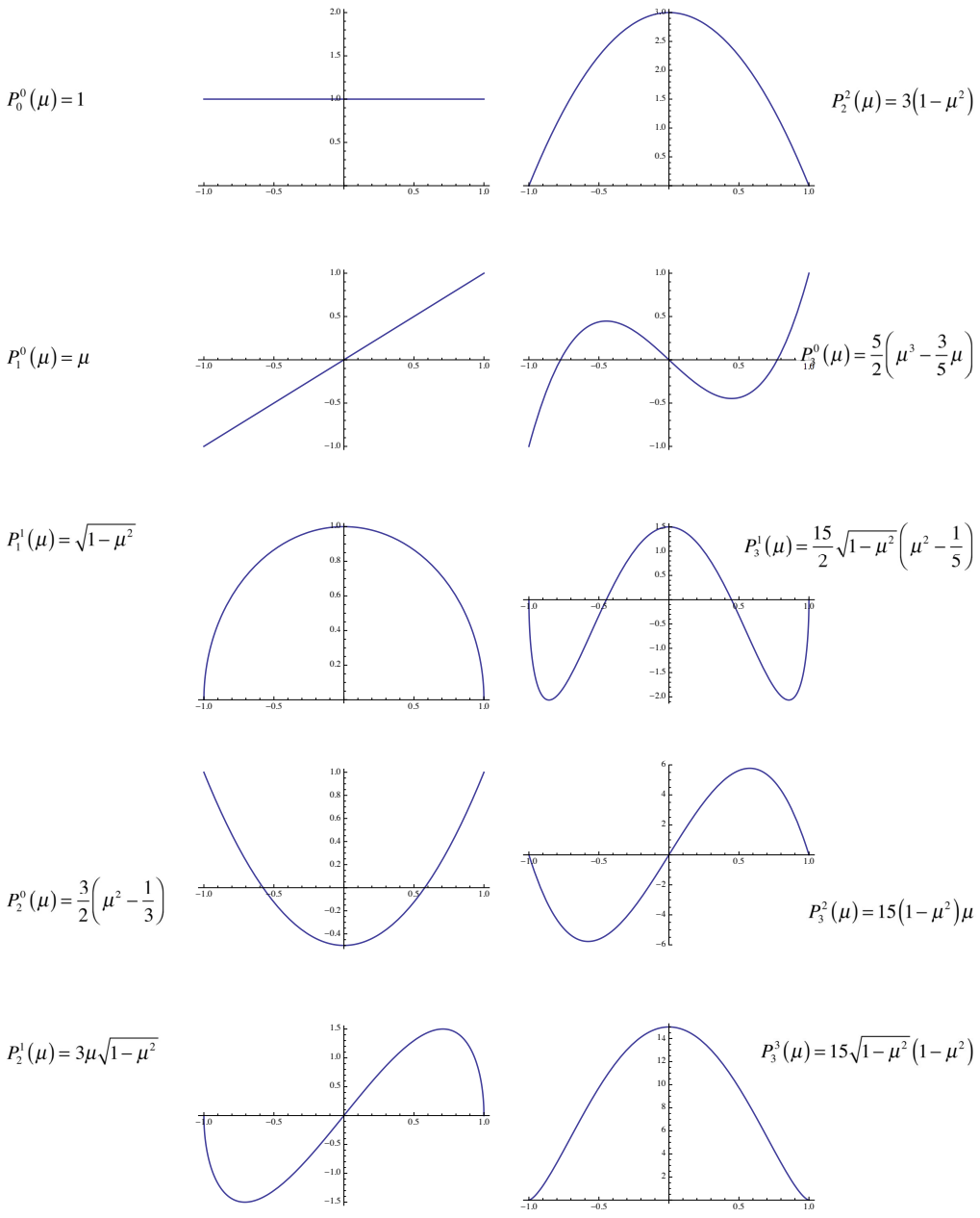


Figure D.1: Algebraic forms and plots of selected associated Legendre functions.

Fig. D.2 shows examples of spherical harmonics of low order, as mapped out onto the longitude-latitude plane. Fig. D.3 gives similar diagrams for $n = 5$ and $m = 0$, 1, 2,...,5, plotted out onto stretched spheres. Fig. D.4 shows some low-order spherical harmonics mapped onto three-dimensional pseudo-spheres, in which the local radius of the surface of the pseudo-sphere is one plus a constant times the local value of the spherical harmonic.

By using the orthogonality condition (D.19) for the associated Legendre functions, and

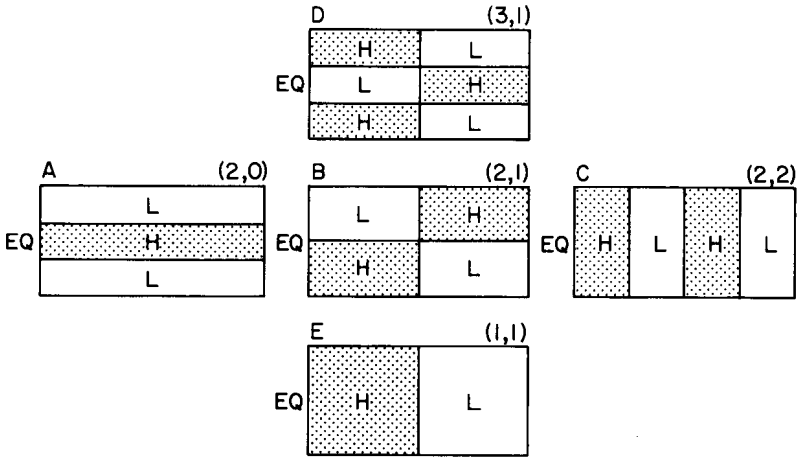


Figure D.2: Examples of low-resolution spherical harmonics, mapped out onto the plane. The horizontal direction in each panel represents longitude, and the vertical direction represents latitude. The numbers in parentheses in each panel are the appropriate values of n and m, in that order. Recall that the number of nodes in the meridional direction is $n-m$. The shading in each panel represents the sign of the field (and all signs can be flipped arbitrarily). You may think of "white" as negative and "stippled" as positive. From Washington and Parkinson (2005).

also the orthogonality properties of the trigonometric functions, we can show that

$$\int_{-1}^{1}\int_{0}^{2\pi} P_n^m(\mu)\exp(im\lambda)\,P_l^{m'}(\mu)\exp\left(im'\lambda\right)\,d\mu\,d\lambda = 0 \qquad \text{unless} \quad n=1 \quad \text{and} \quad m=m'\,, \tag{D.22}$$

The mean value over the surface of a sphere of the square of a spherical surface harmonic is given by

$$\frac{1}{4\pi}\int_{-1}^{1}\int_{0}^{2\pi}\left[P_n^m(\mu)\exp(im\lambda)\right]^2\,d\mu\,d\lambda = \frac{1}{2(2n+1)}\frac{(n+m)!}{(n-m)!} \quad \text{for} \quad s\neq 0\,. \tag{D.23}$$

For the special case $m=0$, the corresponding value is $1/(2n+1)$.

For a given n, the mean values given by (D.23) vary greatly with m, which is inconvenient for the interpretation of data. For this reason, it is customary to replace the P_n^m by the

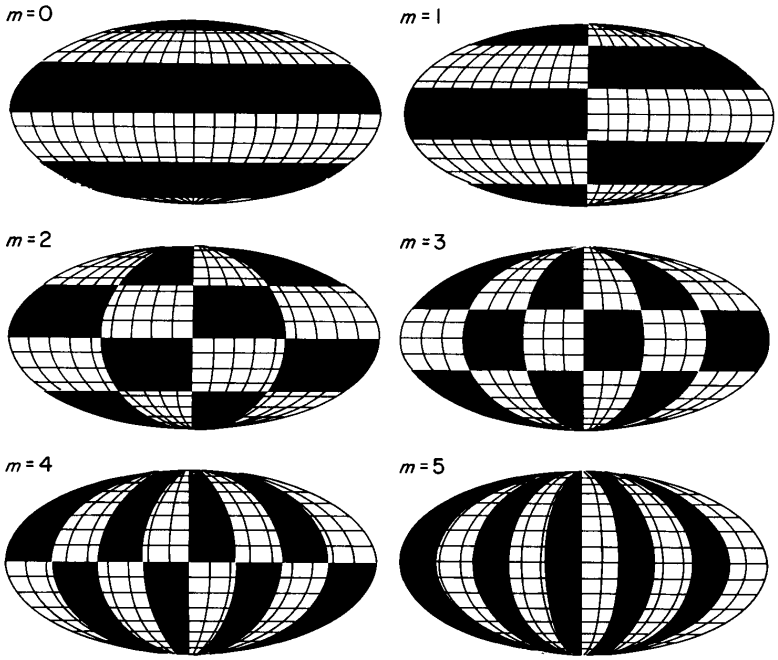


Figure D.3: Alternating patterns of positives and negatives for spherical harmonics with $n = 5$ and $m = 0, 1, 2, \ldots, 5$. From Baer (1972).

"semi-normalized associated Legendre functions," denoted by $\widehat{P}_n^m(\mu)$. These functions are identical with P_n^m when $m = 0$. For $m > 0$, the semi-normalized functions are defined by

$$\widehat{P}_n^m(\mu) \equiv \sqrt{2\frac{(n-m)!}{(n+m)!}}\, P_n^m(\mu) \ . \tag{D.24}$$

The mean value over the sphere of the square of $\widehat{P}_n^m(\mu)$ is then $1/(2n+1)$, for any values of n and m.

The spherical harmonics can be shown to form a complete orthonormal basis, and so can be used to represent an arbitrary function, $F(\lambda,\ \varphi)$ of latitude and longitude:

$$F(\lambda,\ \varphi) = \sum_{m=-\infty}^{\infty} \sum_{n=|m|}^{\infty} T_n^m Y_n^m(\lambda,\ \varphi) \ . \tag{D.25}$$

Here the T_n^m are the expansion coefficients. Note that the sum over m ranges over both positive and negative values, and that the sum over n is taken so that $n - \|m\| \geq 0$.

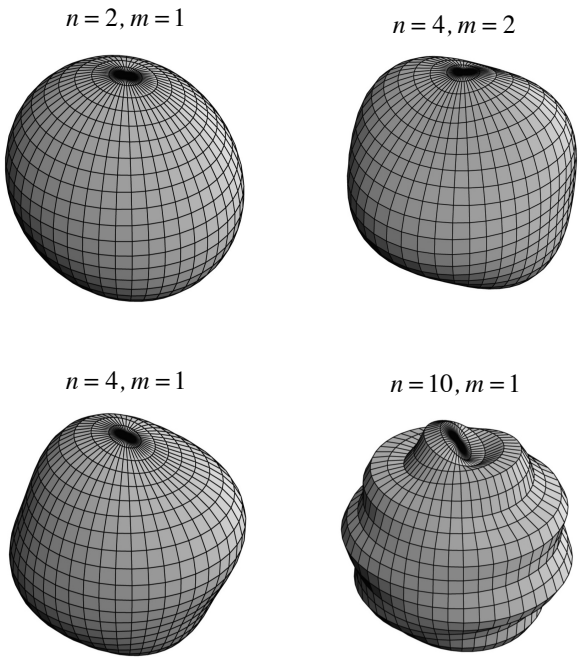


Figure D.4: Selected spherical harmonics mapped onto three-dimensional pseudo-spheres, in which the local radius of the surface of the pseudo-sphere is one plus a constant times the local value of the spherical harmonic.

The sums in (D.25) range over an infinity of terms, but in practice, of course, we have to truncate after a finite number of terms, so that (D.25) is replaced by

$$\overline{F} = \sum_{m=-M}^{M} \sum_{n=|m|}^{n(m)} T_n^m Y_n^m \ . \tag{D.26}$$

Here the overbar is a reminder that the sum is truncated. The sum over n ranges up to $N(m)$, which has to be specified somehow. The sum over m ranges from $-M$ to M. This ensures that the final result is real.

The choice of $N(m)$ fixes what is called the "truncation procedure." There are two commonly used truncation procedures. The first, called "rhomboidal," takes

$$N(m) = M + |m| \ . \tag{D.27}$$

The second, called "triangular," takes

$$N(m) = M \, . \tag{D.28}$$

Triangular truncation has the following beautiful property: In order to actually perform a spherical harmonic transform, it is necessary to adopt a spherical coordinate system (λ, φ). There are of course infinitely many such systems. There is no reason in principle that the coordinates have to be chosen in the conventional way, so that the poles of the coordinate system coincide with the Earth's poles of rotation. The choice of a particular spherical coordinate system is, therefore, somewhat arbitrary. Suppose that we choose two different spherical coordinate systems (tilted with respect to one another in an arbitrary way), perform a triangularly truncated expansion in both, then plot the results. It can be shown that *if triangular truncation is used the two maps will be identical*. This means that the arbitrary orientations of the spherical coordinate systems used had no effect whatsoever on the results obtained. The coordinate system used "disappears" at the end. Triangular truncation is very widely used today, in part because of this nice property.

Fig. D.5 shows an example based on 500 hPa height data, provided originally on a 2.5° longitude-latitude grid. The figure shows how the data look when represented by just a few spherical harmonics (top left), a few more (top right), a moderate number (bottom left) and at full 2.5° resolution. The smoothing effect of severe truncation is clearly visible.

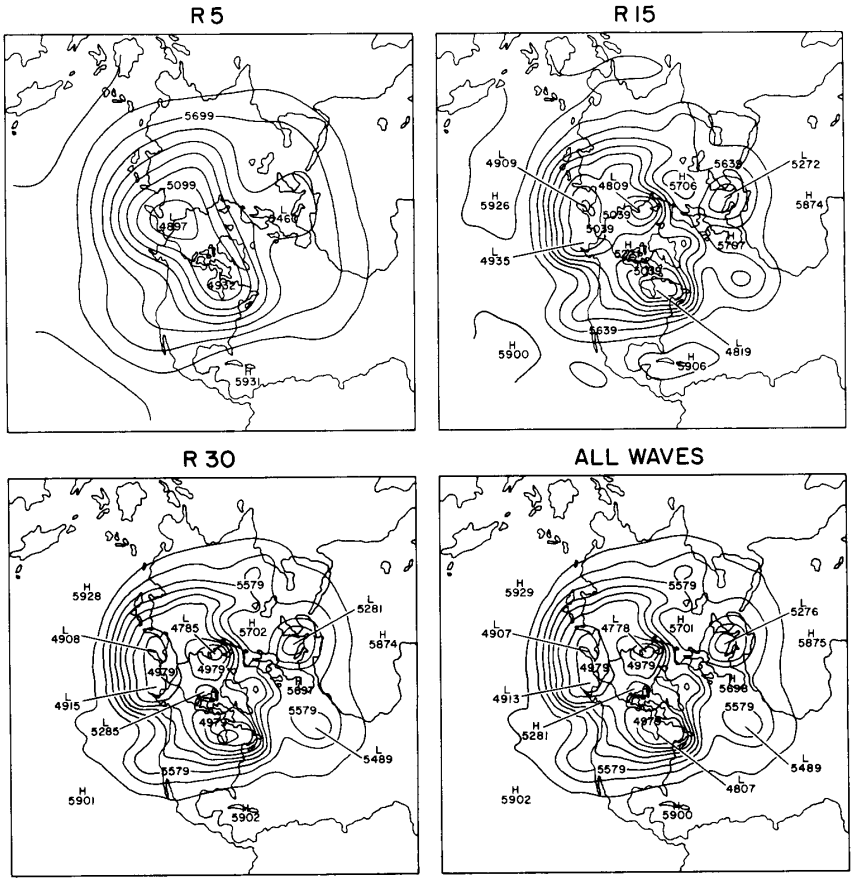


Figure D.5: A demonstration of the effects of various horizontal truncations of 500 hPa geopotential height (m) data. Th e original data are provided on a 2.5° longitude-latitude grid. From Washington and Parkinson (2005).

Bibliography

Abbrescia, M., and Coauthors, 2022: Observation of rayleigh-lamb waves generated by the 2022 hunga-tonga volcanic eruption with the pola detectors at ny-ålesund. *Scientific Reports*, **12**, URL https://api.semanticscholar.org/CorpusID:253710067.

Allen, D. N., 1954: *Relaxation Methods in Engineering and Science*. McGraw-Hill.

Arakawa, A., 1966: Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I. *Journal of Computational Physics*, **1 (1)**, 119–143.

Arakawa, A., and C. S. Konor, 1996: Vertical differencing of the primitive equations based on the charney–phillips grid in hybrid &sigma–p vertical coordinates. *Monthly weather review*, **124 (3)**, 511–528.

Arakawa, A., and C. S. Konor, 2009: Unification of the anelastic and quasi-hydrostatic systems of equations. *Monthly Weather Review*, **137 (2)**, 710–726.

Arakawa, A., and V. R. Lamb, 1977: Computational design of the basic dynamical processes of the UCLA general circulation model. *Methods in computational physics*, **17**, 173–265.

Arakawa, A., and V. R. Lamb, 1981: A potential enstrophy and energy conserving scheme for the shallow water equations. *Monthly Weather Review*, **109 (1)**, 18–36.

Arakawa, A., and S. Moorthi, 1988: Baroclinic instability in vertically discrete systems. *Journal of the atmospheric sciences*, **45 (11)**, 1688–1708.

Arakawa, A., and M. J. Suarez, 1983: Vertical differencing of the primitive equations in sigma coordinates. *Monthly Weather Review*, **111 (1)**, 34–45.

Arfken, G., 1985: *Mathematical methods for physicists*. Academic Press, San Diego, 985 pp.

Asselin, R., 1972: Frequency filter for time integrations. *Mon. Wea. Rev*, **100 (6)**, 487–490.

Baer, F., 1972: An alternate scale representation of atmospheric energy spectra. *Journal of the Atmospheric Sciences*, **29 (4)**, 649–664.

Baer, F., and T. J. Simons, 1970: Computational stability and time truncation of coupled nonlinear equations with exact solutions. *Colorado State University*, Citeseer.

Bannon, P. R., 1995: Hydrostatic adjustment: Lamb's problem. *Journal of the atmospheric sciences*, **52 (10)**, 1743–1752.

Bannon, P. R., 1996: On the anelastic approximation for a compressible atmosphere. *Journal of the Atmospheric Sciences*, **53 (23)**, 3618–3628.

Bates, J., S. Moorthi, and R. Higgins, 1993: A global multilevel atmospheric model using a vector semi-lagrangian finite-difference scheme. part i: Adiabatic formulation. *Monthly weather review*, **121 (1)**, 244–263.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the ruc hybrid isentropic–terrain-following coordinate model. *Monthly Weather Review*, **132 (2)**, 473–494.

Bhattacharya, A., 2021: *The Man from the Future: The Visionary Life of John von Neumann*. Penguin UK.

Bleck, R., 1973: Numerical forecasting experiments based on the conservation of potential vorticity on isentropic surfaces. *Journal of Applied Meteorology*, **12 (5)**, 737–752.

Boris, J. P., and D. L. Book, 1973: Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *Journal of computational physics*, **11 (1)**, 38–69.

Bouteloup, Y., 1995: Improvement of the spectral representation of the earth topography with a variational method. *Monthly weather review*, **123 (5)**, 1560–1574.

Brandt, A., 1973: Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems. *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics*, Springer, 82–89.

Brandt, A., 1977: Multi-level adaptive solutions to boundary-value problems. *Mathematics of computation*, **31 (138)**, 333–390.

Browning, G. L., J. J. Hack, and P. N. Swarztrauber, 1989: A comparison of three numerical methods for solving differential equations on the sphere. *Monthly Weather Review*, **117 (5)**, 1058–1075.

Calabretta, M. R., and B. F. Roukema, 2007: Mapping on the healpix grid. *Monthly Notices of the Royal Astronomical Society*, **381 (2)**, 865–872.

Carley, J., C. Alexander, L. Wicker, C. Jablonowski, A. Clark, J. Nelson, I. Jirak, and K. Viner, 2023: Mitigation efforts to address rapid refresh forecast system (RRFS) v1 dynamical core performance issues and recommendations for RRFS v2. *U.S. Dept. of Commerce, National Centers for Environmental Prediction*, **Office Note 516**, 87.

Chang, A., H. Lee, R. Fu, and Q. Tang, 2023: A seamless approach for evaluating climate models across spatial scales. *Frontiers in Earth Science*, **11 (LLNL-JRNL-847736)**.

Charney, J. G., and N. Phillips, 1953: Numerical integration of the quasi-geostrophic equations for barotropic and simple baroclinic flows. *Journal of Atmospheric Sciences*, **10 (2)**, 71–99.

Colella, P., and P. R. Woodward, 1984: The piecewise parabolic method (ppm) for gas-dynamical simulations. *Journal of computational physics*, **54 (1)**, 174–201.

Cotter, C., J. Frank, and S. Reich, 2007: The remapped particle-mesh semi-lagrangian advection scheme. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **133 (622)**, 251–260.

Courant, R., K. Friedrichs, and H. Lewy, 1928: Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische annalen*, **100 (1)**, 32–74.

Dalal, P., B. Kundu, J. Panda, and S. Jin, 2023: Atmospheric lamb wave pulse and volcanic explosivity index following the 2022 hunga tonga (south pacific) eruption. *Frontiers in Earth Science*, URL https://api.semanticscholar.org/CorpusID:255497658.

Danielsen, E. F., 1961: Trajectories: Isobaric, isentropic and actual. *Journal of Atmospheric Sciences*, **18 (4)**, 479–486.

Davies, T., M. J. Cullen, A. J. Malcolm, M. Mawson, A. Staniforth, A. White, and N. Wood, 2005: A new dynamical core for the met office's global and regional modelling of the atmosphere. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131 (608)**, 1759–1782.

Diamantakis, M., 2013: The semi-lagrangian technique in atmospheric modelling: current status and future challenges. *ECMWF Seminar in numerical methods for atmosphere and ocean modelling*, 183–200.

Dubey, S., R. Mittal, and P. H. Lauritzen, 2014: A flux-form conservative semi-lagrangian multitracer transport scheme (ff-cslam) for icosahedral-hexagonal grids. *Journal of Advances in Modeling Earth Systems*, **6 (2)**, 332–356.

Dubos, T., and F. Voitus, 2014: A semihydrostatic theory of gravity-dominated compressible flow. *Journal of the Atmospheric Sciences*, **71 (12)**, 4621–4638.

Dukowicz, J. K., 2013: Evaluation of various approximations in atmosphere and ocean modeling based on an exact treatment of gravity wave dispersion. *Monthly weather review*, **141 (12)**, 4487–4506.

Dukowicz, J. K., and J. R. Baumgardner, 2000: Incremental remapping as a transport/advection algorithm. *Journal of Computational Physics*, **160 (1)**, 318–335.

Durran, D. R., 1989: Improving the anelastic approximation. *Journal of the atmospheric sciences*, **46 (11)**, 1453–1461.

Durran, D. R., 1991: The third-order adams-bashforth method: An attractive alternative to leapfrog time differencing. *Monthly weather review*, **119 (3)**, 702–720.

Durran, D. R., 2008: A physically motivated approach for filtering acoustic waves from the equations governing compressible stratified flow. *Journal of Fluid Mechanics*, **601**, 365–379.

Durran, D. R., and A. Arakawa, 2007: Generalizing the boussinesq approximation to stratified compressible flow. *Comptes Rendus Mécanique*, **335 (9-10)**, 655–664.

Eliasen, E., B. Machenhauer, and E. Rasmussen, 1970: *On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields*. Kobenhavns Universitet, Institut for Teoretisk Meteorologi.

Eliassen, A., 1956: *AA Procedure for Numerical Integration of the Primitive Equations of the Two-parameter Model of the Atmosphere*. University of California.

Fjørtoft, R., 1953: On the changes in the spectral distribution of kinetic energy for twodimensional, nondivergent flow. *Tellus*, **5 (3)**, 225–230.

Fulton, S. R., P. E. Ciesielski, and W. H. Schubert, 1986: Multigrid methods for elliptic problems: A review. *Monthly Weather Review*, **114 (5)**, 943–959.

Galerkin, B. G., 1915: Rods and plates. series occurring in various questions concerning the elastic equilibrium of rods and plates. *Engineers Bulletin (Vestnik Inzhenerov)*, **19**, 897–908.

Gassner, G. J., and A. R. Winters, 2021: A novel robust strategy for discontinuous galerkin methods in computational fluid mechanics: Why? when? what? where? *Frontiers in Physics*, **8**, 500 690.

Giraldo, F. X., 2020: *An Introduction to Element-Based Galerkin Methods on Tensor-Product Bases: Analysis, Algorithms, and Applications*, Vol. 24. Springer Nature.

Girard, C., and Coauthors, 2014: Staggered vertical discretization of the canadian environmental multiscale (gem) model using a coordinate of the log-hydrostatic-pressure type. *Monthly Weather Review*, **142 (3)**, 1183–1196.

Godunov, S. K., 1959: A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations. *Math. Sbornik*, **47**, 271–306.

Godunov, S. K., and I. Bohachevsky, 1959: Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematičeskij sbornik*, **47 (3)**, 271–306.

Gottlieb, D., and C.-W. Shu, 1997: On the gibbs phenomenon and its resolution. *SIAM review*, **39 (4)**, 644–668.

Haertel, P., 2019: A lagrangian ocean model for climate studies. *Climate*, **7 (3)**, 41.

Haertel, P., W. Boos, and K. Straub, 2017: Origins of moist air in global lagrangian simulations of the madden–julian oscillation. *Atmosphere*, **8 (9)**, 158.

Haertel, P., and A. Fedorov, 2012: The ventilated ocean. *Journal of Physical Oceanography*, **42 (1)**, 141–164.

Haertel, P., and Y. Liang, 2024: Potential strengthening of the madden–julian oscillation modulation of tropical cyclogenesis. *Atmosphere*, **15 (6)**, 655.

Haertel, P., K. Straub, and A. Budsock, 2015: Transforming circumnavigating kelvin waves that initiate and dissipate the madden–julian oscillation. *Quarterly Journal of the Royal Meteorological Society*, **141 (690)**, 1586–1602.

Haertel, P., K. Straub, and A. Fedorov, 2014: Lagrangian overturning and the madden–julian oscillation. *Quarterly Journal of the Royal Meteorological Society*, **140 (681)**, 1344–1361.

Haertel, P. T., and D. A. Randall, 2002: Could a pile of slippery sacks behave like an ocean? *Monthly weather review*, **130 (12)**, 2975–2988.

Haertel, P. T., D. A. Randall, and T. G. Jensen, 2004: Simulating upwelling in a large lake using slippery sacks. *Monthly weather review*, **132 (1)**, 66–77.

Haertel, P. T., and K. H. Straub, 2010: Simulating convectively coupled kelvin waves using lagrangian overturning for a convective parametrization. *Quarterly Journal of the Royal Meteorological Society*, **136 (651)**, 1598–1613.

Haertel, P. T., L. Van Roekel, and T. G. Jensen, 2009: Constructing an idealized model of the north atlantic ocean using slippery sacks. *Ocean Modelling*, **27 (3-4)**, 143–159.

Halem, M., and G. Russell, 1973: A split-grid differencing scheme for the giss model. *NASA Goddard Institute for Space Studies Research Review*, 144–200.

Haltiner, G. J., and R. T. Williams, 1980: *Numerical prediction and dynamic meteorology*. John Wiley & Sons Inc.

Hansen, J., G. Russell, D. Rind, P. Stone, A. Lacis, S. Lebedeff, R. Ruedy, and L. Travis, 1983: Efficient three-dimensional global models for climate studies: Models I and II. *Monthly Weather Review*, **111 (4)**, 609–662.

Harten, A., 1983: High resolution schemes for hyperbolic conservation laws. *Journal of computational physics*, **49 (3)**, 357–393.

Heikes, R., and D. Randall, 1995a: a: Numerical integration of the shallow water equations on a twisted icosahedral grid. part i: Basic design and results of tests. *Mon. Wea. Rev*, **123**, 1862–1880.

Heikes, R., and D. Randall, 1995b: b: Numerical integration of the shallow water equations on a twisted icosahedral grid. part ii: Grid refinement, accuracy and computational performance. *Mon. Wea. Rev*, **123**, 1881–1887.

Heikes, R. P., D. A. Randall, and C. S. Konor, 2013: Optimized icosahedral grids: Performance of finite-difference operators and multigrid solver. *Monthly Weather Review*, **141 (12)**, 4450–4469, doi:10.1175/MWR-D-12-00236.1, URL https://doi.org/10.1175/MWR-D-12-00236.1, https://doi.org/10.1175/MWR-D-12-00236.1.

Herman, G. F., and W. T. Johnson, 1978: The sensitivity of the general circulation to arctic sea ice boundaries: A numerical experiment. *Monthly Weather Review*, **106 (12)**, 1649–1664.

Hirt, C. W., A. A. Amsden, and J. Cook, 1974: An arbitrary lagrangian-eulerian computing method for all flow speeds. *Journal of computational physics*, **14 (3)**, 227–253.

Holton, J. R., 1973: An introduction to dynamic meteorology. *American Journal of Physics*, **41 (5)**, 752–754.

Holzer, M., 1996: Optimal spectral topography and its effect on model climate. *Journal of climate*, **9 (10)**, 2443–2463.

Hortal, M., and A. Simmons, 1991: Use of reduced gaussian grids in spectral models. *Monthly Weather Review*, **119 (4)**, 1057–1074.

Hoskins, B. J., 1980: Representation of the earth topography using spherical harmonics. *Monthly Weather Review*, **108 (1)**, 111–115.

Hoskins, B. J., M. McIntyre, and A. W. Robertson, 1985: On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, **111 (470)**, 877–946.

Hsu, Y.-J. G., and A. Arakawa, 1990: Numerical modeling of the atmosphere with an isentropic vertical coordinate. *Monthly Weather Review*, **118 (10)**, 1933–1959.

Hyman, J. M., R. J. Knapp, and J. C. Scovel, 1992: High order finite volume approximations of differential operators on nonuniform grids. *Physica D: Nonlinear Phenomena*, **60 (1-4)**, 112–138.

James Purser, R., 1988: Accurate numerical differencing near a polar singularity of a skipped grid. *Monthly weather review*, **116 (5)**, 1067–1076.

Janjic, Z., 1977: Pressure gradient force and advection scheme used for forecasting with steep and small scale topography. *Beiträge zur Physik der Atmosphäre*, **50 (1)**, 186–199.

Janjić, Z. I., and F. Mesinger, 1989: Response to small-scale forcing on two staggered grids used in finite-difference models of the atmosphere. *Quarterly Journal of the Royal Meteorological Society*, **115 (489)**, 1167–1176.

Jarraud, M., and A. J. Simmons, 1983: The spectral technique. *Seminar on Numerical Methods for Weather Prediction*, European Centre for Medium Range Weather Prediction, Vol. 2, 15–19.

Jiménez, J., 1994: Hyperviscous vortices. *Journal of Fluid Mechanics*, **279**, 169–176.

Johnson, D. R., and L. W. Uccellini, 1983: A comparison of methods for computing the sigma-coordinate pressure gradient force for flow over sloped terrain in a hybrid theta-sigma model. *Monthly Weather Review*, **111 (4)**, 870–886.

Kageyama, A., 2005: Dissection of a sphere and yin-yang grids. *J. Earth Simulator*, **3**, 20–28.

Kageyama, A., and T. Sato, 2004: "yin-yang grid": An overset grid in spherical geometry. *Geochemistry, Geophysics, Geosystems*, **5 (9)**.

Kalnay, E., and M. Kanamitsu, 1988: Time schemes for strongly nonlinear damping equations. *Monthly weather review*, **116 (10)**, 1945–1958.

Kalnay-Rivas, E., A. Bayliss, and J. Storch, 1977: The 4th order giss model of the global atmosphere. *Contrib. Atmos. Phys.*

Kasahara, A., 1974: Various vertical coordinate systems used for numerical weather prediction. *Monthly Weather Review*, **102 (7)**, 509–522.

Kasahara, A., and W. M. Washington, 1967: Ncar global general circulation model of the atmosphere. *Monthly Weather Review*, **95 (7)**, 389–402, doi: 10.1175/1520-0493(1967)095<0389:NGGCMO>2.3.CO;2, URL http://dx.doi.org/10.1175/1520-0493(1967)095<0389:NGGCMO>2.3.CO;2.

Klemp, J., and W. Skamarock, 2022: A constant pressure upper boundary formulation for models employing height-based vertical coordinates. *Monthly Weather Review*, **150 (8)**, 2175–2186.

Klemp, J. B., 2011: A terrain-following coordinate with smoothed coordinate surfaces. *Monthly weather review*, **139 (7)**, 2163–2169.

Klemp, J. B., and R. B. Wilhelmson, 1978: The simulation of three-dimensional convective storm dynamics. *Journal of Atmospheric Sciences*, **35 (6)**, 1070–1096.

Konor, C. S., 2014: Design of a dynamical core based on the nonhydrostatic "unified system" of equations*. *Monthly Weather Review*, **142**, 364–385, URL https://api.semanticscholar.org/CorpusID:121533621.

Konor, C. S., and A. Arakawa, 1997: Design of an atmospheric model based on a generalized vertical coordinate. *Monthly weather review*, **125 (7)**, 1649–1673.

Konor, C. S., and D. A. Randall, 2018a: Impacts of the horizontal and vertical grids on the numerical solutions of the dynamical equations–part 1: Nonhydrostatic inertia–gravity modes. *Geoscientific Model Development*, **11 (5)**, 1753–1784.

Konor, C. S., and D. A. Randall, 2018b: Impacts of the horizontal and vertical grids on the numerical solutions of the dynamical equations–part 2: Quasi-geostrophic rossby modes. *Geoscientific Model Development*, **11 (5)**, 1785–1797.

Kurihara, Y., 1965: Numerical integration of the primitive equations on a spherical grid. *Mon. Wea. Rev*, **93 (7)**, 399–415.

Laprise, R., 1992a: The euler equations of motion with hydrostatic pressure as an independent variable. *Monthly weather review*, **120 (1)**, 197–207.

Laprise, R., 1992b: The resolution of global spectral models. *Bull. Amer. Meteor. Soc*, **73 (9)**, 1453–1454.

Lauritzen, P. H., and R. D. Nair, 2008: Monotone and conservative cascade remapping between spherical grids (cars): Regular latitude–longitude and cubed-sphere grids. *Monthly Weather Review*, **136 (4)**, 1416–1432.

Lauritzen, P. H., R. D. Nair, and P. A. Ullrich, 2010: A conservative semi-lagrangian multi-tracer transport scheme (cslam) on the cubed-sphere grid. *Journal of Computational Physics*, **229 (5)**, 1401–1424.

Lauritzen, P. H., M. A. Taylor, J. Overfelt, P. A. Ullrich, R. D. Nair, S. Goldhaber, and R. Kelly, 2017: Cam-se–cslam: Consistent coupling of a conservative semi-lagrangian finite-volume method with spectral element dynamics. *Monthly Weather Review*, **145 (3)**, 833–855.

Lax, P., and B. Wendroff, 1960: Systems of conservation laws. *Communications on Pure and Applied mathematics*, **13 (2)**, 217–237.

Leith, C. E., 1965a: Convection in a six-level model atmosphere. Tech. rep., Lawrence Radiation Lab., Univ. of California, Livermore.

Leith, C. E., 1965b: Numerical simulation of the earth's atmosphere, methods. *Comput. Phys.*, **4**, 1–28.

Leonard, B., 1993: Large time-step stability of explicit one-dimensional advection schemes. Tech. Rep. NASA-TM-106203, NASA Lewis Research Center, Cleveland, Ohio.

Leonard, B., A. Lock, and M. MacVean, 1996: Conservative explicit unrestricted-time-step multidimensional constancy-preserving advection schemes. *Monthly weather review*, **124 (11)**, 2588–2606.

Lilly, D. K., 1965: On the computational stability of numerical solutions of time-dependent non-linear geophysical fluid dynamics problems. *Mon. Wea. Rev*, **93 (1)**, 11–26.

Lindberg, C., and A. J. Broccoli, 1996: Representation of topography in spectral climate models and its effect on simulated precipitation. *Journal of climate*, **9 (11)**, 2641–2659.

Lipps, F. B., and R. S. Hemler, 1982: A scale analysis of deep moist convection and some related numerical calculations. *Journal of the Atmospheric Sciences*, **39 (10)**, 2192–2210.

Lorenz, E. N., 1955: Available potential energy and the maintenance of the general circulation. *Tellus*, **7 (2)**, 157–167.

Lorenz, E. N., 1960: Energy and numerical weather prediction. *Tellus*, **12 (4)**, 364–373.

Manabe, S., and T. B. Terpstra, 1974: The effects of mountains on the general circulation of the atmosphere as identified by numerical experiments. *Journal of the atmospheric Sciences*, **31 (1)**, 3–42.

Masuda, Y., and H. Ohnishi, 1986: An integration scheme of the primitive equation model with an icosahedral-hexagonal grid system and its application to the shallow water equations. *Journal of the Meteorological Society of Japan. Ser. II*, **64**, 317–326.

Matsuno, T., 1966: Numerical integrations of the primitive equations by a simulated backward difference method. *METEOROLOGICAL SOCIETY OF JAPAN, JOURNAL*, **44**, 76–84.

Maynard, C., T. Melvin, and E. H. Müller, 2020: Multigrid preconditioners for the mixed finite element dynamical core of the lfric atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, **146 (733)**, 3917–3936.

McGregor, J. L., 1996: Semi-lagrangian advection on conformal-cubic grids. *Monthly weather review*, **124 (6)**, 1311–1322.

McRae, A. T. T., 2015: Compatible finite element methods for atmospheric dynamical cores. Ph.D. thesis, Imperial College London.

Mellor, G. L., T. Ezer, and L.-Y. Oey, 1994: The pressure gradient conundrum of sigma coordinate ocean models. *Journal of atmospheric and oceanic technology*, **11 (4)**, 1126–1134.

Mesinger, F., 1971: Numerical integration of the primitive equations with a floating set of computation points- experiments with a barotropic global model. *Monthly Weather Review*, **99 (1)**.

Mesinger, F., 1982: On the convergence and error problems of the calculation of the pressure gradient force in sigma coordinate models. *Geophysical & Astrophysical Fluid Dynamics*, **19 (1-2)**, 105–117.

Mesinger, F., 1984: A blocking technique for representation of mountains in atmospheric models. *Riv. Meteorol. Aernonaut.*, **44**, 195–202.

Mesinger, F., and Z. I. Janjic, 1985: Problems and numerical methods of the incorporation of mountains in atmospheric models. *Lectures in Applied Mathematics*, **22**, 81–120.

Miller, M. J., 1974: On the use of pressure as vertical co-ordinate in modelling convection. *Quarterly Journal of the Royal Meteorological Society*, **100**, 155–162, URL https://api.semanticscholar.org/CorpusID:123012047.

Moeng, C.-H., 1984: A large-eddy-simulation model for the study of planetary boundary-layer turbulence. *Journal of Atmospheric Sciences*, **41 (13)**, 2052–2062.

Monaghan, J. J., 1992: Smoothed particle hydrodynamics. *Annual review of Astronomy and Astrophysics*, **30**, 543–574.

Murray, R. J., 1996: Explicit generation of orthogonal grids for ocean models. *Journal of Computational Physics*, **126 (2)**, 251–273.

Nair, R. D., S. J. Thomas, and R. D. Loft, 2005: A discontinuous galerkin transport scheme on the cubed sphere. *Monthly Weather Review*, **133 (4)**, 814–828.

Namias, J., and R. G. Stone, 1940: *An introduction to the study of air mass and isentropic analysis*. American Meteorological Society,.

Navarra, A., W. Stern, and K. Miyakoda, 1994: Reduction of the gibbs oscillation in spectral model simulations. *Journal of climate*, **7 (8)**, 1169–1183.

Nitta, T., 1964: On the reflective computational wave caused by the outflow boundary condition. *Journal of the Meteorological Society of Japan. Ser. II*, **42 (4)**, 274–276.

Norris, P. M., 1996: Radiatively driven convection in marine stratocumulus clouds. Ph.D. thesis, University of California, San Diego.

Ogura, Y., and N. A. Phillips, 1962: Scale analysis of deep and shallow convection in the atmosphere. *Journal of the atmospheric sciences*, **19 (2)**, 173–179.

Orszag, S. A., 1970: Transform method for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. *Journal of Atmospheric Sciences*, **27 (6)**, 890–895.

Orszag, S. A., 1974: Fourier series on spheres. *Monthly weather review*, **102 (1)**, 56–75.

Phillips, N. A., 1957: A coordinate system having some special advantages for numerical forecasting. *Journal of Meteorology*, **14 (2)**, 184–185.

Phillips, N. A., 1959a: An example of non-linear computational instability. *The Atmosphere and the Sea in motion*, **501**.

Phillips, N. A., 1959b: Numerical integration of the primitive equations on the hemisphere. *Monthly Weather Review*, **87 (9)**, 333–345, doi:10.1175/1520-0493(1959)087<0333:NIOTPE>2.0.CO;2.

Phillips, N. A., 1974: Application of arakawa's energy conserving layer model to operational numerical weather prediction. *U.S. Dept. of Commerce, NMC*, **Office Note 104**, 40.

Platzman, G., 1954: The computational stability of boundary conditions in numerical integration of the vorticity equation. *Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie A*, **7 (1)**, 29–40.

Prather, M. J., 1986: Numerical advection by conservation of second-order moments. *Journal of Geophysical Research: Atmospheres*, **91 (D6)**, 6671–6681.

Purser, R., and M. Rančcić, 1998: Smooth quasi-homogeneous gridding of the sphere. *Quarterly Journal of the Royal Meteorological Society*, **124 (546)**, 637–647.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, **227 (1)**, 55–78.

Qaddouri, A., C. Girard, S. Z. Husain, and R. Aider, 2021: Implementation of a semi-lagrangian fully implicit time integration of the unified soundproof system of equations for numerical weather prediction. *Monthly Weather Review*, **149 (6)**, 2011–2029.

Randall, D. A., 1994: Geostrophic adjustment and the finite-difference shallow-water equations. *Monthly Weather Review*, **122 (6)**, 1371–1377.

Randall, D. A., and Coauthors, 2019: 100 years of earth system model development. *Meteorological Monographs*, **59**, 12–1.

Reich, S., 2007: An explicit and conservative remapping strategy for semi-lagrangian advection. *Atmospheric Science Letters*, **8 (2)**, 58–63.

Richtmyer, R. D., 1963: *A survey of difference methods for non-steady fluid dynamics*. 63, National Center for Atmospheric Research.

Ringler, T. D., R. P. Heikes, and D. A. Randall, 2000: Modeling the atmospheric general circulation using a spherical geodesic grid: A new class of dynamical cores. *Monthly Weather Review*, **128 (7)**, 2471–2490.

Ringler, T. D., J. Thuburn, J. B. Klemp, and W. C. Skamarock, 2010: A unified approach to energy conservation and potential vorticity dynamics for arbitrarily-structured c-grids. *Journal of Computational Physics*, **229 (9)**, 3065–3090.

Robert, A., T. L. Yee, and H. Ritchie, 1985: A semi-lagrangian and semi-implicit numerical integration scheme for multilevel atmospheric models. *Monthly Weather Review*, **113 (3)**, 388–394.

Robert, A. J., 1966: The integration of a low order spectral form of the primitive meteorological equations. *METEOROLOGICAL SOCIETY OF JAPAN, JOURNAL*, **44**, 237–245.

Roe, P. L., 1981: Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of computational physics*, **43 (2)**, 357–372.

Ronchi, C., R. Iacono, and P. S. Paolucci, 1996: The "cubed sphere": a new method for the solution of partial differential equations in spherical geometry. *Journal of Computational Physics*, **124 (1)**, 93–114.

Rossby, C.-G., and Collaborators, 1937: Isentropic analysis. *Bulletin of the american meteorological society*, **18 (6-7)**, 201–209.

Sadourny, R., 1969: Numerical integration of the primitive equations on a spherical grid with hexagonal cells. *Proceedings of the WMO/IUGG Symposium on Numerical Weather Prediction in Tokyo, Tech. Rep. of JMA, Japan Meteorological Agency*.

Sadourny, R., A. Arakawa, and Y. Mintz, 1968: *Integration of the nondivergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere*. Citeseer.

Sadourny, R., and P. Morel, 1969: A finite-difference approximation of the primitive equations for a hexagonal grid on a plane. *Monthly Weather Review*, **97 (6)**, 439–445.

Satoh, M., T. Matsuno, H. Tomita, H. Miura, T. Nasuno, and S.-I. Iga, 2008: Nonhydrostatic icosahedral atmospheric model (nicam) for global cloud resolving simulations. *Journal of Computational Physics*, **227 (7)**, 3486–3514.

Semtner, A. J., and R. M. Chervin, 1992: Ocean general circulation from a global eddy-resolving model. *Journal of Geophysical Research: Oceans*, **97 (C4)**, 5493–5550.

Shewchuk, J. R., 1994: An introduction to the conjugate gradient method without the agonizing pain. Carnegie-Mellon University. Department of Computer Science, Available on the web at http://www.cs.berkeley.edu/ jrs/.

Shu, C.-W., 2020: Essentially non-oscillatory and weighted essentially non-oscillatory schemes. *Acta Numerica*, **29**, 701–762.

Shukla, J., and Y. Sud, 1981: Effect of cloud-radiation feedback on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, **38 (11)**, 2337–2353.

Silberman, I., 1954: Planetary waves in the atmosphere. *Journal of Meteorology*, **11 (1)**, 27–34.

Simmons, A. J., and D. M. Burridge, 1981: An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. *Monthly Weather Review*, **109 (4)**, 758–766.

Skamarock, W. C., 2008: A linear analysis of the ncar ccsm finite-volume dynamical core. *Monthly Weather Review*, **136 (6)**, 2112–2119.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal voronoi tesselations and c-grid staggering. *Monthly Weather Review*, **140 (9)**, 3090–3105.

Smith, C. J., 2000: The semi-lagrangian method in atmospheric modelling. Ph.D. thesis, Citeseer.

Smolarkiewicz, P., 1991: Nonoscillatory advection schemes. *Numerical Methods in Atmospheric Models, Proceedings of Seminar on Numerical Methods in Atmospheric Models, ECMWF, Reading, UK, 9*, 235.

Southwell, R., 1940: Relaxation methods in engineering science. Oxford University Press.

Southwell, R., 1946: Relaxation Methods in Theoretical Physics. Oxford University Press.

Souza, A., and Coauthors, 2022: The flux-differencing discontinuous galerkin method applied to an idealized fully compressible nonhydrostatic dry atmosphere. *Authorea Preprints*.

Spiegel, E. A., and G. Veronis, 1960: On the boussinesq approximation for a compressible fluid. *The Astrophysical Journal*, **131**, 442.

Staniforth, A., and J. Côté, 1991: Semi-lagrangian integration schemes for atmospheric models-a review. *Monthly weather review*, **119 (9)**, 2206–2223.

Staniforth, A., and J. Thuburn, 2012: Horizontal grids for global weather and climate prediction models: a review. *Quarterly Journal of the Royal Meteorological Society*, **138 (662)**, 1–26.

Staniforth, A. N., 2022: *Global Atmospheric and Oceanic Modelling: Fundamental Equations*. Cambridge University Press.

Starr, V. P., 1945: A quasi-lagrangian system of hydrodynamical equations. *Journal of Meteorology*, **2 (4)**, 227–237.

Strang, G., 2007: *Computational science and engineering*, Vol. 1. Wellesley-Cambridge Press Wellesley.

Swinbank, R., and J. R. Purser, 2006: Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, **132 (619)**, 1769–1793.

Takacs, L. L., 1985: A two-step scheme for the advection equation with minimized dissipation and dispersion errors. *Monthly Weather Review*, **113 (6)**, 1050–1065.

Thuburn, J., 2008: Numerical wave propagation on the hexagonal c-grid. *Journal of Computational Physics*, **227 (11)**, 5836–5858.

Thuburn, J., T. D. Ringler, W. C. Skamarock, and J. B. Klemp, 2009: Numerical representation of geostrophic modes on arbitrarily structured c-grids. *Journal of Computational Physics*, **228 (22)**, 8321–8335.

Tokioka, T., 1978: Some considerations on vertical differencing. *Meteorological Society of Japan Journal*, **56**, 98–111.

Toy, M. D., and D. A. Randall, 2009: Design of a nonhydrostatic atmospheric model based on a generalized vertical coordinate. *Monthly weather review*, **137 (7)**, 2305–2330.

Trease, H. E., 1988: Three-dimensional free-lagrange hydrodynamics. *Computer Physics Communications*, **48 (1)**, 39–50.

Uccellini, L. W., and D. R. Johnson, 1979: The coupling of upper and lower tropospheric jet streaks and implications for the development of severe convective storms. *Monthly Weather Review*, **107 (6)**, 682–703.

Ullrich, P. A., P. H. Lauritzen, and C. Jablonowski, 2009: Geometrically exact conservative remapping (gecore): regular latitude–longitude and cubed-sphere grids. *Monthly Weather Review*, **137 (6)**, 1721–1741.

Van Leer, B., 1974: Towards the ultimate conservative difference scheme. ii. monotonicity and conservation combined in a second-order scheme. *Journal of computational physics*, **14 (4)**, 361–370.

Van Leer, B., 1977a: Towards the ultimate conservative difference scheme iii. upstream-centered finite-difference schemes for ideal compressible flow. *Journal of Computational Physics*, **23 (3)**, 263–275.

Van Leer, B., 1977b: Towards the ultimate conservative difference scheme. iv. a new approach to numerical convection. *Journal of computational physics*, **23 (3)**, 276–299.

Van Leer, B., 2008: Towards the ultimate conservative difference scheme i. the quest of monotonicity. *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics: Vol. I General Lectures. Fundamental Numerical Techniques July 3–7, 1972 Universities of Paris VI and XI*, Springer, 163–168.

Van Roekel, L. P., T. Ito, P. T. Haertel, and D. A. Randall, 2009: Lagrangian analysis of the meridional overturning circulation in an idealized ocean basin. *Journal of Physical Oceanography*, **39 (9)**, 2175–2193.

Voitus, F., P. Bénard, C. Kühnlein, and N. P. Wedi, 2019: Semi-implicit integration of the unified equations in a mass-based coordinate: model formulation and numerical testing. *Quarterly Journal of the Royal Meteorological Society*, **145 (725)**, 3387–3408.

Waruszewski, M., C. Kühnlein, H. Pawlowska, and P. K. Smolarkiewicz, 2018: Mpdata: Third-order accuracy for variable flows. *Journal of Computational Physics*, **359**, 361–379.

Washington, W. M., and C. Parkinson, 2005: *Introduction to three-dimensional climate modeling*. University science books.

Wicker, L. J., 2023: Assessment of convective-scale attributes of the fv3 dycore using idealized simulations. *32nd Conference on Weather Analysis and Forecasting (WAF)/28th Conference on Numerical Weather Prediction (NWP)/20th Conference on Mesoscale Processes*, AMS.

Williams, P. D., 2013: Achieving seventh-order amplitude accuracy in leapfrog integrations. *Monthly Weather Review*, **141 (9)**, 3037–3051.

Williamson, D., 1969: Numerical integration of fluid flow over triangular grids. *Mon. Wea. Rev*, **97 (12)**, 885–895.

Williamson, D. L., 1968: Integration of the barotropic vorticity equation on a spherical geodesic grid. *Tellus*, **20 (4)**, 642–653.

Williamson, D. L., 1970: Integration of the primitive barotropic model over a spherical geodesic grid. *Mon. Wea. Rev*, **98**, 512–520.

Williamson, D. L., J. B. Drake, J. J. Hack, R. Jakob, and P. N. Swarztrauber, 1992: A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of Computational Physics*, **102 (1)**, 211–224.

Williamson, D. L., and J. G. Olson, 1994: Climate simulations with a semi-Lagrangian version of the NCAR Community Climate Model. *Monthly weather review*, **122 (7)**, 1594–1610.

Williamson, D. L., and P. J. Rasch, 1994: Water vapor transport in the NCAR CCM2. *Tellus A*, **46 (1)**, 34–51.

Winninghoff, F. J., 1968: On the adjustment toward a geostrophic balance in a simple primitive equation model with application to the problems of initialization and objective analysis. *Ph.D. thesis, UCLA*.

Wurtele, M., 1961: On the problem of truncation error. *Tellus*, **13 (3)**, 379–391.

Zalesak, S. T., 1979: Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of computational physics*, **31 (3)**, 335–362.

Zhang, Y.-T., and C.-W. Shu, 2016: Eno and weno schemes. *Handbook of numerical analysis*, Vol. 17, Elsevier, 103–122.

Zhu, Z., J. Thuburn, B. J. Hoskins, and P. H. Haynes, 1992: A vertical finite-difference scheme based on a hybrid σ-θ-p coordinate. *Monthly Weather Review*, **120 (5)**, 851–862.