# An Introduction to Numerical Modeling of the Atmosphere

David A. Randall

# Contents

# Preface

Numerical modeling is one of four broad 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 four 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.

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.

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 at UCLA, as they existed in the early 1970s. Arakawa's influence is still apparent throughout the book, but especially in Chapters 2, 3, and 4. A lot of additional material has been incorporated, mainly reflecting developments in the field since the 1970s.

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

# Introduction

## 1.1 What is a model?

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.

The data by themselves are just numbers. In order to make physical sense of the data, some sort of model is needed. It might be a qualitative conceptual model, or it might be an analytical theory, or it might be a numerical model. Models provide a basis for understanding data, and also for making predictions about the outcomes of measurements.

Accordingly, a community of modelers is hard at work developing 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.

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, $\rho$ 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 equa-

tion, in which $\rho$ is a prognostic variable. A model that contains prognostic equations is solved by time integration, and initial conditions are needed for the prognostic variables.

Any variable that is not prognostic is 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 of the Earth. In this sense, the prognostic variables are primary, and the diagnostic variables are secondary.

## 1.2 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 the physical principles listed above 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 suspended in or falling through the air, or the optical properties of ice crystals) so that their direct application to practical atmospheric problems is usually thwarted by the sheer size and complexity of the atmosphere.

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

$$p = \rho RT. \tag{1.2}$$

Eq. (1.2) can be derived using the kinetic theory of gases, which is an analytical model; the ideal gas law can be called a "result" of the model. This simple formula 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.

## 1.3 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 as straightforward as it sounds, 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 represents all possibilities in form of a single (numerically generated) table or plot.

## 1.4 Physical and mathematical errors

All models entail 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 introduce 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. This means that the design of atmospheric models can never be a purely mathematical problem.

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

In short, we have to *choose* the equation system that is used to formulate the model.

Once we have settled on a suitable set of physical equations, we must devise mathemat-

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

ical methods to solve them. The mathematical methods are almost always approximate, which is another way of saying that they introduce errors. This book focuses on the mathematical methods for models in the atmospheric sciences, and especially the mathematical errors that these methods entail. We discuss how the mathematical errors can be identified, analyzed, and minimized. All your life you have been making errors. Now, finally, you get to read a book about errors.

Another source of error is the finite precision of the computer hardware. This *round-off error* is a property of the machine being used (and to some extent the details of the program). Round-off error sometimes causes problems, but usually it doesn't, and we will not worry about it in this book.

## 1.5 Discretization

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

Most atmospheric models include equations that have 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, the same results would be obtained either way, but in a discrete model the results can depend on which variable is chosen. Much more discussion of the choice of prognostic variables is given later.

There are multiple approaches to discretization. This book emphasizes grid-point methods, which are some times called finite-difference methods. The fields of the model are defined at the discrete points of a grid. The grid can and usually does span time as well as space. Derivatives are then approximated in terms of differences involving neighboring grid-point values. A finite-difference equation (or set of equations) that approximates a differential equation (or set of equations) is called a finite-difference scheme, or a grid-point scheme. Grid-point schemes can be derived by various approaches, and the derivation methods themselves are sometimes given names. Examples include finite-volume methods, finite-element methods, and semi-Lagrangian methods. This book emphasizes finite-volume methods, for reasons that will be explained as we go.

A major alternative to the finite-difference method is the spectral method, which involves 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 would 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 used in the global domain, using spherical harmonics as the basis functions.

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 *chosen* a grid-point method or a spectral method or a finite-element 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. Again, there are trade-offs.

Having settled on the shapes of the grid cells, we must *choose* where to locate the predicted quantities on the grid. There can be good reasons to locate different quantities 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. The many meanings of "accuracy" will be discussed later. More accurate schemes have smaller errors, but less accurate schemes are usually simpler and faster. Again, we have to make *choices*.

## 1.6 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 essential physics of the processes represented by the model. This book emphasizes that very basic and inadequately recognized point.

As 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). Integrating (1.1) over the whole atmosphere, with appropriate boundary conditions, we can show that

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

and so (1.1) implies 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, in connection with the continuity equation, we can refuse to accept any error in the conservation of global mass. This means that we can *choose* to design our model so that an appropriate 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 will be surprised to learn how often it is *not* done.

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.

## 1.7 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 experiments are possible, as with some micrometeorological phenomena, it is usually not possible to draw definite conclusions, because of the difficulty of separating any one physical process from the others. 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 hardly provide direct tests of theories, which are inevitably highly idealized.

Numerical modeling is a powerful tool for studying the atmosphere through an experimental approach. A numerical model simulates the physical processes that occur in the atmosphere. There are various types of numerical models, designed for various purposes. One class of models is designed for simulating the actual atmosphere as closely as possible. Examples are numerical weather prediction models and climate simulation models, which include representations of many physical processes. Direct comparisons with observations must be made for evaluation of the model results.

Once such comparisons have given us sufficient confidence that a model behaves like the real atmosphere, we can use the model 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 the Earth's mountains were removed, and of course this has been done (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 jointly in the last several decades, and this will continue for the indefinite future. Observational and theoretical studies guide the design of numerical models. It is also true that numerical simulations can suggest theoretical ideas and can be used to design and make use of 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 the Mathematics Department. Instead, we concentrate on practical aspects of the numerical solution of the specific differential

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

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 discretization get a chapter 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 boundary-value problems.

## 1.8 Where we are going in this book

Chapter 2 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 3 presents approximations to derivatives to construct an approximation to a simple differential equation, and examines some aspects of the solution of the finite-difference equation, including its numerical stability. This leads to a survey of time-differencing methods, in Chapter 4. Next comes 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 6 deals with solving sets of linear equations, a topic that has to be covered in preparation for the next chapter on diffusion. Chapter 8 focuses on waves, with an emphasis on inertia-gravity waves. Chapter 9 gives a first look at momentum advection and related issues. Chapter 10, which is quite long, deals with vertical coordinate systems and vertical differencing. Chapter 11 gives a deeper look at momentum advection, including the issues of energy and enstrophy conservation. Chapter 12 discusses finite-difference methods with spherical geometry. Chapter 13 gives a brief introduction to spectral methods. Chapter 14 introduces finite-element methods. Chapter 15 presents a closing discussion.

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, is presented by Randall et al. (2019).

# Chapter 2

# Finite-Difference Approximations to Derivatives

## 2.1 Finite-difference quotients

Consider the derivative $df/dx$, where $f = f(x)$, and $x$ is the independent variable (which 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. Thus, we usually introduce a "grid" with discrete points where the variable $f$ is defined,

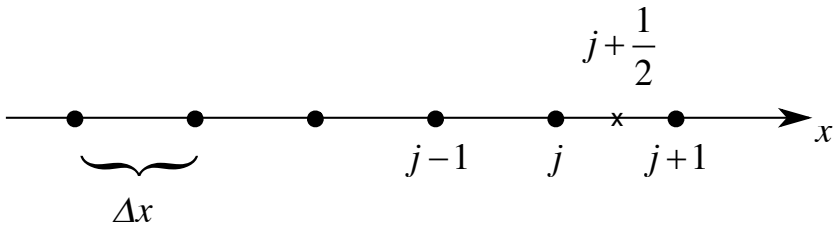



Figure 2.1: An example of a one-dimensional grid, with uniform grid spacing $\Delta x$. The grid points are denoted by the integer index $j$. Half-integer grid points can also be defined, as shown by the x at $j+\frac{1}{2}$.

as shown in Fig. 2.1. Sometimes the words "mesh" or "lattice" are used in place of "grid." The interval $\Delta x$ is called the grid spacing, grid size, mesh size, etc. For the time being, we assume that the grid spacing is constant, 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.

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

$$\text{the } \textit{forward difference} \text{ at the point } j \text{ by } f_{j+1} - f_j \tag{2.1}$$

the *backward difference* at the point $j$ by $f_j - f_{j-1}$, and (2.2)

the *centered difference* at the point $j+\frac{1}{2}$ by $f_{j+1} - f_j$. (2.3)

Note that $f$ itself is not defined at the point $j+\frac{1}{2}$. From (2.1) - (2.3) we can define the following "finite-difference quotients:"

the forward-difference quotient at the point $j$:

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1} - f_j}{\Delta x}; \tag{2.4}$$

the backward-difference quotient at the point $j$:

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_j - f_{j-1}}{\Delta x}; \tag{2.5}$$

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

$$\left(\frac{df}{dx}\right)_{j+\frac{1}{2},\,\text{approx}} = \frac{f_{j+1} - f_j}{\Delta x}. \tag{2.6}$$

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

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1} - f_{j-1}}{2\Delta x}. \tag{2.7}$$

Since (2.4) and (2.5) employ the values of $f$ at two points, and give an approximation to $df/dx$ at one of the same points, they are sometimes called *two-point approximations*. On the other hand, (2.6) and (2.7) 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 equals sign. When $x$ is time, the time point is frequently referred to as a "level." In that case, (2.4) and (2.5) can be referred to as two-level approximations and (2.6) and (2.7) as three-level approximations.

How accurate are these finite-difference approximations? "Accuracy" can be measured in a variety of ways, as we shall see. One measure of accuracy is *truncation error*, which refers to the truncation of an infinite series expansion. As an example, consider the forward difference quotient

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} = \frac{f_{j+1}-f_j}{\Delta x} \equiv \frac{f\left[(j+1)\,\Delta x\right]-f\left(j\Delta x\right)}{\Delta x}. \tag{2.8}$$

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

$$f_{j+1} = f_j + \Delta x\left(\frac{df}{dx}\right)_j + \frac{(\Delta x)^2}{2!}\left(\frac{d^2f}{dx^2}\right)_j + \frac{(\Delta x)^3}{3!}\left(\frac{d^3f}{dx^3}\right)_j + \cdots + \frac{(\Delta x)^{n-1}}{(n-1)!}\left(\frac{d^{n-1}f}{dx^{n-1}}\right)_j + \cdots \tag{2.9}$$

The expansion (2.9) can be derived without any assumptions or approximations except that the indicated derivatives exist (Arfken, 1985). When they do, the expansion is exact. 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. (2.9) can be rearranged to

$$\frac{f_{j+1}-f_j}{\Delta x} = \left(\frac{df}{dx}\right)_j + \varepsilon, \tag{2.10}$$

where

$$\varepsilon \equiv \frac{\Delta x}{2!}\left(\frac{d^2f}{dx^2}\right)_j + \frac{(\Delta x)^2}{3!}\left(\frac{d^3f}{dx^3}\right)_j + \cdots + \frac{\Delta x^{n-2}}{(n-1)!}\left(\frac{d^{n-1}f}{dx^{n-1}}\right)_j + \cdots \tag{2.11}$$

is called the "*truncation error*." If $\Delta x$ is small enough, the leading term on the right-hand side of (2.11) will be the largest part of the error. The lowest power of $\Delta x$ that appears in the truncation error is called the "*order of accuracy*" of the corresponding difference

quotient. For example, the leading term of (2.11) is of order $\Delta x$, abbreviated as $O(\Delta x)$, and so we say that (2.10) is a first-order approximation or an approximation of first-order accuracy. Obviously (2.5) is also first-order accurate.

Just to be as clear as possible, a first-order scheme for the first derivative has the form $(df/dx)_{j,\text{ approx}} = (df/dx)_j + O[\Delta x]$, 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 (2.6) and (2.7) are of second-order accuracy. For example, we can write

$$f_{j-1} = f_j + \left(\frac{df}{dx}\right)_j (-\Delta x) + \left(\frac{d^2f}{dx^2}\right)_j \frac{(-\Delta x)^2}{2!} + \left(\frac{d^3f}{dx^3}\right)_j \left[\frac{-(\Delta x)^3}{3!}\right] + \cdots \tag{2.12}$$

Subtracting (2.12) from (2.9) gives

$$f_{j+1} - f_{j-1} = 2\left(\frac{df}{dx}\right)_j (\Delta x) + \frac{2}{3!}\left(\frac{d^3f}{dx^3}\right)_j \left[(\Delta x)^3\right] + \cdots \text{odd powers only}, \tag{2.13}$$

which can be rearranged to

$$\left(\frac{df}{dx}\right)_j = \frac{f_{j+1} - f_{j-1}}{2\Delta x} - \left(\frac{d^3f}{dx^3}\right)_j \frac{\Delta x^2}{3!} + O\left[(\Delta x)^4\right]. \tag{2.14}$$

Similarly,

$$\left(\frac{df}{dx}\right)_{j+\frac{1}{2}} \cong \frac{f_{j+1} - f_j}{\Delta x} - \left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!} + O\left[(\Delta x)^4\right]. \tag{2.15}$$

From (2.14) and (2.15), we see that

$$\left|\frac{\text{Error of (2.14)}}{\text{Error of (2.15)}}\right| \cong \frac{\left(\frac{d^3f}{dx^3}\right)_j \frac{\Delta x^2}{3!}}{\left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!}} = \frac{4\left(\frac{d^3f}{dx^3}\right)_j}{\left(\frac{d^3f}{dx^3}\right)_{j+\frac{1}{2}}} \cong 4. \tag{2.16}$$

This shows that the error of (2.14) is about four times as large as the error of (2.15), 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

$$\varepsilon \cong C(\Delta x)^p, \tag{2.17}$$

where $C$ is a constant. From (2.17) we see that $\ln(\varepsilon) \cong p\ln(\Delta x) + \ln(C)$, and so

$$\frac{d[\ln(\varepsilon)]}{d[\ln(\Delta x)]} \cong p. \tag{2.18}$$

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 (approximately) a 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 compute the error of the finite-difference approximation, and that can only be done when the exact derivative is known. For that reason, this empirical approach is usually implemented by using an analytical "test function." Can you think of an approximate way to use the empirical approach even when the exact solution is not known?

## 2.2 Quantifying the accuracy of finite-difference quotients

Suppose that we write

$$\frac{f_{j+2} - f_{j-2}}{4\Delta x} = \left(\frac{df}{dx}\right)_j + \frac{1}{3!}\left(\frac{d^3f}{dx^3}\right)_j (2\Delta x)^2 + \cdots \text{even powers only.} \tag{2.19}$$

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 (2.19) is second-order accurate, although for any given value of $\Delta x$ the error of (2.19) is expected to be larger than the error of (2.14). We can combine (2.14) and (2.19) with a weight, $w$, so as to obtain a "hybrid" approximation to $(df/dx)_j$:

$$\begin{aligned}\left(\frac{df}{dx}\right)_j =& w\left(\frac{f_{j+1}-f_{j-1}}{2\Delta x}\right)+(1-w)\left(\frac{f_{j+2}-f_{j-2}}{4\Delta x}\right)\\ &-\frac{w}{3!}\left(\frac{d^3f}{dx^3}\right)_j(\Delta x)^2-\frac{(1-w)}{3!}\left(\frac{d^3f}{dx^3}\right)_j(2\Delta x)^2+O\left[(\Delta x)^4\right].\end{aligned} \tag{2.20}$$

Inspection of (2.20) shows that we can force the coefficient of $(\Delta x)^2$ to vanish by choosing

$$w+(1-w)4=0, \text{ or } w=4/3. \tag{2.21}$$

With this choice, (2.20) reduces to

$$\left(\frac{df}{dx}\right)_j=\frac{4}{3}\left(\frac{f_{j+1}-f_{j-1}}{2\Delta x}\right)-\frac{1}{3}\left(\frac{f_{j+2}-f_{j-2}}{4\Delta x}\right)+O\left[(\Delta x)^4\right]. \tag{2.22}$$

This is a fourth-order accurate scheme.

The derivation given above, in terms of a weighted combination of two second-order schemes, can be interpreted as a linear *extrapolation* of the value of the finite-difference expression to a smaller grid size, as illustrated in Fig. 2.2. 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.

Are there more systematic ways to construct schemes of any desired order of accuracy? The answer is "Yes," and one such approach is as follows. Suppose that we write a finite-difference approximation to $(df/dx)_j$ in the following somewhat generalized form:

$$\left(\frac{df}{dx}\right)_{j,\text{ approx}}\cong\frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty}a_{j'}f\left(x_j+j'\Delta x\right). \tag{2.23}$$

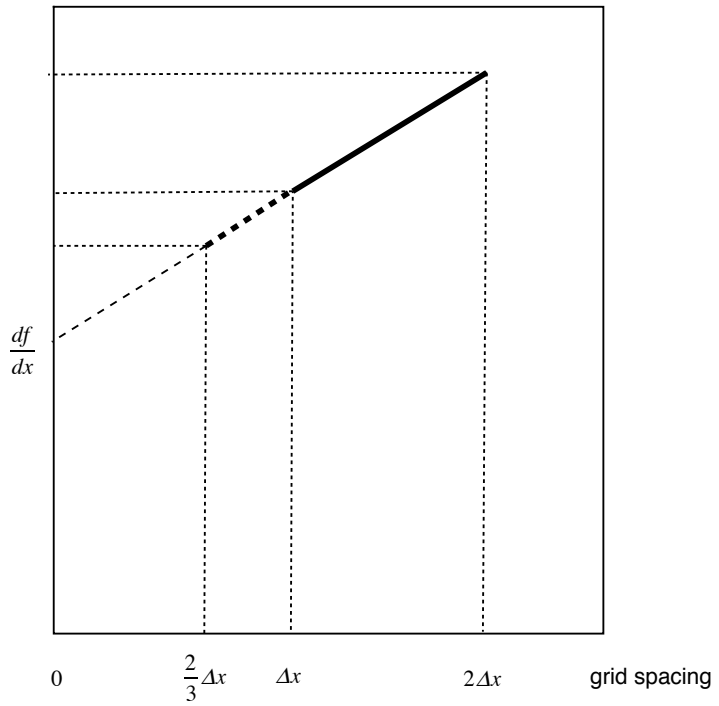



Figure 2.2: Schematic illustrating an *interpretation* of the fourth-order scheme given by (2.22) in terms of an extrapolation from two second-order schemes. The fourth-order scheme with grid spacing $\Delta x$ produces approximately the same accuracy as a second-order scheme with grid spacing $\frac{2}{3}\Delta x$.

Here the $a_{j'}$ are coefficients or "weights," which are undetermined at this point. *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 (2.23) will actually involve only a few (non-zero) terms. In writing (2.23), we have assumed for simplicity that $\Delta x$ is a constant; this assumption will be relaxed soon. The index $j'$ in (2.23) is a counter that is zero at our "home base," grid point $j$. For $j' < 0$ we count to the left, and for $j' > 0$ we count to the right. According to (2.23), 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 vicinity of point $j$. Every finite-difference approximation that we have considered so far does indeed have this form, 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

$$f\left(x_j + j'\Delta x\right) = f_j^0 + f_j^1\left(j'\Delta x\right) + f_j^2\frac{\left(j'\Delta x\right)^2}{2!} + f_j^3\frac{\left(j'\Delta x\right)^3}{3!} + \cdots \tag{2.24}$$

Here we introduce a new short-hand notation: $f_j^n$ is the $n$th derivative of $f$, evaluated at the point $j$. Using (2.24), we can rewrite (2.23) as

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'}\left[f_j^0 + f_j^1\left(j'\Delta x\right) + f_j^2\frac{\left(j'\Delta x\right)^2}{2!} + f_j^3\frac{\left(j'\Delta x\right)^3}{3!} + \cdots\right]. \tag{2.25}$$

Let's consider what happens with each of the various "pieces" of the right-hand side of (2.25). The first piece is

$$\frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'} f_j^0. \tag{2.26}$$

As $\Delta x$ goes to zero, this expression will blow up unless

$$\sum_{j'=-\infty}^{\infty} a_{j'} = 0. \tag{2.27}$$

The condition (2.27) is, therefore, a requirement that any useful scheme has to satisfy. The second piece of the right-hand side of (2.25) is

$$\frac{1}{\Delta x}\sum_{j'=-\infty}^{\infty} a_{j'} f_j^1\left(j'\Delta x\right) = f_j^1\sum_{j'=-\infty}^{\infty} a_{j'} j' \tag{2.28}$$

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

$$\sum_{j'=-\infty}^{\infty} a_{j'} j' = 1. \tag{2.29}$$

In summary, these two results show that in order to have at least first-order accuracy, we need

$$\sum_{j'=-\infty}^{\infty} a_{j'} = 0 \text{ and } \sum_{j'=-\infty}^{\infty} j' a_{j'} = 1. \tag{2.30}$$

To achieve at least second-order accuracy, we must impose an additional requirement, namely that the third piece of the right-hand side of (2.25) vanishes. This leads to

$$\sum_{j'=-\infty}^{\infty} {j'}^2 a_{j'} = 0. \tag{2.31}$$

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

$$\sum_{j'=-\infty}^{\infty} {j'}^m a_{j'} = \delta_{m,1} \text{ for } 0 \leq m \leq n. \tag{2.32}$$

Here $\delta_{m,1}$ is the Kronecker delta. In order to satisfy (2.32), we must solve a system of $n+1$ linear equations for the $n+1$ unknown coefficients $a_{j'}$.

According to (2.32), a scheme of $n$th-order accuracy can be constructed by satisfying $n+1$ equations. In particular, because (2.30) 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 points is called "*compact*."

Here is a simple 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 (2.30), we get $a_0 + a_1 = 0$ and $a_1 = 1$. It follows that we must choose $a_1 = -1$. Substituting into (2.23), 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 requirements for first-order accuracy, i.e., (2.30), will actually give us the centered second-order scheme, i.e., $f_j^1 \cong \dfrac{f\left(x_{j+1}\right) - f\left(x_{j-1}\right)}{2\Delta x}$ because (2.31) is satisfied "accidentally" or "automatically" – by luck,

we manage to satisfy three equations using only two unknowns. 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.

It should now be apparent that (2.32) 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 useful to go beyond 5th-order accuracy.

Next, we work out a generalization of the family of schemes given above, for the case of (possibly) non-uniform grid spacing. Eq. (2.23) is replaced by

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} b_{j,j'} f(x_{j+j'}). \tag{2.33}$$

Note that, since $\Delta x$ is no longer a constant, the factor of $\frac{1}{\Delta x}$ that appears in (2.23) has been omitted in (2.33), and in view of this, in order to avoid notational confusion, we have replaced the symbol $a_{j'}$ by $b_{j,j'}$. Naturally, it is going to turn out that $b_{j,j'} \sim \frac{1}{\Delta x}$. The subscript $j$ is included on $b_{j,j'}$ because for a given value of $k$ the numerical values of the coefficients are different for different values of $j$, i.e., at different places on the non-uniform grid.

Similarly, (2.24) is replaced by

$$f\left(x_{j+j'}\right) = f_j^0 + f_j^1 (\Delta x)_{j,j'} + f_j^2 \frac{(\Delta x)_{j,j'}^2}{2!} + f_j^3 \frac{(\Delta x)_{j,j'}^3}{3!} + \cdots \tag{2.34}$$

Here $(\Delta x)_{j,j'} \equiv x_{j+j'} - x_j$ takes the place of $j'\Delta x$ in (2.24). Note that $(\Delta x)_{j,0} = 0$, and $(\Delta x)_{j,j'} < 0$ for $j' < 0$. Substitution of (2.34) into (2.33) gives

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} b_{j,j'} \left[ f_j^0 + f_j^1 (\Delta x)_{j,j'} + f_j^2 \frac{(\Delta x)_{j,j'}^2}{2!} + f_j^3 \frac{(\Delta x)_{j,j'}^3}{3!} + \cdots \right]. \tag{2.35}$$

To have first-order accuracy with (2.35), we must require that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} = 0 \text{ and } \sum_{j'=-\infty}^{\infty} b_{j,j'} (\Delta x)_{j,j'} = 1, \text{for all } j. \tag{2.36}$$

Compare with (2.30). It may appear that when we require first-order accuracy by enforcing (2.36), the leading term of the error in (2.35), namely $\sum_{j'=-\infty}^{\infty} b_{j,j'} f_j^2 \frac{(\Delta x)_{j,j'}^2}{2!}$, will be of order $(\Delta x)^2$, but this is not true because, as mentioned above and shown below, $b_{j,j'} \sim 1/\Delta x$. To achieve second-order accuracy with (2.35), we must require, in addition to (2.36), that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} (\Delta x)_{j,j'}^2 = 0 \text{ for all } j. \tag{2.37}$$

Eq. (2.37) 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 $n$th-order accuracy, we must require that

$$\sum_{j'=-\infty}^{\infty} b_{j,j'} (\Delta x)_{j,j'}^m = \delta_{m,1} \text{ for all } j, \text{ and for } 0 \le m \le n. \tag{2.38}$$

This is a generalization of Eq. (2.32).

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 (2.36), i.e., $b_{j,0} + b_{j,1} = 0$ and $b_{j,1} = 1/(\Delta x)_{j,1}$. We see immediately that $b_{j,0} = 1/(\Delta x)_{j,1}$. Referring back to (2.33), we see that the scheme is $f_j^1 \cong \left[f\left(x_{j+1}\right) - f\left(x_j\right)\right] / (\Delta x)_{j,1}$, which, not unexpectedly, has the same form as the result that we obtained for the case of the uniform grid. Naturally, the non-uniformity of the grid is irrelevant when we consider only two points.

To obtain a second-order accurate approximation to $f_j^1$ on an arbitrary grid, using the three points $j-1$, $j$, and $j+1$, we must require, from (2.36) that

$$b_{j,-1} + b_{j,0} + b_{j,1} = 0 \quad \text{and} \quad b_{j,-1}(\Delta x)_{j,-1} + b_{j,1}(\Delta x)_{j,1} = 1, \tag{2.39}$$

which suffice for first-order accuracy, and additionally from (2.37) that

$$b_{j,-1} (\Delta x)_{j,-1}^2 + b_{j,1} (\Delta x)_{j,1}^2 = 0. \tag{2.40}$$

Note that $(\Delta x)_{j,-1} \equiv x_{j-1} - x_j = -(\Delta x)_{j-1,1}$. The solution of this system of three linear equations can be written as

$$b_{j,-1} = \frac{(\Delta x)_{j,1}}{(\Delta x)_{j,-1}\left[(\Delta x)_{j,1} - (\Delta x)_{j,-1}\right]}, \tag{2.41}$$

$$b_{j,0} = -\left[\frac{(\Delta x)_{j,1} + (\Delta x)_{j,-1}}{(\Delta x)_{j,1}(\Delta x)_{j,-1}}\right], \tag{2.42}$$

$$b_{j,1} = \frac{-(\Delta x)_{j,-1}}{(\Delta x)_{j,1}\left[(\Delta x)_{j,1} - (\Delta x)_{j,-1}\right]}. \tag{2.43}$$

You should confirm that for the 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, we use the scheme

$$\left(\frac{df}{dx}\right)_{j,\,\text{approx}} \cong \frac{f\left(x_{j+1}\right) - f\left(x_{j-1}\right)}{x_{j+1} - x_{j-1}}. \tag{2.44}$$

By inspection, we have

$$b_{j,-1} = \frac{-1}{x_{j+1} - x_{j-1}}, \tag{2.45}$$

$$b_{j,0} = 0, \tag{2.46}$$

$$b_{j,1} = \frac{1}{x_{j+1} - x_{j-1}}. \tag{2.47}$$

Eqs. (2.45) - (2.47) do satisfy both of the conditions in (2.39), so that *the scheme does have first-order accuracy, even on the non-uniform grid*. Eq. (2.41) - (2.43) 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 (2.44) should be nearly 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 $(\Delta x)_{j+1/2} \equiv x_{j+1} - x_j$ for all $j$, and let $\Delta$ 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:

$$\begin{aligned}\left(\frac{df}{dx}\right)_{j,\text{ approx}} &= \frac{f_{j+1}-f_{j-1}}{x_{j+1}-x_{j-1}} \\ &= \frac{\left[f_j + f_j^1 (\Delta x)_{j+1/2} + \frac{1}{2!} f_j^2 (\Delta x)_{j+1/2}^2 + O\left(\Delta^3\right)\right] - \left[f_j - f_j^1 (\Delta x)_{j-1/2} + \frac{1}{2!} f_j^2 (\Delta x)_{j-1/2}^2 + O\left(\Delta^3\right)\right]}{(\Delta x)_{j+1/2} + (\Delta x)_{j-1/2}}.\end{aligned} \tag{2.48}$$

Here $(\Delta x)_{j+1/2} \equiv x_{j+1} - x_j$ and $(\Delta x)_{j-1/2} \equiv x_j - x_{j-1}$. Eq. (2.48) can be simplified to

$$\left(\frac{df}{dx}\right)_{j,\text{ approx}} = f_j^1 + \frac{1}{2!} f_j^2 \left[(\Delta x)_{j+1/2} - (\Delta x)_{j-1/2}\right] + O\left(\Delta^2\right). \tag{2.49}$$

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, this term will be second-order. For example, if $(\Delta x)_{j+1/2} = \Delta(1 + \alpha x_{j+1/2})$, then

$$\begin{aligned}\left[(\Delta x)_{j+1/2} - (\Delta x)_{j-1/2}\right] &= \Delta(1 + \alpha x_{j+1/2}) - \Delta(1 + \alpha x_{j-1/2}) \\ &= \Delta\alpha\left(x_{j+1/2} - x_{j-1/2}\right) \\ &\sim O\left(\Delta^2\right),\end{aligned} \tag{2.50}$$

provided that $\alpha \sim O(1)$ or smaller.

Next, we observe that (2.33) 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

$$\left(\frac{d^2 f}{dx^2}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} c_{j,j'} f\left(x_{j+j'}\right). \tag{2.51}$$

Obviously, it is going to turn out that $c_{j,j'} \sim \frac{1}{(\Delta x)^2}$. Substitution of (2.34) into (2.51) gives

$$\left(\frac{d^2 f}{dx^2}\right)_{j,\,\text{approx}} = \sum_{j'=-\infty}^{\infty} c_{j,j'} \left[ f_j^0 + f_j^1 (\Delta x)_{j,j'} + f_j^2 \frac{(\Delta x)_{j,j'}^2}{2!} + f_j^3 \frac{(\Delta x)_{j,j'}^3}{3!} + \cdots \right]. \tag{2.52}$$

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

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} = 0\,, \quad \sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'} = 0\,, \text{ and } \sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^2 = 2!\,, \text{ for all } j. \tag{2.53}$$

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

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^3 = 0\,, \text{ for all } j. \tag{2.54}$$

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

$$\sum_{j'=-\infty}^{\infty} c_{j,j'} (\Delta x)_{j,j'}^m = (2!)\,\Delta_{m,2} \text{ for all } j, \text{ and for } 0 \le m \le n+1. \tag{2.55}$$

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

Earlier we showed that, in general, a second-order approximation to the first derivative must involve a minimum of three grid points, because three conditions must be satisfied [i.e., (2.39) and (2.40)]. Now we see that, in general, a second-order approximation to the second derivative must involve four grid points, because four conditions must be satisfied, i.e., (2.53) and (2.54). 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) nonuniform grid, an $n$th-order accurate approximation to the $l$th derivative of $f$ takes the form

$$\left(\frac{d^l f}{dx^l}\right)_{j,\,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} d_{j,j'} f\left(x_{j+j'}\right), \tag{2.56}$$

where

$$\sum_{j'=-\infty}^{\infty} (\Delta x)_{j,j'}^m d_{j,j'} = (l!)\,\delta_{m,l} \text{ for } 0 \le m \le n+l-1. \tag{2.57}$$

This is a total of $n+l$ requirements, so in general a minimum of $n+l$ points will be needed. It is straightforward to write a computer program that will automatically generate the coefficients for a compact $n$th-order-accurate approximation to the $l$th derivative of $f$, using $n+l$ points on a nonuniform grid.

What is the meaning of (2.57) when $l=0$?

## 2.3 Extension to two dimensions

The approach presented above can be 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

$$\frac{\partial f}{\partial t} = K\nabla^2 f, \tag{2.58}$$

where $t$ is time and $K$ is a constant positive diffusion coefficient. A later chapter is entirely devoted to solving equations similar to (2.58).

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

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

$$\left(\nabla^2 f\right)_{j,\text{approx}} \cong \sum_{j'=-\infty}^{\infty} e_{j,j'} f\left(x_{j+j'}, y_{j+k}\right). \tag{2.59}$$

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 $j$ denotes a particular grid point ("home base" for this calculation), whose coordinates are $\left(x_j, y_j\right)$. Similarly, the subscript $j+k$ denotes a grid point *in the neighborhood of point* $j$, whose coordinates are $\left(x_{j+j'}, y_{j+k}\right)$. In practice, a method is needed to compute (and perhaps tabulate for later use) the appropriate values of $k$ for the grid points in the neighborhood of each $j$; for purposes of the present discussion this is an irrelevant detail.

The *two-dimensional* Taylor series is

$$\begin{aligned} f\left(x_{j+j'}, y_{j+k}\right) = f\left(x_j, y_j\right) + \left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y}\right] f + \tfrac{1}{2!}\left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y}\right]^2 f \\ + \tfrac{1}{3!}\left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y}\right]^3 f + \tfrac{1}{4!}\left[(\Delta x)_{j,j'} \frac{\partial}{\partial x} + (\Delta y)_{j,j'} \frac{\partial}{\partial y}\right]^4 f + \cdots \end{aligned} \tag{2.60}$$

which can be written out in gruesome detail as

$$
\begin{aligned}
f\left(x_{j+j'}, y_{j+k}\right) &= f\left(x_j, y_j\right) \\
&+ \left[(\Delta x)_{j,j'} f_x + (\Delta y)_{j,j'} f_y\right] \\
&+ \tfrac{1}{2!}\left[(\Delta x)^2_{j,j'} f_{xx} + 2(\Delta x)_{j,j'}(\Delta y)_{j,j'} f_{xy} + (\Delta y)^2_{j,j'} f_{yy}\right] \\
&+ \tfrac{1}{3!}\left[(\Delta x)^3_{j,j'} f_{xxx} + 3(\Delta x)^2_{j,j'}(\Delta y)_{j,j'} f_{xxy} + 3(\Delta x)_{j,j'}(\Delta y)^2_{j,j'} f_{xyy} + (\Delta y)^3_{j,j'} f_{yyy}\right] \\
&+ \tfrac{1}{4!}\left[(\Delta x)^4_{j,j'} f_{xxxx} + 4(\Delta x)^3_{j,j'}(\Delta y)_{j,j'} f_{xxxy} + 6(\Delta x)^2_{j,j'}(\Delta y)^2_{j,j'} f_{xxyy}\right. \\
&\left. + 4(\Delta x)_{j,j'}(\Delta y)^3_{j,j'} f_{xyyy} + (\Delta y)^4_{j,j'} f_{yyyy}\right] + \cdots
\end{aligned}
\tag{2.61}
$$

Here we use the notation

$$
(\Delta x)_{j,j'} \equiv x_{j+j'} - x_j \text{ and } (\Delta y)_{j,j'} \equiv y_{j+k} - y_j, \tag{2.62}
$$

and it is understood that all of the derivatives are evaluated at the point $\left(x_j, y_j\right)$. Notice the "cross terms" that involve products of $(\Delta x)_k$ and $(\Delta y)_k$, and the corresponding cross-derivatives. A more general form of (2.61) is

$$
f(\mathbf{r}+\mathbf{a}) = f(\mathbf{r}) + \sum_{n=1}^{\infty} \frac{1}{n!}(\mathbf{a}\cdot\nabla)^n f(\mathbf{r}), \tag{2.63}
$$

where $\mathbf{r}$ is a position vector, and $\mathbf{a}$ is a displacement vector (Arfken (1985), p. 309). In (2.63), the operator $(\mathbf{a}\cdot\nabla)^n$ acts on the function $f(\mathbf{r})$. You should confirm for yourself that the general form (2.63) is consistent with (2.61). Eq. (2.63) 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 polar coordinates, by using the form of $\nabla$ in that coordinate system.

Substituting from (2.61) into (2.59), we find that

$$
\begin{aligned}
\left(f_{xx}+f_{yy}\right)_{j,\text{approx}} \cong \quad & \sum_{j'=-\infty}^{\infty} e_{j,j'} \left\{ f\left(x_j,y_j\right) + \left[ (\Delta x)_{j,j'} f_x + (\Delta y)_{j,j'} f_y \right] \right. \\
& + \tfrac{1}{2!} \left[ (\Delta x)^2_{j,j'} f_{xx} + 2(\Delta x)_{j,j'} (\Delta y)_{j,j'} f_{xy} + (\Delta y)^2_{j,j'} f_{yy} \right] \\
& + \tfrac{1}{3!} \left[ (\Delta x)^3_{j,j'} f_{xxx} + 3\,(\Delta x)^2_{j,j'} (\Delta y)_{j,j'} f_{xxy} + 3(\Delta x)_{j,j'} (\Delta y)^2_{j,j'} f_{xyy} + (\Delta y)^3_{j,j'} f_{yyy} \right] \\
& + \tfrac{1}{4!} \left[ (\Delta x)^4_{j,j'} f_{xxxx} + 4\,(\Delta x)^3_{j,j'} (\Delta y)_{j,j'} f_{xxxy} + 6\,(\Delta x)^2_{j,j'} (\Delta y)^2_{j,j'} f_{xxyy} \right. \\
& \left. \left. + 4(\Delta x)_{j,j'} (\Delta y)^3_{j,j'} f_{xyyy} + (\Delta y)^4_{j,j'} f_{yyyy} \right] + \cdots \right\}.
\end{aligned}
\tag{2.64}
$$

Notice that we have expressed the Laplacian on the left-hand side of (2.64) in terms of Cartesian coordinates. The motivation 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 (2.64). *The use of Cartesian coordinates in* (2.64) *does not limit its applicability to Cartesian grids*. In other words, we can use a Cartesian coordinate system even if we are not using a Cartesian grid. Eq. (2.64) 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

$$
\sum_{j'=-\infty}^{\infty} e_{j,j'} = 0, \text{ for all } j, \tag{2.65}
$$

$$
\sum_{j'=-\infty}^{\infty} e_{j,j'} (\Delta x)_{j,j'} = 0, \text{ for all } j, \tag{2.66}
$$

$$
\sum_{j'=-\infty}^{\infty} e_{j,j'} (\Delta y)_{j,j'} = 0, \text{ for all } j, \text{ and} \tag{2.67}
$$

$$
\sum_{j'=-\infty}^{\infty} e_{j,j'} (\Delta x)^2_{j,j'} = 2!, \text{ for all } j, \tag{2.68}
$$

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta x)_{j,j'}(\Delta y)_{j,j'} = 0, \text{ for all } j, \tag{2.69}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta y)_{j,j'}^2 = 2!, \text{ for all } j. \tag{2.70}$$

From (2.59) and (2.64), it is clear that $e_{j,j'}$ is of order $\Delta^{-2}$, where $\Delta$ is a shorthand for "$\Delta x$ or $\Delta y$." Therefore, the quantities inside the sums in (2.66)-(2.67) are of order $\Delta^{-1}$, and those inside the sums in (2.68)-(2.70) are of order one. This is why (2.68)-(2.70) are required, in addition to (2.65)-(2.67), to obtain first-order accuracy.

Eq. (2.65) implies (with (2.59)) that a constant field has a Laplacian of zero, as it should. That's nice.

So far, (2.65)-(2.70) involve six equations. This means that to ensure first-order accuracy for a two-dimensional grid, six grid points are needed for the general case. 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 (2.65)-(2.70) 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:

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta x)_{j,j'}^3 = 0, \text{ for all } j, \tag{2.71}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta x)_{j,j'}^2(\Delta y)_{j,j'} = 0, \text{ for all } j, \tag{2.72}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta x)_{j,j'}(\Delta y)_{j,j'}^2 = 0, \text{ for all } j, \tag{2.73}$$

$$\sum_{j'=-\infty}^{\infty} e_{j,j'}(\Delta y)^3_{j,j'} = 0, \text{ for all } j. \tag{2.74}$$

In general, a total of ten conditions, i.e., Eqs. (2.65)-(2.74), must be satisfied to ensure second-order accuracy on a non-uniform two-dimensional grid.

If we were working in more than two dimensions, we would simply replace (2.61) 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.

## 2.4 Laplacians on rectangular grids

Consider a 3x3 block nine-points on a rectangular grid, as shown in Fig. 2.3. Because this 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, and use a two-dimensional indexing system, with counters $i$ and $j$ in the $x$ and $y$ directions, respectively.

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

$$\left(\nabla^2 f\right)_{i,j,\text{approx}} \cong \frac{f_{i,j+1} + f_{i-1,j} + f_{i,j-1} + f_{i+1,j} - 4f_{i,j}}{d^2} \text{ with the + stencil,} \tag{2.75}$$

and

$$\left(\nabla^2 f\right)_{i,j,\text{approx}} \cong \frac{f_{i+1,j+1} + f_{i-1,j+1} + f_{i-1,j-1} + f_{i+1,j-1} - 4f_{i,j}}{\left(\sqrt{2}d\right)^2} \text{ with the x stencil.} \tag{2.76}$$

Inspection 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

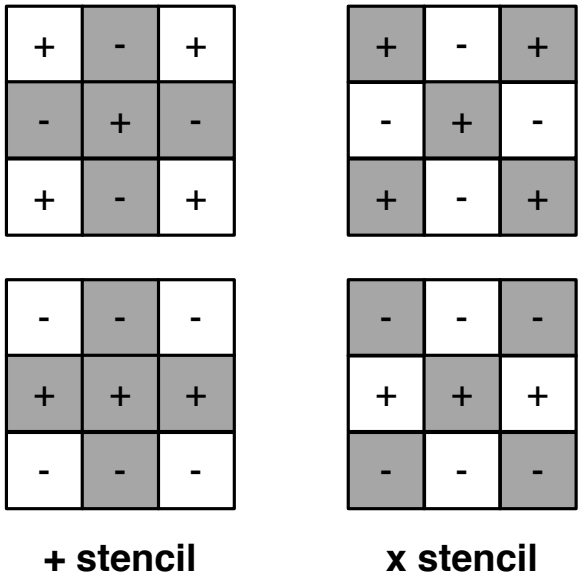



Figure 2.3: The grey shading shows two five-point stencils that can be used to create second-order Laplacians on a rectangular grid. In the upper two panels, the plus and minus symbols represent an input function that has the form of a checkerboard. In the lower two panels the plus and minus symbols represent an input function that has the form of a "rotated checkerboard," which can also be viewed as a set of horizontal stripes. (The checkerboard in the top two panels can be viewed as "rotated stripes.")

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. That's bad.

The + stencil is not necessarily the clear winner here, however. It under-estimates the strength of the "rotated checkerboard" (which can also be called "stripes") shown in the bottom two panels of Fig. 2.3, 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 (2.75) and (2.76). 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.

## 2.5 Integral properties of the Laplacian

Here comes a digression, which deals with something other than the order of accuracy of the scheme.

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:

$$\int_A \left(\nabla^2 f\right) dA = 0, \tag{2.77}$$

$$\int_A f\left(\nabla^2 f\right) dA \leq 0. \tag{2.78}$$

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. These results hold for any sufficiently differentiable function $f$. For the diffusion equation, (2.58), Eq. (2.77) implies that diffusion does not change the area-averaged value of $f$, and the inequality (2.78) implies that diffusion reduces the area-average of the square of $f$.

The finite-difference requirements corresponding to (2.77) and (2.78) are

$$\sum_{\text{all}\,j} \left(\nabla^2 f\right)_{j,\text{approx}} A_j \cong \sum_{\text{all}\,j} \left[\sum_{j'=-\infty}^{\infty} e_{j,j'} f\left(x_{j+j'}, y_{j+k}\right)\right] A_j = 0, \tag{2.79}$$

and

$$\sum_{\text{all}\,j} f_j \left(\nabla^2 f\right)_{j,\text{approx}} A_j \cong \sum_{\text{all}\,j} f_j \left[\sum_{j'=-\infty}^{\infty} e_{j,j'} f\left(x_{j+j'}, y_{j+k}\right)\right] A_j \leq 0, \tag{2.80}$$

where $A_j$ is the area of grid-cell $j$ . Suppose that we want to satisfy (2.79)-(2.80) *regardless of the numerical values assigned to* $f\left(x_j, y_j\right)$. 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 (2.79) 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 sum on the right-hand side of (2.79). We can "collect the coefficients" of each value of $f\left(x_j, y_j\right)$, and require that the sum of the coefficients is zero. To see how this works, define $j' \equiv j + k$, and write

$$\begin{aligned}\sum_{\text{all}\,j}\left[\sum_{j'=-\infty}^{\infty} e_{j,k} f\left(x_{j+j'}, y_{j+k}\right)\right] A_j &= \sum_{\text{all}\,j}\left[\sum_{\text{all}\,j'} e_{j,j'-j} f\left(x_{j'}, y_{j'}\right) A_j\right] \\ &= \sum_{\text{all}\,j'}\left[\sum_{\text{all}\,j} e_{j,j'-j} f\left(x_{j'}, y_{j'}\right) A_j\right] \\ &= \sum_{\text{all}\,j'}\left[f\left(x_{j'}, y_{j'}\right)\left(\sum_{\text{all}\,j} e_{j,j'-j} A_j\right)\right] \\ &= 0 \quad .\end{aligned} \tag{2.81}$$

In the second line above, we simply change the order of summation. The last equality above is a re-statement of the requirement (2.79). The only way to satisfy it for an arbitrary distribution of $f\left(x_j, y_j\right)$ is to write

$$\sum_{\text{all}\,j} e_{j,j'-j} A_j = 0 \text{ for each } j', \tag{2.82}$$

which is equivalent to

$$\sum_{\text{all}\,j} e_{j,j'} A_j = 0 \text{ for each } k. \tag{2.83}$$

If the $e_{j,j'}$ satisfy (2.83) then (2.79) 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 11.

A finite-difference scheme with a property similar to (2.78) is discussed in Chapter 7.

## 2.6 Why be square?

As is well known, only three regular polygons tile the plane: equilateral triangles, squares, and hexagons. Fig. 2.4 shows planar grids made from each of these three possible polygonal elements.

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

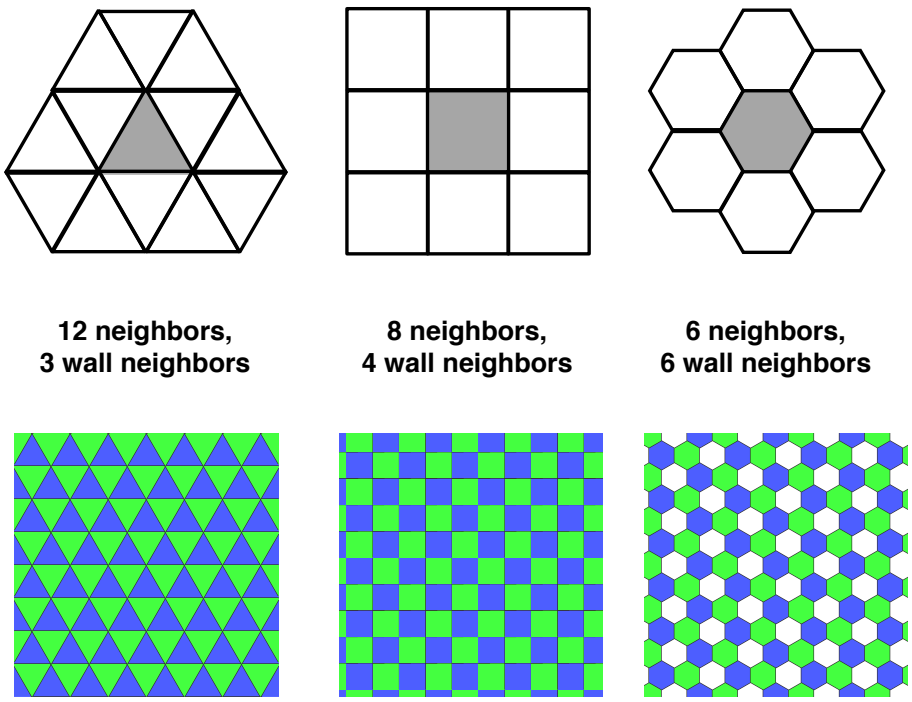



Figure 2.4: The upper row of figures shows small sections of grids made up of equilateral triangles (left), squares (center), and hexagons (right). These are the only regular polygons that tile the plane. The hexagonal grid has the highest symmetry. For example, all neighboring cells of a given hexagonal 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.

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. Although 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 (e.g., the Arakawa Jacobian, as discussed by Arakawa (1966) and in Chapter 11), the essential anisotropies of these grids remain, and are unavoidably manifested in the forms of the finite-difference operators.

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.

Further discussion of non-quadrilateral grids is given later.

## 2.7 Summary

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.

## 2.8 Problems

1. Prove that a finite-difference scheme with errors of order $n$ gives exact derivatives for polynomial functions of degree $n$ or less. For example, a first-order scheme gives exact derivatives for linear functions. Give a completely general proof; avoid making any unnecessary assumptions.

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

   (a) a first-order scheme, and

   (b) a second-order scheme

   to approximate $(df/dx)_{x=x_1}$. For each case, plot the log of the absolute value of the total error of these approximations as a function of the log of $\Delta x$. You can find the total error by subtracting the approximate derivative from the exact derivative. By inspection of the plot, verify that for sufficiently small $\Delta x$ 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 $\Delta x$ decreases. For sufficiently large values of $\Delta x$, 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 $\Delta x$ that are large enough to show the departures from straight lines.

3. Both the + and x stencils 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.

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

5. For this problem, use equations (2.65)-(2.74), which state the requirements for second-order accuracy of the Laplacian.

   *To standardize the notation, in all cases let d be the distance between grid-cell centers, as measured across cell walls. Write each of your solutions in terms of d, as was done in* (2.75) *and* (2.76).

   (a) Consider a plane tiled 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. Make a sketch to explain your notation for distances in the $x$ and $y$ directions. *Hint*: You can drastically simplify the problem by taking advantage of the high degree of symmetry.

   (b) Repeat for a plane tiled with equilateral triangles. You will have to figure out which and how many cells to use, in addition to the central cell. Use the centroids of the triangles to assign their positions. Make a sketch to explain your notation for distances in the $x$ and $y$ directions. *Hint*: You can drastically simplify the problem by taking advantage of the high degree of symmetry.

   (c) Repeat for a plane tiled with squares, using the central cell and the neighboring cells that lie diagonally across cell corners. Make a sketch to explain your notation for distances in the $x$ and $y$ directions. *Hint*: You can drastically simplify the problem by taking advantage of the high degree of symmetry.

   (d) For the triangular and hexagonal grids, and *assuming a doubly periodic domain on a plane*, can you find an example of a "checkerboard" for which the Laplacian as computed by your second-order-accurate scheme is zero throughout the entire domain even when the input field is not constant? Fig. 2.4 may help you with this question. I am not asking you to answer the question for the square grid because the answer is already given in the text.

6. (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 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. Therefore you have to 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. The period in the $x$-direction cannot be exactly the same as the period in the $y$-direction, but it can be pretty close. Make sure that the boundary conditions are truly periodic on your grid, so that no discontinuities occur. Fig. 2.5 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.

7. Consider a scheme for the Laplacian on a square grid that is given by a linear combination of (2.75) and (2.76). These are both second-order schemes. Is it possible to combine them to obtain a fourth-order scheme?

8. 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 $\theta$. Prove by direct calculation that $f_{xx} + f_{yy} = f_{x'x'} + f_{y'y'}$. This demonstrates that the Laplacian is invariant with respect to rotations of a Cartesian coordinate system. In fact, it can be demonstrated the 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 in fact the Laplacian can be defined without using a coordinate system.

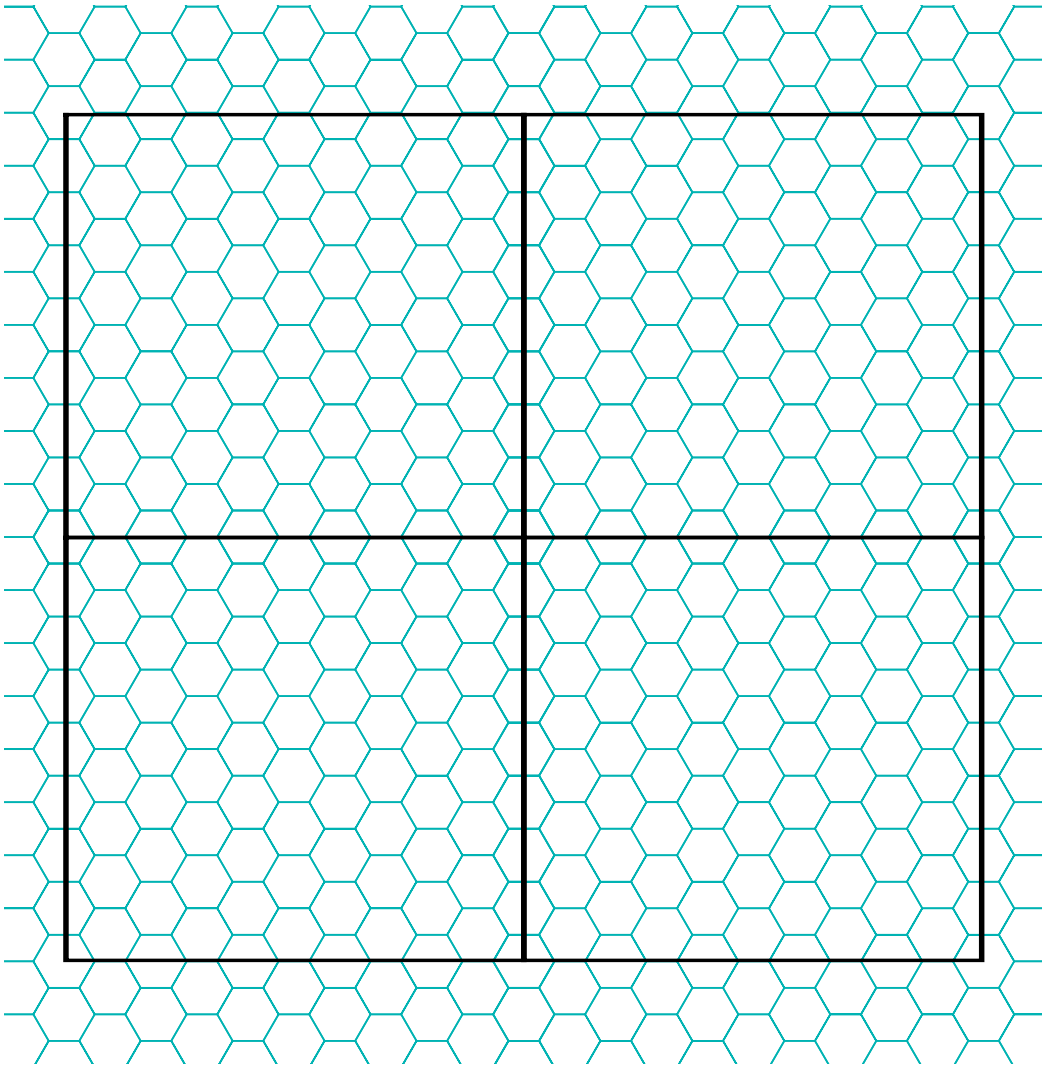


Figure 2.5: Sketch showing four copies of a periodic, nearly square domain on a (low-resolution) hexagonal grid. All four corners of each domain are located in the centers of hexagonal cells. The vertical edges of each domain run through vertical stacks of cells. Obviously the picture could be rotated by 90° (or any other angle) without changing the idea.

# Chapter 3

# Some Time-Differencing Schemes

## 3.1 Introduction

We have already analyzed the accuracy of finite-difference quotients. Now we analyze the accuracy of *finite-difference schemes*. A finite-difference scheme is defined as a finite-difference equation that approximates, term-by-term, a differential equation. Using the methods outlined in Chapter 2, 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:

$$\frac{dq}{dt} = f\left[q(t),\, t\right]. \tag{3.1}$$

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 $t$:

$$\frac{dq}{dt} = -\kappa q + a\sin\left(\omega t\right). \tag{3.2}$$

Here the first term represents decay towards zero (assuming that $\kappa$ is positive), and the second term represents a time-dependent external forcing (e.g., from the sun) that can drive $q$ away from zero.

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

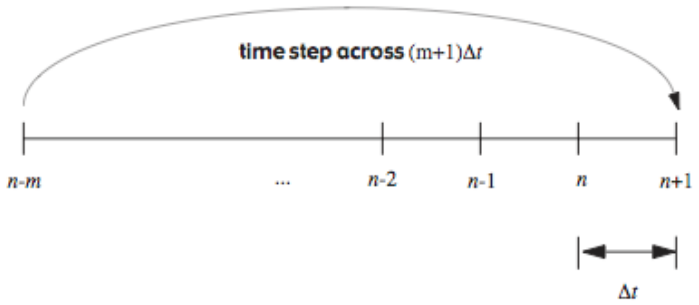



Figure 3.1: In Eq. (3.4), we use a weighted combination of $f^{n+1}$, $f^{n}$, $f^{n-1}$, ... $f^{n-l}$ to compute an average value of $f \equiv dq/dt$ over the time interval $(1+m)\,\Delta t$.

Suppose that we integrate (3.1) 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. 3.1 shows the meaning of $m$ in a graphical way. We also assume that $n \geq m$, which may not be possible close to the initial condition; this point is considered later. Integration of (3.1) gives

$$q[(n+1)\,\Delta t] - q[(n-m)\,\Delta t] = \int_{(n-m)\Delta t}^{(n+1)\Delta t} f(q,t)\,dt. \tag{3.3}$$

For larger values of $m$, the domain of integration "reaches back further" in time, before $t = n\Delta t$. Equation (3.3) is still "exact;" no finite-difference approximations have been introduced.

## 3.2 A family of schemes

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

We now approximate the integral on the right-hand side of (3.3) using the values of $f$ at the discrete time levels. Eq. (3.3), divided by $(1+m)\,\Delta t$, can be approximated by

$$\frac{q^{n+1}-q^{n-m}}{(1+m)\,\Delta t} \cong \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1} + \alpha_{n-2} f^{n-2} + \ldots \alpha_{n-l} f^{n-l}, \qquad (3.4)$$

where $l$ can be zero or a positive integer. A *family of schemes* is defined by (3.4). In Chapter 2, we discussed families of schemes that can be used to approximate finite-difference operators. In (3.4) we are defining a family of schemes that can be used to approximate a finite-difference equation.

Take a minute to look carefully at the form of (3.4), 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. 3.1. 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." The remaining time levels are in "the past." Time level $n-l$ is furthest back in the past; this is essentially the definition of $l$.

In designing the scheme, we get to choose the value of $l$. We then get to choose the values of $l+3$ parameters: namely, $m$, $\beta$, and the $l+1$ (potentially) non-zero values of $\alpha$.

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.

Here is some important and widely used terminology: A scheme is called "implicit" if it has $\beta \neq 0$, and "explicit' for' $\beta = 0$. As discussed later, implicit schemes have 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 $f^{n+1}$.

Later in this chapter we will discuss schemes that do not necessarily fit into the family defined by (3.4).

## 3.3 Discretization error

We need a way to measure the error of the time-differencing schemes defined by (3.4). To start, we need some notation. Let $q(t)$ denote the (exact) solution of the differential equation, so that $q(n\Delta t)$ is the value of this 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 point. 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, and so we cannot substitute $q^n$ into the differential equation. Instead, we substitute the *true solution*, $q(t)$, and the corresponding $f[q(t),\, t]$, into the finite-difference equation (3.4). We then expand the various terms using Taylor series around $t = n\Delta t$. The result is

$$
\begin{aligned}
&\frac{1}{(1+m)\,\Delta t}\left\{\left[q+(\Delta t)\,q'+\frac{(\Delta t)^2}{2!}q''+\frac{(\Delta t)^3}{3!}q'''+\frac{(\Delta t)^4}{4!}q''''+\ldots\right]\right.\\
&\qquad\left.-\left[q-(m\Delta t)\,q'+\frac{(m\Delta t)^2}{2!}q''-\frac{(m\Delta t)^3}{3!}q'''+\frac{(m\Delta t)^4}{4!}q''''\ldots\right]\right\}\\
&\qquad=\beta\left[f+(\Delta t)\,f'+\frac{(\Delta t)^2}{2!}f''+\frac{(\Delta t)^3}{3!}f'''+\ldots\right]\\
&\qquad+\alpha_n f\\
&\qquad+\alpha_{n-1}\left[f-(\Delta t)\,f'+\frac{(\Delta t)^2}{2!}f''-\frac{(\Delta t)^3}{3!}f'''+\ldots\right]\\
&\qquad+\alpha_{n-2}\left[f-(2\Delta t)\,f'+\frac{(2\Delta t)^2}{2!}f''-\frac{(2\Delta t)^3}{3!}f'''+\ldots\right]\\
&\qquad+\alpha_{n-3}\left[f-(3\Delta t)\,f'+\frac{(3\Delta t)^2}{2!}f''-\frac{(3\Delta t)^3}{3!}f'''+\ldots\right]\\
&\qquad+\ldots\\
&\qquad+\alpha_{n-l}\left[f-(l\Delta t)\,f'+\frac{(l\Delta t)^2}{2!}f''-\frac{(l\Delta t)^3}{3!}f'''+\ldots\right]\\
&\qquad+\varepsilon,
\end{aligned}
\tag{3.5}
$$

where $\varepsilon$ is the *discretization error* and a prime denotes a time derivative. Multiplying through by $(1+m)\,\Delta t$, collecting powers of $\Delta t$, and using

$$
q'=f,\; q''=f', \tag{3.6}
$$

etc., we can rearrange (3.5) to obtain an expression for the discretization error:

$$
\begin{aligned}
& q'\left[1-(\beta+\alpha_n+\alpha_{n-1}+\alpha_{n-2}+\alpha_{n-3}+\ldots+\alpha_{n-l})\right] \\
& +\Delta t q''\left\{\frac{1}{2}\left(\frac{1-m^2}{1+m}\right)-\beta+\alpha_{n-1}+2\alpha_{n-2}+3\alpha_{n-3}+\ldots+l\alpha_{n-l}\right\} \\
& +\frac{(\Delta t)^2}{2!}q'''\left\{\frac{1}{3}\left(\frac{1+m^3}{1+m}\right)-\beta-\alpha_{n-1}-4\alpha_{n-2}-9\alpha_{n-3}-\ldots-l^2\alpha_{n-l}\right\} \\
& +\frac{(\Delta t)^3}{3!}q''''\left\{\frac{1}{4}\left(\frac{1-m^4}{1+m}\right)-\beta+\alpha_{n-1}+8\alpha_{n-2}+27\alpha_{n-3}+\ldots+l^3\alpha_{n-1}\right\} \\
& +\ldots \\
& =\varepsilon \quad .
\end{aligned}
\tag{3.7}
$$

Each line on the left-hand side of (3.7) goes to zero "automatically" as $\Delta t \to 0$, except for the first line, which does not involve $\Delta t$ at all. We therefore have to force the first line to be zero, by requiring that

$$
1=\beta+\alpha_n+\alpha_{n-1}+\alpha_{n-2}+\alpha_{n-3}+\ldots+\alpha_{n-l}. \tag{3.8}
$$

Eq. (3.8) simply means that the sum of the coefficients on the right-hand side of (3.4) is equal to one, so that the right-hand side is a kind of "average $f$." Because (3.8) ensures at least first-order accuracy, we call it the "*consistency condition*."

When (3.8) is satisfied, the expression for the discretization error reduces to

$$
\varepsilon=\Delta t q''\left\{\frac{1}{2}\left(\frac{1-m^2}{1+m}\right)-\beta+\alpha_{n-1}+2\alpha_{n-2}+3\alpha_{n-3}+\ldots+l\alpha_{n-l}\right\}+\mathrm{O}\left[(\Delta t)^2\right]. \tag{3.9}
$$

Recall that we started with $l+3$ free parameters. Eq. (3.8) effectively uses up one of those degrees of freedom, but *we are still free to choose $l+2$ coefficients*. In particular, we can require that the coefficient of $\Delta t$ in (3.7) is also zero. This will give us a second-order scheme, i.e., one in which the error, $\varepsilon$, 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.

In summary, the order of accuracy of our time-differencing scheme can be made at least as high as $l+2$ by appropriate choices of the coefficients. One of these coefficients is $\beta$. Recall that $\beta = 0$ for explicit schemes. Generally, then, the accuracy of an explicit scheme can be made at least as high as $l+1$.

With the approach outlined above, schemes of higher-order accuracy are made possible by bringing in more time levels. It is also possible to obtain schemes of higher accuracy in other ways, which will be discussed later.

We now survey a number of time-differencing schemes, without (yet) specifying any particular form for $f$. In this analysis, we can determine the order of accuracy of each scheme. In the next Chapter, we investigate two particular choices for $f$.

## 3.4 Explicit schemes

*$m = 0$, $l = 0$ (Forward scheme or Euler scheme)*

For this simple and already-familiar case we have $\alpha_n \neq 0$, but all of the other $\alpha$'s are zero. The consistency condition, (3.8), immediately forces us to choose $\alpha_n = 1$. The scheme (3.4) then reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = f^n. \tag{3.10}$$

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

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

Better accuracy can be obtained by proper choice of the $\alpha$'s, if we use $l > 0$. For $l = 1$, the scheme reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = \alpha_n f^n + \alpha_{n-1} f^{n-1}, \tag{3.11}$$

the consistency condition, (3.8), reduces to

$$\alpha_n + \alpha_{n-1} = 1, \tag{3.12}$$

and the discretization error is

$$\varepsilon = \Delta t q'' \left( \alpha_{n-1} + \frac{1}{2} \right) + \mathrm{O}\left(\Delta t^2\right). \tag{3.13}$$

If we choose $\alpha_{n-1} = -\frac{1}{2}$ and $\alpha_n = \frac{3}{2}$, the scheme has second-order accuracy. It is called the second-order Adams-Bashforth scheme.

Although the right-hand side of (3.11) involves two different values of $f$, we only have to evaluate $f$ once per time step, if we simply save one "old" time level of $f$ for use on the next time step. Additional memory has to be allocated to save the "old" time level 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.

Table 3.1: Adams-Bashforth Schemes ($\beta = m = 0,\ l > 0$)

| $l$ | $\alpha_n$ | $\alpha_{n-1}$ | $\alpha_{n-2}$ | $\alpha_{n-3}$ | **truncation error** |
|---|---|---|---|---|---|
| 1 | $3/2$ | $-1/2$ | | | $\mathrm{O}\left(\Delta t^2\right)$ |
| 2 | $23/12$ | $-4/3$ | $5/12$ | | $\mathrm{O}\left(\Delta t^3\right)$ |
| 3 | $55/24$ | $-59/24$ | $37/24$ | $-9/24$ | $\mathrm{O}\left(\Delta t^4\right)$ |

In a similar way, we can obtain Adams-Bashforth schemes with higher accuracy by using larger $l$, and choosing the $\alpha$'s accordingly. Table 3.1 shows the results for $l = 1$, 2, and 3. See the paper by Durran (1991) for 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)*

The famous "leap-frog" scheme is given by

$$\frac{1}{2\Delta t}\left(q^{n+1} - q^{n-1}\right) = f^n. \tag{3.14}$$

From (3.7) we can immediately see that the discretization error is $\frac{\Delta t^2}{6} q''' + \mathrm{O}\left(\Delta t^4\right)$. For the leap-frog scheme, the order of accuracy 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 has been very widely used, but it has some serious disadvantages, as will be discussed later.

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

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)*

We can increase the order of accuracy by choosing appropriate values of $\alpha$ if $l > 1$.

$\underline{m > 1}$

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

## 3.5 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. (3.4) reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta f^{n+1} + \alpha_n f^n. \tag{3.15}$$

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

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

These are analogous to the Adams-Bashforth schemes, except that $\beta \neq 0$. Table 3.2 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 $\beta = 0$, which gives the leapfrog scheme.

Table 3.2: Adams-Moulton Schemes

| $l$ | $\beta$ | $\alpha_n$ | $\alpha_{n-1}$ | $\alpha_{n-2}$ | $\alpha_{n-3}$ | **truncation error** |
|---|---|---|---|---|---|---|
| 1 | $5/12$ | $8/12$ | $-1/12$ | | | $\mathrm{O}\left(\Delta t^3\right)$ |
| 2 | $9/24$ | $19/24$ | $-5/24$ | $1/24$ | | $\mathrm{O}\left(\Delta t^4\right)$ |
| 3 | $251/720$ | $646/720$ | $-264/720$ | $106/720$ | $-19/720$ | $\mathrm{O}\left(\Delta t^5\right)$ |

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

Eq. (3.4) reduces to

$$\frac{q^{n+1} - q^{n-1}}{2\Delta t} = \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1}, \tag{3.16}$$

where

$$\beta + \alpha_n + \alpha_{n-1} = 1. \tag{3.17}$$

The discretization error is

$$\varepsilon = \Delta t q'' \left(-\beta + \alpha_{n-1}\right) + \frac{\Delta t^2}{2!} q''' \left(\frac{1}{3} - \beta - \alpha_{n-1}\right) + \frac{\Delta t^3}{3!} q'''' \left(-\beta + \alpha_{n-1}\right) + \mathrm{O}\left(\Delta t^4\right). \tag{3.18}$$

[1] If there is a "Milne corrector," then there must be "Milne predictor." (See Section 4.3 for an explanation of this terminology.) In fact, the Milne predictor is an explicit scheme with

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

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.

## 3.6 Iterative schemes

Iterative schemes are sometimes called "predictor-corrector" schemes. The idea is that we 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."

An advantage of iterative schemes is that we can gain higher accuracy, without increasing $m$. A second advantage is that it may be possible to take longer time steps than with non-iterative schemes.

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

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

Consider (3.15) as an example. Change (3.15) by replacing $f^{n+1} \equiv f\left[q^{n+1}, (n+1)\,\Delta t\right]$ by $(f^{n+1})^* \equiv f\left[(q^{n+1})^*, (n+1)\,\Delta t\right]$, where $(q^{n+1})^*$ is obtained using the Euler scheme:

$$\frac{(q^{n+1})^* - q^n}{\Delta t} = f^n. \tag{3.19}$$

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

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta^* (f^{n+1})^* + \alpha_n f^n, \tag{3.20}$$

where $\beta^*$ and $\alpha_n$ are coefficients that can be chosen to obtain the desired properties of the scheme. When $\beta^* = 1,\ \alpha_n = 0$, Eq. (3.20) is an imitation of the backward (implicit) differ-

ence scheme, and is called the Euler-backward scheme or the Matsuno scheme (Matsuno (1966)). When $\beta^* = \frac{1}{2}$, $\alpha_n = \frac{1}{2}$, Eq. (3.20) 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 (3.20) does not "fit" into the framework (3.4), because $(f^{n+1})^*$ does not appear on the right-hand side of (3.4), and in general $(f^{n+1})^*$ cannot be written as a linear combination of the $f$s 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:

$$\begin{aligned} q^{n+1} &= q^n + \Delta t\,(k_1 + 2k_2 + 2k_3 + k_4/6)\,, \\ k_1 = f\,(q^n, n\Delta t)\,,\ k_2 &= f\left[q^n + k_1\Delta t/2, \left(n+\tfrac{1}{2}\right)\Delta t\right], \\ k_3 = f\left[q^n + k_2\Delta t/2, \left(n+\tfrac{1}{2}\right)\Delta t\right]\,,\ k_4 &= f\left[q^n + k_3\Delta t, (n+1)\,\Delta t\right]. \end{aligned} \tag{3.21}$$

Each of the $k$s can be interpreted as an approximation to $f$. The $k$s have to be evaluated successively, which means that the function $f$ has to be evaluated four times to take one time step. None of these $f$s 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.

Fig. 3.2 provides a simple fortran example to illustrate more clearly how the fourth-order Runge-Kutta scheme actually works. The appendix to this chapter provides a proof that the scheme really does have the advertised fourth-order accuracy.

## 3.7 What's next?

This chapter has presented a survey of some potentially useful time-differencing schemes, including a discussion of their order of accuracy. Two key issues have not yet been addressed, however. These are computational stability and computational modes in time. The next chapter will introduce those topics, in the context of two simple but important ordinary differential equations.

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

# Chapter 4

# The Oscillation and Decay Equations

## 4.1 Introduction

In this Chapter, we focus on two simple ordinary differential equations that are very relevant to atmospheric modeling. We will define and discuss how to test the computational stability of finite-difference schemes for those equations. We will also consider to what extent the solutions of the finite-difference schemes correspond to the solutions of the differential equations.

The two equations are actually quite similar to each other. The first is the "oscillation equation," which is given by

$$\frac{dq}{dt} = i\omega q \,, \tag{4.1}$$

where $i \equiv \sqrt{-1}$, $q$ is complex, and $\omega$ is real. The second is the "decay equation," which is

$$\frac{dq}{dt} = -\kappa q, \tag{4.2}$$

where $q$ and $\kappa$ are both real, and $\kappa$ is positive. The right-hand sides of both (4.1) and (4.2) are proportional to $q$, but the solutions of the two equations are very different, as discussed below. The oscillation equation (4.1) is relevant to advection and wave propagation, among other things, and the decay equation is relevant to many physical parameterizations, including diffusion, radiative transfer, cloud microphysics, and convection.

## 4.2 Computational stability

In almost all physical problems, including both (4.1) and (4.2), the true solution is bounded. We now investigate the behavior of the discretization error, $q^n - q(n\Delta t)$, as $n$ increases, for fixed $\Delta t$. *Does the error remain bounded for any initial condition*? If so, the scheme is said to be stable; otherwise it is unstable.

There are at least three ways in which the stability of a scheme can be tested. These are: 1) the *direct method*, 2) the *energy method*, and 3) *von Neumann's method*.

The direct method establishes stability by demonstrating that the largest absolute value of $q^{n+1}$ on the grid does not increase with time. We ask, "Is $\mid q^{n+1} \mid$ bounded after an arbitrary number of time steps?"

Unfortunately, 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 $\left(q^{n+1}\right)^2$ bounded after an arbitrary number of time steps?" The energy method is only applicable when $q$ is real.

For ordinary differential equations, all schemes can be written in the form

$$q^{n+1} = \lambda q^n, \tag{4.3}$$

where $\lambda$, a very important quantity, is called the *amplification factor*. With von Neumann's method, we work out the form of $\lambda$, and check its absolute value. If $\mid \lambda \mid < 1$, the scheme is stable. If $\mid \lambda \mid < 1$, the scheme is said to be neutral. If $\mid \lambda \mid > 1$, the scheme unstable.

We show in the next chapter that von Neumann's method can also be applied to separable partial differential equations.

For problems that do not involve space differencing, such as the oscillation and decay equations discussed in this chapter, von Neumann's method and the direct method are equivalent. We show in the next chapter that the two methods are quite different when space differencing is included. The energy method is distinct from the direct method and von Neumann's method, even without space differencing.

## 4.3 The oscillation equation

### 4.3.1 The solution of the continuous oscillation equation

The exact solution of (4.1) is

$$q(t) = \hat{q}e^{i\omega t}, \tag{4.4}$$

where the constant $\hat{q}$ is the "initial" value of $q$, at $t = 0$. Eq. (4.4) describes circular motion in the complex plane. The state of the system can be characterized by an amplitude and a phase. We first show that the amplitude of the solution is independent of time. Here and frequently throughout the remainder of this book we will use Euler's formula (I should say *one* of Euler's formulas), which says that

$$e^{i\gamma} = \cos\gamma + i\sin\gamma, \tag{4.5}$$

where $\gamma$ is any real number. Since $\sin^2\gamma + \cos^2\gamma = 1$, we see that

$$\left|e^{i\gamma}\right| = 1\,. \tag{4.6}$$

It follows from (4.4) and (4.6) that

$$|q| = |\hat{q}| \text{ for all time.} \tag{4.7}$$

In view of (4.7), it is reasonable to say that a "good" scheme for the oscillation equation will give a solution that has a time-independent amplitude. From (4.4) we can show that

$$q[(n+1)\Delta t] = e^{i\Omega}q(n\Delta t)\,, \tag{4.8}$$

where

$$\Omega \equiv \omega\Delta t \tag{4.9}$$

is the change in phase over the time interval $\Delta t$.

### 4.3.2 Amplitude errors and phase errors

Comparing (4.3) with (4.8), we see that the exact value of $\lambda$, based on the solution of the differential equation, is given by

$$\lambda_T = e^{i\Omega}. \tag{4.10}$$

As discussed above, Euler tells us that

$$|\lambda_T| = 1. \tag{4.11}$$

If the $|\lambda|$ of the solution of a numerical scheme for (4.1) 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 $\Omega$, 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 they still have phase errors.

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 $\lambda$. We write

$$\begin{aligned} \lambda &= \lambda_r + i\lambda_i \\ &= |\lambda|\, e^{i\theta}, \\ \text{where } \theta = \tan^{-1}\left(\frac{\lambda_i}{\lambda_r}\right), &\lambda_r = |\lambda|\cos\theta, \text{ and } \lambda_i = |\lambda|\sin\theta. \end{aligned} \tag{4.12}$$

Here $\theta$ is the change in phase per time step. Positive $\theta$ (like positive $\omega$) denotes counter-clockwise rotation in the complex plane. For example, if $\theta = \frac{\pi}{2}$, it takes four time steps to complete one oscillation. This is the case in which $\lambda$ is purely imaginary. For the case $\theta = \frac{\pi}{2}$, the discrete numerical solution may look as shown schematically in Fig. 4.1; here the ordinate represents the imaginary part of $q^n$.

From the preceding discussion, we see that the amplitude error per time step is $|\lambda| - 1$, and the phase error per time step is $\theta - \Omega$. Following Takacs (1985), the *relative error*, $\varepsilon$, can be defined by

$$\lambda \equiv (1+\varepsilon)\lambda_T. \tag{4.13}$$

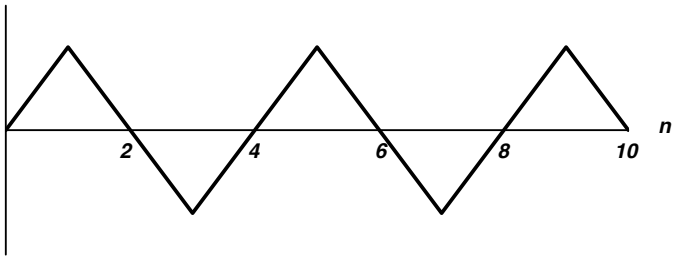



Figure 4.1: Schematic illustration of the solution of the oscillation equation for the case in which $\lambda$ is pure imaginary and the phase changes by $\theta = \frac{\pi}{2}$ on each time step. The horizontal axis is the time step counter, and the vertical axis is the imaginary part of the solution. The initial condition is assumed to be real in this example.

The relative error can be separated into its real and imaginary parts:

$$\varepsilon \equiv \varepsilon_R + i\varepsilon_I. \tag{4.14}$$

If we know $\lambda$ and $\lambda_T$, we can determine $\varepsilon_R$ and $\varepsilon_I$. With these definitions, we can write for the oscillation equation

$$\begin{aligned} |\lambda|^2 \equiv & |1+\varepsilon|^2 \\ = & (1+\varepsilon_R)^2 + \varepsilon_I^2 \\ \equiv & 1 + \varepsilon_{|\lambda|} \quad , \end{aligned} \tag{4.15}$$

where

$$\boxed{\varepsilon_{|\lambda|} = 2\varepsilon_R + \varepsilon_R^2 + \varepsilon_I^2} \tag{4.16}$$

can be interpreted as a measure of the amplitude error. Similarly, we define $\varepsilon_\varphi$ as a measure of the phase error, such that

$$\theta \equiv \Omega + \varepsilon_\varphi, \tag{4.17}$$

and it can be shown, with the use of a trigonometric identity, that

$$\boxed{\varepsilon_\varphi = \tan^{-1}\left(\frac{\varepsilon_I}{\varepsilon_R + 1}\right)}. \tag{4.18}$$

Equations (4.16) and (4.18) will be used later.

### 4.3.3 Non-iterative two-level schemes for the oscillation equation

In principle, any of the schemes described in Chapter 3 is a candidate for application to the oscillation equation, (4.1). Each scheme has its own properties, as discussed below.

A family of (possibly) implicit schemes is given by

$$q^{n+1} - q^n = i\omega\Delta t\left(\alpha q^n + \beta q^{n+1}\right). \tag{4.19}$$

We require $\alpha + \beta = 1$ in order to guarantee consistency. The explicit Euler scheme is obtained (as a special case) with $\alpha = 1,\ \beta = 0$; the backward implicit scheme with $\alpha = 0,\ \beta = 1$; and the trapezoidal-implicit scheme with $\alpha = \beta = \frac{1}{2}$. Eq. (4.19) can easily be solved for $q^{n+1}$:

$$(1 - i\Omega\beta)\, q^{n+1} = (1 + i\Omega\alpha)\, q^n, \tag{4.20}$$

or

$$\begin{aligned} q^{n+1} &= \left(\frac{1 + i\Omega\alpha}{1 - i\Omega\beta}\right) q^n \\ &\equiv \lambda q^n. \end{aligned} \tag{4.21}$$

In the second equality of (4.21) we make use of the definition of $\lambda$.

For the forward (Euler) scheme, $\alpha = 1,\ \beta = 0$, and so from (4.21) we find that

$$\lambda = 1 + i\Omega, \tag{4.22}$$

and

$$|\lambda| = \sqrt{1+\Omega^2} > 1. \tag{4.23}$$

This means that, for the oscillation equation, the forward scheme is *unconditionally unstable*. From (4.12) and (4.22), we see that the phase change per time step, $\theta$, satisfies $\tan\theta = \Omega$, so that $\theta \cong \Omega$ for small $\Delta t$, as expected.

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

$$\begin{aligned} \lambda =& \frac{1}{1-i\Omega} \\ =& \frac{1+i\Omega}{1+\Omega^2}. \end{aligned} \tag{4.24}$$

so that

$$\begin{aligned} |\lambda| =& \frac{\sqrt{1+\Omega^2}}{1+\Omega^2} \\ =& \frac{1}{\sqrt{1+\Omega^2}} < 1. \end{aligned} \tag{4.25}$$

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 $\Delta t$. The real part of $\lambda$ (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 $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. 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 $\alpha = \frac{1}{2}\ \beta = \frac{1}{2}$, we find that

$$\begin{aligned} \lambda =& \frac{1+\frac{i\Omega}{2}}{1-\frac{i\Omega}{2}} \\ =& \frac{\left(1-\frac{\Omega^2}{4}\right)+i\Omega}{1+\frac{\Omega^2}{4}} \end{aligned} \tag{4.26}$$

so that $|\lambda| = 1$. This scheme is *unconditionally stable*; in fact, it has no amplitude error at all. The phase error per time step is small. It is a very nice scheme for the oscillation equation.

### 4.3.4 Iterative schemes for the oscillation equation

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

$$q^{n+1*} - q^n = i\Omega q^n, \tag{4.27}$$

$$q^{n+1} - q^n = i\Omega \left(\alpha q^n + \beta^* q^{n+1*}\right). \tag{4.28}$$

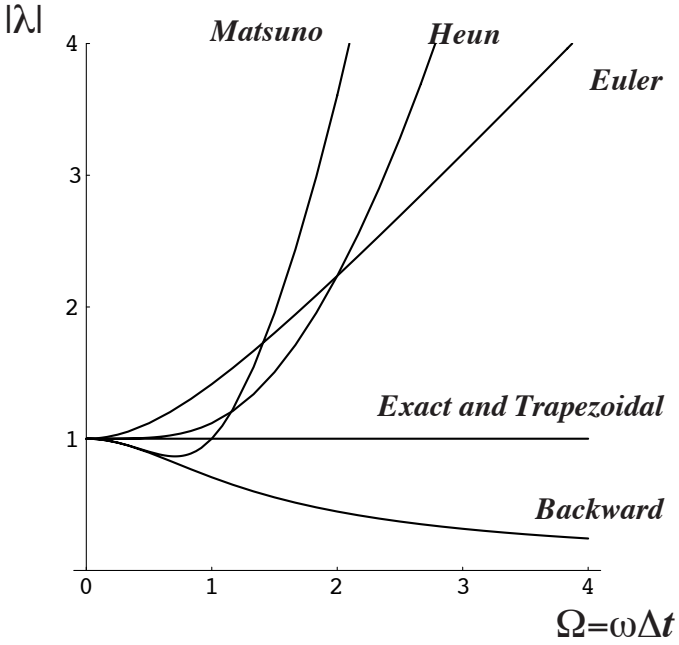



Figure 4.2: The magnitude of the amplification factor $\lambda$ as a function of $\Omega \equiv \omega\Delta t$, for various difference schemes. The Euler, backward, trapezoidal, Matsuno, and Heun schemes are shown. The magnitude of the amplification factor for the trapezoidal scheme coincides with that of the true solution for all values of $\Omega$. Caution: This does not mean that the trapezoidal scheme gives the exact solution!

Recall that $\alpha = 0$, $\beta^* = 1$ gives the Matsuno scheme, and $\alpha = \beta^* = \frac{1}{2}$ gives the Heun scheme. Eliminating $q^{n+1*}$ between (4.27) and (4.28) for the Matsuno scheme, we obtain

$$\lambda = \left(1 - \Omega^2\right) + i\Omega, \tag{4.29}$$

so that

$$|\lambda| = \sqrt{1 - \Omega^2 + \Omega^4} \begin{cases} > 1 & \text{for} \quad \Omega > 1 \\ = 1 & \text{for} \quad \Omega = 1 \\ < 1 & \text{for} \quad \Omega < 1 \, . \end{cases} \tag{4.30}$$

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

$$\lambda = \left(1 - \Omega^2/2\right) + i\Omega \, , \tag{4.31}$$

and

$$|\lambda| = \sqrt{\left(1 - \Omega^2/2\right)^2 + \Omega^2} = \sqrt{1 + \Omega^4/2} > 1 \, . \tag{4.32}$$

This shows that the Heun scheme is *unconditionally unstable* when applied to the oscillation equation, but for small $\Omega$ 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. 4.2 and Fig. 4.3.

### 4.3.5 Computational modes in time

The leapfrog scheme, which is illustrated in Fig. 4.4, is not "self-starting" because when we are sitting at $n = 0$ $q$ at $n = -1$, so we can't step from $n = -1$ to $n = +1$. A special procedure is therefore needed to start the solution. We really need two initial conditions to solve the finite-difference problem, even though only one initial condition is needed to solve the exact equation. A similar problem arises with any scheme that involves more than two time levels. One of the two required initial conditions is the "physical" initial condition

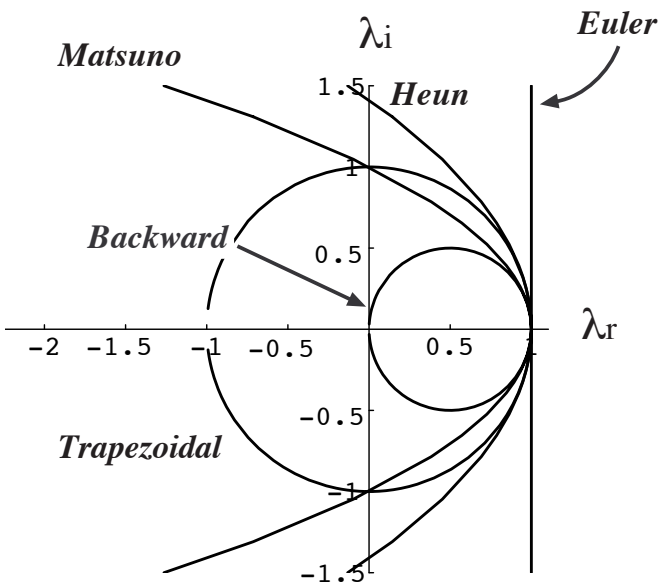



Figure 4.3: Variations of the real and imaginary components of the amplification factor, as $\Omega$ changes "parametrically." The actual values of $\Omega$ are not shown in the figure. Both the exact solution and the trapezoidal scheme lie on the unit circle.

that is 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.

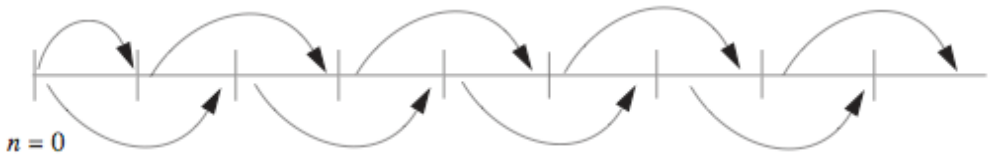



Figure 4.4: The leapfrog scheme.

The leapfrog analogue scheme for the oscillation equation is

$$q^{n+1} - q^{n-1} = 2i\Omega q^n, \tag{4.33}$$

where, as before, $\Omega \equiv \omega\Delta t$. First, consider the special case $\omega = 0$. Then (4.33) gives

$$q^{n+1} - q^{n-1} = 0, \tag{4.34}$$

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 an initial condition at $n = 1$. If you perversely make the two initial conditions different, i.e., $q^1 \neq q^0$, then an oscillation will occur, as shown in Fig. 4.5. On the other hand, if you sensibly assign $q^1 = q^0$, then the solution will be constant. This shows that judicious selection of $q^1$ is essential for schemes with more than two time levels.

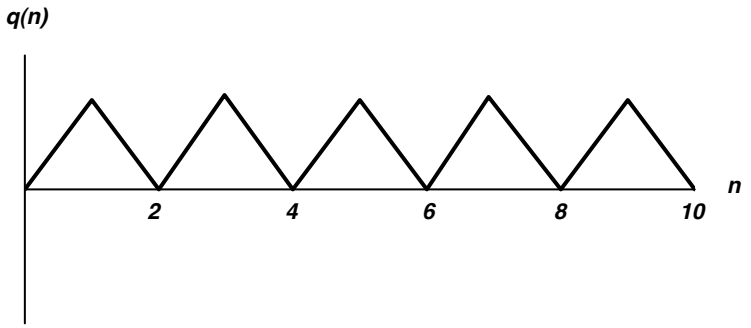



Figure 4.5: An oscillatory solution that arises with the leapfrog scheme for $\frac{dq}{dt} = 0$, in case the two initial values of $q$ are not the same.

Now rewrite (4.23) as

$$q^{n+1} - 2i\Omega q^n - q^{n-1} = 0. \tag{4.35}$$

We look for a solution of the form $q^{n+1} = \lambda q^n$, for all $n$. Then (4.35) reduces to

$$\lambda^2 - 2i\Omega\lambda - 1 = 0. \tag{4.36}$$

Make sure that you understand where (4.36) came from. The solutions of (4.36) are

$$\lambda_1 = i\Omega + \sqrt{1-\Omega^2},\ \lambda_2 = i\Omega - \sqrt{1-\Omega^2}, \tag{4.37}$$

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

$$q_1{}^{n+1} = \lambda_1 q_1{}^n,\ q_2{}^{n+1} = \lambda_2 q_2{}^n. \tag{4.38}$$

*The differential equation only has one solution, so the fact that the finite-difference equation has two solutions is bad.*

Consider the limiting values of $\lambda_1$ and $\lambda_2$ 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 a "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 a computational mode by making the time step smaller. The computational mode arises from the *structure* of the scheme.

Two-level time-differencing schemes do not have computational modes. A scheme that uses $l+1$ time levels (see the right-hand side of Eq. (3.4)) has $l$ computational modes. *The existence of computational modes is a major disadvantage of all schemes that involve more than two time levels.* There are people in the numerical modeling community who advocate using only two-time-level schemes, which are sometimes called "forward-in-time" schemes[1]. We have already seen that two-time-level schemes can be quite accurate; for example, the fourth-order Runge-Kutta scheme is a two-time-level scheme.

The current discussion is about computational modes in time. There are also computational modes in space, which will be explained later.

From (4.38) we see that after $n$ time steps our two solutions will be

$$ {q_1}^n = {\lambda_1}^n {q_1}^0, \text{ and } {q_2}^n = {\lambda_2}^n {q_2}^0. \tag{4.39} $$

The general solution is a linear combination of these two modes, i.e.,

$$ q^n = a{\lambda_1}^n q_1^0 + b{\lambda_2}^n q_2^0, \tag{4.40} $$

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

Because we have two solutions, we have to provide 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 "forward-in-time" scheme. For example, we could use the Euler forward scheme, or the Matsuno scheme, or the fourth-order Runge-Kutta scheme.

[1] This terminology is potentially confusing; it does not necessarily refer to the Euler-forward scheme.

The values of $a$ and $b$, i.e., the amplitudes of the physical and computational modes, depend in part on how we choose the computational initial condition. With reference to (4.40), the two initial conditions can be written this way:

$$q^0 = aq_1^0 + bq_2^0 \text{ for } n = 0, \tag{4.41}$$

and

$$q^1 = \lambda_1 \left(aq_1^0\right) + \lambda_2 \left(bq_2^0\right) \text{ for } n = 1. \tag{4.42}$$

Here we have used $\lambda_1$ and $\lambda_2$ to advance $q_1$ and $q_2$ from time level 0 to time level 1. We can solve (4.41) and (4.42) for $aq_1^0$ and $bq_2^0$ in terms of $q^0$ and $q^1$, and substitute the results back into (4.40). This gives

$$q^n = \frac{\left(q^1 - \lambda_2 q^0\right) {\lambda_1}^n - \left(q^1 - \lambda_1 q^0\right) {\lambda_2}^n}{\lambda_1 - \lambda_2} \,. \tag{4.43}$$

Notice that (4.43) applies for any choice of $f(q,t)$; it is not specific to the oscillation equation.

"By inspection" of (4.43), the initial values of the physical and computational modes are $\left[\frac{q^1-\lambda_2 q^0}{\lambda_1-\lambda_2}\right]$ and $-\left[\frac{q^1-\lambda_1 q^0}{\lambda_1-\lambda_2}\right]$, respectively. Which mode dominates in the numerical solution will depend on how we determine $q^1$. *If we can choose* $q^1$ *such 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 an accurate method to predict $q^1$ by starting from $q^0$.

In this idealized example, 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 uses leapfrog time differencing needs a way to *damp* the computational mode, 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 leapfrog 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

$$\overline{q^n} = q^n + \alpha \left( \overline{q^{n-1}} - 2q^n + q^{n+1} \right) \tag{4.44}$$

Here the overbar denotes a filtered quantity, and $\alpha$ is a 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 apply the filter, the procedure is to predict $q^{n+1}$ in the usual way, then use (4.44) 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 $n+1$. Note that (4.44) 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. (4.32) will give $\overline{q^n} = -1 + \alpha\,(1+2+1) = -1+4\alpha$. If we choose $\alpha = 0.5$, we get $\overline{q^n} = 1$, and the zig-zag is eliminated.

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

### 4.3.6 The leapfrog scheme for the oscillation equation

To evaluate the 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 (4.37) is real, and we find that $|\lambda_1| = |\lambda_2| = 1$. 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 $\theta_1$ and $\theta_2$, respectively. Then

$$\lambda_1 = e^{i\theta_1} \text{ and } \lambda_2 = e^{i\theta_2}. \tag{4.45}$$

Comparing (4.45) with (4.37), and using Euler's formula, we find by inspection that

$$\begin{gathered}\cos\theta_1 = \sqrt{1-\Omega^2}, \cos\theta_2 = -\sqrt{1-\Omega^2}, \\ \sin\theta_1 = \Omega, \sin\theta_2 = \Omega.\end{gathered} \tag{4.46}$$

From (4.46), 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 $\omega$. Then we can write

$$q_1^{n+1} = e^{i\theta_1} q_1^n \tag{4.47}$$

for the physical mode, and

$$q_2^{n+1} = e^{i(\pi-\theta_1)} q_2^n \tag{4.48}$$

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

$$q\left[(n+1)\Delta t\right] = e^{i\Omega} q\left(n\Delta t\right). \tag{4.49}$$

Panels a and b of Fig. 4.6 respectively show plots of $\lambda_1$ and $\lambda_2$ in the complex $\lambda$-plane. The figures have been drawn for the case of $\theta_1 = \frac{\pi}{8}$. The absolute value of $\lambda$ 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 4.7. The physical mode looks nice. The computational mode is ugly.

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

Here $\lambda_1 = \lambda_2 = i\Omega = i$ (see (4.37)), i.e., both $\lambda$'s are purely imaginary with modulus one (i.e., both are neutral), as shown in Fig. 4.8. This means that both solutions rotate through $\frac{\pi}{2}$ 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 $\frac{\pi}{2}$ radians, which implies an error of 57%.

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

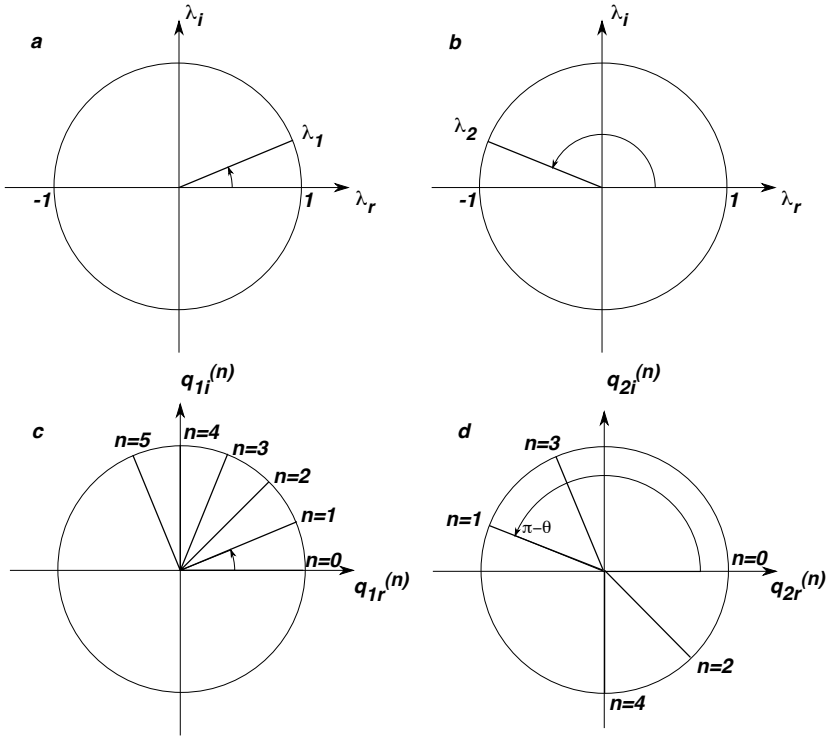



Figure 4.6: Panels a and b show the amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| < 1$. Panel a is for the physical mode, and panel b is for the computational mode. Panels c and d show solutions of the oscillation as obtained with the leapfrog scheme 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 $\theta_1 = \frac{\pi}{8}$.

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

Here again both $\lambda_1$ and $\lambda_2$ are purely imaginary, so again both solutions rotate by $\frac{\pi}{2}$ on each time step, regardless of the value of $\omega$. Eq. (4.38) can be written as

$$\lambda_1 = i\left(\Omega + \sqrt{\Omega^2 - 1}\right) \text{ and } \lambda_2 = i\left(\Omega - \sqrt{\Omega^2 - 1}\right). \tag{4.50}$$

If $\Omega > 1$ then $|\lambda_1| > 1$ and $|\lambda_2| < 1$, and if $\Omega < -1$, then $|\lambda_1| < 1$ and $|\lambda_2| > 1$. In both cases, one of the modes is damped and the other amplifies. Since one of the modes amplifies either way, the scheme is unstable.

Panels a and b of Fig. 4.9 give a graphical representation of $\lambda$ 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. 4.9 shows a plot of the evolving solution, $q^n$, in the complex plane, for the modes corresponding to $\lambda_1$ and $\lambda_2$ for $\Omega > 1$. The phase changes by $\frac{\pi}{2}$ on each step, because $\lambda$

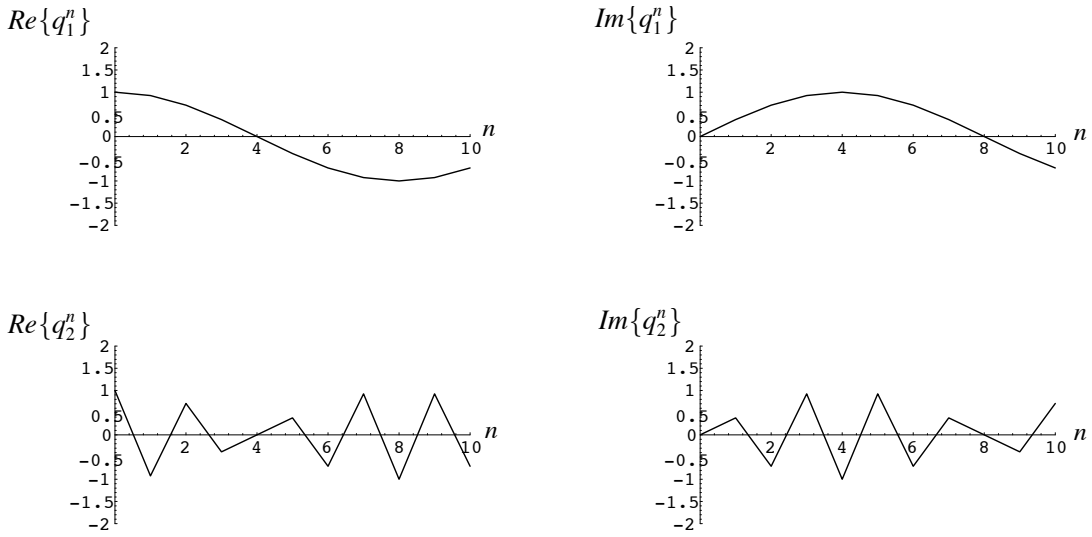



Figure 4.7: Graphs of the real and imaginary parts of the physical and computational modes for the solution of the oscillation equation as obtained with the leapfrog scheme for $|\Omega| < 1$.

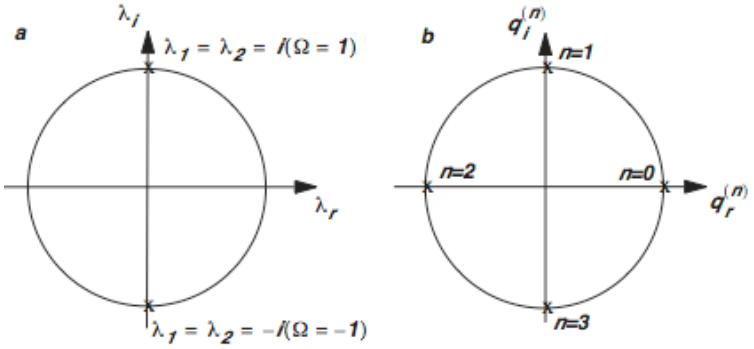



Figure 4.8: Panel a shows the amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| = 1$. Panel b shows the real and imaginary parts of the corresponding solution, for $n = 0,\ 1,\ 2,$ and 3.

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. 4.9 schematically shows $q^n$ as a function of $n$ for the amplifying mode corresponding to $\lambda_1$ 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 $\Delta t$, 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 this reason it has been very widely used. On the other hand, the

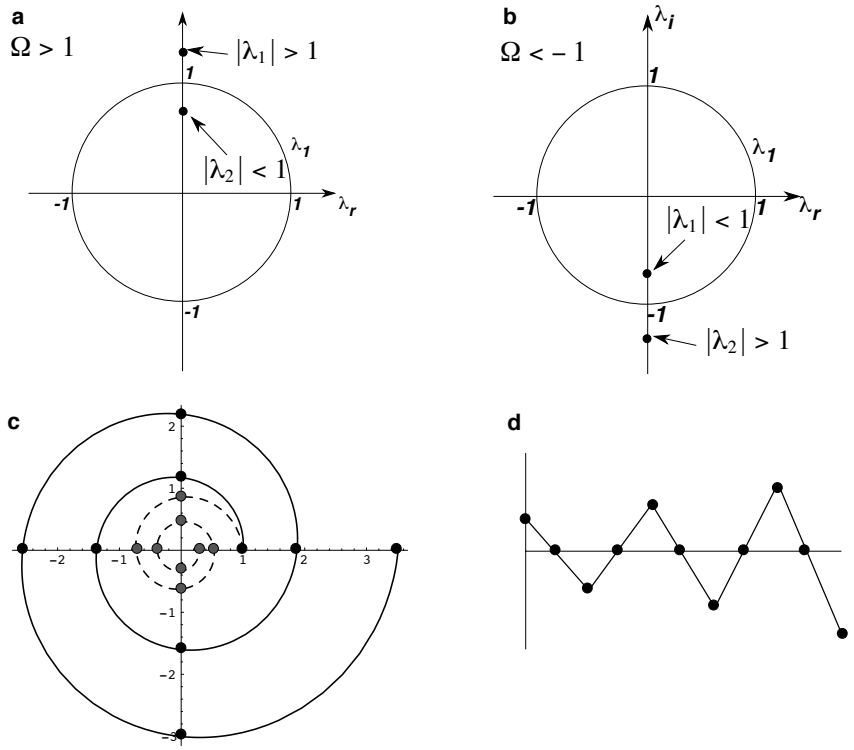



Figure 4.9: Panels a and b show the amplification factors for the oscillation equation with the leapfrog scheme, with $|\Omega| > 1$. Panel c shows the corresponding solution. The solid curve shows the unstable mode, which is actually defined only at the black dots. The dashed curve shows the damped 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.

leapfrog scheme allows a computational mode in time, which is bad. Once again, we have trade-offs.

The trapezoidal implicit scheme is also neutrally stable for the oscillation equation but, compared with an explicit scheme, it is more difficult to use in complicated nonlinear problems.

### 4.3.7 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

$$q^{n+1} - q^n = i\Omega \left( \frac{3}{2} q^n - \frac{1}{2} q^{n-1} \right). \tag{4.51}$$

Like the leapfrog scheme, this is a three-level scheme, so it has a computational mode. The right-hand side of (4.51) represents a linear *extrapolation* (in time) of $q$ from $q^{n-1}$ and $q^n$ to $n+\frac{1}{2}$. It can be interpreted in terms of a scheme centered around time level $n+\frac{1}{2}$. The amplification factor satisfies

$$\lambda^2 - \lambda\left(1+\frac{3}{2}i\Omega\right) + i\frac{1}{2}\Omega = 0. \tag{4.52}$$

The two solutions of (4.52) are

$$\lambda_1 = \frac{1}{2}\left(1+\frac{3}{2}i\Omega + \sqrt{1-\frac{9}{4}\Omega^2+i\Omega}\right), \tag{4.53}$$

and

$$\lambda_2 = \frac{1}{2}\left(1+\frac{3}{2}i\Omega - \sqrt{1-\frac{9}{4}\Omega^2+i\Omega}\right), \tag{4.54}$$

Since $\lambda_1 \to 1$ as $\Omega \to 0$, the first mode is the physical mode and corresponds to the true solution as $\Delta t \to 0$. Note, however, that $\lambda_2 \to 0$ as $\Omega \to 0$. This means that *the "computational" mode is damped*, which is good.

In order to examine $|\lambda|$, we will make some approximations based on the assumption that $\Omega << 1$, because the expressions in (4.54) - (4.66) are complicated and in practice $\Omega$ is usually small. Using the binomial theorem, we can approximate $\lambda_1$ by

$$\lambda_1 \cong 1 + i\Omega - \frac{9}{16}\Omega^2 \cong 1 + i\Omega - \frac{1}{2}\Omega^2, \tag{4.55}$$

so that

$$|\lambda_1| \cong \sqrt{1+\frac{\Omega^4}{4}} \cong 1+\frac{\Omega^4}{8}. \tag{4.56}$$

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

### 4.3.8 A survey of time differencing schemes for the oscillation equation

Baer and Simons (1970) summarized the properties of various explicit and implicit schemes, which are listed in Table 4.1. Properties of these schemes are shown in Fig. 4.10 and Fig. 4.11.

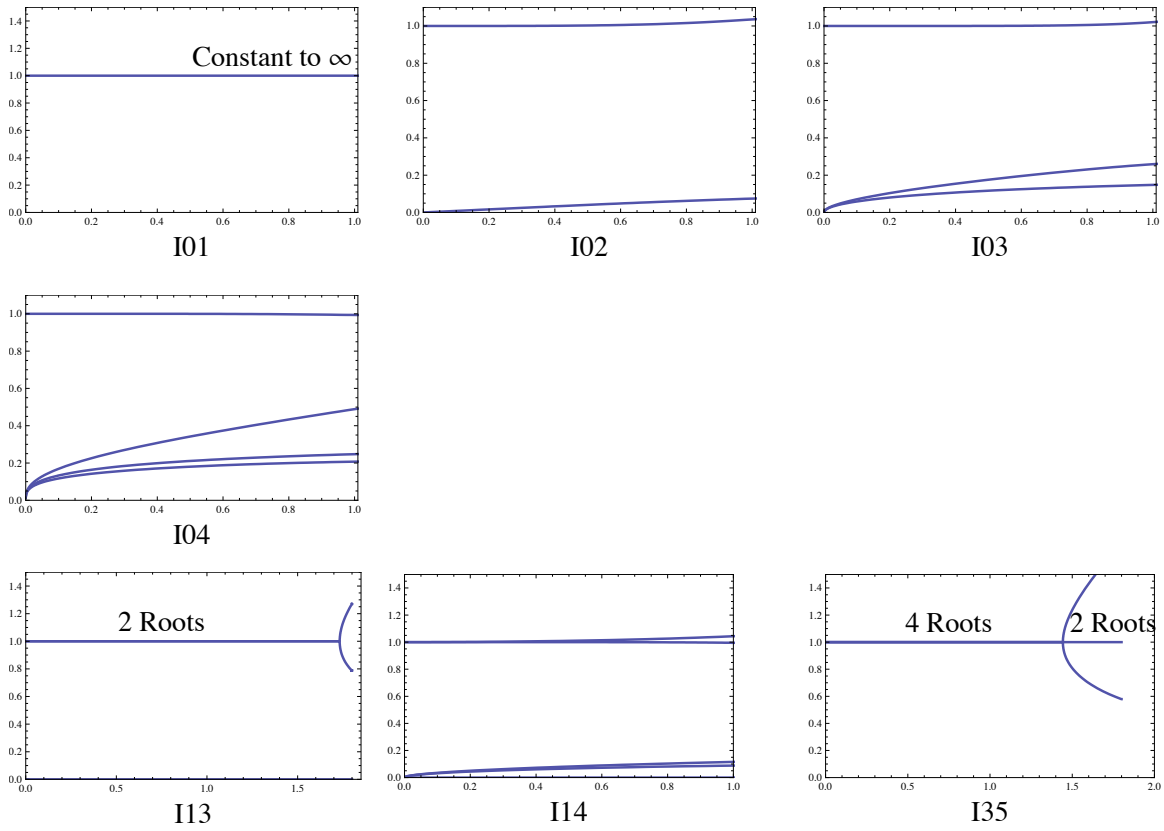



Figure 4.10: Amplification factors of various schemes as applied to the oscillation equation (after Baer and Simons (1970)). The horizontal axis in each panel is $\Omega$. See Table 4.1.

As discussed in books on numerical analysis, 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.

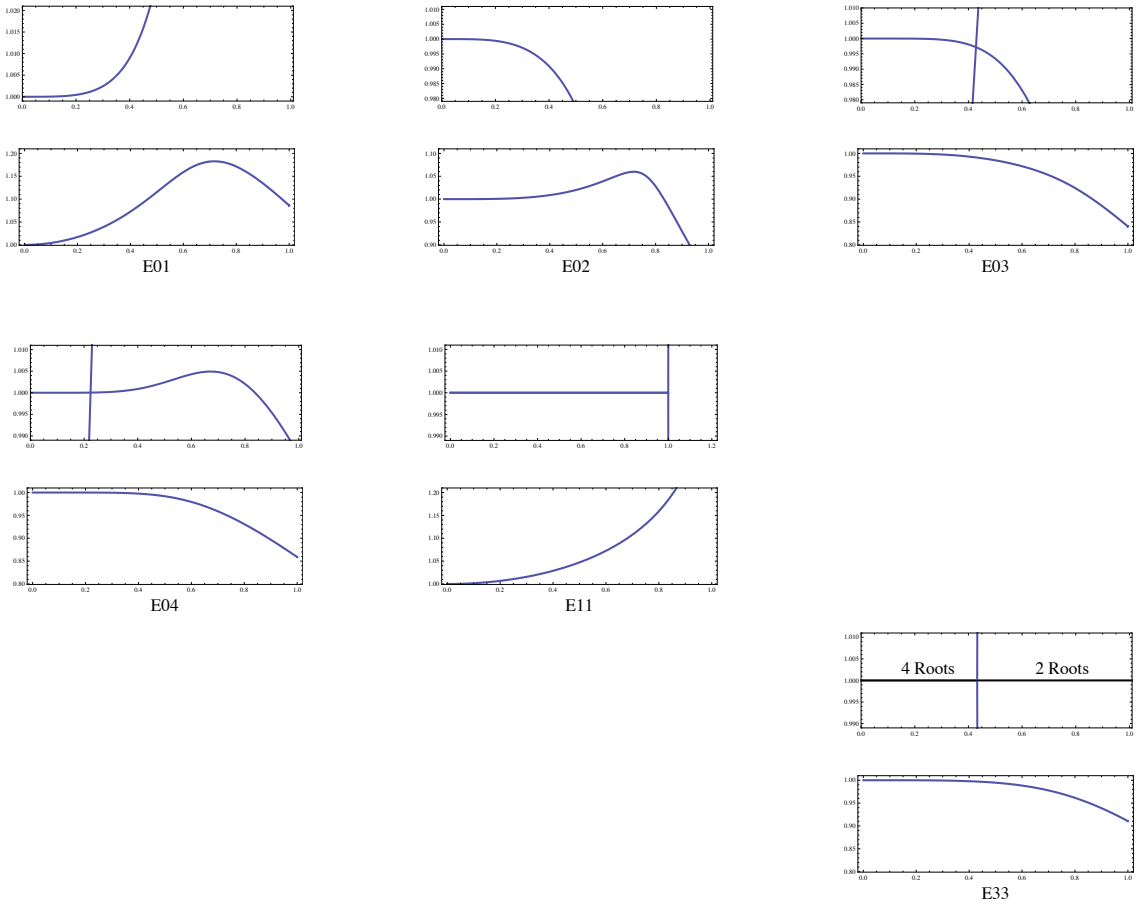



Figure 4.11: Magnitude of the amplification factor and $\theta/\Omega$ for various schemes as applied to the oscillation equation (after Baer and Simons (1970)). The horizontal axis in each panel is $\Omega$. See Table 4.1.

## 4.4 The decay equation

The exact solution of the decay equation is

$$q(t) = q(0)\, e^{-\kappa t}. \tag{4.57}$$

This describes a simple exponential decay with time. For large time, $q \to 0$. A good scheme should therefore give $q^{n+1} \to 0$ as $\kappa \Delta t \to \infty$. The "true" value of $\lambda$ is given by

$$\lambda_T = e^{-\kappa \Delta t} < 1 \tag{4.58}$$

Table 4.1: List of the time differencing schemes 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. $(3.4)$ and Fig. 3.1; and $l$, which controls the number of values of $f$ used, again as defined in $(3.4)$.

| **Scheme identifier (m, l)** | **Name** | $\beta$ | $\alpha_n$ | $\alpha_{n-1}$ | $\alpha_{n-2}$ | $\alpha_{n-3}$ | $\alpha_{n-4}$ | **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 differs from the oscillation equation, for which $\mid \lambda_T \mid = 1$.

For the Euler (forward) scheme, the finite-difference analogue of (4.2) is

$$q^{n+1} - q^n = -Kq^n, \tag{4.59}$$

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

$$q^{n+1} = (1-K)\, q^n \, . \tag{4.60}$$

Note that $\lambda = 1 - K$ is real. For $|1-K| < 1$, which is satisfied for $\kappa$ or $\Delta t$ small enough to give $K \leq 2$, the scheme is stable. This is, therefore, a *conditionally* stable scheme. Note, however, that 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

$$q^{n+1} - q^n = -Kq^{n+1}, \tag{4.61}$$

with solution

$$q^{n+1} = \frac{q^n}{1+K}. \tag{4.62}$$

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 pretty good choice for the decay equation, although of course it has only first-order accuracy.

It is easy to show 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.

The energy method can also be applied to analyze the stability of these schemes for the decay equation.

Finally, the leapfrog scheme for the decay equation is

$$q^{n+1} - q^{n-1} = -2Kq^n, \tag{4.63}$$

and so $\lambda$ satisfies

$$\lambda^2 + 2K\lambda - 1 = 0. \tag{4.64}$$

The two roots are

$$\lambda_1 = -K + \sqrt{K^2+1}, \lambda_2 = -K - \sqrt{K^2+1}. \tag{4.65}$$

Since $0 \leq \lambda_1 \leq 1$, and $\lambda_1 \to 1$ as $K \to 0$, we see that $\lambda_1$ 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 $\lambda_2 \leq -1$ ($\lambda_2 \to -1$ as $\Delta t \to 0$) so the computational mode 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. 4.12. From (4.64) 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. 4.12. 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 "more accurate" leapfrog scheme gives a bad solution. This illustrates again that accuracy is a vague concept.

$$q^0 = 0$$

$$q^1 > 0$$

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

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

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

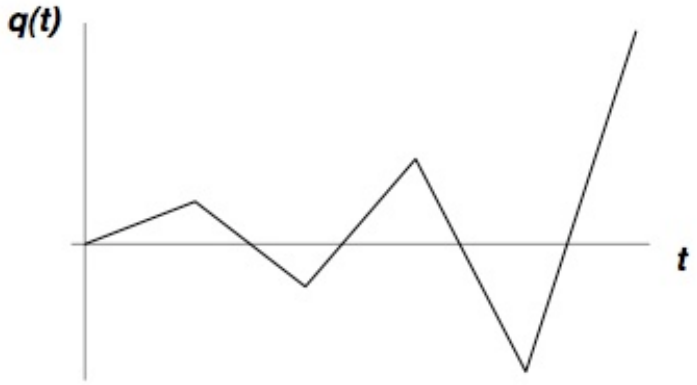



Figure 4.12: An illustration of how the leapfrog scheme leads to instability with the decay equation. The solution plotted represents the computational mode only and would be superimposed on the physical mode.

## 4.5 Damped oscillations

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

$$\frac{dq}{dt} = (i\omega - \kappa)\, q? \tag{4.66}$$

The exact solution of (4.66) 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 (4.66) as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n-1}, \tag{4.67}$$

(decay term forward differenced), or as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1}, \tag{4.68}$$

(decay term backward differenced). The oscillation terms on the right-hand sides of (4.67) and (4.68) are in "centered" form, whereas the damping terms have an uncentered form. These schemes are conditionally stable.

## 4.6 Nonlinear damping

In real applications, it is quite typical that $\kappa$ 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 (4.2), given by

$$\frac{dq}{dt} = \left(-\kappa q^P\right) q + S, \tag{4.69}$$

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 $S = 0$, (4.69) reduces to (4.2).

An example of a real application that gives rise to an equation of the form (4.69) is boundary-layer parameterization. The soil temperature, $T_g$, satisfies an equation roughly of the form

$$C\frac{dT_g}{dt} = -\rho_a c_p c_T V \left(T_g - T_a\right) + S_g, \tag{4.70}$$

where $C$ is the heat capacity of the soil layer, $\rho_a$ 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:

$$\rho_a D c_p \frac{dT_a}{dt} = \rho_a c_p c_T V (T_g - T_a) + S_a. \tag{4.71}$$

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 (4.70) and (4.71).

Comparison of (4.70) with (4.71) shows that

$$\frac{d(T_g - T_a)}{dt} = -\rho_a c_p c_T V (T_g - T_a) \left( \frac{1}{C} + \frac{1}{\rho_a D c_p} \right) + \left( \frac{S_g}{C} - \frac{S_a}{\rho_a D c_p} \right). \tag{4.72}$$

The analogy between (4.69) and (4.72) 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 (4.69), i.e.,

$$\frac{q^{n+1} - q^n}{\Delta t} = -\kappa \left( q^{n+1} \right)^{P+1} + S. \tag{4.73}$$

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, (4.73) may not be considered a viable choice, except where $P$ is a small integer, in which case (4.73) 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 $\bar{q}$ denote an equilibrium solution of (4.69), so that

$$\kappa \bar{q}^{P+1} = S. \tag{4.74}$$

We are assuming for simplicity that $S$ is independent of $q$ and time. Let $q'$ denote a departure from the equilibrium, so that $q = \bar{q} + q'$. Then (4.70) can be linearized as follows:

$$\frac{d}{dt}\left(\bar{q}+q'\right) = -\kappa\bar{q}^{P+1} - \kappa\left(P+1\right)\bar{q}^{P}q' + S, \tag{4.75}$$

which reduces to

$$\frac{dq'}{dt} = -\kappa\left(P+1\right)\left(\bar{q}\right)^{P}q'. \tag{4.76}$$

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

As an example, the forward time-differencing scheme, applied to (4.76), gives

$$q^{n+1} - q^{n} = -\alpha\left(P+1\right)q^{n}, \tag{4.77}$$

where we use the shorthand notation

$$\alpha \equiv \kappa(\bar{q})^{P}\Delta t, \tag{4.78}$$

and we have dropped the "prime" notation for simplicity. We can rearrange (4.77) to

$$q^{n+1} = \left[1 - \alpha\left(P+1\right)\right]q^{n}, \tag{4.79}$$

from which we see that

$$\lambda = 1 - \alpha\left(P+1\right). \tag{4.80}$$

From (4.80), 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 4.2 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 4.2: 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 | $\frac{\hat{q}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$<br>$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left(\hat{q}\right)^P q^{n+1}$ | $\frac{1-\alpha(P-1)+(\alpha P)^2}{(1+\alpha)^2}$ | $\alpha(P-1)<1$ |
| Average coefficient, implicit q | $\frac{\hat{q}-q^n}{\Delta t} = \kappa\left(q^n\right)^P \hat{q}$<br>$\frac{q^{n+1}-q^n}{\Delta t} = -\kappa\left[\frac{\kappa\left(q^n\right)^P+\left(\hat{q}\right)^P}{2}\right]q^{n+1}$ | $\frac{1-\alpha(P-1)-\frac{\alpha^2 P}{2}+\frac{(\alpha P)^2}{2}}{(1+\alpha)^2}$ | $\alpha(P-2)<2$ |
| 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 | $\frac{\hat{q}-q^n}{\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$<br>$q^{n+1} = (1-A)\hat{q}+Aq^n$ | $\frac{(1-A)(1-\alpha P)}{1+\alpha}+A$ | $\alpha\left[P(1-A)-1-A\right]<2$ |
| Double time step, explicit coefficient, implicit q with time average filter | $\frac{\hat{q}-q^n}{2\Delta t} = -\kappa\left(q^n\right)^P \hat{q}$<br>$q^{n+1} = \frac{\hat{q}+q^n}{2}$ | $\frac{1-\alpha(P-1)}{1+2\alpha}$ | $\alpha(P-3)<2$ |
| Linearization of backward implicit scheme | $\frac{q^{n+1}-q^n}{\Delta t} =$<br>$-\kappa\left(q^n\right)^P\left[(P+1)q^{n+1}-Pq^n\right]$ | $\frac{1+\alpha P}{1+\alpha(P+1)}$ | Unconditionally stable |

backward implicit and centered implicit schemes, the amplification factor has been obtained by linearization, but the "linear stability criteria" 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.

## 4.7 Summary

It is possible to construct time-differencing schemes of arbitrary 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, but on the other hand the small time step makes it possible to represent the temporal evolution in more detail.

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. This is an example of how "accuracy" in the Taylor-series sense can be misleading.

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 intrinsically damps the computational mode more than the physical mode, e.g., an Adams-Bashforth scheme.

Finally, we update our list of the 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.

## 4.8 Problems

1. Find the exact solution of

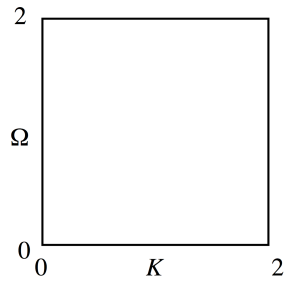
$$\frac{dq}{dt} = i\omega q - \kappa q. \tag{4.81}$$

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

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

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1}. \tag{4.82}$$

   (c) Plot the neutral stability boundary (where $|\lambda| = 1$) as a curve in the $(K, \Omega)$ plane, for $K$ and $\Omega$ 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 $(K, \Omega)$ plot correspond to instability.

   (d) Code the equation given in part (b) above. Use a forward time step for the first step only. Use $q(t=0) = 100$, and $\Delta t = 1$. Plot the solution out to $t = 100$ for the following cases:

$$\frac{\omega}{2\pi} = 0.1,\ \kappa = 0$$
$$\frac{\omega}{2\pi} = 0,\ \kappa = 0.1 \tag{4.83}$$
$$\frac{\omega}{2\pi} = 0.1,\ \kappa = 0.1$$

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

(f) Derive an equation satisfied by the amplification factor for the second-order Adams-Bashforth scheme applied to Eq. (4.82). (The result is quite complicated.) Contour plot $|\lambda|$ as a function of both $\omega$ and $\kappa$. Find an approximate solution valid for sufficiently small $\Delta t$.

2. The trapezoidal-implicit scheme for the oscillation equation is given by

$$q^{n+1} - q^n = \frac{i\Omega}{2}\left(q^n + q^{n+1}\right). \tag{4.84}$$

Analyze the stability of this scheme using von Neumann's method.

3. Determine the orders of the *discretization errors* of the Matsuno and Heun schemes.

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

5. 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.

6. Plot $\theta$ as a function of $\Omega$ 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 $-\pi \leq \Omega \leq \pi$.

7. The equations of Lorenz's famous butterfly model are

$$\begin{aligned}\dot{X} &= -\sigma(X-Y),\\ \dot{Y} &= -XZ+rX-Y,\\ \dot{Z} &= XY-bZ.\end{aligned} \tag{4.85}$$

The model has three equilibria, one of which is $X = Y = Z = 0$.

(a) Linearizing the system about the equilibrium solution mentioned above, analyze the stability for the time continuous case. Assume that $\sigma = 10$, $b = \frac{8}{3}$, and $r = 24.74$.

(b) For the case of the forward time-differencing scheme, use von Neumann's method to find the amplification factor. Compare with the effective amplification factor for the continuous case. Suggest a good choice for the value of the time step.

(c) Program the model twice, using forward time-differencing in one case and the fourth-order Runge-Kutta method in the other. As in part (a), use the parameter settings $\sigma = 10$, $b = \frac{8}{3}$, and $r = 24.74$. Start the model from the initial condition $(X,Y,Z) = (0.1,0,0)$, which is close to the equilibrium mentioned above. Experiment to determine the longest permissible time step with each scheme. Run the model long enough to see the butterfly. Plot the results for both schemes, and compare them.

8. Prove that the phase error $\varepsilon_\varphi$, defined by $\theta = \Omega + \varepsilon_\varphi$, satisfies $\varepsilon_\varphi = \tan^{-1}\left(\frac{\varepsilon_I}{\varepsilon_R+1}\right)$.

9. Find the phase and amplitude errors of the upstream scheme for advection, and plot them as functions of $\mu$ and $k\Delta x$.

# Chapter 5

## Riding along with the air

The purpose of this chapter is to review the physical nature of advection, because the design or choice of a numerical method should always be motivated as far as possible by our understanding of the physical process at hand.

In Lagrangian form, the advection equation, in any number of dimensions, is simply

$$\frac{DA}{Dt} = 0. \tag{5.1}$$

*This 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 (5.1), 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 important 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.

Consider a simple function of $A$, such as $A^2$. Since $A$ is unchanged during advection,

for each particle, $A^2$ will also be unchanged. Obviously, any other 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; a negative mixing ratio would have no physical meaning. Some other variables, such as the zonal component of the wind vector, can be either positive or negative; for the zonal wind, our convention is that positive values denote westerlies and negative values denote easterlies.

Suppose that $A$ is conserved under advection, following each particle. It follows that 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).

Typically the variable $A$ represents an "intensive" property, which is defined per unit mass. An example is the mixing ratio of some trace species, such as water vapor. A second example is temperature, which is proportional to the internal energy per unit mass. In the most troublesome case, $A$ is a component of the wind field itself.

Of course, in general these various quantities are not really conserved following particles, simply 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 radiative heating. To describe more general processes that include not only advection but also sources and sinks, we can replace (5.1) by

$$\frac{DA}{Dt} = S, \tag{5.2}$$

where $S$ is the source of $A$ per unit time. (A negative value of $S$ represents a sink.) We still refer to (5.2) as a "conservation" equation; it says that $A$ is conserved *except* to the extent that sources or sinks come into play.

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

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{V}), \tag{5.3}$$

where $\rho$ is the density (mass per unit volume) and $\mathbf{V}$ is the velocity vector. We can also write (5.3) as

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\mathbf{V}. \tag{5.4}$$

*Eq.* (5.4) *tells us that, in general,* $\rho$ *is not constant following a particle*. If we have an incompressible fluid, i.e., one for which $\rho$ is an immutable property of the fluid so that $D\rho/Dt = 0$, then (5.4) implies that $\nabla\cdot\mathbf{V} = 0$. In other words, incompressibility implies nondivergence. 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).

We can expand (5.2) into the Eulerian advection form of the conservation equation for $A$:

$$\frac{\partial A}{\partial t} = -(\mathbf{V}\cdot\nabla)A + S. \tag{5.5}$$

Multiply (5.3) by $A$, and (5.5) by $\rho$, and add the results to obtain

$$\frac{\partial}{\partial t}(\rho A) = -\nabla\cdot(\rho\mathbf{V}A) + \rho S. \tag{5.6}$$

This is called the *flux form* of the conservation equation for $A$. Notice that if we put $A \equiv 1$ and $S \equiv 0$ (because there is no "source of 1"!) then (5.6) reduces to (5.3). This is an important point that can and should be used in the design of advection schemes, i.e., 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 for the continuity equation.

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. An advection scheme that has this property is called *"compatible."*

Suppose that we integrate (5.3) 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 we find, using Gauss's Theorem, that

$$\frac{d}{dt}\int_R \rho\, dR = 0. \tag{5.7}$$

This simply states that mass is conserved within the domain. Similarly, for the case of closed or periodic boundaries we can integrate (5.6) over $R$ to obtain

$$\frac{d}{dt}\int_R \rho A dR = \int_R \rho S dR. \tag{5.8}$$

This says that the mass-weighted average value of $A$ is conserved within the domain, except for the effects of sources and sinks. We can describe (5.7) and (5.8) as integral forms of the conservation equations for mass and $A$, respectively.

It may seem that the ideal way to simulate advection in a model is to define a collection of particles, to associate various properties of interest with each particle, and to 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., of its longitude, latitude, and height. These would change in response to the predicted velocity field. Such a Lagrangian approach has been demonstrated, and will be discussed later in this chapter.

At the present time, however, virtually all models in atmospheric science are based on Eulerian methods. For the case of vertical advection, the Eulerian vertical coordinate is sometimes permitted to "move" as the circulation evolves (e.g., Phillips (1957); Hsu and Arakawa (1990)). This will be discussed in a later chapter.

# Chapter 6

# The upstream scheme

## 6.1 Introduction

Consider the one-dimensional advection equation, given by

$$\left(\frac{\partial A}{\partial t}\right)_x + c\left(\frac{\partial A}{\partial x}\right)_t = 0, \tag{6.1}$$

where $A = A(x,t)$. The physical meaning of (6.1) is that $A$ remains constant at the position of a particle that moves in the $x$-direction with speed $c$. We can interpret $A$ as a "conserved" property of the particle. We will assume for now that $c$ is a constant. Eq. (6.1) is a first-order linear partial differential equation with a constant coefficient, namely $c$. It looks harmless, but it causes no end of trouble.

Suppose that

$$A(x,0) = F(x) \text{ for } -\infty < x < \infty. \tag{6.2}$$

This is an "initial condition." Our goal is to determine $A(x,t)$. This is a simple example of an initial-value problem. We first work out the analytic solution of (6.1), for later comparison with our numerical solution. Define

$$\xi \equiv x - ct, \text{ so that } \left(\frac{\partial x}{\partial t}\right)_\xi = c. \tag{6.3}$$

The value of $\xi$ does not change as the particle moves. If $\xi = 0$ at the location of the particle, then the value of $\xi$ measures distance from the particle. Using the chain rule, we write

$$\begin{aligned}\left(\frac{\partial A}{\partial x}\right)_\xi &= \left(\frac{\partial A}{\partial x}\right)_t + \left(\frac{\partial A}{\partial t}\right)_x \left(\frac{\partial t}{\partial x}\right)_\xi \\ &= \left(\frac{\partial A}{\partial x}\right)_t + \left(\frac{\partial A}{\partial t}\right)_x \frac{1}{c} \\ &= 0 \quad .\end{aligned} \tag{6.4}$$

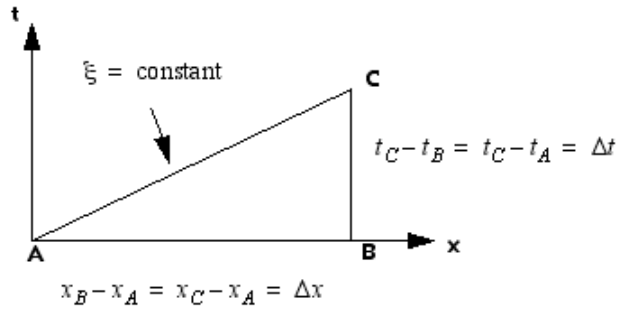



Figure 6.1: Figure used in the derivation of the first line of (6.4). *** Redraw this figure.

The first line of (6.4) can be understood by reference to Fig. 6.1. The third line comes by use of (6.1) or (6.3). Similarly,

$$\begin{aligned}\left(\frac{\partial A}{\partial t}\right)_\xi &= \left(\frac{\partial A}{\partial t}\right)_x + \left(\frac{\partial A}{\partial x}\right)_t \left(\frac{\partial x}{\partial t}\right)_\xi \\ &= \left(\frac{\partial A}{\partial t}\right)_x + \left(\frac{\partial A}{\partial x}\right)_t c \\ &= 0 \quad .\end{aligned} \tag{6.5}$$

Again, the last equality follows from (6.1). We can interpret $(\partial A/\partial t)_\xi$ as the time rate of change of $A$ that we would see if we were riding on the moving particle of air, so the advection equation describes what happens as particles of air move around without changing their values of $A$. From (6.4) and (6.5), we conclude that

$$A = f(\xi) \tag{6.6}$$

is the general solution to (6.1). This means that $A$ depends only on $\xi$, in the sense that if you tell me the value of $\xi$, that's all the information I need to tell you the value of $A$. (Note that "$A$ depends only on $\xi$" does *not* mean that $A$ is independent of $x$ for fixed $t$, or of $t$ for fixed $x$.)

The initial condition is

$$\xi \equiv x \text{ and } A(x) = f(x) \text{ at } t = 0, \tag{6.7}$$

i.e., the shape of $f(\xi)$ is determined by the initial condition. Eq. (6.6) means that $A$ is constant along the line $\xi =$ constant. In order to satisfy the initial condition, we chose $f \equiv F$ [see Eq. (6.2)]. Referring to (6.6) , we see that $A(\xi) = F(\xi) \equiv F(x-ct)$ is the solution to (6.1) that satisfies the initial condition (6.7). An initial value simply "moves along" the lines of constant $\xi$, 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. Partial differential equations whose solutions are constant along characteristics are called *hyperbolic* equations. The advection equation is hyperbolic. Further discussion is given later.

Keeping in mind the exact solution, we now investigate the solution of one possible numerical scheme for (6.1). We construct a grid, as in Fig. 6.2. One of the infinitely many possible finite-difference approximations to (6.1) is

$$\boxed{\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) = 0}. \tag{6.8}$$

Here we have used the forward difference quotient in time and the backward difference quotient in space. If we know $A_j^n$ at some time level $n$ for all $j$, then we can solve (6.8) for $A_j^{n+1}$ at the next time level, $n+1$. For $c > 0$, (6.8) 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 one side and goes to the other, 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 it in a box.

Because

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} \to \frac{\partial A}{\partial t} \text{ as } \Delta t \to 0, \tag{6.9}$$

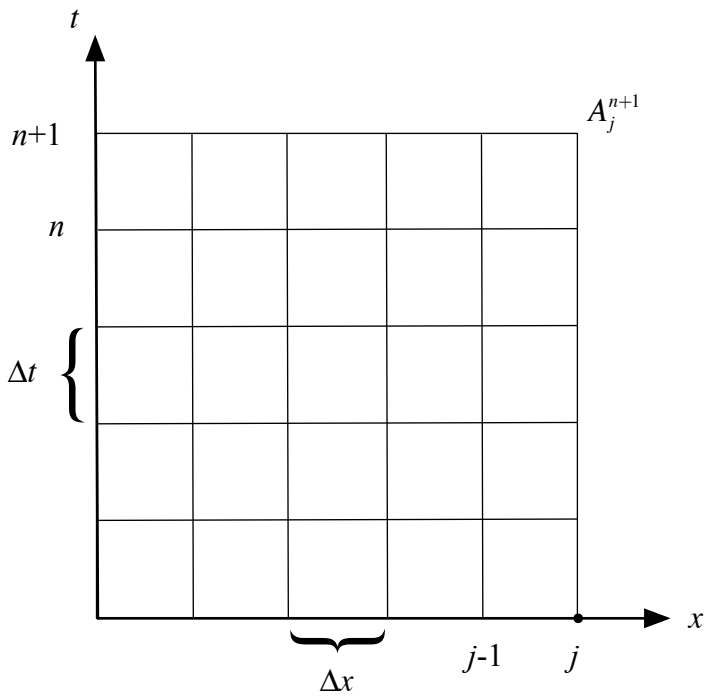



Figure 6.2: A grid for the solution of the one-dimensional advection equation.

and

$$\frac{A_j^n - A_{j-1}^n}{\Delta x} \to \frac{\partial A}{\partial x} \text{ as } \Delta x \to 0, \tag{6.10}$$

we can say that (6.8) does approach (6.1) as $\Delta t$ and $\Delta x$ both approach zero. In view of (6.9) and (6.10), it may seem obvious that the *solution* of (6.8) approaches the *solution* of (6.1) as $\Delta x \to 0$ and $\Delta t \to 0$. Unfortunately, that is not necessarily true.

## 6.2 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. 6.2. 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 (6.8), we get

$$\left\{\frac{A\left[j\Delta x,(n+1)\,\Delta t\right]-A\left(j\Delta x,n\Delta t\right)}{\Delta t}\right\}+c\left\{\frac{A\left(j\Delta x,n\Delta t\right)-A\left[(j-1)\Delta x,n\Delta t\right]}{\Delta x}\right\}=\varepsilon \quad (6.11)$$

The discretization error, $\varepsilon$, is a measure of how accurately the solution $A(x,t)$ of the original differential equation (6.1), satisfies the finite-difference equation, (6.8). It is far from a perfect measure of accuracy, however, as will become evident.

If we obtain the terms of (6.11) 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* (6.1), we find that

$$\varepsilon=\left(\frac{1}{2!}\Delta t\frac{\partial^2 A}{\partial t^2}+\cdots\right)+c\left(-\frac{1}{2!}\Delta x\frac{\partial^2 A}{\partial x^2}+\cdots\right). \quad (6.12)$$

We say this is a "first-order scheme" because the first powers of $\Delta t$ and $\Delta x$ appear in (6.12). The notations $O(\Delta t,\Delta x)$ or $O(\Delta t)+O(\Delta x)$ 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* $\Delta t$ *and* $\Delta x$ *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.

## 6.3 Convergence

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 $\Delta t$ 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. 6.3 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 $\Delta x$ and $\Delta t$ 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, (6.8).

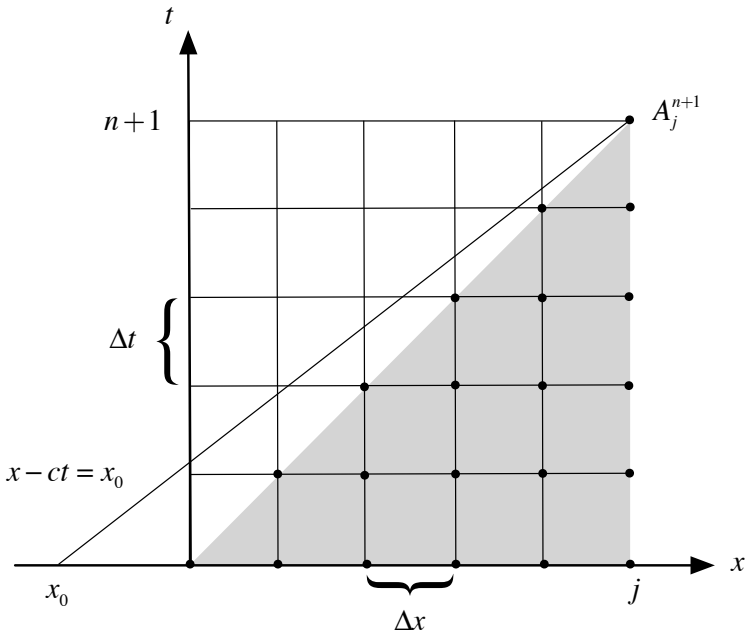



Figure 6.3: The shaded area represents the "domain of dependence" of the solution of the upstream scheme at the point $(j\Delta x, n\Delta t)$.

We could increase the accuracy of the scheme by cutting $\Delta x$ and $\Delta t$ in half, or by a factor of 100, but the domain of dependence would not change as long as the ratio $\frac{c\Delta t}{\Delta x}$ remains the same. This is a clue that $c\Delta t/\Delta x$ is an important quantity. We will give it a name:

$$\mu \equiv \frac{c\Delta t}{\Delta x}. \tag{6.13}$$

Suppose that the line through the point $(j\Delta x, n\Delta t)$, i.e., $x - ct = x_0$, where $x_0$ is a constant, does not lie in the domain of dependence. This is the situation shown in Fig. 6.3. In general, there is no hope of obtaining smaller discretization error, no matter how small $\Delta x$ and $\Delta t$ 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. This illustrates that if the value of $c$ 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. The truncation error goes to zero, but the discretization error does not. Bummer.

The discussion above shows that

$$0 \leq \mu \leq 1 \tag{6.14}$$

is a *necessary* condition for convergence of the upstream scheme.

Notice that if $c$ 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, for $c < 0$ we can use

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^n - A_j^n}{\Delta x}\right) = 0, \tag{6.15}$$

in place of (6.8). For $c < 0$, Eq. (6.15) is the appropriate form of the upstream scheme.

A computer program can have an "if-test" that checks the sign of $c$, and uses (6.8) if $c \geq 0$, and (6.15) if $c < 0$. If-tests can cause slow execution on certain types of computers, and besides, if-tests are ugly and reduce the readability of a code. If we define

$$c_+ \equiv \frac{c + |c|}{2} \geq 0, \text{ and } c_- \equiv \frac{c - |c|}{2} \leq 0, \tag{6.16}$$

then a "generalized" upstream scheme can be written as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c_+\left(\frac{A_j^n - A_{j-1}^n}{\Delta x}\right) + c_-\left(\frac{A_{j+1}^n - A_j^n}{\Delta x}\right) = 0. \tag{6.17}$$

This form avoids the use of *if*-tests and is also convenient for use in pencil-and-paper analysis, as discussed later.

In summary: *Truncation error* measures the accuracy of an approximation to a differential operator or operators. It is a measure of the accuracy with which the terms of a differential equation have been approximated. *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.

## 6.4 Interpolation and extrapolation

Referring back to (6.8), we can rewrite the upstream scheme as

$$A_j^{n+1} = A_j^n(1-\mu) + A_{j-1}^n\mu. \tag{6.18}$$

This scheme has the form of either an *interpolation* or an *extrapolation*, depending on the value of $\mu$. To see this, refer to Figure 6.4. Along the line plotted in the figure

$$\begin{aligned} A &= A_{j-1}^n - \left(x - x_{j-1}\right)\left(\frac{A_j^n - A_{j-1}^n}{x_j - x_{j-1}}\right) \\ &= A_j^n\left[1 - \left(\frac{x - x_{j-1}}{x_j - x_{j-1}}\right)\right]_j + A_{j-1}^n\left(\frac{x - x_{j-1}}{x_j - x_{j-1}}\right), \end{aligned} \tag{6.19}$$

which has the same form as our scheme if we identify

$$A \equiv A_j^{n+1} \text{ and } \mu \equiv \frac{x - x_{j-1}}{x_j - x_{j-1}}. \tag{6.20}$$

For $0 \leq \mu \leq 1$ we have *interpolation*. For $\mu < 0$ or $\mu > 1$ we have *extrapolation*.

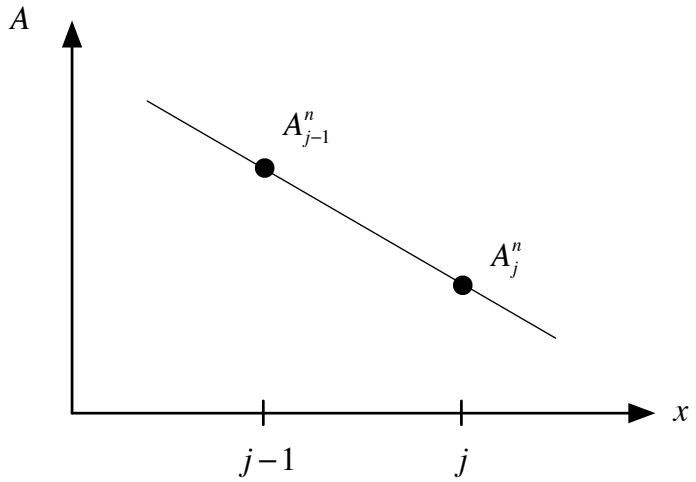



Figure 6.4: Diagram illustrating the concepts of interpolation and extrapolation. See text for details.

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 we take. After a number of time steps, however, the repeated interpolation can produce an unrealistic smoothing of the solution.

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 is not necessarily a problem if we are only taking one (reasonably small) time step, but after a sufficient number of time steps the solution will become useless.

Both interpolation and extrapolation are used extensively in atmospheric modeling. Much more discussion of both is given later.

## 6.5 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.

### 6.5.1 The direct method

As mentioned earlier, the direct method establishes stability by demonstrating that the largest absolute value of $A_j^{n+1}$ on the grid does not increase with time. 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 $0 \le \mu \le 1$ (the necessary condition for convergence according to (6.14)), we can write

$$\left|A_j^{n+1}\right| \le \left|A_j^n\right|(1-\mu) + \left|A_{j-1}^n\right|\mu. \tag{6.21}$$

Therefore,

$$\begin{aligned} \max_{(j)} \left|A_j^{n+1}\right| \le & \max_{(j)} \left|A_j^n\right|(1-\mu) + \max_{(j)} \left|A_{j-1}^n\right|\mu \\ = & \max_{(j)} \left|A_j^n\right| \quad , \end{aligned} \tag{6.22}$$

where $\max_{(j)}$ denotes the largest value at any point on the grid. The second line of (6.22) follows because $\max_{(j)} \left|A_j^n\right| = \max_{(j)} \left|A_{j-1}^n\right|$. Eq. (6.22) demonstrates that the scheme is stable provided that our assumption

$$0 \leq \mu \leq 1 \tag{6.23}$$

is satisfied. This means that the solution remains bounded for all time provided that $0 \leq \mu \leq 1$, and so it 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 (6.22) would hold.

Eq. (6.22), the sufficient condition for stability, is actually obvious from (6.18), because when $0 \leq \mu \leq 1$, $A_j^{n+1}$ is obtained by linear interpolation *in space*, from the available $A_j^n$ to the point $x = j\Delta x - c\Delta t$. This is reasonable, because the nature of advection is such that the time rate of change at a point is closely related to the spatial variations upstream of that point.

### 6.5.2 The energy method

The direct method cannot be used to check the stability of more complicated schemes. The energy method is more widely applicable, even for some nonlinear equations, and it is quite important in practice. We illustrate it here by application to the upstream scheme. With the energy method we ask: "Is $\sum_j \left(A_j^n\right)^2$ 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, $\max_{(j)} \left|A_j^{n+1}\right|$, with the energy method we check $\sum_j \left(A_j^n\right)^2$. The two approaches are somewhat similar.

Returning to (6.18), squaring both sides, and summing over the domain, we obtain

$$\begin{aligned}\sum_j \left(A_j^{n+1}\right)^2 =& \sum_j \left[\left(A_j^n\right)^2(1-\mu)^2 + 2\mu\,(1-\mu)\,A_j^n A_{j-1}^n + \mu^2\left(A_{j-1}^n\right)^2\right] \\ =&(1-\mu)^2\sum_j \left(A_j^n\right)^2 + 2\mu\,(1-\mu)\sum_j A_j^n A_{j-1}^n + \mu^2\sum_j \left(A_{j-1}^n\right)^2.\end{aligned} \tag{6.24}$$

If $A$ is periodic in $x$, then

$$\sum_j \left(A_{j-1}^n\right)^2 = \sum_j \left(A_j^n\right)^2. \tag{6.25}$$

By starting from $\sum_j \left(A_j^n - A_{j-1}^n\right)^2 \geq 0$, and using (6.25), we can show that

$$\sum_j A_j^n A_{j-1}^n \leq \sum_j \left(A_j^n\right)^2. \tag{6.26}$$

Another way to derive (6.26) is to use (6.25) and Schwartz's inequality (e.g., Arfken (1985), p. 527), i.e.,

$$(\sum_j a_j b_j)^2 \leq (\sum_j {a_j}^2)(\sum_j {b_j}^2), \tag{6.27}$$

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 (6.25) and (6.26) in (6.24) gives

$$\sum_j \left(A_j^{n+1}\right)^2 \leq \sum_j \left(A_j^n\right)^2, \tag{6.28}$$

provided that $\mu\left(1-\mu\right) \geq 0$, which follows from $0 \leq \mu \leq 1$. We conclude that

$$\sum_j \left(A_j^{n+1}\right)^2 \leq \sum_j \left(A_j^n\right)^2, \text{ provided that } 0 \leq \mu \leq 1. \tag{6.29}$$

The conclusion is the same as that obtained using the direct method, i.e., $0 \leq \mu \leq 1$ is a sufficient condition for stability.

### 6.5.3 von Neumann's method

A very powerful tool for testing the stability of linear partial difference equations with constant coefficients is *von Neumann's method*. It 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, (6.1). We assume for simplicity that the domain is infinite. First, we look for a solution with the wave form

$$A(x,t) = \mathrm{Re}\left[\widehat{A}(t)\, e^{ikx}\right], \tag{6.30}$$

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 $c$ is constant in space and time. For now, we consider a single wave number, for simplicity, but we can (and soon will) generalize (6.30) by replacing the right-hand side by a sum over a range of wave numbers. Substituting (6.30) into (6.1), we find that

$$\frac{d\widehat{A}}{dt} + ikc\widehat{A} = 0. \tag{6.31}$$

By this substitution, we have converted the partial differential equation (6.1) into an ordinary differential equation, (6.31), whose solution is

$$\widehat{A}(t) = \widehat{A}(0)\, e^{-ikct}, \tag{6.32}$$

where $\widehat{A}(0)$ is the initial value of $\widehat{A}$. Substituting (6.32) back into (6.30), we find that the full solution to (6.31) is

$$A(x,t) = \mathrm{Re}\left[\widehat{A}(0)\, e^{ik(x-ct)}\right], \tag{6.33}$$

provided that $c$ = constant. As can be seen from (6.33), the sign convention used here implies that for $c > 0$ the signal will move towards larger $x$. Note that the exponent in (6.33) is a constant times $\xi = x - ct$, which is the line (the characteristic) along which the solution of (6.1)is expected to be constant.

For a finite-difference equation, the assumed form of the solution, given by Eq. (6.30), is replaced by

$$A_j^n = \mathrm{Re}\left[\widehat{A}^n e^{ikj\Delta x}\right]. \tag{6.34}$$

Here $\left|\widehat{A}^n\right|$ is the amplitude of the wave at time-level $n$. Recall that the wavelength is $2\pi$ 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, $\lambda$, which was defined in Eq. (4.3). We can write

$$\widehat{A}^{n+1} \equiv \lambda \widehat{A}^n \,. \tag{6.35}$$

The amplification factor can be a complex number, but have a special interest in its magnitude, which reveals the stability of a numerical scheme. Note that

$$\left|\widehat{A}^{n+1}\right| = |\lambda| \left|\widehat{A}^n\right| \,. \tag{6.36}$$

In general, $\lambda$ depends on $k$, so we could write $\lambda_k$ or $\lambda\,(k)$, but here we suppress that urge for the sake of keeping the notation simple. The value of $\lambda$ also depends on the size of the time step. As shown below, we can work out the form of $\lambda$ for a particular finite-difference scheme.

Before doing that, we use Eq. (6.35) to identify the effective value of $\lambda$ for the exact solution to the differential equation. From (6.32), we find that

$$\text{for the exact advection equation } \widehat{A}(t+\Delta t) \equiv e^{ikc\Delta t}\widehat{A}\,(t)\,, \tag{6.37}$$

from which it follows "by inspection" that

$$\text{for the exact advection equation } \lambda = e^{ikc\Delta t}. \tag{6.38}$$

Eq. (6.38) implies that

$$\text{for the exact advection equation } |\lambda| = 1, \tag{6.39}$$

*regardless of the value of* $\Delta t$. To go from (6.38) to (6.39), we have used Euler's formula.

It will be shown later that for processes other than advection the exact value of $|\lambda|$ can differ from 1, and can depend on $\Delta t$.

From (6.35) we see that after $n$ time steps, starting from $n = 0$, the solution will be

$$\widehat{A}^n = \widehat{A}^0 \lambda^n. \tag{6.40}$$

From (6.40), the requirement for stability, i.e., that the solution remains bounded after arbitrarily many time steps, implies that

$$\boxed{|\lambda| \leq 1}. \tag{6.41}$$

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 (6.41) is satisfied.

Consider the particular case of the upstream scheme, as given by (6.8). Substituting (6.34) into (6.8) leads to

$$\frac{\widehat{A}^{n+1} - \widehat{A}^n}{\Delta t} + \left(\frac{1 - e^{-ik\Delta x}}{\Delta x}\right) c\widehat{A}^n = 0. \tag{6.42}$$

Notice that the true advection speed, $c$, is multiplied, in (6.42), by the factor $\left(\frac{1-e^{-ik\Delta x}}{\Delta x}\right)$. Comparing (6.42) with (6.31), we see that $\left(\frac{1-e^{-ik\Delta x}}{\Delta x}\right) c$ is "taking the place" of $ikc$ in the exact solution. In fact, you should be able to show that,

$$\lim_{\Delta x \to 0} \left(\frac{1 - e^{-ik\Delta x}}{\Delta x}\right) = ik. \tag{6.43}$$

This is a clue that, for a given value of $k$, the upstream scheme does not advect $A$ at the correct speed. We return to this point later.

For now, we use the definition of $\lambda$, i.e., (6.35), together with (4.5) and (6.42), to infer that

$$\lambda = 1 - \mu\,(1 - \cos k\Delta x + i \sin k\Delta x)\,. \tag{6.44}$$

Note that $\lambda$ is complex. This is to be expected, because the effective value of $\lambda$ for the exact solution of the differential equation is also complex. Computing the square of the modulus of both sides of (6.44), we obtain

$$|\lambda|^2 = 1 + 2\mu\,(\mu - 1)\,(1 - \cos k\Delta x)\,. \tag{6.45}$$

According to (6.45), the amplification factor $|\lambda|$ depends on the wave number, $k$, and also on $\mu$. As an example, for $\mu = \frac{1}{2}$, (6.45) reduces to

$$|\lambda|^2 = 1 - \frac{1}{2}\,(1 - \cos k\Delta x)\,. \tag{6.46}$$

Fig. 3.5 shows that the upstream scheme damps for $0 \leq \mu \leq 1$ and is unstable for $\mu < 0$ and $\mu > 1$. For $\mu$ close to zero, the scheme is close to neutral, but many time steps are needed to complete a given simulation. For $\mu$ close to one, the scheme is again close to neutral, but it is also close to instability. If we choose intermediate values of $\mu$, the shortest modes are strongly damped, and such strong smoothing is usually unacceptable. No matter what we do, there are problems with the scheme.

Although $\lambda$ 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 fact 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.

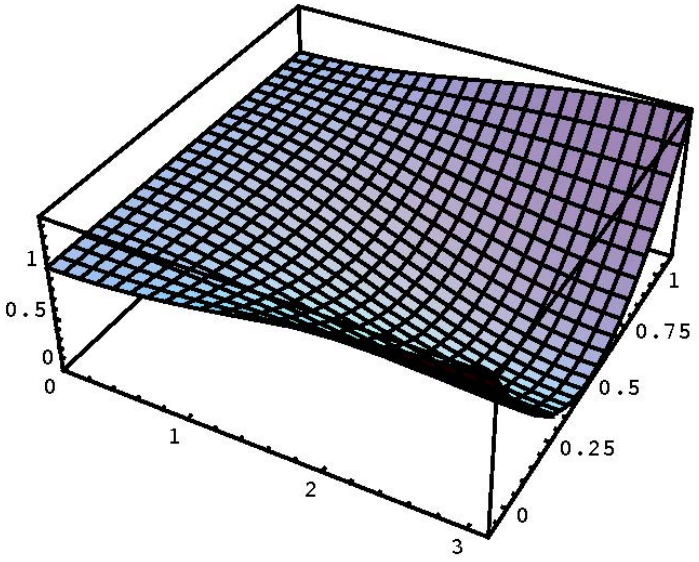



Figure 6.5: The square of the amplification factor for the upstream scheme is shown on the vertical axis. The front axis is $k\Delta x$, and the right-side axis is $\mu$.

If von Neumann's method says that a scheme is *unstable*, you can be confident that it is really is unstable.

As mentioned above, the full solution for $A_j^n$ can be expressed as a Fourier series. For simplicity, we assume that the solution 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

$$\begin{aligned} A_j^n =& \mathrm{Re}\left[\sum_{m=-\infty}^{\infty} \widehat{A}_m^n e^{imk_0 j\Delta x}\right] \\ =& \mathrm{Re}\left[\sum_{m=-\infty}^{\infty} \widehat{A}_m^0 e^{imk_0 j\Delta x}(\lambda_m)^n\right], \end{aligned} \tag{6.47}$$

where

$$k \equiv mk_0, \tag{6.48}$$

$$k_0 \equiv \frac{2\pi}{L_0}, \tag{6.49}$$

and $m$ is a nondimensional integer (i.e., a "pure number"), which is analogous to what we call 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 (6.47), the summation has been formally taken over all integers, although of course only a finite number of $m$'s can be used in a real application. We can interpret $|\lambda_m|$ as the amplification factor for mode $m$. This means that different wave numbers have different amplification factors. If *any* wave number is unstable, then the scheme is unstable. We can write

$$\begin{aligned} \left|A_j^n\right| &\leq \left| \sum_{m=-\infty}^{\infty} \widehat{A}_m^0 e^{imk_0 j\Delta x} (\lambda_m)^n \right| \\ &\leq \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 e^{imk_0 j\Delta x} (\lambda_m)^n \right| \\ &= \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 \right| |\lambda|^n. \end{aligned} \tag{6.50}$$

If $|\lambda| \leq 1$ is satisfied for all $m$, then

$$\left|A_j^n\right| \leq \sum_{m=-\infty}^{\infty} \left| \widehat{A}_m^0 \right|. \tag{6.51}$$

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 $u_{m_1} = 1$ and $u_m = 0$ for all $m \neq m_1$ is unbounded.

From (6.18), $\lambda_m$ for the upstream scheme is given by

$$\lambda = 1 - \mu \left(1 - \cos mk_0\Delta x + i \sin mk_0\Delta x\right). \tag{6.52}$$

The amplification factor is

$$|\lambda| = \sqrt{1 + 2\mu \left(\mu - 1\right)\left(1 - \cos mk_0\Delta x\right)}. \tag{6.53}$$

From (6.53) we can show that $|\lambda| \leq 1$ holds for all $m$, if and only if $\mu(\mu-1) \leq 0$, which is equivalent to $0 \leq \mu \leq 1$. This is the necessary and sufficient condition for the stability of the scheme.

## 6.6 How to take into account periodic boundary conditions

To explicitly allow for a finite periodic domain, the upstream scheme can be written in matrix form as

$$\begin{bmatrix} A_1^{n+1} \\ A_2^{n+1} \\ \cdots \\ A_{j-1}^{n+1} \\ A_j^{n+1} \\ A_{j+1}^{n+1} \\ \cdots \\ A_{J-1}^{n+1} \\ A_J^{n+1} \end{bmatrix} = \begin{bmatrix} 1-\mu & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \mu \\ \mu & 1-\mu & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \mu & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1-\mu & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \mu & 1-\mu & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mu & 1-\mu & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \mu & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1-\mu & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \mu & 1-\mu \end{bmatrix} \begin{bmatrix} A_1^n \\ A_2^n \\ \cdots \\ A_{j-1}^{n1} \\ A_j^n \\ A_{j+1}^n \\ \cdots \\ A_{J-1}^n \\ A_J^n \end{bmatrix}, \tag{6.54}$$

or

$$\left[A_j^{n+1}\right] = [M]\left[A_j^n\right], \tag{6.55}$$

where $[M]$ is the matrix written out on the right-hand side of (6.54). In writing (6.54), the cyclic boundary condition

$$A_1^{n+1} = (1-\mu)A_1^n + \mu A_J^n \tag{6.56}$$

has been assumed, and that is why $\mu$ appears in the top-right corner of the matrix. From the definition of $\lambda$, (6.35), we can write

$$\begin{bmatrix} A_1^{n+1} \\ A_2^{n+1} \\ \cdots \\ A_{j-1}^{n+1} \\ A_j^{n+1} \\ A_{j+1}^{n+1} \\ \cdots \\ A_{J-1}^{n+1} \\ A_J^{n+1} \end{bmatrix} = \begin{bmatrix} \lambda & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \begin{bmatrix} A_1^n \\ A_2^n \\ \cdots \\ A_{j-1}^{n1} \\ A_j^n \\ A_{j+1}^n \\ \cdots \\ A_{J-1}^n \\ A_J^n \end{bmatrix} \tag{6.57}$$

or

$$\left[A_j^{n+1}\right] = \lambda [I] \left[A_j^n\right], \tag{6.58}$$

where $[I]$ is the identity matrix. Comparing (6.55) and (6.58), we see that

$$([M] - \lambda [I]) \left[A_j^n\right] = 0. \tag{6.59}$$

This equation must hold regardless of the values of the $A_j^n$. It follows that the amplification factors, $\lambda$, are the *eigenvalues* of $[M]$, obtained by solving

$$|[M] - \lambda [I]| = 0, \tag{6.60}$$

where the absolute value signs denote the determinant. For the current example, we can use (6.60) to show that

$$\lambda = 1 - \mu \left(1 - e^{i\frac{2m\pi}{J}}\right), m = 0, 1, 2, \ldots, J-1. \tag{6.61}$$

This has essentially the same form as (6.44), and so it turns out that $0 \leq \mu \leq 1$ is the stability condition again.

## 6.7 Does the solution improve if we increase the number of grid points and cut the time step?

Consider what happens when we increase the number of grid points, *while fixing the domain size, D, the wind speed, c, 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 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, which is an error. The increased spatial resolution is a good thing, but the increased number of time steps is a bad thing. Does the solution improve, or not?

Consider grid spacing $\Delta x$, such that

$$D = J\Delta x. \tag{6.62}$$

As we decrease $\Delta x$, we increase $J$ correspondingly, so that $D$ does not change, and

$$k\Delta x = \frac{kD}{J}. \tag{6.63}$$

Substituting this into (6.45), we find that the amplification factor satisfies

$$|\lambda|^2 = 1 + 2\mu\left(\mu - 1\right)\left[1 - \cos\left(\frac{kD}{J}\right)\right]. \tag{6.64}$$

In order to maintain computational stability, *we keep $\mu$ fixed as $\Delta x$ decreases*, so that

$$\begin{aligned} \Delta t &= \frac{\mu\Delta x}{c} \\ &= \frac{\mu D}{cJ}. \end{aligned} \tag{6.65}$$

The time required for the air to flow through the domain is

$$T = \frac{D}{c}. \tag{6.66}$$

Let $N$ be the number of time steps needed for the air to flow through the domain, so that

$$\begin{aligned} N &= \frac{T}{\Delta t} \\ &= \frac{D}{c\Delta t} \\ &= \frac{D}{\mu \Delta x} \\ &= \frac{J}{\mu} \end{aligned} \tag{6.67}$$

To obtain the last line of (6.67), we have substituted from (6.62). The total amount of damping that "accumulates" as the air moves across the domain is given by

$$|\lambda|^N = \left(|\lambda|^2\right)^{N/2} = \left\{1 - 2\mu\left(1-\mu\right)\left[1-\cos\left(kD/J\right)\right]\right\}^{\frac{J}{2\mu}}. \tag{6.68}$$

Here we have used (6.64) and (6.67).

As we increase the resolution with a fixed domain size, $J$ increases. In Fig. 6.6, we show the dependence of $|\lambda|^N$ on $J$, for two different fixed values of $\mu$. The wavelength is assumed to be half the domain width, so that $kD = 4\pi$. This causes the cosine factor in (6.68) to approach 1, which weakens the damping associated with $|\lambda| < 1$; but on the other hand it also causes the exponent in (6.68) to increase, which strengthens the damping. Which effect dominates? The answer can be seen in Fig. 6.6. Increasing $J$ leads to less total damping for a given value of $\mu$, even though the number of time steps needed to cross the domain increases. This is good news.

On the other hand, if we fix $J$ and decrease $\mu$ (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* $\mu$. We minimize the error by living dangerously.

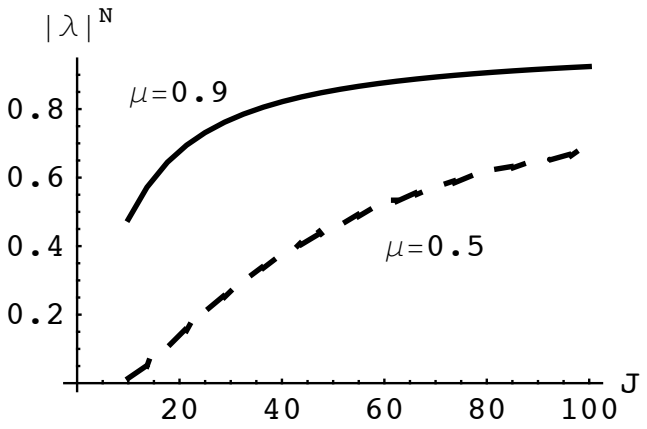



Figure 6.6: "Total" damping experienced by a disturbance crossing the domain, as a function of $J$, the number of grid points across the domain, for two different fixed values of $\mu$. In these examples we have assumed $D/L = 2$, i.e., the wavelength is half the width of the domain.

## 6.8 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 in Chapter 11.

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.

## 6.9 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 $\mu$ = -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. Consider the following pair of equations, which describe inertial oscillations:

$$\frac{du}{dt} = fv, \tag{6.69}$$

$$\frac{dv}{dt} = -fu. \tag{6.70}$$

   (a) Show that kinetic energy is conserved by this system.

   (b) *Using the energy method*, determine the stability of the forward time-differencing scheme as applied to these two equations.

3. Analyze the stability of

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = 0 \tag{6.71}$$

   using von Neumann's method.

# Chapter 7

# "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).

## 7.1 Accuracy and stability of a family of advection schemes

Following Takacs (1985), we define a fairly general family of schemes of the form

$$A_j^{n+1} - \sum_{j'=-\infty}^{\infty} a_{j'} A_{j'}^n = 0. \tag{7.1}$$

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 assume for simplicity that the grid spacing is uniform. Recall from Chapter 3 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 (7.1) 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 (7.1) by the discretization error of the scheme, denoted by $\varepsilon$, we find that

$$\left(A + \Delta t \frac{\partial A}{\partial t} + \frac{\Delta t^2}{2!}\frac{\partial^2 A}{\partial t^2} + \ldots\right) - \sum_{j'=-\infty}^{\infty} a_{j'} \left[A + \left(j'\Delta x\right)\frac{\partial A}{\partial x} + \frac{\left(j'\Delta x\right)^2}{2!}\frac{\partial^2 A}{\partial x^2} + \ldots\right] = \varepsilon, \tag{7.2}$$

where all quantities are evaluated at $(x,t) = \left(x_j, t^n\right)$. We now use the continuous advection equation in the forms

$$\frac{\partial A}{\partial t} = -c\frac{\partial A}{\partial x}, \quad \frac{\partial^2 A}{\partial t^2} = c^2\frac{\partial^2 A}{\partial x^2}, \quad \text{etc.}. \tag{7.3}$$

These relations only hold if $c$ is constant, so the results derived below are only approximately valid when $c$ is variable. For derivatives of order $m$, (7.3) generalizes to

$$\frac{\partial^m A}{\partial t^m} = (-c)^m\frac{\partial^m A}{\partial x^m}. \tag{7.4}$$

This allows us to write

$$\Delta t^m\frac{\partial^m A}{\partial t^m} = (-\mu)^m\Delta x^m\frac{\partial^m A}{\partial x^m}, \tag{7.5}$$

or

$$\Delta x^m\frac{\partial^m A}{\partial x^m} = \left(\frac{\Delta t}{-\mu}\right)^m\frac{\partial^m A}{\partial t^m}, \tag{7.6}$$

where $\mu$ is the usual CFL parameter. With the use of (7.6), we can rewrite (7.2) as

$$\left(A+\Delta t\frac{\partial A}{\partial t}+\frac{\Delta t^2}{2!}\frac{\partial^2 A}{\partial t^2}+\ldots\right)-\sum_{j'=-\infty}^{\infty}a_{j'}\left[A-\left(\frac{j'\Delta t}{\mu}\right)\frac{\partial A}{\partial t}+\frac{1}{2!}\left(\frac{j'\Delta t}{\mu}\right)^2\frac{\partial^2 A}{\partial t^2}+\ldots\right]=\varepsilon, \tag{7.7}$$

Inspection of (7.7) shows that in order to ensure first-order accuracy in both time and space, we need

$$1-\sum_{j'=-\infty}^{\infty}a_{j'}=0 \tag{7.8}$$

and

$$1+\sum_{j'=-\infty}^{\infty} a_{j'}\frac{j'}{\mu}=0. \tag{7.9}$$

To have second-order accuracy in both time and space, we need

$$1-\sum_{j'=-\infty}^{\infty} a_{j'}\left(\frac{j'}{\mu}\right)^2=0. \tag{7.10}$$

In general, to have $m$th-order accuracy in both time and space, we must require (7.8), (7.9), and

$$\sum_{j'=-\infty}^{\infty}\left(j'\right)^l a_{j'}=(-\mu)^l \quad \text{for } l=2 \text{ to } m. \tag{7.11}$$

Using the binomial theorem, it can be shown from (7.8) through (7.11) that for a scheme of $m$th-order accuracy

$$\sum_{j'=-\infty}^{\infty}\left(j'+\mu\right)^l a_{j'}=0 \text{ for } 1\le l\le m\,. \tag{7.12}$$

This will be used later.

We close this section by finding the amplification factor for the family of schemes given by (7.1). As usual, we look for a solution of the form

$$A_j^n=\text{Re}\left[\hat{A}^n e^{ikj\Delta x}\right]. \tag{7.13}$$

It follows from (7.1) and (7.13) that the amplification factor is given by

$$\lambda=\sum_{j'=-\infty}^{\infty} a_{j'}e^{ikj\Delta x}. \tag{7.14}$$

In the following sections, we discuss several schemes to which the preceding analysis is directly applicable.

## 7.2 Matsuno time-differencing with centered space differencing

In the case of the Matsuno scheme, the first approximation to $A_j^{n+1}$ comes from

$$\frac{A_j^{n+1*} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = 0, \tag{7.15}$$

and the final value from

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^{n+1*} - A_{j-1}^{n+1*}}{2\Delta x}\right) = 0. \tag{7.16}$$

Eliminate the terms with $(\,)*$ from (7.16) by using (7.15) twice (first with $j$ replaced by $j+1$, then with $j$ replaced by $j-1$). The result can be written as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = \frac{c^2\Delta t}{(2\Delta x)^2}\left(A_{j+2}^n - 2A_j^n + A_{j-2}^n\right). \tag{7.17}$$

It should be clear that (7.17) is a member of the family given by (7.1). The term on the right-hand side of (7.17) approaches zero as $\Delta t \to 0$, and thus (7.17) 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 $c^2\Delta t$, which goes to zero as $\Delta t \to 0$. We say that the centered-in-space advection scheme with Matsuno time differencing is "diffusive."

## 7.3 The Lax-Wendroff scheme

A similarly diffusive scheme, called the Lax-Wendroff scheme, has second-order accuracy. Consider an explicit two-level scheme of the form:

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + \frac{c}{\Delta x}\left(a_{j-1}A_{j-1}^n + a_j A_j^n + a_{j+1}A_{j+1}^n\right) = 0. \tag{7.18}$$

The scheme given by Eq. (7.22) was proposed by Lax and Wendroff (1960), and recommended by Richtmyer (1963). It is a member of the family given by (7.1). 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

$$a_{j-1} + a_j + a_{j+1} = 0, \text{ and } -a_{j-1} + a_{j+1} = 1. \tag{7.19}$$

To obtain second-order accuracy in both time and space, we must also enforce

$$\mu + a_{j-1} + a_{j+1} = 0. \tag{7.20}$$

Solving (7.19) and (7.20) for the parameters of the scheme, we find that

$$a_{j-1} = \frac{-1-\mu}{2},\ a_j = \mu \text{ and } a_{j+1} = \frac{1-\mu}{2}. \tag{7.21}$$

For $\mu > 0$ (which means $c > 0$), the absolute value of the upstream coefficient, $a_{j-1}$, is larger than the absolute value of the downstream coefficient, $a_{j+1}$. Something similar happens if $\mu < 0$. In short, the scheme is automatically "upstream-weighted" regardless of which way the wind is blowing, even though the stencil is centered. Substituting from (7.21) into (7.18), we can write the scheme as

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} + c\left(\frac{A_{j+1}^n - A_{j-1}^n}{2\Delta x}\right) = \frac{c^2\Delta t}{2\Delta x^2}\left(A_{j+1}^n - 2A_j^n + A_{j-1}^n\right). \tag{7.22}$$

Compare (7.22) with (7.17), which is the corresponding result for the Matsuno scheme. The left-hand side of (7.22) 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 (7.22) is second-order accurate in time, even though it involves only two time levels; this result 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 right-hand-side of (7.22) looks like a diffusion term. This is similar to what happens when the Matsuno time differencing scheme is combined with centered space differencing.

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

$$\frac{A_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}\left(A_{j+1}^{n} + A_{j}^{n}\right)}{\frac{1}{2}\Delta t} = -c\left(\frac{A_{j+1}^{n} + A_{j}^{n}}{\Delta x}\right), \tag{7.23}$$

$$\frac{A_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}\left(A_{j}^{n} + A_{j-1}^{n}\right)}{\frac{1}{2}\Delta t} = -c\left(\frac{A_{j}^{n} - A_{j-1}^{n}}{\Delta x}\right), \tag{7.24}$$

and then use these to obtain $A_j^{n+1}$ from

$$\frac{A_j^{n+1} - A_j^n}{\Delta t} = -c\left(\frac{A_{j+\frac{1}{2}}^{n+\frac{1}{2}} - A_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x}\right). \tag{7.25}$$

Note that (7.25) is "centered in time." If (7.23) and (7.24) are substituted into (7.25), we recover (7.22). 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

$$\lambda = 1 - 2\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) - i\mu \sin\left(k\Delta x\right), \tag{7.26}$$

To obtain (7.26), we have used the trigonometric identity $2\sin^2\left(\frac{\theta}{2}\right) = 1 - \cos\theta$. We find that

$$\begin{aligned} |\lambda|^2 &= \left[1 - 4\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) + 4\mu^4 \sin^4\left(\frac{k\Delta x}{2}\right)\right] + \mu^2 \sin^2\left(k\Delta x\right) \\ &= 1 - 4\mu^2\left(1 - \mu^2\right)\sin^4\left(\frac{k\Delta x}{2}\right). \end{aligned} \tag{7.27}$$

To obtain the second line of (7.27), we have used the trigonometric identity $\sin(2\theta) = 2\sin\theta\cos\theta$. Since (7.27) involves only $\mu^2$, the stability criterion does not depend on the direction of the wind. If $\mu^2 < 1$, then $|\lambda| < 1$ and the scheme is dissipative. Fig. 7.1 shows how $|\lambda|^2$ depends on $\mu$ and $L$, for both the upstream and Lax-Wendroff schemes. The damping of the Lax-Wendroff scheme is more scale selective.

## 7.4 Implicit schemes for the advection equation

There are also various implicit schemes, such as the trapezoidal implicit scheme, which are neutral and unconditionally stable, so that in principle any $\Delta t$ can be used if the phase error can be tolerated. Such schemes are *not* members of the family defined by (7.1). Implicit schemes have the drawback that an iterative procedure is usually needed to solve the system of equations involved. In many cases, the iterative procedure may take as much computer time as a simpler non-iterative scheme with a smaller $\Delta t$.

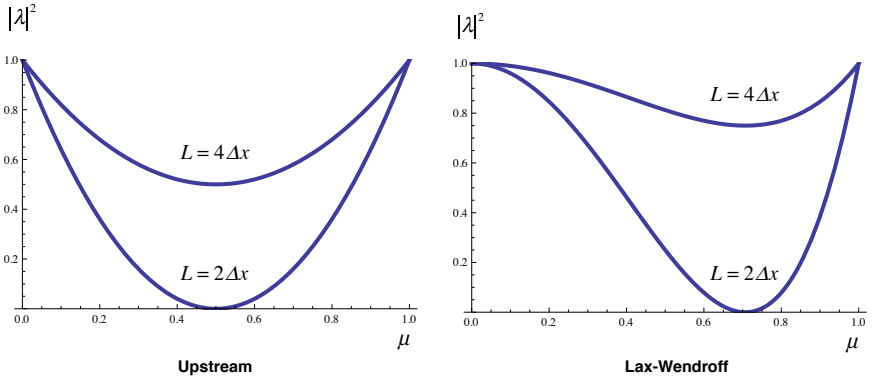



Figure 7.1: The amplification factors for the Lax-Wendroff and upstream schemes, for two different wavelengths, plotted as a function of $\mu^2$.

## 7.5 Two-dimensional advection

Variable currents more or less have to be 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

$$\frac{\partial A}{\partial t} + u\frac{\partial A}{\partial x} + v\frac{\partial A}{\partial y} = 0, \tag{7.28}$$

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

$$\frac{dA_{i,j}}{dt} + u\frac{1}{2\Delta x}\left(A_{i+1,j} - A_{i-1,j}\right) + v\frac{1}{2\Delta y}\left(A_{i,j+1} - A_{i,j-1}\right) = 0. \tag{7.29}$$

Assume that $A$ has the form

$$A_{i,j} = \mathrm{Re}\left\{\widehat{A}(t)\, e^{\underset{\sim}{i}[(ki\Delta x)+(lj\Delta y)]}\right\}, \tag{7.30}$$

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:

$$\frac{d\widehat{A}}{dt} = i\omega\widehat{A}\,, \tag{7.31}$$

where this this time the frequency is given by

$$\omega \equiv -\left[u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right]. \tag{7.32}$$

If we were to use leapfrog time-differencing, the stability criterion would be

$$\left|u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right|\Delta t \le 1. \tag{7.33}$$

Since

$$\begin{aligned}\left|u\frac{\sin(k\Delta x)}{\Delta x} + v\frac{\sin(l\Delta y)}{\Delta y}\right|\Delta t &\le \left(\left|u\frac{\sin(k\Delta x)}{\Delta x}\right| + \left|v\frac{\sin(l\Delta y)}{\Delta y}\right|\right)\Delta t \\ &\le \left(\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y}\right)\Delta t,\end{aligned} \tag{7.34}$$

a *sufficient* condition to satisfy (7.33) is

$$\left(\frac{|u|}{\Delta x}+\frac{|v|}{\Delta y}\right)\Delta t \leq 1. \tag{7.35}$$

If we require the scheme to be stable for all possible $k$ and $l$, and for all combinations of $u$ and $v$, then (7.35) 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

$$|u| \equiv S\cos\alpha \text{ and } |v| \equiv S\sin\alpha, \tag{7.36}$$

where the wind speed $S \geq 0$ and $0 \leq \alpha \leq \pi/2$. For $\alpha = 0$, the flow is zonal, and for $\alpha = \pi/2$ it is meridional. Then (7.35) becomes

$$S\left(\frac{\cos\alpha}{\Delta x}+\frac{\sin\alpha}{\Delta y}\right)\Delta t \leq 1. \tag{7.37}$$

In order for the scheme to be stable *for any orientation of the current*, we must have

$$S\left(\frac{\cos\alpha_m}{\Delta x}+\frac{\sin\alpha_m}{\Delta y}\right)\Delta t \leq 1, \tag{7.38}$$

where $\alpha_m$ is the "worst-case" $\alpha$, which makes the left-hand side of (7.37) a maximum. We can show that $\alpha_m$ satisfies

$$\tan\alpha_m = \frac{\Delta x}{\Delta y}, \text{ so that } \sin\alpha_m = \frac{\Delta x}{\sqrt{(\Delta x)^2+(\Delta y)^2}} \text{ and } \cos\alpha_m = \frac{\Delta y}{\sqrt{(\Delta x)^2+(\Delta y)^2}}. \tag{7.39}$$

As shown in Fig. 7.2, $\alpha_m$ measures the angle of the "diagonal" across a grid box. For example, when $\frac{\Delta y}{\Delta x} << 1$, we get $\alpha_m \to \pi/2$, 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 $\Delta x = \Delta y$, we get $\alpha_m = \pi/4$.

From (7.38) and (7.39) we see that the stability criterion can be written as

$$\frac{C\Delta t}{\sqrt{(\Delta x)^2+(\Delta y)^2}}\left(\frac{\Delta y}{\Delta x}+\frac{\Delta x}{\Delta y}\right)\leq 1. \tag{7.40}$$

In particular, for $\Delta x=\Delta y=d$,

$$\frac{C\Delta t}{d}\leq\frac{1}{\sqrt{2}}<1. \tag{7.41}$$

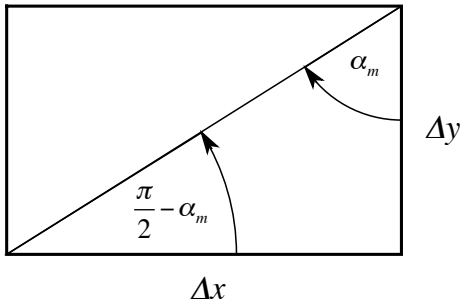



Figure 7.2: Sketch illustrating the angle $\alpha_m$ on a rectangular grid.

# Chapter 8

# Finite-volume methods

## 8.1 Definitions

As discussed in Appendix B, 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 a class of finite-difference methods called "finite-volume methods," in which the definition of the operator is used but the volume takes the form of a finite "grid cell."

For example, the divergence operator can be defined using

$$\nabla \cdot \mathbf{Q} \equiv \lim_{V \to 0} \left[ \frac{1}{V} \oint_S \mathbf{n} \cdot \mathbf{Q} dS \right] \tag{8.1}$$

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 (8.2) is one in which both the volume and the area of its bounding surface shrink to zero.

Eq. (8.1) 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:

$$\nabla A \equiv \lim_{V \to 0} \left[ \frac{1}{V} \oint_S \mathbf{n} A dS \right], \tag{8.2}$$

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. For example, 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:

$$\nabla \times \mathbf{Q} \equiv \lim_{V \to 0} \left[ \frac{1}{V} \oint_S \mathbf{n} \times \mathbf{Q} dS \right] \tag{8.3}$$

This can be used to formulate an approximation to the vorticity in terms of the tangential wind components on the wall of the volume.

Finally, the Jacobian on a two-dimensional surface can be defined by

$$J(A,B) = \lim_{C \to 0} \left[ \oint_C A \nabla B \cdot \mathbf{t} dl \right], \tag{8.4}$$

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 $B$.

## 8.2 How is discrete conservation defined?

When we design finite-difference schemes to represent advection, we strive, as always, for accuracy, stability, simplicity, and computational economy. In addition, it is often required that a finite-difference scheme for advection be conservative in the sense that

$$\sum_j \rho_j^{n+1} dR_j = \sum_j \rho_j^n dR_j, \tag{8.5}$$

and

$$\sum_j (\rho A)_j^{n+1} dR_j = \sum_j (\rho A)_j^n dR_j + \Delta t \sum_j (\rho S)_j^n dR_j. \tag{8.6}$$

These are finite-difference analogs to the integral forms (5.7) and (5.8), respectively. In (8.6), we have assumed for simplicity that the effects of the source, $S$, are evaluated using forward time differencing, although this need not be the case in general.

We may also wish to require conservation 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.

# Chapter 9

# Conservative advection schemes

## 9.1 Continuous advection in one dimension

Let $A$ be a "conservative" variable, satisfying the following one-dimensional conservation law:

$$\frac{\partial}{\partial t}(\rho A) + \frac{\partial}{\partial x}(\rho u A) = 0. \tag{9.1}$$

Here $\rho$ is the density of the air, and $\rho u$ is a mass flux. Putting $A \equiv 1$ in (9.1) gives mass conservation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0. \tag{9.2}$$

By combining (9.1) and (9.2), we can write the "advective form" of the conservation equation for $A$ as

$$\rho\left(\frac{\partial A}{\partial t} + u\frac{\partial A}{\partial x}\right) = 0. \tag{9.3}$$

The continuity equation itself can be rewritten as

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = 0. \tag{9.4}$$

When the wind field is non-divergent, this reduces to an "advection equation" for the density:

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = 0. \tag{9.5}$$

Note, however, that in general the density is *not* conserved following a particle, because in general the wind field is divergent.

## 9.2 Conserving mass

Suppose that we approximate (9.2) by:

$$\frac{d\rho_j}{dt} + \frac{(\rho u)_{j+\frac{1}{2}} - (\rho u)_{j-\frac{1}{2}}}{\Delta x_j} = 0, \tag{9.6}$$

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 $\rho$ is defined at integer points, while $u$ and $\rho u$ are defined at half-integer points. See Fig. 9.1. In order to use this approach, the wind-point quantities $\rho_{j+\frac{1}{2}}$ and $\rho_{j-\frac{1}{2}}$ must be interpolated somehow from the predicted values of $\rho$. This is an example of a "staggered" grid. The properties of staggered grids, in multiple spatial dimensions, are discussed in detail in later chapters.

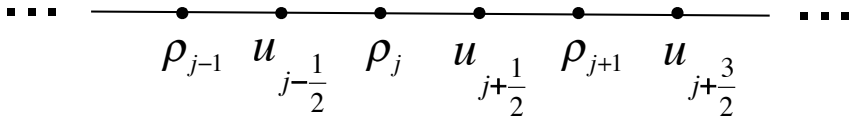



Figure 9.1: The staggered grid used in (9.10) and (9.6).

Multiply (9.6) through by $\Delta x_j$, and sum over the domain, to obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j \Delta x_j\right) + (\rho u)_{J+\frac{1}{2}} - (\rho u)_{-\frac{1}{2}} = 0. \tag{9.7}$$

If

$$(\rho u)_{J+\frac{1}{2}} = (\rho u)_{-\frac{1}{2}} \tag{9.8}$$

(these are periodic boundary conditions), then we obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j \Delta x_j\right) = 0, \tag{9.9}$$

which expresses conservation of mass (9.9). Compare with (8.5). Note that (9.9) holds regardless of the form of the interpolation used for $\rho_{j+\frac{1}{2}}$.

## 9.3 Conserving an intensive scalar

In a similar way, we approximate (9.1) by

$$\frac{d}{dt}\left(\rho_j A_j\right) + \frac{(\rho u)_{j+\frac{1}{2}} A_{j+\frac{1}{2}} - (\rho u)_{j-\frac{1}{2}} A_{j-\frac{1}{2}}}{\Delta x_j} = 0, \tag{9.10}$$

Here $A$, like $\rho$, is defined at integer points. The half-integer values of $A$, i.e., $A_{j+\frac{1}{2}}$, and $A_{j-\frac{1}{2}}$, must be interpolated somehow from the predicted values of $A$. Note that if we put $A \equiv 1$, (9.10) reduces to the finite-difference continuity equation, (9.6). As mentioned earlier, schemes that have this property are called "compatible."

Multiply (9.10) through by $\Delta x_j$, and sum over the domain:

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j A_j \Delta x_j\right) + (\rho u)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} - (\rho u)_{-\frac{1}{2}} A_{-\frac{1}{2}} = 0, \tag{9.11}$$

If

$$(\rho u)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} = (\rho u)_{-\frac{1}{2}} A_{-\frac{1}{2}}, \tag{9.12}$$

(these are periodic boundary conditions), then we obtain

$$\frac{d}{dt}\sum_{j=0}^{J}\left(\rho_j A_j \Delta x_j\right) = 0, \tag{9.13}$$

which expresses conservation of the mass-weighted value of $A$. Compare with (8.6). Note that (9.13) holds regardless of the form of the interpolation used for $A_{j+\frac{1}{2}}$.

A scheme that conserves the mass-weighted value of $A$ will also conserve any linear function of $A$.

## 9.4 An advective form

By combining (9.1) and (9.2), we obtain the advective form of our conservation law:

$$\rho\frac{\partial A}{\partial t} + \rho u\frac{\partial A}{\partial x} = 0. \tag{9.14}$$

From (9.10) and (9.6), we can derive a finite-difference "advective form," analogous to (9.14):

$$\rho_j\frac{dA_j}{dt} + \frac{(\rho u)_{j+\frac{1}{2}}\left(A_{j+\frac{1}{2}} - A_j\right) + (\rho u)_{j-\frac{1}{2}}\left(A_j - A_{j-\frac{1}{2}}\right)}{\Delta x_j} = 0. \tag{9.15}$$

Since (9.15) is consistent with (9.10) and (9.6), use of (9.15) and (9.6) 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 (9.15) gives $\frac{dA_j}{dt} = 0$, which is "the right answer." This is ensured ***because*** (9.10) reduces to (9.6) 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.*

If $\rho$ and $u$ are spatially uniform, then (9.15) reduces to

$$\frac{dA_j}{dt} + u\left(\frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x_j}\right) = 0. \tag{9.16}$$

Schemes of this form are discussed later in this chapter.

## 9.5 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., $A^2$, $A^{17}$, $ln(A)$, etc. This is most easily seen from the Lagrangian form:

$$\frac{DA}{Dt} = 0. \tag{9.17}$$

According to (9.17), $A$ is conserved "following a particle." As discussed earlier, this implies that

$$\frac{D}{Dt}\left[F\left(A\right)\right] = 0, \tag{9.18}$$

where $F(A)$ is an arbitrary function of $A$ only. We can derive (9.18) by multiplying (9.17) 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 $F\left(A_j\right)$, where $F$ is an arbitrary function, and let $F'_j$ denote $\frac{d\left[F\left(A_j\right)\right]}{dA_j}$. Multiplying (9.15) by $F'_j$ gives

$$\rho_j \frac{dF_j}{dt} + \frac{\left(\rho u\right)_{j+\frac{1}{2}} F'_j \left(A_{j+\frac{1}{2}} - A_j\right) + \left(\rho u\right)_{j-\frac{1}{2}} F'_j \left(A_j - A_{j-\frac{1}{2}}\right)}{\Delta x_j} = 0. \tag{9.19}$$

Now use (9.6) to rewrite (9.19) in "flux form":

$$\frac{d}{dt}\left(\rho_j F_j\right) + \frac{1}{\Delta x_j}\left\{\left(\rho u\right)_{j+\frac{1}{2}}\left[F'_j\left(A_{j+\frac{1}{2}} - A_j\right) + F_j\right] - \left(\rho u\right)_{j-\frac{1}{2}}\left[-F'_j\left(A_j - A_{j-\frac{1}{2}}\right) + F_j\right]\right\} = 0. \tag{9.20}$$

Inspection of (9.20) shows that, to ensure conservation of $F(A)$, we must choose

$$F_{j+\frac{1}{2}} = F'_j \left( A_{j+\frac{1}{2}} - A_j \right) + F_j, \tag{9.21}$$

$$F_{j-\frac{1}{2}} = -F'_j \left( A_j - A_{j-\frac{1}{2}} \right) + F_j. \tag{9.22}$$

Let $j \to j+1$ in (9.22), giving

$$F_{j+\frac{1}{2}} = -F'_{j+1} \left( A_{j+1} - A_{j+\frac{1}{2}} \right) + F_{j+1}. \tag{9.23}$$

Eliminating $F_{j+\frac{1}{2}}$ between (9.21) and (9.23), we find that $A_{j+\frac{1}{2}}$ must satisfy

$$\boxed{A_{j+\frac{1}{2}} = \frac{\left( F'_{j+1} A_{j+1} - F_{j+1} \right) - \left( F'_j A_j - F_j \right)}{F'_{j+1} - F'_j}.} \tag{9.24}$$

The conclusion is that by choosing $A_{j+\frac{1}{2}}$ according to (9.24), 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

$$A_{j+\frac{1}{2}} = \frac{\left( 2A^2_{j+1} - A^2_{j+1} \right) - \left( 2A^2_j - A^2_j \right)}{2\left( A_{j+1} - A_j \right)} = \frac{1}{2}\left( A_{j+1} + A_j \right). \tag{9.25}$$

This arithmetic-mean interpolation allows conservation of the square of $A$. It may or may not be an *accurate* interpolation for $A_{j+\frac{1}{2}}$. Note that $x_{j+1}$, $x_j$, and $x_{j+\frac{1}{2}}$ do not appear in (9.25). 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 scheme. If the grid spacing is uniform, (9.25) gives second-order accuracy in space. If the grid spacing is highly nonuniform, the accuracy drops to first-order, but if the grid spacing varies smoothly second-order accuracy can be maintained, as discussed in Chapter 2.

Substituting (9.25) back into (9.15) gives

$$\rho_j \frac{dA_j}{dt} + \frac{1}{2\Delta x_j}\left[(\rho u)_{j+\frac{1}{2}}\left(A_{j+1} - A_j\right) + (\rho u)_{j-\frac{1}{2}}\left(A_j - A_{j-1}\right)\right] = 0. \tag{9.26}$$

This is the advective form that allows conservation of $A^2$ (and of $A$).

## 9.6 Lots of ways to interpolate

There are infinitely many ways to interpolate a variable. We can spatially interpolate $A$ itself in a linear fashion, e.g.,

$$A_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} A_j + \left(1 - \alpha_{j+\frac{1}{2}}\right) A_{j+1}, \tag{9.27}$$

where $0 \leq \alpha_{j+\frac{1}{2}} \leq 1$ is a weighting factor that might be a constant, as in (9.25), or might be a function of $x_j$, $x_{j+1}$ and $x_{j+\frac{1}{2}}$, or a function of $\mu \equiv c\Delta t/\Delta x$. Alternatively, we can interpolate so as to conserve an arbitrary function of $A$, as in (9.24).

Another approach is to compute some function of $f(A)$, interpolate $f(A)$ using a form such as (9.27), 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 (9.25). Examples are the "*geometric mean*,"

$$A_{j+\frac{1}{2}} = \sqrt{A_j A_{j+1}}. \tag{9.28}$$

and the "*harmonic mean*,"

$$A_{j+\frac{1}{2}} = \frac{2A_j A_{j+1}}{A_j + A_{j+1}}, \tag{9.29}$$

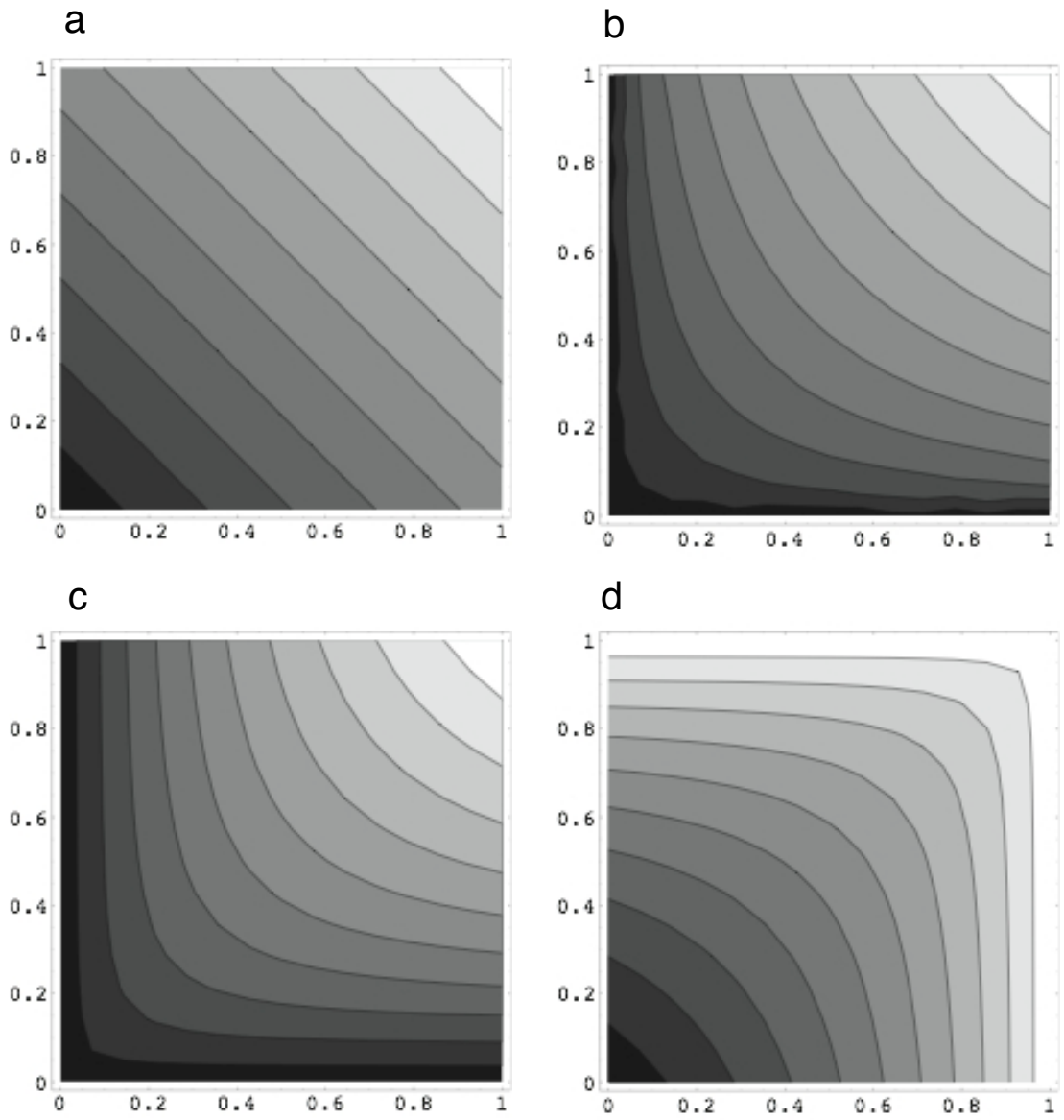



Figure 9.2: Four interpolations as functions of the input values. a) arithmetic mean, b) geometric mean, c) harmonic mean, d) Eq. (9.30), which makes the interpolated value close to the larger of the two input values. In all plots, black is close to zero, and white is close to one.

both of which are plotted in Fig. 9.2. Note that both (9.28) and (9.29) give $A_{j+\frac{1}{2}} = C$ if both $A_{j+1}$ and $A_j$ are equal to $C$, which is what we expect from an interpolation. They are both nonlinear interpolations. For example, 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 both have the potentially useful property that if either $A_{j+1}$ or $A_j$ is equal to zero, then $A_{j+\frac{1}{2}}$ will also be equal to zero. More generally, both (9.28) and (9.29) tend to make the interpolated value close to the smaller of the two

input values.

Here is another interpolation that has the opposite property, i.e., it makes the interpolated value close to the *larger* of the two input values. Define $r$ as a normalized value of $A$, such that $0 \leq r \leq 1$. For example, we could define $r_{j,j+\frac{1}{2}} \equiv \frac{A_j}{\text{Max}\left\{A_j, A_{j+1}\right\}}$ and $r_{j+1j+\frac{1}{2}} \equiv \frac{A_{j+1}}{\text{Max}\left\{A_j, A_{j+1}\right\}}$. We could then interpolate $r$ to the cell wall using

$$r_{j+\frac{1}{2}} = \frac{r_{j,j+\frac{1}{2}} + r_{j+1,j+\frac{1}{2}} - 2r_{j,j+\frac{1}{2}} r_{j+1,j+\frac{1}{2}}}{2 - \left(r_{j,j+\frac{1}{2}} + r_{j+1,j+\frac{1}{2}}\right)}. \tag{9.30}$$

We would then compute the cell-wall value of $A$ using $A_{j+\frac{1}{2}} \equiv r_{j+\frac{1}{2}} \text{Max}\left\{A_j, A_{j+1}\right\}$. It can be confirmed that when $r_{j,j+\frac{1}{2}} = r_{j+1,j+\frac{1}{2}}$ Eq. (9.30) gives $r_{j+\frac{1}{2}} = r_{j,j+\frac{1}{2}} = r_{j+1,j+\frac{1}{2}}$, so (9.30) can be interpreted as an interpolation. Eq. (9.30) is also plotted in Fig. 9.2.

All of the interpolations described above can be reformulated in terms of $r$. The interpolation given by (9.30) can be interpreted as one minus the harmonic mean formulated in terms of $r$.

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* way for a particular application.

## 9.7 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. These *ad hoc* procedures are called "fixers" (not to be confused with shady attorneys). I don't recommend them. For further discussion of fixers, see Takacs (1988) and Diamantakis and Flemming (2014).

## 9.8 A flux form of the upstream scheme

In Chapter 3, 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 $A$ in the flux-form equation (9.10) so that the corresponding advective form is the upstream scheme.

Suppose that $(\rho u)_{j+\frac{1}{2}}$ and $(\rho u)_{j-\frac{1}{2}}$ are both positive. Then the spatial differencing of

(9.15) is consistent with the upstream scheme if we choose $A_{j+\frac{1}{2}} = A_j$ and $A_{j-\frac{1}{2}} = A_{j-1}$, and (9.15) reduces to

$$\rho_j \frac{dA_j}{dt} + \frac{(\rho u)_{j-\frac{1}{2}} \left(A_j - A_{j-1}\right)}{\Delta x_j} = 0. \tag{9.31}$$

On the other hand, if $(\rho u)_{j+\frac{1}{2}}$ and $(\rho u)_{j-\frac{1}{2}}$ are both negative, we can get the upstream scheme with $A_{j+\frac{1}{2}} = A_{j+1}$ and $A_{j-\frac{1}{2}} = A_j$, in which case (9.15) becomes

$$\rho_j \frac{dA_j}{dt} + \frac{(\rho u)_{j+\frac{1}{2}} \left(A_{j+1} - A_j\right)}{\Delta x_j} = 0. \tag{9.32}$$

To see how to proceed more generally, regardless of the signs of the mass fluxes, we go back to the flux form, (9.10), and write

$$\begin{aligned} \frac{d}{dt}\left(\rho_j A_j\right) + \left[\frac{\text{Max}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\} A_j + \text{Min}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\} A_{j+1}}{\Delta x_j}\right] \\ - \left[\frac{\text{Max}\left\{(\rho u)_{j-\frac{1}{2}}, 0\right\} A_{j-1} + \text{Min}\left\{(\rho u)_{j-\frac{1}{2}}, 0\right\} A_j}{\Delta x_j}\right] = 0\,. \end{aligned} \tag{9.33}$$

By setting $A \equiv 1$ in (9.33), we see that a "compatible" form of the continuity equation is

$$\begin{aligned} \frac{d\rho_j}{dt} + \left[\frac{\text{Max}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\} + \text{Min}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\}}{\Delta x_j}\right] \\ - \left[\frac{\text{Max}\left\{(\rho u)_{j-\frac{1}{2}}, 0\right\} + \text{Min}\left\{(\rho u)_{j-\frac{1}{2}}, 0\right\}}{\Delta x_j}\right] = 0\,, \end{aligned} \tag{9.34}$$

which is equivalent to (9.6). Of course, compatibility would also imply the use of forward time-differencing in the continuity equation. That will be unstable in combination with centered-in-space differencing for the density. How do we know that it will be unstable?

Because, as shown in Chapter 4, the oscillation equation is unstable with the Euler forward scheme.

Recall, moreover, that the continuity equation itself reduces to "advection of density" when the fluid is incompressible and the wind field is non-divergent. We are thus led to consider an upstream version of the continuity equation. All we have to do is choose $\rho_{j+\frac{1}{2}} = \rho_j$ when $u_{j+\frac{1}{2}} > 0$, and $\rho_{j+\frac{1}{2}} = \rho_{j+1}$ when $u_{j+\frac{1}{2}} < 0$, with corresponding choices for $\rho_{j-\frac{1}{2}}$. We write

$$\text{Max}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\} = \rho_j \text{Max}\left\{u_{j+\frac{1}{2}}, 0\right\} \tag{9.35}$$

and

$$\text{Min}\left\{(\rho u)_{j+\frac{1}{2}}, 0\right\} = \rho_{j+1} \text{Min}\left\{u_{j+\frac{1}{2}}, 0\right\} \ . \tag{9.36}$$

Eqs. (9.35) and (9.36) can be substituted back into both (9.33) and (9.34).

## 9.9 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

$$\frac{d\rho_j}{dt} + \frac{(\hat{\rho} u)_{j+\frac{1}{2}} - (\hat{\rho} u)_{j-\frac{1}{2}}}{\Delta x} = 0, \tag{9.37}$$

and the advection equation

$$\frac{dA_j}{dt} + \frac{1}{2}\left(u_{j+\frac{1}{2}} + u_{j-\frac{1}{2}}\right)\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0. \tag{9.38}$$

Does this scheme conserve the mass-weighted average value of $A$? Give a proof to support your answer.

3. Program the following one-dimensional shallow-water model with a tracer $A$:

$$\begin{aligned}
&\frac{(hA)_j^{n+1} - (hA)_j^{n-1}}{2\Delta t} + \frac{\left(\hat{h}u\right)^n{}_{j+\frac{1}{2}} \hat{A}^n_{j+\frac{1}{2}} - \left(\hat{h}u\right)^n{}_{j-\frac{1}{2}} \hat{A}^n_{j-\frac{1}{2}}}{\Delta x} = 0, \\
&\frac{h_j{}^{n+1} - h_j{}^{n-1}}{2\Delta t} + \frac{\left(\hat{h}u\right)^n{}_{j+\frac{1}{2}} - \left(\hat{h}u\right)^n{}_{j-\frac{1}{2}}}{\Delta x} = 0, \\
&\frac{u^{n+1}{}_{j+\frac{1}{2}} - u^{n-1}{}_{j+\frac{1}{2}}}{2\Delta t} + \left(\frac{k^n{}_{j+1} - k^n{}_j}{\Delta x}\right) + g\left(\frac{h^n{}_{j+1} - h^n{}_j}{\Delta x}\right) = 0.
\end{aligned} \tag{9.39}$$

Use a forward time step for the first step only. Take

$$\begin{aligned}
\Delta x &= 10^5 \mathrm{m}, \\
g &= 0.1\,\mathrm{m\,s^{-2}}, \\
\hat{h}_{j+\frac{1}{2}} &= \frac{1}{2}\left(h_j + h_{j+1}\right), \\
k_j &= \frac{1}{4}\left(u^2{}_{j+\frac{1}{2}} + u^2{}_{j-\frac{1}{2}}\right).
\end{aligned} \tag{9.40}$$

Use 100 grid points, with periodic boundary conditions. Let the initial condition be

$$\begin{aligned}
u_{j+\frac{1}{2}} &= 0 \quad \text{for all} \quad j, \\
h_j &= 1000 + 50\sin\left(\frac{2\pi j}{20}\right), \\
A_j &= 11 + 10\cos\left(\frac{2\pi j}{4}\right).
\end{aligned} \tag{9.41}$$

(a) Use von Neumann's method to estimate the largest time step that is consistent with numerical stability.

(b) Construct the model, and experiment with time steps "close" to the predicted maximum stable $\Delta t$ (within a factor of 2), in order to find a value that is stable in practice.

(c) Run the model for the following two choices of $\hat{A}_{j+\frac{1}{2}}$:

$$\begin{aligned}\hat{A}_{j+\frac{1}{2}} &= \frac{1}{2}\left(A_j + A_{j+1}\right), \\ \hat{A}_{j+\frac{1}{2}} &= \sqrt{\text{Max}\left\{0, A_j A_{j+1}\right\}}.\end{aligned} \tag{9.42}$$

Run out to $t = 1.5 \times 10^6$ seconds. If you encounter $A < 0$, invent or choose a method to enforce $A \geq 0$ without violating conservation of $A$. Explain your method.

(d) Check the conservation of $A$ and $A^2$ for both cases. Explain how you do this.

4. Consider a periodic domain with cells numbered by $0 \leq j \leq 100$, with initial conditions

$$\begin{aligned}q_j &= 100, 45 \leq j \leq 55, \\ q_j &= 0 \text{ otherwise.}\end{aligned} \tag{9.43}$$

Integrate

$$\frac{\partial q}{\partial t} + c\frac{\partial q}{\partial x} = 0, \tag{9.44}$$

using

(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 once. Plot the results for the end of the run, and also the half-way point. Compare the solutions, with particular attention to amplitude errors, phase errors, dispersion, sign preservation, and monotonicity.

5. Determine the order of accuracy of

$$\left(\frac{\partial A}{\partial x}\right)_j \cong \frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x} \tag{9.45}$$

when

$$A_{j+\frac{1}{2}} = \frac{2A_jA_{j+1}}{A_j + A_{j+1}} \tag{9.46}$$

is used for interpolation to the cell walls. Assume uniform grid spacing.

6. Discuss the meaning of Eq. (9.24) for the special cases $F(A) = A$, and $F(A) = 1$.
7. 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, and the forward scheme is not.
8. Check the stability of the Matsuno scheme with centered space differencing, for the advection equation.
9. Prove that the geometric mean cannot be larger than the arithmetic mean.

# Chapter 10

# Computational dispersion

## 10.1 Centered space differencing and computational dispersion

Consider the centered-difference quotient

$$\left(\frac{\partial A}{\partial x}\right)_j \cong \frac{A_{j+1} - A_{j-1}}{2\Delta x}. \tag{10.1}$$

If $A_j(t)$ has the wave form $A_j(t) = \widehat{A}(t)\, e^{ikj\Delta x}$, where $k$ is the wave number, then

$$\frac{A_{j+1} - A_{j-1}}{2\Delta x} = ik\left(\frac{\sin k\Delta x}{k\Delta x}\right)\widehat{A}(t)\, e^{ikj\Delta x}. \tag{10.2}$$

Therefore, for one particular Fourier mode the advection equation becomes

$$\frac{d\widehat{A}}{dt} + ikc\left(\frac{\sin k\Delta x}{k\Delta x}\right)\widehat{A} = 0. \tag{10.3}$$

If we define $\omega \equiv -kc\frac{\sin k\Delta x}{k\Delta x}$, then (10.3) reduces to the oscillation equation, which was discussed at length in Chapter 4. The function $\sin\alpha/\alpha$ is sometimes written as $\mathrm{sinc}\alpha$. Note that $\mathrm{sinc}(k\Delta x) \to 1$ as $k\Delta x \to 0$. A plot is given in Fig. 10.1.

We are going study the properties of the various space- and time-differencing schemes for the advection scheme, making direct use of some of our results from Chapter 4. For particular space-differencing schemes, we will be able to obtain an explicit relationship between $\Delta x$ and $\Delta t$ as a condition for stability.

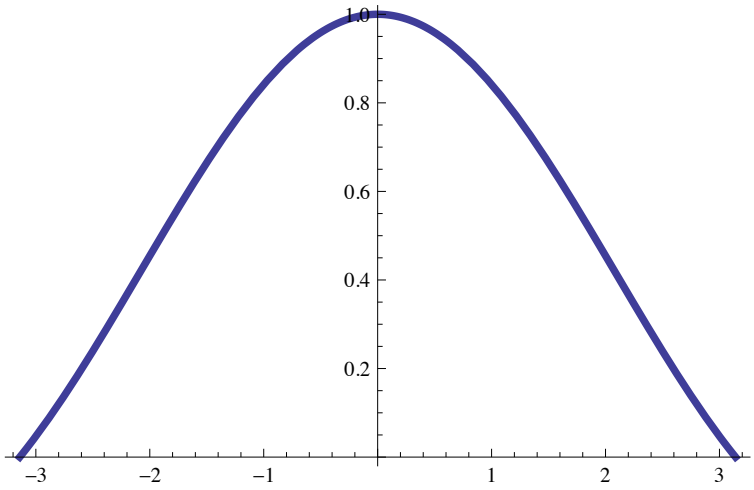



Figure 10.1: A plot of $\mathrm{sinc}\alpha$, for $-\pi \leq \alpha \leq \pi$.

As mentioned in Chapter 3, the Euler forward time scheme is unstable when combined with the centered space scheme for advection. You should prove this fact and remember it.

In the case of leapfrog time differencing and centered second-order space differencing, the one-dimensional advection equation is

$$\frac{{A_j}^{n+1} - {A_j}^{n-1}}{2\Delta t} + c\left(\frac{{A^n}_{j+1} - {A^n}_{j-1}}{2\Delta x}\right) = 0. \tag{10.4}$$

If we assume that ${A_j}^n$ has the wave-form solution for which Eq. (10.2) holds, then (10.4) can be written as

$$\widehat{A}^{n+1} - \widehat{A}^{n-1} = 2i\Omega\widehat{A}^n, \tag{10.5}$$

where

$$\Omega \equiv -kc\,\mathrm{sinc}\,(k\Delta x)\,\Delta t. \tag{10.6}$$

At this point, we recognize Eq. (10.5) as the leapfrog scheme for the oscillation equation. Recall from Chapter 4 that $|\Omega| \leq 1$ is necessary for (10.5) to be stable. Therefore, *we can simply re-use our result from Chapter 4*, i.e.,

$$|kc\text{sinc}\,(k\Delta x)\,\Delta t| = \left| c\frac{\sin(k\Delta x)}{\Delta x}\Delta t \right| \leq 1 \tag{10.7}$$

must hold for stability, for any and all $k$. Because $|\sin k\Delta x| \leq 1$, the "*worst case*" is $|\sin k\Delta x| = 1$. This occurs for $k\Delta x = \pm\frac{\pi}{2}$, which corresponds to the wavelength $L = 4\Delta x$. We conclude that

$$|c|\,\frac{\Delta t}{\Delta x} \leq 1 \tag{10.8}$$

is the necessary condition for stability. *Here "stability" means stability for all modes*. Eq. (10.8) is the famous "CFL" stability criterion associated with the names Courant, Friedrichs and Levy. The stability criterion (10.8) also applies to the upstream scheme, as we saw already in Chapter 3.

Note that the $2\Delta x$ wave is not the main problem here. It is the $4\Delta x$ wave that can most easily become unstable.

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 3:

$$\widehat{A}_1^n = \lambda_1^n \widehat{A}_1^0, \text{ and } \widehat{A}_2^n = \lambda_2^n \widehat{A}_2^0. \tag{10.9}$$

For $|\Omega| \leq 1$, we find, as discussed in Chapter 4, that

$$\lambda_1 = e^{i\theta}, \text{ and } \lambda_2 = e^{i(\pi-\theta)} = -e^{-i\theta}, \text{ where } \theta \equiv \tan^{-1}\left(\frac{\Omega}{\sqrt{1-\Omega^2}}\right). \tag{10.10}$$

Both modes are neutral. For the physical mode,

$$\begin{aligned}\left(A_j^n\right)_1 &= \lambda_1^n \widehat{A}_1^0 e^{ikj\Delta x} \\ &= \widehat{A}_1^0 \exp\left[ik\left(j\Delta x + \frac{\theta}{k\Delta t}n\Delta t\right)\right].\end{aligned} \tag{10.11}$$

Similarly, for the computational mode we obtain

$$\left(A_j{}^n\right)_2 = \widehat{A}_2^0(-1)^n \exp\left[ik\left(j\Delta x - \frac{\theta}{k\Delta t}n\Delta t\right)\right]. \tag{10.12}$$

Note the nasty factor of $(-1)^n$ in (10.12). It comes from the leading minus sign in (10.10). Comparing (10.11) and (10.12) with the expression $A(x,t) = \widehat{A}(0)\, e^{ik(x-ct)}$, 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. It is easy to see that as $(\Delta x, \Delta t) \to 0$, we obtain $\theta \to \Omega \to -kc\Delta t$. This means that the speed of the physical mode approaches $c$ (i.e., the right answer), while the speed of the computational mode approaches $-c$. *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 finite-difference approximation to the phase speed depends on $k$, while the true phase speed, $c$, is independent of $k$. *The shorter waves move more slowly than the longer waves*. The $2\Delta x$ does not move at all! The reason is easy to see in (10.4). The spurious dependence of phase speed on wave number with the finite-difference scheme is an example of *computational dispersion*, which will be discussed in detail later.

## 10.2 More about computational dispersion

Consider the differential-difference equation

$$\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0. \tag{10.13}$$

Using $A_j = \widehat{A}e^{ikj\Delta x}$, as before, we can write (10.13) as

$$\frac{d\widehat{A}_j}{dt} + cik\frac{\sin(k\Delta x)}{k\Delta x}\widehat{A}_j = 0. \tag{10.14}$$

If we had retained the differential form of the advection equation, we would have obtained $\frac{d\widehat{A}}{dt} + cik\widehat{A} = 0$. Comparison with Equation (10.14) shows that the phase speed is not simply $c$, but $c^*$, given by

$$c^* \equiv c\frac{\sin(k\Delta x)}{k\Delta x}. \tag{10.15}$$

Because $c^*$ depends on the wave number $k$, we have *computational dispersion* that arises from the space differencing. Note that the true phase speed, $c$, is independent of $k$. A plot of $c^*/c$ versus $k\Delta x$ is given by the upper curve in Fig. 10.2. (The second, lower curve in the figure, which illustrates the computational group velocity, will be discussed later.)

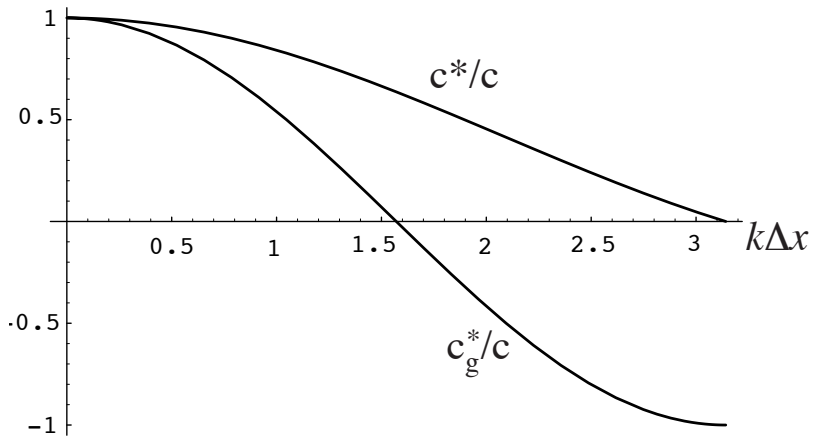



Figure 10.2: The ratio of the computational phase speed to the true phase speed, and also the ratio of the computational group speed to the true group speed, both plotted as functions of wave number.

If $k_s$ is defined by $k_s\Delta x \equiv \pi$, then $L_s \equiv \frac{2\pi}{k_s} = 2\Delta x$ is the smallest wave length that our grid can resolve. Therefore, we need only be concerned with $0 \leq k\Delta x \leq \pi$. Because $k_s\Delta x = \pi$, $c^* = 0$ for this wave, and so *the shortest possible wave is stationary!* This is actually obvious from the form of the space difference. Since $c^* < c$ for all $k > 0$, all waves move slower than they should according to the exact equation. Moreover, if we have a number of wave components superimposed on one another, each component moves with a different phase speed, depending on its wave number. The total "pattern" formed by the superimposed waves will break apart, as the waves separate from each other. This is called a *computational dispersion*.

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 $k_2$, respectively. Define

$$k \equiv \frac{k_1 + k_2}{2},\ c \equiv \frac{c_1 + c_2}{2},\ \Delta k \equiv \frac{k_1 - k_2}{2},\ \Delta(kc) \equiv \frac{k_1 c_1 - k_2 c_2}{2}. \tag{10.16}$$

See Fig. 10.3. Note that $k_1 = k + \Delta k$ and $k_2 = k - \Delta k$. Similarly, $c_1 = c + \Delta c$ and $c_2 = c - \Delta c$. You should be able to show that

$$k_1 c_1 \cong kc + \Delta(kc) \quad \text{and} \quad k_2 c_2 \cong kc - \Delta(kc). \tag{10.17}$$

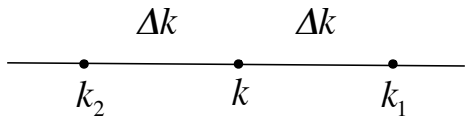



Figure 10.3: Sketch defining notation used in the discussion of the group velocity.

Here we neglect terms involving the product $\Delta k \Delta c$. This is acceptable when $k_1 \cong k_2$ and $c_1 \cong c_2$. Using (10.17), we can write the sum of two waves, each with unit amplitude, as

$$\begin{aligned}
&\exp\left[ik_1(x-c_1t)\right] + \exp\left[ik_2(x-c_2t)\right] \\
&\qquad \cong \exp\left(i\left\{(k+\Delta k)x - \left[kc+\Delta(kc)\right]t\right\}\right) + \exp\left(i\left\{(k-\Delta k)x - \left[kc-\Delta(kc)\right]t\right\}\right) \\
&\qquad = \exp\left[ik(x-ct)\right]\left(\exp\left\{i\left[(\Delta k)x - \Delta(kc)t\right]\right\} + \exp\left\{-i\left[(\Delta k)x - \Delta(kc)t\right]\right\}\right) \\
&\qquad = 2\cos\left[(\Delta k)x - \Delta(kc)t\right]\exp\left[ik(x-ct)\right] \\
&\qquad = 2\cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\}\exp\left[ik(x-ct)\right].
\end{aligned} \tag{10.18}$$

When $\Delta k$ is small, the factor $\cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\}$ behaves like the outer, slowly varying envelope in Fig. 10.4.

The envelope "modulates" wave $k$, which is represented by the inner, rapidly varying curve in the figure. The short waves move with phase speed $c$, but the "wave packets", i.e., the envelopes of the short waves, move with speed $\frac{\Delta(kc)}{\Delta k}$. The differential expression $\frac{d(kc)}{dk} \equiv c_g$ is called the "group velocity." Note that $c_g = c$ if $c$ does not depend on $k$. *For advection, the "right answer" is $c_g = c$, i.e., the group velocity and phase velocity 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 our finite-difference scheme, however, we have

$$c^*{}_g = \frac{d(kc^*)}{dk} = c\frac{d}{dk}\left(\frac{\sin k\Delta x}{\Delta x}\right) = c\cos k\Delta x. \tag{10.19}$$

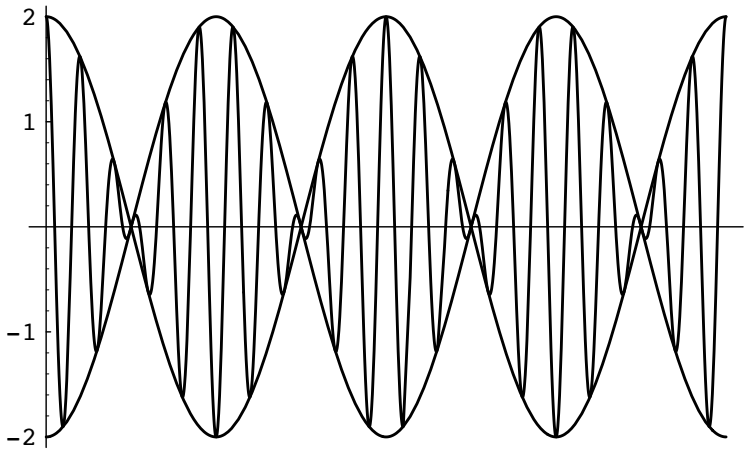



Figure 10.4: Sketch used to illustrate the concept of group velocity. The short waves are modulated by longer waves.

A plot of ${c^*}_g$ versus $k\Delta x$ is given in Fig. 10.2. Note that ${c^*}_g = 0$ for the $4\Delta x$ wave, and is negative 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 $L = 2\Delta x$, ${c^*}_g$ actually approaches $-c$, when in reality it should be equal to $c$ for all wavelengths. For all waves, ${c^*}_g < c^* < c = c_g$. This problem arises from the space differencing; it has nothing to do with time differencing.

Fig. 10.5, which is a modified version of Fig. 10.4, illustrates this phenomenon 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 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 ${c^*}_g$) by $B_j$, we see that

$$B_j = (-1)^j A_j. \tag{10.20}$$

(This is true only for the particular case $L = 2\Delta x$.) Using (10.20), Eq. (10.13) can be rewritten as

$$\frac{dB_j}{dt} + (-c)\left(\frac{B_{j+1} - B_{j-1}}{2\Delta x}\right) = 0. \tag{10.21}$$

Eq. (10.21) shows that the upper solid curve will move with speed $-c$.

Recall that when we introduce time differencing, the computed phase change per time

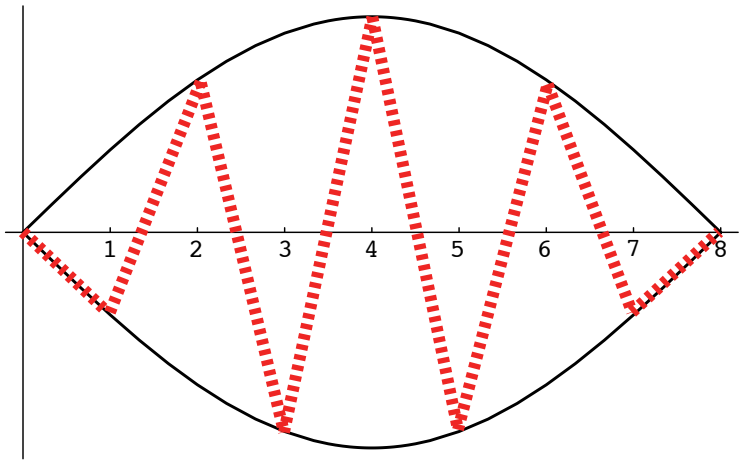



Figure 10.5: Yet another sketch used to illustrate the concept of group velocity. The short wave has wavelength $L = 2\Delta x$.

step is generally not equal to $-kc\Delta t$. This leads to changes in $c^*$ and ${c^*}_g$, although the formulas discussed above remain valid for $\Delta t \to 0$.

We now present an *analytical* solution of (10.13), which illustrates dispersion error in a very clear way, following an analysis by Matsuno (1966). If we write (10.13) in the form

$$2\frac{dA_j}{d\left(\frac{tc}{\Delta x}\right)} = A_{j-1} - A_{j+1}, \tag{10.22}$$

and define a non-dimensional time $\tau \equiv \frac{tc}{\Delta x}$, we obtain

$$2\frac{dA_j}{d\tau} = A_{j-1} - A_{j+1}. \tag{10.23}$$

I don't expect your to know this, but (10.23) happens to have the same form as a recursion formula satisfied by the Bessel functions of the first kind of order $j$, which are usually denoted by $J_j(\tau)$. You can Google it. The $J_j(\tau)$ have the property that $J_0(0) = 1$, and $J_j(0) = 0$ for $j \neq 0$. Because the $J_j(\tau)$ satisfy (10.23), each $J_j(\tau)$ represents the solution at one particular grid point, $j$, as a function of the nondimensional time, $\tau$.

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 (10.23) for the points $j = 0$, 1, and 2 is illustrated in Fig. 10.6.

By using the identity

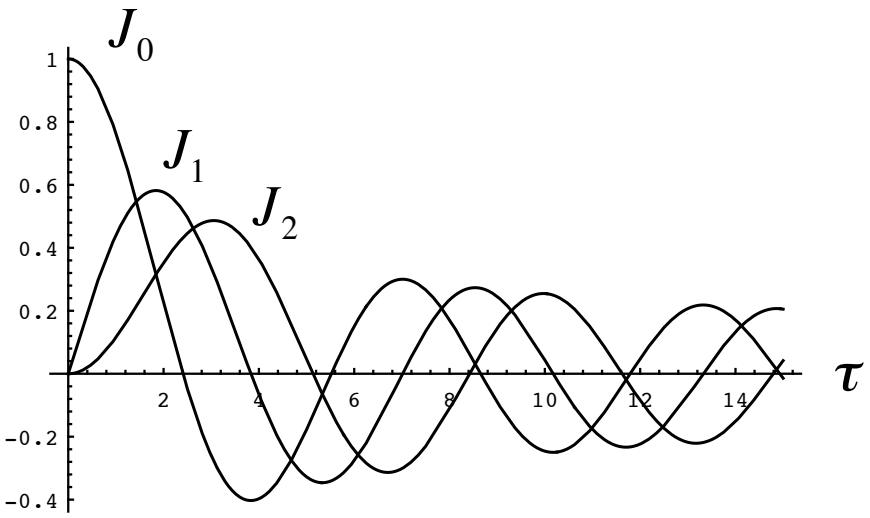



Figure 10.6: The time evolution of the solution of (10.23) at grid points $j = 0$, 1, and 2.

$$J_{(-j)} = (-1)^j J_j, \tag{10.24}$$

we can obtain the solution at 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 issues associated with time differencing.

Fig. 10.7 shows the solution of (10.23) 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. 10.2 and Fig. 10.5, is seen directly here. The figure also shows that $c_g$ is negative for the shortest wave.

A similar type of solution is shown in Fig. 10.8, which is taken from a paper by Wurtele (1961). Here the initial conditions are slightly different, namely,

$$A_{-1} = A_0 = A_1 = 1, \text{ and } A_j = 0 \text{ for } j \leq -2, j \geq 2. \tag{10.25}$$

This is a "top hat" or "box" initial condition. We can construct the initial condition by combining

$$J_{j-1}(0) = 1 \text{ for } j = 1 \text{ and zero elsewhere,} \tag{10.26}$$

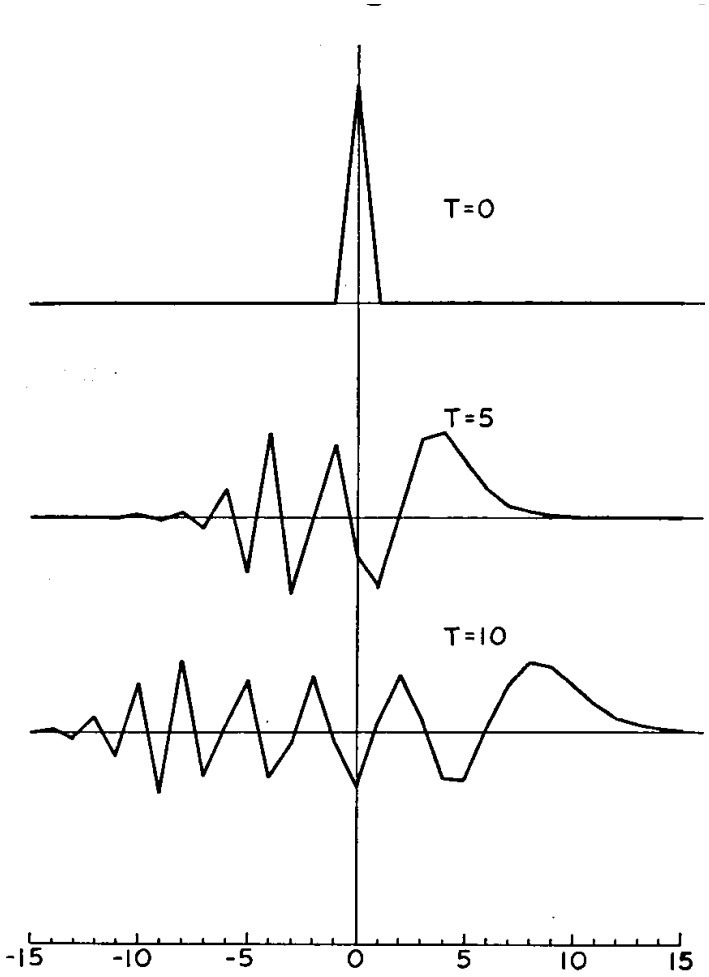



Figure 10.7: The solution of $(10.23)$ for $\tau = 5$ and $\tau = 10$ for $-15 \leq j \leq 15$, with "spike" initial conditions. From Matsuno (1966).

$$J_j(0) = 1 \text{ for } j = 0 \text{ and zero elsewhere, and} \tag{10.27}$$

$$J_{j+1}(0) = 1 \text{ for } j = -1 \text{ and zero elsewhere,} \tag{10.28}$$

so that the full solution is given by

$$A_j(\tau) = J_{j-1}(\tau) + J_j(\tau) + J_{j+1}(\tau). \tag{10.29}$$

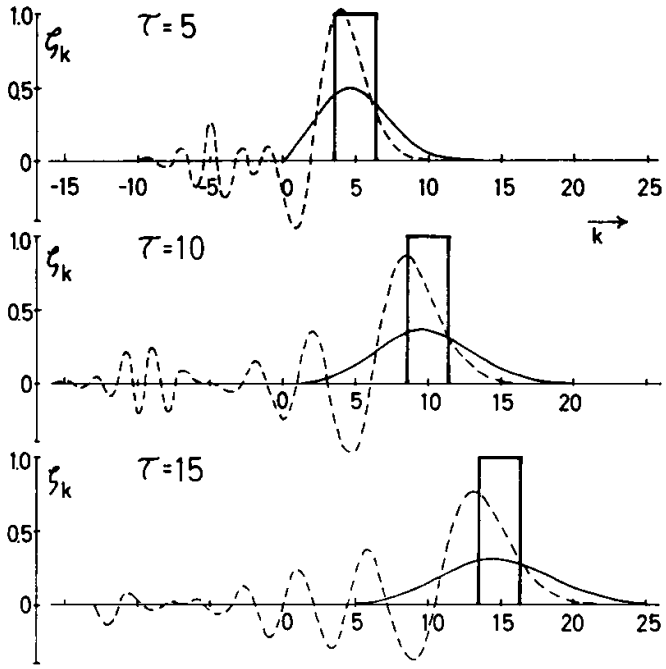



Fig. 1. Three solutions of the advection equation for (non-dimensional) times $\tau = 5, 10, 15$.
——— (exact) solution of continuous equation (4)
— — — — solution (8) of centered differential-difference equation
——— solution (11) of backward differential-difference equation
For typical meteorological values, ten units of non-dimensional time correspond to about 42 hours.

Figure 10.8: The solution of (10.20) with "box" initial conditions. From Wurtele (1961).

Dispersion is evident again in Fig. 10.8. The dashed curve is for centered space differencing, and the solid curve is for the upstream scheme. (The solution for the upstream case is given in terms of the Poisson distribution rather than Bessel functions; see Wurtele's paper for details.) The principal disturbance moves to the right, but the short-wave components move to the left.

*Do not confuse computational dispersion with instability.* Both dispersion and instability can lead to "noise," but a noisy solution is not 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.

## 10.3 The effects of fourth-order space differencing on the phase speed

As discussed in Chapter 2, the fourth-order difference quotient takes the form

$$\left(\frac{\partial A}{\partial x}\right)_j = \frac{4}{3}\left(\frac{A_{j+1}-A_{j-1}}{2\Delta x}\right) - \frac{1}{3}\left(\frac{A_{j+2}-A_{j-2}}{4\Delta x}\right) + O\left[(\Delta x)^4\right]. \tag{10.30}$$

Recall that in our previous discussion concerning the second-order scheme, we derived an expression for the phase speed of the numerical solution given by

$$c^* = c\left(\frac{\sin k\Delta x}{k\Delta x}\right). \tag{10.31}$$

For this fourth-order scheme, the corresponding expression for the phase speed is

$$c^* = c\left(\frac{4}{3}\frac{\sin k\Delta x}{k\Delta x} - \frac{1}{3}\frac{\sin 2k\Delta x}{2k\Delta x}\right). \tag{10.32}$$

Fig. 10.9 shows a graph of $c^*/c$ versus $k\Delta x$ for each scheme. The fourth-order scheme gives a considerable improvement in the accuracy of the phase speed, for long waves. There is no improvement for wavelengths close to $L = 2\Delta x$, however, and the problems that we have discussed in connection with the second-order schemes become more complicated with the fourth-order scheme. This illustrates that increasing the order of accuracy does not help with the errors of the shortest waves.

## 10.4 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. 10.10. The right answer, for $c > 0$, is that the domain of dependence is region II only. For $c < 0$ the correct domain of dependence should be region I only.

The "upstream scheme," given by

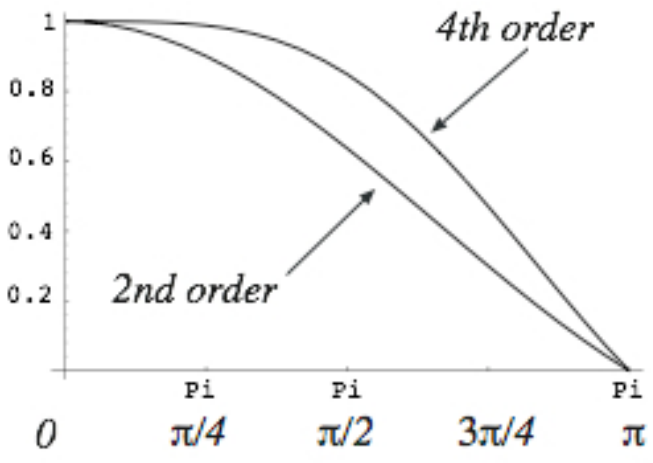



Figure 10.9: The ratio of the computational phase speed, $c^*$, to the true phase speed, $c$, plotted as a function of $k\Delta x$, for the second-order and fourth-order schemes.

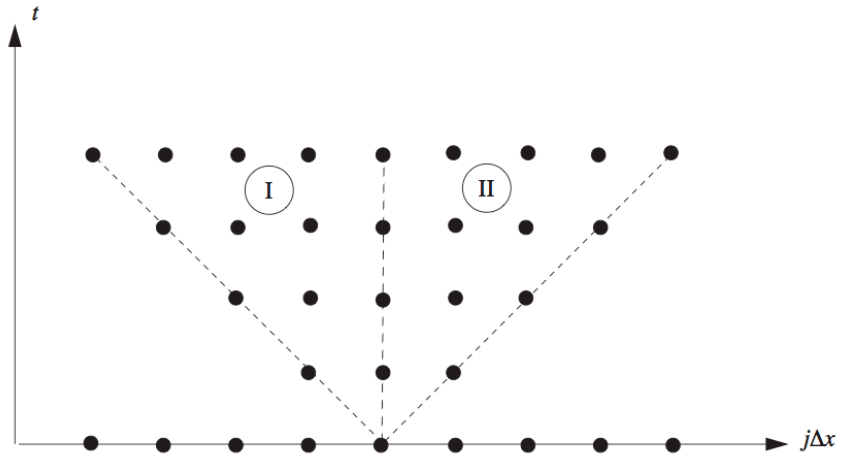



Figure 10.10: The domain of influence for explicit non-iterative space-centered schemes expands in time, as is shown by the union of Regions I and II.

$$\frac{A_j^{n+1}-A_j^n}{\Delta t}+c\left(\frac{A_j^n-A_{j-1}^n}{\Delta x}\right)=0 \quad \text{for} \quad \text{c}>0, \tag{10.33}$$

or

$$\frac{A_j^{n+1}-A_j^n}{\Delta t}+c\left(\frac{A_{j+1}^n-A_j^n}{\Delta x}\right)=0 \quad \text{for} \quad \text{c}<0, \tag{10.34}$$

is the simplest example of a space-uncentered scheme. As shown earlier, we can write (10.33) as

$$A_j^{n+1} = (1-\mu)A_j^n + \mu A_{j-1}^n, \tag{10.35}$$

which has the form of an interpolation, and (10.34) can be written in a similar way. Obviously, for the upstream scheme, Region II alone is the domain of influence when $c > 0$, and Region I alone is the domain of influence when $c < 0$. This is good. The scheme produces strong damping, however, as shown in Fig. 10.8. The damping results from the linear 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.

Takacs (1985) proposed a forward-in-time space-uncentered advection scheme of the form

$$A_j^{n+1} = a_1 A_{j+1}^n + a_0 A_j^n + a_{-1} A_{j-1}^n + a_{-2} A_{j-2}^n. \tag{10.36}$$

Here we assume that the wind is spatially uniform and blowing towards larger values of $j$, so that points $j-1$ and $j-2$ are in the upstream direction. The scheme given by (10.36) has four undetermined coefficients. Recall from Chapter 2 that with four coefficients we can, in general, achieve third-order accuracy. Takacs (1985) took a different approach, however. Requiring only second-order accuracy, he found that

$$a_1 = \mu(\mu-1)/2 - a_{-2}, \tag{10.37}$$

$$a_0 = 1 - \mu^2 + 3a_{-2}, \tag{10.38}$$

$$a_{-1} = \mu(\mu+1)/2 - 3a_{-2}, \tag{10.39}$$

and

$$a_{-2} = \alpha\mu\left(\mu - 1\right), \tag{10.40}$$

where $\mu$ is the usual CFL parameter, and $\alpha$ is a free (i.e., as yet unchosen) parameter that is discussed below. The value $a_{-2}$, the "upstream" coefficient, is directly proportional to $\alpha$. For $\alpha = 0$ the Takacs scheme reduces to the second-order Lax-Wendroff scheme discussed earlier in this chapter, which uses only three grid-points and so can be described as spatially centered. Takacs showed that with $\alpha = \left(\mu + 1\right)/6$ the scheme has third-order accuracy. He demonstrated that the value of $\alpha$ has little effect on the amplitude error of the scheme, but strongly affects the dispersion error, which is the largest part of the total error. Takacs's explanation for this interesting fact is explained below. He concluded that the best choice for $\alpha$ is the one that gives third-order accuracy.

Takacs (1985) made some important and general points 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.

In Chapter 4, we defined the relative error, $\varepsilon$, by

$$\lambda = \left(1 + \varepsilon\right)\lambda_T, \tag{10.41}$$

where

$$\lambda_T = e^{-i\mu k\Delta x} \tag{10.42}$$

is the amplification factor for the solution to the differential equation. Comparing (10.41) with (7.14), and using (10.42), we find that the real part of the relative error is given by

$$\varepsilon_R = \sum_{j'=-\infty}^{\infty} a_k \cos\left[\left(j' + \mu\right)k\Delta x\right] e^{ikj\Delta x} - 1. \tag{10.43}$$

Similarly, the imaginary part of the relative error satisfies

$$\varepsilon_I = \sum_{j'=-\infty}^{\infty} a_k \sin\left[\left(j' + \mu\right)k\Delta x\right] e^{ikj\Delta x}. \tag{10.44}$$

Now comes the key step. We want to understand how $\varepsilon_R$ and $\varepsilon_I$ depend on $k\Delta x$. To do this, we expand (10.43) and (10.44) as Taylor series in powers of $k\Delta x$. Because $\varepsilon_R$ involves the cosine of $(j+\mu)\,k\Delta x$, its expansion involves only even powers of $k\Delta x$:

$$\varepsilon_R = -\frac{(k\Delta x)^2}{2!}\sum_{j'=-\infty}^{\infty}\left(j'+\mu\right)^2 a_{j'} + \frac{(k\Delta x)^4}{4!}\sum_{j'=-\infty}^{\infty}\left(j'+\mu\right)^4 a_{j'} + \ldots. \tag{10.45}$$

Here we have used (7.8). Similarly, because $\varepsilon_I$ involves the sine of $(j+\mu)\,k\Delta x$, its expansion involves only odd powers of $k\Delta x$:

$$\varepsilon_I = (k\Delta x)\sum_{j'=-\infty}^{\infty}\left(j'+\mu\right) a_{j'} + \frac{(k\Delta x)^3}{3!}\sum_{j'=-\infty}^{\infty}\left(j'+\mu\right)^3 a_{j'} + \ldots. \tag{10.46}$$

What do these two results mean? Expressions of the form $\sum_{j'=-\infty}^{\infty}(j'+\mu)^l a_{j'}$ appear repeatedly in (10.45) and (10.46). According to (7.12), for a scheme of $m$th-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 decrease, as more and more terms of (10.45) and (10.46) drop out. *But the real and imaginary parts take turns*. Only even-order schemes will reduce $\varepsilon_R$ (relative to the next-lower odd-order scheme), and only odd-order schemes will reduce $\varepsilon_I$ (relative to the next-lower even-order scheme).

As an example, suppose that we have a first-order scheme, like the upstream scheme. Then the leading term of $\varepsilon_R$ is second-order, but the leading term of $\varepsilon_I$ is third-order. When we go to a second-order scheme, the leading term of $\varepsilon_I$ becomes third-order, but the leading term of $\varepsilon_R$ stays the same. If we then go to a third-order scheme, the leading term of t$\varepsilon_R$ will stay the same, but the leading term of $\varepsilon_I$ will decrease.

Now recall from Chapter 4 that the amplitude error is proportional to $\varepsilon_R$, while the phase error is proportional to $\varepsilon_I$. We conclude that the amplitude error decreases as we go from and odd order to the next even order, while the phase error decreases as we go from and even order to the next odd order. This is why odd-order schemes have smaller phase errors than even-order schemes.

## 10.5 Sign-preserving and monotone schemes

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, since negative values that arise through truncation errors 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 (10.35) guarantees that

$$\text{Min}\left\{A_j^n, A_{j-1}^n\right\} \leq A_j^{n+1} \leq \text{Max}\left\{A_j^n, A_{j-1}^n\right\}. \tag{10.47}$$

This means that $A_j^{n+1}$ cannot be smaller than the smallest value of $A_j^n$, or larger than the largest value of $A_j^n$. The finite-difference advection process associated with the upstream scheme cannot produce any new maxima or minima. As discussed in the introduction to this chapter, real advection also has this property.

In particular, 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. A monotone scheme is one that cannot produce new maxima or minima, like those associated with the dispersive ripples seen in Fig. 10.8. The property of monotonicity is expressed by (10.47).

All monotone schemes are sign-preserving schemes. The converse is not true.

Sign preservation makes sense for both the continuity equation and the tracer equation. This is not true of monotonicity. Recall from (5.4) that the density of a compressible fluid is not constrained to be constant following a particle. *For this reason, monotonicity is physically wrong for the continuity equation, although it is physically right for the tracer equation*.

As discussed by Smolarkiewicz (1991), sign-preserving schemes tend to be stable. To see why, suppose that we have a linearly conservative scheme for a variable $A$, such that

$$\sum_j A_j^n = \sum_j A_j^0, \tag{10.48}$$

where the sum represents a sum over the whole spatial domain, the superscripts $n > 0$ and 0 represent two time levels. For simplicity we use only a single spatial dimension here, but the following argument holds for any number of dimensions. Suppose that the scheme that takes us from $A_j^0$ to $A_j^n$ through $n$ time steps is sign-preserving and conserves $A$, as in (10.48). If $A_j^0$ is of one sign everywhere, it follows trivially from (10.48) that

$$\sum_j |A_j^n| = \sum_j |A_j^0|. \tag{10.49}$$

We can also show that

$$\sum_j (A_j)^2 \leq \left( \sum_j |A_j| \right)^2, \tag{10.50}$$

Then from (10.49) and (10.50) we see that

$$\sum_j (A_j^n)^2 \leq \left( \sum_j |A_j^0| \right)^2. \tag{10.51}$$

The quantity on the right-hand side of (10.51) is a property of the initial condition, so it is independent of time. Eq. (10.51) therefore demonstrates that $\left(A_j^n\right)^2$ is bounded for all time, and so it proves stability by the energy method discussed in Chapter 2. The essence of (10.51) is that there is an upper bound on $\sum_j \left(A_j^n\right)^2$. The bound is very weak, however; try some numbers to see for yourself. So, although (10.51) does demonstrate absolute stability, it does not ensure good behavior!

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.,

$$A = A^+ + A^-. \tag{10.52}$$

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 (10.52). 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.

Although the upstream scheme is sign-preserving, it is only first-order accurate and strongly damps, as we have seen. Can we find 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 (9.10) with either the geometric mean given by (9.28), or the harmonic mean given by (9.29). 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 minimize it.

## 10.6 Hole filling

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 have insufficient 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" is a particularly simple and computationally fast algorithm. The first step is to add up all of the positive water on the mesh:

$$P \equiv \sum_{\text{where} A_j \geq 0} m_j A_j \geq 0. \tag{10.53}$$

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

$$T \equiv \sum_{\text{allpoints}} m_j A_j. \tag{10.54}$$

Both $T$ and $P$ have the dimensions of mass. Define the nondimensional ratio

$$\Phi \equiv \frac{T}{P} \leq 1; \tag{10.55}$$

normally $\Phi$ 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

$$A_j^{new} = \Phi A_j. \tag{10.56}$$

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 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.

Note that the algorithm 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, and 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.

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.

## 10.7 Flux-corrected transport

The upstream scheme is monotone and sign-preserving, but, unfortunately, as we have seen, it is strongly damping. Damping is in fact characteristic of 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

$$A_j^{n+1} = A_j^n - \left( FH_{j+\frac{1}{2}} - FH_{j-\frac{1}{2}} \right). \tag{10.57}$$

Here $FH$ represents the "high-order" fluxes associated with the scheme. Note that (10.57) 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.") We can write

$$FH_{j+\frac{1}{2}} \equiv FL_{j+\frac{1}{2}} + FC_{j+\frac{1}{2}}. \tag{10.58}$$

Here $FC_{j+\frac{1}{2}}$ is a "corrective" flux, sometimes called an "anti-diffusive" flux. Eq. (10.58) is essentially the definition of $FC_{j+\frac{1}{2}}$. According to (10.58), 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.,

$$A_j^{n+1*} = A_j^n - \left( FL_{j+\frac{1}{2}} - FL_{j-\frac{1}{2}} \right). \tag{10.59}$$

Since, by assumption, the low-order scheme is monotone, we know that

$$A_j^{MAX} \geq A_j^{n+1*} \geq A_j^{MIN} \text{ for all } j. \tag{10.60}$$

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 (10.60). 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 (10.57), (10.58), and Equation (10.59) it is easy to see that

$$A_j^{n+1} = A_j^{n+1*} - \left( FC_{j+\frac{1}{2}} - FC_{j-\frac{1}{2}} \right). \tag{10.61}$$

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

$$\widehat{FC}_{j+\frac{1}{2}} \equiv C_{j+\frac{1}{2}} FC_{j+\frac{1}{2}}. \tag{10.62}$$

In place of (10.61), we use

$$A_j^{n+1} = A_j^{n+1*} - \left( \widehat{FC}_{j+\frac{1}{2}} - \widehat{FC}_{j-\frac{1}{2}} \right). \tag{10.63}$$

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 (10.63) will reduce to (10.61) 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 (10.63) will simply give the low-order solution. We enforce

$$0 \leq C_{j+\frac{1}{2}} \leq 1 \text{ for all } j, \tag{10.64}$$

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

$$A_j^{MAX} \geq A_j^{n+1} \geq A_j^{MIN} \text{ for all } j \tag{10.65}$$

be satisfied. Compare (10.65) with (10.60). We can always ensure that (10.65) will be satisfied by taking $C_{j+\frac{1}{2}} = 0$; this is the "worst case." Quite often it may happen that (10.65) 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 $A_j^{n+1}$.

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 (10.65) 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 $C$s for each cell wall. To ensure that all constraints are simultaneously satisfied, we choose the smaller of the two provisional $C$s for each cell wall. *Voila!* All constraints are satisfied.

The procedure outlined above is not unique. Other approaches are described in the literature.

It remains to choose the upper and lower bounds $A_j^{MAX}$ and $A_j^{MIN}$ that appear in (10.65) and (10.60). 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

$$A_j^{MAX} = Max\left\{A_{j-1}^n, A_j^n, A_{j+1}^n, A_{j-1}^{n+1*}, A_j^{n+1*}, A_{j+1}^{n+1*}\right\}, \tag{10.66}$$

and

$$A_j^{MIN} = Min\left\{A_{j-1}^n, A_j^n, A_{j+1}^n, A_{j-1}^{n+1*}, A_j^{n+1*}, A_{j+1}^{n+1*}\right\}. \tag{10.67}$$

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.

## 10.8 A survey of some advection schemes that you might run into out there

In closing this long chapter, we list here some advection schemes that are well known and widely used. The purpose of the list is to help you find information when you need it.

## 10.9 Summary

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.

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.

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

| Name of Scheme | Description | Properties | Reference |
|---|---|---|---|
| Upstream | | | |
| Lax-Wendroff | | | |
| Lin-Rood | | | |
| MPDATA | | | |
| Piecewise Parabolic | | | |
| Prather | | | |
| ULTIMATE | | | |
| | | | |

even negative, when it should be positive. Short waves generally move slower than longer waves. The phase 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. Space-uncentered schemes are well suited to advection, which is a spatially asymmetric process, and they can minimize the effects of computational dispersion.

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.

# Chapter 11

# Lagrangian and semi-Lagrangian advection schemes

## 11.1 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 physics (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" is put in quotation marks because of course the result is actually approximate in the sense that, in practice, only a finite number of particles can be tracked.
- As a consequence of the first point mentioned above, Lagrangian schemes are monotone and sign-preserving.
- Time steps can be very long without triggering computational instability, although of course long time steps still lead to large truncation errors.
- Aliasing instability does not occur with Lagrangian schemes. Aliasing instability will be discussed later.

Alas, there are always trade-offs. Lagrangian schemes suffer from a number of practical difficulties. Some of these have to do with "voids" that develop, i.e., regions with few particles. 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 from the equation of motion) on the basis of a collection of particles that can travel literally anywhere within the domain, in an uncontrolled way.

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 approach is to invent a way to compute a given field at all points through-

out the domain, given the value of the field at a finite number of arbitrarily located points that happen to be occupied by particles. The point is that once the field has been defined everywhere, we can differentiate it.

For an arbitrary field $A$, let

$$A(\mathbf{r}) = \int A(\mathbf{r}') W(\mathbf{r}-\mathbf{r}', h)\, d\mathbf{r}', \tag{11.1}$$

where the integration is over the whole domain (e.g., the whole atmosphere), and $W(\mathbf{r}-\mathbf{r}', h)$ is an interpolating "kernel" such that

$$\int W(\mathbf{r}-\mathbf{r}', h)\, d\mathbf{r}' = 1 \tag{11.2}$$

and

$$\lim_{h\to 0} W(\mathbf{r}-\mathbf{r}', h) = \delta(\mathbf{r}-\mathbf{r}'), \tag{11.3}$$

where $\delta(\mathbf{r}-\mathbf{r}')$ is the Dirac delta function. In (11.1) – (11.3), $h$ is a parameter, which is a measure of the "width" of $W$. We can interpret $W\delta(\mathbf{r}-\mathbf{r}', h)$ as a "weighting function" that is strongly peaked at $(\mathbf{r}-\mathbf{r}') = 0$. For example, we might use the Gaussian weighting function given by

$$W(\mathbf{r}-\mathbf{r}', h) = \frac{e^{-\left[x(\mathbf{r}-\mathbf{r}')^2/h^2\right]}}{h\sqrt{\pi}}, \tag{11.4}$$

which can be shown to satisfy (11.3). With these definitions, if $A(\mathbf{r}')$ is a constant field, then $A(\mathbf{r})$ is given by the same constant. For $h \to 0$, (11.1) gives $A(\mathbf{r}) = A(\mathbf{r}')$ when $\mathbf{r} = \mathbf{r}'$.

In a discrete model, we replace (11.1) by

$$A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} W(\mathbf{r}-\mathbf{r}_b, h). \tag{11.5}$$

Here the index $b$ denotes a particular particle, $m_b$ is the mass of the particle, and $\rho_b$ is the density of the particle. To see what is going on in (11.5), consider the case $A \equiv \rho$. Then (11.5) reduces to

$$\rho(\mathbf{r}) = \sum_b m_b W(\mathbf{r} - \mathbf{r}_b, h), \tag{11.6}$$

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 – which makes sense.

We can now perform spatial differentiation simply by taking the appropriate derivatives of $W(\mathbf{r} - \mathbf{r}_b, h)$, e.g.,

$$\nabla A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} \nabla W(\mathbf{r} - \mathbf{r}_b, h). \tag{11.7}$$

This follows because $m_b$, $A_b$ and $\rho_b$ 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.

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 also be parameterized in a straightforward way. The first applications of the method were to lakes and then 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 parcels can move for hundreds of years in the deep ocean circulation, with literally no computational diffusion. 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 now been coupled, and used to simulate climate change (P. Haertel, personnal communication, 2019). This fully Lagrangian global modeling system is truly unique, and it will be interesting to see how it evolves from here.

## 11.2 Semi-Lagrangian schemes

"Semi-Lagrangian" schemes (e.g., Robert et al. (1985); Staniforth and Côté (1991); Bates et al. (1993)) 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, $c$. A particle that resides at $x = x_j$ at time level $t = t^{n+1}$ has a departure point given by

$$\left(x_{departure}\right)^n{}_j = x_j - c\Delta t. \tag{11.8}$$

Here the superscript $n$ is used to indicate that the departure point is the location of the particle at time level $n$. Suppose that $c > 0$, and that the departure point is less than one $\Delta x$ away from $x_j$, so that

$$x_{j-1} < \left(x_{departure}\right)^n{}_j < x_j. \tag{11.9}$$

Then the simplest linear interpolation for $A$ at the departure point is

$$\begin{aligned}\left(A_{departure}\right)_j^n &= A_{j-1}^n + \left[\frac{\left(x_{departure}\right)_j^n - x_{j-1}}{\Delta x}\right]\left(A_j^n - A_{j-1}^n\right) \\ &= A_{j-1}^n + \left(\frac{\Delta x - c\Delta t}{\Delta x}\right)\left(A_j^n - A_{j-1}^n\right) \\ &= A_{j-1}^n + (1-\mu)\left(A_j^n - A_{j-1}^n\right) \\ &= \mu A_{j-1}^n + (1-\mu)A_j^n.\end{aligned} \tag{11.10}$$

Here we assume for simplicity that the mesh is spatially uniform, and $\mu \equiv c\Delta t/\Delta x$, as usual. The semi-Lagrangian scheme uses

$$A_j^{n+1} = \left(A_{departure}\right)_j^n, \tag{11.11}$$

so we find that

$$A_j^{n+1} = \mu A_{j-1}^n + (1-\mu)A_j^n. \tag{11.12}$$

This is the familiar upstream scheme. Note that (11.9), which was used in setting up the spatial interpolation, is equivalent to

$$0 < \mu < 1. \tag{11.13}$$

As shown earlier, this is the condition for stability of the upstream scheme.

What if (11.9) 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,

$$x_{j-a} < \left(x_{departure}\right)_j^n < x_{j-a+1}. \tag{11.14}$$

where $a$ is an *integer* greater than 1. For this case, we find in place of (11.10) that

$$\left(A_{departure}\right)_j^n = \hat{\mu} A_{j-a}^n + (1-\hat{\mu}) A_{j-a+1}^n. \tag{11.15}$$

where

$$\hat{\mu} \equiv 1 - a + \mu. \tag{11.16}$$

Notice that we have assumed again here, for simplicity, that both the mesh and the advecting current are spatially uniform. It should be clear that

$$0 < \hat{\mu} < 1. \tag{11.17}$$

For $a = 1$, $\mu = \hat{\mu}$. Eq. (11.11) gives

$$A_j^{n+1} = \hat{\mu} A_{j-a}^n + (1-\hat{\mu}) A_{j-a+1}^n. \tag{11.18}$$

*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," and we know that the upstream scheme is very diffusive. By using higher-order interpolations (e.g., cubic interpolations), the strength of this computational diffusion can be reduced, but it cannot be eliminated completely.

Is the semi-Langrangian scheme conservative? To prove that the scheme is conservative, it would suffice to show that it can be written in a "flux form." Note, however, that in deriving the 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." An example is CSLAM (Lauritzen et al., 2010; Dubey et al., 2014; Lauritzen et al., 2017), which stands for "Conservative Semi-Lagrangian Multitracer."

# Chapter 12

# Just relax

## 12.1 Introduction

Systems of linear equations frequently arise in atmospheric science. The systems can involve thousands or even millions of unknowns, which must be solved for simultaneously. They can be solved by a wide variety of methods, which are discussed in standard texts on numerical analysis. The solution of linear systems is conceptually simple, but may nevertheless present challenges in practice. *The main issue is how 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. For the problems that arise in atmospheric science, and considering the characteristics of modern computers, minimizing computational work is often more of a concern than minimizing storage.

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. Two-dimensional linear boundary-value problems occur quite often. Here is an example: Consider a two-dimensional flow. Let $\zeta$ and $\delta$ be the vorticity and divergence, respectively. We can define a stream function, $\psi$, and a velocity potential, $\chi$, by

$$\mathbf{V}_r = \mathbf{k} \times \nabla \psi, \tag{12.1}$$

and

$$\mathbf{V}_d = \nabla \chi, \tag{12.2}$$

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

$$\mathbf{V} = \mathbf{V}_r + \mathbf{V}_d. \tag{12.3}$$

The vorticity and divergence then satisfy

$$\zeta = \nabla^2 \psi. \tag{12.4}$$

and

$$\delta = \nabla^2 \chi, \tag{12.5}$$

respectively. Suppose that we are given the distributions of $\zeta$ and $\delta$, and want to determine the wind vector. This can be done by first solving the two boundary-value problems represented by (12.4) - (12.5), with suitable boundary conditions, then using (12.1) - (12.2) to obtain $\mathbf{V}_r$ and $\mathbf{V}_d$, and finally using Eq. (12.3) 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 (see the next Chapter) or the shallow -water equations.

## 12.2 Solution of one-dimensional boundary-value problems

As a simple one-dimensional example, consider

$$\frac{d^2 q(x)}{dx^2} = f(x), \tag{12.6}$$

on a periodic domain. Here $f(x)$ is a *given* periodic function of $x$. Solution of (12.6) requires two boundary conditions. One of these can be the condition of periodicity, which we have already specified. We assume that a second boundary condition is also given. For example, the average of $q$ over the domain may be prescribed.

The exact solution of (12.6) can be obtained by expanding $q(x)$ and $f(x)$ in an infinite Fourier series. The individual Fourier modes will satisfy

$$-k^2\widehat{q}_k = \hat{f}_k, \tag{12.7}$$

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 (12.6) 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 called the "spectral" method. Like everything else, the spectral method has both strengths and weaknesses. It will be discussed in a later chapter.

Suppose, however, that the problem posed by (12.6) 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 finite differences. In that case, the need for consistency with the other equations of the model may dictate that the spatial derivatives in (12.6) be approximated by a finite-difference method, such as

$$\frac{q_{i+1} - 2q_i + q_{i-1}}{d^2} = f_i. \tag{12.8}$$

Here $d$ is the grid spacing. We have used centered second-order spatial differences in (12.8). Assuming a periodic, wave-like solution for $q_i$, and correspondingly expanding $f_i$, we obtain, in the usual way,

$$-k^2\widehat{q}_k \left( \frac{\sin \frac{kd}{2}}{\frac{kd}{2}} \right)^2 = \hat{f}_k. \tag{12.9}$$

Note the similarity between (12.9) and (12.7). Clearly (12.9) 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* (12.8). This example illustrates that Fourier solution methods can be used even in combination with finite-difference approximations. For each $k$, the factor of $-k^2\left(\frac{\sin kd/2}{kd/2}\right)^2$

in (12.9) can be evaluated once and stored for use later in the simulation. This is advantageous if (12.9) must be solved on each of many time steps. Fourier methods may or may not be applicable, depending on the geometry of the domain.

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 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.

There are other ways to solve (12.8). 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 x 6 problem shown below:

$$\begin{bmatrix} d_1 & a_2 & 0 & 0 & 0 & b_6 \\ b_1 & d_2 & a_3 & 0 & 0 & 0 \\ 0 & b_2 & d_3 & a_4 & 0 & 0 \\ 0 & 0 & b_3 & d_4 & a_5 & 0 \\ 0 & 0 & 0 & b_4 & d_5 & a_6 \\ a_1 & 0 & 0 & 0 & b_5 & d_6 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{bmatrix}. \tag{12.10}$$

Here each element of the 6 x 6 matrix is labeled with a single subscript, indicating 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.

The solution of tri-diagonal linear systems is very fast and easy. For instance, the first of the six equations represented by (12.10) can be solved for $q_1$ as a function of $q_2$, and $q_6$, provided that $d_1 \neq 0$. This solution 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, we can obtain the other unknowns 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.

The issue of "scaling" deals with the change in the amount of work required as the problem size increases. The best we can hope for is that the amount of work needed to solve the system is simply proportional to the number of unknowns. It should be clear that this is the case for the tri-diagonal solver described above. In other words, 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 are very inefficient (i.e., 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.

Finally, we could solve (12.8) 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), Southwell (1946); Allen (1954)], and are now very widely used (e.g., Strang (2007)). Several relaxation methods are discussed below.

## 12.3 Jacobi relaxation

Starting from this point, most of the discussion in this chapter is a condensed version of that found in the paper by Fulton et al. (1986).

Consider the boundary-value problem

$$\begin{aligned} -\nabla^2 q = f \text{ on a two-dimensional domain, and} \\ q = g \text{ on the boundary of the domain,} \end{aligned} \tag{12.11}$$

where $f$ and $g$ are known. We approximate (12.11) 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:

$$\begin{aligned} d^{-2}\left(4q_{j,k} - q_{j-1,k} - q_{j+1,k} - q_{j,k-1} - q_{j,k+1}\right) = f_{j,k} \text{ for } 0 < (j,k) < N, \\ q_{j,k} = g_{j,k}, \text{ for } j = 0,\ N \text{ and } k = 0, N. \end{aligned} \tag{12.12}$$

In order to explore relaxation methods for the solution of (12.12), we need a notation that allows us to distinguish approximate solutions from exact solutions. Here by "exact" solution we mean an exact solution to the *finite-difference problem* posed in (12.12). We let $\widehat{q}_{j,k}$ 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 $\widehat{q}^{new}_{j,k}$ by applying (12.12) with the new value at the point $(j,k)$ and the "old" values at the neighboring points, i.e.,

$$d^{-2}\left(4\widehat{q}^{new}_{j,k} - \widehat{q}_{j-1,k} - \widehat{q}_{j+1,k} - \widehat{q}_{j,k-1} - \widehat{q}_{j,k+1}\right) = f_{j,k}, \tag{12.13}$$

so that

$$\widehat{q}^{new}_{j,k} = \frac{1}{4}\left(d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1}\right). \tag{12.14}$$

With this approach, we compute $\widehat{q}^{new}_{j,k}$ at all interior points using (12.13), 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 methods to determine the speed of convergence are discussed below.

Jacobi relaxation is well suited to parallelization, because we do the exactly 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 (12.14) will all be the same number, and those four terms alone will try to make ${q^{new}}_{j,k}$ 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 [1]

[1] 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.

$$\varepsilon_{j,k} \equiv \widehat{q}_{j,k} - q_{j,k}. \tag{12.15}$$

Here again $q_{j,k}$ is the exact solution of the finite-difference system.Using (12.15) to eliminate all values of $\widehat{q}$ in (12.14), we find that

$$\varepsilon_{j,k}^{new} + q_{j,k} = \frac{1}{4}\left[d^2 f_{j,k} + \left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1}\right) + \left(q_{j-1,k} + q_{j+1,k} + q_{j,k-1} + q_{j,k+1}\right)\right]. \tag{12.16}$$

Since the exact solution satisfies (12.12), we can simplify (12.16) to

$$\varepsilon_{j,k}^{new} = \frac{1}{4}\left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1}\right). \tag{12.17}$$

Eq. (12.17) 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 (12.17) 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. This means that the checkerboard error can never be reduced to zero by Jacobi iteration. That's bad.

Here is a more general way to analyze the method. First, rewrite (12.17) as

$$\varepsilon_{j,k}^{new} = \varepsilon_{j,k} + \frac{1}{4}\left(\varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1} - 4\varepsilon_{j,k}\right). \tag{12.18}$$

The quantity in parentheses in (12.18) is an "increment" which, when added to the "old" error, $\varepsilon_{j,k}$, gives the new error, $\varepsilon_{j,k}^{new}$. Eq. (12.18) looks like time differencing, which we have already discussed in some detail. We can use von Neumann's method to analyze the decrease in the error from one sweep to the next. First, write

$$\varepsilon_{j,k} = \varepsilon_0 e^{i(jld+kmd)}, \tag{12.19}$$

where $l$ and $m$ are the wave numbers in the $x$ and $y$ directions, respectively. We also define an "amplification factor" by

$$\varepsilon_{j,k}{}^{new} \equiv \lambda \varepsilon_{j,k}. \tag{12.20}$$

Substituting (12.19) and (12.20) into (12.18), we find that

$$\begin{aligned} \lambda &= 1 + \frac{1}{4}\left(e^{i(j-1)ld} + e^{i(j+1)ld} + e^{i(k-1)md} + e^{i(k+1)md} - 4\right) \\ &= \frac{1}{2}\left[\cos(ld) + \cos(md)\right]. \end{aligned} \tag{12.21}$$

If $\lambda$ is negative, the sign of the error will oscillate from one sweep to the next. To have rapid, monotonic convergence, we want $\lambda$ to be positive and considerably less than one. Eq. (12.21) shows that for "long" modes, with $ld \ll 1$ and $md \ll 1$, $\lambda$ is just slightly less than one. This means that the error in the long modes goes away slowly; it will take lots of iterations 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 checkerboard is an exception.

A strategy to overcome the checkerboard problem is to "*under-relax*." To understand this approach, we first re-write (12.14) as

$$\widehat{q}_{j,k}^{new} = \widehat{q}_{j,k} + \left[\frac{1}{4}\left(d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1}\right) - \widehat{q}_{j,k}\right]. \tag{12.22}$$

This simply says that $\widehat{q}_{j,k}^{new}$ 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 $\widehat{q}_{j,k}^{new}$ flips from one iteration to the next. This simple obsevation suggests that it would be useful to reduce the increment by multiplying it by a factor less than one, which we will call $\omega$, i.e., we replace (12.18) by

$$\widehat{q}_{j,k}^{new} = \widehat{q}_{j,k} + \omega \left[ \frac{1}{4} \left( d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1} \right) - \widehat{q}_{j,k} \right]. \tag{12.23}$$

where $0 < \omega < 1$. For $\omega = 0$, a sweep does nothing. For $\omega = 1$, (12.23) reverts to (12.22). We can rearrange (12.18) to

$$\widehat{q}_{j,k}^{new} = \widehat{q}_{j,k} (1 - \omega) + \frac{\omega}{4} \left( d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1} \right). \tag{12.24}$$

Substitution of (12.15) into (12.24), and use of (12.12), gives

$$\varepsilon_{j,k}^{new} = \varepsilon_{j,k} (1 - \omega) + \frac{\omega}{4} \left( \varepsilon_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon_{j,k-1} + \varepsilon_{j,k+1} \right). \tag{12.25}$$

From (12.25), you should be able to see that, if we choose $\omega = 0.5$, 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 $\omega$ is in the range $0.5 < \omega < 1$.

Suppose that, on a particular sweep, the error is spatially uniform over the grid. Then, according to (12.17), the error will never change under Jacobi relaxation, and this is true even with under-relaxation, as can be seen from (12.25). 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.

*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.

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.

The long modes are the ones most accurately resolved on the grid, so it is ironic that they "cause trouble" by limiting the speed of convergence.

For errors of intermediate spatial scale, Jacobi relaxation works reasonably well.

## 12.4 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 $q$s that have already been updated, so that (12.14) is replaced by

$$\widehat{q}^{new}_{j,k} = \frac{1}{4}\left(d^2 f_{j,k} + \widehat{q}^{new}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}^{new}_{j,k-1} + \widehat{q}_{j,k+1}\right). \tag{12.26}$$

This 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.

Obviously (12.26) 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 (12.14). It is only for the second and later *rows* of points that (12.26) 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 (12.12) exactly, it will be independent of where the sweep starts.

For Gauss-Seidel relaxation, the error-reduction formula corresponding to (12.17) is

$$\varepsilon^{new}_{j,k} = \frac{1}{4}\left(\varepsilon^{new}_{j-1,k} + \varepsilon_{j+1,k} + \varepsilon^{new}_{j,k-1} + \varepsilon_{j,k+1}\right), \tag{12.27}$$

and the amplification factor defined by (12.20) turns out to be a complex number, which means that the error will oscillate 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:

$$\widehat{q}^{0}_{j,k} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}. \tag{12.28}$$

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:

$$\widehat{q}^{1,\text{partial}}_{j,k} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ -0.5 & 0.25 & -0.281 & -1 & 1 & -1 \\ 1 & -0.5 & 0.625 & -0.593 & 0.602 & -0.60 \end{bmatrix}. \tag{12.29}$$

Although the solution is flipping sign as a result of the sweep, the amplitude of the checkerboard is decreasing noticeably. You can finish the exercise for yourself.

Inspection of (12.29) shows that the errors have been reduced after one (partial) sweep. Convergence can be speeded up by multiplying the increment by a factor *greater* than one, i.e., by "*over-relaxation*." By analogy with (12.24), we replace (12.26) by

$$\widehat{q}^{new}_{j,k} = \widehat{q}_{j,k}(1-\omega) + \frac{\omega}{4}\left(d^2 f_{j,k} + \widehat{q}^{new}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}^{new}_{j,k-1} + \widehat{q}_{j,k+1}\right), \tag{12.30}$$

where this time we choose $\omega > 1$. Choosing $\omega$ too large will cause the iteration to diverge. It can be shown that the convergence of (12.30) is optimized (i.e., made as rapid as possible) if we choose

$$\omega = \frac{2}{1+\sin\left(\frac{\pi d}{L}\right)}, \tag{12.31}$$

where $L$ is the total width of the domain. According to (12.31), $\omega$ approaches 2 on very fine grids. In practice, some experimentation may be needed to find the best value of $\omega$.

The algorithm represented by (12.30) and (12.31) is called "successive over-relaxation," or SOR.

## 12.5 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 (12.12) as

$$\left(-q_{j-1,k}+2q_{j,k}-q_{j+1,k}\right)+\left(-q_{j,k-1}+2q_{j,k}-q_{j,k+1}\right)=d^2 f_{j,k}. \tag{12.32}$$

The first quantity in parentheses on the left-hand side of (12.32) 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

$$\left[-\widehat{q}^{\,\text{int}}_{j-1,k}+(2+r)\,\widehat{q}^{\,\text{int}}_{j,k}-\widehat{q}^{\,\text{int}}_{j+1,k}\right]=d^2 f_{j,k}-\left[-\widehat{q}_{j,k-1}+(2-r)\,\widehat{q}_{j,k}-\widehat{q}_{j,k+1}\right] \tag{12.33}$$

for the values with superscript "int." Here $r$ is a parameter used to control convergence, as discussed below. Eq. (12.33) is a tri-diagonal system, which can easily be solved. The sweep is completed by solving

$$\left[-\widehat{q}^{\,new}_{j,k-1}+(2-r)\,\widehat{q}^{\,new}_{j,k}-\widehat{q}^{\,new}_{j,k+1}\right]=d^2 f_{j,k}-\left[-\widehat{q}^{\,\text{int}}_{j-1,k}+(2+r)\,\widehat{q}^{\,\text{int}}_{j,k}-\widehat{q}^{\,\text{int}}_{j+1,k}\right] \tag{12.34}$$

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

$$r = 2\sin\left(\frac{\pi d}{L}\right). \tag{12.35}$$

## 12.6 Multigrid methods

Fulton et al. (1986) summarize the multi-grid approach to solving boundary-value problems, which was developed by Achi Brandt (Brandt (1973), Brandt (1977); see additional references in Fulton's paper). The basic idea is very simple and elegant, although implementation can be complicated.

As we have already discussed, with Gauss-Seidel relaxation the small-scale errors are eliminated quickly, while the large-scale errors disappear more slowly. As the iteration proceeds, the distribution of the error over the domain becomes smoother, because it is approaching zero everywhere. A key observation is that, essentially by the definition of "large-scale," *the large-scale part of the error can be represented on a relatively coarse grid*. On such a coarse grid, the large-scale errors are represented using fewer grid points, and so 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 relax away the large-scale errors, and a fine grid to relax away the small-scale errors. In practice, we introduce *as many "nested" grids as possible*; the "multi" in the multi-grid method is quite important. Each coarse grid is composed of a subset of the points used in the finer grids. We move back and forth between the grids, from coarse to fine by interpolation, and from fine to coarse by simply copying the fine-grid values onto the corresponding points of the coarse grid. A relaxation (e.g., Jacobi or Gauss-Seidel) is done on each grid in turn. The sweeps on the coarser grids remove the large-scale part of the error, while the sweeps on the finer grids remove the small-scale part of the error.

Although the transfers between grids involve some computational work, the net effect is to speed up the solution (for a given degree of error) considerably beyond what can be achieved through relaxation exclusively on the finest grid. In addition, the scaling improves.

Here is a brief summary of how a multigrid method can be implemented in practice. Suppose that our unknown, $q$, satisfies

$$\begin{aligned} -L^M q^M &= f^M, \text{ and} \\ q &= g \text{ on the boundary of the domain,} \end{aligned} \tag{12.36}$$

where $L$ is a linear operator (which could be the Laplacian). Eq. (12.36) is a generalization of (12.11). The superscripts $M$ in (12.36) denote the fine grid on which we want to obtain the solution of (12.36). We need this grid-naming notation because the multigrid method involves multiple grids. The approximate solution on grid $M$ is given by $\hat{q}^M$, and the corresponding error is denoted by

$$\hat{\varepsilon}^M \equiv \hat{q}^M - q^M. \tag{12.37}$$

This is all the same as in our earlier discussion, except for the new superscripts $M$.

Now we add some new ideas. Using (12.37) to eliminate $q^M$ in (12.36), we find that

$$-L^M \left(\hat{q}^M - \hat{\varepsilon}^M\right) = f^M. \tag{12.38}$$

Since $L$ is linear (by assumption), we know that

$$L^M \left(\hat{q}^M - \hat{\varepsilon}^M\right) = L^M \hat{q}^M - L^M \hat{\varepsilon}^M. \tag{12.39}$$

With the use of (12.39), we can rewrite (12.38) as

$$\boxed{L^M \hat{\varepsilon}^M = r^M,} \tag{12.40}$$

where

$$r^M \equiv f^M + L^M \hat{q}^M \tag{12.41}$$

is called the "*residual*," and Eq. (12.40) is called the "*residual equation*." The residual is the what comes out when the operator $L^M$ is applied to the error. Eq. (12.40) shows that

when the error is zero everywhere, the residual is also zero. We can say that the residual is a measure of the error.

As can be seen from (12.41), the quantities needed to compute $r^M$ are known. They are the forcing function, $f^M$, and the approximate (partially converged) solution, $\hat{q}^M$. If we have a current guess for $\hat{q}^M$, then we can calculate the corresponding $r^M$. If $\hat{q}^M$ changes on the next sweep, then $r^M$ will also change.

In contrast, the error itself is not known. If the error were known, we could just use (12.37) to compute the exact solution, i.e.,

$$q^M = \hat{q}^M - \hat{\varepsilon}^M \tag{12.42}$$

and we would be done!

Since the residual is known, *the unknown in* (12.40) *is the error*, $\hat{\varepsilon}^M$. Instead of solving (12.36) for $q^M$, we can solve the residual equation (12.40) for $\hat{\varepsilon}^M$.

But, you may be wondering, how does this help? Either way, we still have to solve a linear system. The reason that it is advantageous to solve for $\hat{\varepsilon}^M$ instead of $q^M$ is that, during the iteration, the high-wave-number part of the error is quickly eliminated, so that part-way through the solution procedure $\hat{\varepsilon}^M$ is smooth, even if the final solution for $q^M$ is not smooth. This is where the multigrid method comes in. *Because* $\hat{\varepsilon}^M$ *is smooth, we can represent it on a coarser grid*, which will be denoted by superscript $l$. This is the motivation for solving for the smooth $\hat{\varepsilon}^M$ rather than for (the possibly noisy) $q^M$.

On a given grid, the procedure is as follows. We have a current guess, $\hat{q}^M$. This allows us to calculate the residual, $r^M$, from (12.41). Then we do a sweep with (12.40) to obtain an updated error field, $\varepsilon^M$.

Typically each coarser grid is chosen to have half as many points (in each direction) as the next finer grid, so, with a two-dimensional domain, grid $l$ would have $1/4$ as many points as grid $M$. We replace (12.40) by

$$L^l \hat{\varepsilon}^l = r^l, \tag{12.43}$$

where superscript $l$ denotes the coarser grid. The process by which quantities are transferred from the finer grid to the coarser grid is called "*restriction*." In many cases, the points that comprise grid $l$ are a subset of the points on grid $M$, in which case we can just copy the appropriate values; no interpolation is necessary. This special case of restriction is called "*injection*."

You can probably see where this is going. After doing a sweep (or possibly more than one) on grid $l$, the error has been further smoothed, and a new residual equation can be solved on an even coarser grid. This process can be repeated until reaching the coarsest possible grid – say a 2 x 2 grid (Fig. 12.1). On the coarsest grids, direct solves (e.g., matrix inversion) may be preferable to iterative methods.

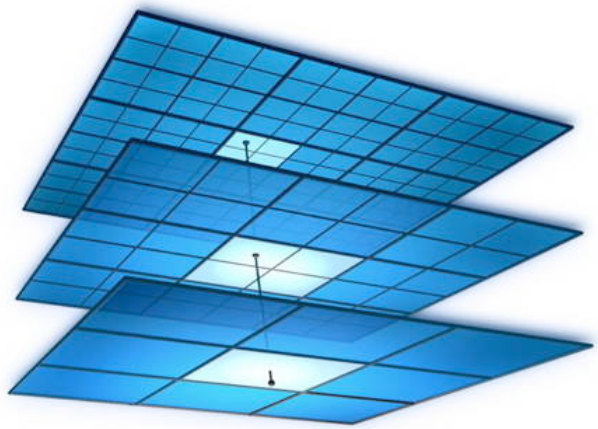


Figure 12.1: Schematic illustrating three grids used by a multigrid solver, with the finest grid on top. Source: `http://ccma.math.psu.edu/ccma-wp/?page_id=237`.

Having worked our way down to the coarsest possible grid, we then start back the other way, towards the finer grids. The error on a coarse grid is *interpolated* to construct the error on the next finer grid. This is called "*prolongation*." The error on the finer grid is then subtracted from the previous estimate of $\hat{q}$ on that grid, essentially following (12.42). This gives a revised (i.e., improved) estimate of $\hat{q}$ on that grid, which can be used to compute a new residual using (12.41). A sweep is performed to obtain a new estimate of the error on the finer grid, and the result is interpolated to the next finer grid, and so on, until we arrive back at the finest grid.

A sequence of sweeps on successively coarser grids, followed by a sequence of sweeps on the successively finer grids, is called a *V-cycle*.

For further discussion of multi-grid methods, see the paper by Fulton et al. (1986).

## 12.7 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 12.1 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 12.1: 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 | $N^2$ | $N^2$ |

## 12.8 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

$$\widehat{q}_{j,k}^{new} = \widehat{q}_{j,k} + \omega \left[ \frac{1}{4} \left( d^2 f_{j,k} + \widehat{q}_{j-1,k} + \widehat{q}_{j+1,k} + \widehat{q}_{j,k-1} + \widehat{q}_{j,k+1} \right) - \widehat{q}_{j,k} \right]. \tag{12.44}$$

3. Consider a square domain, of width $L$, with periodic boundary conditions in both $x$ and $y$ directions. We wish to solve

$$\nabla^2 q = f(x,y) = \left( \sin \frac{4\pi x}{L} \right) \left( \cos \frac{4\pi y}{L} \right) \tag{12.45}$$

for the unknown function $q$, where

$$\begin{aligned} 0 \leq & x \leq L, \\ 0 \leq & y \leq L. \end{aligned} \tag{12.46}$$

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 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 $\omega$;
- Gauss-Seidel relaxation;
- Gauss-Seidel over-relaxation, with a suitable choice of the parameter $\omega$.

For each of the relaxation methods, try the following two initial guesses:

$$\begin{array}{l} 1)\ q_{j,k} = (-1)^{j+k}, \\ 2)\ q_{j,k} = 0 \text{ everywhere .} \end{array} \tag{12.47}$$

(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

$$\varepsilon^n_{j,k} \equiv \left(\widehat{q}^n_{j,k}\right) - f_{j,k}. \tag{12.48}$$

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 (12.20), as evaluated on the grid. Let the convergence criterion be

$$\text{Max}\forall(j,k)\left\{\left|\varepsilon^n_{j,k}\right|\right\} < 10^{-2}\text{Max}\forall(j,k)\left\{\left|f_{j,k}\right|\right\}. \tag{12.49}$$

How many iterations are needed to obtain convergence with Jacobi, Gauss-Seidel, and SOR?

(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 2. 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.

   Note that in all cases the domain-average of the specified vorticity "forcing" must be zero. Explain why that is true. Require 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 (12.6) on a domain with 128 equally spaced grid points and periodic boundary conditions. For use in the test, invent a forcing $f(x)$ that is sufficiently complicated to be interesting. 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 13

# It's only dissipation (but I like it)

## 13.1 Introduction

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

$$\frac{\partial q}{\partial t} = -\frac{\partial F_q}{\partial x}. \tag{13.1}$$

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

$$F_q = -K\frac{\partial q}{\partial x}, \tag{13.2}$$

where $K$ is a "diffusion coefficient," which must be determined somehow. Physically meaningful applications of are possible when

$$K \geq 0. \tag{13.3}$$

Substitution of (13.2) into (13.1) gives

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial x}\left(K\frac{\partial q}{\partial x}\right). \tag{13.4}$$

Because (13.4) 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 $q$ is periodic. It then follows immediately from (13.1) that the spatially averaged value of $q$ does not change with time:

$$\frac{d}{dt}\left(\int_{\text{spatial domain}} q dx\right) = 0. \tag{13.5}$$

When (13.3) is satisfied, (13.4) 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

$$\frac{d}{dt}\left(\int_{\text{spatial domain}} q^2 dx\right) \le 0. \tag{13.6}$$

To prove this, multiply both sides of (13.4) by $q$:

$$\begin{aligned} \frac{\partial}{\partial t}\left(\frac{q^2}{2}\right) &= q\frac{\partial}{\partial x}\left(K\frac{\partial q}{\partial x}\right) \\ &= \frac{\partial}{\partial x}\left(qK\frac{\partial q}{\partial x}\right) - K\left(\frac{\partial q}{\partial x}\right)^2. \end{aligned} \tag{13.7}$$

When we integrate the second line of (13.7) over a periodic domain, the first term vanishes and the second is negative (or possibly zero). The result (13.6) follows immediately.

With the assumed periodic boundary conditions, we can expand $q$ in a Fourier series:

$$q(x,t) = \sum_k \mathrm{Re}\left[\hat{q}_k(t)\, e^{ikx}\right]. \tag{13.8}$$

Substituting into (13.4), and (temporarily) assuming spatially constant $K$, we find that the amplitude of a particular Fourier mode satisfies

$$\frac{d\hat{q}_k}{dt} = -k^2 K \hat{q}_k, \tag{13.9}$$

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 (13.9) is

$$\hat{q}_k(t) = \hat{q}_k(0)\, e^{-k^2 K t}. \tag{13.10}$$

Note that higher wave numbers decay more rapidly, for a given value of $K$. Since

$$\hat{q}_k(t+\Delta t) = \hat{q}_k(0)\, e^{-k^2 K(t+\Delta t)} = \hat{q}_k(t)\, e^{-k^2 K \Delta t}, \tag{13.11}$$

we see that, for the exact solution, the amplification factor is given by

$$\lambda = e^{-k^2 K \Delta t} < 1. \tag{13.12}$$

## 13.2 A simple explicit scheme

A finite-difference analog of (13.1) is

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left(q_{j+1}^n - q_j^n\right) - \kappa_{j-\frac{1}{2}} \left(q_j^n - q_{j-1}^n\right), \tag{13.13}$$

where for convenience we define the nondimensional combination

$$\kappa_{j+\frac{1}{2}} \equiv \frac{K_{j+\frac{1}{2}} \Delta t}{(\Delta x)^2}. \tag{13.14}$$

Here we have assumed for simplicity that $\Delta x$ is a constant. It should be obvious that, with periodic boundary conditions, (13.13) guarantees conservation of $q$ in the sense that

$$\sum_j q_j^{n+1} \Delta x = \sum_j q_j^n \Delta x. \tag{13.15}$$

The scheme given by (13.13) 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 (13.13) using von Neumann's method, we have to assume that $\kappa$ is a constant. Then (13.13) yields

$$(\lambda - 1) = \kappa \left[ \left( e^{ik\Delta x} - 1 \right) - \left( 1 - e^{-ik\Delta x} \right) \right], \tag{13.16}$$

which is equivalent to

$$\lambda = 1 - 4\kappa \sin^2 \left( \frac{k\Delta x}{2} \right) \leq 1. \tag{13.17}$$

Note that $\lambda$ is real and less than one. Instability occurs for $\lambda < -1$, which is equivalent to

$$\kappa \sin^2 \left( \frac{k\Delta x}{2} \right) > \frac{1}{2}. \tag{13.18}$$

The worst case is $\sin^2 \left( \frac{k\Delta x}{2} \right) = 1$, which occurs for $\left( \frac{k\Delta x}{2} \right) = \frac{\pi}{2}$, or $k\Delta x = \pi$. This is the $2\Delta x$ wave. We conclude that with (13.13)

$$\kappa \leq \frac{1}{2} \text{ is required for stability.} \tag{13.19}$$

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

$$\sum_j \left( q_j^{n+1} \right)^2 < \sum_j \left( q_j^n \right)^2; \tag{13.20}$$

this is the condition for stability according to the energy method discussed in Chapter 2. Eq. (13.20) is analogous to (13.6).

## 13.3 An implicit scheme

We can obtain unconditional stability through the use of an implicit scheme, but at the cost of some additional complexity. Replace (13.13) by

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left( q_{j+1}^{n+1} - q_j^{n+1} \right) - \kappa_{j-\frac{1}{2}} \left( q_j^{n+1} - q_{j-1}^{n+1} \right). \tag{13.21}$$

We use the *energy method* to analyze the stability of (13.21), for the case of spatially variable but non-negative $K$. Multiplying (13.21) by $q_j^{n+1}$, we obtain:

$$\left( q_j^{n+1} \right)^2 - q_j^{n+1} q_j^n = \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \kappa_{j+\frac{1}{2}} \left( q_j^{n+1} \right)^2 - \kappa_{j-\frac{1}{2}} \left( q_j^{n+1} \right)^2 + \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} \tag{13.22}$$

Summing over the domain gives

$$\begin{aligned} \sum_j \left( q_j^{n+1} \right)^2 - \sum_j q_j^{n+1} q_j^n &= \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} \left( q_j^{n+1} \right)^2 \\ &\quad - \sum_j \kappa_{j-\frac{1}{2}} \left( q_j^{n+1} \right)^2 + \sum_j \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} \\ &= \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} \left( q_j^{n+1} \right)^2 \\ &\quad - \sum_j \kappa_{j+\frac{1}{2}} \left( q_{j+1}^{n+1} \right)^2 + \sum_j \kappa_{j+\frac{1}{2}} q_j^{n+1} q_{j+1}^{n+1} \\ &= - \sum_j \kappa_{j+\frac{1}{2}} \left( q_{j+1}^{n+1} - q_j^{n+1} \right)^2, \end{aligned} \tag{13.23}$$

which can be rearranged to

$$\sum_j q_j^{n+1} q_j^n = \sum_j \left[ \left( q_j^{n+1} \right)^2 + \kappa_{j+\frac{1}{2}} \left( q_{j+1}^{n+1} - q_j^{n+1} \right)^2 \right]. \tag{13.24}$$

Next, note that

$$\sum_j \left(q_j^{n+1} - q_j^n\right)^2 = \sum_j \left[\left(q_j^{n+1}\right)^2 + \left(q_j^n\right)^2 - 2q_j^{n+1}q_j^n\right] \geq 0. \tag{13.25}$$

Substitute (13.24) into (13.25), to obtain

$$\sum_j \left\{\left(q_j^{n+1}\right)^2 + \left(q_j^n\right)^2 - 2\left[\left(q_j^{n+1}\right)^2 + \kappa_{j+\frac{1}{2}}\left(q_{j+1}^{n+1} - q_j^{n+1}\right)^2\right]\right\} \geq 0, \tag{13.26}$$

which can be simplified and rearranged to

$$\sum_j \left[\left(q_j^{n+1}\right)^2 - \left(q_j^n\right)^2\right] \leq -2\sum_j \left[\kappa_{j+\frac{1}{2}}\left(q_{j+1}^{n+1} - q_j^{n+1}\right)^2\right] \leq 0. \tag{13.27}$$

Eq. (13.27) 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,

$$\sum_j \left[\left(q_j^{n+1}\right)^2 - \left(q_j^n\right)^2\right] \leq 0. \tag{13.28}$$

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. (13.21) 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.

## 13.4 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

$$\frac{q_j^{n+1}-q_j^{n-1}}{2\Delta t}=\frac{1}{(\Delta x)^2}\left[K_{j+\frac{1}{2}}\left(q_{j+1}^n-q_j^{n+1}\right)-K_{j-\frac{1}{2}}\left(q_j^{n-1}-q_{j-1}^n\right)\right]. \tag{13.29}$$

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* (13.29) *can be obtained without solving a system of simultaneous equations:*

$$q_j^{n+1}=\frac{q_j^{n-1}+2\left[\kappa_{j+\frac{1}{2}}q_{j+1}^n-\kappa_{j-\frac{1}{2}}\left(q_j^{n-1}-q_{j-1}^n\right)\right]}{1+2\kappa_{j+\frac{1}{2}}}. \tag{13.30}$$

To apply von Neumann's method, we assume spatially constant $\kappa$, and for convenience define

$$\alpha\equiv 2\kappa\geq 0. \tag{13.31}$$

The amplification factor satisfies

$$\lambda^2-1=\alpha\left(\lambda e^{ik\Delta x}-\lambda^2-1+\lambda e^{-ik\Delta x}\right), \tag{13.32}$$

which is equivalent to

$$\lambda^2(1+\alpha)-\lambda 2\alpha\cos(k\Delta x)-(1-\alpha)=0. \tag{13.33}$$

The solutions are

$$\begin{aligned}\lambda &= \frac{\alpha\cos(k\Delta x) \pm \sqrt{\alpha^2\cos^2(k\Delta x) + (1-\alpha^2)}}{1+\alpha} \\ &= \frac{\alpha\cos(k\Delta x) \pm \sqrt{1-\alpha^2\sin^2(k\Delta x)}}{1+\alpha}.\end{aligned} \tag{13.34}$$

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 $\lambda$ is real, and by considering the two solutions separately it is easy to show that

$$|\lambda| \leq \frac{1+|\alpha\cos(k\Delta x)|}{1+\alpha} \leq 1. \tag{13.35}$$

Second, if $\alpha^2\sin^2(k\Delta x) > 1$, which implies that $\alpha > 1$, then $\lambda$ is complex, and we find that

$$|\lambda| = \frac{\sqrt{\alpha^2\cos^2(k\Delta x) + \alpha^2\sin^2(k\Delta x) - 1}}{1+\alpha} = \frac{\sqrt{\alpha^2-1}}{1+\alpha} = \sqrt{\frac{\alpha-1}{\alpha+1}} < 1. \tag{13.36}$$

We conclude that the scheme is unconditionally stable.

It does not follow, however, that the scheme gives a good solution for large $\Delta t$. For $\alpha \to \infty$ (strong diffusion and/or a long time step), (13.36) gives

$$|\lambda| \to 1. \tag{13.37}$$

We conclude that the Dufot-Frankel scheme does not damp when the diffusion coefficient is large or the time step is large. This is very bad behavior.

## 13.5 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 $\Delta x$.

## 13.6 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 both the explicit and implicit versions of the diffusion equation, for a periodic domain consisting of 100 grid points, with constant $K = 1$ and $\Delta x = 1$. Also program the DuFort-Frankel scheme. Let the initial condition be

$$q_j = 100,\ j \in [1,\ 50]\,,\ \text{and}\ q_j = 110\ \text{for}\ j \in [51, 100]\,. \tag{13.38}$$

   Compare the three solutions for different choices of the time step.

3. Use the energy method to evaluate the stability of

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}} \left(q_{j+1}^n - q_j^n\right) - \kappa_{j-\frac{1}{2}} \left(q_j^n - q_{j-1}^n\right). \tag{13.39}$$

   Do not assume that $\kappa$ is spatially constant.

# Chapter 14

# Making Waves

## 14.1 The shallow-water equations

In most of this chapter we will discuss the shallow-water equations, which can be written as

$$\frac{\partial \mathbf{v}}{\partial t} + (\zeta + f)\,\mathbf{k} \times \mathbf{v} = -\nabla \left[ g\,(h + h_S) + K \right], \tag{14.1}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (\mathbf{v} h) = 0. \tag{14.2}$$

Here $\mathbf{v}$ is the horizontal velocity vector, $\zeta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{v})$ is the vertical component of the vorticity, $f$ is the Coriolis parameter, $h$ is the depth of the fluid, $h_S$ is the height of the "bottom topography," $g$ is the acceleration of gravity, and $K \equiv \frac{1}{2}\mathbf{v} \cdot \mathbf{v}$ is the kinetic energy per unit mass. In (14.1), all frictional effects have been neglected, for simplicity. Although the shallow water equations are highly idealized, they are extremely useful for testing the horizontal (and temporal) discretizations of numerical models that are used to simulate atmospheric dynamics.

For the special case of a one-dimensional, non-rotating small-amplitude gravity wave on a resting basic state, without topography, Eqs. (14.1) and (14.2) become

$$\frac{\partial u}{\partial t} + g \frac{\partial h}{\partial x} = 0, \tag{14.3}$$

and

$$\frac{\partial h}{\partial t} + H\frac{\partial u}{\partial x} = 0, \tag{14.4}$$

respectively. The equations have been linearized, and $H$ is the mean depth of the fluid. We refer to (14.3) - (14.4) as "the gravity wave equations." Let

$$c^2 \equiv gH. \tag{14.5}$$

Note that the phase speed $c$ is independent of wave number for pure gravity waves. As discussed later, including the Coriolis terms will cause the phase speed to depend on wave number. By combining (14.3) - (14.4) we can derive

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \tag{14.6}$$

and

$$\frac{\partial^2 h}{\partial t^2} = c^2\frac{\partial^2 h}{\partial x^2}, \tag{14.7}$$

which are both examples of "the wave equation."

Assuming solutions of the form, $e^{i(kx\pm\sigma t)}$, and substituting into either (14.6) or (14.7), we obtain the dispersion equation

$$\sigma^2 = c^2k^2. \tag{14.8}$$

There are two waves, one propagating in the positive $x$-direction, and the other in the negative $x$-direction. Remember that $c$ is a constant, given by (14.5). Therefore (14.8) tells us that the frequency $\sigma$ is a constant times the wave number $k$. The physical meaning should be clear.

## 14.2 The normal forms

We demonstrate in this section that the solutions of (14.6) are constant along space-time lines (or surfaces) called "characteristics." A solution is fully determined if $u$ and $\frac{\partial u}{\partial t}$ are specified somewhere on each characteristic. The characteristics can, and generally do, intersect boundaries. As with the advection equation, $f(x-ct)$ is a solution of the wave equation (14.6), but $g(x+ct)$ is a second solution. Here we can assume $c>0$ without loss of generality. Note that in this discussion we are not assuming a single Fourier mode; the functions $f$ and $g$ can include many Fourier components in the $x$ direction. The general solution of (14.6) is given by a super-position of the eastward- and westward propagating solutions. In particular, for $u(x,t)$ we can write

$$u(x,t) = f(x-ct) + g(x+ct). \tag{14.9}$$

The forms of $f(x-ct)$ and $g(x+ct)$ are completely determined by the initial conditions, which can be written as

$$\begin{aligned} u_{t=0} &= F(x), \\ \left(\frac{\partial u}{\partial t}\right)_{t=0} &= G(x). \end{aligned} \tag{14.10}$$

Note that $\left(\frac{\partial u}{\partial t}\right)_{t=0} = G(x)$ contains information about $h(x,0)$, so together these two initial conditions contain information about both the mass field and the wind field at $t=0$.

Substituting (14.9) into (14.10), we find that at $t=0$

$$\begin{aligned} f(x) + g(x) &= F(x), \\ -cf'(x) + cg'(x) &= G(x). \end{aligned} \tag{14.11}$$

Here a prime denotes differentiation. Differentiating the first of (14.11), and then using the second, we can solve for $f'(x)$ and $g'(x)$:

$$\begin{aligned} f'(x) &= \frac{1}{2}\left[F'(x) - \frac{G(x)}{c}\right], \\ g'(x) &= \frac{1}{2}\left[F'(x) + \frac{G(x)}{c}\right]. \end{aligned} \tag{14.12}$$

These can be integrated to obtain $f(x)$ and $g(x)$:

$$\begin{aligned} f(x) &= \frac{1}{2}\left[F(x) - \frac{1}{c}\int_0^x G(\xi)\,d\xi\right] + C_1, \\ g(x) &= \frac{1}{2}\left[F(x) + \frac{1}{c}\int_0^x G(\xi)\,d\xi\right] + C_2. \end{aligned} \tag{14.13}$$

Here $C_1$ and $C_2$ are constants of integration. Finally, we obtain $u(x,t)$ by replacing $x$ by $x-ct$ and $x+ct$, respectively, in the expressions for $f(x)$ and $g(x)$ in (14.13), and then substituting back into (14.9). This gives

$$u(x,t) = \frac{1}{2}\left[F(x-ct) + F(x+ct) + \frac{1}{c}\int_{x-ct}^{x+ct} G(\xi)\,d\xi\right]. \tag{14.14}$$

Here $\xi$ is a dummy variable of integration, and 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 (14.14).

Once $u(x,t)$ is known from (14.14), we can obtain $\frac{\partial h}{\partial x}$ from (14.3). The constant "background" value of $h$ cannot be determined without additional information.

In order to make connections between the wave equation and the advection equation that we have already analyzed, we reduce (14.6) to a pair of first-order equations by defining

$$p \equiv \frac{\partial u}{\partial t} \text{ and } q \equiv -c\frac{\partial u}{\partial x}. \tag{14.15}$$

Substitution of (14.15) into the wave equation (14.6) gives

$$\frac{\partial p}{\partial t} + c\frac{\partial q}{\partial x} = 0, \tag{14.16}$$

and differentiation of the second of (14.15) with respect to $t$, with the use of the first of (14.15), gives

$$\frac{\partial q}{\partial t} + c\frac{\partial p}{\partial x} = 0. \tag{14.17}$$

If we alternately add (14.16) and (14.17), and subtract (14.17) from (14.16), we obtain

$$\frac{\partial P}{\partial t} + c\frac{\partial P}{\partial x} = 0, \text{ where } P \equiv p + q, \text{ and} \tag{14.18}$$

$$\frac{\partial Q}{\partial t} - c\frac{\partial Q}{\partial x} = 0, \text{ where } Q \equiv p - q, \tag{14.19}$$

respectively. Now we have a system of two first-order equations, each of which "looks like" the advection equation. 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 (14.18) and (14.19), it is clear that $P$ is constant along the line $x - ct =$ constant, and $Q$ is constant along the line $x + ct =$ constant. Eqs. (14.18) and (14.19) are called the *normal forms* of (14.16) and (14.17).

These concepts are applicable, with minor adjustments, to any hyperbolic system of equations. The curves $x - ct =$ constant and $x + ct =$ constant are called "characteristics." The wave equation is characterized, so to speak, by two such families of curves, while the advection equation has only one family. Because the phase speeds of the pure gravity waves analyzed here are constant, the characteristics that we have found for the waves are straight, parallel lines, but in general they can have any shape so long as the curves within a family do not intersect each other.

## 14.3 Staggered grids for the shallow water equations

Now we discuss the differential-difference equations

$$\frac{du_j}{dt} + g\left(\frac{h_{j+1} - h_{j-1}}{2d}\right) = 0, \tag{14.20}$$

$$\frac{dh_j}{dt} + H\left(\frac{u_{j+1} - u_{j-1}}{2d}\right) = 0, \tag{14.21}$$

where $d$ is the grid spacing. which are, of course, differential-difference analogs of the one-dimensional shallow water equations, (14.3) - (14.4). We keep the time derivatives continuous because the issues that we will discuss next have to do with space differencing only. Consider a distribution of the dependent variables on the grid as shown in Fig. 14.1. Notice that from (14.20) and (14.21) *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.

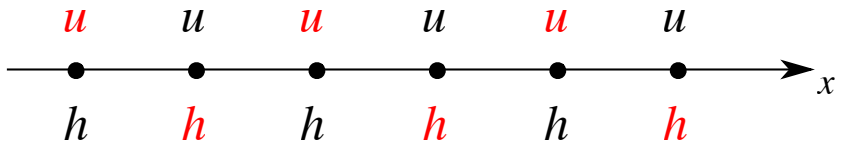



Figure 14.1: A-grid for solution of the one-dimensional shallow water equations.

Here is a mathematical way to draw the same conclusion. The wave solutions of (14.20) and (14.21) are

$$\left(u_j, h_j\right) \sim e^{i(kjd - \sigma t)}, \tag{14.22}$$

giving

$$\begin{aligned} \sigma u_j - g h_j \frac{\sin(kd)}{d} &= 0, \\ \sigma h_j - H u_j \frac{\sin(kd)}{d} &= 0. \end{aligned} \tag{14.23}$$

Provided that $u_j$ and $h_j$ are not both identically zero, we obtain the dispersion relation

$$\sigma^2 = k^2 g H \left( \frac{\sin kd}{kd} \right)^2. \tag{14.24}$$

The phase speed satisfies $c^2 = gH\left(\frac{\sin kd}{kd}\right)^2$. In the exact solution, the phase speed $c = \pm\sqrt{gH}$ is independent of wave number, but 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.

Suppose that $\sigma$ 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 $-\pi \leq kd \leq \pi$.

For convenience, define $p \equiv kd$. If $p = p_0$ satisfies (14.24), then $p = -p_0$, $p = \pi - p_0$ and $p = -(\pi - p_0)$ also satisfy it. This shows that *there are four possible modes for the given frequency, although physically there should only be two. The "extra" pair of modes comes from the redundancy of the grid. The extra modes are computational modes in space.* In chapter 5, we encountered computational modes in space in the context of advection with a boundary condition. 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. In the case of advection, the envelope of a computational mode moves toward the downstream direction. In the case of the wave equation, there is no "downstream" direction.

For a given $\sigma$, the general solution for $u_j$ is a linear combination of the four modes, and can be written as

$$u_j = \left[ A e^{i p_0 j} + B e^{-i p_0 j} + C e^{i(\pi - p_0) j} + D e^{-i(\pi - p_0) j} \right] e^{-i\sigma t}. \tag{14.25}$$

By substituting (14.25) into the second of (14.23), we find that $h_j$ satisfies

$$h_j = \frac{H \sin p_0}{\omega \Delta x} \left[ A e^{i p_0 j} - B e^{-i p_0 j} + C e^{i(\pi - p_0) j} - D e^{-i(\pi - p_0) j} \right] e^{-i\sigma t}. \tag{14.26}$$

Remember that we are assuming $\sigma \geq 0$, so that $\sin(p_0) = \frac{\omega d}{\sqrt{gH}}$ [see (14.24)]. Then (14.26) reduces to

$$h_j = \sqrt{\frac{H}{g}} \left[ A e^{ip_0 j} - B e^{-ip_0 j} + C e^{i(\pi - p_0) j} - D e^{-i(\pi - p_0) j} \right] e^{-i\sigma t}. \tag{14.27}$$

We now repeat the analysis for the case in which the red variables are omitted in Fig. 14.1. The governing equations can be written as

$$\frac{du_{j+1/2}}{dt} + g \left( \frac{h_{j+1} - h_j}{d} \right) = 0, \tag{14.28}$$

$$\frac{dh_j}{dt} + H \left( \frac{u_{j+1/2} - u_{j-1/2}}{d} \right) = 0. \tag{14.29}$$

Here we use half-integer subscripts for the wind points, and integer subscripts for the mass points. The solutions of (14.28) and (14.29) are assumed to have the form

$$\begin{aligned} u_{j+1/2} &= u_0 e^{i[k(j+1/2)d - \sigma t]}, \\ h_j &= h_0 e^{i(kjd - \sigma t)}. \end{aligned} \tag{14.30}$$

Substitution into (14.28) and (14.29) gives

$$\begin{aligned} -\sigma d u_0 + 2g h_0 \sin\left(\frac{kd}{2}\right) &= 0, \\ -\sigma d h_0 + 2H u_0 \sin\left(\frac{kd}{2}\right) &= 0. \end{aligned} \tag{14.31}$$

The resulting dispersion equation is

$$\sigma^2 = k^2 gH \left[ \frac{\sin(kd/2)}{kd/2} \right]^2. \tag{14.32}$$

For a given $\sigma$, there are only two solutions of (14.32), because $kd/2$ ranges only between $-\pi/2$ and $\pi/2$. *This demonstrates that omitting the red variables eliminates the spurious computational modes.*

For a given $\sigma$, the solution for $u_{j+\frac{1}{2}}$ is a linear combination of the two modes, and can be written as

$$u_{j+\frac{1}{2}} = \left[ Ae^{ip_0\left(j+\frac{1}{2}\right)} + Be^{-ip_0\left(j+\frac{1}{2}\right)} \right] e^{-i\sigma t}. \tag{14.33}$$

By substituting (14.33) into (14.29), we find that $h_j$ satisfies

$$h_j = -\frac{H}{\sigma d}\left(Ae^{ip_0 j} - Be^{-ip_0 j}\right) 2\sin\left(\frac{p_0}{2}\right) e^{-i\sigma t}. \tag{14.34}$$

Since we are assuming $\sigma \geq 0$, so that $2\sin\left(\frac{p_0}{2}\right) = \frac{\sigma d}{\sqrt{gH}}$, (14.34) reduces to

$$h_j = -\sqrt{\frac{H}{g}}\left(Ae^{ip_0 j} - Be^{-ip_0 j}\right) e^{-i\sigma t}. \tag{14.35}$$

## 14.4 Dispersion properties as a guide to grid design

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.

AL considered the shallow water equations linearized about a resting basic state, in the following form:

$$\frac{\partial u}{\partial t} - fv + g\frac{\partial h}{\partial x} = 0, \tag{14.36}$$

$$\frac{\partial v}{\partial t} + fu + g\frac{\partial h}{\partial y} = 0, \tag{14.37}$$

$$\frac{\partial h}{\partial t} + H\delta = 0. \tag{14.38}$$

Here $H$ is the constant depth of the "water" in the basic state, $\delta \equiv \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$ is the divergence, and all other symbols have their conventional meanings. From (14.36) - (14.38), we can derive an equivalent set in terms of vorticity, $\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$, and divergence:

$$\frac{\partial \delta}{\partial t} - f\zeta + g\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)h = 0, \tag{14.39}$$

$$\frac{\partial \zeta}{\partial t} + f\delta = 0, \tag{14.40}$$

$$\frac{\partial h}{\partial t} + H\delta = 0. \tag{14.41}$$

Of course, (14.41) is identical to (14.38). It is convenient to eliminate the vorticity and mass in (14.39) by using (14.40) and (14.41), respectively. Then by assuming wave solutions, we obtain the dispersion relation:

$$\left(\frac{\sigma}{f}\right)^2 = 1 + \lambda^2\left(k^2 + l^2\right). \tag{14.42}$$

Here $\sigma$ is the frequency, $\lambda \equiv \frac{\sqrt{gH}}{f}$ is the radius of deformation, 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 (14.42) are important for the geostrophic adjustment process.

In their discussion of various numerical representations of (14.29) - (14.31), AL defined five-grids denoted by "A" through "E," as shown in Fig. 14.2. The figure also shows the

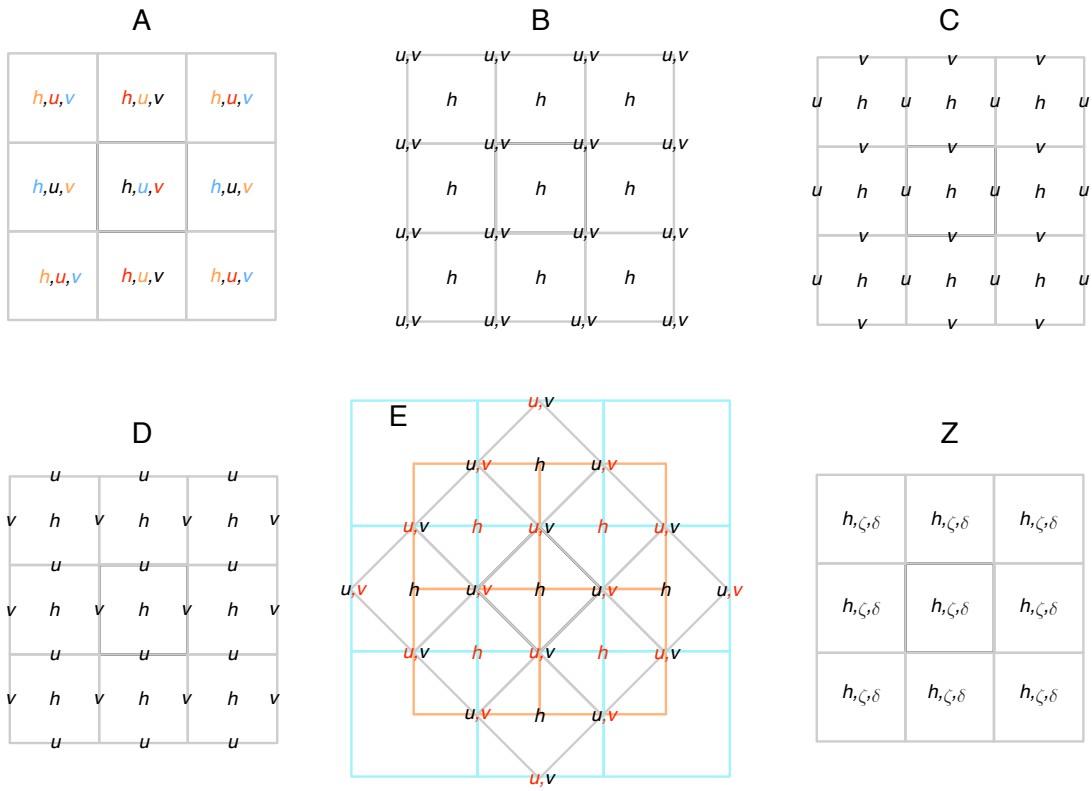



Figure 14.2: Grids A-E and Z, on a square mesh. The A-grid is equivalent to four shifted C-grids. The E-grid is equivalent to two shifted C-grids, which live in the overlapping blue and orange boxes. The E-grid can also be interpreted as a B-grid that has been rotated by 45°, but note that the directions 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 grey boxes.

Z-grid, which will be discussed later. AL also gave the simplest centered finite-difference approximations to of (14.29) - (14.31), for each of the five grids; 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. 14.3, which also gives a plot of the nondimensional frequency, $\sigma/f$, as a function of $kd$ and $ld$, for the special case $\lambda/d = 2$, where $d$ is the grid spacing, assumed to be the same in the $x$ and $y$ directions. The particular choice $\lambda/d = 2$ means that the radius of deformation is larger than the grid spacing, although only a little bit larger. The significance of this choice is discussed later. The plots in Fig. 14.3 show how the nondimensional frequency varies out to $kd = \pi$ and $ld = \pi$; these wave numbers correspond to the shortest waves that can be represented on the grid.

The A-grid may appear to be the simplest, since 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. There is a serious problem, however. In the sketch of the A-grid in Fig. 14.2, some of the variables are colored red, some blue, some orange, and some

black black. The red winds are used to predict the red masses, and vice versa. Similarly, the blue winds interact with the blue masses, the orange winds interact with the orange masses, and the black winds interact with the black masses. There are four independent models, one for each color. This means the the A-grid can support four solutions at once, and those solutions can differ considerably. 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. A plot of the dispersion equation for the A-grid, shown in Fig. 14.3, 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. The problem of the A-grid is obviously closely analogous to that of the unstaggered one-dimensional grid discussed above, but worse.

Next, consider the B-grid. Fig. 14.2 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. As on the A-grid, the coriolis terms are easily evaluated, without averaging, since $u$ and $v$ are defined at the same points. On the other hand, the pressure-gradient terms must be averaged, again as on the A-grid. There is an important difference, however. On the A-grid, the averaging used to approximate the $x$-component of the pressure-gradient force, $\partial h/\partial x$, is averaging *in the x-direction*. On the B-grid, the corresponding averages are in the *y-direction*. On the B-grid, 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. A similar conclusion holds for the convergence / divergence terms of the continuity equation. For example, the averaging in the $y$-direction does no harm for solutions that are uniform in the $y$-direction. Nevertheless, it does do some harm, as is apparent in the plot of the B-grid dispersion equation, as shown in Fig. 14.3. 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.

Now consider the C-grid. The pressure gradient terms are easily evaluated, 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. On the other hand, averaging is needed to obtain the coriolis terms, since $u$ and $v$ are defined at different points. For very small-scale inertia-gravity waves, the coriolis terms are negligible; we essentially have pure gravity waves. This suggests that the C-grid will perform well if the horizontal resolution of the model is high 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 $\lambda$, the radius of deformation. A plot of the dispersion equation,

given in Fig. 14.3, shows that the frequency increases monotonically with wave number, as in the exact solution, although not as rapidly. Recall, however, that this plot is for the special case $\lambda/d = 2$. We return to this point later.

From Fig. 14.1, 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. Fig. 14.2 shows that the two-dimensional A-grid is equivalent to a super-position of four (shifted) two-dimensional C-grids (blue, red, orange, and black).

Next, we turn to the D grid. Inspection of the stencil shown in Fig. 14.2 reveals that the D grid allows a simple evaluation of the geostrophic wind. In view of the importance of geostrophic balance for large-scale motions, this may appear to be an attractive property. 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. 14.3, is very badly behaved, giving zero phase speed for the shortest represented waves, and also giving a zero group speed for some modes.

Finally, consider the the E-grid. As shown in Fig. 14.2, the E-grid can be viewed as a modified B-grid, rotated by 45°. Note, however, that the wind components $u$ and $v$ are *not* rotated. Because the grid has been rotated, the grid spacing for the E-grid is $d^* \equiv \sqrt{2}d$, for the same "density" of $h$ points as in the other four grids. The mass can be considered to live inside the rotated grey cells. The rotation of the grid does not change the meaning of the $u$ and $v$ components of the wind, however, which point in the same directions as before, diagonally across the corners of the grey 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 E-grid at first 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 14.3, the nondimensional frequency for the E-grid is plotted as a function of $kd^*$ and $ld^*$, out to a value of $2\pi$; this corresponds to the shortest "one-dimensional" mode. The group speed is zero for some combinations of $k$ and $l$.

Fig. 14.2 also shows that the A-grid can 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 super-imposed 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.

Now recall the conclusion of AL, mentioned earlier, that the C-grid gives a good simulation of geostrophic adjustment provided that $\lambda/d > 1$. Large-scale modelers are never happy to choose $d$ and $\lambda$ so that $\lambda/d$ can be less than one. Nevertheless, in practice modes for which $\lambda/d << 1$ can be unavoidable, at least for some situations. For example, Hansen et al. (1983) described a low-resolution atmospheric GCM, 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.

Ocean models must contend with small radii of deformation, so that very fine-grids are needed to ensure that $\lambda/d > 1$, even for external modes. For this reason, ocean models tend to use the B-grid (e.g., Semtner and Chervin (1992)).

In addition, three-dimensional models of the atmosphere and ocean generate internal modes. With vertical structures typical of current general 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 $\lambda/d$ can be on the order of 0.1. Fig. 14.4 demonstrates that the C-grid behaves very badly for $\lambda/d = 0.1$. 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. In contrast, the dispersion relation for the B-grid is qualitatively insensitive to the value of $\lambda/d$. The B-grid has moderate problems for $\lambda/d = 2$, but these problems do not become significantly worse for $\lambda/d = 0.1$.

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.*

Now consider an unstaggered grid for the integration of (14.32) - (14.34), which was called the Z-grid by Randall (1994). This grid is also illustrated in Fig. 14.2. 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 (14.32) - (14.34) is the Laplacian, $\nabla^2(\ )$, 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 at the same point as $h$ itself. An unstaggered grid is thus a natural

choice for the numerical integration of (14.32) - (14.34).

Fig. 14.4 shows that the dispersion relation for the Z-grid is very close to that of the C-grid, for $\lambda/d = 2$, but is drastically different for $\lambda/d = 0.1$. Whereas the C-grid behaves very badly for $\lambda/d = 0.1$, the dispersion relation obtained with the Z-grid is qualitatively insensitive to the value of $\lambda/d$; 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.

The discussion presented above suggests that geostrophic adjustment in shallow water is well simulated on an unstaggered grid when the vorticity and divergence equations are used. The vorticity and divergence equations are routinely used in global spectral models, but are rarely used in global finite-difference models. The reason seems to be that it is necessary to solve elliptic equations to obtain the winds from the vorticity and divergence, e.g., to evaluate the advection terms of the nonlinear primitive equations. Experience shows that this is not a major practical problem.

## 14.5 Other meshes

In order to define the grids A-E 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. We now generalize the discussion to include the triangular and hexagonal meshes, applying the definitions consistently in all cases.

The A-grid has the both velocity components co-located with the mass.

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 define and predict two mutually orthogonal components of the horizontal wind vector. 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 vector is 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 use 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 it to have the normal component of the velocity on the edges of all mass-cells.

The generalized D-grid has the tangential velocity component on the edges of mass-

cells.

As with 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 grey cells in the illustration of the E-grid in Fig. 14.2. 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 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. 14.2. It is not possible to create triangular or hexagonal E-grids in this way, so again we find that the E-grid cannot be defined for triangular or hexagonal meshes.

From this point of view, the E-grid is really only compatible with quadrilateral 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.

Table 14.1: 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 |

Table 14.1 lists the numbers of corners and edges per face, on the triangular, square, and hexagonal meshes. Table 14.2 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 per mass point in the horizontal wind for all three mesh shapes.

Table 14.2 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.

Table 14.2: 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 |

## 14.6 Time-differencing schemes for the shallow-water equations

In this section we will consider both space and time differencing for the linearized shallow water equations.

We begin our discussion with the one-dimensional shallow-water equations. The spatial coordinate is $x$, and the single velocity component is $u$. We consider the non-rotating case with $v \equiv 0$. We have divergence (i.e., $\partial u/\partial x$), but no vorticity. Linearizing about a state of rest, the continuous equations are (14.3) and (14.4).

We use a staggered one-dimensional (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 is actually 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:

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n-1}_{j+\frac{1}{2}}}{2\Delta t} + g\left(\frac{h^n_{j+1} - h^n_j}{d}\right) = 0, \tag{14.43}$$

$$\frac{h_j^{n+1}-h_j^{n-1}}{2\Delta t}+H\left(\frac{u^n_{j+\frac{1}{2}}-u^n_{j-\frac{1}{2}}}{d}\right)=0. \tag{14.44}$$

Compare with (14.20) - (14.21). With assumed solutions of the form $u^n{}_j=\hat{u}^n\exp(ikjd)$, $h^n{}_j=\hat{h}^n\exp(ikjd)$ and the usual definition of the amplification factor, we find that

$$\left(\lambda^2-1\right)\hat{u}^n+\lambda\frac{g\Delta t}{d}4i\sin\left(\frac{kd}{2}\right)\hat{h}^n=0, \tag{14.45}$$

$$\lambda\frac{H\Delta t}{d}4i\sin\left(\frac{kd}{2}\right)\hat{u}^n+\left(\lambda^2-1\right)\hat{h}^n=0. \tag{14.46}$$

Non-trivial solutions occur for

$$\left(\lambda^2-1\right)^2+\lambda^2\left(\frac{4c_{GW}\Delta t}{d}\right)^2\sin^2\left(\frac{kd}{2}\right)=0, \tag{14.47}$$

where $c_{GW}\equiv\sqrt{gH}$. 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 (14.47) as a quadratic equation for $\lambda^2$. As a first step, rewrite it as

$$\left(\lambda^2\right)^2+\lambda^2\left(-2+b\right)+1=0, \tag{14.48}$$

where, for convenience, we define

$$b\equiv\left(\frac{4c_{GW}\Delta t}{d}\right)^2\sin^2\left(\frac{kd}{2}\right)\geq 0. \tag{14.49}$$

Obviously, for $\Delta t\to 0$ with fixed $d$ we get $b\to 0$. The solution of (14.41) is

$$\begin{aligned}\lambda^2 &= \frac{-(b-2)\pm\sqrt{(b-2)^2-4}}{2}\\ &= \frac{-(b-2)\pm\sqrt{b(b-4)}}{2}.\end{aligned} \tag{14.50}$$

Inspection of (14.43) shows that for $b \to 0$, we get $|\lambda| \to 1$, as expected. For $\lambda = |\lambda| e^{i\theta}$ we see that

$$|\lambda|^2 [\cos(2\theta) + i\sin(2\theta)] = \frac{-(b-2)\pm\sqrt{b(b-4)}}{2}. \tag{14.51}$$

First consider the case $b \leq 4$. It follows from (14.51) that

$$|\lambda|^2 \cos(2\theta) = -\left(\frac{b-2}{2}\right), \text{ and } |\lambda|^2 \sin(2\theta) = \frac{\pm\sqrt{b(4-b)}}{2}, \text{ for } b \leq 4, \tag{14.52}$$

from which we obtain

$$\tan(2\theta) = \frac{\sqrt{b(4-b)}}{2-b} \text{ for } b \leq 4, \tag{14.53}$$

and

$$|\lambda|^4 = \left(\frac{b-2}{2}\right)^2 + \frac{b(4-b)}{4} = 1 \text{ for } b \leq 4. \tag{14.54}$$

The scheme is thus neutral for $b \leq 4$, as was anticipated from our earlier analysis of the oscillation equation.

Next, consider the case $b > 4$. Returning to (14.44), we find that

$$\sin(2\theta) = 0,\ \cos(2\theta) = \pm 1 \text{ and } |\lambda|^2 = \frac{-(b-2)\pm\sqrt{b(b-4)}}{2} \text{ for } b > 4. \tag{14.55}$$

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

$$\left(\frac{c_{GW}\Delta t}{d}\right)\left|\sin\left(\frac{kd}{2}\right)\right| \leq \frac{1}{2} \tag{14.56}$$

The worst case occurs for $\left|\sin\left(\frac{kd}{2}\right)\right| = 1$, which corresponds to $kd = \pi$, i.e., the shortest wave that can be represented on the grid. It follows that

$$\frac{c_{GW}\Delta t}{d} < \frac{1}{2} \text{ is required for stability,} \tag{14.57}$$

and that the shortest wave will be the first to become unstable.

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., (14.49), 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 (14.36) - (14.37) by

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n}_{j+\frac{1}{2}}}{\Delta t} + g\left(\frac{h^{n+1}_{j+1} - h^{n+1}_{j}}{d}\right) = 0, \tag{14.58}$$

$$\frac{h^{n+1}_{j} - h^{n}_{j}}{\Delta t} + H\left(\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n+1}_{j-\frac{1}{2}}}{d}\right) = 0. \tag{14.59}$$

This leads to

$$(\lambda - 1)\,\hat{u}^n + \lambda \frac{g\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right) \hat{h}^n = 0, \tag{14.60}$$

$$\lambda \frac{H\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right) \hat{u}^n + (\lambda - 1)\,\hat{h}^n = 0. \tag{14.61}$$

The condition for non-trivial solutions is

$$(\lambda - 1)^2 + \lambda^2 4 \left(\frac{c_{GW}\Delta t}{d}\right)^2 \sin^2\left(\frac{kd}{2}\right) = 0, \tag{14.62}$$

which, using the definition (14.49), is equivalent to

$$\lambda^2 \left(1 + \frac{b}{4}\right) - 2\lambda + 1 = 0. \tag{14.63}$$

This time there are no computational modes; the two physical modes satisfy

$$\lambda^2 = \frac{2 \pm \sqrt{4 - 4\left(1 + \frac{b}{4}\right)}}{2\left(1 + \frac{b}{4}\right)} = \frac{1 \pm i\sqrt{\frac{b}{4}}}{1 + \frac{b}{4}}. \tag{14.64}$$

The solutions are always oscillatory, and

$$|\lambda|^2 = \frac{1 + \frac{b}{4}}{\left(1 + \frac{b}{4}\right)^2} = \frac{4}{4 + b} \leq 1, \tag{14.65}$$

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 (14.50) - (14.51) by

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n}_{j+\frac{1}{2}}}{\Delta t} + \frac{g}{d}\left[\left(\frac{h^{n}_{j+1} + h^{n+1}_{j+1}}{2}\right) - \left(\frac{h^{n}_{j} + h^{n+1}_{j}}{2}\right)\right] = 0, \tag{14.66}$$

$$\frac{h^{n+1}_{j} - h^{n}_{j}}{\Delta t} + \frac{H}{d}\left[\left(\frac{u^{n}_{j+\frac{1}{2}} + u^{n+1}_{j+\frac{1}{2}}}{2}\right) - \left(\frac{u^{n}_{j-\frac{1}{2}} + u^{n+1}_{j-\frac{1}{2}}}{2}\right)\right] = 0. \tag{14.67}$$

This leads to

$$(\lambda - 1)\,\hat{u}^{n} + \left(\frac{1+\lambda}{2}\right)\frac{g\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right)\hat{h}^{n} = 0, \tag{14.68}$$

$$\left(\frac{1+\lambda}{2}\right)\frac{H\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right)\hat{u}^{n} + (\lambda - 1)\,\hat{h}^{n} = 0. \tag{14.69}$$

For non-trivial solutions, we need

$$(\lambda - 1)^2 + (1+\lambda)^2\left(\frac{c_{GW}\Delta t}{d}\right)^2 \sin^2\left(\frac{kd}{2}\right) = 0. \tag{14.70}$$

Using (14.49) we can show that this is equivalent to

$$\lambda^2 - 2\lambda\left(\frac{16-b}{16+b}\right) + 1 = 0. \tag{14.71}$$

The solutions are

$$\lambda = \left(\frac{16-b}{16+b}\right) \pm i\sqrt{1 - \left(\frac{16-b}{16+b}\right)^2}. \tag{14.72}$$

It follows that $|\lambda|^2 = 1$ for all modes, i.e., the trapezoidal scheme is unconditionally neutral.

The disadvantage of such 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

$$\frac{u^{n+1}_{j+\frac{1}{2}} - u^{n}_{j+\frac{1}{2}}}{\Delta t} + g\left(\frac{h^{n+1}_{j+1} - h^{n+1}_{j}}{d}\right) = 0, \tag{14.73}$$

$$\frac{h^{n+1}_{j} - h^{n}_{j}}{\Delta t} + H\left(\frac{u^{n}_{j+\frac{1}{2}} - u^{n}_{j-\frac{1}{2}}}{d}\right) = 0. \tag{14.74}$$

This scheme can be called "partially implicit," because the end-of-time-step mass field predicted using (14.74) is used to compute the pressure-gradient force in (14.73). 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 (14.66) and (14.67), we get

$$(\lambda - 1)\,\hat{u}^n + \lambda \frac{g\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right)\hat{h}^n = 0, \tag{14.75}$$

$$\frac{H\Delta t}{d} 2i \sin\left(\frac{kd}{2}\right)\hat{u}^n + (\lambda - 1)\,\hat{h}^n = 0. \tag{14.76}$$

This leads to

$$(\lambda - 1)^2 + 4\lambda\left(\frac{c_{GW}\Delta t}{d}\right)^2 \sin^2\left(\frac{kd}{2}\right) = 0, \tag{14.77}$$

which is equivalent to

$$\lambda^2 + \left(\frac{b}{4} - 2\right)\lambda + 1 = 0. \tag{14.78}$$

The solutions are

$$\lambda = \frac{\left(2 - \frac{b}{4}\right) \pm \sqrt{\left(2 - \frac{b}{4}\right)^2 - 4}}{2} = \left(1 - \frac{b}{8}\right) \pm i\sqrt{\frac{b}{4} - \left(\frac{b}{8}\right)^2}. \tag{14.79}$$

The discriminant is non-negative for

$$b \leq 16, \tag{14.80}$$

which corresponds to

$$\left(\frac{c_{GW}\Delta t}{d}\right)^2 \sin^2\left(\frac{kd}{2}\right) \leq 1. \tag{14.81}$$

It follows that

$$\frac{c_{GW}\Delta t}{d} \leq 1 \text{ is required for stability.} \tag{14.82}$$

The time step can thus be twice as large as with the leapfrog scheme. When (14.73) 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.

## 14.7 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:

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)u+g\frac{\partial h}{\partial x}=0, \tag{14.83}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)v=0, \tag{14.84}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)h+H\frac{\partial u}{\partial x}=0. \tag{14.85}$$

We have also added a velocity component, $v$, in the $y$-direction. The dependent perturbation quantities $u$, $v$, and $h$ are assumed to be constant in $y$. In this sense the problem is one-dimensional, even though $v \neq 0$ is allowed.

Since (15.88) through (14.83) are hyperbolic, we can write them in the normal form discussed earlier in this chapter:

$$\left[\frac{\partial}{\partial t}+(U+c)\frac{\partial}{\partial x}\right]\left(u+\sqrt{\frac{g}{H}}h\right)=0, \tag{14.86}$$

$$\left(\frac{\partial}{\partial t}+U\frac{\partial}{\partial x}\right)v=0, \tag{14.87}$$

$$\left[\frac{\partial}{\partial t}+(U-c)\frac{\partial}{\partial x}\right]\left(u-\sqrt{\frac{g}{H}}h\right)=0. \tag{14.88}$$

Here $c \equiv \sqrt{gH}$. We assume that $c > |U|$, which is often true in the atmosphere. For (14.86) and (14.88), the lines $x-(U+c)t =$ constant and $x-(U-c)t =$ constant are the characteristics, and are shown as the solid lines in Fig. 14.5. Everything is similar to the

case without advection, except that now the slopes of the two characteristics which involve $c$ differ not only in sign but also in magnitude.

We also have an additional equation, namely (14.87). This, of course, is an advection equation, and so $v$ is a constant along the lines $x - Ut =$ constant, which are shown schematically by the broken lines in Fig. 14.5. We should then specify $v$ only on the inflow boundary. The divergence is given by $\frac{\partial u}{\partial x}$, and the vorticity by $\frac{\partial v}{\partial x}$. We conclude that for this one-dimensional case the normal or divergent component of the wind ($u$) can be specified at both boundaries, but the tangential or rotational component ($v$) can be specified only at the inflow boundary.

## 14.8 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.

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.

## 14.9 Problems

1. Derive the dispersion equation for the C-grid, as given in Fig. 14.2.

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 leapfrog time differencing and centered space differencing. Derive the stability criteria for both cases, and compare the two results.

3. Write down differential-difference equations for the linearized (about a resting basic state) one-dimensional shallow water equations without rotation on an *unstaggered* grid (the A-grid), using fourth-order accuracy for the spatial derivatives. (Just use the fourth-order scheme discussed in Chapter 2; you are not required to prove the order of accuracy in this problem.) Perform an analysis to determine whether or not the scheme has computational modes. Compare with the corresponding second-order scheme.

4. 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 leapfrog time differencing. Set $f = 10^{-4}$ s$^{-1}$, $g = 0.1$ m s$^{-1}$, $H = 10^3$ m, and $d = 10^5$ m. In the square region

$$\begin{aligned} 45 \le i \le 55, \\ 45 \le j \le 55, \end{aligned} \tag{14.89}$$

apply a forcing in the continuity equation, of the form

$$\left(\frac{\partial h}{\partial t}\right)_{\text{noise}} = (-1)^{i+j} N \sin(\omega_N t), \tag{14.90}$$

and set $\left(\frac{\partial h}{\partial t}\right)_{\text{noise}} = 0$ 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 forcing to the entire domain of the form

$$\left(\frac{\partial h}{\partial t}\right)_{\text{smooth}} = S \sin\left(\frac{2\pi x}{L}\right) \sin\left(\frac{2\pi y}{L}\right) \sin(\omega_S t) \tag{14.91}$$

with $\omega_S = \frac{2\pi\sqrt{gH}}{L}$ s$^{-1}$ and $S = 10^{-4}$ m s$^{-1}$. Here $L$ is $101 \times d$ the width of the domain. Finally, include friction in the momentum equations, of the form

$$\begin{aligned} \left(\frac{\partial u}{\partial t}\right)_{\text{fric}} &= -Ku, \\ \left(\frac{\partial v}{\partial t}\right)_{\text{fric}} &= -Kv, \end{aligned} \tag{14.92}$$

where $K = 2 \times 10^{-5}$ s$^{-1}$. *Use forward time differencing (instead of leapfrog 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. Your analysis should compare various aspects of the solutions, in light of the discussion given in this chapter.

(c) Repeat your runs using $f = 3 \times 10^{-3}$ s$^{-1}$. Discuss the changes in your results.

5. Show that there is no pressure-gradient term in the vorticity equation for the shallow-water system on the C-grid.

6. 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.

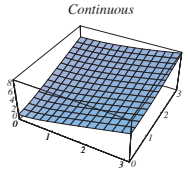
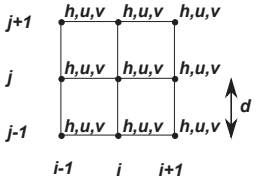
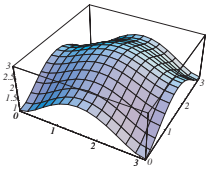
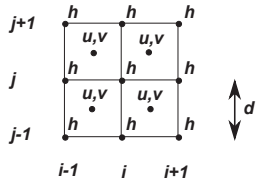
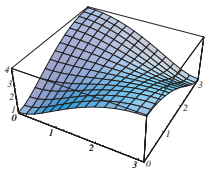
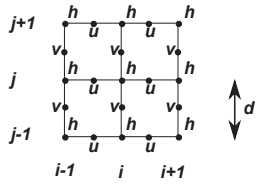
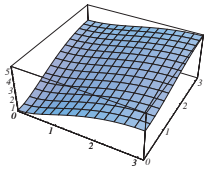
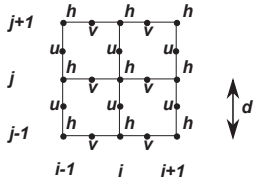
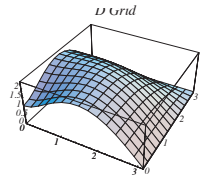
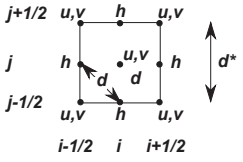
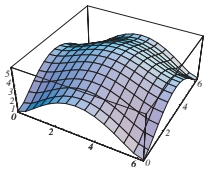
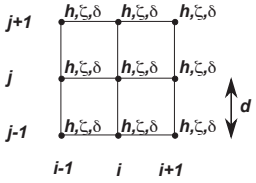
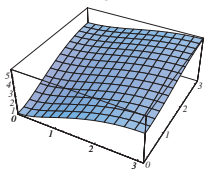
| | | | |
|---|---|---|---|
| | ***Continuous*** | $\left(\frac{\sigma}{f}\right)^2 = 1 + \left(\frac{\lambda}{d}\right)^2\left[(kd)^2 + (ld)^2\right]$ |   |
| |   | $\left(\frac{\sigma}{f}\right)^2 = 1 + \left(\frac{\lambda}{d}\right)^2\left[\sin^2(kd) + \sin^2(ld)\right]$ |  |
| |   | $\left(\frac{\sigma}{f}\right)^2 = 1 + 2\left(\frac{\lambda}{d}\right)^2[1 - \cos(kd)\cos(ld)]$ |  |
| |   | $\left(\frac{\sigma}{f}\right)^2 = ?$<br>$\frac{1}{4}[1 + \cos(kd) + \cos(ld) + \cos(kd)\cos(ld)]$<br>$? + 4\left(\frac{\lambda}{d}\right)^2\left[\sin^2\left(\frac{kd}{2}\right) + \sin^2\left(\frac{ld}{2}\right)\right]$ |  |
| |   | $\left(\frac{\sigma}{f}\right)^2 =$<br>$\frac{1}{4}[1 + \cos(kd) + \cos(ld) + \cos(kd)\cos(ld)]$<br>$+ \left(\frac{\lambda}{d}\right)^2\left[\cos^2(kd)\sin^2(ld) + \cos^2(ld)\sin^2(kd)\right]$ |   |
| |   | $\left(\frac{\sigma}{f}\right)^2 = 1 + 4\left(\frac{\lambda}{d^\times}\right)^2\left[\sin^2\left(\frac{kd^\times}{2}\right) + \sin^2\left(\frac{ld^\times}{2}\right)\right]$ |  |
| |   | $\left(\frac{\sigma}{f}\right)^2 = 1 + 4\left(\frac{\lambda}{d}\right)^2\left[\sin^2\left(\frac{kd}{2}\right) + \sin^2\left(\frac{ld}{2}\right)\right]$ |  |

Figure 14.3: 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 $\lambda/d = 2$.

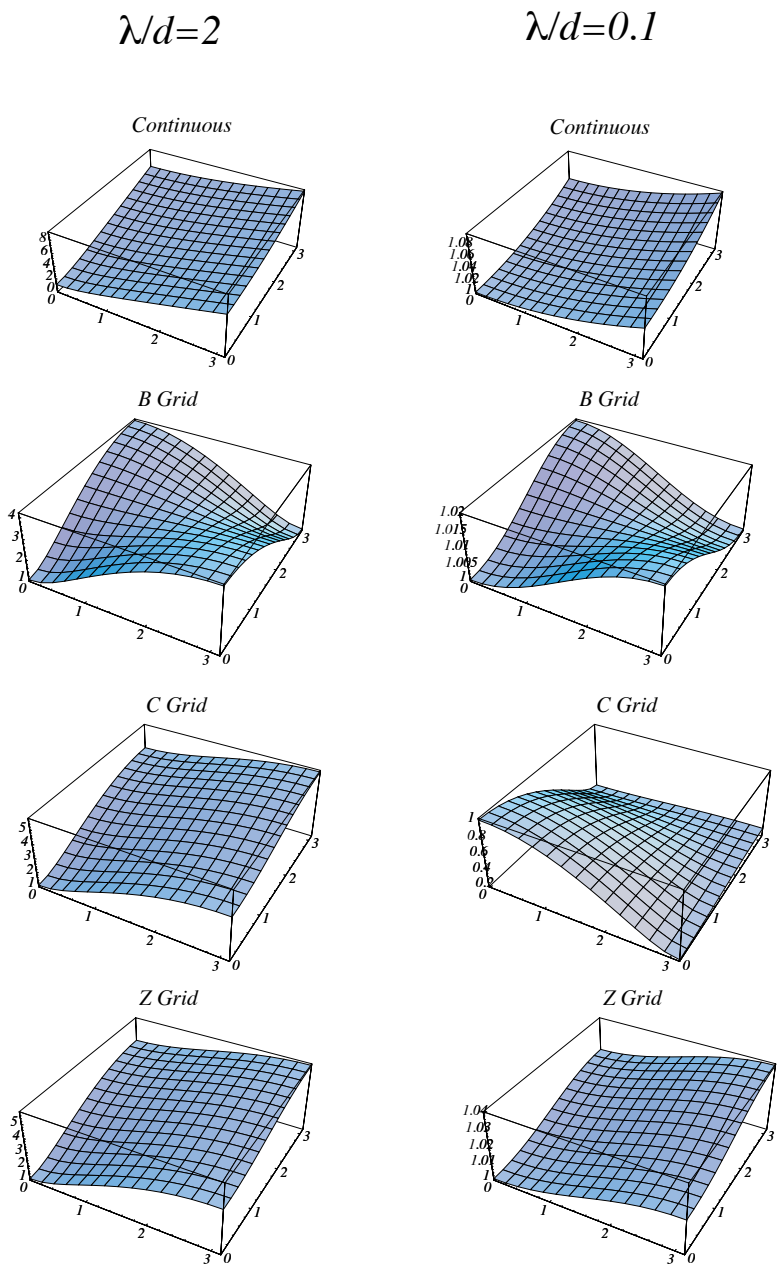



Figure 14.4: Dispersion relations for the continuous shallow water equations, and for finite-difference approximations based on the B, C, and Z-grids. The horizontal coordinates in the plots are $kd$ and $ld$, respectively, except for the E-grid, for which $kd^*$ and $ld^*$ are used. The vertical coordinate is the normalized frequency, $\sigma/f$. For the E-grid, the results are meaningful only in the triangular region for which $kd^* + ld^* \leq 2\pi$. The left column shows results for $\lambda/d = 2$, and the right column for $\lambda/d = 0.1$.

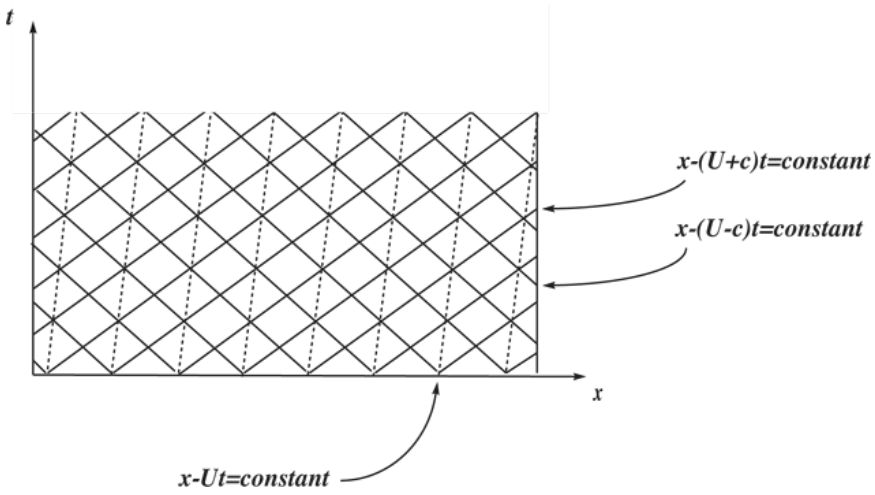



Figure 14.5: Characteristics for the case of shallow water wave propagation with an advecting current $U$.

# Chapter 15

# The Wall

## 15.1 Introduction

Boundary conditions can be real or fictitious, because boundaries can be real or fictitious.

In the simple case of a real wall, the normal component of the velocity vanishes, which implies that no mass crosses the wall. The Earth's surface is a "real wall" at the lower boundary of the atmosphere. 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.

Most numerical models have "fictitious walls" or "lids" at their tops. 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.

## 15.2 Inflow boundaries

Suppose that the initial condition is given only in a certain limited domain. To illustrate, in Fig. 15.1, lines of constant $x-ct$ are shown, for $c>0$. If the initial condition is specified at $t=0$, 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 $x-ct=x_0$, we need an "inflow boundary condition" at $t>0$ at $x=x_0$. 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 $(x_0\leq x\leq x_1)$. If the subsidiary conditions are given so that the solution exists and is determined uniquely, we have a well-posed problem. Note that a boundary condition at $x=x_1$ is of no use.

At an inflow boundary, say at $x=0$, we have to prescribe $A(0,t)$ for all time. As an

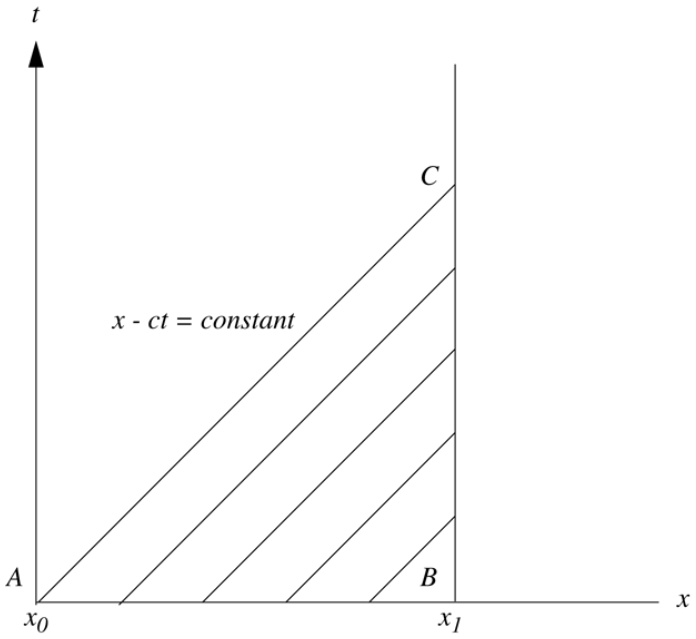



Figure 15.1: An initial condition that is specified between $x = x_0$ and $x = x_1$ determines the solution only along the characteristics shown.

example, suppose that $A(0,t)$ is a simple harmonic function, with frequency $\omega$. The value of $\omega$ is determined by the upstream boundary condition. We can choose $\omega$ any way we please.

Referring again to the advection equation, i.e.,

$$\frac{\partial A}{\partial t} + c\frac{\partial A}{\partial x} = 0, \tag{15.1}$$

we assume $c > 0$ and write

$$A(x,t) = \mathrm{Re}\left[\widehat{A}(x)\, e^{-i\omega t}\right] \text{ for } \omega \neq 0 \tag{15.2}$$

where $\widehat{A}(0)$ is a real constant. Then

$$A(0,t) = \widehat{A}(0)\,\mathrm{Re}\left(e^{-i\omega t}\right) = \widehat{A}(0)\cos\omega t. \tag{15.3}$$

Since $c > 0$, we have effectively prescribed an "inflow" or "upstream" boundary condition. Use of (15.2) in (15.1) gives

$$-i\omega\widehat{A} + c\frac{d\widehat{A}}{dx} = 0, \tag{15.4}$$

which has the solution

$$\widehat{A}(x) = \widehat{A}(0)\, e^{ikx}. \tag{15.5}$$

The dispersion equation is obtained by substituting (15.5) into (15.4):

$$\omega = ck. \tag{15.6}$$

The full solution is thus

$$\begin{aligned} A(x,t) &= \widehat{A}_0 \mathrm{Re}\left\{\exp\left[i\left(kx - \omega t\right)\right]\right\} \\ &= \widehat{A}_0 \mathrm{Re}\left\{\exp\left[ik\left(x - ct\right)\right]\right\}. \end{aligned} \tag{15.7}$$

Now consider the same problem again, this time as represented through the differential-difference equation

$$\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0. \tag{15.8}$$

We assume a solution of the form

$$A_j = \mathrm{Re}\left[\widehat{A}_j e^{-i\omega t}\right], \tag{15.9}$$

we obtain the now-familiar dispersion relation

$$\omega = c\left[\frac{\sin(k\Delta x)}{\Delta x}\right]. \tag{15.10}$$

Compare (15.10) with (15.6). Fig. 15.2 gives a schematic plot, with $\omega\Delta x/c$ and $k\Delta x$ as coordinates, for the true dispersion equation (15.6) and the approximate dispersion equation (15.10). *For a given* $\omega$ we have one $k$ in the exact solution. In the numerical solution, however, we have two $k$s, which we are going to call $k_1$ and $k_2$. As discussed below, for $\omega > 0$, $k_1$ corresponds to the exact solution. The figure makes it clear that

$$k_2\Delta x = \pi - k_1\Delta x. \tag{15.11}$$

The group velocity is positive, as it should be, for $k\Delta x < \pi/2$ and negative for $k\Delta x > \pi/2$.

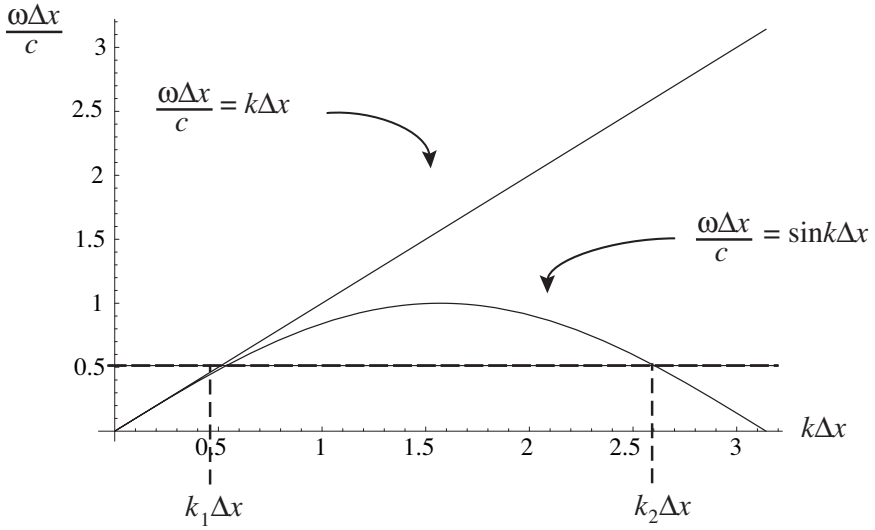



Figure 15.2: A schematic plot, with $\omega\Delta x/c$ and $k\Delta x$ as coordinates, for the true solution (5.2) and the approximate solution (8.5). The dashed line illustrates that, for a given $\omega$, the approximate solution is consistent with two different wave numbers.

When we studied the leapfrog time-differencing scheme in Chapter 4, we found two solutions with different frequencies for a given wavelength. Here we have two solutions with different wavelengths 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 $k\Delta x = \frac{\pi}{2}$). For $k\Delta x > \frac{\pi}{2}$, there really is no physical mode. In other words, *the physical mode exists only for* $L \geq 4\Delta x$.

In view of (15.10), the condition $k_1\Delta x < \frac{\pi}{2}$, which is required for a physical mode to exist, corresponds to $\sin^{-1}\left(\frac{\omega\Delta x}{c}\right) < \frac{\pi}{2}$. This condition can be satisfied by choosing $\Delta x$ small enough, for given values of $\omega$ 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*. Put this way, the conclusion seems like common sense.

Referring back to (15.7), we see that the two modes can be written as

$$\text{Physical mode}: \quad A_j = \widehat{A}_0 \text{Re}\left\{\exp\left[ik_1\left(j\Delta x - \frac{\omega}{k_1}t\right)\right]\right\}. \tag{15.12}$$

$$\begin{aligned}\text{Computational mode}: \quad A_j &= \widehat{A}_0 \text{Re}\left\{\exp\left[ik_2\left(j\Delta x - \frac{\omega}{k_2}t\right)\right]\right\} \\ &= \widehat{A}_0 \text{Re}\left\{\exp\left[i\left(j\pi - k_1 j\Delta x - \omega t\right)\right]\right\} \\ &= (-1)^j \widehat{A}_0 \text{Re}\left\{\exp\left[-ik_1\left(j\Delta x + \frac{\omega}{k_1}t\right)\right]\right\}.\end{aligned} \tag{15.13}$$

Here we have used (15.11) 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.

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. 15.3 shows this schematically for some arbitrary time. The dashed line for $j < 0$ represents (15.13), 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 $\Delta x$), the computational mode will appear as an oscillation from point to point, i.e., a wave of length $2\Delta x$.

According to (15.10), the apparent phase change per grid interval, denoted by $\Omega$, is related to the wave number by

$$\Omega \equiv \frac{\omega\Delta x}{c} = \sin\left(k\Delta x\right). \tag{15.14}$$

With the exact equations, $\Omega = k\Delta x$. Since we control $\omega$, $\Delta x$, and $c$, we effectively control $\Omega$.

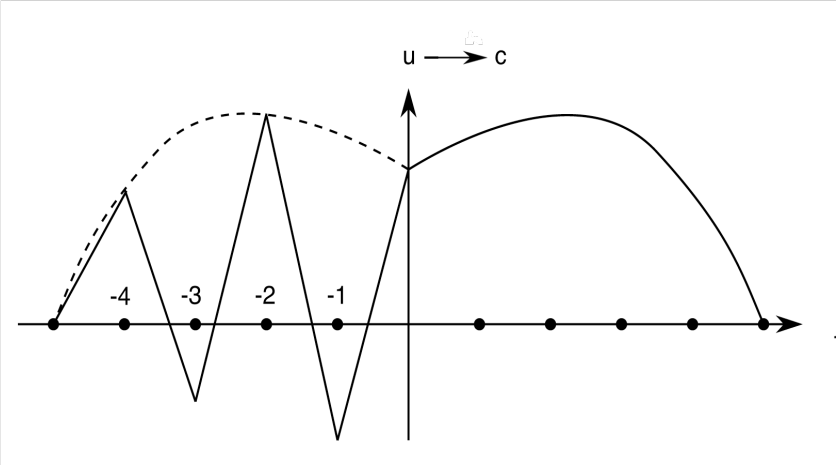



Figure 15.3: Schematic illustration of the solution of the advection equation in the neighborhood of a "smokestack," which acts as a time-dependent source of the advected quantity. In the numerical solution, a computational mode appears in the domain $j < 0$, and a physical mode appears in the domain $j > 0$.

Suppose that we give $\Omega > 1$ by choosing a large value of $\Delta x$. In that case, $k$ has to be complex:

$$k = k_r + ik_i. \tag{15.15}$$

To see what this means, we use the mathematical identity

$$\sin(k\Delta x) = \sin(k_r\Delta x)\cosh(k_i\Delta x) + i\cos(k_r\Delta x)\sinh(k_i\Delta x), \tag{15.16}$$

which can be derived from Euler's formula. Substituting from (15.16) into the right-hand side of (15.14), and equating real and imaginary parts, we find that

$$\begin{aligned} \Omega &= \sin(k_r\Delta x)\cosh(k_i\Delta x), \\ 0 &= \cos(k_r\Delta x)\sinh(k_i\Delta x). \end{aligned} \tag{15.17}$$

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

$$\cos(k_r \Delta x) = 0, \text{ which implies that } k_r \Delta x = \frac{\pi}{2}. \tag{15.18}$$

This is the $4\Delta x$ wave, for which

$$\sin(k_r \Delta x) = 1, \tag{15.19}$$

and so from (15.17) we find that

$$k_i \Delta x = \cosh^{-1}(\Omega) > 0. \tag{15.20}$$

The inequality follows because we have assumed that $\Omega > 1$. We can now write (15.9) as

$$A_j = \widehat{A}_0 \mathrm{Re}\left[e^{-i\omega t} e^{ik_r j\Delta x} e^{-k_i j\Delta x}\right]. \tag{15.21}$$

Since $k_i > 0$, the signal dies out downstream, which is obviously unrealistic but will not make the model crash.

Suppose that we use an uncentered scheme instead of the centered scheme (15.8), e.g.,

$$\frac{dA_j}{dt} + \frac{c}{\Delta x}\left(A_j - A_{j-1}\right) = 0, \tag{15.22}$$

with $c > 0$. This eliminates the computational mode in space. We will show that the uncentered scheme damps regardless of the values of $\omega$, $\Delta x$, and $c$. Let

$$A_j = \widehat{A} e^{-i\omega t} e^{ikj\Delta x}, \tag{15.23}$$

$$A_{j-1} = \widehat{A} e^{-i\omega t} e^{ik(j-1)\Delta x}. \tag{15.24}$$

Then we obtain the dispersion equation

$$-i\omega + \frac{c}{\Delta x}\left(1 - e^{-ik\Delta x}\right) = 0. \tag{15.25}$$

First, suppose that $k$ is real. Setting the real and imaginary parts of (15.25) to zero gives

$$\cos(k\Delta x) = 1, \text{ and } -\omega + ck\left[\frac{\sin(k\Delta x)}{k\Delta x}\right] = 0. \tag{15.26}$$

Since $\cos(k\Delta x) = 1$ implies that $k\Delta x = 0$, this solution is not acceptable. We conclude that $k$ must be complex.

We therefore use (15.15) to obtain

$$-i\omega + \frac{c}{\Delta x}\left(1 - e^{-ik_r\Delta x}e^{k_i\Delta x}\right) = 0. \tag{15.27}$$

Setting the real part to zero gives

$$1 - e^{k_i\Delta x}\cos(k_r\Delta x) = 0, \tag{15.28}$$

and setting the imaginary part to zero gives

$$\Omega + e^{k_i\Delta x}\sin(k_r\Delta x) = 0. \tag{15.29}$$

These two equations can be solved for the two unknowns $k_r$ and $k_i$. Let

$$\begin{aligned} X &\equiv e^{k_i\Delta x}, \\ Y &\equiv k_r\Delta x. \end{aligned} \tag{15.30}$$

We will use $X$ and $Y$ as proxies for $k_r$ and $k_i$, respectively. Eqs. (15.28) and (15.29) become

$$\begin{aligned} 1 - X\cos(Y) &= 0, \\ -\Omega &= X\sin(Y). \end{aligned} \tag{15.31}$$

We conclude that

$$\begin{aligned} X &= \sec Y, \\ -\tan Y &= \Omega, \end{aligned} \tag{15.32}$$

from which it follows that

$$X \equiv e^{k_i \Delta x} = \sec\left[\tan^{-1}(\Omega)\right] > 1. \tag{15.33}$$

From (15.33), we see that $k_i > 0$. Substituting back, we obtain

$$A_j = \widehat{A}\mathrm{Re}\left[e^{-i\omega t} e^{ik_r j\Delta x} e^{-k_i j\Delta x}\right]. \tag{15.34}$$

This shows that, as $j \to \infty$, the signal weakens.

## 15.3 Outflow boundaries

Suppose that we are carrying out our numerical solution of the one-dimensional advection equation over the region between $j = 0$ and $j = J$, as shown in Fig. 15.4, using centered space-differencing with a continuous time derivative, i.e.,

$$\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0, \tag{15.35}$$

and that we are given $A$ at $j = 0$ as a function of time. At $j = 1$ we can write, using centered space differencing,

$$\frac{dA_1}{dt} + c\left(\frac{A_2 - A_0}{2\Delta x}\right) = 0. \tag{15.36}$$

At $j = J-1$ we have

$$\frac{dA_{J-1}}{dt} + c\left(\frac{A_J - A_{J-2}}{2\Delta x}\right) = 0, \tag{15.37}$$

and at $j = J$ we have

$$\frac{dA_J}{dt} + c\left(\frac{A_{J+1} - A_{J-1}}{2\Delta x}\right) = 0. \tag{15.38}$$

Eq. (15.38) 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, or in other words, a "*computational boundary condition*" at the fictitious outflow boundary. Ideally, this artificial boundary condition should not affect the solution in the interior in any way, since its only purpose is to limit (for computational purposes) the size of the domain.


Figure 15.4: An outflow boundary condition must be specified at $j = J$ in this finite and non-periodic domain.

For the continuous advection equation, we need to give and can give boundary conditions only at the inflow point, but for the finite-difference equation we also need a computational boundary condition at the outflow point. With the leapfrog scheme, we needed two initial conditions. The current situation is somewhat analogous. Essentially, both problems arise because of the three-level differences (one in time, the other in space) used in the respective schemes. If the computational boundary condition is not given properly, there is a possibility of exciting a strong computational mode.

Table 15.1: A summary of the computational boundary conditions studied by Nitta.

| Identifier | Form of Scheme |
|---|---|
| Method 1 | $A_J =$ constant in time |
| Method 2 | $A_J^n = A_{J-1}^n$ |
| Method 3 | $\left(\frac{dA}{dt}\right)_J = \left(\frac{dA}{dt}\right)_{J-1}$ |
| Method 4 | $A_J^n = A_{J-2}^n$ |
| Method 5 | $A_J^n = 2A_{J-1}^n - A_{J-2}^n$ |
| Method 6 | $\left(\frac{dA}{dt}\right)_J = 2\left(\frac{dA}{dt}\right)_{J-1} - \left(\frac{dA}{dt}\right)_{J-2}$ |
| Method 7 | $\left(\frac{dA}{dt}\right)_J = -c\left(\frac{A_J^n - A_{J-1}^n}{\Delta x}\right)$ |
| Method 8 | $\left(\frac{dA}{dt}\right)_J = -\frac{c}{2\Delta x}\left(3A_J^n - 4A_{J-1}^n + A_{J-2}^n\right)$ |

Nitta (1964) presented some results of integrating the advection equation with leapfrog time differencing, using various methods of specifying the computational boundary condition. Nitta's paper deals mainly with space differencing, but as discussed later his conclusions are influenced by his choice of leapfrog time differencing. Table 15.1 summarizes the boundary conditions or "Methods" that Nitta considered. The results that he obtained are shown in Fig. 15.5.

With 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.

Method 5, on the other hand, sets $A_J^n = 2A_{J-1}^n - A_{J-2}^n$, a linear extrapolation of the two

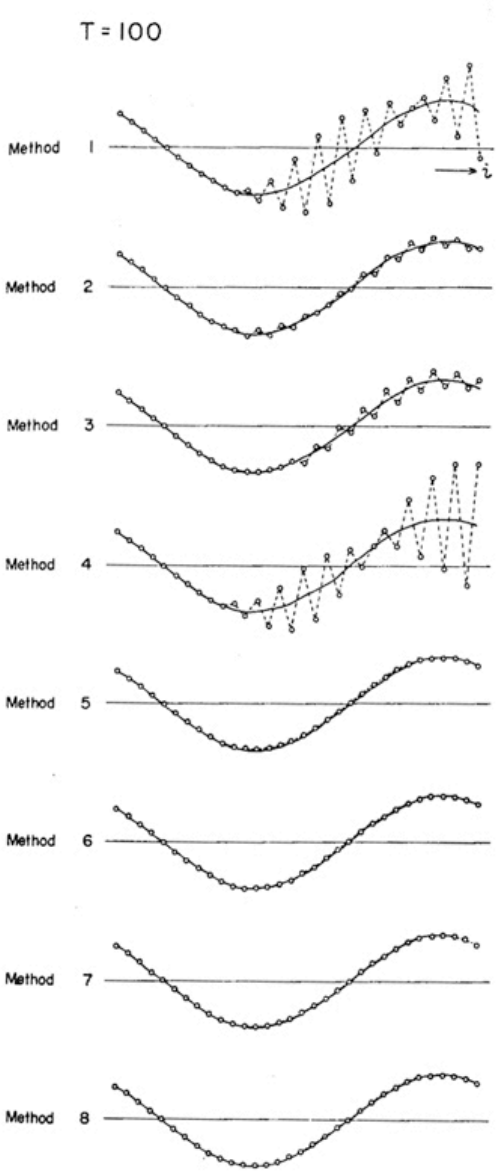



Figure 15.5: A summary of Nitta's numerical results, with various computational boundary conditions. Here leapfrog time differencing was used.

interior points to $A_J^n$. This is equivalent to setting the second derivative to zero at the wall. It can also be interpreted as a linear extrapolation to the wall. Method 6 is similar to Method 5, but uses a linear extrapolation of the time derivative.

Method 7 predicts $A_j$ by means of

$$\frac{dA_J}{dt} + c\left(\frac{A^n{}_J - A^n{}_{J-1}}{\Delta x}\right) = 0, \tag{15.39}$$

which uses uncentered differencing in space. This is equivalent to using $A_{J+1} = 2A_{J-1}$, which is very similar to Method 5. For this reason, Methods 5 and 7 give very similar results.

Method 8 is similar to Method 7, but has higher-order accuracy.

We now perform an analysis to try 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 $A_j \sim \exp\left[ik_1\left(j\Delta x - \frac{\omega}{k_1}t\right)\right]$ and the computational mode is given by $A_j \sim (-1)^j \exp\left[-ik_1\left(j\Delta x + \frac{\omega}{k_1}t\right)\right]$. 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 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 (15.12) and (15.13), we can write

$$A_j = \hat{A}\mathrm{Re}\left\{\exp\left[ik\left(j\Delta x - \frac{\omega}{k}t\right)\right] + r(-1)^j \exp\left[-ik\left(j\Delta x + \frac{\omega}{k}t\right)\right]\right\}, \tag{15.40}$$

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 want 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 (15.40),

$$\begin{aligned} A_J &= \hat{A}\mathrm{Re}\left\{\exp\left[ik\left(J\Delta x - \frac{\omega}{k}t\right)\right] + r\exp\left[-ik\left(J\Delta x + \frac{\omega}{k}t\right)\right]\right\} \\ &= \hat{A}\left[\exp\left(ikJ\Delta x\right) + r\exp\left(-ikJ\Delta x\right)\right]\exp\left(-i\omega t\right) \\ &= 0. \end{aligned} \tag{15.41}$$

Since $e^{-i\omega t} \neq 0$, we conclude that

$$r = \frac{-\exp(ikJ\Delta x)}{\exp(-ikJ\Delta x)} = -\exp(2ikJ\Delta x), \tag{15.42}$$

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. 15.5.

With Method 2, and still assuming that $J$ is even, we put $u_J = u_{J-1}$. Then we obtain

$$\exp(ikJ\Delta x) + r\exp(-ikJ\Delta x) = \exp[ik(J-1)\Delta x] - r\exp[-ik(J-1)\Delta x], \tag{15.43}$$

which leads to $|r| = \tan\frac{k\Delta x}{2}$. 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., very long waves are not falsely reflected. In Fig. 15.5, the incident mode is relatively long.

With Method 5, it turns out that $|r| = \tan^2\left(\frac{k\Delta x}{2}\right)$. Fig. 15.6 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 only shows 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.

With Method 1, the computational mode is "neutral," in the sense that $|r| = 1$. With all of the seven other methods, the computational mode is damped. Recall from Chapter 4 that any damping process is unstable with the leapfrog scheme, i.e., when the leapfrog scheme is used there will in fact be an "energy source" for the computational model. 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 use the leapfrog scheme.

If we do 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 the velocity $c$), but once the noise becomes

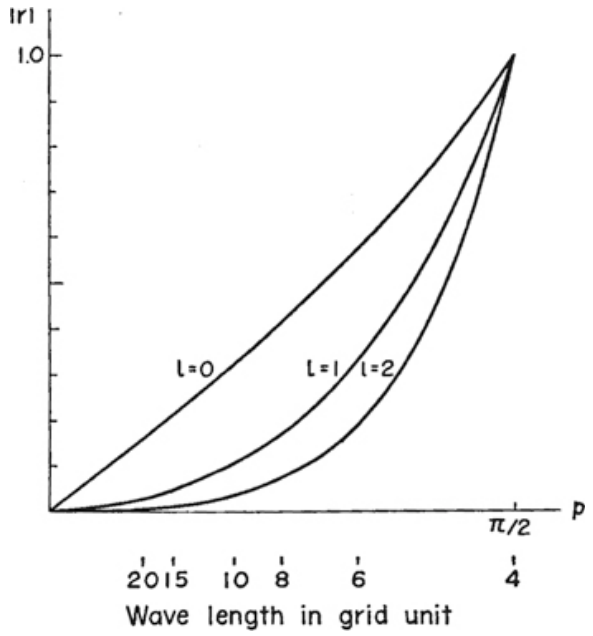



Figure 15.6: A graph of $|r|$ versus $k\Delta x$ for Methods 2 (labeled $l = 0$) and 5 (labeled $l = 1$). The curve labeled $l = 2$ shows the results obtained when the finite-difference approximation to the third derivative is set to 0. From Matsuno (1966).

noticeable, it will amplify and the model will blow up. Why? There is an energy loss in the solution domain through the boundaries when using Method 5 or 7. With Method 1, all of the energy is held, whereas 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 friction term, which was discussed in Chapter 4. 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 time-differencing scheme.

Nitta's schemes can also be analyzed in terms of the energy flux in the neighborhood of the outflow boundary. Multiplying the one-dimensional advection equation by $2A$, we obtain

$$\frac{\partial A^2}{\partial t} + \frac{\partial}{\partial x}\left(cA^2\right) = 0. \tag{15.44}$$

This shows that $A^2$ is advected. Defining $A^2$ as the "energy," we see that $cA^2$ is the energy

flux and $\frac{\partial}{\partial x}\left(cA^2\right)$ is the energy flux divergence. Now suppose that the advection equation is approximated by the differential difference equation:

$$\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0. \tag{15.45}$$

Multiplying (15.45) by $2A_j$, we obtain

$$\frac{d}{dt}A_j^2 + \left(\frac{cA_jA_{j+1} - cA_{j-1}A_j}{\Delta x}\right) = 0. \tag{15.46}$$

Comparing (15.46) with (15.44), we see that $cA_jA_{j+1}$ and $cA_{j-1}A_j$ are the energy fluxes from grid point $j$ to grid point $j+1$, and from grid point $j-1$ to grid point $j$, respectively, as shown in Fig. 15.7. Applying (15.46) to the grid point $j+1$ gives

$$\frac{d}{dt}A_{j+1}^2 + \left(\frac{cA_{j+1}A_{j+2} - cA_jA_{j+1}}{\Delta x}\right) = 0. \tag{15.47}$$

Inspection shows that the energy flux between $j$ and $j+1$ is given by $cA_jA_{j+1}$. In the differential case, the sign of the energy flux is the same as the sign of $c$. 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 $c$. This implies negative ${c_g}^*$ 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.

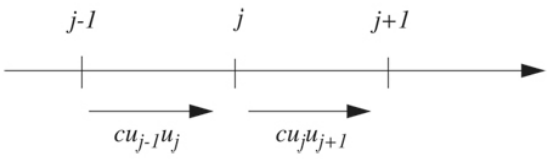



Figure 15.7: Sketch illustrating the energy fluxes that appear in (13.76).

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

$$cA_J A_{J-1} = cA_J^2 > 0. \tag{15.48}$$

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

$$cA_{J-1}A_J = cA_{J-1}A_{J-2} < 0. \tag{15.49}$$

Short-wave energy moves back upstream, and the computational mode is strongly excited.

For Nitta's Method 5,

$$A_J = 2A_{J-1} - A_{J-2}, \tag{15.50}$$

so

$$\begin{aligned} cA_J A_{J-1} &= cA_{J-1}\left(2A_{J-1} - A_{J-2}\right) \\ &= c\left(2A_{J-1}^2 - A_{J-1}A_{J-2}\right). \end{aligned} \tag{15.51}$$

For very short waves,

$$A_{J-1}A_{J-2} < 0, \tag{15.52}$$

so that the flux given by (15.51) is positive, as it should be. For very long waves,

$$A_{J-1}A_{J-2} \cong A_{J-1}A_{J-1}, \tag{15.53}$$

so the flux is approximately

$$cA_JA_{J-1} \cong cA_{J-1}^2 > 0. \tag{15.54}$$

For Nitta's Method 7,

$$\frac{dA_J}{dt} + c\left(\frac{A_J - A_{J-1}}{\Delta x}\right) = 0. \tag{15.55}$$

so we find that

$$\frac{dA_J^2}{dt} + 2c\left(\frac{A_J^2 - A_JA_{J-1}}{\Delta x}\right) = 0. \tag{15.56}$$

The energy flux "into $J$" is $A_JA_{J-1}$, while that "out of $J$" is $A_J^2 > 0$. Applying (15.49) to $J-1$,

$$\frac{d}{dt}\left(A_{J-1}^2\right) + c\left(\frac{A_{J-1}A_J - A_{J-2}A_{J-1}}{\Delta x}\right) = 0. \tag{15.57}$$

This shows that the energy flux out of $J-1$ is the same as the flux into $J$, which is good.

## 15.4 Nested grids

If we use an inhomogeneous grid (one in which the grid size varies), we will encounter a problem similar to the one met at the boundaries; a reflection occurs because the fine portion of the grid permits short waves that cannot be represented on the coarse portion of the grid. This is a serious difficulty with all models that have variable spatial resolution. The problem can be minimized by various techniques, but it cannot be eliminated.

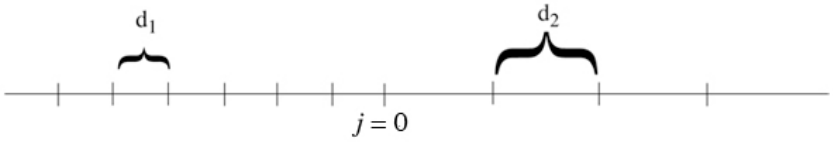



Figure 15.8: Schematic showing a change in the grid size at $j = 0$.

Consider the one-dimensional advection equation with a positive advecting curren $c$. We wish to solve this equation on a grid in which the grid size changes suddenly at $j = 0$, from $d_1$ to $d_2$, in Fig. 15.8. The figure shows $d_2 > d_1$, but we will also consider the opposite case. We use the following differential-difference equations:

$$\begin{gathered}
\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2d_1}\right) = 0, \text{for} j < 0\ , \\
\frac{dA_0}{dt} + c\left[\alpha\left(\frac{A_0 - A_{-1}}{d_1}\right) + \beta\left(\frac{A_1 - A_0}{d_2}\right)\right] = 0, \alpha+\beta = 1, \text{for} j = 0\ , \\
\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2d_2}\right) = 0, \text{for} j > 0.
\end{gathered} \tag{15.58}$$

The method used here for the point $j = 0$ is not completely general, but it allows us to consider various choices for the weights $\alpha$ and $\beta$. If we choose $d_1 = d_2$ and $\alpha = \beta = \frac{1}{2}$, then scheme used at $j = 0$ is the same as the scheme used at all of the other points on the grid.

As a shorthand notation, define

$$p_1 = k_1 d_1, \text{ and } p_2 = k_2 d_2, \tag{15.59}$$

where $k_1$ is wave number of the incoming signal, moving in from the left, and $k_2$ is the wave number for $j \geq 0$. From (15.40), the solution for $j \leq 0$ is given by: fred

$$A_j = e^{i(jp_1 - \omega t)} + r(-1)^j e^{-i(jp_1 + \omega t)}, \tag{15.60}$$

where

$$\omega = c\frac{\sin p_1}{d_1}. \tag{15.61}$$

We have assumed for simplicity that the incident wave has unit amplitude. The solution for $j \geq 0$ is

$$A_j = Re^{i(jp_2 - \omega t)}, \tag{15.62}$$

where

$$\omega = c\frac{\sin p_2}{d_2}, \tag{15.63}$$

and $R$ is the amplitude of the transmitted wave. *The frequency, $\omega$, must be the same throughout the domain*. Eliminating $\omega$ between (15.61) and (15.63), gives

$$\sin p_2 = \frac{d_2}{d_1}\sin p_1. \tag{15.64}$$

This relates $p_2$ to $p_1$, or $k_2$ to $k_1$.

Since the incident wave must have ${c_g}^* > 0$ (this is what is meant by "incident"), we know that

$$0 < p_1 < \frac{\pi}{2}, \tag{15.65}$$

i.e., the wavelength of the incident wave is longer than $4d_1$.

Now consider several cases:

1. $d_2/d_1 > 1$.

   Suppose that advecting current blows from a finer grid onto a coarser grid, which can be expected to cause problems. Define

   $$\sin p_2 = \frac{d_2}{d_1}\sin p_1 \equiv a. \tag{15.66}$$

   This implies that

   $$e^{ip_2} = ia \pm \sqrt{1-a^2}. \tag{15.67}$$

Since we can choose $d_1$, $d_2$, and $k_0$ any way we want, it is possible to make $\sin p_2 > 1$ or $\leq 1$. We consider these two possibilities separately.

(a) $a \equiv \sin p_2 = \frac{d_2}{d_1} \sin p_1 > 1$.

In this case $p_2$ has to be complex. From (15.67), we find that

$$\begin{aligned} e^{ip_2} &= i\left(a \pm \sqrt{a^2-1}\right) \\ &= e^{i\frac{\pi}{2}}\left(a \pm \sqrt{a^2-1}\right). \end{aligned} \tag{15.68}$$

The solution for $j \geq 0$ is then

$$A_j = R\left(a \pm \sqrt{a^2-1}\right)^j e^{i\left(\frac{\pi}{2}j - \omega t\right)} \text{ for } j \geq 0. \tag{15.69}$$

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 must choose the minus sign. Then

$$A_j = R\left(a - \sqrt{a^2-1}\right)^j e^{i\left(\frac{\pi}{2}j - \omega t\right)} \text{ for} j \geq 0. \tag{15.70}$$

This is a damped oscillation. The wavelength is $4d_2$, and the amplitude decays as $j$ increases, as shown in Fig. 15.9.

(b) $a \equiv \sin p_2 = \frac{d_2}{d_1} \sin p_1 \leq 1$.

In this case $p_2$ is real, and

$$|p_2| < \frac{\pi}{2}. \tag{15.71}$$

This implies that $L = \frac{2\pi}{k} > 4d_2$, i.e., the transmitted wave has a wavelength longer than four times the grid spacing. The solution is

$$A_j = Re^{i\left(j\sin^{-1}a - \omega t\right)}. \tag{15.72}$$

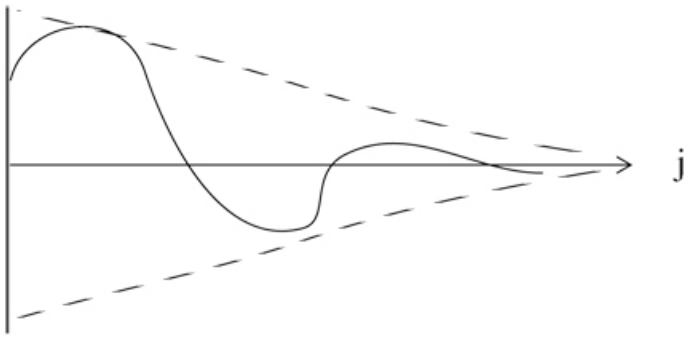



Figure 15.9: Sketch of the solution given by (15.70).

From (15.64), since we are presently considering $d_2/d_1 > 1$, we have $p_2 > p_1$. We also have from (15.64) that

$$\frac{k_2}{k_1} = \frac{\left(\frac{\sin p_1}{p_1}\right)}{\left(\frac{\sin p_2}{p_2}\right)}. \tag{15.73}$$

Recall that $\sin x/x$ 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 less than that of the incident wave, even though it 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, $p_2$ 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 $p_2 \leq \sin^{-1}(d_2/d_1)$, it can be shown that the minimum wavelength of the transmitted wave, which occurs for $p_1 = \frac{\pi}{2}$, is

$$L_{\min} = \frac{2\pi d_2}{\sin^{-1}(d_2/d_1)}. \tag{15.74}$$

As an example, when $d_2/d_1 = 1/2$, the minimum wavelength is $12d_2$.

Next, we find $R$ and $r$. At $j = 0$, the solutions given by (15.60) and (15.62) must agree. It follows that

$$1 + r = R. \tag{15.75}$$

We can substitute (15.60) and (15.62) into the middle equation of (15.58), giving

$$-i\omega R + c\left\{\frac{\alpha}{d_1}\left[1 - e^{-ip_1} + r\left(1 + e^{ip_1}\right)\right] + \frac{\beta}{d_2}R\left(e^{ip_2} - 1\right)\right\} = 0. \tag{15.76}$$

Use (15.75) to eliminate $r$ in (7.24), and solve for $R$:

$$R = \frac{c\frac{\alpha}{d_1}2\cos p_1}{-i\omega + c\left[\frac{\alpha}{d_1}\left(1 + e^{ip_1}\right) + \frac{\beta}{d_2}\left(e^{ip_2} - 1\right)\right]}. \tag{15.77}$$

Now use (15.61) to eliminate $\omega$. Also use $\alpha + \beta = 1$ and (15.64). The result is

$$R = \frac{2\cos p_1}{1 + \cos p_1 - \gamma(1 - \cos p_2)}, \tag{15.78}$$

where we have defined

$$\gamma \equiv \left(\frac{\beta}{\alpha}\right)\left(\frac{d_1}{d_2}\right). \tag{15.79}$$

Substituting (15.78) back into (15.75) gives the reflection coefficient as

$$r = -\left[\frac{1 - \cos p_1 - \gamma(1 - \cos p_2)}{1 + \cos p_1 - \gamma(1 - \cos p_2)}\right]. \tag{15.80}$$

Eq.s (15.78) and (15.80) are basic results. Ideally, we want to have $R = 1$ and $r = 0$.

As a check, suppose that $d_1 = d_2$ and $\alpha = \beta = \frac{1}{2}$. 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 (15.64), we see that for this case $k_2 = k_1$ and $p_1 = p_2$. Then (15.78) and (15.80) give $R = 1$, $r = 0$, i.e., complete transmission and no reflection, as expected. So it works.

For $\alpha \to 0$ with finite $d_1/d_2$, we get $\gamma \to \infty$, $R \to 0$, and $|r| \to 1$, unless $\cos p_2 = 1$, which is the case of an infinitely long wave, i.e., $p_2 = 0$. This is like Nitta's Method 1.

For $\beta \to 0$ with finite $d_1/d_2$, $\gamma \to \infty$, so that

$$\begin{aligned} R &\to \frac{2\cos p_1}{1+\cos p_1} = 1 - \tan^2\left(\frac{p_1}{2}\right), \\ r &\to -\left(\frac{1-\cos p_1}{1+\cos p_1}\right) = -\tan^2\left(\frac{p_1}{2}\right). \end{aligned} \tag{15.81}$$

This is like Nitta's Method 5.

As $p_1$ and $p_2$ both $\to 0$, we get $R \to 1$ and $r \to 0$, regardless of the value of $\gamma$. When $p_1$ and $p_2$ are small but not zero,

$$\cos p_1 \cong 1 - \frac{{p_1}^2}{2},\ \cos p_2 \cong 1 - \frac{{p_2}^2}{2}, \text{ and } p_2 \cong \left(\frac{d_2}{d_1}\right) p_1. \tag{15.82}$$

Then we find that

$$\begin{aligned} R \cong \frac{2\left(1-\frac{{p_1}^2}{2}\right)}{2-\frac{{p_1}^2}{4}+\gamma\frac{{p_2}^2}{4}} &\cong \left(1-\frac{{p_1}^2}{2}\right)\left(1+\frac{{p_1}^2}{4}+\gamma\frac{{p_2}^2}{4}\right) \\ &\cong 1-\frac{{p_1}^2}{4}+\gamma\frac{{p^2}_2}{4} \\ &\cong 1-\frac{{p_1}^2}{4}\left[1-\gamma\left(\frac{d_2}{d_1}\right)^2\right]. \end{aligned} \tag{15.83}$$

Choosing

$$\gamma = (d_1/d_2)^2 \tag{15.84}$$

gives $R = 1 + O\left(p_2^4\right)$. Referring back to (15.64), we see that this choice of $\gamma$ corresponds to

$$\frac{\beta}{\alpha} = \frac{d_1}{d_2}, \text{ or } -\alpha d_1 + \beta d_2 = 0. \tag{15.85}$$

This gives $R$ close to one and $|r|$ close to zero. It can be shown that (15.85) is the requirement for second-order accuracy at the "joint" that connects the two grids. Since the given equations have second-order accuracy elsewhere, (15.85) essentially expresses the requirement that the order of accuracy be spatially homogeneous.

## 15.5 Physical and computational reflection of gravity waves at a wall

At a real, physical wall, the normal component of the velocity has to be zero. If we use the C-grid, for example, we should position and orient the walls so that they are located at wind points, and perpendicular to one of the locally defined velocity components. This means that the walls correspond to the edges of mass boxes on the C-grid. See Fig. 15.10 for a one-dimensional example.

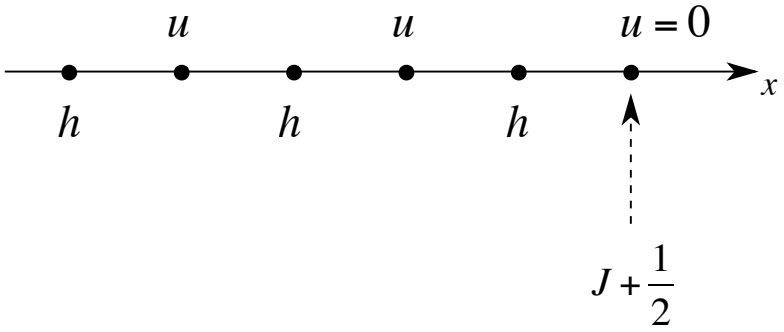



Figure 15.10: A one-dimensional staggered grid for solution of the shallow water equations, near a wall where $j = J + \frac{1}{2}$.

Consider an incident wave traveling toward the right with a certain wave number $k_0$, such that $0 < p_0\,(\equiv k_0 d) < \pi$. Since we are assuming $\sigma \geq 0$, $e^{ip_0 j}e^{-i\sigma t}$ represents such a wave. An additional wave, with $p = -p_0$, can be produced by reflection at a boundary. We assume that the amplitude of the incident wave with $p = p_0$ is 1, and let $R$ denote the amplitude of the reflected wave. In other words, we take $A = 1$ and $B = R$. Then (14.33) and (14.35) can be written as

$$u_{j+\frac{1}{2}} = \left[e^{ip_0\left(j+\frac{1}{2}\right)} + Re^{-ip_0\left(j+\frac{1}{2}\right)}\right] e^{-i\sigma t}, \tag{15.86}$$

$$h_j = -\sqrt{\frac{H}{g}}\left(e^{ip_0 j} - Re^{-ip_0 j}\right)e^{-i\sigma t}. \tag{15.87}$$

Suppose that at $j = J + \frac{1}{2}$ we have a rigid wall (a real, physical wall), as shown in Fig. 15.10. 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 (15.86) reduces to

$$0 = e^{ip_0\left(J+\frac{1}{2}\right)} + Re^{-ip_0\left(J+\frac{1}{2}\right)}, \tag{15.88}$$

which implies that

$$|R| = 1. \tag{15.89}$$

Because $|R| = 1$, the reflected wave has the same amplitude as the incident wave. The reflection is "complete."

The case of a fictitious or "open" boundary is more difficult. We want incident wave to propagate out of the domain, with no reflection. In Chapter 5, we discussed a similar problem for the case of advection towards an open boundary. Obviously, fixing $u_{J+\frac{1}{2}}$ at zero or some other constant value is not going to allow the wave to radiate out of the computational domain. An almost equally simple alternative is to set $u_{J+\frac{1}{2}} = u_{J-\frac{1}{2}}$. The continuity equation shows that $h_J$ would never change in that case.

*** ADD MORE HERE.

## 15.6 Problems

1. Program the discontinuous-grid model given discussed in the text, i.e.,

$$\begin{aligned}
\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2d_1}\right) &= 0, \text{for} j < 0\,, \\
\frac{dA_0}{dt} + c\left[\alpha\left(\frac{A_0 - A_{-1}}{d_1}\right) + \beta\left(\frac{A_1 - A_0}{d_2}\right)\right] &= 0, \alpha+\beta = 1, \text{for} j = 0\,, \\
\frac{dA_j}{dt} + c\left(\frac{A_{j+1} - A_{j-1}}{2d_2}\right) &= 0, \text{for} j > 0,
\end{aligned} \tag{15.90}$$

where

$$-\alpha d_1 + \beta d_2 = 0 \text{ and } \alpha + \beta = 1. \tag{15.91}$$

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 16

# Conservative Schemes for the One-Dimensional Nonlinear Shallow-Water Equations

## 16.1 Properties of the continuous equations

In this chapter, we consider a highly idealized version of the momentum equation: Shallow water, one dimension, no rotation. In a later chapter, we will go to two dimensions with rotation, and bring in the effects of vorticity, which are extremely important.

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 exact equations are

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \tag{16.1}$$

and

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left[K + g\left(h + h_S\right)\right] = 0. \tag{16.2}$$

Here

$$K \equiv \frac{1}{2}u^2 \tag{16.3}$$

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. (16.2), 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 a later chapter.

*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. (16.2) is one possible choice for the continuous form of the momentum equation. An alternative choice is

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(huu) + gh\frac{\partial}{\partial x}(h + h_S) = 0, \tag{16.4}$$

i.e., the flux form of the momentum equation, which can be derived by combining (16.1) and (16.2).

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 (16.1) with respect to $x$, over a closed or periodic domain, we obtain

$$\frac{d}{dt}\left(\int_{\text{domain}} h dx\right) = 0, \tag{16.5}$$

which means that mass is conserved.

Using

$$h\frac{\partial h}{\partial x} = \frac{\partial}{\partial x}\left(\frac{h^2}{2}\right), \tag{16.6}$$

we can rewrite (16.4) as

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}\left(huu + g\frac{h^2}{2}\right) = -gh\frac{\partial h_S}{\partial x}. \tag{16.7}$$

The momentum per unit area is $hu$. If we integrate with respect to $x$, over a periodic domain, we obtain

$$\frac{d}{dt}\left(\int\limits_{\text{domain}} hu dx\right) = -\int\limits_{\text{domain}} gh\frac{\partial h_S}{\partial x} dx. \tag{16.8}$$

This shows that in the absence of topography, i.e., if $\frac{\partial h_S}{\partial x} = 0$ 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 pressure force.

The flux form of the kinetic energy equation can be derived by multiplying (16.1) by $K$ and (16.2) by $hu$, and adding the results, to obtain

$$\frac{\partial}{\partial t}(hK) + \frac{\partial}{\partial x}(huK) + hu\frac{\partial}{\partial x}[g(h+h_S)] = 0. \tag{16.9}$$

The last term of (16.9) represents conversion between potential and kinetic energy.

The potential energy equation can be derived by multiplying (16.1) by $g(h+h_S)$ to obtain

$$\frac{\partial}{\partial t}\left[hg\left(h_S+\frac{1}{2}h\right)\right] + g(h+h_S)\frac{\partial}{\partial x}(hu) = 0, \tag{16.10}$$

or

$$\frac{\partial}{\partial t}\left[hg\left(h_S+\frac{1}{2}h\right)\right] + \frac{\partial}{\partial x}[hug(h+h_S)] - hu\frac{\partial}{\partial x}[g(h+h_S)] = 0. \tag{16.11}$$

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 (16.9). In deriving (16.10), we have assumed that $h_S$ is independent of time. This assumption can easily be relaxed, at the cost of an additional term in (16.11).

When we add (16.9) and (16.11), the energy conversion terms cancel, and we obtain a statement of the conservation of total energy, i.e.,

$$\frac{\partial}{\partial t}\left\{h\left[K+g\left(h_S+\frac{1}{2}h\right)\right]\right\}+\frac{\partial}{\partial x}\left\{hu\left[K+g\left(h+h_S\right)\right]\right\}=0. \tag{16.12}$$

The integral of (16.12) over a closed or periodic domain gives

$$\frac{d}{dt}\int\limits_{\text{domain}} h\left[K+g\left(h_S+\frac{1}{2}h\right)\right]dx=0, \tag{16.13}$$

which shows that the domain-integrated total energy is conserved.

## 16.2 The spatially discrete case

Now consider finite-difference approximations to (16.1) and (16.2). *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, $\Delta x$, 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

$$\frac{dh_i}{dt}+\left[\frac{(hu)_{i+\frac{1}{2}}-(hu)_{i-\frac{1}{2}}}{\Delta x}\right]=0. \tag{16.14}$$

It should be understood that

$$h_{i+\frac{1}{2}}u_{i+\frac{1}{2}}\equiv(hu)_{i+\frac{1}{2}}. \tag{16.15}$$

The "wind-point masses," e.g., $h_{i+\frac{1}{2}}$, 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 (16.14) 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 (16.14) makes it possible for the model to conserve mass, i.e.,

$$\frac{d}{dt}\left(\sum_{\text{domain}} h_i\right) = 0, \tag{16.16}$$

and this is true regardless of how $h_{i+\frac{1}{2}}$ is defined. Eq. (16.16) is analogous to (16.5).

A finite-difference momentum equation that is modeled after (16.2) is

$$\frac{du_{i+\frac{1}{2}}}{dt} + \left(\frac{K_{i+1} - K_i}{\Delta x}\right) + g\left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x}\right] = 0. \tag{16.17}$$

The kinetic energy per unit mass, $K_i$, is undefined at this stage, but resides at mass points. The finite-difference approximations used in (16.17) 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. (16.17) does not have 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 is discussed further below.

Multiply (16.17) by $h_{i+\frac{1}{2}}$ to obtain

$$h_{i+\frac{1}{2}}\frac{du_{i+\frac{1}{2}}}{dt} + h_{i+\frac{1}{2}}\left(\frac{K_{i+1} - K_i}{\Delta x}\right) + gh_{i+\frac{1}{2}}\left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x}\right] = 0. \tag{16.18}$$

In order to mimic the differential relationship (16.6), we must require that

$$h_{i+\frac{1}{2}}\left(\frac{h_{i+1} - h_i}{\Delta x}\right) = \left(\frac{h^2{}_{i+1} - h_i{}^2}{2\Delta x}\right), \tag{16.19}$$

which can by satisfied by choosing

$$h_{i+\frac{1}{2}} = \frac{h_{i+1} + h_i}{2}. \tag{16.20}$$

This choice is required for ensure that the pressure-gradient force does not produce any net source or sink of momentum in the absence of topography. In view of (16.20), we can write

$$(hu)_{i+\frac{1}{2}} = \left(\frac{h_{i+1}+h_i}{2}\right) u_{i+\frac{1}{2}}. \tag{16.21}$$

Combining (16.20) with the continuity equation (16.14), we see that we can write a *continuity equation for the wind points*, as follows:

$$\frac{dh_{i+\frac{1}{2}}}{dt} + \frac{1}{2\Delta x}\left[(hu)_{i+\frac{3}{2}} - (hu)_{i-\frac{1}{2}}\right] = 0. \tag{16.22}$$

It should be clear from the form of (16.22) that the "wind-point mass" is actually conserved by the model. Of course, we do not actually use (16.22) when we integrate the model; instead we use (16.14). Nevertheless, (16.22) will be satisfied, because it can be derived from (16.14) and (16.20). An alternative form of (16.22) is

$$\frac{dh_{i+\frac{1}{2}}}{dt} + \frac{1}{\Delta x}\left[(hu)_{i+1} - (hu)_i\right] = 0, \tag{16.23}$$

where

$$(hu)_{i+1} \equiv \frac{1}{2}\left[(hu)_{i+\frac{3}{2}} + (hu)_{i+\frac{1}{2}}\right] \text{ and } (hu)_i \equiv \frac{1}{2}\left[(hu)_{i+\frac{1}{2}} + (hu)_{i-\frac{1}{2}}\right]. \tag{16.24}$$

Now add (16.18) and $u_{i+\frac{1}{2}}$ times (16.23), and use (16.20), to obtain what "should be" the flux form of the momentum equation, analogous to (16.4):

$$\begin{aligned} &\frac{d}{dt}\left(h_{i+\frac{1}{2}}u_{i+\frac{1}{2}}\right) + h_{i+\frac{1}{2}}\left(\frac{K_{i+1}-K_i}{\Delta x}\right) + \frac{u_{i+\frac{1}{2}}\left[(hu)_{i+1} - (hu)_i\right]}{\Delta x} + g\left(\frac{h_{i+1}^2 - h_i^2}{2\Delta x}\right) \\ &= -gh_{i+\frac{1}{2}}\left[\frac{(h_S)_{i+1} - (h_S)_i}{\Delta x}\right]. \end{aligned} \tag{16.25}$$

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

$$K_i \equiv \frac{1}{2} u_{i+\frac{1}{2}} u_{i-\frac{1}{2}}. \tag{16.26}$$

Other possible definitions of $K_i$ will be discussed later. Using (16.26) and (16.24), we can write

$$\begin{aligned}
&h_{i+\frac{1}{2}} \left( \frac{K_{i+1} - K_i}{\Delta x} \right) + u_{i+\frac{1}{2}} \frac{1}{\Delta x} \left[ (hu)_{i+1} - (hu)_i \right] \\
&= \frac{1}{2\Delta x} \left\{ h_{i+\frac{1}{2}} \left( u_{i+\frac{3}{2}} u_{i+\frac{1}{2}} - u_{i+\frac{1}{2}} u_{i-\frac{1}{2}} \right) + u_{i+\frac{1}{2}} \left[ (hu)_{i+\frac{3}{2}} - (hu)_{i-\frac{1}{2}} \right] \right\} \\
&= \frac{1}{\Delta x} \left[ \left( \frac{h_{i+\frac{1}{2}} + h_{i+\frac{3}{2}}}{2} \right) u_{i+\frac{3}{2}} u_{i+\frac{1}{2}} - \left( \frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2} \right) u_{i-\frac{1}{2}} u_{i+\frac{1}{2}} \right].
\end{aligned} \tag{16.27}$$

This is a flux form. The momentum flux at the point $i$ is $\left( \frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2} \right) u_{i-\frac{1}{2}} u_{i+\frac{1}{2}}$, and the momentum flux at the point $i+1$ is $\left( \frac{h_{i+\frac{1}{2}} + h_{i+\frac{3}{2}}}{2} \right) u_{i+\frac{3}{2}} u_{i+\frac{1}{2}}$. Because (16.27) is a flux form, momentum will be conserved by the scheme *if we define the kinetic energy by* (16.26).

Note, however, that there are two problems with (16.26). When $u$ is dominated by the $2\Delta x$-mode, (16.26) will give a negative value of $K_i$, which is unphysical. In addition, when $u$ is dominated by the $2\Delta x$-mode, the momentum flux will always be negative for the $2\Delta x$-mode, 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 (16.26) can be considered non-viable.

OK, that's not good, but let's see what happens with the kinetic energy equation. For this purpose, we return to general form of $K_i$; Eq. (16.26) will not be used. Recall that the kinetic energy is defined at mass points. To begin the derivation, multiply (16.17) by $(hu)_{i+\frac{1}{2}}$ to obtain

$$(hu)_{i+\frac{1}{2}} \frac{du_{i+\frac{1}{2}}}{dt} + (hu)_{i+\frac{1}{2}} \left( \frac{K_{i+1} - K_i}{\Delta x} \right) + g(hu)_{i+\frac{1}{2}} \left[ \frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x} \right] = 0. \tag{16.28}$$

Rewrite (16.28) for grid point $i - \frac{1}{2}$, simply by subtracting one from each subscript:

$$(hu)_{i-\frac{1}{2}}\frac{du_{i-\frac{1}{2}}}{dt}+(hu)_{i-\frac{1}{2}}\left(\frac{K_i-K_{i-1}}{\Delta x}\right)+g(hu)_{i-\frac{1}{2}}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]=0. \quad (16.29)$$

Now add (16.28) and (16.29), and multiply the result by $\frac{1}{2}$ to obtain the arithmetic mean:

$$\begin{aligned}
&\frac{1}{2}\left[(hu)_{i+\frac{1}{2}}\frac{du_{i+\frac{1}{2}}}{dt}+(hu)_{i-\frac{1}{2}}\frac{du_{i-\frac{1}{2}}}{dt}\right]\\
&+\frac{1}{2}\left[(hu)_{i+\frac{1}{2}}\left(\frac{K_{i+1}-K_i}{\Delta x}\right)+(hu)_{i-\frac{1}{2}}\left(\frac{K_i-K_{i-1}}{\Delta x}\right)\right]\\
&+\frac{g}{2}\left\{(hu)_{i+\frac{1}{2}}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-\frac{1}{2}}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}=0.
\end{aligned} \quad (16.30)$$

You should be able to see that this is an advective form of the kinetic energy equation.

Next we try to derive, from (16.30) and (16.14), a *flux form* of the kinetic energy equation. If we can do that, then kinetic energy will be conserved. Begin by multiplying (16.14) by $K_i$:

$$K_i\left\{\frac{dh_i}{dt}+\left[\frac{(hu)_{i+\frac{1}{2}}-(hu)_{i-\frac{1}{2}}}{\Delta x}\right]\right\}=0. \quad (16.31)$$

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 (16.31) and (16.30) to obtain

$$\begin{aligned}
&K_i\frac{dh_i}{dt}+\frac{1}{2}\left[(hu)_{i+\frac{1}{2}}\frac{du_{i+\frac{1}{2}}}{dt}+(hu)_{i-\frac{1}{2}}\frac{du_{i-\frac{1}{2}}}{dt}\right]\\
&+\left\{\frac{(hu)_{i+\frac{1}{2}}}{\Delta x}\left[K_i+\frac{1}{2}(K_{i+1}-K_i)\right]-\frac{(hu)_{i-\frac{1}{2}}}{\Delta x}\left[K_i-\frac{1}{2}(K_i-K_{i-1})\right]\right\}\\
&+\frac{g}{2}\left\{(hu)_{i+\frac{1}{2}}\frac{\left[(h+h_S)_{i+1}-(h+h_S)_i\right]}{\Delta x}+(hu)_{i-\frac{1}{2}}\frac{\left[(h+h_S)_i-(h+h_S)_{i-1}\right]}{\Delta x}\right\}=0.
\end{aligned} \quad (16.32)$$

Eq. (16.32) "should be" a flux form of the kinetic energy equation. Is it really? To answer this question, we analyze the various terms of (16.32) one by one.

The advection terms on the second line of (16.32) are very easy to deal with. They can be rearranged to

$$\frac{1}{\Delta x}\left[(hu)_{i+\frac{1}{2}}\frac{1}{2}(K_{i+1}+K_i)-(hu)_{i-\frac{1}{2}}\frac{1}{2}(K_i+K_{i-1})\right]. \tag{16.33}$$

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 (16.32), i.e.,

$$\frac{g}{2}\left\{(hu)_{i+\frac{1}{2}}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-\frac{1}{2}}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}. \tag{16.34}$$

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 (16.14) by $g(h+h_S)_i$:

$$\frac{d}{dt}\left[h_i g\left(h_S+\frac{1}{2}h\right)_i\right]+g(h+h_S)_i\left[\frac{(hu)_{i+\frac{1}{2}}-(hu)_{i-\frac{1}{2}}}{\Delta x}\right]=0. \tag{16.35}$$

Eq. (16.35) is analogous to (16.10). We want to recast (16.35) so that we see advection of potential energy, as well as the energy conversion term corresponding to (16.34); compare with (16.11). To accomplish this, we "put in" the energy conversion term by hand, and write the advection term symbolically, like this:

$$\begin{aligned}&\frac{d}{dt}\left[h_i g\left(h_S+\frac{1}{2}h\right)_i\right]+ADV_i\\&-\frac{g}{2}\left\{(hu)_{i+\frac{1}{2}}\left[\frac{(h+h_S)_{i+1}-(h+h_S)_i}{\Delta x}\right]+(hu)_{i-\frac{1}{2}}\left[\frac{(h+h_S)_i-(h+h_S)_{i-1}}{\Delta x}\right]\right\}=0.\end{aligned} \tag{16.36}$$

Here "$ADV_i$" represents the advection of potential energy, 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 (16.36) is a copy of the energy conversion terms of (16.32), but with the sign reversed. We require that (16.36) be equivalent to (16.35), and ask what form of $ADV_i$ is implied by this requirement. The answer is:

$$ADV_i = \frac{g}{2\Delta x}\left\{(hu)_{i+\frac{1}{2}}\left[(h+h_S)_{i+1}+(h+h_S)_i\right]-(hu)_{i-\frac{1}{2}}\left[(h+h_S)_i+(h+h_S)_{i-1}\right]\right\}. \tag{16.37}$$

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 have both turned out to be 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 (16.32). Obviously, the first line of (16.32) must be analogous to $\frac{\partial}{\partial t}(hK)$. For convenience, we define

$$(\text{KE tendency})_i \equiv K_i\frac{dh_i}{dt}+\frac{1}{2}\left[(hu)_{i+\frac{1}{2}}\frac{d}{dt}u_{i+\frac{1}{2}}+(hu)_{i-\frac{1}{2}}\frac{d}{dt}u_{i-\frac{1}{2}}\right]. \tag{16.38}$$

Substituting for the mass fluxes from (16.21), we can write (16.38) as

$$(\text{KE tendency})_i \equiv K_i\frac{dh_i}{dt}+\frac{1}{8}\left[(h_{i+1}+h_i)\frac{d}{dt}\left(u^2{}_{i+\frac{1}{2}}\right)+(h_i+h_{i-1})\frac{d}{dt}\left(u^2{}_{i-\frac{1}{2}}\right)\right]. \tag{16.39}$$

The requirement for kinetic energy conservation is

$$\sum_{\text{domain}}(\text{KE tendency})_i = \sum_{\text{domain}}\frac{d}{dt}(h_iK_i). \tag{16.40}$$

Note that *only the sums over i must agree*; it is not necessary that

$$K_i \frac{dh_i}{dt} + \frac{1}{8}\left[(h_{i+1}+h_i)\frac{d}{dt}\left(u^2{}_{i+\frac{1}{2}}\right) + (h_i+h_{i-1})\frac{d}{dt}\left(u^2{}_{i-\frac{1}{2}}\right)\right] \text{ be equal to } \frac{d}{dt}(h_i K_i) \text{ for each } i. \tag{16.41}$$

To complete our check of kinetic energy conservation, we substitute for $K_i$ on the right-hand side of (16.40), and check to see whether the resulting equation is actually satisfied.

The bad news is that, if we use (16.26), Eq. (16.40) is not satisfied. This means that, when we start from the continuous form of (16.2), we cannot have both momentum conservation under advection and kinetic energy conservation. On the other hand, we didn't like (16.26) 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 (16.40). Two alternative definitions of the kinetic energy are

$$K_i \equiv \frac{1}{4}\left(u^2{}_{i+\frac{1}{2}} + u^2{}_{i-\frac{1}{2}}\right), \tag{16.42}$$

$$h_i K_i \equiv \frac{1}{4}\left(h_{i+\frac{1}{2}} u^2{}_{i+\frac{1}{2}} + h_{i-\frac{1}{2}} u^2{}_{i-\frac{1}{2}}\right). \tag{16.43}$$

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 (16.42) is equal to the sum over the domain of $h_i K_i$ given by (16.43). Either choice allows (16.40) 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 (16.2), 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 (16.4). In that case, we can conserve *both* momentum under advection and kinetic energy under advection. You are asked to demonstrate this in Problem 2 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.

## 16.3 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 introduced several new things. This is the first time that we have considered the advection terms of the momentum equation. 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.

## 16.4 Problems

1. Show that if we use (16.26) it is not possible to conserve kinetic energy under advection.

2. Starting from a finite-difference form that mimics (16.4), show that it is possible to conserve both momentum and total energy. Use the C grid, and keep the time derivatives continuous.

# Chapter 17

# Stairways to Heaven

## 17.1 Introduction to vertical coordinate systems

Up to now we have ignored the vertical structure of the atmosphere. Simulating the vertical structure is very different from simulating the horizontal structure, for three reasons.

- Gravitational effects strongly control vertical motions, and gravitational potential energy is an important 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. In models, the strong vertical gradients require a small vertical grid spacing. The small vertical grid spacing can require small 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.

The most obvious choice of vertical coordinate system, and one of the least useful, is height. As you probably already know, the equations of motion are frequently expressed using vertical coordinates other than height. The most basic requirement for a variable to be used as a vertical coordinate is that it vary monotonically with height. Even this requirement can be relaxed; e.g., 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 condition (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 (simpler is better);
- the intensity of the "vertical motion" as seen in the coordinate system (weaker vertical motion is simpler and better);
- the method used to compute the vertical motion (simpler is better).

Each of these factors will be discussed below, for specific vertical coordinates. We begin, however, by presenting the basic governing equations, *for quasi-static motions*, using a general vertical coordinate.

To construct a vertically discrete 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 exact equations that we want the discrete equations to mimic.

As usual, these choices will involve trade-offs. Each possible choice will have strengths and weaknesses. We must also be aware of possible interactions between the vertical differencing and the horizontal and temporal differencing.

This chapter deals with the first two choices above, which can be discussed in the context of the continuous system. The next chapter deals with the remaining two choices.

## 17.2 Choice of equation set

The speed of sound in the Earth's atmosphere is about 300 m $s^{-1}$. If we permit 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, if a model has 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.

There are four possible ways around this problem. One approach is to use a set of equations that filters sound waves, i.e., "anelastic" equations. There are several varieties of anelastic systems, developed over a period of forty years or so (Ogura and Phillips (1962); Lipps and Hemler (1982); Durran (1989); Bannon (1996); Arakawa and Konor (2009)). Some 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. The most recent formulations are quite

accurate. Anelastic models of various types have been very widely used, especially for high-resolution modeling.

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 is 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. The main disadvantage is complexity.

The fourth approach is to "sub-cycle." Small time steps can be used to integrate the terms of the equations that govern sound waves, while longer time steps are used for the remaining terms.

*** Add lecture on "sound-proof" systems.

## 17.3 The basic equations in height coordinates

The basic equations in height coordinates, without rotation and friction, are

$$\frac{D\mathbf{V}_h}{Dt} + 2\boldsymbol{\Omega} \times \mathbf{V}_h = -\frac{1}{\rho}\nabla_z p - \mathbf{F}_h, \tag{17.1}$$

$$\frac{Dw}{Dt} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g - F_v, \tag{17.2}$$

$$\left(\frac{\partial \rho}{\partial t}\right)_z + \nabla_z \cdot (\rho \mathbf{V}_h + \frac{\partial}{\partial z}(\rho w) = 0, \tag{17.3}$$

$$\dot{\theta} \equiv \frac{D\theta}{Dt} = \frac{Q}{\Pi}. \tag{17.4}$$

Here $\frac{D\mathbf{V}_h}{Dt}$ is the Lagrangian time derivative, $\mathbf{V}_h$ is the horizontal velocity, $\rho$ is density, $p$ is pressure, $z$ is height, $w$ is the vertical velocity, and $\mathbf{F}_h$ is the horizontal friction vector. In (17.2), $g$ is the acceleration of gravity. In (17.4), $\theta$ is the potential temperature, $Q$ is the heating rate per unit mass, and $\Pi$ is the Exner function, which satisfies

$$c_p T = \Pi\theta, \tag{17.5}$$

where $c_p$ is the heat capacity of air at constant pressure, $T$ is temperature, and

$$\Pi = c_p\left(\frac{p}{p_0}\right)^{\kappa}, \tag{17.6}$$

where

$$\kappa \equiv \frac{R}{c_p}, \tag{17.7}$$

and $R$ is the specific gas constant. We will also need the ideal gas law, which can be written as

$$p = \rho RT. \tag{17.8}$$

Finally, we include a prognostic equation for an arbitrary intensive scalar $A$:

$$\left[\frac{\partial}{\partial t}(\rho A)\right]_z + \nabla_z \cdot (\rho \mathbf{V}_h A) + \frac{\partial}{\partial z}(\rho w A) = \rho S_A. \tag{17.9}$$

Here $S_A$ is a source of $A$, per unit mass.

## 17.4 Transformation to generalized 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 section, we derive the *nonhydrostatic* system of equations using a general vertical coordinate. The quasi-static limit is easily recovered as a special case.

Consider an arbitrary vertical coordinate denoted by $\hat{z}$. We assume that $\hat{z}$ is a monotonic function of $z$, and in the analysis below we also assume that $\hat{z}$ can be differentiated with respect to time and spatial coordinates. The dimensions of $\hat{z}$ can be different from those of $z$. For example, $\hat{z}$ could be pressure or potential temperature.

As shown in Appendix B, the horizontal pressure gradient force can be transformed to the generalized coordinate as follows:

$$\begin{aligned}\frac{1}{\rho}\nabla_z p &= \frac{1}{\rho}\nabla_{\hat{z}} p - \frac{1}{\rho}\frac{\partial p}{\partial z}\nabla_{\hat{z}} z \\ &= \frac{1}{\rho}\nabla_{\hat{z}} p - \frac{1}{\rho}\frac{\partial \hat{z}}{\partial z}\frac{\partial p}{\partial \hat{z}}\nabla_{\hat{z}} z \\ &= \frac{1}{\rho}\nabla_{\hat{z}} p - \frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}}\nabla_{\hat{z}} z,\end{aligned} \tag{17.10}$$

where

$$\hat{\rho} \equiv \rho\frac{\partial z}{\partial \hat{z}} \tag{17.11}$$

is the "pseudodensity" in $\hat{z}$ coordinates, i.e., the amount of mass (per unit horizontal area) between two $\hat{z}$-surfaces. Note that

$$\frac{1}{\rho}\frac{\partial p}{\partial z} = \frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}}. \tag{17.12}$$

Eq. 17.10 can be rearranged to

$$\begin{aligned}\frac{1}{\rho}\nabla_z p &= \frac{1}{\rho}\nabla_{\hat{z}} p - \frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}}\nabla_{\hat{z}} z + g\nabla_{\hat{z}} z - g\nabla_{\hat{z}} z \\ &= \left(\frac{1}{\rho}\nabla_{\hat{z}} p + g\nabla_{\hat{z}} z\right) - \left(\frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}} + g\right)\nabla_{\hat{z}} z\end{aligned} \tag{17.13}$$

In the hydrostatic limit, the second term on the bottom line of (17.13) vanishes.

We now write

$$\begin{aligned}\frac{1}{\rho}\nabla_{\hat{z}}p =& RT\frac{\nabla_{\hat{z}}p}{p}\\ =&\frac{RT}{\kappa}\frac{\nabla_{\hat{z}}\Pi}{\Pi}\\ =&\frac{c_pT}{\Pi}\nabla_{\hat{z}}\Pi\\ =&\theta\nabla_{\hat{z}}\Pi.\end{aligned} \tag{17.14}$$

The bottom line of (17.14) is a particularly useful form. Continuing from (17.14), we can obtain a second useful form as follows:

$$\begin{aligned}\frac{1}{\rho}\nabla_{\hat{z}}p =&\nabla_{\hat{z}}(\Pi\theta)-\Pi\nabla_{\hat{z}}\theta\\ =&\nabla_{\hat{z}}(c_pT)-\Pi\nabla_{\hat{z}}\theta.\end{aligned} \tag{17.15}$$

Substituting (17.14) back into (17.13), we find that

$$\begin{aligned}\frac{1}{\rho}\nabla_z p =&[\nabla_{\hat{z}}(c_pT)-\Pi\nabla_{\hat{z}}\theta+g\nabla_{\hat{z}}z]-\left(\frac{1}{\hat{\rho}}\frac{\partial p}{\partial\hat{z}}+g\right)\nabla_{\hat{z}}z\\ =&\nabla_{\hat{z}}s-\Pi\nabla_{\hat{z}}\theta-\left(\frac{1}{\hat{\rho}}\frac{\partial p}{\partial\hat{z}}+g\right)\nabla_{\hat{z}}z,\end{aligned} \tag{17.16}$$

where $s$ is the dry static energy, defined by

$$s\equiv c_pT+gz. \tag{17.17}$$

In the hydrostatic limit, (17.16) reduces to

$$\frac{1}{\rho}\nabla_z p=\nabla_{\hat{z}}s-\Pi\nabla_{\hat{z}}\theta. \tag{17.18}$$

For the special case $\hat{z} \equiv \theta$, (17.18) further simplifies to

$$\frac{1}{\rho}\nabla_z p = \nabla_\theta s. \tag{17.19}$$

The vertical pressure-gradient force can also be written in various ways. Two of the possibilities are shown in (17.12). Here are some more examples:

$$\begin{aligned}
\frac{1}{\rho}\frac{\partial p}{\partial z} &= \frac{RT}{p}\frac{\partial p}{\partial z} \\
&= \frac{RT}{\kappa\Pi}\frac{\partial \Pi}{\partial z} \\
&= \frac{c_p T}{\Pi}\frac{\partial \Pi}{\partial z} \\
&= \theta\frac{\partial \Pi}{\partial z}\,.
\end{aligned} \tag{17.20}$$

The form shown on the bottom line of (17.20) is analogous to (17.14). A second useful form can be obtained by further manipulation starting from (17.20), as follows:

$$\begin{aligned}
\frac{1}{\rho}\frac{\partial p}{\partial z} &= \frac{\partial}{\partial z}(\Pi\theta) - \Pi\frac{\partial \theta}{\partial z} \\
&= \frac{\partial}{\partial z}(c_p T) - \Pi\frac{\partial \theta}{\partial z} \\
&= \frac{\partial s}{\partial z} - g - \Pi\frac{\partial \theta}{\partial z} \\
&= \frac{\partial \theta}{\partial z}\left(\frac{\partial s}{\partial \theta} - \Pi\right) - g.
\end{aligned} \tag{17.21}$$

In the hydrostatic limit, (17.21) reduces to

$$\frac{1}{\rho}\frac{\partial p}{\partial z} = -g. \tag{17.22}$$

With the use of (17.16) and (17.12), we can now rewrite (17.1) and (17.2) as

$$\frac{D\mathbf{V}_h}{Dt} = -\left(\nabla_{\hat{z}} s - \Pi \nabla_{\hat{z}} \theta\right) - \left(\frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}} + g\right)\nabla_{\hat{z}} z, \tag{17.23}$$

and

$$\frac{Dw}{Dt} = -\frac{1}{\hat{\rho}}\frac{\partial p}{\partial \hat{z}} - g, \tag{17.24}$$

respectively.

Using methods discussed in Appendix B, the continuity equation, (17.3), can be written as

$$\left(\frac{\partial \rho}{\partial t}\right)_{\hat{z}} - \frac{\partial \hat{z}}{\partial z}\frac{\partial \rho}{\partial \hat{z}}\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} + \nabla_{\hat{z}} \cdot (\rho \mathbf{V}_h) - \frac{\partial \hat{z}}{\partial z}\left[\frac{\partial}{\partial \hat{z}}(\rho \mathbf{V}_h)\right] \cdot \nabla_{\hat{z}} z + \frac{\partial \hat{z}}{\partial z}\frac{\partial}{\partial \hat{z}}(\rho w) = 0, \tag{17.25}$$

or

$$\frac{\partial z}{\partial \hat{z}}\left(\frac{\partial \rho}{\partial t}\right)_{\hat{z}} - \frac{\partial \rho}{\partial \hat{z}}\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} + \frac{\partial z}{\partial \hat{z}}\nabla_{\hat{z}} \cdot (\rho \mathbf{V}_h) - \left[\frac{\partial}{\partial \hat{z}}(\rho \mathbf{V}_h)\right] \cdot \nabla_{\hat{z}} z + \frac{\partial}{\partial \hat{z}}(\rho w) = 0. \tag{17.26}$$

This result can be simplified using

$$\begin{aligned}
\left(\frac{\partial \hat{\rho}}{\partial t}\right)_{\hat{z}} &= \left[\frac{\partial}{\partial t}\left(\rho \frac{\partial z}{\partial \hat{z}}\right)\right]_{\hat{z}} \\
&= \left(\frac{\partial \rho}{\partial t}\right)_{\hat{z}} \frac{\partial z}{\partial \hat{z}} + \rho\left[\frac{\partial}{\partial t}\left(\frac{\partial z}{\partial \hat{z}}\right)\right]_{\hat{z}} \\
&= \left(\frac{\partial \rho}{\partial t}\right)_{\hat{z}} \frac{\partial z}{\partial \hat{z}} + \rho \frac{\partial}{\partial \hat{z}}\left(\frac{\partial z}{\partial t}\right)_{\hat{z}} \\
&= \left(\frac{\partial \rho}{\partial t}\right)_{\hat{z}} \frac{\partial z}{\partial \hat{z}} + \frac{\partial}{\partial \hat{z}}\left(\rho \frac{\partial z}{\partial t}\right)_{\hat{z}} - \frac{\partial \rho}{\partial \hat{z}}\left(\frac{\partial z}{\partial t}\right)_{\hat{z}}
\end{aligned} \tag{17.27}$$

and

$$\begin{aligned}\nabla_{\hat{z}} \cdot (\hat{\rho}\mathbf{V}_h) =& \nabla_{\hat{z}} \cdot \left( \rho \frac{\partial z}{\partial \hat{z}} \mathbf{V}_h \right) \\ =& \frac{\partial z}{\partial \hat{z}} \nabla_{\hat{z}} \cdot (\rho \mathbf{V}_h) + \rho \mathbf{V}_h \cdot \nabla_{\hat{z}} \left( \frac{\partial z}{\partial \hat{z}} \right) \\ =& \frac{\partial z}{\partial \hat{z}} \nabla_{\hat{z}} \cdot (\rho \mathbf{V}_h) + \rho \mathbf{V}_h \frac{\partial}{\partial \hat{z}} (\nabla_{\hat{z}} z) .\end{aligned} \tag{17.28}$$

Using (17.27) and (17.28) in (17.26), we can show, after some algebra, that

$$\left( \frac{\partial \hat{\rho}}{\partial t} \right)_{\hat{z}} + \nabla_{\hat{z}} \cdot (\hat{\rho}\mathbf{V}_h) + \frac{\partial}{\partial \hat{z}} (\hat{\rho}\hat{w}) = 0, \tag{17.29}$$

where

$$\hat{w} \equiv -\frac{\partial \hat{z}}{\partial z} \left[ \left( \frac{\partial z}{\partial t} \right)_{\hat{z}} + \mathbf{V}_h \cdot \nabla_{\hat{z}} z - w \right]. \tag{17.30}$$

Note that for $\hat{z} \equiv z$ Eq. (17.30) reduces to $\hat{w} = w$.

There are two ways to use (17.30):

1. Specify a formula to determine $\hat{w}$, and use (17.30) to predict $z$ on surfaces of constant $\hat{z}$. For example, if $\hat{z}$ is potential temperature, then $\hat{w}$ is proportional to the parameterized heating rate. Given the heating rate, we can use (17.30) to predict the height of $\theta$ surfaces.
2. Specify a (possibly complicated) rule to determine $(\partial z/\partial t)_{\hat{z}}$, and use (17.30) to determine $\hat{w}$. This is called the "Arbitrary Lagrangian-Eulerian" (ALE) method. It is "arbitrary" in the sense that any physically reasonable rule can be used to determine $(\partial z/\partial t)_{\hat{z}}$.

It is possible (and common) to use the two approaches in combination. For example, if $\hat{z}$ is defined in such a way that the boundary-layer top is a surface of constant $\hat{z}$ (e.g., Suarez et al., 1983), then the height of the boundary-layer top can be predicted using (17.30), given a parameterization of $\hat{w}$ (which would involve the entrainment rate and the cumulus mass

flux). This is another example of the first approach. But we might choose to impose upper and/or lower limits on the height of the boundary-layer top. That can be done by using (17.30) to determine the value of $\hat{w}$ required to avoid violating the limits. This would be an application of the second approach. Toy and Randall (2009) present a second example of the two approaches in combination.

By analogy with (17.29), we can transform (17.9), the conservation equation for arbitrary intensive scalar, to

$$\left[\frac{\partial}{\partial t}\left(\hat{\rho}A\right)\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot \left(\hat{\rho}\mathbf{V}_h A\right) + \frac{\partial}{\partial \hat{z}}\left(\hat{\rho}\hat{w}A\right) = \hat{\rho}S_A. \tag{17.31}$$

The advective form can be obtained by combining (26) with (25):

$$\left(\frac{\partial A}{\partial t}\right)_{\hat{z}} + \mathbf{V}_h \cdot \nabla_{\hat{z}} A + \hat{w}\frac{\partial A}{\partial \hat{z}} = S_A. \tag{17.32}$$

From (17.32), we see that the Lagrangian time-rate-of-change operator can be written as

$$\frac{D}{Dt} = \left(\frac{\partial}{\partial t}\right)_{\hat{z}} + \mathbf{V}_h \cdot \nabla_{\hat{z}} + \hat{w}\frac{\partial}{\partial \hat{z}}. \tag{17.33}$$

In particular,

$$\frac{D\hat{z}}{Dt} = \hat{w}\,. \tag{17.34}$$

If we set $A \equiv \theta$ and $S_A = \dot{\theta}$, then (17.31) becomes the thermodynamic energy equation in the form

$$\left[\frac{\partial}{\partial t}\left(\hat{\rho}\theta\right)\right]_{\hat{z}} + \nabla_{\hat{z}} \cdot \left(\hat{\rho}\mathbf{V}_h \theta\right) + \frac{\partial}{\partial \hat{z}}\left(\hat{\rho}\hat{w}\theta\right) = \hat{\rho}\dot{\theta}. \tag{17.35}$$

Let $\hat{z}_S$ and $\hat{z}_\infty$ be the values of $\hat{z}$ at the Earth's surface and at the "top of the atmosphere,"respectively. Integration of (17.29) gives

$$\begin{aligned}&\frac{\partial}{\partial t}\int_{\hat{z}_S}^{\hat{z}_\infty}\hat{\rho}\;d\hat{z}-\hat{\rho}_\infty\frac{\partial\hat{z}_\infty}{\partial t}+\hat{\rho}_S\frac{\partial\hat{z}_S}{\partial t}\\ +\nabla\cdot\int_{\hat{z}_S}^{\hat{z}_\infty}&(\hat{\rho}\mathbf{V}_h)\,d\hat{z}-(\hat{\rho}\mathbf{V}_h)_\infty\cdot\nabla\hat{z}_\infty+(\hat{\rho}\mathbf{V}_h)_\S\cdot\nabla\hat{z}_\S\\ &+(\hat{\rho}\hat{w})_\infty-(\hat{\rho}\hat{w})_S=0\,.\end{aligned}\tag{17.36}$$

The condition that no mass crosses the top of the atmosphere can be written as

$$\hat{\rho}_\infty\frac{\partial\hat{z}_\infty}{\partial t}+(\hat{\rho}\mathbf{V}_h)_\infty\cdot\nabla\hat{z}_\infty-(\hat{\rho}\hat{w})_\infty=0.\tag{17.37}$$

If the top of the model is assumed to be a surface of constant $\hat{z}$, which is usually the case, then (17.37) reduces to

$$\hat{w}_T=0.\tag{17.38}$$

Similarly, the condition that no mass crosses the Earth's surface is expressed by

$$\hat{\rho}_S\frac{\partial\hat{z}_S}{\partial t}+(\hat{\rho}\mathbf{V}_h)_S\cdot\nabla\hat{z}_S-(\hat{\rho}\hat{w})_S=0.\tag{17.39}$$

With the use of (17.37) and (17.39), Eq. (17.36) simplifies to

$$\frac{\partial}{\partial t}\int_{\hat{z}_S}^{\hat{z}_\infty}\hat{\rho}\;d\hat{z}+\nabla\cdot\int_{\hat{z}_S}^{\hat{z}_\infty}(\hat{\rho}\mathbf{V}_h)\,d\hat{z}=0\,,\tag{17.40}$$

which expresses conservation of mass for the entire column of air.

## 17.5 Vertical coordinates for quasi-static models

# Chapter 18

# Vertical coordinates for quasi-static models

## 18.1 Introduction

With a general vertical coordinate, $\hat{z}$, the hydrostatic equation can be expressed as

$$\frac{\partial p}{\partial \hat{z}} = -\hat{\rho} g \tag{18.1}$$

In view of (18.1), Eq. (17.40) is equivalent to

$$\frac{\partial p_S}{\partial t} = -\nabla \cdot \left( \int_{\hat{z}_T}^{\hat{z}_S} \hat{\rho} \mathbf{V} d\hat{z} \right) + \frac{\partial p_T}{\partial t} \,, \tag{18.2}$$

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 $\partial p_T / \partial t = 0$, as an upper boundary condition. This is discussed later. Corresponding to (18.2), we can show that the pressure tendency on an arbitrary $\hat{z}$-surface satisfies

$$\left( \frac{\partial p}{\partial t} \right)_{\hat{z}} = \frac{\partial p_T}{\partial t} - \nabla \cdot \left( \int_{\hat{z}_T}^{\hat{z}} \hat{\rho} \mathbf{V} d\hat{z} \right) + (\hat{\rho} \hat{w})_{\hat{z}}. \tag{18.3}$$

The thermodynamic equation can be written as

$$c_p \left[ \left( \frac{\partial T}{\partial t} \right)_{\hat{z}} + \mathbf{V} \cdot \nabla_{\hat{z}} T + \hat{w} \frac{\partial T}{\partial \hat{z}} \right] = \omega \alpha + Q, \tag{18.4}$$

where $c_p$ is the specific heat of air at constant pressure, $\alpha$ is the specific volume, and $Q$ is the heating rate per unit mass. An alternative form of the thermodynamic equation is

$$\left(\frac{\partial\theta}{\partial t}\right)_{\hat{z}} + \mathbf{V}\cdot\nabla_{\hat{z}}\theta + \hat{w}\frac{\partial\theta}{\partial\hat{z}} = \frac{Q}{\Pi}, \tag{18.5}$$

where

$$\begin{aligned}\Pi \equiv & c_p\frac{T}{\theta}\\ = & c_p\left(\frac{p}{p_0}\right)^{\kappa}\end{aligned} \tag{18.6}$$

is the Exner function. In (18.6), $\theta$ is the potential temperature; $p_0$ is a positive, constant reference pressure, usually taken to be 1000 hPa, and $\kappa \equiv R/c_p$, where $R$ is the gas constant.

## 18.2 The equation of motion and the horizontal pressure-gradient force

The horizontal momentum equation can be written as

$$\left(\frac{\partial\mathbf{V}}{\partial t}\right)_{\hat{z}} + [f + \mathbf{k}\cdot(\nabla_{\hat{z}}\times\mathbf{V})]\,\mathbf{k}\times\mathbf{V} + \nabla_{\varsigma}K + \hat{w}\frac{\partial\mathbf{V}}{\partial\hat{z}} = -\nabla_p\phi + \mathbf{F}. \tag{18.7}$$

Here $-\nabla_p\phi$ is the horizontal pressure-gradient force (hereafter abbreviated as HPGF), which is expressed as minus the gradient of the geopotential *along an isobaric surface*, and $\mathbf{F}$ is the friction vector. Also, $\mathbf{k}$ is a unit vector pointing upward, and it is important to remember that *the meaning of* $\mathbf{k}$ *is not affected by the choice of vertical coordinate system*. Similarly, $\mathbf{V}$ is the horizontal component of the velocity, and *the meaning of* $\mathbf{V}$ *is not affected by the choice of the vertical coordinate system*. Using the relation

$$\begin{aligned}\nabla_p = & \nabla_{\hat{z}} - \nabla_{\hat{z}}p\frac{\partial}{\partial p}\\ = & \nabla_{\hat{z}} + \frac{\nabla_{\hat{z}}p}{\hat{\rho}}\frac{\partial}{\partial\hat{z}},\end{aligned} \tag{18.8}$$

we can rewrite the HPGF as

$$-\nabla_p \phi = -\nabla_{\hat{z}} \phi - \frac{1}{\hat{\rho}} \frac{\partial \phi}{\partial \hat{z}} \nabla_{\hat{z}} p. \tag{18.9}$$

In view of (18.1), this can be expressed as

$$-\nabla_p \phi = -\nabla_{\hat{z}} \phi - \alpha \nabla_{\hat{z}} p. \tag{18.10}$$

Eq. (18.10) is a nice result. For $\hat{z} \equiv z$ it reduces to $-\nabla_p \phi = -\alpha \nabla_z p$, and for $\hat{z} = p$ it becomes $-\nabla_p \phi = -\nabla_p \phi$. These special cases are both very familiar.

Another useful form of the HPGF is expressed in terms of the dry static energy, which is defined by

$$s \equiv c_p T + \phi. \tag{18.11}$$

For the special case in which $\hat{z} \equiv \theta$, which will be discussed in detail later, the hydrostatic equation (18.1) can be written as

$$\frac{\partial s}{\partial \theta} = \Pi. \tag{18.12}$$

With the use of (18.11) and (18.12), Eq. (18.10) can be expressed as

$$-\nabla_p \phi = -\nabla_{\hat{z}} s + \Pi \nabla_{\hat{z}} \theta. \tag{18.13}$$

This form of the HPGF will be discussed later.

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

$$\left(\frac{\partial q_{\hat{z}}}{\partial t}\right)_{\hat{z}} + (\mathbf{V}\cdot\nabla_{\hat{z}})\,q_{\hat{z}} + \hat{w}\frac{\partial q_{\hat{z}}}{\partial \hat{z}} = -q_{\hat{z}}(\nabla_{\hat{z}}\cdot\mathbf{V}) + \frac{\partial \mathbf{V}}{\partial \hat{z}}\times(\nabla_{\hat{z}}\hat{w}) - \mathbf{k}\cdot[\nabla_{\hat{z}}\times(\nabla_p\phi)] + \mathbf{k}\cdot(\nabla_{\hat{z}}\times\mathbf{F})\,. \tag{18.14}$$

The first term on the right-hand side of (18.14) 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*. It is apparent from (18.10) and (18.13), however, that in general 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. From (18.10) we see that the HPGF is a pure gradient for $\hat{z}\equiv p$, and from (18.13) we see that the HPGF is a pure gradient for $\hat{z}\equiv\theta$. This is an advantage shared by the pressure and theta coordinates.

The *vertically integrated* HPGF has a very important property that can be used in the design of vertical differencing schemes. With the use of (18.1) and (17.11), we can rewrite (18.9) as follows:

$$\begin{aligned}
\hat{\rho}\mathbf{HPGF} &= -\hat{\rho}\nabla_{\hat{z}}\phi - \frac{\partial\phi}{\partial\hat{z}}\nabla_{\hat{z}}p \\
&= -\nabla_{\hat{z}}(\hat{\rho}\phi) + \phi\nabla_{\hat{z}}\hat{\rho} - \frac{\partial\phi}{\partial\hat{z}}\nabla_{\hat{z}}p \\
&= -\nabla_{\hat{z}}(\hat{\rho}\phi) - \phi\nabla_{\hat{z}}\left(\frac{\partial p}{\partial\hat{z}}\right) - \frac{\partial\phi}{\partial\hat{z}}\nabla_{\hat{z}}p \\
&= -\nabla_{\hat{z}}(\hat{\rho}\phi) - \phi\frac{\partial}{\partial\hat{z}}(\nabla_{\hat{z}}p) - \frac{\partial\phi}{\partial\hat{z}}\nabla_{\hat{z}}p \\
&= -\nabla_{\hat{z}}(\hat{\rho}\phi) - \frac{\partial}{\partial\hat{z}}(\phi\nabla_{\hat{z}}p)\,.
\end{aligned} \tag{18.15}$$

Vertically integrating with respect to mass, we find that

$$\int_{\hat{z}_T}^{\hat{z}_S}\hat{\rho}\mathbf{HPGF}\,d\hat{z} = -\nabla\left(\int_{\hat{z}_T}^{\hat{z}_S}\hat{\rho}\phi\,d\hat{z}\right) + (\hat{\rho}\phi)_S\nabla\hat{z}_S - (\hat{\rho}\phi)_T\nabla\hat{z}_T - \phi_S\left(\nabla_\varsigma p\right)_S + \phi_T\left(\nabla_\varsigma p\right)_T. \tag{18.16}$$

Here we have included the $(\hat{\rho}\phi)_S \nabla \hat{z}_S$ and $-(\hat{\rho}\phi)_T \nabla \hat{z}_T$ terms to allow for the possibility that $\hat{z}_T$ and $\hat{z}_S$ are spatially variable.

Consider a line integral of the vertically integrated HPGF, i.e., $\int_{\hat{z}_T}^{\hat{z}_S} \hat{\rho} \nabla_p \phi d\hat{z}$, along a *closed path*. Because the term $\nabla_{\hat{z}} \int_{\hat{z}_T}^{\hat{z}_S} \hat{\rho}\phi d\hat{z}$ is a gradient, its line integral is zero. The line integral of $\phi_S \nabla p_S$ will also be zero if either $\phi_S$ or $p_S$ is constant along the path of integration, which is not likely with realistic geography. On the other hand, if either $\phi_T$ or $p_T$ is constant along the path of integration, then the line integral of $\phi_T \nabla p_T$ will vanish, and this can easily be arranged. *This is a motivation to choose either $\phi_T$ = constant or $p_T$ = constant, regardless of the choice of vertical coordinate*. In addition, it is almost always possible (and advisable) to choose $\hat{z}_T$ = constant. Further discussion is given later.

To see how (18.16) plays out, let's consider two examples. For the case of pressure coordinates, with $\rho_p = -1$ and $p_T$ = constant, the last two terms of (18.16) vanish, because they are proportional to the gradient of pressure on pressure surfaces. We get

$$\int_{p_T}^{p_S} \mathbf{HPGF}\, dp = -\nabla \left( \int_{p_T}^{p_S} \phi dp \right) + \phi_S \nabla p_S. \tag{18.17}$$

Note that $\nabla p_S$ is *not* the same as $(\nabla_p p)_S$ (which is equal to zero).

For the case of height coordinates, with $\rho_z = \rho g$ and $z_T$ = constant, we get

$$\int_{z_T}^{z_S} \rho g \mathbf{HPGF}\, dz = -\nabla \left[ \int_{z_T}^{z_S} \rho g \phi dz \right] + (\rho g \phi)_S \nabla z_S - \phi_S (\nabla_z p)_S + \phi_T (\nabla_z p)_T. \tag{18.18}$$

Swapping the limits of integration, and flipping signs to compensate, we get

$$\begin{aligned}
\int_{z_S}^{z_T} \rho g \mathbf{HPGF}\, dz &= -\nabla \left( \int_{z_S}^{z_T} \rho g \phi dz \right) - (\rho g \phi)_S \nabla z_S + \phi_S (\nabla_z p)_S - \phi_T (\nabla_z p)_T \\
&= -\nabla \left( \int_{z_S}^{z_T} \rho g \phi dz + \phi_T p_T \right) + \phi_S \left[ -(\rho g)_S \nabla z_S + (\nabla_z p)_S \right] \\
&= -\nabla \left( \int_{z_S}^{z_T} \rho g \phi dz - \phi_T p_T \right) + \phi_S \left[ \left( \frac{\partial p}{\partial z} \right)_S \nabla z_S + (\nabla_z p)_S \right] \\
&= -\nabla \left( \int_{z_S}^{z_T} \rho g \phi dz - \phi_T p_T \right) + \phi_S \nabla p_S.
\end{aligned} \tag{18.19}$$

In the final line above, we have used a coordinate transformation.

We conclude that, *in the absence of topography along the path of integration, and with either either $\phi_T$ = constant or $p_T$ = constant, there cannot be any net spin-up or spin-down of a circulation in the region bounded by a closed path.* This conclusion is independent of the choice of vertical coordinate system. Later we will show how this important constraint can be mimicked in a vertically discrete model.

## 18.3 Vertical mass flux for 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.,

$$\hat{z} \equiv F(\theta, p, p_S). \tag{18.20}$$

We get 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. (18.20) does include a variety of interesting cases, which will be discussed below, 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 is not included in (18.20).

By taking the partial derivative (18.20) with respect to time, on a surface of constant $\hat{z}$, we find that

$$0 = \left[\frac{\partial}{\partial t} F(\theta, p, p_S)\right]_{\hat{z}}. \tag{18.21}$$

The chain rule tells us that this is equivalent to

$$\frac{\partial F}{\partial \theta}\left(\frac{\partial \theta}{\partial t}\right)_{\hat{z}} + \frac{\partial F}{\partial p}\left(\frac{\partial p}{\partial t}\right)_{\hat{z}} + \frac{\partial F}{\partial p_S}\left(\frac{\partial p_S}{\partial t}\right) = 0. \tag{18.22}$$

Substituting from (18.5), (18.3), and (18.2), we obtain

$$\begin{aligned} &\frac{\partial F}{\partial \theta}\left[-\left(\mathbf{V}\cdot\nabla_{\hat{z}}\theta+\hat{w}\frac{\partial\theta}{\partial\hat{z}}\right)+\frac{Q}{\Pi}\right] \\ &+\frac{\partial F}{\partial p}\left[\frac{\partial p_T}{\partial t}-\nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}}\hat{\rho}\mathbf{V}d\hat{z}\right)+(\hat{\rho}\hat{w})_{\hat{z}}\right] \\ &+\frac{\partial F}{\partial p_S}\left[\frac{\partial p_T}{\partial t}-\nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}_S}\hat{\rho}\mathbf{V}d\hat{z}\right)\right]=0. \end{aligned} \tag{18.23}$$

Eq. (18.23) can be solved for the vertical velocity, $\hat{w}$:

$$\boxed{\hat{w}=\frac{\frac{\partial F}{\partial\theta}\left(-\mathbf{V}\cdot\nabla_{\hat{z}}\theta+\frac{Q}{\Pi}\right)+\frac{\partial F}{\partial p}\left[\frac{\partial p_T}{\partial t}-\nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}}\hat{\rho}\mathbf{V}d\hat{z}\right)\right]+\frac{\partial F}{\partial p_S}\left[\frac{\partial p_T}{\partial t}-\nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}_S}\hat{\rho}\mathbf{V}d\hat{z}\right)\right]}{\left\{\frac{\partial\theta}{\partial\hat{z}}\frac{\partial F}{\partial\theta}-\hat{\rho}\frac{\partial F}{\partial p}\right\}}}. \tag{18.24}$$

Here we have assumed that the heating rate, $Q$, is *not* formulated as an explicit function of $\hat{w}$; this is generally the case in modern numerical models, but not in some older theoretical models. With this assumption, the model can be constructed so that $Q$ is computed before determining the vertical velocity, which means that it can be considered "known" in (18.24).

As a check of (18.24), consider the special case $F\equiv p$, so that $\rho_{\hat{z}}=-1$, and assume that $\frac{\partial p_T}{\partial t}=0$, as would be natural for the case of pressure coordinates. Then (18.24) reduces to

$$\dot{p}\,(\equiv\omega)=-\nabla\cdot\left(\int_{p_T}^{p}\mathbf{V}dp\right). \tag{18.25}$$

As a second special case, suppose that $F\equiv\theta$. Then (18.24) becomes

$$\dot{\theta}=Q/\Pi\,. \tag{18.26}$$

Both of these results are as expected.

We assume that the model top is a surface of constant $\hat{z}$, i.e., $\hat{z}_T=$ constant. Then (18.22) must apply at the model top, so that we can write

$$\left(\frac{\partial F}{\partial \theta}\right)_{\theta_T,p_T}\frac{\partial \theta_T}{\partial t}+\left(\frac{\partial F}{\partial p}\right)_{\theta_T,p_T}\frac{\partial p_T}{\partial t}+\left(\frac{\partial F}{\partial p_S}\right)_{\theta_T,p_T}\frac{\partial p_S}{\partial t}=0. \tag{18.27}$$

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. Then Eq. (18.27) simplifies to

$$\left(\frac{\partial F}{\partial \theta}\right)_{\theta_T,p_T}\frac{\partial \theta_T}{\partial t}+\left(\frac{\partial F}{\partial p}\right)_{\theta_T,p_T}\frac{\partial p_T}{\partial t}=0. \tag{18.28}$$

Now consider two possibilities. If we make the top of the model an isobaric surface, so that $\partial p_T/\partial t=0$, then the last term of (18.28) 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 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* $(\partial F/\partial p_S)_{\theta_T,p_T}=0$.

As a second possibility, *when the model top is an isentropic surface,* $\partial\theta_T/\partial t=0$*, and the form of* $F(\theta,p,p_S)$ *must be chosen so that* $(\partial F/\partial p_S)_{\theta_T,p_T}=0$.

Further discussion is given later.

## 18.4 Survey of particular vertical coordinate systems

We now discuss the following nine particular choices of vertical coordinate:

- height, $z$
- pressure, $p$
- log-pressure, $z^*$, which is used in many theoretical studies and some numerical models (e.g., Girard et al., 2014)
- $\sigma$, defined by

$$\sigma=\frac{p-p_T}{p_S-p_T}, \tag{18.29}$$

  which is designed to simplify the lower boundary condition
- a "hybrid," or "mix," of $\sigma$ and $p$ coordinates, used in many global circulation models, including the model of the European Centre for Medium Range Weather Forecasts

- $\eta$, which is a modified $\sigma$ coordinate, defined by

$$\eta \equiv \left( \frac{p - p_T}{p_S - p_T} \right) \eta_S, \tag{18.30}$$

  where $\eta_S$ is a time-independent function of the horizontal coordinates
- potential temperature, $\theta$, which has many attractive properties
- entropy, $\varepsilon = c_p \ln \frac{\theta}{\theta_0}$, where $\theta_0$ is a constant reference value of $\theta$
- a hybrid sigma-theta coordinate, which behaves like $\sigma$ near the Earth's surface, and like $\theta$ away from the Earth's surface.

Of these nine possibilities, all except the height coordinate and the $\eta$ coordinate are members of the family of coordinates given by (18.20).

### 18.4.1 Height

In height coordinates, the hydrostatic equation is

$$\frac{\partial p}{\partial z} = -\rho g \, . \tag{18.31}$$

The continuity equation in height coordinates is

$$\left( \frac{\partial \rho}{\partial t} \right)_z + \nabla_z \cdot (\rho \mathbf{V}) + \frac{\partial}{\partial z} (\rho w) = 0. \tag{18.32}$$

This equation is easy to interpret, but it is mathematically complicated, because it is nonlinear and involves the time derivative of a quantity that varies with height, namely the density.

The lower boundary condition in height coordinates is

$$\frac{\partial z_S}{\partial t} + \mathbf{V}_S \cdot \nabla z_S - w_S = 0. \tag{18.33}$$

Normally we can assume that $z_S$ is independent of time, but (18.33) 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.

The thermodynamic energy equation in height coordinates can be written as

$$c_p \rho \left( \frac{\partial T}{\partial t} \right)_z = -c_p \rho \left( \mathbf{V} \cdot \nabla_z T + w \frac{\partial T}{\partial z} \right) + \omega + \rho Q. \tag{18.34}$$

Here

$$\begin{aligned} \omega &= \left( \frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p + w \frac{\partial p}{\partial z} \\ &= \left( \frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p - \rho g w. \end{aligned} \tag{18.35}$$

By using (18.35) in (18.34), we find that

$$c_p \rho \left( \frac{\partial T}{\partial t} \right)_z = -c_p \rho \mathbf{V} \cdot \nabla_z T - \rho w c_p (\Gamma_d - \Gamma) + \left[ \left( \frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p \right] + \rho Q, \tag{18.36}$$

where the actual lapse rate and the dry-adiabatic lapse rate are given by

$$\Gamma \equiv -\frac{\partial T}{\partial z}, \tag{18.37}$$

and

$$\Gamma_d \equiv \frac{g}{c_p}, \tag{18.38}$$

respectively. Eq. (18.36) 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 (18.3), which is

$$\left(\frac{\partial p}{\partial t}\right)_z = -g\nabla_z \cdot \int_z^\infty (\rho \mathbf{V})\, dz + g\rho(z)\, w(z) + \frac{\partial p_T}{\partial t}. \tag{18.39}$$

Substitution into (18.36) gives

$$\begin{aligned} c_p \rho \left(\frac{\partial T}{\partial t}\right)_z &= -c_p \rho \mathbf{V} \cdot \nabla_z T - \rho w c_p (\Gamma_d - \Gamma) \\ &\quad + \left[-g\nabla_z \cdot \int_z^\infty (\rho \mathbf{V})\, dz + g\rho(z)\, w(z) + \frac{\partial p_T}{\partial t}\right] + \mathbf{V} \cdot \nabla_z p + \rho Q. \end{aligned} \tag{18.40}$$

According to (18.40), the time rate of change of the temperature at a given height is influenced by the convergence of the horizontal wind field through a deep layer. 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

$$\left(\frac{\partial \theta}{\partial t}\right)_z = -\left(\mathbf{V} \cdot \nabla_z \theta + w\frac{\partial \theta}{\partial z}\right) + \frac{Q}{\Pi}. \tag{18.41}$$

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 $w$? The height coordinate is not a member of the family of schemes defined by (18.20), and so (18.24), the formula for the vertical mass flux derived from (18.20), does not apply. Instead, $w$ 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. Richardson's equation is actually closely analogous to (18.24), but somewhat more complicated. The derivation of Richardson's equation is also more complicated than the derivation of (18.24). 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

$$p = \rho RT. \tag{18.42}$$

"Logarithmic differentiation" of (18.42) with respect to time gives

$$\frac{1}{p}\left(\frac{\partial p}{\partial t}\right)_z = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial t}\right)_z + \frac{1}{T}\left(\frac{\partial T}{\partial t}\right)_z. \tag{18.43}$$

The time derivatives can be eliminated by using continuity (18.32), the thermodynamic energy equation (18.36) and the pressure tendency equation (18.39). Note that the derivation of (18.39) involves use of the hydrostatic equation. After some manipulation, we find that

$$\begin{aligned} c_p T \frac{\partial}{\partial z}(\rho w) + \rho w\left[g\frac{c_v}{R} + c_p(\Gamma_d - \Gamma)\right] &= (-c_p \rho \mathbf{V}\cdot\nabla_z T + \mathbf{V}\cdot\nabla_z p) \\ &\quad - c_p T \nabla_z\cdot(\rho\mathbf{V}) + g\frac{c_v}{R}\nabla_z\cdot\int_z^\infty (\rho\mathbf{V})\,dz' + \rho Q. \end{aligned} \tag{18.44}$$

where

$$c_v \equiv c_p - R \tag{18.45}$$

is the specific heat of air at constant volume.

Eq. (18.44) 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:

$$c_p T \frac{\partial(\rho w)}{\partial z} = \rho c_p T \frac{\partial w}{\partial z} + w c_p T \frac{\partial \rho}{\partial z}, \tag{18.46}$$

A second logarithmic differentiation of (18.42), this time with respect to height, gives

$$\frac{1}{p}\frac{\partial p}{\partial z} = \frac{1}{\rho}\frac{\partial \rho}{\partial z} + \frac{1}{T}\frac{\partial T}{\partial z}. \tag{18.47}$$

Using the hydrostatic equation again, we can rewrite (18.47) as

$$\begin{aligned} \frac{1}{\rho}\frac{\partial \rho}{\partial z} &= -\frac{\rho g}{p} + \frac{\Gamma}{T} \\ &= \frac{1}{T}\left(-\frac{g}{R} + \Gamma\right). \end{aligned} \tag{18.48}$$

Substitution of (18.48) into (18.46) gives

$$c_p T \frac{\partial (\rho w)}{\partial z} = \rho c_p T \frac{\partial w}{\partial z} + \rho w c_p \left(-\frac{g}{R} + \Gamma\right). \tag{18.49}$$

Finally, substitute (18.49) into (18.44), and combine terms, to obtain

$$\begin{aligned} \frac{\partial w}{\partial z} = &\left(\frac{-c_p \rho \mathbf{V}\cdot\nabla_z T + \mathbf{V}\cdot\nabla_z p}{\rho c_p T}\right) - \frac{1}{\rho}\nabla_z\cdot(\rho\mathbf{V}) \\ &+ \frac{c_v}{c_p p}\left[g\nabla_z\cdot\int_z^\infty (\rho\mathbf{V})\,dz' - \frac{\partial p_T}{\partial t}\right] + \frac{Q}{c_p T}. \end{aligned} \tag{18.50}$$

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 (18.50) 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 of (18.50) may (individually) tend to upset that balance.

As a very simple illustration of the use of (18.50), suppose that we have no horizontal motion. Then (18.50) drastically simplifies to

$$\frac{\partial w}{\partial z} = \frac{1}{c_p}\left(\frac{Q}{T} - \frac{c_v}{p}\frac{\partial p_T}{\partial t}\right). \tag{18.51}$$

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

$$\frac{\partial p_T}{\partial t} = \frac{\int_{z_S}^{z_T} \frac{Q}{c_v T} dz'}{\int_{z_S}^{z_T} \frac{1}{p} dz'}, \tag{18.52}$$

This shows that heating causes the pressure at the model top to increase with time, like in a pressure cooker. The vertical velocity satisfies

$$w(z) = \int_0^z \left( \frac{Q}{c_p T} - \frac{c_v}{c_p p} \frac{\partial p_T}{\partial t} \right) dz', \tag{18.53}$$

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.

### 18.4.2 Pressure

The hydrostatic equation in pressure coordinates is

$$\frac{\partial \phi}{\partial p} = -\alpha \, . \tag{18.54}$$

Eq. (17.11) reduces to

$$\rho_p = -1/g. \tag{18.55}$$

The continuity equation in pressure coordinates is relatively simple; it is linear and does not involve a time derivative:

$$\nabla_p \cdot \mathbf{V} + \frac{\partial \omega}{\partial p} = 0. \tag{18.56}$$

On the other hand, the lower boundary condition is complicated in pressure coordinates:

$$\frac{\partial p_S}{\partial t} + \mathbf{V}_S \cdot \nabla p_S - \omega_S = 0. \tag{18.57}$$

Recall that $p_S$ can be predicted using the surface pressure-tendency equation, (18.2). Substitution from (18.2) into (18.57) gives

$$\omega_S = \frac{\partial p_T}{\partial t} - \nabla \cdot \left( \int_{p_T}^{p_S} \mathbf{V} dp \right) + \mathbf{V}_S \cdot \nabla p_S, \tag{18.58}$$

which can be used to diagnose $\omega_S$. 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 Leith at the Lawrence National Laboratory (now the Lawrence Livermore National Laboratory).

With the pressure coordinate, we can write

$$\left[ \frac{\partial}{\partial t} \left( \frac{\partial \phi}{\partial p} \right) \right]_p = -\frac{R}{p} \left( \frac{\partial T}{\partial t} \right)_p. \tag{18.59}$$

This allows us to eliminate the temperature in favor of the geopotential, which is often done in theoretical studies.

### 18.4.3 Log-pressure

Obviously a surface of constant $p$ is also a surface of constant $\ln p$. Nevertheless, the equations take different forms in the $p$ and $\ln p$ coordinate systems.

Let $T_0$ be a *constant* reference temperature, and $H \equiv \dfrac{RT_0}{g}$ the corresponding scale height. *Define* the "log-pressure coordinate," denoted by $z^*$, in terms of the differential relationship

$$dz^* \equiv -H\frac{dp}{p}. \tag{18.60}$$

Note that $z^*$ has the units of length (i.e., it is "like" height). It is easy to show that

$$dz^* = \frac{T_0}{T}dz, \tag{18.61}$$

so that

$$dz^* = dz \text{ when } T(p) = T_0. \tag{18.62}$$

Although generally $z \neq z^*$, we can force $z(p = p_S) = z^*(p = p_S)$. From (18.60), we see that

$$\frac{\partial \phi^*}{\partial p} = -\frac{RT_0}{p}, \tag{18.63}$$

where

$$\phi^* \equiv gz^*. \tag{18.64}$$

Although (18.63) looks like the hydrostatic equation, it is really nothing more than the definition of the log pressure coordinate. Since $z^*$ is a constant along a log-pressure surface (i.e., along a pressure surface), $\phi^*$ is also constant.

We do of course have the true hydrostatic equation, which can be written as

$$\frac{\partial \phi}{\partial p} = -\frac{RT}{p}. \tag{18.65}$$

Here the true (non-constant) temperature appears. Subtracting (18.63) from (18.65), we obtain a useful form of the hydrostatic equation:

$$\frac{\partial\left(\phi-\phi^*\right)}{\partial p}=-\frac{R\left(T-T_0\right)}{p}. \tag{18.66}$$

Since $\phi^*$ and $T_0$ are independent of time on $z^*$ surfaces, we see that

$$\frac{\partial}{\partial t}\left(\frac{\partial\phi}{\partial p}\right)_{z^*}=-\frac{R}{p}\left(\frac{\partial T}{\partial t}\right)_{z^*}. \tag{18.67}$$

The pseudo-density in log-pressure coordinates is given by

$$\rho_{z^*}=\frac{p}{RT_0}. \tag{18.68}$$

The vertical velocity in log-pressure coordinates is

$$\begin{aligned} w^* &\equiv \frac{Dz^*}{Dt} \\ &= -\frac{H}{p}\frac{Dp}{Dt} \\ &= -\frac{H}{p}\omega. \end{aligned} \tag{18.69}$$

The continuity equation in log-pressure coordinates is given by

$$\left(\frac{\partial\rho_{z^*}}{\partial t}\right)_{z^*}+\nabla_{z^*}\cdot\left(\rho_{z^*}\mathbf{V}\right)+\frac{\partial}{\partial z^*}\left(\rho_{z^*}w^*\right)=0. \tag{18.70}$$

### 18.4.4 Terrain-following coordinates

The $\sigma$-coordinate of Phillips (1957) is defined by

$$\sigma\equiv\frac{p-p_T}{\pi}, \tag{18.71}$$

where

$$\pi \equiv p_S - p_T, \tag{18.72}$$

which is independent of height. From (18.71) and (18.72), it is clear that

$$\sigma_S = 1 \text{ and } \sigma_T = 0. \tag{18.73}$$

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) chose $p_T = 0$.

Rearranging (18.71), we can write

$$p = p_T + \sigma\pi. \tag{18.74}$$

For a fixed value of $\sigma$, i.e., along a surface of constant $\sigma$, this implies that

$$dp = \sigma d\pi, \tag{18.75}$$

where the differential can represent a fluctuation in either time or horizontal position. Also,

$$\frac{\partial}{\partial p}(\,) = \frac{1}{\pi}\frac{\partial}{\partial \sigma}(\,). \tag{18.76}$$

Here the partial derivatives are evaluated at fixed horizontal position and time.

The pseudodensity in $\sigma$-coordinates is

$$\rho_\sigma = \pi, \tag{18.77}$$

which is independent of height. Here we choose not to use the minus sign in (17.11). The continuity equation in $\sigma$-coordinates can therefore be written as

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{V}) + \frac{\partial (\pi \dot{\sigma})}{\partial \sigma} = 0. \tag{18.78}$$

Although this equation does contain a time derivative, the differentiated quantity, $\pi$, is independent of height, which makes it considerably simpler than than the continuity equation in height coordinates.

The lower boundary condition in $\sigma$-coordinates is very simple:

$$\dot{\sigma} = 0 \text{ at } \sigma = 1. \tag{18.79}$$

This simplicity was in fact Phillips' motivation for the invention of $\sigma$-coordinates. The upper boundary condition is similar:

$$\dot{\sigma} = 0 \text{ at } \sigma = 0. \tag{18.80}$$

The continuity equation in $\sigma$-coordinates plays a dual role. First, it is used to predict $\pi$. This is done by integrating (18.78) through the depth of the vertical column and using the boundary conditions (18.79) and (18.80), to obtain the surface pressure-tendency equation in the form

$$\frac{\partial \pi}{\partial t} = -\nabla \cdot \left( \int_0^1 \pi \mathbf{V} d\sigma \right). \tag{18.81}$$

The continuity equation is also used to determine $\pi\dot{\sigma}$. Once $\frac{\partial \pi}{\partial t}$ has been evaluated using (18.81), which does not involve $\pi\dot{\sigma}$, we can substitute back into (18.78) to obtain

$$\frac{\partial}{\partial \sigma} (\pi \dot{\sigma}) = \nabla \cdot \left( \int_0^1 \pi \mathbf{V} d\sigma \right) - \nabla_\sigma \cdot (\pi \mathbf{V}). \tag{18.82}$$

This can be integrated vertically to obtain $\pi\dot{\sigma}$ as a function of $\sigma$, 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, and the result is consistent with Eq. (18.24).

The hydrostatic equation in $\sigma$-coordinates is simply

$$\frac{1}{\pi}\frac{\partial \phi}{\partial \sigma} = -\alpha, \tag{18.83}$$

which is closely related to the hydrostatic equation in pressure coordinates.

Finally, the horizontal pressure-gradient force takes a relatively complicated form:

$$\mathbf{HPGF} = -\sigma\alpha\nabla\pi - \nabla_\sigma\phi, \tag{18.84}$$

which can easily be obtained from (18.10). Consider the two contributions to the HPGF when evaluated near a mountain, as illustrated in Fig. 18.1. Near steep topography, the spatial variations of $p_S$ and the near-surface value of $\phi$ along a $\sigma$-surface are strong and of opposite sign. For example, moving uphill $p_S$ decreases while $\phi_S$ increases. As a result, the two terms on the right-hand side of (18.83) are individually large and opposing, and the HPGF 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 (18.83) can be as large as the HPGF.

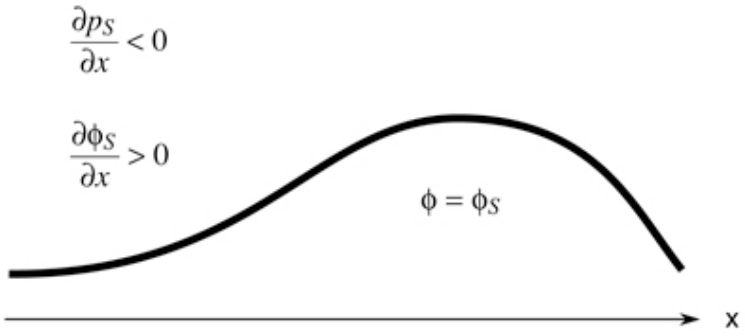



Figure 18.1: Sketch illustrating the opposing terms of the horizontal pressure gradient force as measured in $\sigma$-coordinates.

Using the hydrostatic equation, (18.83), we can rewrite (18.84) as

$$\mathbf{HPGF} = \sigma\left(\frac{1}{\pi}\frac{\partial \phi}{\partial \sigma}\right)\nabla\pi - \nabla_\sigma\phi. \tag{18.85}$$

Rearranging, we find that

$$\begin{aligned}\pi\,(\mathbf{HPGF}) &= \sigma\frac{\partial\phi}{\partial\sigma}\nabla\pi - \pi\nabla_\sigma\phi \\ &= \left[\frac{\partial\,(\sigma\phi)}{\partial\sigma} - \phi\right]\nabla\pi - \pi\nabla_\sigma\phi \\ &= \frac{\partial\,(\sigma\phi)}{\partial\sigma}\nabla\pi - \nabla_\sigma\,(\pi\phi)\,.\end{aligned} \tag{18.86}$$

This is a special case of (18.15).

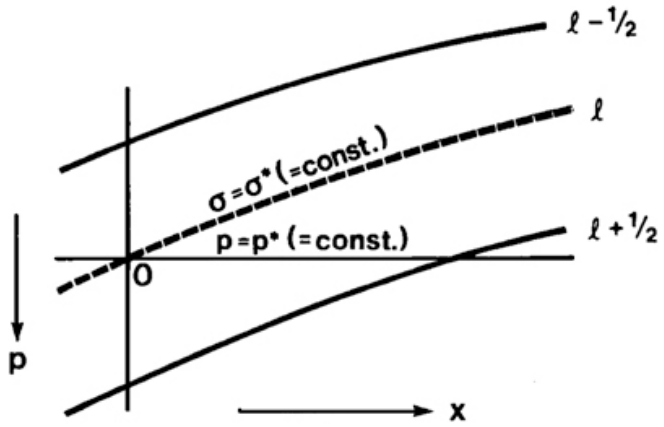



Figure 18.2: Sketch illustrating the pressure-gradient force as seen in $\sigma$-coordinates and pressure coordinates.

The problem with the hydrostatic equation in $\sigma$-coordinates may appear to be an issue mainly with horizontal differencing, because the HPGF involves only horizontal derivatives, but vertical differencing also comes in. To see how, consider Fig. 18.2. At the point O, the $\sigma = \sigma*$ and $p = p*$ surfaces intersect. As we move away from point O, the two surfaces separate. By a coordinate transformation, we can write

$$\begin{aligned}\mathbf{HPGF} &= -\nabla_p\phi \\ &= -\nabla_\sigma\phi + \frac{\partial\phi}{\partial p}\nabla_\sigma p.\end{aligned} \tag{18.87}$$

This second line of (18.87) expresses the HPGF in terms of both the horizontal change in $\phi$ along a $\sigma$-surface, say between two neighboring horizontal grid points (mass points), and the vertical change in $\phi$ between neighboring model layers. The latter depends, hydrostatically, on the temperature. Using hydrostatics, the ideal gas law, and the definition of $\sigma$, we can rewrite (18.87) as

$$\mathbf{HPGF} = -\nabla_\sigma \phi - \left(\frac{RT}{p}\right)\sigma\nabla\pi. \tag{18.88}$$

Compare with (18.83).

If the $\sigma$-surfaces are very steeply tilted relative to constant height surfaces, which is expected near steep mountains, the temperature needed on the right-hand side of (18.87) will be representative of two or more $\sigma$-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

$$\frac{\delta\sigma}{\delta x} \geq \frac{\left|\left(\frac{\delta\phi}{\delta x}\right)_\sigma\right|}{\left|\left(\frac{\delta\phi}{\delta\sigma}\right)_x\right|}. \tag{18.89}$$

This is a condition on the aspect ratio of the grid cells. It means that $\delta\sigma$ must be coarse enough for a given $\delta x$, or that $\delta x$ must be small enough for a given $\delta\sigma$. This implies that an increase in the vertical resolution without a corresponding increase in the horizontal resolution can cause problems. The numerator of the right-hand side of (18.89) 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 (18.89) near the surface.

Finally, the Lagrangian time derivative of pressure can be expressed in $\sigma$-coordinates as

$$\begin{aligned}\omega \equiv \frac{Dp}{Dt} &= \left(\frac{\partial p}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma p + \dot{\sigma}\frac{\partial p}{\partial\sigma} \\ &= \sigma\left(\frac{\partial\pi}{\partial t} + \mathbf{V}\cdot\nabla\pi\right) + \pi\dot{\sigma}.\end{aligned} \tag{18.90}$$

### 18.4.5 Hybrid sigma-pressure coordinates

The advantage of the sigma coordinate is realized in the lower boundary condition. The disadvantage, in terms of the complicated and poorly behaved pressure-gradient force, is realized at all levels. This has motivated the use of hybrid coordinates that reduce to sigma

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

$$\sigma_p(p, p_S) \equiv \left(\frac{p - p_T}{p_0 - p_T}\right)\left(\frac{p_S - p}{p_S - p_T}\right) + \left(\frac{p - p_T}{p_S - p_T}\right)^2, \tag{18.91}$$

where $p_0$ is a positive constant. You can confirm that $\sigma_p$ is monotonic with pressure, provided that $p_0 > p_S/2$. Inspection of (18.91) shows that

$$\sigma_p = 0 \text{ for } p = 0, \text{ and } \sigma_p = 1 \text{ for } p = p_S. \tag{18.92}$$

It can be shown that $\sigma_p$-surfaces are nearly parallel to isobaric surfaces in the upper troposphere and stratosphere, despite possible variations of the surface pressure in the range ~1000 mb to ~500 mb. When we evaluate the HPGF with the $\sigma_p$-coordinate, there are still two terms, as with the $\sigma$-coordinate, but above the lower troposphere one of the terms is strongly dominant.

### 18.4.6 The eta coordinate

As a solution to the problem with the HPGF in $\sigma$-coordinates, Mesinger and Janjic (1985) proposed the $\eta$-coordinate, which was used operationally at NCEP (the National Centers for Environmental Prediction). The coordinate is defined by

$$\eta \equiv \sigma \eta_S, \tag{18.93}$$

where

$$\eta_S = \frac{p_{rf}(z_S) - p_T}{p_{rf}(0) - p_T}. \tag{18.94}$$

Whereas $\sigma = 1$ at the Earth's surface, Eq. (18.93) shows that $\eta = \eta_S$ at the Earth's surface. According to (18.94), $\eta_S = 1$ (just as $\sigma = 1$) if $z_S = 0$. Here $z_S = 0$ is chosen to be at or near "sea level." The function $p_{rf}(z_S)$ is pre-specified so that it gives typical surface pressures for the full range of possible values of $z_S$. Because $z_S$ depends on the horizontal

coordinates, $\eta_S$ does too. In fact, after choosing the function $p_{rf}(z_S)$ 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 $\sigma$-coordinate model, we must specify (i.e., choose) values of $\sigma$ to serve as layer-edges and/or layer centers. These values are constant in the horizontal and time. Similarly, when we build an $\eta$-coordinate model, we must specify fixed values of $\eta$ to serve as layer edges and/or layer centers. The values of $\eta$ to be chosen include the possible values of $\eta_S$. This means that *only a finite number of discrete (and constant) values of $\eta_S$ are permitted*; the number increases as the vertical resolution of the model increases. It follows that only a finite number of discrete values of $z_S$ are permitted: *Mountains must come in a few discrete sizes, like off-the-rack clothing!* This is sometimes called the "step-mountain" approach. Fig. 18.3 shows how the $\eta$-coordinate works near mountains. Note that, unlike $\sigma$-surfaces, $\eta$-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 mountains.

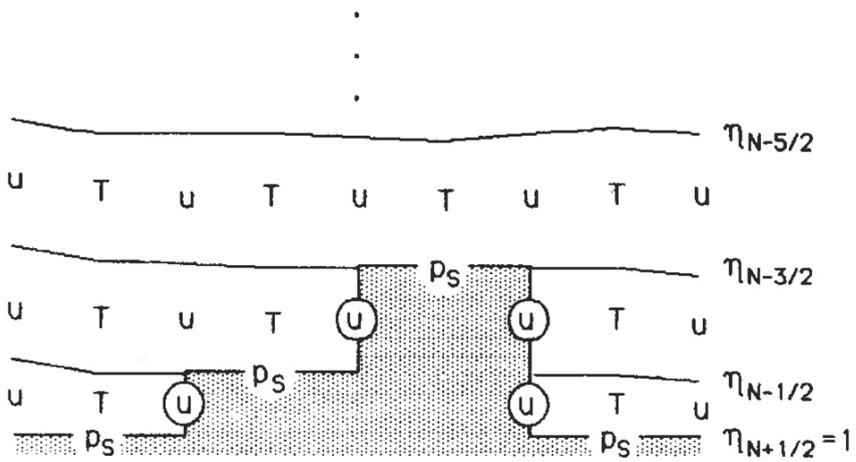



Figure 18.3: Sketch illustrating the $\eta$-coordinate.

In $\eta$-coordinates, the HPGF still consists of two terms:

$$-\nabla_p \phi = -\nabla_\eta \phi - \alpha \nabla_\eta p. \tag{18.95}$$

Because the $\eta$-surfaces are nearly flat, however, these two terms are each comparable in magnitude to the HPGF itself, even near mountains, so the problem of near-cancellation is greatly mitigated.

### 18.4.7 Potential temperature

The potential temperature is defined by

$$\theta \equiv T\left(\frac{p_0}{p}\right)^{\kappa}. \tag{18.96}$$

The potential temperature increases upwards in a statically stable atmosphere, so that there is a monotonic relationship between $\theta$ 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 $\theta$-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, $\theta$ is conserved following a particle. This means that the vertical motion in $\theta$-coordinates is proportional to the heating rate:

$$\dot{\theta} = \frac{\theta}{c_p T} Q; \tag{18.97}$$

in the absence of heating, there is "no vertical motion," from the point of view of $\theta$-coordinates; we can also say that, in the absence of heating, a particle that is on a given $\theta$-surface remains on that surface. Eq. (18.97) is equivalent to (18.5), and is an expression of the thermodynamic energy equation in $\theta$-coordinates. In fact, $\theta$-coordinates provide an especially simple pathway for the derivation of many important results, including the conservation equation for the Ertel potential vorticity. In addition, $\theta$-coordinates prove to have some important advantages for the design of numerical models (e.g., Eliassen and Raustein (1968); Bleck (1973); Johnson and Uccellini (1983); Hoskins et al. (1985); Hsu and Arakawa (1990)).

The continuity equation in $\theta$-coordinates is given by

$$\left(\frac{\partial \rho_\theta}{\partial t}\right)_\theta + \nabla_\theta \cdot (\rho_\theta \mathbf{V}) + \frac{\partial}{\partial \theta}\left(\rho_\theta \dot{\theta}\right) = 0\,. \tag{18.98}$$

Note, however, that $\dot{\theta} = 0$ in the absence of heating; in such case, (18.98) reduces to

$$\left(\frac{\partial \rho_\theta}{\partial t}\right)_\theta + \nabla_\theta \cdot (\rho_\theta \mathbf{V}) = 0, \tag{18.99}$$

which is closely analogous to the continuity equation of a shallow-water model. In the absence of heating, a model that uses $\theta$-coordinates behaves like "a stack of shallow-water models."

The lower boundary condition in $\theta$-coordinates is

$$\frac{\partial \theta_S}{\partial t} + \mathbf{V} \cdot \nabla \theta_S - \dot{\theta}_S = 0. \tag{18.100}$$

As a reminder, this means that mass does not cross the Earth's surface. Eq. (18.100) can be used to predict $\theta_S$. The complexity of the lower boundary condition in $\theta$-coordinates is one of its chief drawbacks. This will be discussed further below.

For the case of $\theta$-coordinates, the hydrostatic equation, (18.1), reduces to

$$\frac{\partial \phi}{\partial \theta} = \alpha \frac{\partial p}{\partial \theta}. \tag{18.101}$$

Logarithmic differentiation of (18.94) gives

$$\frac{d\theta}{\theta} = \frac{dT}{T} - \kappa \frac{dp}{p}. \tag{18.102}$$

It follows that

$$\alpha \frac{\partial p}{\partial \theta} = c_p \frac{\partial T}{\partial \theta} - c_p \frac{T}{\theta}. \tag{18.103}$$

Substitution of (18.103) into (18.101) gives

$$\frac{\partial s}{\partial \theta} = \Pi, \tag{18.104}$$

where $s$ was defined in (18.11).

Following (18.10), the HPGF in $\theta$-coordinates can be written as

$$\mathbf{HPGF} = -\alpha \nabla_\theta p - \nabla_\theta \phi. \tag{18.105}$$

From (18.102) it follows that

$$\nabla_\theta p = c_p \left( \frac{p}{RT} \right) \nabla_\theta T. \tag{18.106}$$

Substitution of (18.106) into (18.105) gives

$$\mathbf{HGPF} = -\nabla_\theta s. \tag{18.107}$$

This can also be obtained directly from (18.13).

Of course, $\theta$-surfaces can intersect the lower boundary, but following Lorenz (1955) we can consider 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. 18.4.

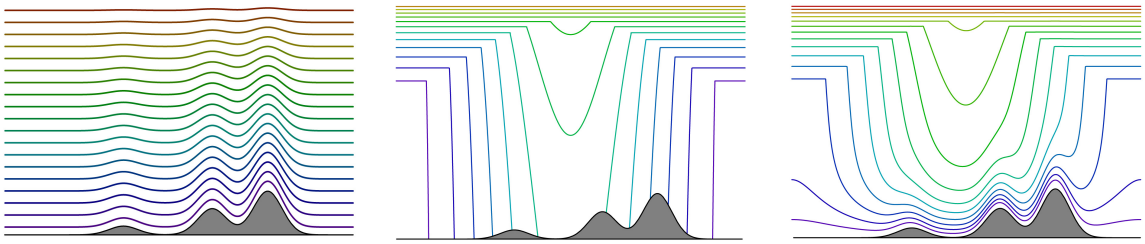


Figure 18.4: Coordinate surfaces with topography: Left, the $\sigma$-coordinate. Center, the $\theta$-coordinate. Right, a hybrid $\sigma$-$\theta$ coordinate.

Obviously, a model that follows the massless-layer approach has to avoid producing negative mass. This can be done, for example, through the use of flux-corrected transport. Nevertheless, this practical difficulty has led most modelers to avoid $\theta$-coordinates up to this time.

The massless layer approach leads us to use values of $\theta$ that are colder than any actually present in an atmospheric column, particularly in the tropics of a global model. The coldest possible value of $\theta$ is zero Kelvin. Consider the lower boundary condition on the hydrostatic equation, (18.104). We can write

$$s(\theta) - s(0) = \int_0^\theta \Pi\left(\theta'\right) d\theta', \tag{18.108}$$

where $\theta'$ 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 $\Pi_S$, i.e., the surface value of $\Pi$. We can therefore write

$$\begin{aligned} s(\theta) &= \phi_S + \int_0^{\theta_S} \Pi\left(\theta'\right) d\theta' + \int_{\theta_S}^\theta \Pi\left(\theta'\right) d\theta' \\ &= \phi_S + \Pi_S \theta_S + \int_{\theta_S}^\theta \Pi\left(\theta'\right) d\theta' \\ &= \phi_S + c_p T_S + \int_{\theta_S}^\theta \Pi\left(\theta'\right) d\theta'. \end{aligned} \tag{18.109}$$

It follows that

$$s(\theta) = s_S + \int_{\theta_S}^\theta \Pi\left(\theta'\right) d\theta', \tag{18.110}$$

as expected.

The dynamically important isentropic potential vorticity, $q$, is easily constructed in $\theta$-coordinates, since it involves the curl of $\mathbf{V}$ on a $\theta$-surface:

$$q \equiv \left(\mathbf{k} \cdot \nabla_\theta \times \mathbf{V} + f\right) \frac{\partial \theta}{\partial p}. \tag{18.111}$$

The available potential energy is also easily obtained, since it involves the distribution of pressure on $\theta$-surfaces.

### 18.4.8 Entropy

The entropy coordinate is very similar to the $\theta$-coordinate. We define the entropy by

$$\varepsilon \equiv c_p \ln \frac{\theta}{\theta_0}, \tag{18.112}$$

where $\theta_0$ is a constant reference value of $\theta$. It follows that

$$d\varepsilon = c_p \frac{d\theta}{\theta}. \tag{18.113}$$

The hydrostatic equation can then be written as

$$\frac{\partial s}{\partial \varepsilon} = T. \tag{18.114}$$

This is a particularly attractive form because the "thickness" (in terms of $s$) between two entropy surfaces is simply the temperature.

### 18.4.9 Hybrid sigma-theta coordinates

Konor and Arakawa (1997) discuss a hybrid vertical coordinate, which we will call $\hat{z}_{KA}$, that reduces to $\theta$ away from the surface, and to $\sigma$ near the surface. This hybrid coordinate is a member of the family of schemes given by (18.20). It is designed to combine the strengths of $\theta$ and $\sigma$ coordinates, while avoiding their weaknesses. Hybrid 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 (18.20). Following Konor and Arakawa (1997), define

$$\hat{z}_{KA} = F(\theta, p, p_S) \equiv f(\sigma) + g(\sigma)\,\theta, \tag{18.115}$$

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.,

$$\sigma \equiv \frac{p_S - p}{p_S}. \tag{18.116}$$

If we specify $f(\sigma)$ and $g(\sigma)$, then the hybrid coordinate is fully determined.

We require, of course, that $\hat{z}_{KA}$ itself increases upwards, so that

$$\frac{\partial \hat{z}_{KA}}{\partial \sigma} > 0. \tag{18.117}$$

We also require that

$$\hat{z}_{KA} = \text{constant for } \sigma = \sigma_S, \tag{18.118}$$

which means that $\hat{z}_{KA}$ is $\sigma$-like at the Earth's surface, and that

$$\hat{z}_{KA} = \theta \text{ for } \sigma = \sigma_T, \tag{18.119}$$

which means that $\hat{z}_{KA}$ becomes $\theta$ at the model top (or lower). These conditions imply, from (18.115), that

$$g(\sigma) \to 0 \text{ as } \sigma \to \sigma_S, \tag{18.120}$$

$$f(\sigma) \to 0 \text{ and } g(\sigma) \to 1 \text{ as } \sigma \to \sigma_T. \tag{18.121}$$

Now substitute (18.115) into (18.117), to obtain

$$\frac{df}{d\sigma} + \frac{dg}{d\sigma}\theta + g\frac{\partial \theta}{\partial \sigma} > 0. \tag{18.122}$$

This is the requirement that $\hat{z}_{KA}$ increases monotonically upward. Any choices for $f$ and $g$ that satisfy (18.120) - (18.122) 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 $\sigma$, i.e.,

$$\frac{dg}{d\sigma} > 0 \text{ for all } \sigma. \tag{18.123}$$

Since $\theta$ also increases upward, the condition (18.123) simply ensures that $g(\sigma)$ and $\theta$ change in the same sense, and the middle term on the left-hand side of (18.122) is guaranteed to be positive. We also choose $g(\sigma)$ so that the conditions (18.120) - (18.121) are satisfied. There are many possible choices for $g(\sigma)$ that meet these requirements.

Next, define $\theta_{\min}$ and $\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min}$ as lower bounds on $\theta$ and $\frac{\partial\theta}{\partial\sigma}$, respectively, so that

$$\theta > \theta_{\min} \text{ and } \frac{\partial\theta}{\partial\sigma} > \left(\frac{\partial\theta}{\partial\sigma}\right)_{\min}. \tag{18.124}$$

*When we choose the value of $\theta_{\min}$, we are saying that we have no interest in simulating situations in which $\theta$ 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 $\theta$ 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 $\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min}$, we are saying that we have no interest in simulating situations in which $\frac{\partial\theta}{\partial\sigma}$ is actually less stable (or more unstable) than $\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min}$.* We can choose $\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min} < 0$, i.e., a value of $\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min}$ that corresponds to a statically unstable sounding. Further discussion is given below.

Now, with reference to the *inequality* (18.122), we write the following *equation*:

$$\frac{df}{d\sigma} + \frac{dg}{d\sigma}\theta_{\min} + g\left(\frac{\partial\theta}{\partial\sigma}\right)_{\min} = 0. \tag{18.125}$$

Remember that $g(\sigma)$ will be specified in such a way that (18.119) is satisfied. You should be able to see that if the *equality* (18.125) is satisfied, then the *inequality* (18.122) will also be satisfied, i.e., $\hat{z}_{KA}$ will increase monotonically upward. *This will be true even if the sounding is statically unstable in some regions, provided that* (18.120) *is satisfied.*

Eq. (18.125) is a first-order ordinary differential equation for $f(\sigma)$, which can be solved subject to the boundary condition (18.121).

That's all there is to it. Amazingly, the scheme does not involve any "if-tests." It is simple and fairly flexible.

The vertical velocity is obtained using (18.24).

### 18.4.10 Summary of vertical coordinate systems

Table 18.1 summarizes key properties of some important vertical coordinate systems. All of the systems discussed here (with the exception of the entropy coordinate) have been used in many theoretical and numerical studies. Each system has its advantages and disadvantages, which must be weighed with a particular application in mind.

Table 18.1: Summary of properties of some vertical coordinate systems.

| Coordinate | Hydrostatics | HPGF | Vertical velocity | Continuity | LBC |
|---|---|---|---|---|---|
| $z$ | $\frac{\partial p}{\partial z}=-\rho g$ | $-\alpha\nabla_z p$ | $w\equiv\frac{Dz}{Dt}$ | $\frac{\partial\rho}{\partial t}+\nabla_z\cdot(\rho\mathbf{V})$ $+\frac{\partial(\rho w)}{\partial z}=0$ | $\mathbf{V}_S\cdot\nabla z_S-w_S=0$ |
| $p$ | $\frac{\partial\phi}{\partial p}=-\alpha$ | $-\nabla_p\phi$ | $\omega\equiv\frac{Dp}{Dt}$ | $\nabla_p\cdot(\mathbf{V})+\frac{\partial\omega}{\partial p}=0$ | $\frac{\partial p_S}{\partial t}+\mathbf{V}_S\cdot\nabla p_S$ $-\omega_S=0$ |
| $z^*\equiv -H\ln\left(\frac{p}{p_0}\right)$ | $\frac{\partial z}{\partial z^*}=-\frac{T}{T_0}$ | $-\nabla_{z^*}\phi$ | $w\equiv\frac{Dz^*}{Dt}$ $=-\frac{H\omega}{p}$ | $\nabla_{z^*}\cdot\mathbf{V}+\frac{\partial w^*}{\partial z^*}$ $-\frac{w^*}{H}=0$ | $\frac{\partial z_S^*}{\partial t}+\mathbf{V}_S\cdot\nabla z_S^*$ $-w_S^*=0$ |
| $\sigma\equiv\frac{p-p_T}{\pi}$ | $\frac{1}{\pi}\frac{\partial\phi}{\partial\sigma}=-\alpha$ | $-\nabla_\sigma\phi$ $-\sigma\alpha\pi$ | $\dot\sigma\equiv\frac{D\sigma}{Dt}$ | $\frac{\partial\pi}{\partial t}+\nabla_\sigma\cdot(\pi\mathbf{V})$ $+\frac{\partial(\pi\dot\sigma)}{\partial z}=0$ | $-\dot\sigma_S=0$ |
| $\theta$ | $\frac{\partial M}{\partial\theta}=\pi$ | $-\nabla_\theta M$ | $\dot\theta\equiv\frac{D\theta}{Dt}$ | $\frac{\partial m}{\partial t}+\nabla_\theta\cdot(m\mathbf{V})$ $+\frac{\partial(m\dot\theta)}{\partial\theta}=0$ | $\frac{\partial\theta_S}{\partial t}+\mathbf{V}_S\cdot\nabla\theta_S$ $-\dot\theta_S=0$ |
| $s$ | $\frac{\partial\psi}{\partial s}=T$ | $-\nabla_s M$ | $\dot s\equiv\frac{Ds}{Dt}$ | $\frac{\partial\mu}{\partial t}+\nabla_s\cdot(\mu\mathbf{V})$ $+\frac{\partial(\mu\dot s)}{\partial s}=0$ | $\frac{\partial s_S}{\partial t}+\mathbf{V}_S\cdot\nabla s_S$ $-\dot s_S=0$ |

## 18.5 Problems

1. Derive the form of the HPGF in $\sigma_p$ coordinates, and compare with the corresponding formula in $\sigma$ 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 $\sigma$ coordinate, i.e.,

$$\frac{\partial}{\partial \sigma}(\pi\dot{\sigma}) = \nabla \cdot \left( \int_0^1 \pi \mathbf{V} d\sigma \right) - \nabla_\sigma \cdot (\pi \mathbf{V}) \tag{18.126}$$

   is consistent with the method to determine the vertical velocity for a general family of schemes (that includes the $\sigma$ coordinate), i.e.,

$$\hat{w} = \frac{\frac{\partial F}{\partial \theta}\left(-\mathbf{V}\cdot\nabla_{\hat{z}}\theta + \frac{Q}{\Pi}\right) + \frac{\partial F}{\partial p}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}} \hat{\rho}\mathbf{V}d\hat{z}\right)\right] + \frac{\partial F}{\partial p_S}\left[\frac{\partial p_T}{\partial t} - \nabla\cdot\left(\int_{\hat{z}_T}^{\hat{z}_S} \hat{\rho}\mathbf{V}dp\right)\right]}{\left\{\frac{\partial \theta}{\partial \hat{z}}\frac{\partial F}{\partial \theta} - \hat{\rho}\frac{\partial F}{\partial p}\right\}}. \tag{18.127}$$

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) A method to determine the vertical velocity, modeled after the method used with $\sigma$-coordinates. Write down a "recipe" explaining how you would program the calculation of the vertical velocity.
   (c) The form of the horizontal pressure-gradient force.

# Chapter 19

# Vertical differencing

## 19.1 Vertical staggering

### 19.1.1 Lorenz vs. Charney-Phillips

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. 19.1. These are the "Lorenz" or "L" staggering, and the "Charney-Phillips" or "CP" staggering. Suppose that both grids have $N$ wind-levels. The L-grid also has $N$ $\theta$-levels, while the CP-grid has $N+1$ $\theta$-levels. On both grids, $\phi$ is hydrostatically determined on the wind-levels, and

$$\phi_l - \phi_{l+1} \sim \theta_{l+\frac{1}{2}}. \tag{19.1}$$

On the CP-grid, $\theta$ is located between $\phi$-levels, so (19.1) is convenient. With the L-grid, $\theta$ must be interpolated. For example, we might choose

$$\phi_l - \phi_{l+1} \sim \frac{1}{2}\left(\theta_l + \theta_{l+1}\right). \tag{19.2}$$

Because (19.2) involves averaging, an oscillation in $\theta$ is not "felt" by $\phi$, and so has no effect on the winds; it is dynamically inert. 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 (18.111) is

$$q_l \equiv \left(\mathbf{k}\cdot\nabla_\theta \times \mathbf{V}_l + f\right)\left(\frac{\partial\theta}{\partial p}\right)_l. \tag{19.3}$$

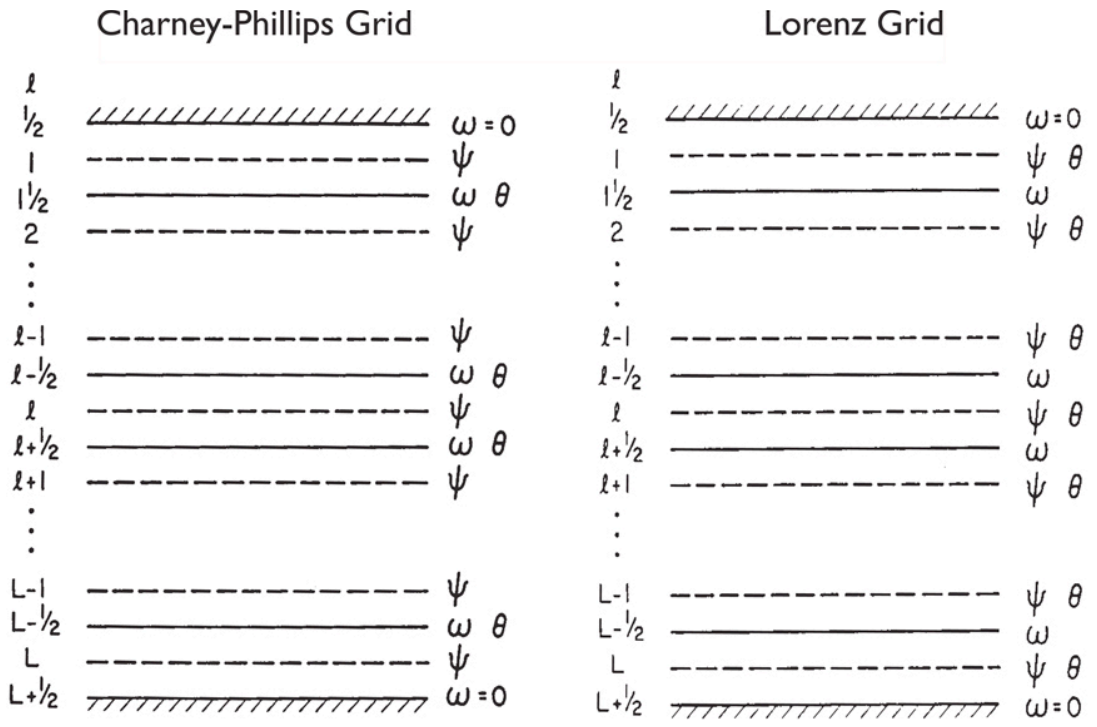



Figure 19.1: A comparison of the Lorenz and Charney-Phillips staggering methods.

Inspection shows that (19.3) "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.

### 19.1.2 The continuity equation at layer edges

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. The discsusion in this subsection very closely follows the ideas presented in Section 3b of Arakawa and Konor (1996).

Let the total number of model layers be $K$, so that $1 \leq k \leq K$; as mentioned earlier, $k$ increases downward. Here we assume for simplicity that the vertical velocity vanishes at the top and bottom edges of the model, so that

$$(\rho\sigma w)_j^{i,1/2} = (\rho\sigma w)_j^{i,K+1/2} = 0. \tag{19.4}$$

Consider a layer, $k$, with its upper edge at $k-1/2$, and its lower edge at $k+1/2$. There is a "half layer" at the top of the model, between $k=1/2$ and $k=1$, and another half layer at the bottom of the model, between $k=K$ and $k=K+1/2$. Following Arakawa and Konor (1996), we set

$$(\rho\sigma)_j^{i,k+1/2} = a^{k+1/2}(\rho\sigma)_j^{i,k} + b^{k+1/2}(\rho\sigma)_j^{i,k+1}, \tag{19.5}$$

where

$$a^{k+1/2} + b^{k+1/2} = 1 \qquad \text{for } 0 \le k \le K, \tag{19.6}$$

and

$$a^{1/2} = b^{K+1/2} = 0, \qquad \text{so that} \qquad b^{1/2} = a^{K+1/2} = 1. \tag{19.7}$$

The interpolation weights $a$ and $b$ are assumed to be independent of time and the horizontal coordinates. Eq. (19.7) makes it possible to apply (19.5) for all $k$ in the range $1 \le k \le K$. Defining

$$\bar{\rho}_j^{i,k+1/2} \equiv \sum_j (\rho\sigma)_j^{i,k+1/2} \tag{19.8}$$

and

$$\hat{\sigma}_j^{i,k+1/2} \equiv \frac{(\rho\sigma)_j^{i,k+1/2}}{\bar{\rho}_j^{i,k+1/2}}, \tag{19.9}$$

we see that

$$\sum_j \hat{\sigma}_j^{i,k+1/2} = 1. \tag{19.10}$$

With these preparations, and using (19.9), we can rewrite (19.5) as

$$(\bar{\rho}\hat{\sigma})_j^{i,k+1/2} = a^{k+1/2}(\bar{\rho}\hat{\sigma})_j^{i,k} + b^{k+1/2}(\bar{\rho}\hat{\sigma})_j^{i,k+1}, \tag{19.11}$$

Now we take $\dfrac{\partial}{\partial t}$ of (19.5) and substitute from XXX to obtain

$$\frac{\partial}{\partial t}(\bar{\rho}\hat{\sigma})_j^{i,k+1/2} = \frac{1}{A^i}\left(\sum_{j'\neq j} E_{j',j}^{i,k+1/2} - \sum_{j'\neq j} E_{j,j'}^{i,k+1/2} - G_{tot,j}^{i,k+1/2}\right)$$
$$-\frac{a^{k+1/2}}{\delta z^k}\left[(\bar{\rho}\hat{\sigma}w)_j^{i,k-1/2} - (\bar{\rho}\hat{\sigma}w)_j^{i,k+1/2}\right] - \frac{b^{k+1/2}}{\delta z^{k+1}}\left[(\bar{\rho}\hat{\sigma}w)_j^{i,k+1/2} - (\bar{\rho}\hat{\sigma}w)_j^{i,k+3/2}\right]. \tag{19.12}$$

Here we define

$$E_{j',j}^{i,k+1/2} \equiv a^{k+1/2}E_{j',j}^{i,k} + b^{k+1/2}E_{j',j}^{i,k+1}, \tag{19.13}$$

and

$$G_{j',j}^{i,k+1/2} \equiv a^{k+1/2}G_{j,j'}^{i,k} + b^{k+1/2}G_{j',j}^{i,k+1}, \tag{19.14}$$

From (19.7), it follows that

$$(\bar{\rho}\hat{\sigma})_j^{i,1/2} = (\bar{\rho}\hat{\sigma})_j^{i,1} \text{ and } (\bar{\rho}\hat{\sigma})_j^{i,K+1/2} = (\bar{\rho}\hat{\sigma})_j^{i,K}, \tag{19.15}$$

$$E_{j',j}^{i,1/2} = E_{j',j}^{i,1} \text{ and } E_{j',j}^{i,K+1/2} = E_{j',j}^{i,K}, \tag{19.16}$$

and

$$G^{i,1/2}_{j',j} = G^{i,1}_{j',j} \text{ and } G^{i,K+1/2}_{j',j} = G^{i,K}_{j',j}. \tag{19.17}$$

Inspection of (19.12) shows that the vertical mass flux *difference* centered on layer edge $k+1/2$ satisfies

$$(\bar{\rho}\hat{\sigma}w)^{i,k}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+1}_j = \\ \delta z^{k+1/2}\left\{\frac{a^{k+1/2}}{\delta z^k}\left[(\bar{\rho}\hat{\sigma}w)^{i,k-1/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+1/2}_j\right] - \frac{b^{k+1/2}}{\delta z^{k+1}}\left[(\bar{\rho}\hat{\sigma}w)^{i,k+1/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+3/2}_j\right]\right\}. \tag{19.18}$$

The analogous formulae for layer edges $k-1/2$ and $k+3/2$ are

$$(\bar{\rho}\hat{\sigma}w)^{i,k-1}_j - (\bar{\rho}\hat{\sigma}w)^{i,k}_j = \\ \delta z^{k-1/2}\left\{\frac{a^{k-1/2}}{\delta z^{k-1}}\left[(\bar{\rho}\hat{\sigma}w)^{i,k-3/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k-1/2}_j\right] - \frac{b^{k-1/2}}{\delta z^k}\left[(\bar{\rho}\hat{\sigma}w)^{i,k-1/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+1/2}_j\right]\right\}, \tag{19.19}$$

and

$$(\bar{\rho}\hat{\sigma}w)^{i,k+1}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+2}_j = \\ \delta z^{k+3/2}\left\{\frac{a^{k+3/2}}{\delta z^{k+1}}\left[(\bar{\rho}\hat{\sigma}w)^{i,k+1/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+3/2}_j\right] - \frac{b^{k+3/2}}{\delta z^{k+2}}\left[(\bar{\rho}\hat{\sigma}w)^{i,k+3/2}_j - (\bar{\rho}\hat{\sigma}w)^{i,k+5/2}_j\right]\right\}. \tag{19.20}$$

respectively. The expressions on the left-hand sides of (19.18), (19.19), and (19.20) are flux differences, which will cancel when we sum over $k$. We require that the expressions on the right-hand sides of (19.18), (19.19), and (19.20) *also* have the form of flux differences,

so that the layer-edge continuity equation will have a flux-divergence form. With this in mind, we require that when summed over $k$ in the range 1 to $K-1$ the *total* coefficient of each layer-edge value of $\bar{\rho}\hat{\sigma}w$ is zero (a very Arakawa move). Inspection shows that $(\bar{\rho}\hat{\sigma}w)_j^{i,k+1/2}$ appears on the right-hand sides of all three equations: (19.18), (19.19), and (19.20). Setting the sum of the three coefficients of $(\bar{\rho}\hat{\sigma}w)_j^{i,k+1/2}$ to zero gives

$$\delta z^{k+1/2}\left[-\frac{a^{k+1/2}}{\delta z^k}+\frac{b^{k+1/2}}{\delta z^{k+1}}\right]-\delta z^{k-1/2}\left(\frac{b^{k-1/2}}{\delta z^k}\right)+\delta z^{k+3/2}\left(\frac{a^{k+3/2}}{\delta z^{k+1}}\right)=0 \quad \text{for } 2\le k\le K-2. \tag{19.21}$$

As indicated, this condition must be satisfied for $2 \le k \le K-2$. Substitution shows that (19.21) can be satisfied by setting

$$a^{k+1/2}=\frac{1}{2}\left(\frac{\delta z^k}{\delta z^{k+1/2}}\right) \text{ and } b^{k+1/2}=\frac{1}{2}\left(\frac{\delta z^{k+1}}{\delta z^{k+1/2}}\right) \text{ for } 2\le k\le K-2. \tag{19.22}$$

We also use (19.21) for $k=1$ and $k=K-1$. Using (19.6) and (19.22), we find that the thickness of the "layer" centered on the layer edge is given by

$$\delta z^{k+1/2}=\frac{1}{2}\left(\delta z^k+\delta z^{k+1}\right). \tag{19.23}$$

Substitution of these various results into the vertical-velocity terms of (19.12) allows us to rewrite it as

$$\begin{aligned}\frac{\partial}{\partial t}\left(\bar{\rho}\hat{\sigma}\right)_j^{i,k+1/2}=&\frac{1}{A^i}\left(\sum_{j'\neq j}E_{j',j}^{i,k+1/2}-\sum_{j'\neq j}E_{j,j'}^{i,k+1/2}-\sum_{i'}\left(G_j^{i,i'}\right)^{i,k+1/2}\right)\\&-\frac{1}{\delta z^{k+1/2}}\left[\left(\bar{\rho}\hat{\sigma}w\right)_j^{i,k}-\left(\bar{\rho}\hat{\sigma}w\right)_j^{i,k+1}\right] \text{ for } 1\le k\le K-1,\end{aligned} \tag{19.24}$$

where

$$(\bar{\rho}\hat{\sigma}w)_j^{i,k} \equiv \frac{1}{2}\left[(\bar{\rho}\hat{\sigma}w)_j^{i,k-1/2} + (\bar{\rho}\hat{\sigma}w)_j^{i,k+1/2}\right] \text{ for } 2 \le k \le K-1. \tag{19.25}$$

The thicknesses of the half-layers satisfy

$$\delta z^{K+1/2} = \frac{1}{2}\delta z^K \text{ and } \delta z^{1/2} = \frac{1}{2}\delta z^1. \tag{19.26}$$

The vertical mass fluxes at levels 1 and $K$ are given by

$$(\bar{\rho}\hat{\sigma}w)_j^{i,1} = \frac{1}{2}(\bar{\rho}\hat{\sigma}w)_j^{i,3/2} \text{ and } (\bar{\rho}\hat{\sigma}w)_j^{i,K} = \frac{1}{2}(\bar{\rho}\hat{\sigma}w)_j^{i,K-1/2}. \tag{19.27}$$

These definitions for the top and bottom half-layers, along with the boundary conditions (19.4), allow us to apply (19.24) at all layer edges:

$$\begin{aligned}\frac{\partial}{\partial t}(\bar{\rho}\hat{\sigma})_j^{i,k+1/2} = \frac{1}{A^i}\left[\sum_{j'\neq j} E_{j',j}^{i,k+1/2} - \sum_{j'\neq j} E_{j,j'}^{i,k+1/2} - \sum_{i'}\left(G_j^{i,i'}\right)^{i,k+1/2}\right] \\ -\frac{1}{\delta z^{k+1/2}}\left[(\bar{\rho}\hat{\sigma}w)_j^{i,k} - (\bar{\rho}\hat{\sigma}w)_j^{i,k+1}\right] \text{ for } 0 \le k \le K.\end{aligned} \tag{19.28}$$

This is the layer-edge continuity equation.

We have shown how 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.

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 and the Canadian Environmental Multiscale model (Girard et al. (2014)), both of which use the CP-grid.

## 19.2 Conservation of total energy with continuous pressure coordinates

Even with the continuous equations, the derivation of the total energy equation in $\sigma$ coordinates is a bit complicated. It may be helpful to see the simpler derivation in pressure coordinates first.

In pressure coordinates, the starting points are the following equations, which have appeared earlier but are repeated here for convenience: Continuity is

$$\nabla_p \cdot \mathbf{V} + \frac{\partial \omega}{\partial p} = 0, \tag{19.29}$$

where

$$\omega \equiv \frac{Dp}{Dt} \tag{19.30}$$

the Lagrangian time derivative of pressure. The momentum equation is

$$\left(\frac{\partial \mathbf{V}}{\partial t}\right)_p + [f + \mathbf{k} \cdot (\nabla_p \times \mathbf{V})]\,\mathbf{k} \times \mathbf{V} + \omega \frac{\partial \mathbf{V}}{\partial p} + \nabla_p K = -\nabla_p \phi. \tag{19.31}$$

We have assumed no friction for simplicity. Potential temperature conservation is expressed by

$$\left(\frac{\partial \theta}{\partial t}\right)_p + \nabla_p \cdot (\mathbf{V}\theta) + \frac{\partial}{\partial p}(\omega\theta) = 0. \tag{19.32}$$

Here we have omitted the heating term, for simplicity. Hydrostatics is

$$\frac{\partial \phi}{\partial p} = -\alpha, \tag{19.33}$$

where $\alpha \equiv 1/\rho$ is the specific volume. Finally, we need the definition of $\theta$ and the equation of state:

$$c_p T = \Pi\theta, \tag{19.34}$$

$$p = \rho R T. \tag{19.35}$$

In (19.34),

$$\Pi \equiv c_p \left(\frac{p}{p_0}\right)^{\kappa} \tag{19.36}$$

is the Exner function, with $\kappa \equiv R/c_p$.

Using continuity, (19.32) can be expressed in advective form:

$$\left(\frac{\partial\theta}{\partial t}\right)_p + \mathbf{V}\cdot\nabla_p\theta + \omega\frac{\partial\theta}{\partial p} = 0. \tag{19.37}$$

By logarithmic differentiation of (19.34), and with the use of (19.33), (19.35), and (19.36), we can write (19.37) in terms of temperature, as follows:

$$\begin{aligned} c_p\left[\left(\frac{\partial T}{\partial t}\right)_p + \mathbf{V}\cdot\nabla_p T + \omega\frac{\partial T}{\partial p}\right] &= \frac{c_p T}{\Pi}\left[\left(\frac{\partial \Pi}{\partial t}\right)_p + \mathbf{V}\cdot\nabla_p \Pi + \omega\frac{\partial \Pi}{\partial p}\right] \\ &= \frac{c_p T\kappa}{p}\left[\left(\frac{\partial p}{\partial t}\right)_p + \mathbf{V}\cdot\nabla_p p + \omega\frac{\partial p}{\partial p}\right] \\ &= \omega\alpha. \end{aligned} \tag{19.38}$$

Continuity then allows us to transform (19.38) to the flux form:

$$\boxed{c_p\left[\left(\frac{\partial T}{\partial t}\right)_p + \nabla_p\cdot(\mathbf{V}T) + \frac{\partial}{\partial p}(\omega T)\right] = \omega\alpha}. \tag{19.39}$$

The $\omega\alpha$ term on the right-hand side of (19.39) represents conversion between thermodynamic energy and mechanical energy.

Next, we derive a suitable form of the kinetic energy equation. Dotting the equation of horizontal motion (19.31) with the horizontal wind vector, we obtain

$$\left(\frac{\partial K}{\partial t}\right)_p + \mathbf{V}\cdot\nabla_p K + \omega\frac{\partial K}{\partial p} = -\mathbf{V}\cdot\nabla_p\phi. \tag{19.40}$$

where

$$K \equiv \frac{1}{2}(\mathbf{V}\cdot\mathbf{V}) \tag{19.41}$$

is the kinetic energy per unit mass. The corresponding flux form is

$$\left(\frac{\partial K}{\partial t}\right)_p + \nabla_p\cdot(\mathbf{V}K) + \frac{\partial}{\partial p}(\omega K) = -\mathbf{V}\cdot\nabla_p\phi. \tag{19.42}$$

The pressure-work term on the right-hand side of (19.42) has to be manipulated to facilitate comparison with (19.39). We write

$$\begin{aligned} -\mathbf{V}\cdot\nabla_p\phi &= -\nabla_p\cdot(\mathbf{V}\phi) + \phi\nabla_p\cdot\mathbf{V} \\ &= -\nabla_p\cdot(\mathbf{V}\phi) - \phi\frac{\partial\omega}{\partial p} \\ &= -\nabla_p\cdot(\mathbf{V}\phi) - \frac{\partial}{\partial p}(\omega\phi) + \omega\frac{\partial\phi}{\partial p} \\ &= -\nabla_p\cdot(\mathbf{V}\phi) - \frac{\partial}{\partial p}(\omega\phi) - \omega\alpha. \end{aligned} \tag{19.43}$$

Here we have used first continuity and then hydrostatics. Substituting (19.43) back into (19.42), and collecting terms, we obtain the kinetic energy equation in the form

$$\boxed{\left(\frac{\partial K}{\partial t}\right)_p + \nabla_p \cdot [\mathbf{V}(K+\phi)] + \frac{\partial}{\partial p}[\omega(K+\phi)] = -\omega\alpha}. \tag{19.44}$$

Finally, adding (19.39) and (19.44) gives a statement of the conservation of total energy:

$$\boxed{\left[\frac{\partial}{\partial t}(K+c_pT)\right]_p + \nabla_p \cdot [\mathbf{V}(K+\phi+c_pT)] + \frac{\partial}{\partial p}[\omega(K+\phi+c_pT)] = 0}. \tag{19.45}$$

Note that the energy conversion terms of (19.39) and (19.44) have cancelled.

## 19.3 Conservation of total energy with continuous sigma coordinates

We now present the corresponding derivation using $\sigma$ coordinates. The steps involved are basically "the same" as those used in $p$ coordinates, but a little more complicated. The starting equations are

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{V}) + \frac{\partial(\pi\dot{\sigma})}{\partial \sigma} = 0, \tag{19.46}$$

$$\begin{aligned}\omega &\equiv \frac{Dp}{Dt} \\ &= \left(\frac{\partial p}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma p + \dot{\sigma}\frac{\partial p}{\partial \sigma} \\ &= \sigma\left(\frac{\partial \pi}{\partial t} + \mathbf{V}\cdot\nabla\pi\right) + \pi\dot{\sigma},\end{aligned} \tag{19.47}$$

$$\left(\frac{\partial \mathbf{V}}{\partial t}\right)_\sigma + [f + \mathbf{k}\cdot(\nabla_\sigma \times \mathbf{V})]\,\mathbf{k}\times\mathbf{V} + \dot{\sigma}\frac{\partial \mathbf{V}}{\partial \sigma} + \nabla_\sigma K = -\sigma\alpha\nabla\pi - \nabla_\sigma\phi, \tag{19.48}$$

$$\left[\frac{\partial}{\partial t}(\pi\theta)\right]_\sigma + \nabla_\sigma \cdot (\pi \mathbf{V}\theta) + \frac{\partial}{\partial \sigma}(\pi\dot{\sigma}\theta) = 0, \tag{19.49}$$

$$\frac{\partial \phi}{\partial \sigma} = -\pi\alpha. \tag{19.50}$$

Using continuity, (19.49) can be expressed in advective form:

$$\left(\frac{\partial \theta}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma\theta + \dot{\sigma}\frac{\partial\theta}{\partial\sigma} = 0. \tag{19.51}$$

By logarithmic differentiation of (19.34), and with the use of (19.35), (19.36), and (19.50), we can write (19.51) in terms of temperature, as follows:

$$\begin{aligned} c_p\left[\left(\frac{\partial T}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma T + \dot{\sigma}\frac{\partial T}{\partial\sigma}\right] &= \frac{c_p T}{\Pi}\left[\left(\frac{\partial \Pi}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma \Pi + \dot{\sigma}\frac{\partial \Pi}{\partial\sigma}\right] \\ &= \frac{c_p T\kappa}{p}\left[\left(\frac{\partial p}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma p + \dot{\sigma}\frac{\partial p}{\partial\sigma}\right] \\ &= \sigma\alpha\left(\frac{\partial\pi}{\partial t} + \mathbf{V}\cdot\nabla\pi + \pi\dot{\sigma}\right) \\ &= \omega\alpha. \end{aligned} \tag{19.52}$$

Continuity allows us to rewrite (19.52) in flux form:

$$\boxed{\left[\frac{\partial}{\partial t}(\pi c_p T)\right]_\sigma + \nabla_\sigma\cdot(\pi\mathbf{V}c_pT) + \frac{\partial}{\partial\sigma}(\pi\dot{\sigma}c_pT) = \sigma\pi\alpha\left(\frac{\partial\pi}{\partial t} + \mathbf{V}\cdot\nabla\pi + \pi\dot{\sigma}\right)}. \tag{19.53}$$

Here we have used

$$\pi\omega\alpha = \sigma\pi\alpha\left(\frac{\partial\pi}{\partial t} + \mathbf{V}\cdot\nabla\pi + \pi\dot{\sigma}\right). \tag{19.54}$$

The product "sigma pi alpha" before the parentheses on the right-hand side of (19.54) has been called "SPA" in model codes that I have worked with. At first sight, I thought it was pretty mysterious.

To derive the kinetic energy equation in $\sigma$ coordinates, we dot (19.48) with $\mathbf{V}$ to obtain

$$\left(\frac{\partial K}{\partial t}\right)_\sigma + \mathbf{V}\cdot\nabla_\sigma K + \dot{\sigma}\frac{\partial K}{\partial\sigma} = -\mathbf{V}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi). \tag{19.55}$$

The corresponding flux form is

$$\left[\frac{\partial(\pi K)}{\partial t}\right]_\sigma + \nabla_\sigma\cdot(\pi\mathbf{V}K) + \frac{\partial(\pi\dot{\sigma}K)}{\partial\sigma} = -\pi\mathbf{V}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi). \tag{19.56}$$

The pressure-work term on the right-hand side of (19.56) has to be manipulated to facilitate comparison with (19.53). Begin as follows:

$$\begin{aligned}
-\pi\mathbf{V}\cdot(\nabla_\sigma\phi + \sigma\alpha\nabla\pi) &= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi) + \phi\nabla_\sigma\cdot(\pi\mathbf{V}) - \pi\sigma\alpha\mathbf{V}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi) - \phi\left[\frac{\partial\pi}{\partial t} + \frac{\partial(\pi\dot{\sigma})}{\partial\sigma}\right] - \pi\sigma\alpha\mathbf{V}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi) - \frac{\partial(\pi\dot{\sigma}\phi)}{\partial\sigma} + \pi\dot{\sigma}\frac{\partial\phi}{\partial\sigma} - \phi\frac{\partial\pi}{\partial t} - \pi\sigma\alpha\mathbf{V}\cdot\nabla\pi \\
&= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi) - \frac{\partial(\pi\dot{\sigma}\phi)}{\partial\sigma} - \left(\pi\dot{\sigma}\alpha\pi + \phi\frac{\partial\pi}{\partial t} + \pi\sigma\alpha\mathbf{V}\cdot\nabla\pi\right).
\end{aligned} \tag{19.57}$$

To get the second line of (19.57) 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 $\omega$ is given by (19.47). With this in mind, we rewrite the last three terms (in parentheses) on the bottom line of (19.57) as follows:

$$\begin{aligned}\pi\dot{\sigma}\alpha\pi+\phi\frac{\partial\pi}{\partial t}+\pi\sigma\alpha\mathbf{V}\cdot\nabla\pi &= \pi\omega\alpha-\pi\alpha\left[\sigma\left(\frac{\partial\pi}{\partial t}+\mathbf{V}\cdot\nabla\pi\right)+\pi\dot{\sigma}\right]+\pi\dot{\sigma}\alpha\pi+\phi\frac{\partial\pi}{\partial t}+\pi\sigma\alpha\mathbf{V}\cdot\nabla\pi\\ &= \pi\omega\alpha-\pi\alpha\sigma\left(\frac{\partial\pi}{\partial t}\right)+\phi\frac{\partial\pi}{\partial t}\\ &= \pi\omega\alpha+\left(\frac{\partial\phi}{\partial\sigma}\sigma+\phi\right)\frac{\partial\pi}{\partial t}\\ &= \pi\omega\alpha+\frac{\partial}{\partial\sigma}\left(\phi\sigma\frac{\partial\pi}{\partial t}\right).\end{aligned} \tag{19.58}$$

What is the $\frac{\partial}{\partial\sigma}\left(\phi\sigma\frac{\partial\pi}{\partial t}\right)$ term doing on the last line of (19.58)? It is a contribution to the vertical pressure-work term. Substituting (19.58) back into (19.57), we conclude that

$$\begin{aligned}-\pi\mathbf{V}\cdot(\nabla_\sigma\phi+\sigma\alpha\nabla\pi) &= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi)-\frac{\partial(\pi\dot{\sigma})}{\partial\sigma}-\left[\pi\omega\alpha+\frac{\partial}{\partial\sigma}\left(\phi\sigma\frac{\partial\pi}{\partial t}\right)\right]\\ &= -\nabla_\sigma\cdot(\pi\mathbf{V}\phi)-\frac{\partial}{\partial\sigma}\left[\phi\left(\pi\dot{\sigma}+\sigma\frac{\partial\pi}{\partial t}\right)\right]-\pi\omega\alpha.\end{aligned} \tag{19.59}$$

This can be compared with (19.43). Using (19.59) in (19.56), we obtain the kinetic energy equation in the form

$$\boxed{\left[\frac{\partial(\pi K)}{\partial t}\right]_\sigma+\nabla_\sigma\cdot[\pi\mathbf{V}(K+\phi)]+\frac{\partial}{\partial\sigma}\left[\pi\dot{\sigma}K+\phi\left(\pi\dot{\sigma}+\sigma\frac{\partial\pi}{\partial t}\right)\right]=-\sigma\pi\alpha\left(\frac{\partial\pi}{\partial t}+\mathbf{V}\cdot\nabla\pi+\pi\dot{\sigma}\right)}, \tag{19.60}$$

where (19.54) has been used.

We can now add (19.60) and (19.53) to obtain the total energy equation in $\sigma$ coordinates:

$$\boxed{\left\{\frac{\partial}{\partial t}[\pi(K+c_pT)]\right\}_\sigma+\nabla_\sigma\cdot[\pi\mathbf{V}(K+c_pT+\phi)]+\frac{\partial}{\partial\sigma}\left[\pi\dot{\sigma}(K+c_pT)+\phi\left(\pi\dot{\sigma}+\sigma\frac{\partial\pi}{\partial t}\right)\right]=0}. \tag{19.61}$$

Compare with (19.45).

## 19.4 Total energy conservation as seen in generalized coordinates

Finally, we do the derivation using the generalized $\zeta$ coordinate. The starting equations are

$$\left(\frac{\partial \rho_\zeta}{\partial t}\right)_\zeta + \nabla_\zeta \cdot (\rho_\zeta \mathbf{V}) + \frac{\partial}{\partial \zeta}\left(\rho_\zeta \dot{\zeta}\right) = 0, \tag{19.62}$$

$$\begin{aligned} \omega &\equiv \frac{Dp}{Dt} \\ &= \left(\frac{\partial p}{\partial t}\right)_\zeta + \mathbf{V}\cdot\nabla_\zeta p + \dot{\zeta}\frac{\partial p}{\partial \zeta} \\ &= \left(\frac{\partial p}{\partial t}\right)_\zeta + \mathbf{V}\cdot\nabla_\zeta p - \rho_\zeta \dot{\zeta}, \end{aligned} \tag{19.63}$$

$$\left(\frac{\partial \mathbf{V}}{\partial t}\right)_\zeta + \left[f + \mathbf{k}\cdot\left(\nabla_\varsigma \times \mathbf{V}\right)\right]\mathbf{k}\times\mathbf{V} + \nabla_\varsigma K + \dot{\zeta}\frac{\partial \mathbf{V}}{\partial \zeta} = -\nabla_\zeta \phi - \frac{1}{\rho_\zeta}\frac{\partial \phi}{\partial \zeta}\nabla_\zeta p, \tag{19.64}$$

$$\left[\frac{\partial\left(\rho_\zeta \theta\right)}{\partial t}\right]_\zeta + \nabla_\zeta \cdot \left(\rho_\zeta \mathbf{V}\theta\right) + \frac{\partial}{\partial \zeta}\left(\rho_\zeta \dot{\zeta}\theta\right) = 0, \tag{19.65}$$

$$\frac{\partial \phi}{\partial \zeta} = \alpha \rho_\zeta. \tag{19.66}$$

Using continuity, (19.65) can be expressed in advective form:

$$\left(\frac{\partial \theta}{\partial t}\right)_\zeta + \mathbf{V}\cdot\nabla_\zeta \theta + \dot{\zeta}\frac{\partial \theta}{\partial \zeta} = 0. \tag{19.67}$$

By logarithmic differentiation of (19.34), and with the use of (19.35), (19.36), and (19.66), we can write the thermodynamic energy equation in terms of temperature, as follows:

$$\begin{aligned} c_p\left[\left(\frac{\partial T}{\partial t}\right)_\varsigma + \mathbf{V}\cdot\nabla_\varsigma T + \dot{\varsigma}\frac{\partial T}{\partial \varsigma}\right] &= \frac{c_p T}{\Pi}\left[\left(\frac{\partial \Pi}{\partial t}\right)_\varsigma + \mathbf{V}\cdot\nabla_\varsigma \Pi + \dot{\varsigma}\frac{\partial \Pi}{\partial \varsigma}\right] \\ &= \frac{c_p T\kappa}{p}\left[\left(\frac{\partial p}{\partial t}\right)_\varsigma + \mathbf{V}\cdot\nabla_\varsigma p + \dot{\varsigma}\frac{\partial p}{\partial \varsigma}\right] \\ &= \omega\alpha. \end{aligned} \tag{19.68}$$

Continuity allows us to rewrite (19.68) in flux form:

$$\boxed{\left[\frac{\partial}{\partial t}\left(\rho_\zeta c_p T\right)\right]_\varsigma + \nabla_\varsigma\cdot\left(\rho_\zeta \mathbf{V} c_p T\right) + \frac{\partial}{\partial \varsigma}\left(\rho_\zeta \dot{\varsigma} c_p T\right) = \rho_\zeta \omega\alpha}. \tag{19.69}$$

To derive the kinetic energy equation in $\zeta$ coordinates, we dot (19.64) with $\mathbf{V}$ to obtain

$$\left(\frac{\partial K}{\partial t}\right)_\varsigma + \mathbf{V}\cdot\nabla_\varsigma K + \dot{\varsigma}\frac{\partial K}{\partial \varsigma} = \mathbf{V}\cdot\left(-\nabla_\zeta\phi - \frac{1}{\rho_\zeta}\frac{\partial\phi}{\partial\zeta}\nabla_\zeta p\right). \tag{19.70}$$

The corresponding flux form is

$$\left[\frac{\partial\left(\rho_\zeta K\right)}{\partial t}\right]_\varsigma + \nabla_\varsigma\cdot\left(\rho_\zeta \mathbf{V} K\right) + \frac{\partial\left(\rho_\zeta\dot{\varsigma}K\right)}{\partial\varsigma} = \rho_\zeta\mathbf{V}\cdot\left(-\nabla_\zeta\phi - \frac{1}{\rho_\zeta}\frac{\partial\phi}{\partial\zeta}\nabla_\zeta p\right). \tag{19.71}$$

The pressure-work term on the right-hand side of (19.71) has to be manipulated to facilitate comparison with (19.69). Begin as follows:

$$\begin{aligned}\rho_\zeta \mathbf{V}\cdot\left(-\nabla_\zeta\phi-\frac{1}{\rho_\zeta}\frac{\partial\phi}{\partial\zeta}\nabla_\zeta p\right) &= -\nabla_\zeta\cdot\left(\rho_\zeta\mathbf{V}\phi\right)+\phi\nabla_\zeta\cdot\left(\rho_\zeta\mathbf{V}\right)-\frac{\partial\phi}{\partial\zeta}\mathbf{V}\cdot\nabla_\zeta p\\ &= -\nabla_\varsigma\cdot\left(\rho_\zeta\mathbf{V}\phi\right)-\phi\left[\left(\frac{\partial\rho_\zeta}{\partial t}\right)_\zeta+\frac{\partial}{\partial\zeta}\left(\rho_\zeta\dot{\zeta}\right)\right]-\frac{\partial\phi}{\partial\zeta}\mathbf{V}\cdot\nabla_\zeta p\\ &= -\nabla_\varsigma\cdot\left(\rho_\zeta\mathbf{V}\phi\right)-\frac{\partial\left(\rho_\zeta\dot{\zeta}\phi\right)}{\partial\zeta}+\rho_\zeta\dot{\zeta}\frac{\partial\phi}{\partial\zeta}-\phi\left(\frac{\partial\rho_\zeta}{\partial t}\right)_\zeta-\frac{\partial\phi}{\partial\zeta}\mathbf{V}\cdot\nabla_\zeta p.\end{aligned} \tag{19.72}$$

To complete the derivation, we write

$$\begin{aligned}\phi\left(\frac{\partial\rho_\zeta}{\partial t}\right)_\zeta &= \phi\left[\frac{\partial}{\partial t}\left(-\frac{\partial p}{\partial\zeta}\right)\right]_\zeta\\ &= -\phi\frac{\partial}{\partial\zeta}\left[\left(\frac{\partial p}{\partial t}\right)_\zeta\right]\\ &= -\frac{\partial}{\partial\zeta}\left[\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]+\frac{\partial\phi}{\partial\zeta}\left(\frac{\partial p}{\partial t}\right)_\zeta.\end{aligned} \tag{19.73}$$

Substituting (19.73) into (19.72), and rearranging, we obtain

$$\begin{aligned}\rho_\zeta \mathbf{V}\cdot\left(-\nabla_\zeta\phi-\frac{1}{\rho_\zeta}\frac{\partial\phi}{\partial\zeta}\nabla_\zeta p\right) &= -\nabla_\zeta\cdot\left(\rho_\zeta\mathbf{V}\phi\right)-\frac{\partial}{\partial\zeta}\left[\rho_\zeta\dot{\zeta}\phi-\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]\\ &\quad+\frac{\partial\phi}{\partial\zeta}\left[\rho_\zeta\dot{\zeta}-\left(\frac{\partial p}{\partial t}\right)_\zeta-\mathbf{V}\cdot\nabla_\zeta p\right]\\ &= -\nabla_\zeta\cdot\left(\rho_\zeta\mathbf{V}\phi\right)-\frac{\partial}{\partial\zeta}\left[\rho_\zeta\dot{\zeta}\phi-\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]\\ &\quad-\alpha\rho_\zeta\left[\left(\frac{\partial p}{\partial t}\right)_\zeta+\mathbf{V}\cdot\nabla_\zeta p+\frac{\partial p}{\partial\zeta}\dot{\zeta}\right]\\ &= -\nabla_\zeta\cdot\left(\rho_\zeta\mathbf{V}\phi\right)-\frac{\partial}{\partial\zeta}\left[\rho_\zeta\dot{\zeta}\phi-\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]-\rho_\zeta\omega\alpha.\end{aligned} \tag{19.74}$$

Substituting back into (19.71), we obtain the kinetic energy equation in the form

$$\boxed{\left[\frac{\partial\left(\rho_\zeta K\right)}{\partial t}\right]_\varsigma+\nabla_\varsigma\cdot\left[\rho_\zeta\mathbf{V}\left(K+\phi\right)\right]+\frac{\partial}{\partial\varsigma}\left[\rho_\zeta\dot{\zeta}\left(K+\phi\right)-\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]=-\rho_\zeta\omega}. \tag{19.75}$$

We can now add (19.75) and (19.69) to obtain the total energy equation in $\zeta$ coordinates:

$$\boxed{\left\{\frac{\partial}{\partial t}\left[\rho_\zeta\left(K+c_pT\right)\right]\right\}_\zeta+\nabla_\zeta\cdot\left[\rho_\zeta\mathbf{V}\left(K+c_pT+\phi\right)\right]+\frac{\partial}{\partial\varsigma}\left[\rho_\zeta\dot{\zeta}\left(K+c_pT+\phi\right)-\phi\left(\frac{\partial p}{\partial t}\right)_\zeta\right]=0}. \tag{19.76}$$

## 19.5 Conservation properties of vertically discrete models using sigma-coordinates

We now investigate conservation properties of the vertically discrete equations, using $\sigma$-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

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{V}_l) + \left[ \frac{\delta (\pi \dot{\sigma})}{\delta \sigma} \right]_l = 0, \tag{19.77}$$

where

$$[\delta ( )]_l \equiv ( )_{l+\frac{1}{2}} - ( )_{l-\frac{1}{2}}. \tag{19.78}$$

Similarly, conservation of potential temperature is expressed, in flux form, by

$$\frac{\partial (\pi \theta_l)}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{V}_l \theta_l) + \left[ \frac{\delta (\pi \dot{\sigma} \theta)}{\delta \sigma} \right]_l = 0. \tag{19.79}$$

Here we omit the heating term, for simplicity. In order to use (19.79) it is necessary to define values of $\theta$ at the layer edges, via an interpolation. In Chapter 4 we discussed the interpolation issue in the context of horizontal advection, and that 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

$$\frac{\delta \phi}{\delta \sigma} = \pi \alpha_l. \tag{19.80}$$

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

$$\frac{\partial \mathbf{V}_l}{\partial t} + [f + \mathbf{k} \cdot (\nabla_\sigma \times \mathbf{V}_l)] \mathbf{k} \times \mathbf{V}_l + \left( \dot{\sigma} \frac{\partial \mathbf{V}}{\partial \sigma} \right)_l + \nabla K_l = -\nabla \phi_l - (\sigma \alpha)_l \nabla \pi. \tag{19.81}$$

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

$$\dot{\sigma}_{\frac{1}{2}} = \dot{\sigma}_{L+\frac{1}{2}} = 0. \tag{19.82}$$

We define the vertical coordinate, $\sigma$, at layer edges, which are denoted by half-integer subscripts. The change in $\sigma$ across layer $l$ is written as $\delta\sigma_l$. Note that

$$\sum_{l=1}^{L} \delta \sigma_l = 1, \tag{19.83}$$

$$p_{l+\frac{1}{2}} = \pi \sigma_{l+\frac{1}{2}} + p_T, \tag{19.84}$$

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. (19.84) tells how to compute layer-edge pressures. A method to determine layer-center pressures is also needed, and will be discussed later.

By summing (19.77) over all layers, and using (19.82), we obtain

$$\frac{\partial \pi}{\partial t} + \nabla \cdot \left\{ \sum_{l=1}^{L} \left[ (\pi \mathbf{V}_l) (\delta \sigma_l) \right] \right\} = 0, \tag{19.85}$$

which is the vertically discrete form of the surface pressure tendency equation. From (19.85), 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 (19.85) with (19.77) 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.

### 19.5.1 The horizontal pressure-gradient force

Consider the HPGF, in connection with (18.84) and (18.86). A finite-difference analog of (18.86) is

$$\pi(\mathbf{HPGF})_l = \left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l \nabla\pi - \nabla(\pi\phi_l). \tag{19.86}$$

Multiplying (19.86) by $\delta\sigma_l$, and summing over all layers, we obtain

$$\begin{aligned}\sum_{l=1}^{L} \pi(\mathbf{HPGF})_l(\delta\sigma)_l &= \sum_{l=1}^{L} [\delta(\sigma\phi)]_l \nabla\pi - \sum_{l=1}^{L} [\nabla(\pi\phi_l)(\delta\sigma)_l] \\ &= \phi_S\nabla\pi - \nabla\left\{\sum_{l=1}^{L} [(\pi\phi_l)(\delta\sigma)_l]\right\}.\end{aligned} \tag{19.87}$$

This is analogous to Eq. (18.16), which applies in the continuous system. Inspection of (19.82) shows that, if we use the form of the HPGF given by (19.81), the vertically summed HPGF cannot spin up or spin down a circulation inside a closed path, in the absence of topography (Arakawa and Lamb (1977)). A vertical differencing scheme of this type is often said to be "angular-momentum conserving" (e.g., Simmons and Burridge (1981)).

*The idea outlined above provides a rational way to choose which of the many possible forms of the HPGF 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 $\phi_l$ or the layer-edge values of $\phi$ that appear in (19.86).

Eq. (19.86) is equivalent to

$$\pi(\mathbf{HPGF})_l = \left\{\left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l - \phi_l\right\}\nabla\pi - \pi\nabla\phi_l. \tag{19.88}$$

By comparison with (18.84), we identify

$$\pi(\sigma\alpha)_l = \phi_l - \left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l. \tag{19.89}$$

An analogous equation is true in the continuous case. This allows us to write (19.88) as

$$\pi(\mathbf{HPGF})_l = -\pi(\sigma\alpha)_l \nabla\pi - \pi\nabla\phi_l. \tag{19.90}$$

Eq. (19.90) will be used later.

### 19.5.2 The thermodynamic energy equation

Suppose that we choose to predict $\theta_l$ by using (19.79), because we want to conserve the globally mass-integrated value of $\theta$ in the absence of heating. We relate the temperature to the potential temperature using

$$c_p T_l = \Pi_l \theta_l, \tag{19.91}$$

which corresponds to (19.34). In order to use (19.91), we need a way to determine

$$\Pi_l \equiv c_p \left(\frac{p_l}{p_0}\right)^{\kappa}. \tag{19.92}$$

Norman Phillips (1974) suggested

$$\Pi_l = \left(\frac{1}{1+\kappa}\right)\left[\frac{\delta(\Pi p)}{\delta p}\right]_l, \tag{19.93}$$

on the grounds that this form leads to a good simulation of vertical wave propagation. Eq. (19.93) 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 (19.93), 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 (19.79) with (19.77):

$$\pi\left(\frac{\partial\theta_l}{\partial t}+\mathbf{V}_l\cdot\nabla\theta_l\right)+\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(\theta_{l+\frac{1}{2}}-\theta_l\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(\theta_l-\theta_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]=0. \quad (19.94)$$

A similar manipulation was shown way back in Chapter 5. Substitute (19.91) into (19.94), to obtain the corresponding prediction equation for $T_l$:

$$\begin{aligned} &c_p\pi\left(\frac{\partial T_l}{\partial t}+\mathbf{V}_l\cdot\nabla T_l\right)-\pi\theta_l\frac{\partial\Pi_l}{\partial\pi}\left(\frac{\partial\pi}{\partial t}+\mathbf{V}_l\cdot\nabla\pi\right)\\ &+\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(\Pi_l\theta_{l+\frac{1}{2}}-c_pT_l\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(c_pT_l-\Pi_l\theta_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]=0. \end{aligned} \quad (19.95)$$

The derivative $\frac{\partial\Pi_l}{\partial\pi}$ can be evaluated using (19.93). We now introduce the terms that represent the vertical advection of temperature, modeled after the corresponding terms of (19.94). 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 (19.95) as

$$\begin{aligned} &c_p\pi\left(\frac{\partial T_l}{\partial t}+\mathbf{V}_l\cdot\nabla T_l\right)+c_p\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(T_{l+\frac{1}{2}}-T_l\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(T_l-T_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]\\ &=\pi\theta_l\frac{\partial\Pi_l}{\partial\pi}\left(\frac{\partial\pi}{\partial t}+\mathbf{V}_l\cdot\nabla\pi\right)+\left[\frac{(\pi\dot{\sigma})_{l+\frac{1}{2}}\left(c_pT_{l+\frac{1}{2}}-\Pi_l\theta_{l+\frac{1}{2}}\right)+(\pi\dot{\sigma})_{l-\frac{1}{2}}\left(\Pi_l\theta_{l-\frac{1}{2}}-c_pT_{l-\frac{1}{2}}\right)}{(\delta\sigma)_l}\right]. \end{aligned} \quad (19.96)$$

The layer-edge temperatures can simply be cancelled out in (19.90) to recover (19.89). Obviously, the left-hand side of (19.96) can be rewritten in flux form through the use of the vertically discrete continuity equation:

$$c_p \left\{ \frac{\partial}{\partial t}(\pi T_l) + \nabla \cdot (\pi \mathbf{V}_l T_l) + \left[ \frac{\delta(\pi \sigma T)}{\delta \sigma} \right]_l \right\}$$
$$= \pi \theta_l \frac{\partial \Pi_l}{\partial \pi} \left( \frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) + \frac{1}{(\delta \sigma)_l} \left[ (\pi \dot{\sigma})_{l+\frac{1}{2}} \left( c_p T_{l+\frac{1}{2}} - \Pi_l \theta_{l+\frac{1}{2}} \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left( \Pi_l \theta_{l-\frac{1}{2}} - c_p T_{l-\frac{1}{2}} \right) \right]. \tag{19.97}$$

We now observe, by comparison of (19.97) with the continuous form (19.52), that the expression on the right-hand side of (19.97) must be a form of $\pi\omega\alpha$, i.e.,

$$\boxed{\pi(\omega\alpha)_l = \pi \theta_l \frac{\partial \Pi_l}{\partial \pi} \left( \frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) + \left[ \frac{(\pi \dot{\sigma})_{l+\frac{1}{2}} \left( c_p T_{l+\frac{1}{2}} - \Pi_l \theta_{l+\frac{1}{2}} \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left( \Pi_l \theta_{l-\frac{1}{2}} - c_p T_{l-\frac{1}{2}} \right)}{(\delta \sigma)_l} \right]}. \tag{19.98}$$

Eq. (19.98) is a finite-difference analog of the not-so-obvious continuous equation

$$\pi\omega\alpha = \pi\theta \frac{\partial \Pi}{\partial \pi} \left( \frac{\partial \pi}{\partial t} + \mathbf{V} \cdot \nabla \pi \right) + \frac{\partial (\pi \dot{\sigma} c_p T)}{\partial \sigma} - \Pi \frac{\partial (\pi \dot{\sigma} \theta)}{\partial \sigma}, \tag{19.99}$$

which you should be able to prove is correct. We will return to (19.98) below, after deriving the corresponding expression from the mechanical energy side of the problem.

### 19.5.3 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 HPGF for layer $l$, we write, closely following the continuous case,

$$
\begin{aligned}
-\pi \mathbf{V}_l \cdot [\nabla \phi_l + (\sigma\alpha)_l \nabla \pi] &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) + \phi_l \nabla \cdot (\pi \mathbf{V}_l) - \pi (\sigma\alpha)_l \mathbf{V}_l \cdot \nabla \pi \\
&= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \phi_l \left\{ \frac{\partial \pi}{\partial t} + \left[ \frac{\delta (\pi \dot{\sigma} \phi)}{\delta \sigma} \right]_l \right\} - \pi (\sigma\alpha)_l \mathbf{V}_l \cdot \nabla \pi \\
&= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \left[ \frac{\delta (\pi \dot{\sigma} \phi)}{\delta \sigma} \right]_l + \left[ \frac{(\pi\sigma)_{l+\frac{1}{2}} \left( \phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left( \phi_l - \phi_{l-\frac{1}{2}} \right)}{(\delta\sigma)_l} \right] \\
&\quad - \phi_l \frac{\partial \pi}{\partial t} - \pi (\sigma\alpha)_l \mathbf{V}_l \cdot \nabla \pi.
\end{aligned}
\tag{19.100}
$$

Continuing down this path, we construct the terms that we need by adding and subtracting

$$
\begin{aligned}
-\pi \mathbf{V}_l [\nabla \phi_l + (\sigma\alpha)_l \nabla \pi] &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \left[ \frac{\delta (\pi \dot{\sigma} \phi)}{\delta \sigma} \right]_l + [\pi (\sigma\alpha)_l - \phi_l] \frac{\partial \pi}{\partial t} \\
&\quad - \pi \left\{ (\sigma\alpha)_l \left( \frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) - \left[ \frac{(\pi\dot{\sigma})_{l+\frac{1}{2}} \left( \phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left( \phi_l - \phi_{l-\frac{1}{2}} \right)}{\pi (\delta\sigma)_l} \right] \right\}.
\end{aligned}
\tag{19.101}
$$

Using the continuity equation (19.77), we can rewrite (19.101) as

$$
\begin{aligned}
-\pi \mathbf{V}_l \cdot [\nabla \phi_l + (\sigma\alpha)_l \nabla \pi] &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \left\{ \frac{\delta \left[ \left( \pi \dot{\sigma} + \sigma \frac{\partial \pi}{\partial t} \right) \phi \right]}{\delta \sigma} \right\}_l \\
&\quad - \pi \left\{ (\sigma\alpha)_l \left( \frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) - \left[ \frac{(\pi\dot{\sigma})_{l+\frac{1}{2}} \left( \phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left( \phi_l - \phi_{l-\frac{1}{2}} \right)}{\pi (\delta\sigma)_l} \right] \right\}.
\end{aligned}
\tag{19.102}
$$

By comparing with the continuous form, (19.59), we infer that

$$
\boxed{\pi(\omega\alpha)_l = \pi(\sigma\alpha)_l \left( \frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) - \frac{1}{(\delta\sigma)_l} \left[ (\pi\dot{\sigma})_{l+\frac{1}{2}} \left( \phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi\dot{\sigma})_{l-\frac{1}{2}} \left( \phi_l - \phi_{l-\frac{1}{2}} \right) \right]}.
\tag{19.103}
$$

### 19.5.4 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* (19.103) *must match that given by* (19.98). Comparison of the two equations shows that this can be accomplished by setting:

$$(\sigma\alpha)_l = \theta_l \frac{\partial \Pi_l}{\partial \pi}, \tag{19.104}$$

$$\phi_l - \phi_{l+\frac{1}{2}} = \left( c_p T_{l+\frac{1}{2}} - \Pi_l \theta_{l+\frac{1}{2}} \right), \tag{19.105}$$

and

$$\phi_{l-\frac{1}{2}} - \phi_l = \left( \Pi_l \theta_{l-\frac{1}{2}} - c_p T_{l-\frac{1}{2}} \right). \tag{19.106}$$

As discussed below, all three of these equations are vertically discrete forms of the hydrostatic equation.

Eq. (19.104) gives us an expression for $(\sigma\alpha)_l$. We already had one, though, in Eq. (19.89). By requiring that these two expressions to agree, we obtain

$$\boxed{\phi_l - \left[ \frac{\delta(\sigma\phi)}{\delta\sigma} \right]_l = \pi\theta_l \frac{\partial \Pi_l}{\partial \pi}}. \tag{19.107}$$

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 (19.107).

By adding $\Pi_l \theta_l$ to both sides of both (19.105) and (19.106), and using (19.91), we find that

$$\left( c_p T_{l+\frac{1}{2}} + \phi_{l+\frac{1}{2}} \right) - (c_p T_l + \phi_l) = \Pi_l \left( \theta_{l+\frac{1}{2}} - \theta_l \right), \tag{19.108}$$

and

$$(c_p T_l + \phi_l) - \left(c_p T_{l-\frac{1}{2}} + \phi_{l-\frac{1}{2}}\right) = \Pi_l \left(\theta_l - \theta_{l-\frac{1}{2}}\right), \tag{19.109}$$

respectively. These finite-difference analogs of the hydrostatic equation have the familiar form $\frac{\partial M}{\partial \theta} = \Pi$. Add one to each subscript in (19.109), and add the result to (19.108). This yields

$$\boxed{\phi_{l+1} - \phi_l = -\theta_{l+\frac{1}{2}} \left(\Pi_{l+1} - \Pi_l\right)}. \tag{19.110}$$

This is a finite-difference version of yet another form of the hydrostatic equation, namely $\frac{\partial \phi}{\partial \Pi} = -\theta$. What have we gained by the manipulation just performed? If the forms of $\Pi_l$ and $\theta_{l+\frac{1}{2}}$ are specified, we can use (19.110) to integrate the hydrostatic equation upward from level $l+1$ to level $l$.

### 19.5.5 The problem with the L grid

*In* (19.110)*, 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 $\theta$. The interpolation will "hide" a vertical zig-zag in $\theta$, if one is present in the solution. A hidden zig-zag cannot influence the pressure-gradient force, so it cannot participate in the model's dynamics. Therefore 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 (19.110) to integrate the hydrostatic equation upward, it is still necessary to provide a boundary condition to determine the starting value, $\phi_L$, i.e., the layer-center geopotential for the lowest layer. This can be done by first summing $(\delta\sigma)_l$ times (19.107) over all layers:

$$\sum_{l=1}^{L} \phi_l (\delta\sigma)_l - \phi_S = \sum_{l=1}^{L} \pi\theta_l \frac{\partial \Pi_l}{\partial \pi} (\delta\sigma)_l. \tag{19.111}$$

Now we use the mathematical identity

$$\begin{aligned}\sum_{l=1}^{L}\phi_l(\delta\sigma)_l &= \sum_{l=1}^{L}\phi_l\left(\sigma_{l+\frac{1}{2}}-\sigma_{l-\frac{1}{2}}\right)\\ &= \phi_L+\sum_{l=1}^{L-1}\sigma_{l+\frac{1}{2}}\left(\phi_l-\phi_{l+1}\right).\end{aligned} \tag{19.112}$$

Substitution of (19.112) into the left-hand side of (19.111), and use of (19.110), gives

$$\phi_L=\phi_S+\sum_{l=1}^{L}\pi\theta_l\frac{\partial\Pi_l}{\partial\pi}(\delta\sigma)_l-\sum_{l=1}^{L-1}\sigma_{l+\frac{1}{2}}\left(\Pi_{l+1}-\Pi_l\right)\theta_{l+\frac{1}{2}}, \tag{19.113}$$

which can be used to determine the geopotential height at the lowest layer center. We can then use (19.110) to determine the geopotential for the remaining layers above.

Eq. (19.113) 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 $\theta_l$, throughout the entire column*. An interpretation is that all values of $\theta_l$ are being used to estimate the effective value of $\theta$ between the surface and level $L$. Since we start from $\phi_L$ to determine $\phi_l$ for $l<L$, *all values of $\theta_l$ are being used to determine each value of $\phi_l$ 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 $\theta_L$ 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

$$\phi_l-\phi_{l+1}=c_p\left(A_{l+\frac{1}{2}}\theta_l+B_{l+\frac{1}{2}}\theta_{l+1}\right), \tag{19.114}$$

where $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ are non-dimensional parameters to be determined. Comparing with (19.110), we see that

$$\boxed{\left(\Pi_{l+1}-\Pi_l\right)\theta_{l+\frac{1}{2}}=A_{l+\frac{1}{2}}\theta_l+B_{l+\frac{1}{2}}\theta_{l+1}}\,. \tag{19.115}$$

In order that (19.115) have the form of an interpolation, we must choose $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ so that

$$\frac{A_{l+\frac{1}{2}}+B_{l+\frac{1}{2}}}{\Pi_{l+1}-\Pi_l}=1. \tag{19.116}$$

Eq. (19.115) 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.

*If we use* (19.115)*, we are not free to use the methods of Chapter 5 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 (19.115).

After substitution from (19.114), Eq. (19.113) becomes

$$\phi_L-\phi_S=\sum_{l=1}^{L}\pi\theta_l\frac{\partial\Pi_l}{\partial\pi}(\delta\sigma)_l-\sum_{l=1}^{L-1}\sigma_{l+\frac{1}{2}}\left(A_{l+\frac{1}{2}}\theta_l+B_{l+\frac{1}{2}}\theta_{l+1}\right). \tag{19.117}$$

Every term on the right-hand-side of (19.117) involves a layer-center value of $\theta$. *To eliminate any dependence of* $\phi_L$ *on the values of* $\theta$ *above the lowest layer, we "collect terms" around individual values of* $\theta_l$*, and force the coefficients to vanish for* $l<L$. This leads to

$$\pi\frac{\partial\Pi_l}{\partial\pi}(\delta\sigma)_l=\sigma_{l+\frac{1}{2}}A_{l+\frac{1}{2}}+\sigma_{l-\frac{1}{2}}B_{l-\frac{1}{2}}\text{ for }l<L. \tag{19.118}$$

With the use of (19.118), (19.117) simplifies to

$$\boxed{\phi_L=\phi_S+\left[\pi\frac{\partial\Pi_l}{\partial\pi}(\delta\sigma)_L-\sigma_{L-\frac{1}{2}}B_{L-\frac{1}{2}}\right]c_p\theta_L}. \tag{19.119}$$

Because the coefficient of each $\theta_l$ has been forced to vanish for all $l<L$, only $\theta_L$ influences $\phi_L$. 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 (19.119), we can use either (19.110) or (19.114) 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

$$A_{l+\frac{1}{2}} = \Pi_{l+\frac{1}{2}} - \Pi_l \text{ and } B_{l+\frac{1}{2}} = \Pi_{l+1} - \Pi_{l+\frac{1}{2}}, \tag{19.120}$$

which satisfy (19.116).

## 19.6 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.

## 19.7 Problems

1. Assume that the surface elevation is given by $\frac{\partial z_S}{\partial x} = A\left[1+\sin\left(\frac{2\pi x}{L}\right)\right]$. Also assume that the temperature is independent of height, and equal to 250 K, and that the horizontal temperature gradient is given by $\frac{\partial T}{\partial x} = \left(10^{-5}\ \mathrm{K\,m^{-1}}\right)\sin\left(\frac{2\pi x}{M}\right)$ 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), for all

# Chapter 20

# When the advector is the advectee

## 20.1 Introduction

If the wind field is specified, as for example in the discussion of Chapter 5, 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 9, 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, in which the important new physical ingredient that must be considered is rotation, including both Earth-rotation, $f$, and the relative vorticity, $\zeta$, that is associated with the wind field. Vorticity is key to almost all atmospheric dynamics, on both large and small scales.

## 20.2 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

$$A\left(x_j\right) = \hat{A}e^{ikj\Delta x} \text{ and } B\left(x_j\right) = \hat{B}e^{ilj\Delta x}, \tag{20.1}$$

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

$$\alpha A + \beta B, \tag{20.2}$$

where $\alpha$ and $\beta$ 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, $k+l$:

$$AB = \hat{A}\hat{B}e^{i(k+l)j\Delta x}, \tag{20.3}$$

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.*

### 20.2.1 Aliasing error

Suppose that we have a wave given by the continuous solid line in Fig. 20.1. 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. 20.1, 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.*

### 20.2.2 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. 20.2 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

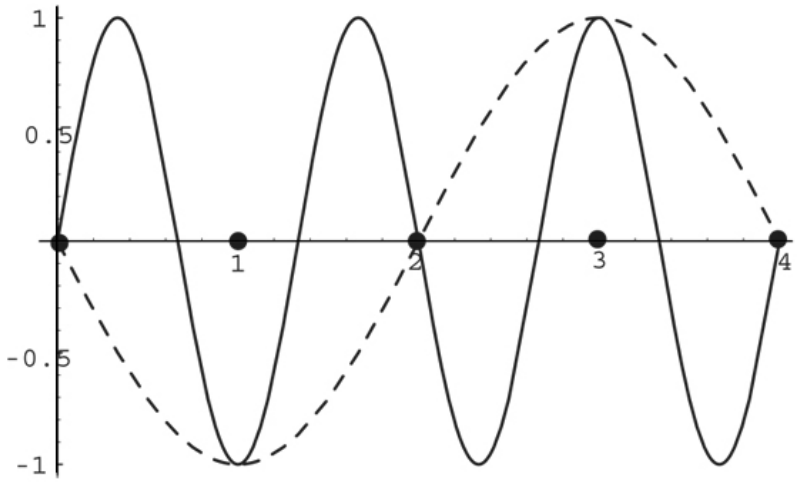



Figure 20.1: An example of aliasing error. Distance along the horizontal axis is measured in units of $\Delta x$. The wave given by the solid line has a wave length of $(4/3)\,\Delta x$. This is shorter than $2\Delta x$, and so the wave cannot be represented on the grid. Instead, the grid "sees" a wave of wavelength $4\Delta x$, as indicated by the dashed line. Note that the $4\Delta x$-wave is "upside-down."

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. 20.2, 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.

### 20.2.3 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 $k_{\max} \equiv \pi/\Delta x$. What happens when a wave with $k > k_{\max}$ is produced through nonlinear interactions? Since $2k_{\max}\Delta x = 2\pi$, a wave with $k > 2k_{\max} = \frac{2\pi}{\Delta x}$ "folds back." We can therefore assume that

$$2k_{\max} > k > k_{\max}. \tag{20.4}$$

The expression $\sin(kj\Delta x)$ can be written as as

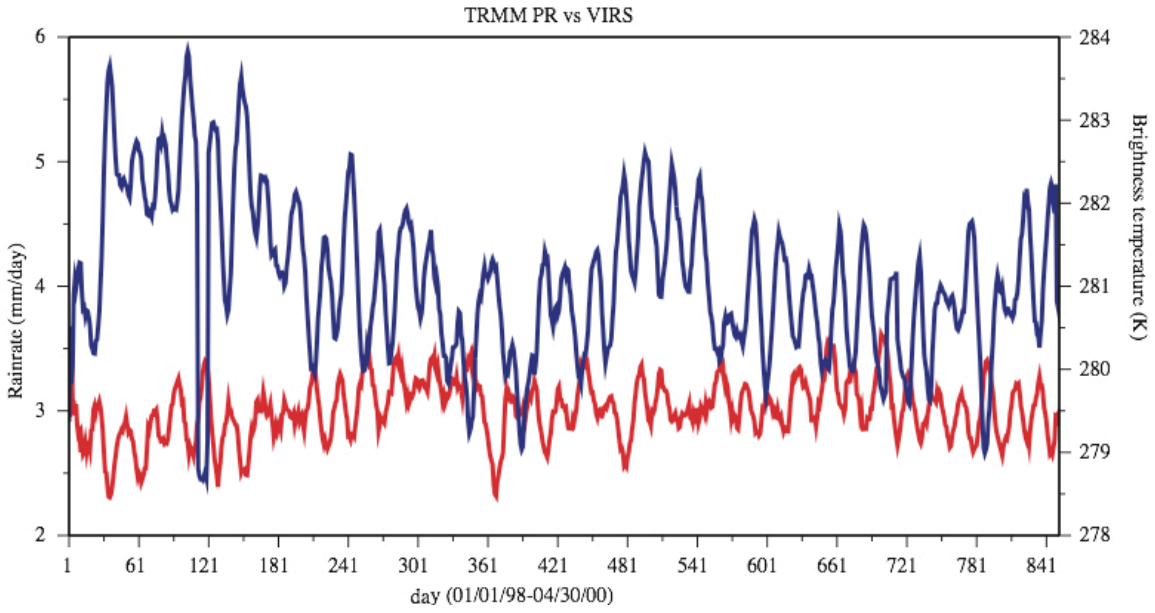



Figure 20.2: An example of aliasing in the analysis of observations. The blue curve shows the precipitation rate, averaged over the global tropics (20° S to 20° N), and the red curve shows a the thermal radiation in the 11.8 $\mu$m band, averaged over the same region. The horizontal axis is time, and 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.

$$\begin{aligned}
\sin\left[k\left(j\Delta x\right)\right] &= \sin\left[\left(2k_{\max} - 2k_{\max} + k\right) j\Delta x\right] \\
&= \sin\left[2\pi j - \left(2k_{\max} - k\right) j\Delta x\right] \\
&= \sin\left[-\left(2k_{\max} - k\right) j\Delta x\right] \\
&= \sin\left[k^*\left(j\Delta x\right)\right],
\end{aligned} \tag{20.5}$$

where

$$k^* \equiv -\left(2k_{\max} - k\right) \text{ for } |k| > k_{\max}. \tag{20.6}$$

Note that $0 < |k^*| < k_{\max}$ because, by assumption, $2k_{\max} > k > k_{\max}$. Similarly,

$$\cos\left[k\left(j\Delta x\right)\right] = \cos\left[k^*\left(j\Delta x\right)\right]. \tag{20.7}$$

Eqs. (20.5) and (20.7) 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 $\Delta x$ is reversed, or "backwards."

Fig. 20.3 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.

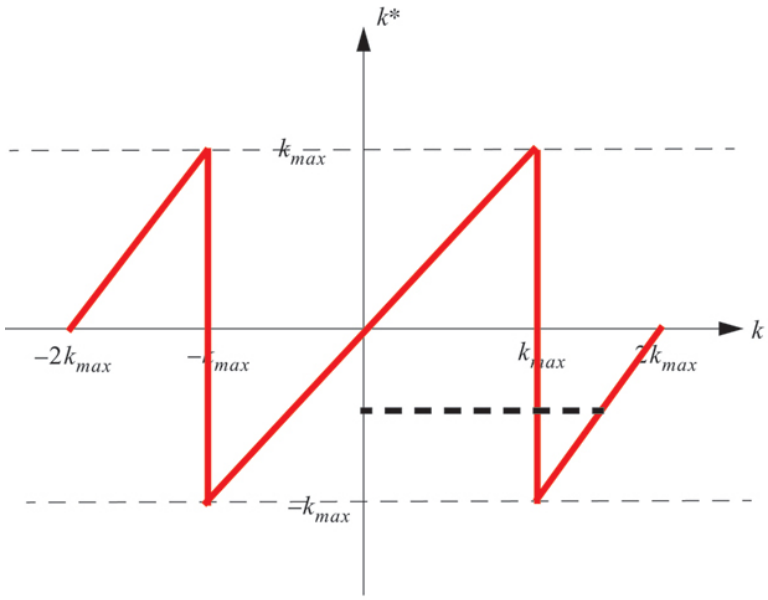



Figure 20.3: The red line is a plot of $k^*$ on the vertical axis, versus $k$ on the horizontal axis. The dashed black line connects $k = \frac{3\pi}{2\Delta x}$ with $k^* = -\frac{\pi}{2\Delta x}$, corresponding to the example of Fig. 20.1.

In the example shown in Fig. 20.1, $L = (4/3)\Delta x$ so $k = \frac{2\pi}{L} = \frac{3\pi}{2\Delta x}$. Therefore $k^* \equiv -(2k_{\max} - k) = \frac{2\pi}{\Delta x} - \frac{3\pi}{2\Delta x} = \frac{\pi}{2\Delta x}$, 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 $\pi$. This is shown in Fig. 20.4 a. For $k > k_{\max}$, the phase change, as $j$ increases by one, is greater than $\pi$. This is shown in Fig. 20.4 b. For $k > k_{\max}$, the dot in the figure appears to move clockwise, i.e., "backwards." This is a manifestation of aliasing that is familiar from the movies, in which wheels appear to spin backwards when the frame rate is too slow to capture the true motion. It also helps in understanding the minus sign that appears in Eq. (20.6).

## 20.3 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

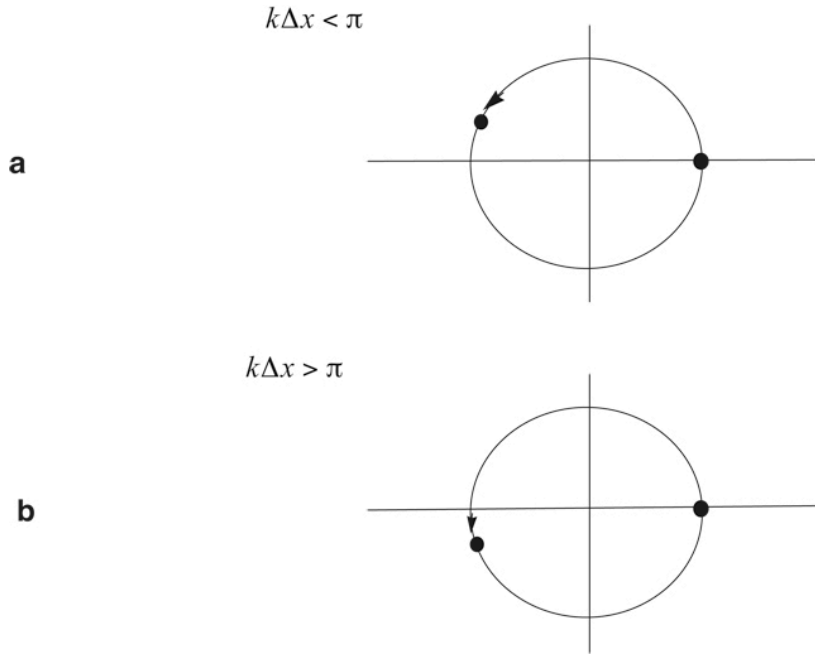



Figure 20.4: The phase change per grid point for: a) $k\Delta x < \pi$, and b) $k\Delta x > \pi$.

$$\frac{\partial q}{\partial t} + \mathbf{V} \cdot \nabla q = 0, \tag{20.8}$$

where the flow is assumed to be non-divergent, i.e.,

$$\nabla \cdot \mathbf{V} = 0. \tag{20.9}$$

Two-dimensional non-divergent flow is a not-too-drastic idealization of the large-scale circulation of the atmosphere. In view of (20.9), we can describe $\mathbf{V}$ in terms of a stream function $\psi$, such that

$$\mathbf{V} = \mathbf{k} \times \nabla \psi. \tag{20.10}$$

In Cartesian coordinates, $u = -\frac{\partial \psi}{\partial y}$, and $v = \frac{\partial \psi}{\partial x}$. (These sign conventions are arbitrary, but is essential that one of the derivatives has a plus sign and the other a minus sign.) Substituting (20.10) into (20.8), we get

$$\frac{\partial q}{\partial t} + (\mathbf{k} \times \nabla \psi) \cdot \nabla q = 0. \tag{20.11}$$

Using the identity

$$(\mathbf{V}_1 \times \mathbf{V}_2) \cdot \mathbf{V}_3 = \mathbf{V}_2 \cdot (\mathbf{V}_3 \times \mathbf{V}_1), \tag{20.12}$$

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

$$(\mathbf{k} \times \nabla \psi) \cdot \nabla q = \mathbf{k} \cdot (\nabla \psi \times \nabla q). \tag{20.13}$$

With the use of (20.13), we can re-write (20.11) as

$$\frac{\partial q}{\partial t} + J(\psi, q) = 0, \tag{20.14}$$

or alternatively as

$$\frac{\partial q}{\partial t} = J(q, \psi). \tag{20.15}$$

Here $J$ is the *Jacobian* operator, which is defined by

$$J(A,B) \equiv \mathbf{k} \cdot (\nabla A \times \nabla B) \tag{20.16}$$

$$= -\mathbf{k} \cdot \nabla \times (A \nabla B) \tag{20.17}$$

$$= \mathbf{k} \cdot \nabla \times (B \nabla A), \tag{20.18}$$

for arbitrary $A$ and $B$. Note that

$$J(A,B) = -J(B,A), \tag{20.19}$$

which can be deduced from (20.16). Eq. (20.19) has been used to go from (20.14) to (20.15). From the definition of the Jacobian, it follows that $J(p,q) = 0$ if either $A$ or $B$ is constant.

In Cartesian coordinates, we can write $J(A,B)$, in the following three alternative forms, which are suggested by the forms of (20.16)-(20.18):

$$J(A,B) = \frac{\partial A}{\partial x}\frac{\partial B}{\partial y} - \frac{\partial A}{\partial y}\frac{\partial B}{\partial x} \tag{20.20}$$

$$= \frac{\partial}{\partial y}\left(B\frac{\partial A}{\partial x}\right) - \frac{\partial}{\partial x}\left(B\frac{\partial A}{\partial y}\right) \tag{20.21}$$

$$= \frac{\partial}{\partial x}\left(A\frac{\partial B}{\partial y}\right) - \frac{\partial}{\partial y}\left(A\frac{\partial B}{\partial x}\right). \tag{20.22}$$

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:

$$\overline{J(A,B)} = 0, \tag{20.23}$$

$$\overline{AJ(A,B)} = 0, \tag{20.24}$$

$$\overline{BJ(A,B)} = 0. \tag{20.25}$$

Multiplying both sides of the advection equation (20.15) by $q$, we obtain

$$\begin{aligned} \frac{1}{2}\frac{\partial q^2}{\partial t} &= qJ(q,\psi) \\ &= J\left(\frac{1}{2}q^2,\psi\right). \end{aligned} \tag{20.26}$$

Integrating over the entire area, we find that

$$\begin{aligned} \int \frac{1}{2}\frac{\partial q^2}{\partial t}\,ds &= \int J\left(\frac{1}{2}q^2,\psi\right) ds \\ &= -\int \mathbf{V}\cdot\nabla\frac{1}{2}q^2 ds \\ &= -\int \nabla\cdot\left(\mathbf{V}\frac{1}{2}q^2\right)\,ds = 0. \end{aligned} \tag{20.27}$$

if the domain is surrounded by a rigid boundary where the normal component of $\mathbf{V}$ is zero, or if the domain is periodic.

## 20.4 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 can be called "aliasing instability." It is more often called "non-linear" instability, but this is somewhat misleading because the instability can also occur in the numerical integration of a linear equation with spatially variable coefficients. The instability is caused by a spurious growth of small-scale features of the flow, due in part to the aliasing error arising from the multiplication of the finite-difference analogs of *any* two spatially varying quantities.

### 20.4.1 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) (also see Lilly (1965)). We keep the

time derivatives continuous. This is not done merely for simplicity. The fact that the instability can be seen with continuous time derivatives makes the point that the mechanism of the instability comes purely from space differencing and has nothing to do with time differencing. It is quite different from the linear computational instability discussed in earlier chapters.

We begin by writing down a differential-difference version of (20.15), on a plane, using a simple finite-difference approximation for the Jacobian. For simplicity, we take $\Delta x = \Delta y = d$. We investigate the particular choice

$$\frac{dq_{i,j}}{dt} = \left[J_1\left(q,\psi\right)\right]_{i,j} \tag{20.28}$$

where we define $J_1$ as a particular finite-difference form of the Jacobian:

$$\left[J_1\left(q,\psi\right)\right]_{i\cdot j} \equiv \frac{1}{4d^2}\left[\left(q_{i+1,j}-q_{i-1,j}\right)\left(\psi_{i,j+1}-\psi_{i,j-1}\right)-\left(q_{i,j+1}-q_{i,j-1}\right)\left(\psi_{i+1,j}-\psi_{i-1,j}\right)\right]. \tag{20.29}$$

The finite-difference Jacobian given in (20.29) is based on (20.20). Later we are going to discuss several other finite-difference approximations to the Jacobian. Combining (20.28) and (20.29), we obtain

$$\frac{dq_{i,j}}{dt} \equiv \frac{1}{4d^2}\left[\left(q_{i+1,j}-q_{i-1,j}\right)\left(\psi_{i,j+1}-\psi_{i,j-1}\right)-\left(q_{i,j+1}-q_{i,j-1}\right)\left(\psi_{i+1,j}-\psi_{i-1,j}\right)\right]. \tag{20.30}$$

We *assume* that the solution, $q_{i,j}\left(t\right)$, has the form

$$q_{i,j}\left(t\right) = \left[C\left(t\right)\cos\left(\frac{\pi i}{2}\right)+S\left(t\right)\sin\left(\frac{\pi i}{2}\right)\right]\sin\left(\frac{2\pi j}{3}\right). \tag{20.31}$$

This strange-looking assumption will be justified later. For all $t$, we *prescribe* the stream function $\psi_{i,j}$ as

$$\psi_{i,j} = \Psi\cos\left(\pi i\right)\sin\left(\frac{2\pi j}{3}\right). \tag{20.32}$$

The wavelength in the $x$-direction is $2d$. By means of (20.32), we are prescribing a time-independent but *spatially variable* advecting current. The advecting current has almost always been 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 (20.31) and (20.32) are plotted in Fig. 20.5. They are nasty functions.

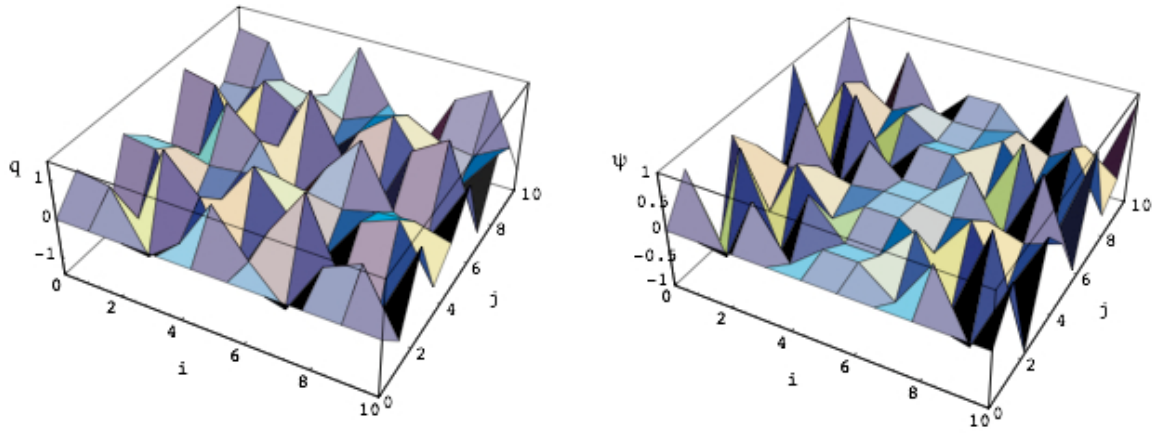



Figure 20.5: Plots of the functions $q_{i,j}(t=0)$ and $\psi_{i,j}$ given by (20.46) and (20.47), respectively. For plotting purposes, we have used $C = S = \Psi = 1$. The functions have been evaluated only for integer values of $i$ and $j$, which gives them a jagged appearance. Nevertheless it is fair 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 (20.32) says that $\psi_{i,j}$ has a wavelength of $2d$ in the $x$-direction, we can simplify (20.30) to

$$\frac{\partial q_{i,j}}{\partial t} = \frac{1}{4d^2}\left(q_{i+1,j} - q_{i-1,j}\right)\left(\psi_{i,j+1} - \psi_{i,j-1}\right). \tag{20.33}$$

From (20.31), we see that

$$\begin{aligned} q_{i+1,j} - q_{i-1,j} &= \left\{ C\left[\cos\frac{\pi(i+1)}{2} - \cos\frac{\pi(i-1)}{2}\right] + S\left[\sin\frac{\pi(i+1)}{2} - \sin\frac{\pi(i-1)}{2}\right]\right\}\sin\left(\frac{2\pi j}{3}\right) \\ &= 2\left(-C\sin\frac{\pi i}{2} + S\cos\frac{\pi i}{2}\right)\sin\left(\frac{2\pi j}{3}\right). \end{aligned} \tag{20.34}$$

Here we have used some trigonometric identities. Similarly, we can use some trig identities with (20.32) to show that

$$\begin{aligned}\psi_{i,j+1} - \psi_{i,j-1} &= \Psi \cos(\pi i)\, 2\cos\left(\frac{2\pi j}{3}\right)\sin\left(\frac{2\pi}{3}\right) \\ &= \sqrt{3}U\cos(\pi i)\cos\left(\frac{2\pi j}{3}\right).\end{aligned} \tag{20.35}$$

As already mentioned, (20.35) holds for all $t$. The product of (20.34) and (20.35) gives the right-hand side of (20.33), which can be written, again using trigonometric identities, as

$$\begin{aligned}\frac{dq_{i,j}}{dt} &= \frac{\sqrt{3}}{4d^2}\Psi\left\{-C\left[\sin\left(\frac{3\pi i}{2}\right) - \sin\left(\frac{\pi i}{2}\right)\right] + S\left[\cos\left(\frac{3\pi i}{2}\right) + \cos\left(\frac{\pi i}{2}\right)\right]\right\}\sin\left(\frac{4\pi j}{3}\right) \\ &= \frac{\sqrt{3}}{4d^2}\Psi\left[C\sin\left(\frac{\pi i}{2}\right) + S\cos\left(\frac{\pi i}{2}\right)\right]\sin\left(\frac{4\pi j}{3}\right).\end{aligned} \tag{20.36}$$

Now we make the important observation that the wave number in the $y$-direction, denoted by $l$, satisfies

$$ly = l(jd) = \frac{4\pi j}{3}. \tag{20.37}$$

By inspection of (20.37),

$$ld = \frac{4\pi}{3} > \pi. \tag{20.38}$$

*Eq.* (20.38) *shows that the product on the right-hand side of* (20.33) *has produced a wave number in the y-direction that is too short to be represented on the grid*. In other words, aliasing occurs. According to our earlier analysis, this wave will be interpreted by the grid as having the smaller wave number $l^* = -(2l_{\max} - l) = -\frac{2\pi}{3d}$. Therefore (20.36) can be re-written as

$$\frac{dq_{i,j}}{dt} = -\frac{\sqrt{3}\Psi}{4d^2}\left(C\sin\frac{\pi i}{2} + S\cos\frac{\pi i}{2}\right)\sin\frac{2\pi j}{3}. \tag{20.39}$$

Rewriting (20.36) as (20.39) is a key step in the analysis, because this is where aliasing enters. Because we are doing the problem algebraically, we have to put in the aliasing "by hand."

Next, we observe that the spatial form of $\frac{dq_{i,j}}{dt}$, as given by (20.39), agrees with the assumed form of $q_{i,j}$, given by (20.31). 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* (20.31)*, 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 (20.31) with respect to time, and substitute the result into the left-hand side of (20.39), we find that

$$\frac{dC}{dt}\cos\frac{\pi i}{2}\sin\frac{2\pi j}{3} + \frac{dS}{dt}\sin\frac{\pi i}{2}\sin\frac{2\pi j}{3} = -\frac{\sqrt{3}}{4d^2}\Psi\left(C\sin\frac{\pi i}{2} + S\cos\pi\frac{i}{2}\right)\sin\frac{2\pi j}{3}. \tag{20.40}$$

Note that time derivatives of $C(t)$ and $S(t)$ appear on the left-hand side of (20.40). Using the linear independence of the sine and cosine functions, we find by inspection of (20.40) that

$$\frac{dC}{dt} = -\frac{\sqrt{3}}{4d^2}\Psi S, \text{ and } \frac{dS}{dt} = -\frac{\sqrt{3}}{4d^2}\Psi C. \tag{20.41}$$

From (20.41), it follows that

$$\frac{d^2C}{dt^2} = \sigma^2 C \text{ and } \frac{d^2S}{dt^2} = \sigma^2 S, \tag{20.42}$$

where

$$\sigma \equiv \frac{\sqrt{3}\Psi}{4d^2}. \tag{20.43}$$

According to (20.42), $C$ and $S$ will grow exponentially, and the growth rate actually increases as the grid spacing becomes finer. This demonstrates that the finite-difference scheme is unstable. The unstable modes will have the ugly form given by (20.31) and plotted in Fig. 20.5.

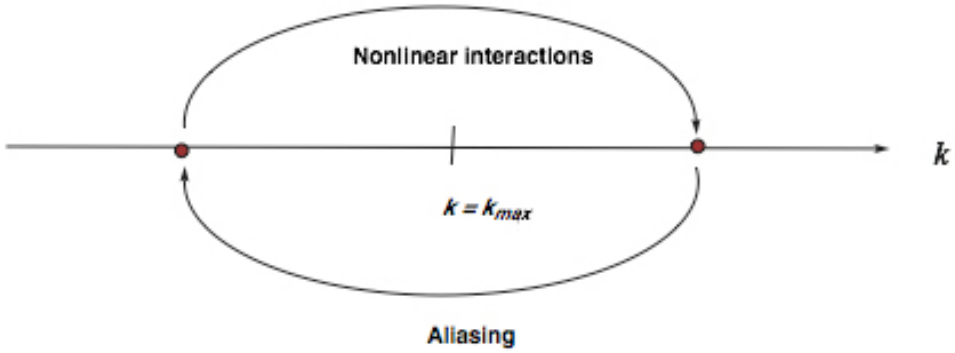



Figure 20.6: Sketch illustrating the mechanism of aliasing instability.

Fig. 20.6 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 would not occur with those Jacobians.

### 20.4.2 Analysis in terms of discretization error

Further general insight into this type of instability can be obtained by investigating the discretization error of (20.30). This can be expressed as

$$\begin{aligned}\left(\frac{dq}{dt}\right)_{i,j} &= \left[J_1(q,\psi)\right]_{i,j} \\ &= \left[J(q,\psi)\right]_{i,j} + \frac{d^2}{6}\left[\frac{\partial q}{\partial x}\frac{\partial^3\psi}{\partial y^3} - \frac{\partial q}{\partial y}\frac{\partial^3\psi}{\partial x^3} + \frac{\partial^3 q}{\partial x^3}\frac{\partial\psi}{\partial y} - \frac{\partial^3 q}{\partial y^3}\frac{\partial\psi}{\partial x}\right]_{i,j} + O\left(d^4\right).\end{aligned} \tag{20.44}$$

The second line is obtained by Taylor series expansion. Here $[J(q,\psi)]_{i,j}$ is the exact Jacobian at the point $(i,j)$. Because $[J_1(q,\psi)]_{i,j}$ has second-order accuracy, the leading term in the error is proportional to $d^2$. Multiplying (20.44) by $q$, integrating over the whole domain, and using (20.27), 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

$$\frac{1}{2}\frac{d}{dt}\int q^2 ds = \frac{d^2}{4}\int \frac{\partial^2\psi}{\partial x\partial y}\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right] ds + \int O\left(d^4\right) ds. \tag{20.45}$$

Note that $\frac{\partial^2\psi}{\partial x\partial y} = -\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$. Eq. (20.45) means that, for $\frac{\partial^2\psi}{\partial x\partial y} > 0$, $q^2$ will falsely grow with time if $\left(\frac{\partial q}{\partial x}\right)^2$ is bigger than $\left(\frac{\partial q}{\partial y}\right)^2$, 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 $\left[\left(\frac{\partial q}{\partial x}\right)^2 - \left(\frac{\partial q}{\partial y}\right)^2\right] < 0$.

Now look at Fig. 20.7. 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 (20.45), as illustrated schematically in Fig. 20.7. 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 (20.45), 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, *aliasing 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.

### 20.4.3 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

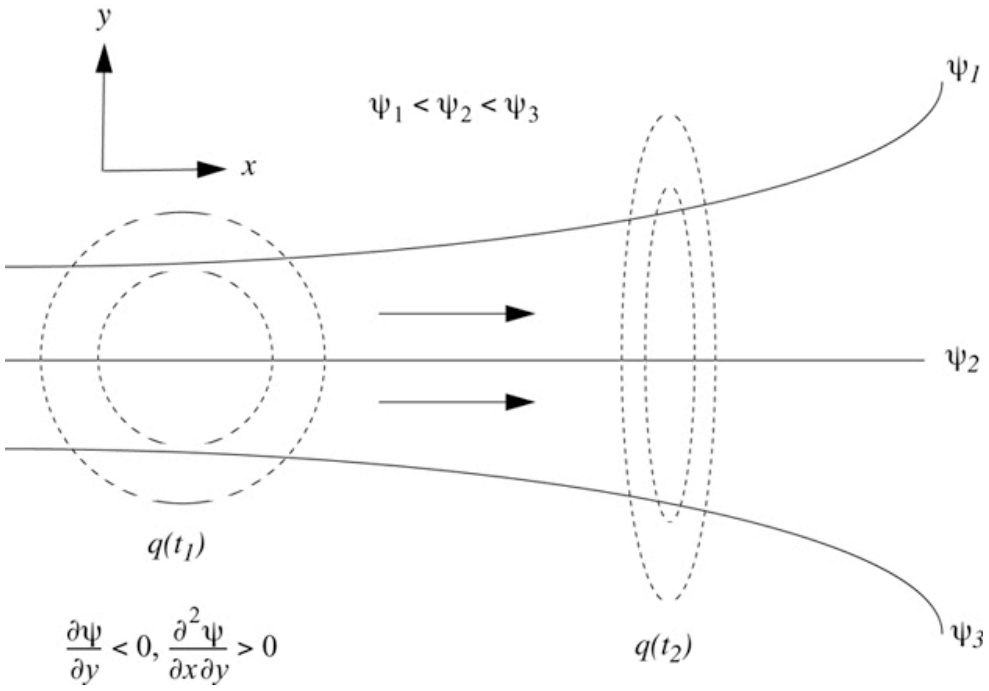



Figure 20.7: Schematic illustration of the mechanism of aliasing instability. The solid lines denote the stream function, which is independent of time. The implied mostly zonal flow is from left to right in the figure, and becomes weaker towards the right. The dashed lines show contours of the advected 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.

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, such as the Lax-Wendroff scheme, can overcome aliasing instability. Experience shows that this is not true. The damping of a dissipative scheme depends on the value of $\frac{c\Delta t}{\Delta x}$, but aliasing instability can occur even for $\frac{c\Delta t}{\Delta x} \to 0$.

A third approach is to use a sign-preserving advection scheme, as discussed in Chapter 4, and advocated by Smolarkiewicz (1991).

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 developed by Arakawa (1966). It will be explained below, after a digression in which we discuss a little more about the nature of two-dimensional non-divergent flows.

## 20.5 Fjortoft's Theorem

When the flow is non-divergent, so that (20.2) is satisfied, the vorticity equation, (20.3), reduces to

$$\frac{\partial}{\partial t}(\zeta + f) = -\mathbf{V} \cdot \nabla (\zeta + f). \tag{20.46}$$

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 $\zeta$, for simplicity. The vorticity and the stream function are related by

$$\zeta \equiv \mathbf{k} \cdot (\nabla \times \mathbf{V}) = \nabla^2 \psi. \tag{20.47}$$

(This relationship was used as an example of a boundary-value problem, back in the chapter on relaxation methods.) Eq. (20.46) can be rewritten as

$$\frac{\partial \zeta}{\partial t} = -\nabla \cdot (\mathbf{V}\zeta), \tag{20.48}$$

or, alternatively, as

$$\frac{\partial \zeta}{\partial t} = J(\zeta, \psi). \tag{20.49}$$

From (20.49) and (20.23) we see that the domain-averaged vorticity is conserved:

$$\frac{d\bar{\zeta}}{dt} = \overline{\frac{\partial \zeta}{\partial t}} = 0. \tag{20.50}$$

By combining (20.49) and (20.24), we can show that

$$\overline{\zeta \frac{\partial \zeta}{\partial t}} = 0, \tag{20.51}$$

from which it follows that the domain-average of the enstrophy is also conserved:

$$\frac{d}{dt}\left(\frac{1}{2}\overline{\zeta^2}\right) = 0. \tag{20.52}$$

Similarly, from (20.49) and (20.25) we find that

$$\overline{\psi \frac{\partial \zeta}{\partial t}} = 0. \tag{20.53}$$

To see what Eq. (20.53) implies, substitute (20.47) into (20.53), to obtain

$$\overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} = 0. \tag{20.54}$$

This is equivalent to

$$
\begin{aligned}
0 &= \overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} \\
&= \overline{\psi \frac{\partial}{\partial t} \left[ \nabla \cdot (\nabla \psi) \right]} \\
&= \overline{\psi \nabla \cdot \frac{\partial}{\partial t} \nabla \psi} \\
&= \overline{\nabla \cdot \left( \psi \frac{\partial}{\partial t} \nabla \psi \right)} - \overline{\nabla \psi \cdot \frac{\partial}{\partial t} \nabla \psi} \\
&= -\overline{\frac{\partial}{\partial t} \left( \frac{1}{2} |\nabla \psi|^2 \right)}.
\end{aligned}
\tag{20.55}
$$

Eq. (20.55) demonstrates that (20.49) implies kinetic energy conservation. It can also be shown that, for a purely rotational flow,

$$
\bar{K} = -\overline{\psi \zeta}. \tag{20.56}
$$

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:

$$
\frac{\text{energy}}{\text{enstrophy}} \sim \frac{L^2 t^{-2}}{t^{-2}} = L^2. \tag{20.57}
$$

This length can be interpreted as the scale of the most energetic vortices, and (20.57) 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.

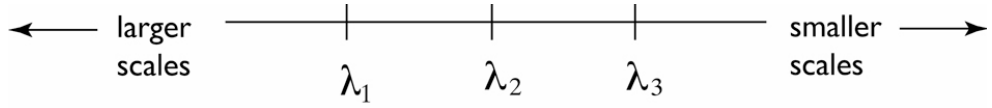



Figure 20.8: Diagram used in the explanation of Fjørtoft (1953) analysis of the exchanges of energy and enstrophy among differing scales in two-dimensional motion

The exchanges of energy and enstrophy among different scales in two-dimensional turbulence were studied by Ragnar Fjørtoft (1953), a Norwegian meteorologist who obtained some very fundamental and famous results, which can be summarized in a simplified way

as follows. Consider three equally spaced wave numbers, as shown in Fig. (20.8). By "equally spaced," we mean that

$$\lambda_2 - \lambda_1 = \lambda_3 - \lambda_2 = \Delta\lambda. \tag{20.58}$$

The enstrophy, $E$, is

$$E = E_1 + E_2 + E_3, \tag{20.59}$$

and the kinetic energy is

$$K = K_1 + K_2 + K_3. \tag{20.60}$$

It can be shown that

$$E_n = \lambda^2{}_n K_n, \tag{20.61}$$

where $\lambda_n$ is a wave number, and the subscript $n$ denotes a particular Fourier component.

Suppose that nonlinear processes redistribute kinetic energy among the three wave numbers, i.e.,

$$K_n \to K_n + \delta K_n, \tag{20.62}$$

while conserving both kinetic energy and enstrophy so that

$$\sum \delta K_n = 0, \tag{20.63}$$

and

$$\sum \delta E_n = 0. \tag{20.64}$$

From (20.61), we see that

$$\delta E_n = \lambda_n^2 \delta K_n. \tag{20.65}$$

It follows from (20.63) that

$$\delta K_1 + \delta K_3 = -\delta K_2, \tag{20.66}$$

and from (20.64) and (20.65) that

$$\begin{aligned} \lambda_1^2 \delta K_1 + \lambda_3^2 \delta K_3 &= -\lambda_2^2 \delta K_2 \\ &= {\lambda_2}^2 (\delta K_1 + \delta K_3). \end{aligned} \tag{20.67}$$

Collecting terms in (20.67), we find that

$$\frac{\delta K_3}{\delta K_1} = \frac{\lambda_2^2 - \lambda_1^2}{\lambda_3^2 - \lambda_2^2} > 0. \tag{20.68}$$

The fact that $\frac{\delta K_3}{\delta K_1}$ is positive means that either both $\delta K_3$ and $\delta K_1$ are positive, or both are negative. Using (20.58), Eq. (20.68) can be simplified to

$$0 < \frac{\delta K_3}{\delta K_1} = \frac{\lambda_2 + \lambda_1}{\lambda_3 + \lambda_2} < 1. \tag{20.69}$$

Eq. (20.69) shows that the energy transferred to or from higher wave numbers $(\delta K_3)$ is less than the energy transferred to or from lower wave numbers $(\delta K_1)$. This conclusion is based on both (20.63) and (20.64), 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 (20.65) and (20.69) to write

$$\begin{aligned}\frac{\delta E_3}{\delta E_1} &= \frac{\lambda_3^2}{\lambda_1^2}\left(\frac{\lambda_2+\lambda_1}{\lambda_3+\lambda_2}\right) \\ &= \frac{(\lambda_2+\Delta\lambda)^2}{(\lambda_2-\Delta\lambda)^2}\frac{\left(\lambda_2-\frac{1}{2}\Delta\lambda\right)}{\left(\lambda_2+\frac{1}{2}\Delta\lambda\right)} > 1.\end{aligned} \tag{20.70}$$

To show that this ratio is actually greater than one, as indicated above, we demonstrate that $\frac{\delta E_3}{\delta E_1} = a \cdot b \cdot c$, where $a$, $b$, and $c$ are each greater than one. We can choose:

$$a = \frac{\lambda_2+\Delta\lambda}{\lambda_2-\Delta\lambda} > 1,\ b = \frac{\lambda_2-\frac{1}{2}\Delta\lambda}{\lambda_2-\Delta\lambda} > 1,\ \text{and } c = \frac{\lambda_2+\Delta\lambda}{\lambda_2+\frac{1}{2}\Delta\lambda} > 1. \tag{20.71}$$

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 (20.57) 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 from kinetic to internal by viscosity. This is relevant to small-scale atmospheric circulations, such as boundary-layer eddies and cumulus 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.

## 20.6 Kinetic energy and enstrophy conservation in two-dimensional non-divergent flow

Lorenz (1960) suggested that energy-conserving finite-difference schemes would be advantageous in efforts to produce realistic numerical simulations of the general circulation of the atmosphere. 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.

We begin by writing down a spatially discrete version of (20.63), keeping the time derivative in continuous form:

$$\begin{aligned} \sigma_{\boldsymbol{i}} \frac{d\zeta_{\boldsymbol{i}}}{dt} &= \sigma_{\boldsymbol{i}} J_{\boldsymbol{i}}(\zeta, \psi) \\ &= \sum_{\boldsymbol{i}'} \sum_{\boldsymbol{i}''} c_{\boldsymbol{i},\boldsymbol{i}',\boldsymbol{i}''} \zeta_{\boldsymbol{i}+\boldsymbol{i}'} \psi_{\boldsymbol{i}+\boldsymbol{i}''}. \end{aligned} \tag{20.72}$$

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 2. 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 2. For later reference, with double subscripts, (20.72) would become

$$\sigma_{i,j} \frac{d\zeta_{i,j}}{dt} = \sum_{j'} \sum_{i'} \sum_{j''} \sum_{i''} \left( c_{i,j;i+i',j+j';i+i'',j+j''} \right) \zeta_{i+i',j+j'} \psi_{i+i'',j+j''}. \tag{20.73}$$

The second line of (20.72) looks a little mysterious and requires some explanation. As can be seen by inspection of (20.16), 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 (20.29). Such products appear in (20.72). The neighboring grid points can be specified in (20.72) 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 $c_{\boldsymbol{i},\boldsymbol{i}',\boldsymbol{i}''}$ can be called "interaction coefficients." Their

form will be chosen later. By making appropriate choices of the $c_{i,i',i''}$ we can construct an approximation to the Jacobian. The double sum in (20.72) essentially picks out the combinations of $\zeta$ and $\psi$, 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 2, but a bit more complicated.

Of course, there is nothing about the form of (20.72) that shows that it is actually a consistent finite-difference approximation to the Jacobian operator; all we can say at this point is that (20.72) 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 2.

The form of (20.72) 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 the next chapter).

As an example, consider the finite-difference Jacobian $J_1$, introduced in Eq. (20.29). Applying $J_1$ to vorticity advection on a square grid with grid spacing $d$, we can write, corresponding to (20.73),

$$\begin{aligned} d^2 \frac{d\zeta_{i,j}}{dt} &= d^2 \left\{ \frac{1}{4d^2} \left[ (\zeta_{i+1,j} - \zeta_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) - (\zeta_{i,j+1} - \zeta_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j}) \right] \right\} \\ &= \frac{1}{4} \left( \zeta_{i+1,j}\psi_{i,j+1} - \zeta_{i+1,j}\psi_{i,j-1} - \zeta_{i-1,j}\psi_{i,j+1} + \zeta_{i-1,j}\psi_{i,j-1} - \zeta_{i,j+1}\psi_{i+1,j} + \zeta_{i,j+1}\psi_{i-1,j} \right. \\ &\quad \left. + \zeta_{i,j-1}\psi_{i+1,j} - \zeta_{i,j-1}\psi_{i-1,j} \right). \end{aligned} \tag{20.74}$$

By inspection of the second line of (20.74), and comparing with (20.73), we see that for $J_1$ each value of $c_{i,j;i+i',j+j';i+i'',j+j''}$ is either $1/4$ or $-1/4$. The values are can be listed as follows:

$$
\begin{aligned}
c_{i,j;i+1,j;i,j+1} &= +\frac{1}{4}, \\
c_{i,j;i+1,j;i,j-1} &= -\frac{1}{4}, \\
c_{i,j;i-1,j;i,j+1} &= -\frac{1}{4}, \\
c_{i,j;i-1,j;i,j-1} &= +\frac{1}{4}, \\
c_{i,j;i,j+1;i+1,j} &= -\frac{1}{4}, \\
c_{i,j;i,j+1;i-1,j} &= +\frac{1}{4}, \\
c_{i,j;i,j-1;i+1,j} &= +\frac{1}{4}, \\
c_{i,j;i,j-1;i-1,j} &= -\frac{1}{4}.
\end{aligned}
\tag{20.75}
$$

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 $(i,j)$. 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 (20.75). The same methods can be applied to very different cases, however, such as nonuniform grids on a sphere.

Any finite-difference Jacobian should give zero if both of the input fields are spatially constant, so we require that

$$
\boxed{0 = \sum_{\boldsymbol{i}'} \sum_{\boldsymbol{i}''} c_{\boldsymbol{i},\boldsymbol{i}',\boldsymbol{i}''} \text{ for all } \boldsymbol{i}} \,, \tag{20.76}
$$

i.e., the sum of the coefficients is zero for all $\boldsymbol{i}$. The requirement (20.76) would emerge automatically if we enforced, for example, second-order accuracy of the Jacobian. You can easily confirm that $J_1$ satisfies (20.76).

Similarly, in case the vorticity is spatially constant, it should remain so for all time. From (20.72), this requirement takes the form

$$\boxed{0 = \sum_{i'} \sum_{i''} c_{i,i',i''} \psi_{i+i''} \text{ for all } i} \,. \tag{20.77}$$

Eq. (20.77) 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* (20.77) *must be true regardless of how the stream function varies in space*. This is only possible if each grid-point value of $\psi_{i+i''}$ appears more than once (in other words, at least twice) in the sum. Then we can arrange that the "total coefficient" multiplying $\psi_{i+i''}$, i.e., the sum of the two or more $c_{i,i',i''}$'s that multiply $\psi_{i+i''}$, is zero. In that case, the actual values of $\psi_{i+i''}$ have no effect on the sum in (20.77). You can confirm that $J_1$ satisfies (20.77).

In order to ensure conservation of the domain-averaged vorticity under advection, we must require that

$$\boxed{0 = \sum_{i} \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_{i+i''}} \,. \tag{20.78}$$

Here we have a triple sum, because we are taking a spatial average. In this respect, Eq. (20.78) is different in kind from (20.76) and (20.77). Eq. (20.78) has to be true *regardless of how the vorticity and stream function vary in space*. This is only possible if each product $\zeta_{i+i''} \psi_{i+i''}$ appears more than once in the sum, such that the sum of the two or more $c_{i,i',i''}$s that multiply each $\zeta_{i+i''} \psi_{i+i''}$, is zero. In that case, the actual values of $\zeta_{i+i''} \psi_{i+i''}$ have no effect on the sum in (20.78).

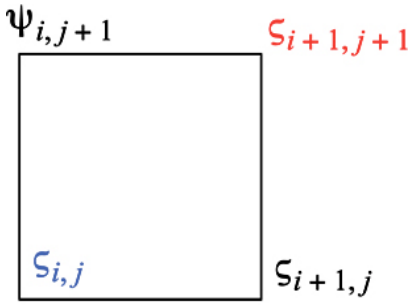



Figure 20.9: Stencil used in the discussion of vorticity conservation for $J_1$. See text for details.

With a little work, we can show that $J_1$ satisfies (20.78).

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. 11.9. 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 $i+1, j$ 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. Cancellation will occur if the coefficients sum to zero, or in other words if they satisfy

$$c_{i,j;i+1,j;i,j+1} = -c_{i+1,j+1;i+1,j;i,j+1}. \tag{20.79}$$

How can we use (20.79) to choose the forms of the coefficients? The key is to use the essential fact 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 (20.79), has to remain unchanged if we subtract one from each $i$ subscript and one from each $j$ subscript. In other words, it must be true that

$$c_{i+1,j+1;i+1,j;i,j+1} = c_{i,j;i,j-1;i-1,j}. \tag{20.80}$$

Eq. (20.80) allows us to rewrite (20.79) as

$$\boxed{c_{i,j;i+1,j;i,j+1} = -c_{i,j;i,j-1;i-1,j}}. \tag{20.81}$$

What has been accomplished here is that in (20.81), *both of the coefficients are associated with the time-rate of change of the vorticity at the point $i, j$*, and so, for $J_1$, both of them are explicitly listed in (20.75). The meaning of (20.81) is that the "right, up" coefficient is equal to minus the "down, left" coefficient. Inspection of (20.75) shows that (20.81) is indeed satisfied. Similar results apply for the remaining terms. In this way, we can satisfy ourselves that $J_1$ conserves vorticity.

Returning to the general problem, what we are going to do now is find a way to enforce finite-difference analogs of (20.51) and (20.53), namely:

$$\boxed{\begin{aligned} 0 &= \sum_i \sigma_i \zeta_i J_i(\zeta,\psi) \\ &= \sum_i \zeta_i \left( \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_{i+i''} \right) \\ &= \sum_i \left( \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_i \zeta_{i+i'} \psi_{i+i''} \right), \end{aligned}} \tag{20.82}$$

$$\boxed{\begin{aligned} 0 &= \sum_i \sigma_i \psi_i J_i(\zeta,\psi) \\ &= \sum_i \psi_i \left( \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_{i+i''} \right) \\ &= \sum_i \left( \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_i \psi_{i+i''} \right). \end{aligned}} \tag{20.83}$$

By enforcing these two requirements, we can ensure conservation of enstrophy and kinetic energy in the finite-difference model (although to ensure kinetic energy conservation we also need to attend to one additional requirement, discussed later). Eqs. (20.82) and (20.83) can be satisfied, as shown below, by suitable choices of the interaction coefficients. They look daunting, though, because they involve triple sums. How in the world are we ever going to figure this out?

Inspection of (20.82) 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 (20.72) and rewrite the scheme as

$$\begin{aligned} \sigma_i J_i(\zeta,\psi) &= \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_{i+i''} \\ &= \sum_{i'} a_{i,i+i'} \zeta_{i+i'}, \end{aligned} \tag{20.84}$$

where, by definition,

$$a_{i,i+i'} \equiv \sum_{i''} c_{i,i',i''} \psi_{i+i''}. \tag{20.85}$$

This definition simplifies things because, in going from the first line of (20.84) to the second line, we replace a double sum involving three subscripts to a single sum involving two subscripts. We can now write (20.84) times $\zeta_i$ as

$$\sigma_i \zeta_i J_i(\zeta, \psi) = \sum_{i'} a_{i,i+i'} \zeta_i \zeta_{i+i'}. \tag{20.86}$$

Here we have simply taken $\zeta_i$ 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 (20.82), which can be rewritten using our new notation as

$$0 = \sum_{i} \left( \sum_{i'} a_{i,i+i'} \zeta_i \zeta_{i+i'} \right). \tag{20.87}$$

The value of $a_{i,i+i'}$ (which involves a weighted sum of stream functions) measures the influence of $\zeta_{i+i'}$ on the enstrophy at the point $\boldsymbol{i}$. Similarly, the value of $a_{i+i',i}$ measures the influence of $\zeta_i$ on the enstrophy at the point $\boldsymbol{i} + \boldsymbol{i'}$. 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

$$a_{i,i+i'} = -a_{i+i',i} \text{ for all } \boldsymbol{i} \text{ and } \boldsymbol{i'}. \tag{20.88}$$

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. (20.88) implies that

$$a_{i,i} = 0 \text{ for all } \boldsymbol{i}. \tag{20.89}$$

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 (20.85), we can rewrite the non-divergence condition (20.77) as

$$0 = \sum_{i'} a_{i,i+i'} \text{ for all } \boldsymbol{i}. \tag{20.90}$$

Any scheme that satisfies (20.88) and (20.89) will also satisfy (20.90). 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 (20.72) as

$$\begin{aligned}\sigma_i J_i(\zeta,\psi) &= \sum_{i'}\sum_{i''} c_{i,i',i''}\zeta_{i+i'}\psi_{i+i''} \\ &= \sum_{i''} b_{i,i+i''}\psi_{i+i''},\end{aligned} \tag{20.91}$$

where

$$b_{i,i+i''} \equiv \sum_{i'} c_{i,i',i''}\zeta_{i+i'}. \tag{20.92}$$

The requirement for kinetic energy conservation, (20.83), can then be written as

$$0 = \sum_i \left( \sum_{i''} b_{i,i+i''}\psi_i\psi_{i+i''} \right), \tag{20.93}$$

which is analogous to (20.87). Kinetic energy conservation can be achieved by requiring that

$$b_{i,i+i''} = -b_{i+i'',i} \text{ all } \boldsymbol{i} \text{ and } \boldsymbol{i''}, \tag{20.94}$$

which is analogous to (20.88).

Actually, that's not quite true. As mentioned earlier, there is one more thing to check. In order to ensure kinetic energy conservation, we have to make sure that the finite-difference analog of (20.55) holds, i.e.,

$$\sum_i \left( \sigma_i \psi_i \frac{d\zeta_i}{dt} \right) = -\sum_i \left[ \sigma_i \frac{d}{dt} \left( \frac{1}{2} |\nabla\psi|^2{}_i \right) \right], \tag{20.95}$$

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. As an example, consider a square grid with grid spacing $d$:

$$\begin{aligned}\zeta_{i,j} &= \left(\nabla^2\psi\right)_{i,j} \\ &\equiv \frac{1}{d^2}\left(\psi_{i+1,j}+\psi_{i-1,j}+\psi_{i,j+1}+\psi_{i,j-1}-4\psi_{i,j}\right).\end{aligned} \tag{20.96}$$

Here we have reverted to a conventional double-subscripting scheme, for clarity. We also define a finite-difference kinetic energy by

$$\begin{aligned}K_{i,j} &\equiv \frac{1}{2}|\nabla\psi|^2_{i,j} \\ &\equiv \frac{1}{4d^2}\left[\left(\psi_{i+1,j}-\psi_{i,j}\right)^2+\left(\psi_{i,j+1}-\psi_{i,j}\right)^2+\left(\psi_{i,j}-\psi_{i-1,j}\right)^2+\left(\psi_{i,j}-\psi_{i,j-1}\right)^2\right].\end{aligned} \tag{20.97}$$

Because the right-hand side of (20.97) is a sum of squares, we are guaranteed that kinetic energy is non-negative. With the use of (20.96) and (20.97), it can be demonstrated, after a little algebra, that (20.95) is actually satisfied.

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.

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 (20.72) 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 (20.72). 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:

$$\left(J_1\right)_{i,j} = \frac{1}{4d^2}\left[\left(\zeta_{i+1,j}-\zeta_{i-1,j}\right)\left(\psi_{i,j+1}-\psi_{i,j-1}\right)-\left(\zeta_{i,j+1}-\zeta_{i,j-1}\right)\left(\psi_{i+1,j}-\psi_{i-1,j}\right)\right], \tag{20.98}$$

$$(J_2)_{i,j} = \frac{1}{4d^2}\left[-\left(\zeta_{i+1,j+1} - \zeta_{i+1,j-1}\right)\psi_{i+1,j} + \left(\zeta_{i-1,j+1} - \zeta_{i-1,j-1}\right)\psi_{i-1,j} + \left(\zeta_{i+1,j+1} - \zeta_{i-1,j+1}\right)\psi_{i,j+1} - \left(\zeta_{i+1,j-1} - \zeta_{i-1,j-1}\right)\psi_{i,j-1}\right], \tag{20.99}$$

$$(J_3)_{i,j} = \frac{1}{4d^2}\left[\zeta_{i+1,j}\left(\psi_{i+1,j+1} - \psi_{i+1,j-1}\right) - \zeta_{i-1,j}\left(\psi_{i-1,j+1} - \psi_{i-1,j-1}\right) - \zeta_{i,j+1}\left(\psi_{i+1,j+1} - \psi_{i-1,j+1}\right) + \zeta_{i,j-1}\left(\psi_{i+1,j-1} - \psi_{i-1,j-1}\right)\right]. \tag{20.100}$$

These can be interpreted as finite-difference analogs to the right-hand sides of (20.20) - (20.22), respectively. We can show that all three of these finite-difference Jacobians vanish if either of the input fields is spatially constant, and all three conserve vorticity, i.e., they all satisfy (20.78).

What we need to do next is identify ("by inspection") the coefficients $a$ and $b$ for each of (20.98) - (20.100), and then check each scheme to see whether the requirements (20.88) and (20.94) are satisfied. In order to understand more clearly what these requirements actually mean, look at Fig. 20.10. 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 $a_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{i}'}$ denotes $\zeta$-interactions of point $\boldsymbol{i}$ with point $\boldsymbol{i}+\boldsymbol{i}'$, while $a_{\boldsymbol{i}+\boldsymbol{i}',\boldsymbol{i}}$ denotes $\zeta$-interactions of point $\boldsymbol{i}+\boldsymbol{i}'$ with point $\boldsymbol{i}$. When we compare $a_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{i}'}$ with $a_{\boldsymbol{i}+\boldsymbol{i}',\boldsymbol{i}}$, it is like peering along one of the red (or purple) lines in Fig. 20.10, first outward from the point $(i, j)$, to one of the other points, and then back toward the point $(i, j)$. The condition (20.88) on the $a$s essentially means that all such interactions are "equal and opposite," thus allowing algebraic cancellations to occur when we sum over all points. The condition (20.94) on the $b$s has a similar interpretation.

To check whether $(J_1)_i$ conserves enstrophy and kinetic energy, we begin by rewriting (20.100) using the double-subscript notation, and equating it to $(J_1)_i$:

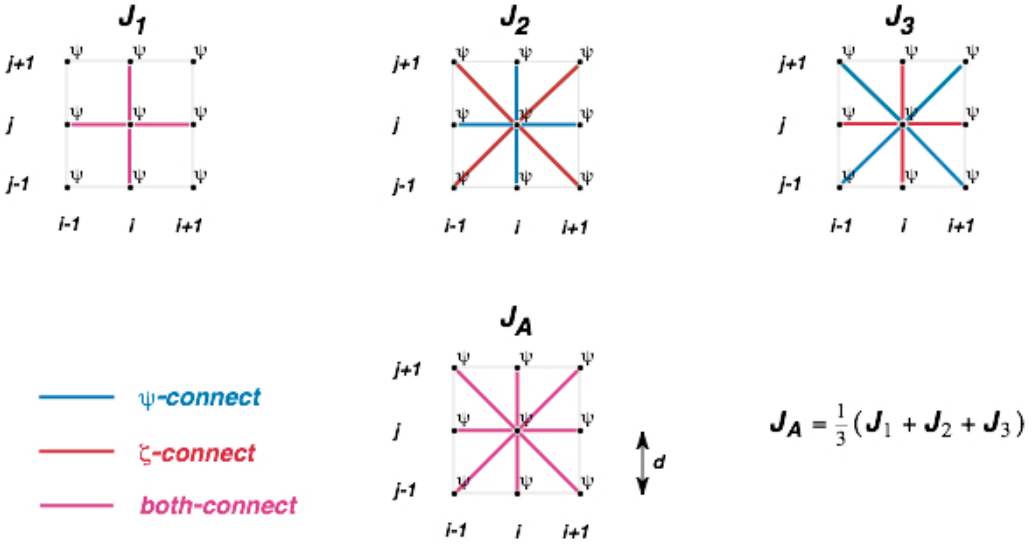



Figure 20.10: The central point in each figure is $(i, j)$. Stream function and vorticity are both defined at each of the mesh points indicated by the black dots. The colored lines represent contributions to $J_{i,j}$ from $\psi$, $\zeta$, or both, from the various neighboring points. For $J_1$ and $J_A$, the red and blue lines overlap, so you only see the pink lines.

$$\begin{aligned}
\sigma_{i,j}(J_1)_{i,j}(\zeta,\psi) &= \sum_{i'}\sum_{j'}\sum_{i''}\sum_{j''} c_{i',j';i'',j''}\,\zeta_{i,j;i+i',j+j'}\,\psi_{i+i'',j+j''} \\
&= \sum_{i'}\sum_{j'} a_{i,j;i+i',j+j'}\,\zeta_{i,j;i+i',j+j'} \\
&= \frac{1}{4}\left[\left(\zeta_{i+1,j}-\zeta_{i-1,j}\right)\left(\psi_{i,j+1}-\psi_{i,j-1}\right)-\left(\zeta_{i,j+1}-\zeta_{i,j-1}\right)\left(\psi_{i+1,j}-\psi_{i-1,j}\right)\right] \\
&= \frac{1}{4}\left[\zeta_{i+1,j}\left(\psi_{i,j+1}-\psi_{i,j-1}\right)-\zeta_{i-1,j}\left(\psi_{i,j+1}-\psi_{i,j-1}\right)\right. \\
&\quad \left.-\zeta_{i,j+1}\left(\psi_{i+1,j}-\psi_{i-1,j}\right)+\zeta_{i,j-1}\left(\psi_{i+1,j}-\psi_{i-1,j}\right)\right].
\end{aligned} \tag{20.101}$$

Here we have used

$$\sigma_{i,j} = d^2. \tag{20.102}$$

In the last line of (20.101), we have collected the coefficients of each distinct value of the vorticity. By inspection of (20.101) and comparison with (20.85), we can read off the

expressions for the $a$s of $J_1$:

$$a_{i,j;i+1,j} = \frac{1}{4}\left(\psi_{i,j+1} - \psi_{i,j-1}\right), \tag{20.103}$$

$$a_{i,j;i-1,j} = -\frac{1}{4}\left(\psi_{i,j+1} - \psi_{i,j-1}\right), \tag{20.104}$$

$$a_{i,j;i,j+1} = -\frac{1}{4}\left(\psi_{i+1,j} - \psi_{i-1,j}\right), \tag{20.105}$$

$$a_{i,j;i,j-1} = \frac{1}{4}\left(\psi_{i+1,j} - \psi_{i-1,j}\right). \tag{20.106}$$

Are these consistent with (20.88)? To find out, replace $i$ by $i+1$ in (20.104); this gives:

$$a_{i+1,j;i,j} = -\frac{1}{4}\left(\psi_{i+1,j+1} - \psi_{i+1,j-1}\right). \tag{20.107}$$

Now simply compare (20.107) with (20.103), to see that the requirement (20.88) is *not* satisfied by $J_1$. This shows 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:

$$J_A = \alpha J_1 + \beta J_2 + \gamma J_3, \tag{20.108}$$

where

$$\alpha + \beta + \gamma = 1. \tag{20.109}$$

With three unknown coefficients, and only one constraint, namely (20.109), we are free to satisfy two additional constraints; and we take these to be (20.88) and (20.94). In this way, we can show that $J_A$ will conserve both enstrophy and kinetic energy if we choose

$$\alpha = \beta = \gamma = 1/3. \tag{20.110}$$

The composite Jacobian, $J_A$, is often called the "Arakawa Jacobian." It is also called $J_7$.

Fig. 20.11 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 will be discussed later; 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 5, 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

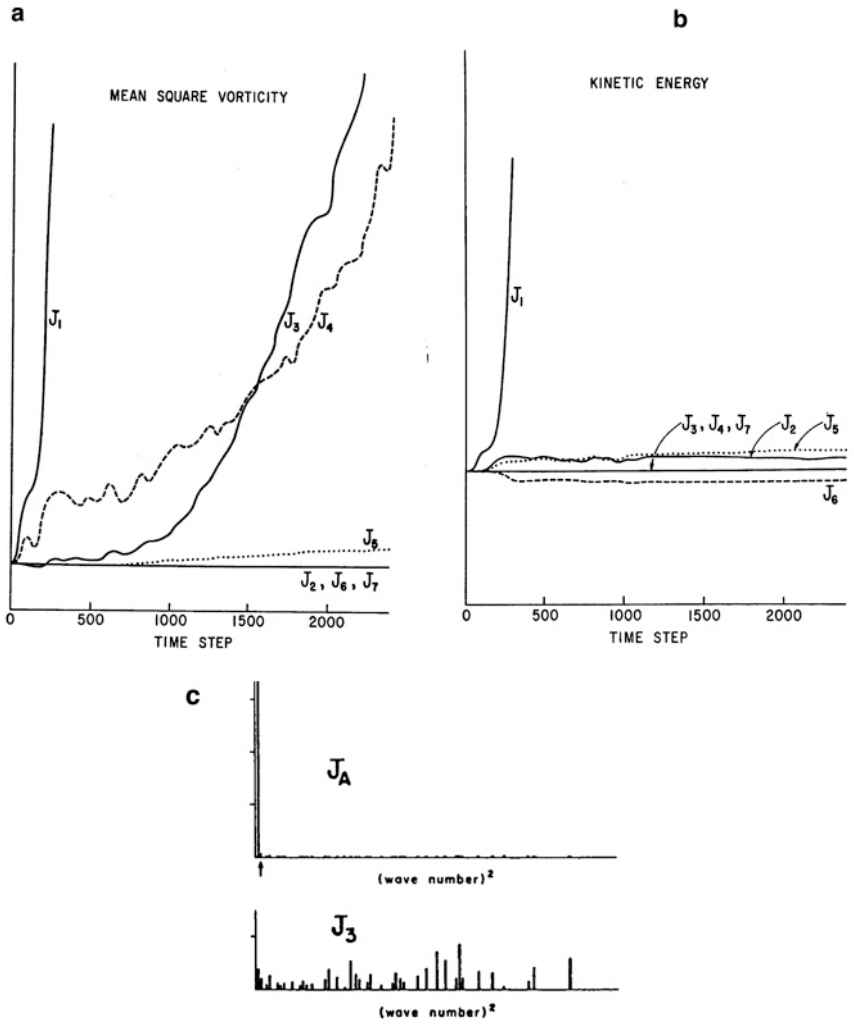



Figure 20.11: Results of tests with the various finite-difference Jacobians. Panel c shows that the initial kinetic energy is at a low wave number.

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.

## 20.7 The effects of time differencing on conservation of squares

When time differencing is included, a family of finite-difference schemes for (20.15) can be written in the generic form

$$\frac{q_{i,j}^{n+1} - q_{i,j}^{n}}{\Delta t} = J_{i,j}\left(q^*, \psi\right), \tag{20.111}$$

where $J_{i,j}$ is a finite difference analog to the Jacobian at the point $(i, j)$, and *different choices of $q^*$ give different time-differencing schemes*. Here $q$ is a generic scalar. Examples are given in Table 20.1.

Multiplying (20.111) by $q^*$, we get

$$q^*\left(q^{n+1} - q^n\right) = \Delta t q^* J\left(q^*, \psi\right). \tag{20.112}$$

Here we drop the subscripts for simplicity. Eq. (20.112) can be rearranged to

$$\left(q^{n+1}\right)^2 - \left(q^n\right)^2 = 2\left(\frac{q^{n+1} + q^n}{2} - q^*\right)\left(q^{n+1} - q^n\right) + 2\Delta t q^* J\left(q^*, \psi\right). \tag{20.113}$$

The left-hand side of (20.113) represents 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 (20.113), we find that

$$\overline{\left(q^{n+1}\right)^2} - \overline{\left(q^n\right)^2} = 2\overline{\left(\frac{q^{n+1} + q^n}{2} - q^*\right)\left(q^{n+1} - q^n\right)} + 2\Delta t\overline{q^* J\left(q^*, \psi\right)}. \tag{20.114}$$

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 (20.114) 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., $q^* = \frac{q^{n+1}+q^n}{2}$, 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 energy conservation.

Table 20.1: 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 $\Delta t / 2$ | $q^* = q^{n+\frac{1}{2}}$ |
| 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 | $q^* = q^n + \Delta t J\left(q^n, \psi\right)$ |

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.

## 20.8 Conservative schemes for the two-dimensional shallow water equations with rotation

We now present a generalization of the ideas discussed in Chapter 9. The two-dimensional shallow-water equations with rotation can be written as

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\mathbf{V}) = 0, \tag{20.115}$$

and

$$\frac{\partial \mathbf{V}}{\partial t} + \left(\frac{\zeta + f}{h}\right)\mathbf{k} \times (h\mathbf{V}) + \nabla\left[K + g\left(h + h_S\right)\right] = 0, \tag{20.116}$$

where $K \equiv \frac{1}{2}\mathbf{V} \cdot \mathbf{V}$ and

$$f \equiv 2\Omega \sin\varphi \tag{20.117}$$

is the Coriolis parameter. In (20.116), we have multiplied and divided the vorticity term by $h$, for reasons to be explained later. The corresponding equations for the zonal and meridional wind components are

$$\frac{\partial u}{\partial t} - \left(\frac{\zeta + f}{h}\right)(hv) + \frac{\partial}{\partial x}\left[K + g\left(h + h_S\right)\right] = 0, \tag{20.118}$$

and

$$\frac{\partial v}{\partial t} + \left(\frac{\zeta + f}{h}\right)(hu) + \frac{\partial}{\partial y}\left[K + g\left(h + h_S\right)\right] = 0, \tag{20.119}$$

respectively. Here

$$\mathbf{V} = u\mathbf{i} + v\mathbf{j}, \tag{20.120}$$

where **i** and **j** are the unit vectors in the zonal and meridional directions, respectively.

When we take the dot product of (20.116) with $h\mathbf{V}$, we obtain the advective form of the kinetic energy equation, i.e.,

$$h\frac{\partial K}{\partial t}+(h\mathbf{V})\cdot\nabla\left[K+g\left(h+h_S\right)\right]=0. \tag{20.121}$$

Here we have used

$$(h\mathbf{V})\cdot\left[\left(\frac{\zeta+f}{h}\right)\mathbf{k}\times(h\mathbf{V})\right]=0, \tag{20.122}$$

which is based on a *vector identity*, i.e., the cross product of two vectors is perpendicular to both. Combining (20.121) with the continuity equation (20.115), we can obtain the flux form of the kinetic energy equation:

$$\frac{\partial}{\partial t}\left(hK\right)+\nabla\cdot\left(h\mathbf{V}K\right)+\left(h\mathbf{V}\right)\cdot\nabla\left[g\left(h+h_S\right)\right]=0. \tag{20.123}$$

Similarly, the flux form of the potential energy equation is

$$\frac{\partial}{\partial t}\left[h\left(gh_S+\frac{1}{2}h\right)\right]+\nabla\cdot\left[\left(h\mathbf{V}\right)g\left(h+h_S\right)\right]-\left(h\mathbf{V}\right)\cdot\nabla\left[g\left(h+h_S\right)\right]=0. \tag{20.124}$$

By adding (20.122) and (20.124), we obtain conservation of total energy.

By taking the curl of (20.116) we can obtain the vorticity equation

$$\frac{\partial\zeta}{\partial t}+\mathbf{V}\cdot\nabla\left(\zeta+f\right)+\left(\zeta+f\right)\nabla\cdot\mathbf{V}=0. \tag{20.125}$$

To derive (20.125), we have used

$$\mathbf{k}\cdot\nabla\times\left\{\nabla\left[K+g\left(h+h_S\right)\right]\right\}=0, \tag{20.126}$$

which is based on another *vector identity*, i.e., the curl of any gradient is zero. Eq. (20.125) can be rearranged to

$$\frac{\partial(\zeta+f)}{\partial t}+\nabla\cdot[\mathbf{V}(\zeta+f)]=0. \tag{20.127}$$

The combination $\left(\frac{\zeta+f}{h}\right)$ is the potential vorticity for the shallow-water system. As you know, conservation of potential vorticity is key to the dynamics of balanced flows. The flux form of the potential vorticity equation for shallow water can be obtained simply by rewriting (20.127) as

$$\frac{\partial}{\partial t}\left[h\left(\frac{\zeta+f}{h}\right)\right]+\nabla\cdot\left[h\mathbf{V}\left(\frac{\zeta+f}{h}\right)\right]=0. \tag{20.128}$$

Earlier in this chapter, 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.

The approach outlined below follows Arakawa and Lamb (1981). We adopt the C-grid, as shown in Fig. 20.12. 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 $(i,j)$. Note that the vorticity is at the integer points, which is a departure from the system used earlier.

The finite-difference form of the continuity equation is

$$\frac{dh_{i+\frac{1}{2},j+\frac{1}{2}}}{dt}=\frac{(hu)_{i,j+\frac{1}{2}}-(hu)_{i+1,j+\frac{1}{2}}}{\Delta x}+\frac{(hv)_{i+\frac{1}{2},j}-(hv)_{i+\frac{1}{2},j+1}}{\Delta y}. \tag{20.129}$$

The various mass fluxes that appear in (20.129) 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

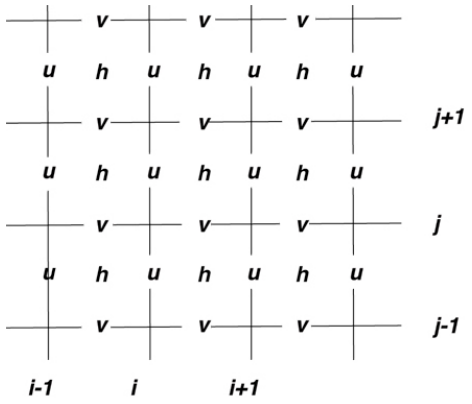



Figure 20.12: The arrangement of the mass, zonal wind, and meridional wind on the C grid. In this setup, the vorticity is at the integer points.

$$\frac{du_{i,j+\frac{1}{2}}}{dt} - \left[\left(\frac{\zeta+f}{h}\right)(hv)\right]_{i,j+\frac{1}{2}} + \left(\frac{K_{i+\frac{1}{2},j+\frac{1}{2}} - K_{i-\frac{1}{2},j+\frac{1}{2}}}{\Delta x}\right) + g\left[\frac{(h+h_S)_{i+\frac{1}{2},j+\frac{1}{2}} - (h+h_S)_{i-\frac{1}{2},j+\frac{1}{2}}}{\Delta x}\right] = 0, \tag{20.130}$$

and

$$\frac{dv_{i+\frac{1}{2},j}}{dt} + \left[\left(\frac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},j} + \left(\frac{K_{i+\frac{1}{2},j+\frac{1}{2}} - K_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y}\right) + g\left[\frac{(h+h_S)_{i+\frac{1}{2},j+\frac{1}{2}} - (h+h_S)_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y}\right] = 0, \tag{20.131}$$

respectively. As in the one-dimensional case discussed in Chapter 9, the kinetic energy per unit mass, $K_{i+\frac{1}{2},j+\frac{1}{2}}$, is undefined at this stage, but resides 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 (20.130) and (20.131) are in the "wrong" places; the 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 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 (20.130) and (20.131) 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 (20.130) and (20.131). Note, however, that it is actually *products* of mass fluxes and vorticities that are needed.

One obvious and important question is: *Is there a finite-difference scheme that allows us to "mimic" the vector identity* (20.122)*?* Since (20.122) is based on a purely mathematical identity, the input variables are irrelevant, and our goal is to mimic the identity itself. Arakawa and Lamb (1981) constructed the finite-difference vorticity terms in such a way that a finite-difference analog to (20.122) is satisfied, regardless of the specific forms of the mass fluxes and potential vorticities that are chosen. Their approach is to write:

$$\begin{aligned}\left[\left(\tfrac{\zeta+f}{h}\right)(hv)\right]_{i,j+\frac{1}{2}} = \quad & \alpha_{i,j+\frac{1}{2};i+\frac{1}{2},j+1}(hv)_{i+\frac{1}{2},j+1} + \beta_{i,j+\frac{1}{2};i-\frac{1}{2},j+1}(hv)_{i-\frac{1}{2},j+1} \\ & + \gamma_{i,j+\frac{1}{2};i-\frac{1}{2},j}(hv)_{i-\frac{1}{2},j} + \delta_{i,j+\frac{1}{2};i+\frac{1}{2},j}(hv)_{i+\frac{1}{2},j}\end{aligned} \tag{20.132}$$

and

$$\begin{aligned}\left[\left(\tfrac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},j} = \quad & \gamma_{i+\frac{1}{2},j;i+1,j+\frac{1}{2}}(hu)_{i+1,j+\frac{1}{2}} + \delta_{i+\frac{1}{2},j;i,j+\frac{1}{2}}(hu)_{i,j+\frac{1}{2}} \\ & + \alpha_{i+\frac{1}{2},j;i,j-\frac{1}{2}}(hu)_{i,j-\frac{1}{2}} + \beta_{i+\frac{1}{2},j;i+1,j-\frac{1}{2}}(hu)_{i+1,j-\frac{1}{2}}.\end{aligned} \tag{20.133}$$

In reality, the forms assumed by Arakawa and Lamb are slightly more general and slightly more complicated than these; we simplify here for ease of exposition. In (20.132) and (20.133), the $\alpha$'s, $\beta$'s, $\gamma$'s, and $\delta$'s obviously represent 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 $\alpha$'s, $\beta$'s, $\gamma$'s, and $\delta$'s are somewhat analogous to the $a$'s and $b$'s that were defined in the discussion of two-dimensional non-divergent flow, in that the $a$'s and $b$'s also linked pairs of points. In (20.132), the interpolated potential vorticities multiply the mass fluxes $hv$ at the four $v$-wind points surrounding the $u$-wind point $\left(i, j+\frac{1}{2}\right)$, and similarly in (20.133), the interpolated potential vorticities multiply the mass fluxes $hu$ at the four $u$-wind points surrounding the $v$-wind point $\left(i+\frac{1}{2}, j\right)$.

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 (20.133) by $(hu)_{i+\frac{1}{2},j}$ and (20.133) by $(hv)_{i,j+\frac{1}{2}}$, and add the results. With the forms given by (20.132) and (20.133), the vorticity terms will sum to

$$
\begin{aligned}
&-(hu)_{i,j+\frac{1}{2}}\left[\left(\tfrac{\zeta+f}{h}\right)(hv)\right]_{i,j+\frac{1}{2}}+(hv)_{i+\frac{1}{2},j}\left[\left(\tfrac{\zeta+f}{h}\right)(hu)\right]_{i+\frac{1}{2},j}\\
&=-(hu)_{i,j+\frac{1}{2}}\left[\alpha_{i,j+\frac{1}{2};i+\frac{1}{2},j+1}(hv)_{i+\frac{1}{2},j+1}+\beta_{i,j+\frac{1}{2};i-\frac{1}{2},j+1}(hv)_{i-\frac{1}{2},j+1}+\gamma_{i,j+\frac{1}{2};i-\frac{1}{2},j}(hv)_{i-\frac{1}{2},j}+\delta_{i,j+\frac{1}{2};i+\frac{1}{2},j}(hv)_{i+\frac{1}{2},j}\right]\\
&+(hv)_{i+\frac{1}{2}\cdot j}\left[\gamma_{i+\frac{1}{2},j;i+1,j+\frac{1}{2}}(hu)_{i+1,j+\frac{1}{2}}+\delta_{i+\frac{1}{2},j;i,j+\frac{1}{2}}(hu)_{i,j+\frac{1}{2}}+\alpha_{i+\frac{1}{2},j;i,j-\frac{1}{2}}(hu)_{i,j-\frac{1}{2}}+\beta_{i+\frac{1}{2},j;i+1,j-\frac{1}{2}}(hu)_{i+1,j-\frac{1}{2}}\right].
\end{aligned}
\tag{20.134}
$$

An analysis of (20.134) shows that cancellation will occur *when we sum over the whole grid*. This means that 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 will occur regardless of the expressions that we choose of the mass fluxes, and regardless of the expressions that we choose for the $\alpha$'s, $\beta$'s, $\gamma$'s, and $\delta$'s. The cancellation arises purely from the forms of (20.132) and (20.133), and is analogous to the cancellation that makes (20.126) work, i.e.,

$$
A\mathbf{V}\cdot(\mathbf{k}\times\mathbf{V})=A\left(u\mathbf{i}+v\mathbf{j}\right)\cdot\left(-v\mathbf{i}+u\mathbf{j}\right)=A\left(-uv+uv\right)=0, \tag{20.135}
$$

regardless of the input quantities $A$ and $\mathbf{V}$.

The above discussion shows that the finite-difference momentum equations represented by (20.130) and (20.131) with the use of (20.132) and (20.133), will guarantee kinetic energy conservation under advection, regardless of the forms chosen for the mass fluxes and the interpolated potential vorticities $\alpha$, $\beta$, $\gamma$, and $\delta$. From this point, the methods used in the discussion of the one-dimensional purely divergent flow will carry through essentially without change to give us conservation of mass, potential energy, and total energy. Arakawa and Lamb (1981) went much further, however, showing how the finite-difference momentum equations presented above (or, actually, slightly generalized versions of these equations) allow conservation of both potential vorticity and potential enstrophy. The details are rather complicated and will not be presented here.

## 20.9 Angular momentum conservation

Define the relative angular momentum per unit mass, $M$, by

$$
M_{rel}\equiv ua\cos\varphi. \tag{20.136}
$$

This is actually the component of the relative angular momentum vector in the direction of the axis of the Earth's rotation. Here we consider motion on the sphere, $a$ is the radius of the Earth, and $u$ is the zonal component of the wind. From the momentum equation, we can show that in the absence of pressure-gradient forces and friction,

$$\frac{\partial M}{\partial t} = -\left(\mathbf{V}\cdot\nabla\right)M, \tag{20.137}$$

where $\lambda$ is longitude, and

$$M \equiv M_{rel} + \Omega a^2 \cos\varphi \tag{20.138}$$

is the component of the absolute angular momentum vector in the direction of the axis of the Earth's rotation. From (20.137) it follows that the absolute angular momentum is conserved under advection.

Using integration by parts, it can be demonstrated that

$$\overline{M_{rel}} = a^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} \zeta \sin\varphi \cos\varphi d\lambda d\varphi. \tag{20.139}$$

This demonstrates that the globally averaged relative angular momentum is a constant times the global average of $\zeta \sin\varphi$. We can also prove that

$$\frac{d}{dt}\overline{M_{rel}} = a^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} \frac{\partial \zeta}{\partial t} \sin\varphi \cos\varphi d\lambda d\varphi = 0. \tag{20.140}$$

*** MORE TO BE ADDED HERE.

## 20.10 Summary

We began this chapter by discussing two-dimensional advection. 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, 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." 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.

## 20.11 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 $F > 0$.

2. Prove that $J_2$ conserves vorticity.

3. Prove that $J_3$ conserves kinetic energy.

4. Work out the *continuous* form of the Jacobian for the case of spherical coordinates (longitude, $\lambda$, and latitude, $\varphi$).

5. 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.

6. Using the form of the Laplacian for the hexagonal grid that you worked out earlier in the semester, show that

$$\sum_i \left( \sigma_i \psi_i \frac{d\zeta_i}{dt} \right) = -\sum_i \left[ \sigma_i \frac{d}{dt} \left( \frac{1}{2} |\nabla \psi|^2{}_i \right) \right] \tag{20.141}$$

   can be satisfied. Note that this condition only has to hold *for the sum over all grid points*, as shown.

7. For a hexagonal grid on a plane, show that a finite-difference Jacobian of the form

$$\frac{d\zeta_0}{dt} = \frac{1}{A} \sum_{i=1}^{6} \left( \frac{\psi_{i+1} - \psi_{i-1}}{\Delta s} \right) \left( \frac{\zeta_0 + \zeta_i}{2} \right) \Delta s \tag{20.142}$$

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 $\Delta s$ is the length of a side.

8. (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 2 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 21

# Finite Differences on the Sphere

## 21.1 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. Perhaps the most obvious possibility is to generate a grid using lines of constant latitude and longitude, i.e., a "spherical coordinate system." On such a grid, 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, which are structured grids derived from cartesian coordinates.

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 called "unstructured grids," although the term is not very descriptive, and sounds 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. For example, a model that uses an unstructured C-grid can predict the normal component of the wind on each cell wall.

## 21.2 Spherical coordinates

### 21.2.1 Vector calculus in spherical coordinates

In three-dimensional spherical coordinates $(\lambda, \varphi, r)$, i.e., longitude, latitude, and radius, the gradient, divergence, and curl operators take the following forms:

$$\nabla A = \left( \frac{1}{r\cos\varphi}\frac{\partial A}{\partial \lambda}, \frac{1}{r}\frac{\partial A}{\partial \varphi}, \frac{\partial A}{\partial r} \right), \tag{21.1}$$

$$\nabla \cdot \mathbf{V} = \frac{1}{r\cos\varphi}\frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{r\cos\varphi}\frac{\partial}{\partial \varphi}\left(V_\varphi \cos\varphi\right) + \frac{1}{r^2}\frac{\partial}{\partial r}\left(V_r r^2\right) \tag{21.2}$$

$$\nabla \times \mathbf{V} = \left\{ \frac{1}{r}\left[ \frac{\partial V_r}{\partial \varphi} - \frac{\partial}{\partial r}\left(rV_\varphi\right) \right], \frac{1}{r}\frac{\partial}{\partial r}\left(rV_\lambda\right) - \frac{1}{r\cos\varphi}\frac{\partial V_r}{\partial \lambda}, \frac{1}{r\cos\varphi}\left[ \frac{\partial V_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi}\left(V_\lambda \cos\varphi\right) \right] \right\}. \tag{21.3}$$

For use with the two-dimensional shallow-water equations, we can simplify these to

$$\nabla A = \left( \frac{1}{a\cos\varphi}\frac{\partial A}{\partial \lambda}, \frac{1}{a}\frac{\partial A}{\partial \varphi} \right), \tag{21.4}$$

$$\nabla \cdot \mathbf{V} = \frac{1}{a\cos\varphi}\frac{\partial V_\lambda}{\partial \lambda_\lambda} + \frac{1}{a\cos\varphi}\frac{\partial}{\partial \varphi}\left(V_\varphi \cos\varphi\right), \tag{21.5}$$

$$\mathbf{k} \cdot (\nabla \times \mathbf{V}) = \frac{1}{a\cos\varphi}\left[ \frac{\partial V_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi}\left(V_\lambda \cos\varphi\right) \right]. \tag{21.6}$$

Here $a$ is the radius of the spherical planet.

In a spherical coordinate system, the lines of constant longitude converge at the poles, so longitude is multivalued at the poles. The components of the wind vector (and all other vectors) are discontinuous at the poles, although the 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. 21.1. 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.

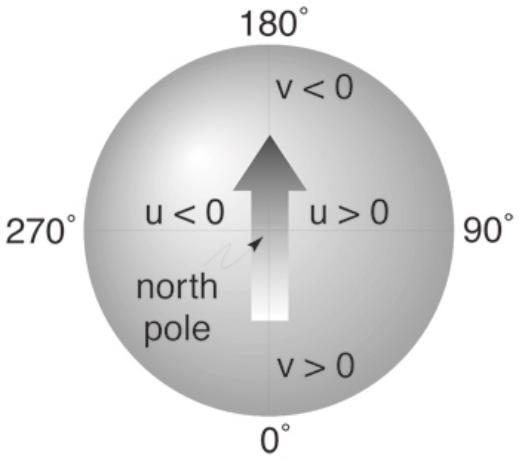



Figure 21.1: For the wind vector shown in the sketch, points along the prime meridian have a strong northward component. There is a discontinuity at the pole, and points along international date line have a strong southward component. Points near $90^\circ$ longitude have a strong positive zonal component, while points near $270^\circ$ longitude have a strong negative zonal component.

### 21.2.2 The shallow water equations in spherical coordinates

In spherical coordinates, the shallow water equations can be written as

$$\frac{\partial u}{\partial t}+\frac{u}{a\cos\varphi}\frac{\partial u}{\partial \lambda}+\frac{v}{a}\frac{\partial u}{\partial \varphi}-\left(f+\frac{u}{a}\tan\varphi\right)v+\frac{g}{a\cos\varphi}\frac{\partial}{\partial \lambda}\left(h+h_S\right)=0, \tag{21.7}$$

$$\frac{\partial v}{\partial t}+\frac{u}{a\cos\varphi}\frac{\partial v}{\partial \lambda}+\frac{v}{a}\frac{\partial v}{\partial \varphi}+\left(f+\frac{u}{a}\tan\varphi\right)u+\frac{g}{a}\frac{\partial}{\partial \varphi}\left(h+h_S\right)=0, \tag{21.8}$$

$$\frac{\partial h}{\partial t}+\frac{u}{a\cos\varphi}\frac{\partial h}{\partial \lambda}+\frac{v}{a}\frac{\partial h}{\partial \varphi}+\frac{h}{a\cos\varphi}\left[\frac{\partial u}{\partial \lambda}+\frac{\partial}{\partial \varphi}\left(v\cos\varphi\right)\right]=0. \tag{21.9}$$

Here $h$ is the depth of the fluid, and $h_S$ is the height of the bottom topography.

### 21.2.3 The "pole problem"

One eighth of a uniform latitude-longitude grid is shown in Fig. 21.2. 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 in shape. There are other ways to deal with the polar regions, e.g., by defining local Cartesian coordinates at the poles.

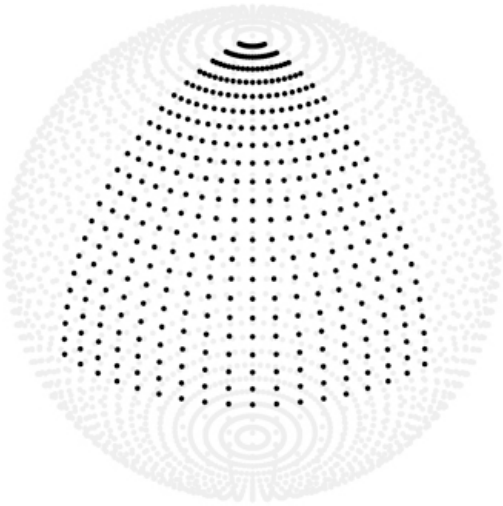


Figure 21.2: One octant of the latitude-longitude grid used by Arakawa and Lamb (1981). In the example shown, there are 72 grid points around a latitude circle and 44 latitude bands from pole to pole. The longitudinal grid spacing is globally uniform, and in this example is $5^\circ$. The latitudinal grid spacing is globally uniform except for "pizza slices" ringing each pole, which are 1.5 times as "tall" as the other grid cells. The reason for this is explained by Arakawa and Lamb (1981). In the example shown here, the latitudinal grid spacing is $4^\circ$ except that the pizza slices are $6^\circ$ tall.

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 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 5 (for advection) and 14 (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 can also avoid the pole problem for wave propagation, and in addition spectral models make it easy to implement semi-implicit time differencing for wave propagation. Further discussion is given in chapter 22.

To derive the stability criterion for wave propagation in the shallow water equations on the sphere, following Arakawa and Lamb (1977), we begin by linearizing (21.7), (21.8), and (21.9) about a state of rest, as follows:

$$\frac{\partial u}{\partial t} + \frac{g}{a\cos\varphi}\frac{\partial h}{\partial \lambda} = 0, \tag{21.10}$$

$$\frac{\partial v}{\partial t} + \frac{g}{a}\frac{\partial h}{\partial \varphi} = 0, \tag{21.11}$$

$$\frac{\partial h}{\partial t} + \frac{H}{a\cos\varphi}\left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \varphi}\left(v\cos\varphi\right)\right] = 0. \tag{21.12}$$

Here we have neglected rotation and bottom topography, for simplicity, and $H$ denotes the mean depth of the fluid. We spatially discretize (21.10) - (21.12) using the C-grid, as follows:

$$\frac{du_{j+\frac{1}{2},k}}{dt} + \frac{g\left(h_{j+1,k} - h_{j,k}\right)}{a\cos\varphi\Delta\lambda} = 0, \tag{21.13}$$

$$\frac{dv_{j,k+\frac{1}{2}}}{dt} + \frac{g\left(h_{j,k+1} - h_{j,k}\right)}{a\Delta\varphi} = 0, \tag{21.14}$$

$$\frac{dh_{j,k}}{dt} + H\left\{\left(\frac{u_{j+\frac{1}{2},k} - u_{j-\frac{1}{2},k}}{a\cos\varphi_j\Delta\lambda}\right) + \left[\frac{\left(v\cos\varphi\right)_{j,k+\frac{1}{2}} - \left(v\cos\varphi\right)_{j,k-\frac{1}{2}}}{a\cos\varphi_j\Delta\varphi}\right]\right\} = 0. \tag{21.15}$$

Here $j$ is the zonal index, and $k$ is the meridional index, and the time derivatives have been left in continuous form. We look for solutions of the form

$$u_{j+\frac{1}{2},k} = \mathrm{Re}\left\{\hat{u}_k \exp\left[im\left(j+\frac{1}{2}\right)\Delta\lambda + i\sigma t\right]\right\}, \tag{21.16}$$

$$v_{j,k+\frac{1}{2}} = \mathrm{Re}\left\{\hat{v}_{k+\frac{1}{2}} \exp\left[i\left(mj\Delta\lambda + \sigma t\right)\right]\right\}, \tag{21.17}$$

$$h_{j,k} = \mathrm{Re}\left\{\hat{h}_k \exp\left[i\left(mj\Delta\lambda + \sigma t\right)\right]\right\}. \tag{21.18}$$

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 (21.16) - (21.18) into (21.13) - (21.15), we obtain

$$\sigma\hat{u}_k + \frac{m}{a\cos\varphi_j}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2} g\left[S_k\left(m\right)\hat{h}_k\right] = 0, \tag{21.19}$$

$$i\sigma\hat{v}_{k+\frac{1}{2}} + g\left(\frac{\hat{h}_{k+1} - \hat{h}_k}{a\Delta\varphi}\right) = 0, \tag{21.20}$$

$$i\sigma\hat{h}_k + H\left\{\frac{im}{a\cos\varphi_k}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2} S_k\left(m\right)\hat{u}_k + \left[\frac{\left(\hat{v}\cos\varphi\right)_{k+\frac{1}{2}} - \left(\hat{v}\cos\varphi\right)_{k-\frac{1}{2}}}{a\cos\varphi_k\Delta\varphi}\right]\right\} = 0, \tag{21.21}$$

*where $S_k\left(m\right)$ is an artificially inserted "smoothing parameter" that depends on wave number and latitude*. The smoothing parameter has been inserted into the term of (21.19) corresponding to the *zonal* pressure gradient force, and also into the term of (21.21) corresponding to the *zonal* mass flux divergence. These are the key terms for zonally propagating gravity waves. Later in this discussion, $S_k\left(m\right)$ will be set to values less than one, in order to achieve computational stability with a "large" time step near the pole. For now, just consider it to be equal to one.

By eliminating $\hat{u}_k$ and $\hat{v}_{k+\frac{1}{2}}$ in (21.19) - (21.21), we can obtain the "meridional structure equation" for $\hat{h}_k$:

$$c^2\left[\frac{m}{a\cos\varphi_k}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2}S_k\left(m\right)\right]^2\hat{h}_k+\frac{c^2}{\left(a\Delta\varphi\right)^2}\left[\left(\hat{h}_k-\hat{h}_{k-1}\right)\frac{\cos\varphi_{k-\frac{1}{2}}}{\cos\varphi_k}-\left(\hat{h}_{k+1}-\hat{h}_k\right)\frac{\cos\varphi_{k+\frac{1}{2}}}{\cos\varphi_k}\right]=\sigma^2\hat{h}_k. \tag{21.22}$$

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 (21.22) dominates the second, and we obtain

$$\begin{aligned}\sigma&\cong\left|c\right|\left[\frac{m}{a\cos\varphi_k}\frac{\sin\left(m\Delta\lambda/2\right)}{m\Delta\lambda/2}S_k\left(m\right)\right]\\&=\frac{2\left|c\right|S_k\left(m\right)\sin\left(\frac{m\Delta\lambda}{2}\right)}{a\cos\varphi_k\Delta\lambda}.\end{aligned} \tag{21.23}$$

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

$$\sigma\Delta t<\varepsilon. \tag{21.24}$$

Using (21.23), this criterion can be written as

$$\frac{\left|c\right|\Delta t}{a\cos\varphi_k\Delta\lambda}\left[2S_k\left(m\right)\sin\left(\frac{m\Delta\lambda}{2}\right)\right]<\varepsilon, \tag{21.25}$$

where $\varepsilon$ is a constant of order one. In view of (21.23) and (21.24), the CFL criterion will place more stringent conditions on $\Delta t$ 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 $\sin\left(\frac{m\Delta\lambda}{2}\right)=1$, for which (21.25) reduces to

$$\frac{\left|c\right|\Delta t}{a\cos\varphi_k\Delta\lambda}2S_k\left(m\right)<\varepsilon. \tag{21.26}$$

This shows that, as expected, the time step required for stability depends on latitude. For the grid shown in Fig. 21.2, 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 (21.26), 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.

### 21.2.4 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 almost 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. 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 $S_k(m)$. We want to make the CFL criterion independent of latitude. Inspection of (21.23) shows that this can be accomplished by choosing the smoothing parameter $S_k(m)$ so that

$$\frac{S_k(m)\sin\left(\frac{m\Delta\lambda}{2}\right)}{a\cos\varphi_k\Delta\lambda} = \frac{1}{d^*}, \tag{21.27}$$

where $d^*$ is a suitably chosen length, comparable to the zonal grid spacing at the Equator. When $S_k(m)$ satisfies (21.27), the stability criterion (21.23) reduces to

$$\sigma = \frac{2|c|}{d^*}, \tag{21.28}$$

and the CFL condition reduces to

$$\frac{|c|\Delta t}{d^*} < \frac{\varepsilon}{2}, \tag{21.29}$$

so that *the time step required is independent of latitude*, as desired. If we choose

$$d^* \equiv a\Delta\varphi, \tag{21.30}$$

i.e., the distance between grid points in the meridional direction, then, referring back to (21.27), we see that $S_k(m)$ must satisfy

$$S_k(m) = \left(\frac{\Delta\lambda}{\Delta\varphi}\right)\frac{\cos\varphi_k}{\sin\left(\frac{m\Delta\lambda}{2}\right)}. \tag{21.31}$$

Of course, at low latitudes (21.31) can give values of $S_k(m)$ which are greater than one; these should be replaced by one, so that we actually use

$$\boxed{S_k(s) = \text{Min}\left\{\left(\frac{\Delta\lambda}{\Delta\varphi}\right)\frac{\cos\varphi_k}{\sin\left(\frac{m\Delta\lambda}{2}\right)},\ 1\right\}}. \tag{21.32}$$

A plot of (21.32) is given in Fig. 21.3, for the case of the shortest zonal mode. The plot shows that some smoothing is needed all the way down into the subtropics.

Polar filters are a partial solution to the pole problem for wave propagation. They work, but they are not ideal.

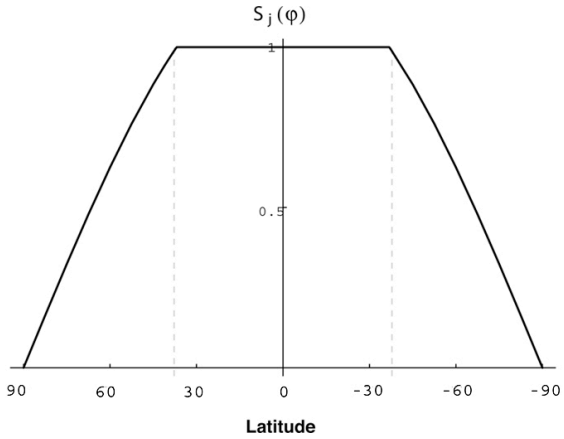



Figure 21.3: A plot of the smoothing parameter $S_k(m)$ as given by (21.32), for the "worst case" of the shortest zonal mode. The dashed vertical lines demarcate the belt of latitude centered on the Equator for which no smoothing is needed. In making the plot, it has been assumed that $\Delta\lambda = (5/4)\Delta\varphi$, which is true for the grid shown in Fig. 21.2.

## 21.3 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. 21.4; compare with Fig. 21.2. 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.

We can measure the homogeneity of the grid by examining the ratio of the zonal distance, $a\cos\varphi_j\Delta\lambda_j$, and the meridional distance $a\Delta\varphi$, for a grid point at latitude $\varphi_j$. Here, $\Delta\varphi \equiv \frac{\pi}{2}\frac{1}{N}$ and $\Delta\lambda_j \equiv \frac{1}{j-1}$. At $j = N+1$, the Equator, the ratio is one, and near the pole the ratio approaches $\pi/2 \cong 1.57$.

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. 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). The 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

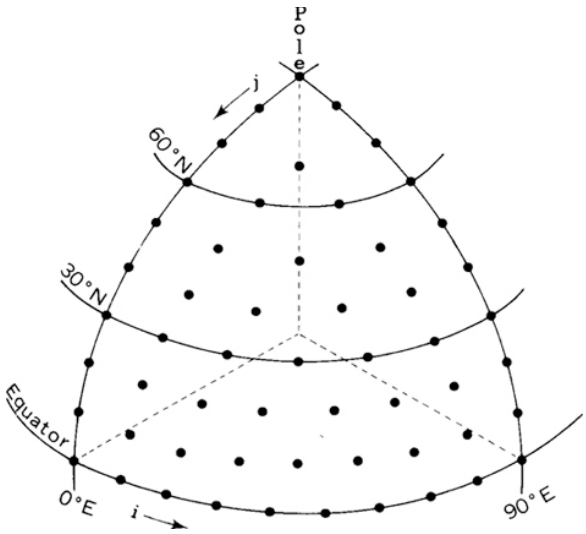



Figure 21.4: Kurihara grid on one octant of the sphere. Compare with Fig. 21.2.

wave should move from west to east, without distortion. In several of Kurihara's runs, however, the wave 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).

## 21.4 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 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 $\alpha_x$ and $\alpha_y$ so that the distance increment satisfies

$$dl^2 = \alpha_x{}^2 dx^2 + \alpha_y{}^2 dy^2. \tag{21.33}$$

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

$$U = \alpha_x \frac{dx}{dt}, \tag{21.34}$$

$$V = \alpha_y \frac{dy}{dt}. \tag{21.35}$$

Williamson (1969) gives the equations of motion for the general velocity components:

$$\frac{DU}{Dt} - \left[ f + \frac{1}{\alpha_x \alpha_y} \left( V \frac{\partial \alpha_y}{\partial x} - U \frac{\partial \alpha_x}{\partial y} \right) \right] V + \frac{g}{\alpha_x} \frac{\partial}{\partial x} (h + h_S) = 0, \tag{21.36}$$

$$\frac{DV}{Dt} + \left[ f + \frac{1}{\alpha_x \alpha_y} \left( V \frac{\partial \alpha_y}{\partial x} - U \frac{\partial \alpha_x}{\partial y} \right) \right] U + \frac{g}{\alpha_y} \frac{\partial}{\partial y} (h + h_S) = 0. \tag{21.37}$$

The Lagrangian time derivative is given by

$$\frac{D}{Dt}(\,) = \frac{\partial}{\partial t}(\,) + \frac{U}{\alpha_x} \frac{\partial}{\partial x}(\,) + \frac{V}{\alpha_y} \frac{\partial}{\partial y}(\,). \tag{21.38}$$

The continuity equation can be written as

$$\frac{Dh}{Dt} + \frac{h}{\alpha_x \alpha_y} \left[ \frac{\partial}{\partial x} (\alpha_y U) + \frac{\partial}{\partial y} (\alpha_x V) \right] = 0. \tag{21.39}$$

As an example, with spherical coordinates we have

$$x = \lambda \text{ and } y = \varphi, \tag{21.40}$$

and the corresponding metric coefficients are

$$\alpha_x = a\cos\varphi \text{ and } \alpha_y = a. \tag{21.41}$$

Then, by (21.34) and (21.35), we have

$$U = u \equiv a\cos\varphi\frac{D\lambda}{Dt} \text{ and } V = v \equiv a\frac{D\varphi}{Dt}. \tag{21.42}$$

Substituting (21.41) and (21.42) into (21.36), (21.37) and (21.39) gives (21.7), (21.8) and (21.9), 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. 21.5.

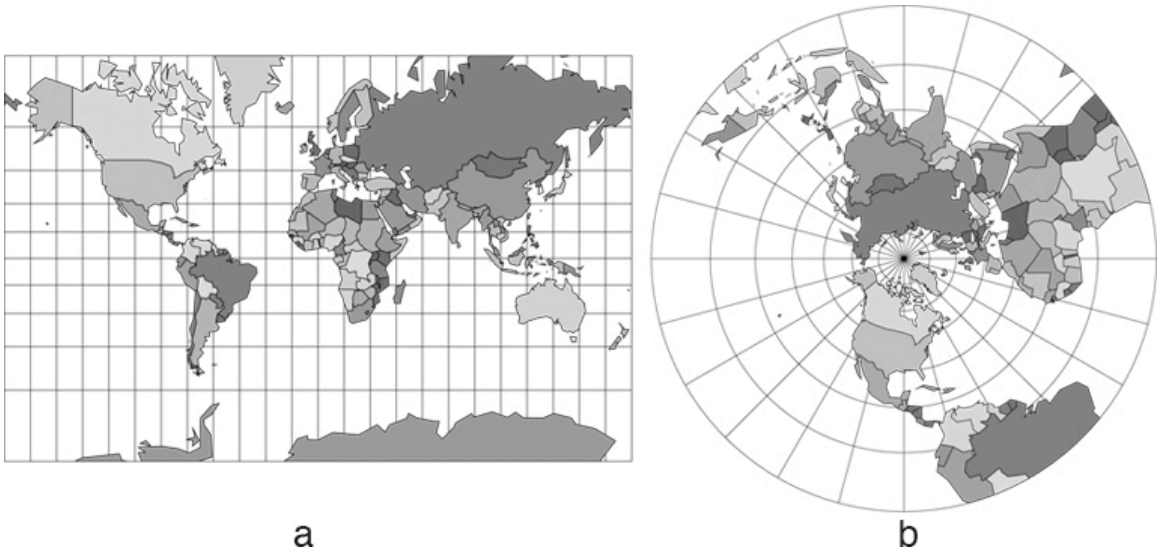



Figure 21.5: Map projections of the continents: a.) Mercator projection, in which the surface of the sphere is projected onto a cylinder that has its axis aligned with the poles, and the poles are stetched out into lines. b.) North polar stereographic projection, in which a hemisphere is projected onto a plane centered at the pole.

The polar stereographic projection is still used in a few models. It 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 $(\lambda,\varphi)$:

$$X = \frac{2a\cos\varphi\cos\lambda}{1+\sin\varphi}, \tag{21.43}$$

$$Y = \frac{2a\cos\varphi\sin\lambda}{1+\sin\varphi}. \tag{21.44}$$

Note that there is a problem at the South Pole, where the denominators of (21.43) and (21.44) go to zero. From (21.43) and (21.44) we find that

$$\begin{bmatrix} dX \\ dY \end{bmatrix} = \left(\frac{2a}{1+\sin\varphi}\right) \begin{bmatrix} -\cos\varphi\sin\lambda & -\cos\lambda \\ \cos\varphi\cos\lambda & -\sin\lambda \end{bmatrix} \begin{bmatrix} d\lambda \\ d\varphi \end{bmatrix}. \tag{21.45}$$

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 (21.33) gives

$$dl^2 = (a\cos\varphi)^2 d\lambda^2 + a^2 d\varphi^2. \tag{21.46}$$

Solving the linear system (21.45) for $d\varphi$, and $d\lambda$, and substituting the results into (21.46), we obtain

$$dl^2 = \left(\frac{1+\sin\varphi}{2}\right)^2 dX^2 + \left(\frac{1+\sin\varphi}{2}\right)^2 dY^2. \tag{21.47}$$

Comparing (21.47) with (21.33), we see that metric coefficients for the polar stereographic projection are given by

$$\alpha_x = \alpha_y = \frac{1+\sin\varphi}{2}. \tag{21.48}$$

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 (21.36), (21.37), and (21.39), we can write the shallow water equations in north polar stereographic coordinates:

$$\frac{dU}{dt} - \left(f + \frac{UY - VX}{2a^2}\right) V + gm(\varphi)\frac{\partial}{\partial X}(h + h_S) = 0, \tag{21.49}$$

$$\frac{dV}{dt} + \left(f + \frac{UY - VX}{2a^2}\right) U + gm(\varphi)\frac{\partial}{\partial Y}(h + h_S) = 0, \tag{21.50}$$

$$\frac{dh}{dt} + m^2(\varphi)\, h \left\{ \frac{\partial}{\partial X}\left[\frac{U}{m(\varphi)}\right] + \frac{\partial}{\partial Y}\left[\frac{V}{m(\varphi)}\right] \right\} = 0. \tag{21.51}$$

The total derivative is given by (21.38).

## 21.5 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.

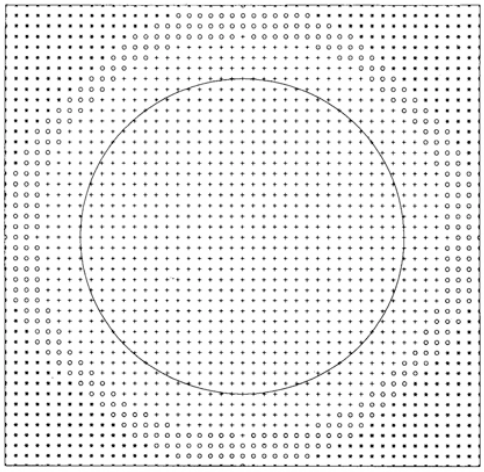


Figure 21.6: Composite grid method grid. Two such grids are used to cover the sphere. Points labeled with are the boundary conditions for the points labeled with +. Values at the points are obtained by interpolation from the other grid. The big circle is the image of the Equator. Points labeled * are not used.

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 (21.49), (21.50), and (21.51). 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. 21.6. 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. Browning's model is also easier to write computer code for because the equations are only expressed in the polar stereographic coordinate systems.

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.

## 21.6 Unstructured spherical grids

This is just for your amusement: One idea for constructing a mesh of grid points that homogeneously covers a sphere is to model 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 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 advan-

tageous 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 a great circle. The remaining electrons can be paired so that each has a mirror image in the opposite hemisphere. We can also fix an electron at each of the poles. Experience shows that unless the positions of some of the electrons are fixed, their positions will to wander indefinitely. Fig. 21.7 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.

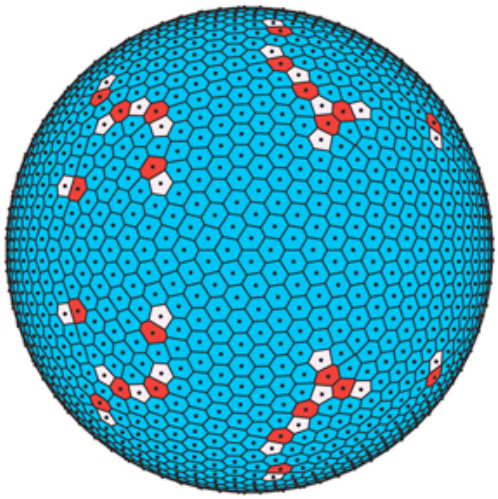


Figure 21.7: Wandering electron grid. White cells have five edges, blue cells have six edges, and red cells have seven edges.

Fig. (21.8) shows five alternative discretizations of the sphere. The left-most panel shows the latitude-longitude grid. The second and third panels show triangular and hexagonal-pentagonal grids, respectively, both generated by starting from the icosahedron.

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 last 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, something like the methods, discussed earlier, that can be used to patch together two polar-stereographic grids. Overlapping grids of this type are sometimes called "overset grids." There have been attempts to use grids based on octahedrons (e.g., McGregor (1996); Purser and Rančcić (1998)). A spiraling "Fibonacci grid" has also been suggested (Swinbank and James Purser, 2006).

Grids based on icosahedra offer an attractive framework for simulation of the global

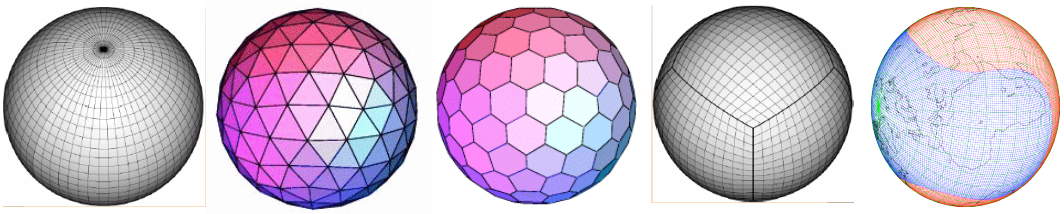



Figure 21.8: Various ways of discretizing the sphere. This figure was made by Bill Skamarock of NCAR.

circulation of the atmosphere. 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 introduced a new approach to more homogeneously discretize the sphere. They constructed grids using spherical triangles which are equilateral and 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).

The grids are constructed from an icosahedron (20 faces and 12 vertices), which is one of the five Platonic solids. 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. 21.9. 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.

Williamson (1968) chose the non-divergent shallow water equations to test the new grid. He solved the two-dimensional non-divergent vorticity equation

$$\frac{\partial \zeta}{\partial t} = J(\eta, \psi), \tag{21.52}$$

where $\zeta$ is relative vorticity, $\eta = \zeta + f$ is absolute vorticity and $\psi$ is the stream function, such that

$$\zeta = \nabla^2 \psi. \tag{21.53}$$

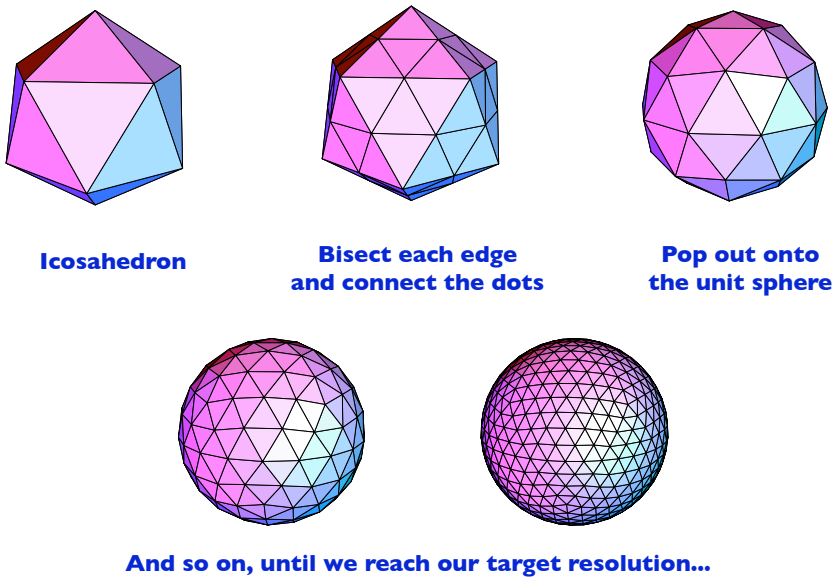



Figure 21.9: A spherical geodesic grid is generated recursively by starting from an icosahedron.

For arbitrary functions $\alpha$ and $\beta$, it follows from the form $J(\alpha,\beta) = \mathbf{k}\cdot\nabla\times(\alpha\nabla\beta)$ that the Jacobian satisfies

$$J(\alpha,\beta) = \lim_{A\to 0}\left\{\frac{1}{A}\oint_S \alpha\frac{\partial\beta}{\partial s}ds\right\}, \tag{21.54}$$

where $A$ is a small area, i.e., the area of a grid cell, and $m$ measures distance along the curve bounding $A$, i.e., the perimeter of the grid cell. Integrating (21.52) over the area $A$, and using (21.54), we get

$$\frac{d}{dt}\int_A \zeta dA = \oint_S \eta\frac{\partial\psi}{\partial s}ds. \tag{21.55}$$

This can be discretized with reference to Fig. 21.10. We approximate the line integral along the polygon defined by the path $P_1,P_2,\ldots\ P_5,P_1$. Let $\zeta_0$ be the relative vorticity defined at the point $P_0$, and let $\eta_i$ be the absolute vorticity defined at the point $P_i$. We can approximate (21.55) by

$$\frac{d\zeta_0}{dt} = \frac{1}{A}\sum_{i=1}^{\mathrm{K}} \left(\frac{\psi_{i+1} - \psi_{i-1}}{\Delta s}\right)\left(\frac{\eta_0 + \eta_i}{2}\right)\Delta s. \tag{21.56}$$

We must also discretize the Laplacian. Consider the smaller, inner polygon in Fig. 21.10. Its walls are formed from the perpendicular bisectors of the line segments $\overline{P_0P_i}$. We can use Gauss's Theorem to write

$$\int_a \zeta dA = -\oint_{s'} (\nabla\psi)\cdot\mathbf{n}ds', \tag{21.57}$$

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. (21.57) is approximated by

$$a\zeta_0 = \sum_{i=1}^{K} \frac{l_i}{|P_0P_i|}(\psi_i - \psi_0), \tag{21.58}$$

where $|P_0P_i|$ is the distance from $P_0$ to $P_i$, and $l_i$ is the length of wall $i$. Eq. (21.58) can be solved for $\psi_i$ by relaxation, using the methods discussed in Chapter 12.

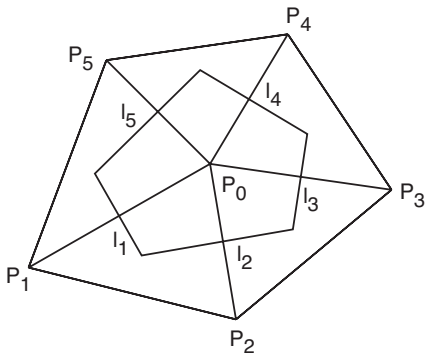



Figure 21.10: Configuration of grid triangles for the case of a pentagon.

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 14). 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.

## 21.7 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. The pole problem can be dealt with by using filters, but these approaches suffer from some problems of their own. Semi-implicit differencing could 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.

## 21.8 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 22

# Spectral Methods

## 22.1 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 actually 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.

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:

$$q(x,t) = \sum_{k=-\infty}^{\infty} \hat{q}_k(t)\, e^{ikx}. \tag{22.1}$$

The complex coefficients $\hat{q}_k(t)$ can be computed using

$$\hat{q}_k(t) = \frac{1}{L} \int_{x-L/2}^{x+L/2} q\left(x',t\right) e^{-ikx'} dx'. \tag{22.2}$$

Recall that the proof of (22.1) and (22.2) involves use of the orthogonality condition

$$\frac{1}{L}\int_{x-L/2}^{x+L/2} e^{-ikx'}e^{ilx'}dx' = \delta_{k,l}, \tag{22.3}$$

where

$$\delta_{k,l} \equiv \begin{cases} 1, \; k = l \\ 0, \; k \neq l \end{cases} \tag{22.4}$$

is the Kronecker delta. Eqs. (22.1) and (22.2) are a "transform pair." They can be used to go back and forth between physical space and wave-number space.

From (22.1), we see that the $x$-derivative of $q$ satisfies

$$\frac{\partial q}{\partial x}(x,t) = \sum_{k=-\infty}^{\infty} ik\hat{q}_k(t)\, e^{ikx}. \tag{22.5}$$

Inspection of (22.5) shows that $\frac{\partial q}{\partial x}$ does not receive a contribution from $\hat{q}_0$; the reason for this should be clear.

A spectral model uses equations similar to (22.1), (22.2), and (22.5), but with a finite set of wave numbers, and with $x$ defined on a finite mesh:

$$q\left(x_j,t\right) \cong \sum_{k=-n}^{n} \hat{q}_k(t)\, e^{ikx_j}, \tag{22.6}$$

$$\hat{q}_k(t) = \frac{1}{M}\sum_{j=1}^{M} q\left(x_j,t\right) e^{-ikx_j}, -n \leq k \leq n, \tag{22.7}$$

$$\frac{\partial q}{\partial x}\left(x_j,t\right) \cong \sum_{k=-n}^{n} ik\hat{q}_k(t)\, e^{ikx_j}. \tag{22.8}$$

The sums in (22.6) and (22.8) are *truncated*, in that they do not include wave numbers outside the range $\pm n$. The value of $n$ is chosen by the modeler. Note that we have used "approximately equal signs" in (22.6) and (22.8), but not in (22.7). The sum that appears in (22.7) is over a grid with $M$ points.

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 (22.6) is *exactly* reversible by the inverse transform (22.7), 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 (22.6) into (22.7) to obtain

$$\hat{q}_k(t) = \frac{1}{M}\sum_{j=1}^{M}\left\{\left[\sum_{l=-n}^{n}\hat{q}_l(t)\,e^{ilx_j}\right]e^{-ikx_j}\right\} \text{ for } -n \le k \le n. \tag{22.9}$$

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 (22.9) 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 (22.9); the range for wave number $k$ is understood because, as indicated, we wish to evaluate the left-hand side of (22.9) for $k$ in the range $-n$ to $n$. Because each term on the right-hand side of (22.9) 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 $\widehat{q}_k(t)$. In general, this is the equivalent of $4n+2$ real numbers, suggesting 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

$$\boxed{\hat{q}_{-k} = \hat{q}_k^*}\,, \tag{22.10}$$

where the $*$ 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 (22.10) is true, consider the *combined* contributions of the $+k$ and $-k$ coefficients to the sum in (22.6). Define $T_k\left(x_j,t\right)$ by

$$\begin{aligned} T_k\left(x_j,t\right) &\equiv \hat{q}_k\left(t\right)e^{ikx_j} + \hat{q}_{-k}\left(t\right)e^{-ikx_j} \\ &\equiv R_k e^{i\theta}e^{ikx_j} + R_{-k}e^{i\mu}e^{-ikx_j} \\ &= R_k e^{i\left(\theta+kx_j\right)} + R_{-k}e^{i\left(\mu-kx_j\right)}. \end{aligned} \tag{22.11}$$

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 (22.6) as

$$q\left(x_j,t\right) \cong \sum_{k=0}^{n} T_k\left(x_j,t\right). \tag{22.12}$$

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

$$R_k \sin\left(\theta + kx_j\right) + R_{-k}\sin\left(\mu - kx_j\right) = 0 \text{ for all } x_j. \tag{22.13}$$

The only way to satisfy (22.13) for all $x_j$ is to set,

$$R_k = R_{-k} \tag{22.14}$$

and

$$\theta + kx_j = -\left(\mu - kx_j\right) = -\mu + kx_j. \tag{22.15}$$

From (22.15), we see that $\theta = -\mu$. Eq. (22.10) follows from (22.14) and (22.15).

As meniontedabove, 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 (22.10) 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 $q\left(x,t\right)$ 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 $M = 2n+1$, the transform pair (22.6) - (22.7) 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.

Substituting (22.7) into (22.8) gives

$$\frac{\partial q}{\partial x}\left(x_j,t\right) \cong \sum_{k=-n}^{n} \left[ \frac{ik}{M} \sum_{l=1}^{M} q\left(x_l,t\right) e^{-kx_l} \right] e^{ikx_j}. \tag{22.16}$$

Reversing the order of summation leads to

$$\boxed{\frac{\partial q}{\partial x}\left(x_j,t\right) \cong \sum_{l=1}^{M} \alpha_j^l q\left(x_l,t\right)}, \tag{22.17}$$

where we define

$$\alpha_j^l \equiv \sum_{k=-n}^{n} \frac{ik}{M} e^{ik\left(x_j - x_l\right)}. \tag{22.18}$$

The point of this little exercise is that (22.17) can be interpreted as a finite-difference approximation. It is a member of the family of approximations discussed many times in this course, but it is 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.

XXX Add a discussion of Gibbs' phenomenon.

## 22.2 Solving linear equations with the spectral method

Now consider the one-dimensional advection equation with a constant current, $c$:

$$\frac{\partial q}{\partial t} = -c\frac{\partial q}{\partial x}. \tag{22.19}$$

Substituting (22.6) and (22.8) into (22.19) gives

$$\sum_{k=-n}^{n} \frac{d\hat{q}_k}{dt} e^{ikx} = -c \sum_{k=-n}^{n} ik\hat{q}_k e^{ikx}. \tag{22.20}$$

By linear independence, we obtain

$$\frac{d\hat{q}_k}{dt} = -ikc\hat{q}_k \text{ for } -n \leq k \leq n. \tag{22.21}$$

Note that $\frac{d\hat{q}_0}{dt}$ will be equal to zero; the interpretation of this should be clear. We can use (22.21) to predict $\hat{q}_k(t)$. When we need to know $q\left(x_j,t\right)$, we can get it from (22.6).

Compare (22.21) with

$$\frac{d\hat{q}_k}{dt} = -ikc\left[\frac{\sin(k\Delta x)}{k\Delta x}\right]\hat{q}_k, \tag{22.22}$$

which, as discussed in earlier chapters, is obtained by using centered second-order space differencing. The spectral method gives the *exact* advection 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 the 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

$$\nabla^2 q = f(x,y), \tag{22.23}$$

as a problem to determine $q$ for given $f(x,y)$. In one dimension, (22.23) becomes

$$\frac{d^2q}{dx^2} = f(x). \tag{22.24}$$

We assume periodic boundary conditions and compute the Fourier coefficients of both $q$ and $f$, using (22.7). Then (22.24) can be written as

$$\sum_{k=-n}^{n} \left(-k^2\right) \hat{q}_k e^{ikx} = \sum_{k=-n}^{n} \hat{f}_k e^{ikx}, \tag{22.25}$$

Invoking linear independence to equate coefficients of $e^{ikx}$, we find that

$$\hat{q}_k = \frac{-\hat{f}_k}{k^2} \text{ for } -n \leq k \leq n \text{ (unless } k = 0). \tag{22.26}$$

Eq. (22.26) can be used to obtain $\hat{q}_k$, for $k = 1, n$. Then $q(x)$ can be constructed using (22.6). This completes the solution of (22.24), 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.

One issue with spectral models involves the representation of topography. In many models, e.g., those that use the sigma 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. With bumpy continents and flat oceans, as schematically shown in Fig. 22.1, truncation leads to "bumpy" oceans. Various approaches have been suggested to alleviate this problem (Hoskins (1980); Navarra et al. (1994); Bouteloup (1995); Holzer (1996); Lindberg and Broccoli (1996)).

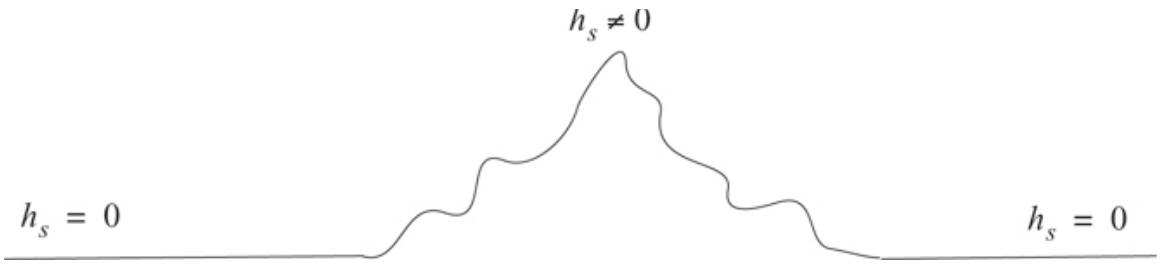



Figure 22.1: The Earth is bumpy.

## 22.3 Solving nonlinear equations with the spectral method

Now consider a *nonlinear* problem, such as momentum advection, i.e.,

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x}, \tag{22.27}$$

again with a periodic domain. Fourier expansion gives

$$\sum_{k=-n}^{n} \frac{d\hat{u}_k}{dt} e^{ikx} = -\left(\sum_{l=-n}^{n} \hat{u}_l e^{ilx}\right)\left(\sum_{m=-n}^{n} im\hat{u}_m e^{imx}\right). \tag{22.28}$$

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 the definition of our chosen truncation. The right-hand-side of (22.28) 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 this range are simply neglected, i.e., they are not included when we evaluate the left-hand side of (22.28).

For a given Fourier mode, (22.28) implies that

$$\frac{d\hat{u}_k}{dt} = -\left\{\sum_{l=-\alpha}^{\alpha}\sum_{m=-\alpha}^{\alpha} im\left[\hat{u}_l\hat{u}_m e^{i(l+m)x}\right]\right\} e^{-ikx}, \text{ for } -n \le k \le n. \tag{22.29}$$

*In* (22.29)*, 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 (22.29) are over the range $\pm\alpha$.

In order to get the exact value of $\frac{d\hat{u}_k}{dt}$ for all $k$ in the range $-n$ to $n$, we must choose $\alpha$ large enough so that we pick up all possible combinations of $l$ and $m$ that lie in the range $-n$ to $n$. See Fig. 22.2. The circled $X$s in the figure denote excluded triangular regions. The number of points in each triangular region is

$$1+2+3\ldots+(n-1) = \frac{n(n-1)}{2}. \tag{22.30}$$

The number of points *retained* is therefore given by

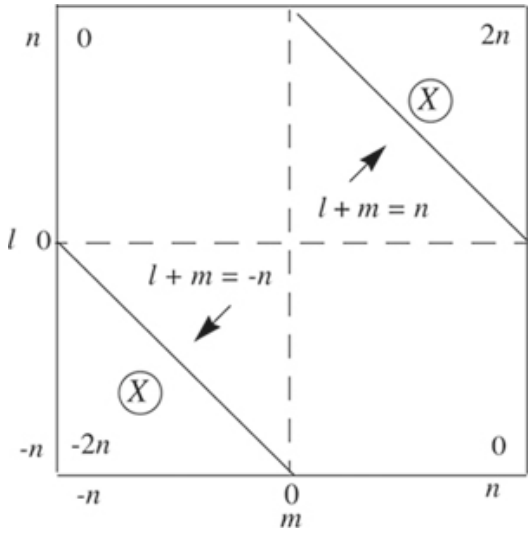



Figure 22.2: Cartoon "table" of $l+m$, showing which $(l,m)$ pairs can contribute to wave numbers $k$ in the range $-n$ to $n$ in Eq. (22.29). The pairs in the triangular regions marked by $X$'s do not contribute.

$$\begin{aligned}(2n+1)^2 - 2\left[\frac{n(n-1)}{2}\right] &= \left(4n^2+4n+1\right) - \left(n^2-n\right) \\ &= 3n^2+5n+1.\end{aligned} \tag{22.31}$$

This is the number of terms that must be evaluated inside the square brackets in (22.29). The number of terms is thus of order $n^2$, i.e., it grows very rapidly as $n$ increases. As a result, the amount of computation 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.

A way around this practical difficulty was proposed by Orszag, and independently by Eliassen et al., both in 1970. They suggested a "transform method" in which (22.6) and (22.8) are used to evaluate both $u$ and $\frac{\partial u}{\partial x}$ on a grid. We then compute the product $u\frac{\partial u}{\partial x}$ on the grid, and transform the result back into wavenumber 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 $u\frac{\partial u}{\partial x}$. Of course, 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 $u\frac{\partial u}{\partial x}$. Aliasing can still arise, however, from "cubic" or higher-order nonlinearities.

To investigate the transform method, we proceed as follows. By analogy with (22.7), we can write

$$\left(\widehat{u\frac{\partial u}{\partial x}}\right)_k = \frac{1}{M}\sum_{j=1}^{M}\left\{\left[u\left(x_j\right)\frac{\partial u}{\partial x}\left(x_j\right)\right]e^{-ikx_j}\right\}, -n \le k \le n. \tag{22.32}$$

Here we are computing the spectral coefficients of the *product* $u\frac{\partial u}{\partial x}$, by starting from $u\left(x_j\right)\frac{\partial u}{\partial x}\left(x_j\right)$, which is the grid-point representation of the same quantity. Using (22.6) and (22.8), we can express $u\left(x_j\right)$ and $\frac{\partial u}{\partial x}\left(x_j\right)$ in terms of Fourier series:

$$\left(\widehat{u\frac{\partial u}{\partial x}}\right)_k = \frac{1}{M}\sum_{j=1}^{M}\left[\left(\sum_{l=-n}^{n}\hat{u}_l e^{ilx_j}\right)\left(\sum_{m=-n}^{n} im\hat{u}_m e^{imx_j}\right)e^{-ikx_j}\right], -n \le k \le n. \tag{22.33}$$

It is important to note that, in (22.33), *the derivative, i.e.,* $\frac{\partial u}{\partial x}$*, 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. (22.33) is analogous to (22.9). When expanded, each term on the right-hand side of (22.33) 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 $u\frac{\partial u}{\partial x}$ is real, those $3n+1$ complex coefficients actually correspond to $3n+1$ independent real numbers. Therefore, we need

$$M \ge 3n+1 \tag{22.34}$$

grid points to represent $u\frac{\partial u}{\partial x}$ 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 $n$. This larger grid is sometimes called a "non-aliasing" grid.

In practice, the transform method to solve (22.27) 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 $\frac{\partial u}{\partial x}$ *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. (22.8), and the result obtained is used to compute grid-point values of $\frac{\partial u}{\partial x}$.

3. Form the product $u\frac{\partial u}{\partial x}$ *on the grid*.

4. Using (22.33), transform $u\frac{\partial u}{\partial x}$ back into wave-number space, for $-n \le k \le n$. This gives the coefficients $\widehat{\left(u\frac{\partial u}{\partial x}\right)}_k$.

5. Predict new values of the $\hat{u}_k$, using $\frac{d\hat{u}_k}{dt} = -\widehat{\left(u\frac{\partial u}{\partial x}\right)}_k$.

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.

## 22.4 Spectral methods on the sphere

Spectral methods on the sphere were first advocated by Silberman (1954). A function $F$ that is defined on the sphere can be represented by

$$F(\lambda,\varphi) = \sum_{m=-\infty}^{\infty} \sum_{n=|m|}^{\infty} F_n^m Y_n^m(\lambda,\varphi), \tag{22.35}$$

where the

$$Y_n^m(\lambda,\varphi) = e^{im\lambda} P_n^m(\sin\varphi) \tag{22.36}$$

are spherical harmonics, and the $P_n^m(\sin\varphi)$ are the associated Legendre functions of the first kind, which satisfy

$$P_n^m(\sin\varphi) = \frac{(2n)!}{2^n n!\,(n-m)!}\left(1-x^2\right)^{\frac{m}{2}}\left[x^{n-m} - \frac{(n-m)(n-m-1)}{2(2n-1)}\,x^{n-m-2} + \frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2\cdot 4\,(2n-1)(2n-3)}x^{n-m-4} - \ldots\right] \quad . \tag{22.37}$$

Here $m$ is the zonal wave number and $n-m$ is the "meridional nodal number." As discussed in the Appendix on spherical harmonics, it has to be true that $n \geq m$. The spherical harmonics ${Y_n}^m$ are the eigenfunctions of the Laplacian on the sphere:

$$\boxed{\nabla^2 Y_n^m = \frac{-n(n+1)}{a^2} Y_n^m}\,. \tag{22.38}$$

Here $a$ is the radius of the sphere. See the Appendix for further explanation.

We can approximate $F$ by a truncated sum:

$$\bar{F} = \sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} F_n^m Y_n^m. \tag{22.39}$$

Here the overbar indicates that $\bar{F}$ is an approximation to $F$. Recall that $m$ is the zonal wave number. In (22.39), the sum over $m$ from $-M$ to $M$ ensures that $\bar{F}$ is real. The choice of $N(m)$ is discussed below. For smooth $F$, $\bar{F}$ 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.

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.

$$\text{Rhomboidal truncation: } N = M + |m|\,, \text{ and} \tag{22.40}$$

$$\text{Triangular truncation: } N = M. \tag{22.41}$$

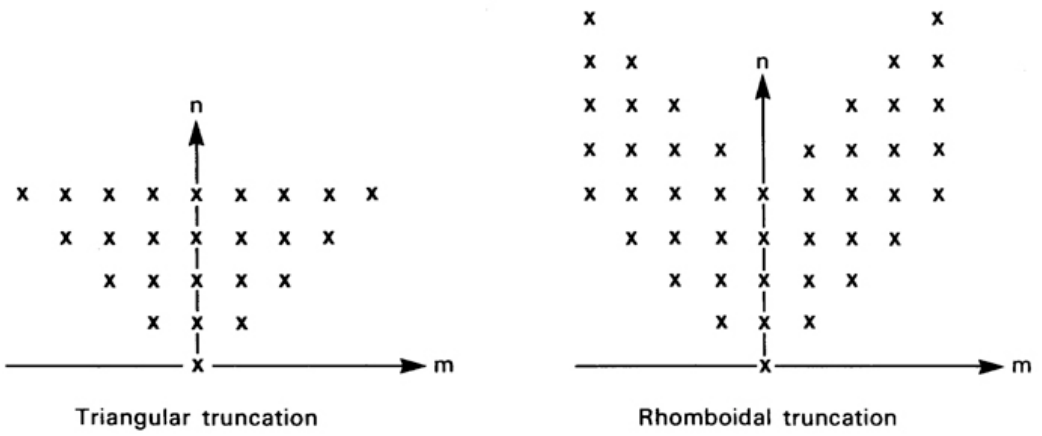



Figure 22.3: Rhomboidal and triangular truncation, both illustrated for the case $M = 4$. From Jarraud and Simmons (1983). In both cases, $n \geq m$. For a given value of $M$, rhomboidal truncation includes more spherical harmonics than triangular truncation.

Rhomboidal and triangular truncation are illustrated in Fig. 22.3. 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 $m$.

The figure 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

$$(M+1)^2 + M^2 + M \text{ for rhomboidal truncation,} \tag{22.42}$$

and

$$(M+1)^2 \text{ for triangular truncation.} \tag{22.43}$$

You can count the $x$'s in Fig. 22.3 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 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 $(\lambda, \varphi)$. 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.,

$$\bar{F}(\lambda_1, \varphi_1) = \bar{F}(\lambda_2, \varphi_2), \tag{22.44}$$

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 nice property, which is not shared by rhomboidal truncation.

As shown in Fig. 22.4, 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. With rhomboidal truncation the thin lines would be horizontal, and more modes would be kept, but they would not add much useful information.

## 22.5 Spherical harmonic transforms

In order to use (22.39) we need a "spherical harmonic transform," analogous to a Fourier transform. From (22.36), 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 $-1 \le x \le 1$, and we wish to evaluate

$$I = \int_{-1}^{1} f(x)\, dx, \tag{22.45}$$

by a numerical method. If $f(x)$ is defined at a finite number of points, denoted by $x_j$, then

$$I \cong \sum_{i=1}^{N} f(x_i)\, w_i, \tag{22.46}$$

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 (22.44) 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 (22.44) to evaluate the integral (22.43) *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.

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 $\ge 3M+1$. The number of latitude circles must be $\ge \frac{(3M+1)}{2}$ for triangular truncation, and so the total number of grid points needed is $\ge \frac{(3M+1)^2}{2}$. Referring back to (22.46), 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.

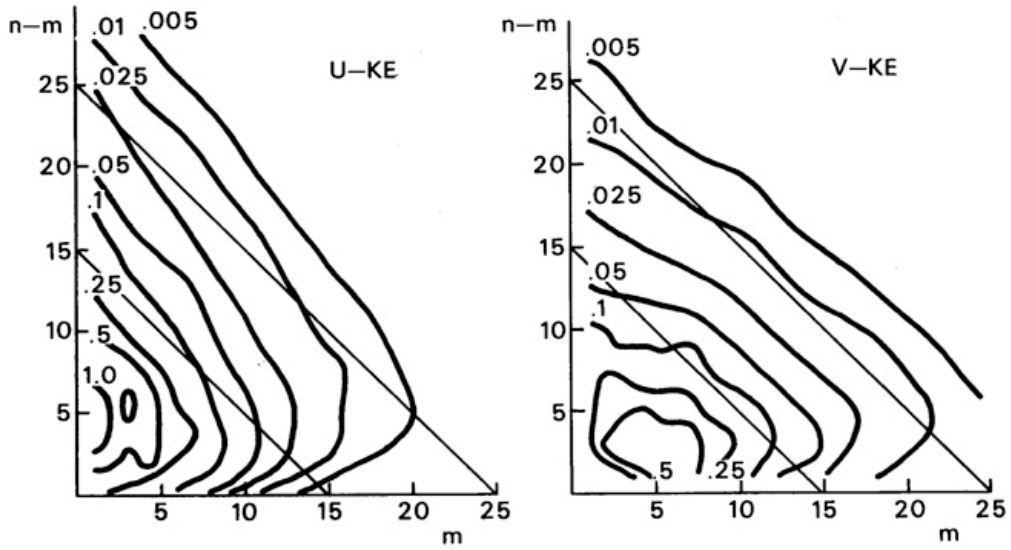



Figure 22.4: Percentage of total kinetic energy in each spectral component. From Jarraud and Simmons (1983), based on Baer (1972).

## 22.6 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. These are called the "Gaussian latitudes." 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 with an inverse Legendre transform. 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 the 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 $O\left(N^3\right)$, where $N$ is the number of spherical harmonics used. Finite-difference methods are, in effect, of $O\left(N^2\right)$. This means that spectral models become increasingly expensive, relative to grid-point models, at high resolution. Further comments are given later in this chapter.

## 22.7 Semi-implicit time differencing

As we have already discussed in Chapters 5 and 8, 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 16. 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. Such problems can be solved, of course, but they can be computationally expensive. 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.

## 22.8 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 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 artificial damping at all.

## 22.9 The "equivalent grid resolution" of spectral models

Laprise (1992) distinguishes four possible ways to answer the following very natural question: "What is the equivalent grid-spacing of a spectral model?"

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.

## 22.10 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 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.

## 22.11 Moisture advection

The mixing ratio of water vapor is non-negative. In Chapter 5, 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."

## 22.12 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.

## 22.13 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 an earlier chapter 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.

## 22.14 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.

## 22.15 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

$$q\left(x_j\right) = 14\cos\left(k_0 x_j\right) + 6i\cos\left(k_1 x_j\right) + 5, \tag{22.47}$$

where

$$\begin{aligned} k_0 &= \frac{2\pi}{L_0} \text{ and } L_0 = \frac{X}{4}, \\ k_1 &= \frac{2\pi}{L_1} \text{ and } L_1 = \frac{X}{8}. \end{aligned} \tag{22.48}$$

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, defined by

$$\begin{aligned} H(x) &= -1 \text{when the integer part of } x \text{ is odd}, \\ H(x) &= +1 \text{when the integer part of } x \text{ is even}. \end{aligned} \tag{22.49}$$

With this definition, $H(x)$ is discontinuous at integer points on the real line, and infinitely differentiable elsewhere. Sample $H(x)$ on the domain $-1 \leq x \leq 1$, using $M$ evenly spaced points, for $M = 11$, $M = 101$, and $M = 1001$, and feel free to go to larger values if you like. Plot the results for $-1 \leq x \leq 1$, for each value of $M$. Discuss in terms of the Gibbs phenomenon.

3. 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* nonlinearities?

4. 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 subproblems, use $n = 10$.

(a) List the spectral coefficients of $u$ and $\frac{\partial u}{\partial x}$.

(b) Work out the numerical values of the finite-difference coefficients that appear in Eq. (22.17).

(c) Show a plot that compares $\frac{\partial u}{\partial x}$ obtained using the spectral method with the corresponding result based on second-order centered finite-differences.

(d) Compute $-u\frac{\partial u}{\partial x}$ with both the direct method (22.28) and the transform method. Do the two methods agree?

# Chapter 23

# Finite-Element Methods

An explanation of finite-element methods.

$$a + b = c. \tag{23.1}$$

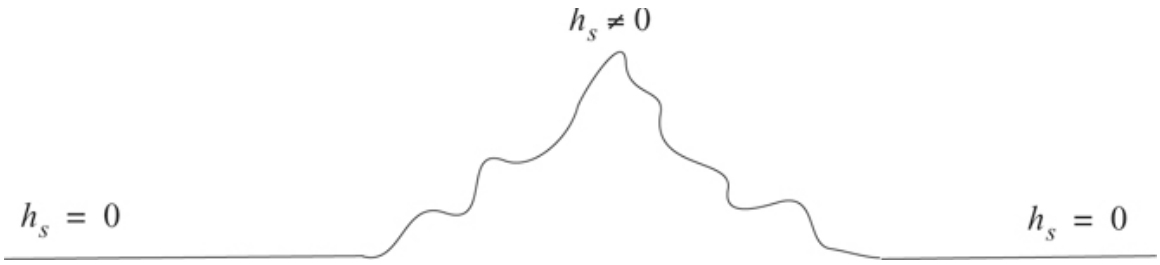



Figure 23.1: A figure about finite-element methods.

Our goal is

$$\begin{aligned} (2n+1)^2 - 2\left[\frac{n(n-1)}{2}\right] &= \left(4n^2+4n+1\right) - \left(n^2-n\right) \\ &= 3n^2+5n+1. \end{aligned} \tag{23.2}$$

Last bit of text.

## 23.1 Problems

1. Write subroutines to compute Fourier Let

$$a = b + c. \tag{23.3}$$

2. Consider a periodic step function, defined by

# Chapter 24

## Concluding discussion

The end

# Appendix A

## 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

$$\frac{dq}{dt} = f(q,t). \tag{A.1}$$

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

$$\frac{q^{n+1}-q^n}{\Delta t} = \frac{1}{6}(k_1, +2k_2 + 2k_3 + k_4), \tag{A.2}$$

where

$$\begin{aligned} k_1 &= f(q^n, n\Delta t), \\ k_2 &= f\left[q^n + \frac{k_1\Delta t}{2}, \left(n+\frac{1}{2}\right)\Delta t\right], \\ k_3 &= f\left[q^n + \frac{k_2\Delta t}{2}, \left(n+\frac{1}{2}\right)\Delta t\right], \\ k_4 &= f[q^n + k_3\Delta t, (n+1)\Delta t]. \end{aligned} \tag{A.3}$$

To demonstrate that A.2 has fourth-order accuracy, we substitute the exact solution into A.2 - A.3, and rearrange the result, to obtain

$$\frac{q^{n+1}-q^n}{\Delta t}-\frac{1}{6}\left(k_1,+2k_2+2k_3+k_4\right)=\varepsilon \tag{A.4}$$

where $\varepsilon$ is the discretization error of the scheme.

Taylor-series expansion allows us to write

$$\frac{q^{n+1}-q^n}{\Delta t}=\frac{dq}{dt}+\frac{1}{2!}\left(\Delta t\right)\frac{d^2q}{dt^2}+\frac{1}{3!}(\Delta t)^2\frac{d^3q}{dt^3}+\frac{1}{4!}(\Delta t)^3\frac{d^4q}{dt^4}+\mathrm{O}\left[(\Delta t)^4\right]. \tag{A.5}$$

Each term on the right-hand side of of A.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

$$\begin{aligned}\frac{dA}{dt}&\equiv\frac{\partial A}{\partial t}\frac{dt}{dt}+\frac{\partial A}{\partial q}\frac{dq}{dt}\\&=\frac{\partial A}{\partial t}+f\frac{\partial A}{\partial q}\\&\equiv\delta\left(A\right).\end{aligned} \tag{A.6}$$

Here a partial derivative with respect to $t$ is taken while holding $q$ constant, and vice versa. As a special case of A.6,

$$\delta f\equiv\frac{\partial f}{\partial t}+f\frac{\partial f}{\partial q}. \tag{A.7}$$

We can now write

$$\boxed{\frac{dq}{dt}=f}, \tag{A.8}$$

$$\boxed{\frac{d^2q}{dt^2} = \delta f}\,, \tag{A.9}$$

and

$$\begin{aligned}
\frac{d^3q}{dt^3} &= \delta\left(\delta f\right) \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right)\frac{dq}{dt} + \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial q}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)\frac{dq}{dt} + \frac{\partial^2 f}{\partial t^2} \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f + \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial q}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) f + \frac{\partial^2 f}{\partial t^2} \\
&= \left[\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)\right] f + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) f + \frac{\partial^2 f}{\partial t^2} \\
&= \frac{\partial}{\partial q}\left(\frac{\partial f}{\partial q}\right) f^2 + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \left(\frac{\partial f}{\partial q}\right)(\delta f) + \frac{\partial^2 f}{\partial t^2} \\
&= f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2} + \left(\frac{\partial f}{\partial q}\right)(\delta f) \\
&= \left(f\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)^2 f + \left(\frac{\partial f}{\partial q}\right)(\delta f)\,,
\end{aligned} \tag{A.10}$$

so that

$$\boxed{\frac{d^3q}{dt^3} = \left(\delta^2 f\right) + f_q\left(\delta f\right)}\,. \tag{A.11}$$

Here we have used the notation $f_q \equiv \frac{\partial f}{\partial q}$. Finally, $\frac{d^4q}{dt^4}$ is given by

$$\frac{d^4q}{dt^4} = \delta\left[\left(\delta^2 f\right) + f_q\left(\delta f\right)\right]$$
$$= \left[f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right)\right] + \left[f\frac{\partial}{\partial q}\left(f_q\right) + \frac{\partial}{\partial t}\left(f_q\right)\right]\left(\delta f\right) + f_q\left[f\frac{\partial}{\partial q}\left(\delta f\right) + \frac{\partial}{\partial t}\left(\delta f\right)\right]. \tag{A.12}$$

This is a bit messy. To break the analysis of A.12 into steps, we first manipulate the first term in square brackets, separately. Expanding, we find that

$$\begin{aligned}
&f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right) \\
&= f\frac{\partial}{\partial q}\left[f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2}\right] + \frac{\partial}{\partial t}\left[f^2\left(\frac{\partial^2 f}{\partial q^2}\right) + 2f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial t^2}\right] \\
&= f\left[2f\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q^2}\right) + f^2\left(\frac{\partial^3 f}{\partial q^3}\right) + 2\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q\partial t}\right) + 2f\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + \left(\frac{\partial^3 f}{\partial t^2\partial q}\right)\right] \\
&+ 2f\frac{\partial f}{\partial t}\left(\frac{\partial^2 f}{\partial q^2}\right) + f^2\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + 2\left(\frac{\partial f}{\partial t}\right)\left(\frac{\partial^2 f}{\partial q\partial t}\right) + 2f\left(\frac{\partial^3 f}{\partial t^2\partial q}\right) + \frac{\partial^3 f}{\partial t^3}
\end{aligned} \tag{A.13}$$

The terms can be collected and grouped as follows:

$$\begin{aligned}
f\frac{\partial}{\partial q}\left(\delta^2 f\right) + \frac{\partial}{\partial t}\left(\delta^2 f\right) &= \left[f^3\left(\frac{\partial^3 f}{\partial q^3}\right) + 3f^2\left(\frac{\partial^3 f}{\partial q^2\partial t}\right) + 3f\left(\frac{\partial^3 f}{\partial t^2\partial q}\right) + \frac{\partial^3 f}{\partial t^3}\right] \\
&+ f\left[2f\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q^2}\right) + 2\frac{\partial f}{\partial q}\left(\frac{\partial^2 f}{\partial q\partial t}\right)\right] + 2f\frac{\partial f}{\partial t}\left(\frac{\partial^2 f}{\partial q^2}\right) + 2\left(\frac{\partial f}{\partial t}\right)\left(\frac{\partial^2 f}{\partial q\partial t}\right) \\
&= \delta^3 f + 2\left(\delta f\right)\left(\delta f_q\right).
\end{aligned} \tag{A.14}$$

Substituting into A.2, we find that

$$\begin{aligned}
\frac{d^4q}{dt^4} &= \left[f\frac{\partial}{\partial q}\left(\delta^2 f\right)+\frac{\partial}{\partial t}\left(\delta^2 f\right)\right]+\left[f\frac{\partial}{\partial q}\left(f_q\right)+\frac{\partial}{\partial t}\left(f_q\right)\right](\delta f)+f_q\left[f\frac{\partial}{\partial q}(\delta f)+\frac{\partial}{\partial t}(\delta f)\right] \\
&= \left[\left(\delta^3 f\right)+2(\delta f)\left(\delta f_q\right)\right]+\left(\delta f_q\right)(\delta f)+f_q\left\{f\frac{\partial}{\partial q}\left[f\left(f_q\right)+\frac{\partial f}{\partial t}\right]+\frac{\partial}{\partial t}\left[f\left(f_q\right)+\frac{\partial f}{\partial t}\right]\right\} \\
&= \left(\delta^3 f\right)+3(\delta f)\left(\delta f_q\right)+f_q\left[\left(f\frac{\partial f}{\partial q}\right)\left(f_q\right)+f\left(f\frac{\partial f_q}{\partial q}\right)+f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)+\left(\frac{\partial f}{\partial t}\right)\left(f_q\right)+f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right] \\
&= \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2\left(f\frac{\partial f}{\partial q}+\frac{\partial f}{\partial t}\right)+f_q\left[f\left(f\frac{\partial f_q}{\partial q}\right)+f\frac{\partial}{\partial q}\left(\frac{\partial f}{\partial t}\right)+f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right] \\
&= \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2(\delta f)+f_q\left[f\left(f\frac{\partial f_q}{\partial q}\right)+2f\frac{\partial f_q}{\partial t}+\frac{\partial^2 f}{\partial t^2}\right],
\end{aligned} \tag{A.15}$$

which can be written as

$$\boxed{\frac{d^4q}{dt^4} = \left(\delta^3 f\right)+3\left(\delta f_q\right)(\delta f)+\left(f_q\right)^2(\delta f)+f_q\left(\delta^2 f\right)}. \tag{A.16}$$

Next, we express $k_1$ - $k_4$ in terms of $f(q,t)$ and its derivatives. We write

$$\boxed{k_1 = f}, \tag{A.17}$$

and

$$\begin{aligned}
k_2 &= f\left[q^n+k_1\frac{\Delta t}{2},\left(n+\frac{1}{2}\right)\Delta t\right] \\
&= f+\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]f \\
&+\frac{1}{2!}\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]^2 f+\frac{1}{3!}\left[\left(\frac{\Delta t}{2}\right)\left(f\frac{\partial}{\partial q}+\frac{\partial}{\partial t}\right)\right]^3 f+\mathrm{O}\left[(\Delta t)^4\right],
\end{aligned} \tag{A.18}$$

which is equivalent to

$$\boxed{k_2 = f + \left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 \delta^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 \delta^3 f + \mathrm{O}\left[(\Delta t)^4\right]} \tag{A.19}$$

Here we have used a two-dimensional Taylor's series expansion, because we have two independent variables, namely $q$ and $t$. Similarly,

$$\begin{aligned} k_3 &= f\left[q^n + k_2\frac{\Delta t}{2}, \left(n+\frac{1}{2}\right)\Delta t\right] \\ &= f + \left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right] f + \frac{1}{2!}\left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^2 f \\ &+ \frac{1}{3!}\left[\frac{\Delta t}{2}\left(k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^3 f + \mathrm{O}\left[(\Delta t)^4\right]. \end{aligned} \tag{A.20}$$

From A.6 and A.20, we see that

$$k_2\frac{\partial}{\partial q} + \frac{\partial}{\partial t} = \delta + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 \delta^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 \delta^3 f\right]\frac{\partial}{\partial q} + \mathrm{O}\left[(\Delta t)^4\right]. \tag{A.21}$$

Substituting A.9 into A.10, we obtain

$$\begin{aligned} k_3 = f &+ \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2 (\delta^2 f)\right]\frac{\partial f}{\partial q}\right\} \\ &+ \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left[\delta f + \left(\frac{\Delta t}{2}\right)\delta f\frac{\partial}{\partial q}\right]^2 f + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3 (\delta^3 f) + \mathrm{O}\left[(\Delta t)^4\right]. \end{aligned} \tag{A.22}$$

Expand, and combine terms:

$$
\begin{aligned}
k_3 &= f + \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left(\delta^2 f\right)\right] f_q\right\} \\
&\quad + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left\{\Delta t\left(\delta f\right)\left(\delta f_q\right) + \left(\delta^2 f\right) + \mathrm{O}\left[(\Delta t)^2\right]\right\} + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3\left(\delta^3 f\right) + \mathrm{O}\left[(\Delta t)^4\right] \\
&= f + \left(\frac{\Delta t}{2}\right)\left\{\delta f + \left[\left(\frac{\Delta t}{2}\right)\delta f + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left(\delta^2 f\right)\right] f_q\right\} \\
&\quad + \frac{1}{2!}\left(\frac{\Delta t}{2}\right)^2\left[\Delta t\left(\delta f\right)\left(\delta f_q\right) + \left(\delta^2 f\right)\right] + \frac{1}{3!}\left(\frac{\Delta t}{2}\right)^3\left(\delta^3 f\right) + \mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}
\tag{A.23}
$$

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 A.23. Collecting powers of $\Delta t$, we conclude that

$$
\boxed{
\begin{aligned}
k_3 &= f + \Delta t\left(\frac{\delta f}{2}\right) + \frac{(\Delta t)^2}{2!}\left[\frac{\left(\delta f\right) f_q}{2} + \frac{\delta^2 f}{4}\right] \\
&\quad + \frac{(\Delta t)^3}{3!}\left[\frac{3\left(\delta^2 f\right) f_q}{8} + \frac{3\left(\delta f\right)\left(\delta f_q\right)}{4} + \frac{\left(\delta^3 f\right)}{8}\right] + \mathrm{O}\left[(\Delta t)^4\right]
\end{aligned}
}.
\tag{A.24}
$$

It remains to assemble $k_4$. We start with the two-dimensional Taylor series expansion:

$$
\begin{aligned}
k_4 &= f\left[q^n + k_3\Delta t, (n+1)\Delta t\right] \\
&= f + \Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right) f + \frac{1}{2!}\left[\Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^2 f + \frac{1}{3!}\left[\Delta t\left(k_3\frac{\partial}{\partial q} + \frac{\partial}{\partial t}\right)\right]^3 f + \mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}
\tag{A.25}
$$

Next, use A.6 and A.14 to write

$$k_3\frac{\partial}{\partial q}+\frac{\partial}{\partial t}=$$
$$\delta+\left\{\Delta t\left(\frac{\delta f}{2}\right)+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]+\frac{(\Delta t)^3}{3!}\left[\frac{3\left(\delta^2 f\right)f_q}{8}+\frac{3(\delta f)\left(\delta f_q\right)}{4}+\frac{\left(\delta^3 f\right)}{8}\right]\right\}$$
$$\frac{\partial}{\partial q}+\mathrm{O}\left[(\Delta t)^4\right]. \tag{A.26}$$

Substituting A.26 into A.25, we obtain

$$\begin{aligned}
k_4 &= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left[\delta+\Delta t\left(\frac{\delta f}{2}\right)\frac{\partial}{\partial q}\right]^2 f+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right]\\
&= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left\{\left[\left(\delta^2 f\right)+\Delta t(\delta f)\left(\delta f_q\right)\right]+\mathrm{O}\left[(\Delta t)^2\right]\right\}+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right]\\
&= f+\Delta t\left\{\delta f+\Delta t\left(\frac{\delta f}{2}\right)f_q+\frac{(\Delta t)^2}{2!}\left[\frac{(\delta f)f_q}{2}+\frac{\delta^2 f}{4}\right]f_q\right\}\\
&+\frac{(\Delta t)^2}{2!}\left[\left(\delta^2 f\right)+\Delta t(\delta f)\left(\delta f_q\right)\right]+\frac{(\Delta t)^3}{3!}\left(\delta^3 f\right)+\mathrm{O}\left[(\Delta t)^4\right].
\end{aligned} \tag{A.27}$$

As before, we have written only terms that contribute up to $O\left[(\Delta t)^3\right]$. Now collect powers of $\Delta t$:

$$\boxed{\begin{aligned}
k_4 &= f+(\Delta t)\,\delta f+\frac{(\Delta t)^2}{2!}\left[(\delta f)f_q+\left(\delta^2 f\right)\right]\\
&+\frac{(\Delta t)^3}{3!}\left[\frac{3(\delta f)\left(f_q\right)^2}{2}+\frac{3\left(\delta^2 f\right)f_q}{4}+3(\delta f)\left(\delta f_q\right)+\left(\delta^3 f\right)\right]+\mathrm{O}\left[(\Delta t)^4\right].
\end{aligned}} \tag{A.28}$$

The final step is to substitute the various boxed relations above into A.4. Everything cancels out, and we are left with $\varepsilon = O\left[(\Delta t)^4\right]$. This completes the proof.

# Appendix B

# Vectors, Coordinates, and Coordinate Transformations

## B.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 in a discussion of physics. For obvious reasons, spherical coordinates are particularly useful in geophysics.

Any physical law should be expressible in a form that is invariant with respect to our choice of coordinate systems; we certainly do not expect that the laws of physics change when we switch from spherical coordinates to cartesian coordinates! It follows that we should be able to express physical laws without making reference to any coordinate system. 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.

## B.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, $(\lambda, \varphi, r)$, where $\lambda$ is longitude, $\varphi$ 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

$$\mathbf{V}' = \mathbf{M}\mathbf{V}. \tag{B.1}$$

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 (B.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 (B.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

$$\mathbf{V} = \dot{x}\mathbf{i} + \dot{y}\mathbf{j} + \dot{z}\mathbf{k} \tag{B.2}$$

$$\mathbf{V} = \dot{\lambda} r \cos\varphi \mathbf{e}_\lambda + r\dot{\varphi}\mathbf{e}_\varphi + \dot{r}\mathbf{e}_r \tag{B.3}$$

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. (B.2) and (B.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 flux of momentum. The momentum flux, also called a "stress," and equivalent to a force per unit area, has a magnitude and "two directions." One of the directions is associated with the force vector itself, and the other is associated with the normal vector to the unit area in question. The momentum flux tensor can be written as $\rho\mathbf{V}\otimes\mathbf{V}$, where $\rho$ 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

$$\mathbf{T}' = \mathbf{MTM}^{-1} \tag{B.4}$$

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.

## B.3 Differential operators

Several familiar differential operators can be defined without reference to any coordinate system. These operators are 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 too) are:

- the gradient, denoted by $\nabla A$, where $A$ 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

$$\begin{aligned} J(\alpha,\beta) \equiv &\quad \mathbf{k} \cdot (\nabla\alpha \times \nabla\beta) \\ = &\quad \mathbf{k} \cdot \nabla \times (\alpha \nabla\beta) \\ = &- \mathbf{k} \cdot \nabla \times (\beta \nabla\alpha) . \end{aligned} \tag{B.5}$$

Here the gradient operators are understood to produce vectors in the two-dimensional space, $\alpha$ and $\beta$ are arbitrary scalars, and $\mathbf{k}$ is a unit vector perpendicular to the two-dimensional surface. The second and third lines of (B.5) can be derived with the use of vector identities found in a table later in this QuickStudy.

A definition of the gradient operator that does not make reference to any coordinate system is:

$$\nabla A \equiv \lim_{V\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}A dS\right], \tag{B.6}$$

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 (B.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 and in the and directions, respectively. Consider a "box" of width $\Delta x$ and height $\Delta y$, as shown in Figure B.1. We can write

$$\begin{aligned}\nabla A \equiv \lim_{(\Delta x,\Delta y)\to 0}\Bigg\{\frac{1}{\Delta x\Delta y}\Bigg[&A\left(x_0+\frac{\Delta x}{2},y_0\right)\Delta y\mathbf{i}+A\left(x_0,y_0+\frac{\Delta y}{2}\right)\Delta x\mathbf{j}\\ &-A\left(x_0-\frac{\Delta x}{2},y_0\right)\Delta y\mathbf{i}-A\left(x_0,y_0-\frac{\Delta y}{2}\right)\Delta x\mathbf{j}\Bigg]\Bigg\}\\ &=\frac{\partial A}{\partial x}\mathbf{i}+\frac{\partial A}{\partial y}\mathbf{j}\,.\end{aligned} \tag{B.7}$$

This is the answer that we expect.

Definitions of the divergence and curl operators that do not make reference to any coordinate system are:

$$\nabla\cdot\mathbf{Q} \equiv \lim_{V\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\cdot\mathbf{Q} dS\right] \tag{B.8}$$

$$\nabla\times\mathbf{Q} \equiv \lim_{V\to 0}\left[\frac{1}{V}\oint_S \mathbf{n}\times\mathbf{Q} dS\right] \tag{B.9}$$

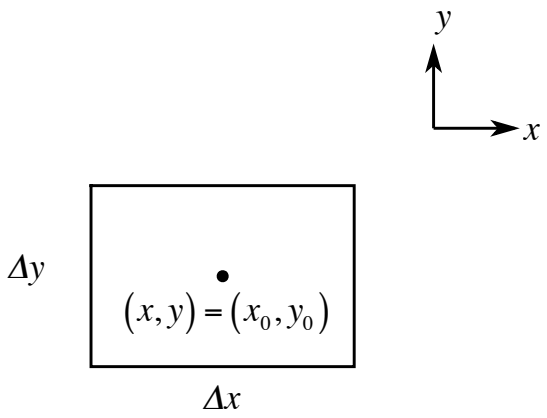



Figure B.1: A rectangular box in a planar two-dimensional space, with center at $(x_0, y_0)$, width $\Delta x$, and height $\Delta y$.

It is possible to work through exercises similar to (B.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

$$J(A,B) = \lim_{C \to 0} \left[ \oint_C A \nabla B \cdot \mathbf{t} dl \right], \tag{B.10}$$

where $\mathbf{t}$ is a unit vector that is tangent to the bounding curve $C$.

## B.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 $\alpha$ and $\beta$ be arbitrary scalars, let $\mathbf{V}$, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ be arbitrary vectors, and let $\mathbf{T}$ be an arbitrary tensor of rank 2. Then:

$$\nabla \times (\nabla \alpha) = 0 \tag{B.11}$$

$$\nabla \cdot (\nabla \times \mathbf{V}) = 0 \tag{B.12}$$

$$\mathbf{A}\times\mathbf{B}=-\mathbf{B}\times\mathbf{A} \tag{B.13}$$

$$\nabla\cdot(\alpha\mathbf{V})=\alpha(\nabla\cdot\mathbf{V})+\mathbf{V}\cdot\nabla\alpha \tag{B.14}$$

$$\nabla\cdot(\mathbf{A}\times\mathbf{B})=(\nabla\times\mathbf{A})\cdot\mathbf{B}-(\nabla\times\mathbf{B})\cdot\mathbf{A} \tag{B.15}$$

$$\nabla\times(\alpha\mathbf{V})=\nabla\alpha\times\mathbf{V}+\alpha(\nabla\times\mathbf{V}) \tag{B.16}$$

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})=(\mathbf{A}\times\mathbf{B})\cdot\mathbf{C}=\mathbf{B}\cdot(\mathbf{C}\times\mathbf{A}) \tag{B.17}$$

$$\mathbf{A}\times(\mathbf{B}\times\mathbf{C})=\mathbf{B}(\mathbf{C}\cdot\mathbf{A})-\mathbf{C}(\mathbf{A}\cdot\mathbf{B}) \tag{B.18}$$

$$\nabla\times(\mathbf{A}\times\mathbf{B})=\mathbf{A}(\nabla\cdot\mathbf{B})-\mathbf{B}(\nabla\cdot\mathbf{A})-(\mathbf{A}\cdot\nabla)\mathbf{B}+(\mathbf{B}\cdot\nabla)\mathbf{A} \tag{B.19}$$

$$\nabla(\mathbf{A}\cdot\mathbf{B})=(\mathbf{A}\cdot\nabla)\mathbf{B}+(\mathbf{B}\cdot\nabla)\mathbf{A}+\mathbf{A}\times(\nabla\times\mathbf{B})+\mathbf{B}\times(\nabla\times\mathbf{A}) \tag{B.20}$$

$$\begin{aligned} J(\alpha,\beta)\equiv\mathbf{k}\cdot(\nabla\alpha\times\nabla\beta)&=\quad\mathbf{k}\cdot\nabla\times(\alpha\nabla\beta)\\ &=-\mathbf{k}\cdot\nabla\times(\beta\nabla\alpha)\\ &=-\mathbf{k}\cdot(\nabla\beta\times\nabla\alpha) \end{aligned} \tag{B.21}$$

$$\nabla^2 \mathbf{V} \equiv (\nabla \cdot \nabla)\,\mathbf{V} = \nabla\,(\nabla \cdot \mathbf{V}) - \nabla \times (\nabla \times \mathbf{V}) \tag{B.22}$$

$$\nabla \cdot (\mathbf{A} \otimes \mathbf{B}) = (\mathbf{A} \cdot \nabla)\,\mathbf{B} + (\mathbf{B} \cdot \nabla)\,\mathbf{A} \tag{B.23}$$

$$\nabla \cdot (\alpha \mathbf{T}) = (\nabla \alpha) \cdot \mathbf{T} + \alpha\,(\nabla \cdot \mathbf{T}) \tag{B.24}$$

In (B.23), $\otimes$ denotes the outer or dyadic product of two vectors, which yields a tensor of rank 2.

A special case of (B.20) is

$$\frac{1}{2}\nabla\,(\mathbf{V} \cdot \mathbf{V}) = (\mathbf{V} \cdot \nabla)\,\mathbf{V} + \mathbf{V} \times (\nabla \times \mathbf{V}) \tag{B.25}$$

This identity is used to write the advection terms of the momentum equation in alternative forms.

Identity (B.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. (B.22) can be used, for example, in a parameterization of momentum diffusion.

## B.5 Spherical coordinates

### B.5.1 Vector operators in spherical coordinates

The gradient, divergence, curl, Laplacian, and Jacobian operators can be expressed in spherical coordinates as follows:

$$\nabla \alpha = \left( \frac{1}{r \cos\varphi} \frac{\partial \alpha}{\partial \lambda},\ \frac{1}{r} \frac{\partial \alpha}{\partial \varphi},\ \frac{\partial \alpha}{\partial r} \right) \tag{B.26}$$

$$\nabla \cdot \mathbf{V} = \frac{1}{r\cos\varphi}\frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{r\cos\varphi}\frac{\partial}{\partial \varphi}\left(V_\varphi \cos\varphi\right) + \frac{1}{r^2}\frac{\partial}{\partial r}\left(V_r r^2\right) \tag{B.27}$$

$$\nabla \times \mathbf{V} = \left\{ \frac{1}{r}\left[\frac{\partial V_r}{\partial \varphi} - \frac{\partial}{\partial r}\left(rV_\varphi\right)\right],\ \frac{1}{r}\frac{\partial}{\partial r}\left(rV_\lambda\right) - \frac{1}{r\cos\varphi}\frac{\partial V_r}{\partial \lambda},\ \frac{1}{r\cos\varphi}\left[\frac{\partial V_\varphi}{\partial \lambda} - \frac{\partial}{\partial \varphi}\left(V_\lambda \cos\varphi\right)\right]\right\} \tag{B.28}$$

$$\nabla^2 \alpha = \frac{1}{r^2\cos\varphi}\left[\frac{\partial}{\partial \lambda}\left(\frac{1}{\cos\varphi}\frac{\partial \alpha}{\partial \lambda}\right) + \frac{\partial}{\partial \varphi}\left(r^2\cos\varphi\frac{\partial \alpha}{\partial r}\right)\right] \tag{B.29}$$

$$J(\alpha,\beta) = \frac{1}{r^2\cos\varphi}\left(\frac{\partial \alpha}{\partial \lambda}\frac{\partial \beta}{\partial \varphi} - \frac{\partial \beta}{\partial \lambda}\frac{\partial \alpha}{\partial \varphi}\right) \tag{B.30}$$

Here $\alpha$ is an arbitrary scalar, and $\mathbf{V}$ is an arbitrary vector.

### B.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 (B.19) is

$$\mathbf{e}_r \cdot [\nabla \times (\mathbf{e}_r \times \mathbf{H})] = \nabla \cdot \mathbf{H}, \tag{B.31}$$

where $\mathbf{e}_r$ is the unit vector pointing upward, and $\mathbf{H}$ is an arbitrary horizontal vector. In words, the curl of $\mathbf{e}_r \times \mathbf{H}$ is equal to the divergence of $\mathbf{H}$. Similarly, a useful special case of (B.15) is

$$\nabla \cdot (\mathbf{e}_r \times \mathbf{H}) = -\mathbf{e}_r \cdot (\nabla \times \mathbf{H}) \tag{B.32}$$

This means that the divergence of $\mathbf{e}_r \times \mathbf{H}$ is equal to minus the curl of $\mathbf{H}$.

If $\mathbf{V}$ is separated into a horizontal vector and a vertical vector, as in

$$\mathbf{V} = \mathbf{V}_h + V_r \mathbf{e}_r, \tag{B.33}$$

then (B.28)) can be written as

$$\boxed{\nabla \times (\mathbf{V}_h + V_r \mathbf{e}_r) = \nabla_r \times \mathbf{V}_h + \mathbf{e}_r \times \left[ \frac{1}{r} \frac{\partial}{\partial r} (r \mathbf{V}_h) - \nabla_r V_r \right].} \tag{B.34}$$

In case $\mathbf{V}$ is the velocity, the first term on the right-hand side of (B.34) is the vertical component of the vorticity, and the second term is the horizontal vorticity vector. Eq. (B.34) 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 $\nabla_r \times \mathbf{V}_h$, and a horizontal part, given by $\mathbf{e}_r \times \left[ \frac{1}{r} \frac{\partial}{\partial r} (r \mathbf{V}_h) - \nabla_r V_r \right]$. The *two-dimensional* curl of a horizontal vector has only a vertical component, namely $\nabla_r \times \mathbf{V}_h$.

Finally, the 3D curl of the 3D curl of a 3D vector is given by

$$\nabla \times [\nabla \times (\mathbf{V}_h + V_r \mathbf{e}_r)] = \nabla_r \times \boldsymbol{\eta} + \mathbf{e}_r \times \left[ \frac{1}{r} \frac{\partial}{\partial r} (r \boldsymbol{\eta}) - \nabla_r \zeta \right], \tag{B.35}$$

where

$$\boldsymbol{\eta} \equiv \mathbf{e}_r \times \left[ \frac{1}{r} \frac{\partial}{\partial r} (r \mathbf{V}_h) - \nabla_r V_r \right], \tag{B.36}$$

and

$$\zeta \equiv \mathbf{e}_r \cdot (\nabla_r \times \mathbf{V}_h). \tag{B.37}$$

### B.5.3 Derivation of the gradient operator in spherical coordinates

Consider how the two-dimensional version of (B.26) can be derived from (B.6). Figure B.2 illustrates the problem. Here we have replaced $r$ by $a$, the radius of the Earth. The angle $\theta$ 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 $\nabla A$ is defined, are different from their respective directions on either east-west wall of the area element. Inspection of Figure B.2 shows that $\theta$ satisfies

$$\begin{aligned}\sin\theta &= \frac{-\frac{1}{2}\left[a\cos\left(\varphi+d\varphi\right)-a\cos\varphi\right]d\lambda}{ad\varphi}\\ &\to -\frac{1}{2}\left(\frac{\partial}{\partial\varphi}\cos\varphi\right)d\lambda\\ &= \frac{1}{2}\sin\varphi d\lambda\,.\end{aligned} \tag{B.38}$$

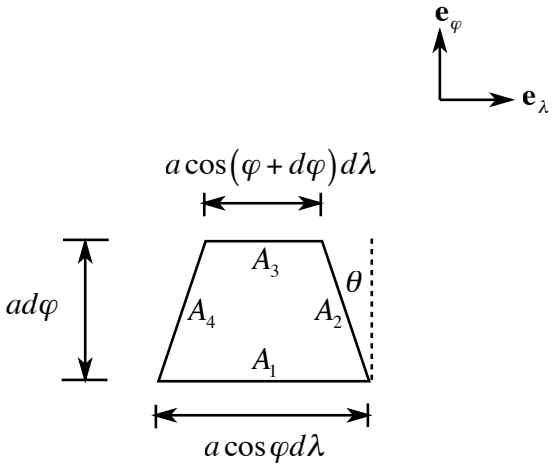



Figure B.2: A patch of the sphere, with longitudinal width $a\cos\varphi d\lambda$, and latitudinal height $ad\varphi$.

The angle $\theta$ is of "differential" or infinitesimal size. Nevertheless, it is needed in the derivation of (B.26). The line integral in (B.6) can be expressed as

$$\begin{aligned}\frac{1}{\text{Area}}\oint A\mathbf{n}dl =&\frac{1}{a^2\cos\varphi d\lambda d\varphi}\left[-\mathbf{e}_\varphi A_1 a\cos\varphi d\lambda+\mathbf{e}_\lambda A_2\cos\theta a d\varphi+\mathbf{e}_\varphi A_2\sin\theta a d\varphi\right.\\&\left.+\mathbf{e}_\varphi A_3 a\cos(\varphi+d\varphi)-\mathbf{e}_\lambda A_4\cos\theta a d\varphi+\mathbf{e}_\varphi A_4\sin\theta a d\varphi\right]\\=&\mathbf{e}_\lambda\frac{(A_2-A_4)\cos\theta}{a\cos\varphi d\lambda}\\&+\mathbf{e}_\varphi\left\{\frac{[A_3\cos(\varphi+d\varphi)-A_1\cos\varphi]d\lambda+(A_2+A_4)\sin\theta d\varphi}{a\cos\varphi d\lambda d\varphi}\right\}\end{aligned} \tag{B.39}$$

Note how the angle $\theta$ has entered here. Put $\cos\theta\to 1$ and $\sin\theta\to\frac{1}{2}\sin\varphi d\lambda$ to obtain

$$\begin{aligned}\frac{1}{\text{Area}}\oint\alpha\mathbf{n}dl=&\mathbf{e}_\lambda\frac{(A_2-A_4)}{a\cos\varphi d\lambda}+\mathbf{e}_\varphi\left\{\left[\frac{A_3\cos(\varphi+d\varphi)-A_1\cos\varphi}{a\cos\varphi d\varphi}\right]+\left(\frac{A_2+A_4}{2}\right)\frac{\sin\varphi}{a\cos\varphi}\right\}\\\to&\mathbf{e}_\lambda\frac{1}{a\cos\varphi}\frac{\partial A}{\partial\lambda}+\mathbf{e}_\varphi\left[\frac{1}{a\cos\varphi}\frac{\partial}{\partial\varphi}(A\cos\varphi)+\frac{A\sin\varphi}{a\cos\varphi}\right]\\=&\mathbf{e}_\lambda\frac{1}{a\cos\varphi}\frac{\partial A}{\partial\lambda}+\mathbf{e}_\varphi\frac{1}{a}\frac{\partial A}{\partial\varphi},\end{aligned} \tag{B.40}$$

which agrees with the two-dimensional version of (B.26).

Similar (but more straightforward) derivations can be given for (B.27) - (B.30).

### B.5.4 Applying vector operators to the unit vectors in spherical coordinates

Using (B.11) - (B.13), we can prove the following about the unit vectors in spherical coordinates:

$$\nabla\cdot\mathbf{e}_\lambda=0, \tag{B.41}$$

$$\nabla\cdot\mathbf{e}_\varphi=-\frac{\tan\varphi}{r}, \tag{B.42}$$

$$\nabla \cdot \mathbf{e}_r = \frac{2}{r}, \tag{B.43}$$

$$\nabla \times \mathbf{e}_\lambda = \frac{\mathbf{e}_\varphi}{r} + \frac{\tan\varphi}{r}\mathbf{e}_r, \tag{B.44}$$

$$\nabla \times \mathbf{e}_\varphi = -\frac{\mathbf{e}_\lambda}{r}, \tag{B.45}$$

$$\nabla \times \mathbf{e}_r = 0. \tag{B.46}$$

The following relations are useful when working with the momentum equation in spherical coordinates:

$$(\mathbf{V}_h \cdot \nabla)\,\mathbf{e}_\lambda = \frac{u\sin\varphi}{r}\mathbf{e}_\varphi - \frac{u\cos\varphi}{r}\mathbf{e}_r\,, \tag{B.47}$$

$$(\mathbf{V}_h \cdot \nabla)\,\mathbf{e}_\varphi = -\frac{u\sin\varphi}{r}\mathbf{e}_\lambda - \frac{v\sin\varphi}{r}\mathbf{e}_r, \tag{B.48}$$

$$(\mathbf{V}_h \cdot \nabla)\,\mathbf{e}_r = \frac{\mathbf{V}_h}{r}\,. \tag{B.49}$$

Here $\mathbf{V}_h$ is the horizontal wind vector.

## B.6 Solid body rotation

As an example of the application of (B.28), the vertical component of the vorticity is

$$\zeta = \frac{1}{r\cos\varphi}\left[\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \varphi}(u\cos\varphi)\right] \tag{B.50}$$

For the case of pure solid body rotation of the atmosphere about the Earth's axis of rotation, we have

$$u = \dot{\lambda} r\cos\varphi \text{ and } v = 0, \tag{B.51}$$

where $\dot{\lambda}$ is independent of $\varphi$. Substitution of (B.51) into (B.50) gives

$$\begin{aligned}\zeta =& \frac{-1}{r\cos\varphi}\frac{\partial}{\partial\varphi}\left(\dot{\lambda} r\cos^2\varphi\right)\\ =& 2\dot{\lambda}\sin\varphi\ .\end{aligned} \tag{B.52}$$

This is the expected form of the vorticity associated with the vertical component of the Earth's rotation vector. In the conventional notation, $\dot{\lambda}$ is replaced by $\Omega$.

## B.7 Formulas that are useful for two-dimensional flow

Consider the special case of two-dimensional flow. Two useful identities are

$$\nabla_r \times (\mathbf{e}_r \times \nabla_r A) = \mathbf{e}_r \nabla_r^2 A, \tag{B.53}$$

and

$$\nabla_r \cdot (\mathbf{e}_r \times \nabla_r A) = 0. \tag{B.54}$$

Also for two-dimensional flow, the Laplacian of a vector can be written in a very simple way. Let $\zeta \mathbf{e}_r \equiv \nabla_r \times \mathbf{V}_h$ and $\delta \equiv \nabla_r \cdot \mathbf{V}_h$. Then (B.22) reduces to

$$\nabla_r^2 \mathbf{V}_h = \nabla_r \delta - \nabla_r \times (\zeta \mathbf{e}_r) \tag{B.55}$$

Using (B.28), we can write

$$\begin{aligned}\nabla_r \times (\zeta \mathbf{e}_r) &= \left\{ \frac{1}{r}\frac{\partial \zeta}{\partial \varphi},\ -\frac{1}{r\cos\varphi}\frac{\partial \zeta}{\partial \lambda},\ 0 \right\} \\ &= -\,\mathbf{e}_r \times \nabla_r \zeta\ .\end{aligned} \tag{B.56}$$

Then (B.55) becomes

$$\nabla_r^2 \mathbf{V}_h = \nabla_r \delta + \mathbf{e}_r \times \nabla_r \zeta. \tag{B.57}$$

## B.8 Basics of 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 B.3. Our goal is to relate $\left(\frac{\partial f}{\partial x}\right)_{\hat{z}}$ to $\left(\frac{\partial f}{\partial x}\right)_z$. With reference to Figure B.3, we can write

$$\begin{aligned}\frac{f_B - f_A}{x_2 - x_1} &= \left(\frac{f_C - f_A}{x_2 - x_1}\right) - \left(\frac{f_C - f_B}{x_2 - x_1}\right) \\ &= \left(\frac{f_C - f_A}{x_2 - x_1}\right) - \left(\frac{z_2 - z_1}{x_2 - x_1}\right)\left(\frac{f_C - f_B}{z_2 - z_1}\right).\end{aligned} \tag{B.58}$$

Taking the limit as the increments become small, we obtain

$$\left(\frac{\partial f}{\partial x}\right)_z = \left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial z}{\partial x}\right)_{\hat{z}}\left(\frac{\partial f}{\partial z}\right)_x. \tag{B.59}$$

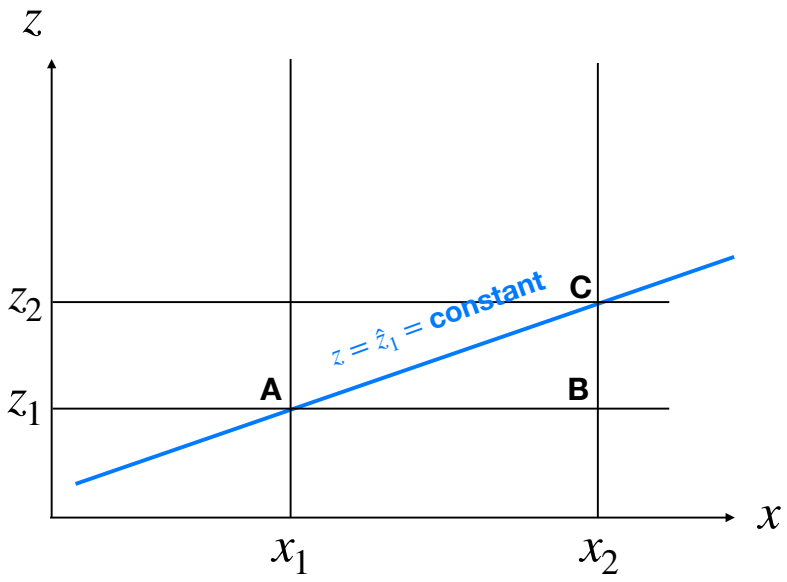



Figure B.3: A sketch used in the derivation of (B.58).

Since we are using $\hat{z}$ as our vertical coordinate, a better way to write (B.59) is

$$\boxed{\left(\frac{\partial f}{\partial x}\right)_z = \left(\frac{\partial f}{\partial x}\right)_{\hat{z}} - \left(\frac{\partial z}{\partial x}\right)_{\hat{z}} \frac{\partial \hat{z}}{\partial z} \left(\frac{\partial f}{\partial \hat{z}}\right)_x .} \tag{B.60}$$

The quantity $\dfrac{\partial \hat{z}}{\partial z}$ encodes the variations of $\hat{z}$ with $z$. It plays the role of a "metric" factor that converts from the dimensions of $\hat{z}$ to the dimension of length, in the same way that the metric factor $a\cos\varphi$ converts the nondimensional coordinate of longitude to a length.

Naturally, the derivation given above works in exactly the same way if $x$ is time, rather than a horizontal coordinate.

## B.9 Some useful operators

Starting from (B.60), we can show that the horizontal gradient satisfies

$$\nabla_z f = \nabla_{\hat{z}} f - \nabla_{\hat{z}} z \frac{\partial \hat{z}}{\partial z} \left(\frac{\partial f}{\partial \hat{z}}\right) . \tag{B.61}$$

Analogous identities apply with other operators. For example, for an arbitrary horizontal vector $\mathbf{V}_h$, we can write

$$\nabla_z \times \mathbf{V}_h = \nabla_{\hat{z}} \times \mathbf{V}_h - \nabla_{\hat{z}} z \times \frac{\partial \hat{z}}{\partial z} \frac{\partial \mathbf{V}_h}{\partial \hat{z}}. \tag{B.62}$$

and

$$\nabla_z \cdot \mathbf{V}_h = \nabla_{\hat{z}} \cdot \mathbf{V}_h - \nabla_{\hat{z}} z \cdot \frac{\partial \hat{z}}{\partial z} \frac{\partial \mathbf{V}_h}{\partial \hat{z}}. \tag{B.63}$$

## B.10 Concluding summary

Scalars, vectors, and tensors have meanings without reference to any particular coordinate system, although they can be expressed in 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.

# Bibliography


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.

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.

Diamantakis, M., and J. Flemming, 2014: Global mass fixer algorithms for conservative tracer transport in the ecmwf model. *Geoscientific Model Development*, **7 (3)**, 965–979.

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.

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.

Eliassen, A., 1956: *AA Procedure for Numerical Integration of the Primitive Equations of the Two-parameter Model of the Atmosphere*. University of California.

Eliassen, A., and E. Raustein, 1968: A numerical integration experiment with a model atmosphere based on isentropic surfaces. *Meteorologiske Annaler*, **5**, 45–63.

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.

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.

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., 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.

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.

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.

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.

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.

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.

Kurihara, Y., 1965: Numerical integration of the primitive equations on a spherical grid. *Mon. Wea. Rev*, **93 (7)**, 399–415.

Laprise, R., 1992: 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.

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.

McGregor, J. L., 1996: Semi-lagrangian advection on conformal-cubic grids. *Monthly weather review*, **124 (6)**, 1311–1322.

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., and Z. I. Janjic, 1985: Problems and numerical methods of the incorporation of mountains in atmospheric models. *Lectures in Applied Mathematics*, **22**, 81–120.

Monaghan, J. J., 1992: Smoothed particle hydrodynamics. *Annual review of Astronomy and Astrophysics*, **30**, 543–574.

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.

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.

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, URL https://doi.org/10.1175/1520-0493(1959)087<0333:NIOTPE>2.0.CO;2, https://doi.org/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.

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.

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.

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.

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.

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.

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.

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/.

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.

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.

Staniforth, A., and J. Côté, 1991: Semi-lagrangian integration schemes for atmospheric models-a review. *Monthly weather review*, **119 (9)**, 2206–2223.

Strang, G., 2007: *Computational science and engineering*, Vol. 1. Wellesley-Cambridge Press Wellesley.

Suarez, M. J., A. Arakawa, and D. A. Randall, 1983: The parameterization of the planetary boundary layer in the ucla general circulation model: Formulation and results. *Monthly weather review*, **111 (11)**, 2224–2243.

Swinbank, R., and R. James 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.

Takacs, L. L., 1988: Effects of using a posteriori methods for the conservation of integral invariants. *Monthly weather review*, **116 (3)**, 525–545.

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.

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 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.

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.

Zhu, Z., J. Thuburn, B. J. Hoskins, and P. H. Haynes, 1992: A vertical finite-difference scheme based on a hybrid $\sigma$-$\theta$-p coordinate. *Monthly Weather Review*, **120 (5)**, 851–862.