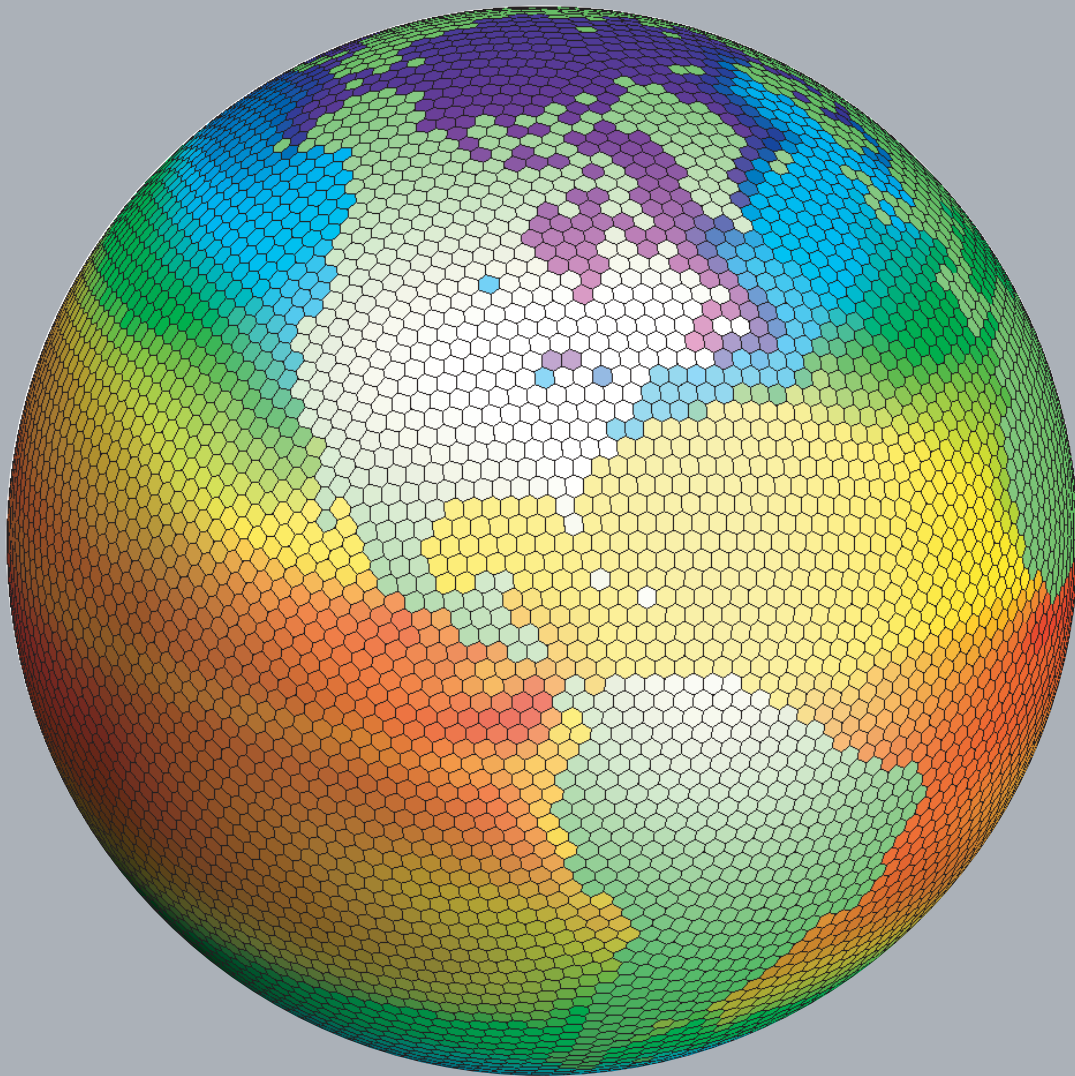


An Introduction to Atmospheric Modeling

Instructor: D. Randall

**AT604
Department of Atmospheric Science
Colorado State University**

Fall, 2004



Announcements

| | |
|------------------------------|--|
| <i>Subject:</i> | A practical introduction to numerical modeling of the atmosphere. |
| <i>Text:</i> | Class notes, available at the class website: http://kiwi.atmos.colostate.edu/group/dave/at604.html |
| <i>Course grade:</i> | 1/4 on homework, 1/4 on each of two midterms (closed book, in class), and 1/4 on final (closed book, in class) The final will emphasize the latter part of the course, and will be held during finals week. |
| <i>Access to instructor:</i> | As you may know, I have posted office hours, but students in this class are welcome to come to me with questions any time, provided only that I am not actually busy with someone else. |
| <i>Teaching assistant:</i> | We are fortunate to have Jonathan Vigh as a TA for this course. He will grade the homework and will be available to answer questions on a schedule which he will make known to you. He may also organized other activities, which will be announced separately. |
| <i>Computing:</i> | Some of the homework will involve writing computer programs, plotting results, etc. You can use any computing language or plotting software you want. Although you are certainly encouraged to ask questions about the homework, neither I nor the TA will help with debugging your programs. |
| <i>Auditing:</i> | Auditing is permitted, provided that you audit officially by filling out the appropriate form. Auditors are required to attend class but are not required to hand in homeworks or take exams. Keep in mind, however, that, like skiing or swimming or bicycling, numerical modeling is learned largely by doing. |
| <i>Schedule:</i> | Classes will be missed occasionally. A calendar will be distributed. |

General References

- Arakawa, A., 1988: Finite-difference methods in climate modeling. *Physically-based modelling and simulation of climate and climatic change - Part I*, M. E. Schlesinger (ed.), 79-168.
- Arfken, G., 1985: *Mathematical methods for physicists*. Academic Press, 985 pp.
- Chang, J., 1977: General circulation models of the atmosphere. *Meth. Comp. Phys.*, **17**, Academic Press, 337 pp.
- Durran, D. R., 1999: *Numerical methods for wave equations in geophysical fluid dynamics*. Springer, 465 pp.
- Haltiner, G. J., and R. T. Williams, 1980: *Numerical prediction and dynamic meteorology*. J. Wiley and Sons, 477 pp.
- Kalnay, E., 2003: *Atmospheric modeling, data assimilation, and predictability*. Cambridge Univ. Press, 341 pp.
- Manabe, S., ed., 1985: Issues in atmospheric and oceanic modeling, Part A: Climate dynamics. *Adv. in Geophys.*, **28**, 591 pp.
- Manabe, S., ed., 1985: Issues in atmospheric and oceanic modeling, Part B: Weather dynamics. *Adv. in Geophys.*, **28**, 432 pp.
- Mesinger, F., and A. Arakawa, 1976: Numerical methods used in atmospheric models. *GARP Publ. Ser. No. 17*, 64 pp.
- Randall, D. A., Ed., 2000: *General Circulation Model Development. Past, Present, and Future*. Academic Press, 807 pp.
- Richtmeyer, R. D., and K. W. Morton, 1967: *Difference methods for initial value problems*. Wiley Interscience Publishers, New York, 405 pp.
- Washington, W. M., and C. L. Parkinson, 1986: *An introduction to three-dimensional climate modeling*. University Science Books, Mill Valley, New York, 422 pp.

Preface

The purpose of this course is to provide an introduction to the methods used in numerical modeling of the atmosphere. The ideas presented are relevant to both large-scale and small-scale models.

Numerical modeling is one of several 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 in most cases no closed form solution is possible. 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 course is to make it less likely that you, the students, will use a model without having any understanding of it.

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 that they may encounter out there in the world.

The first version of these notes, put together in 1991, was heavily based on the class notes developed by Prof. A. Arakawa at UCLA, as they existed in the early 1970s, and this influence is still apparent in the current version, particularly in Chapters 2 and 3. A lot of additional material has been incorporated, mainly reflecting developments in the field since the 1970s. The explanations and problems have also been considerably revised and updated.

The teaching assistants for this 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 by extending and refining these notes, and also through questions and feedback from the students. The course has certainly benefitted

considerably from such student input.

Finally, Michelle McDaniel has spent countless hours patiently assisting in the production of these notes. She created the formatting that you see, and organized the notes into a “book.”

| | |
|--|------------------|
| Preliminaries | <i>i</i> |
| CHAPTER 1 <i>Introduction</i> | <i>1</i> |
| What is a model? | 1 |
| Fundamental physics, mathematical methods, and physical parameterizations | 3 |
| Numerical experimentation | 5 |
| CHAPTER 2 <i>Basic Concepts</i> | <i>7</i> |
| Finite-difference quotients | 7 |
| Difference quotients of higher accuracy | 11 |
| Extension to two dimensions | 18 |
| An example of a finite difference-approximation to a differential equation | 21 |
| Accuracy and truncation error of a finite-difference scheme. | 24 |
| Discretization error and convergence | 25 |
| Interpolation and extrapolation | 28 |
| Stability | 29 |
| The effects of increasing the number of grid points | 38 |
| Summary | 39 |
| Problems | 42 |
| CHAPTER 3 <i>A Survey of Time-Differencing Schemes for the Oscillation and Decay Equations</i> | <i>43</i> |
| Introduction | 43 |
| Non-iterative schemes. | 43 |
| Explicit schemes () | 47 |
| Implicit schemes | 49 |
| Iterative schemes | 51 |
| Finite-difference schemes applied to the oscillation equation | 52 |
| Non-iterative two-level schemes for the oscillation equation | 54 |
| Iterative two-level schemes for the oscillation equation | 57 |
| The leapfrog scheme for the oscillation equation | 58 |
| The second-order Adams Bashforth Scheme (m=0, l=1) for the oscillation equation | 67 |
| A survey of time differencing schemes for the oscillation equation | 68 |
| Finite-difference schemes for the decay equation | 69 |
| Damped oscillations | 72 |
| Nonlinear damping | 72 |

| | |
|---|----|
| Summary | 77 |
| A Proof that the Fourth-Order Runge-Kutta Scheme has Fourth-Order Accuracy | 78 |
| Problems | 83 |

CHAPTER 4 ***A closer look at the advection equation*** **85**

| | |
|--|-----|
| Introduction | 85 |
| Conservative finite-difference methods | 88 |
| Examples of schemes with centered space differencing | 93 |
| Computational dispersion | 100 |
| The effect s of fourth-order space differencing on the phase speed | 107 |
| Space-uncentered schemes | 108 |
| Hole filling | 112 |
| Flux-corrected transport | 113 |
| Lagrangian schemes | 116 |
| Semi-Lagrangian schemes | 118 |
| Two-dimensional advection | 120 |
| Summary | 123 |
| Problems | 123 |

CHAPTER 5 ***Boundary-value problems*** **127**

| | |
|---|-----|
| Introduction | 127 |
| Solution of one-dimensional boundary-value problems | 128 |
| Jacobi relaxation | 130 |
| Gauss-Seidel relaxation | 133 |
| Over-relaxation | 134 |
| The alternating-direction implicit method | 135 |
| Multigrid methods | 135 |
| Summary | 136 |

CHAPTER 6 ***Diffusion*** **141**

| | |
|---------------------------------|-----|
| Introduction | 141 |
| A simple explicit scheme | 143 |
| An implicit scheme | 144 |
| The DuFort-Frankel scheme | 146 |
| Summary | 147 |

| | |
|----------------|-----|
| Problems | 148 |
|----------------|-----|

| | | |
|------------------|----------------------------|------------|
| CHAPTER 7 | <i>Making Waves</i> | 149 |
|------------------|----------------------------|------------|

| | |
|---|-----|
| The shallow-water equations | 149 |
| The wave equation | 150 |
| Staggered grids | 152 |
| Numerical simulation of geostrophic adjustment, as a guide to grid design | 154 |
| Time-differencing schemes for the shallow-water equations | 160 |
| Summary and conclusions | 167 |
| Problems | 168 |

| | | |
|------------------|---|------------|
| CHAPTER 8 | <i>Schemes for the one-dimensional nonlinear shallow-water equations</i> | 169 |
|------------------|---|------------|

| | |
|--|-----|
| Properties of the continuous equations | 169 |
| Space differencing | 171 |
| Summary | 178 |
| Problems | 180 |

| | | |
|------------------|---|------------|
| CHAPTER 9 | <i>Vertical Differencing for Quasi-Static Models</i> | 183 |
|------------------|---|------------|

| | |
|---|-----|
| Introduction | 183 |
| Choice of equation set | 183 |
| General vertical coordinate | 184 |
| The equation of motion and the HPGF | 188 |
| Vertical mass flux for a family of vertical coordinates | 189 |
| Discussion of particular vertical coordinate systems | 191 |
| Height | 192 |
| Pressure | 196 |
| Log-pressure | 197 |
| The σ -coordinate | 197 |
| More on the HPGF in σ -coordinates | 200 |
| Hybrid sigma-pressure coordinates | 201 |
| The η -coordinate | 202 |
| Potential temperature | 203 |
| Entropy | 206 |
| Hybrid σ - θ coordinates | 206 |
| Summary of vertical coordinate systems | 206 |
| Vertical staggering | 208 |
| Conservation properties of vertically discrete models using σ -coordinates | 210 |

| | |
|-------------------------------|-----|
| Summary and conclusions | 221 |
|-------------------------------|-----|

| | | |
|-------------------|------------------------------------|------------|
| CHAPTER 10 | <i>Aliasing instability</i> | 223 |
|-------------------|------------------------------------|------------|

| | |
|---|-----|
| Aliasing error | 223 |
| Advection by a variable, non-divergent current | 227 |
| Fjortoft's Theorem | 236 |
| Kinetic energy and enstrophy conservation in two-dimensional non-divergent flow | 241 |
| Angular momentum conservation | 251 |
| Conservative schemes for the two-dimensional shallow water equations with rotation | 252 |
| The effects of time differencing on energy conservation | 257 |
| Summary | 259 |
| Problems | 260 |

| | | |
|-------------------|--|------------|
| CHAPTER 11 | <i>Finite Differences on the Sphere</i> | 261 |
|-------------------|--|------------|

| | |
|---|-----|
| Introduction | 261 |
| Coordinate systems and map projections | 262 |
| Latitude-longitude grids and the "pole problem" | 267 |
| Kurihara's grid | 273 |
| The Wandering Electron Grid | 274 |
| Spherical geodesic grids | 274 |
| Summary | 280 |

| | | |
|-------------------|--------------------------------|------------|
| CHAPTER 12 | <i>Spectral Methods</i> | 281 |
|-------------------|--------------------------------|------------|

| | |
|---|-----|
| Introduction | 281 |
| Spectral methods on the sphere | 289 |
| The "equivalent grid resolution" of spectral models | 294 |
| Semi-implicit time differencing | 295 |
| Conservation properties and computational stability | 296 |
| Moisture advection | 296 |
| Physical parameterizations | 297 |
| Summary | 297 |
| Problems | 299 |

CHAPTER 13 *Boundary conditions and nested grids* 301

| | |
|---|-----|
| Introduction | 301 |
| Inflow Boundaries | 301 |
| Outflow boundaries | 308 |
| Advection on nested grids | 314 |
| Analysis of boundary conditions for the advection equation using the energy method | 320 |
| Physical and computational reflection of gravity waves at a wall | 323 |
| Boundary conditions for the gravity wave equations with an advection term | 326 |
| The energy method as a guide in choosing boundary conditions for gravity waves | 327 |
| Summary | 330 |
| Problem | 330 |

***References and Bibliography* 341**

CHAPTER 1

Introduction

Copyright 2004 David A. Randall

1.1 What is a model?

The atmospheric science community includes a large and energetic group of researchers who devise and carry out measurements in the atmosphere. This work involves 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. This might be a qualitative conceptual model, or it might be an analytical theory, or it might take the form of a computer program.

Accordingly, a community of modelers is hard at work developing models, performing calculations, and analyzing the results by comparison with data. 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.

A model essentially embodies a theory. A model (or a theory) provides a basis for making predictions about the outcomes of measurements. The disciplines of fluid dynamics, radiative transfer, atmospheric chemistry, and cloud microphysics all make use of models that are essentially direct applications of basic physical principles to phenomena that occur in the atmosphere. Many of these “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 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.

A model that predicts the deterministic evolution of the atmosphere or some macroscopic portion of it can be called a “forecast model.” A forecast model could be, as the name suggests, a model that is used to conduct weather prediction, but there are other possibilities, e.g. it could be used to predict the deterministic evolution of an individual turbulent eddy. Forecast models can be tested against real data, documenting for example the observed development of a synoptic-scale system or the observed growth of an individual convective cloud, assuming of course that the requisite data can be collected.

We are often interested in computing the *statistics* of an atmospheric phenomenon, e.g. the statistics of the general circulation. It is now widely known that there are fundamental

limits on the deterministic predictability of the atmosphere, due to sensitive dependence on initial conditions (e.g. Lorenz, 1969). For the global-scale circulation of the atmosphere, the limit of predictability is thought to be on the order of a few weeks, but for a cumulus-scale circulation it is on the order of a few minutes. For time scales longer than the deterministic limit of predictability for the system in question, only the statistics of the system can be predicted. These statistics can be generated by brute-force simulation, using a forecast model but pushing the forecast beyond the deterministic limit, and then computing statistics from the results. The obvious and most familiar example is simulation of the atmospheric general circulation (e.g. Smagorinski 1963). Additional examples are large eddy simulations of atmospheric turbulence (e.g. Moeng 1984), and simulations of the evolution of an ensemble of clouds using a space and time domains much larger than the space and time scales of individual clouds (e.g. Krueger 1988).

Forecast models are now also being used to make predictions of the time evolution of the *statistics* of the weather, far beyond the limit of deterministic predictability for individual weather systems. Examples are seasonal weather forecasts, which deal with the statistics of the weather rather than day-to-day variations of the weather and are now being produced by several operational centers; and climate change forecasts, which deal with the evolution of the climate over the coming decades and longer. In the case of seasonal forecasting, the predictability of the statistics of the atmospheric circulation beyond the two-week deterministic limit arises primarily from the predictability of the sea surface temperature, which has a much longer memory of its initial conditions than does the atmosphere.

In the case of climate change predictions, the time evolution of the statistics of the climate system are predictable to the extent that they are driven by predictable changes in some external forcing. For example, projected increases in greenhouse gas concentrations represent a time-varying external forcing whose effects on the time evolution of the statistics of the climate system may be predictable. Over the next several decades measurements will make it very clear to what extent these predictions are right or wrong. A more mundane example is the seasonal cycle of the atmospheric circulation, which represents the response of the statistics of the atmospheric general circulation to the movement of the Earth in its orbit; because the seasonal forcing is predictable many years in advance, the seasonal cycle of the statistics of the atmospheric circulation is also highly predictable, far beyond the two-week limit of deterministic predictability for individual weather systems.

Some models predict statistics directly; the dependent variables are the statistics themselves, and there is no need to average the model results to generate statistics after the fact. For example, radiative transfer models describe the statistical behavior of extremely large numbers of photons. “Higher-order closure models” have been developed to simulate directly the statistics of small-scale atmospheric turbulence (e.g., Mellor and Yamada, 1974). Analogous models for direct simulation of the statistics of the large-scale circulation of the atmosphere may be possible (e.g., Green, 1970).

Finally, we also build highly idealized models that are not intended to provide quantitatively accurate or physically complete descriptions of natural phenomena, but rather to encapsulate our physical understanding of a complex phenomenon in the simplest and most compact possible form, as a kind of modeler’s haiku. For example, North (1975) discusses the application of this approach to climate modeling. Toy models are intended primarily as educational tools; the results that they produce can be compared with measurements only in qualitative or semi-quantitative ways.

This course deals with numerical methods that can be used with any of the model “types” discussed above, but for the most part we will be thinking of “forecast models.”

1.2 Fundamental physics, mathematical methods, and physical parameterizations

Most models in atmospheric science are formulated by starting from basic physical principles, such as conservation of mass, conservation of momentum, conservation of thermodynamic energy, and the radiative transfer equation. In principle, these equations can describe the evolution of the atmosphere in extreme detail, down to spatial and temporal scales far below the range of meteorological interest.

Even if such detailed models were technologically feasible, we would still choose to average or aggregate the output produced by the models so as to depict the evolution of the scales of primary interest, e.g. thunderstorms, tropical cyclones, baroclinic waves, and the global circulation. In addition, we would want to *explain* why the solutions of the models turn out as they do. In practice, of course, we cannot use such high spatial and temporal resolution, and so we must represent some important processes parametrically. Such parametric representations, or “parameterizations”, are a practical necessity in models of limited resolution, but even if we were using models with arbitrarily high resolution we would still need parameterizations to understand what the model results mean. Parameterizations are not dealt with in this course, but you can learn about them in courses on cloud physics, radiation, turbulence, and chemistry.

Obviously, mathematical methods are needed to solve the equations of a model, and in practice the methods are almost always approximate, which means that they entail errors. It is useful to distinguish between physical errors and mathematical errors. Suppose that we have 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 as applied to the global circulation of the atmosphere. We simplify the problem by making physical approximations to the equations. For example, we may treat the motion as quasi-static, or we may introduce Reynolds averaging along with closures that can be used to determine turbulent and convective fluxes. In the course of making these physical approximations, we do two important things:

- We introduce errors. That is why the physical approximations are called “approximations.”
- We change the physical model. After making the physical approximations, we no longer have the Navier-Stokes equations.

Unfortunately, even after the various physical approximations have been made, it is still (usually) impossible for us to obtain exact solutions to the modified model. We therefore introduce further approximations that are purely mathematical (rather than physical) in character. For example, we may replace derivatives by finite differences. Solutions of the resulting models take the form of numbers, rather than formulas, so the models are described as “numerical.”

In this course, we deal primarily with the mathematical approximations that are used to convert (already approximate) physical models into numerical models. We focus on the errors involved and how they can be anticipated, analyzed, and minimized. This is a course about errors. All your life you have been making errors. Now, finally, you get to take a course on errors. It’s about time.

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

Having spent a page or so making the distinction between physical errors and mathematical errors, I will now 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 course will emphasize 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, which can be written as

$$\frac{\partial \rho}{\partial t} = -\nabla \bullet (\rho \mathbf{V}), \quad (1.1)$$

where ρ is the density of dry air, and \mathbf{V} is the three-dimensional velocity vector. When (1.1) is integrated over the whole atmosphere, with appropriate boundary conditions, we find that

$$\int_{\text{whole atmosphere}} \nabla \bullet (\rho \mathbf{V}) d^3 \mathbf{x} = 0, \quad (1.2)$$

and so we conclude that

$$\frac{d}{dt} \left\{ \int_{\text{whole atmosphere}} \rho d^3 \mathbf{x} \right\} = 0. \quad (1.3)$$

Eq. (1.3) is a statement of global mass conservation; in order to obtain (1.3), we had to use (1.2), which is a property of the divergence operator with the appropriate boundary conditions.

In a numerical model, we replace (1.1) by an approximation; examples are given later. The approximate form of (1.1) entails an approximation to the divergence operator. These approximations inevitably involve 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. In particular, we refuse to accept any error in the global conservation of mass. This means that we can design our model so that an appropriate analog of (1.3) is satisfied *exactly*.

In order to derive an analog of (1.3), we have to enforce an analog of (1.2); 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 representing the global integral) is exactly zero. This can be done, quite easily. You would 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. These include conservation of energy and conservation of potential vorticity. More

discussion is given later.

1.3 Numerical experimentation

A serious difficulty in the geophysical sciences such as atmospheric science is that it is usually impossible (perhaps fortunately) to perform controlled experiments using the Earth. 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. For a long time, 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. These are intended to be substitutes for the actual atmosphere and, therefore, include representations of many physical processes. Direct comparisons with observations must be made for evaluation of the model results. Unfortunately (or perhaps fortunately), the design of such models can never be a purely mathematical problem. In practice, the models include many simplifications and parameterizations, but still they have to be realistic. To meet this requirement, 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. Once we have gained sufficient confidence that a model is reasonably realistic, it can be used 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. A model can also be used as a purely experimental tool. Predictability experiments are examples.

Simpler numerical models are also very useful for studying individual phenomena, insofar as these phenomena 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 dynamics. Theories, numerical simulations and observational studies have been developed jointly in the last several decades, and this will continue indefinitely. Observational and theoretical studies guide the design of numerical models, and numerical simulations supply theoretical ideas and suggest efficient observational systems.

We do not attempt, in this course, 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 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. These 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. Emphasis is placed on time-dependent equations, but we also briefly discuss boundary-value problems. 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.

CHAPTER 2**Basic Concepts**

Copyright 2004 David A. Randall

2.1 Finite-difference quotients

Consider the derivative $\frac{du}{dx}$, where $u = u(x)$, and x is the independent variable (which could be either space or time). Finite-difference methods represent the continuous function $u(x)$ by a set of values defined at a number of discrete points in a specified region. Thus, we usually introduce a “grid” with discrete points at which the variable u is carried, as shown in Fig. 2.1. Sometimes the words “mesh” or “lattice” are used in place of

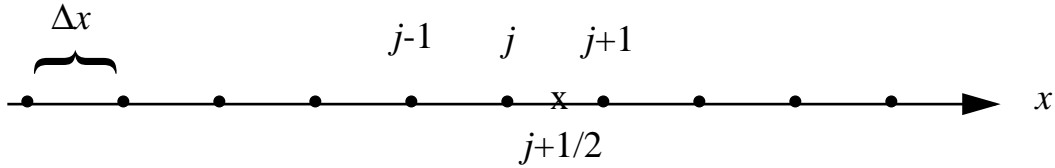


Figure 2.1: An example of a grid, with uniform grid spacing Δx . The grid points are denoted by the integer index j . Half-integer points can also be defined.

“grid.” The interval Δx is called the grid interval, grid size, mesh size, etc. We assume that the grid interval is constant for the time being; then $x_j = j\Delta x$, where j is the “index” used to identify the grid points. Note that u is defined only at the grid points denoted by the integers j , $j + 1$, etc.

Using the notation $u_j = u(x_j) = u(j\Delta x)$, we can define the *forward difference* at the point j by

$$(\Delta u)_j \equiv u_{j+1} - u_j, \quad (2.1)$$

the *backward difference* at the point j by

$$(\nabla u)_j \equiv u_j - u_{j-1}, \quad (2.2)$$

and the *centered difference* at the point $j + \frac{1}{2}$ by

$$(\delta u)_{j+\frac{1}{2}} \equiv u_{j+1} - u_j. \quad (2.3)$$

Note that u 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 :

$$\left(\frac{du}{dx}\right)_j \equiv \frac{u_{j+1} - u_j}{\Delta x}; \quad (2.4)$$

the backward difference quotient at the point j :

$$\left(\frac{du}{dx}\right)_j \equiv \frac{u_j - u_{j-1}}{\Delta x}; \quad (2.5)$$

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

$$\left(\frac{du}{dx}\right)_{j+\frac{1}{2}} \equiv \frac{u_{j+1} - u_j}{\Delta x} = \frac{(\delta u)_{j+\frac{1}{2}}}{\Delta x}. \quad (2.6)$$

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

$$\left(\frac{du}{dx}\right)_j \equiv \frac{u_{j+1} - u_{j-1}}{2\Delta x} = \frac{1}{\Delta x} \frac{1}{2} [(\delta u)_{j+\frac{1}{2}} + (\delta u)_{j-\frac{1}{2}}]. \quad (2.7)$$

Since (2.4) and (2.5) employ the values of u at two points, they are sometimes referred to as two-point approximations. On the other hand, (2.6) and (2.7) are three-point approximations. 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? We now introduce the concepts of accuracy and truncation error. As an example, consider the forward difference quotient

$$\left(\frac{du}{dx}\right)_j \cong \frac{u_{j+1} - u_j}{\Delta x} = \frac{u[(j+1)\Delta x] - u(j\Delta x)}{\Delta x}, \quad (2.8)$$

and expand u in a Taylor series about the point x_j to obtain

$$u_{j+1} = u_j + \Delta x \left(\frac{du}{dx}\right)_j + \frac{(\Delta x)^2}{2!} \left(\frac{d^2 u}{dx^2}\right)_j + \frac{(\Delta x)^3}{3!} \left(\frac{d^3 u}{dx^3}\right)_j + \dots + \frac{(\Delta x)^{n-1}}{(n-1)!} \left(\frac{d^{n-1} u}{dx^{n-1}}\right)_j + \dots, \quad (2.9)$$

which can be rearranged to

$$\frac{u_{j+1} - u_j}{\Delta x} = \left(\frac{du}{dx}\right)_j + \varepsilon, \quad (2.10)$$

where

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

is the error. The expansion (2.9) can be derived without any assumptions or approximations except that the indicated derivatives exist (Arfken, 1985; for a quick review see the Appendix on Taylor's Series). The terms in (2.10) that are lumped into ε are called the "truncation error." *The lowest power of Δx that appears in the truncation error is called the order of accuracy of the corresponding difference quotient.* For example, the leading term of (2.11) is of order Δx or $O(\Delta x)$, 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.

Expansion of (2.6) and (2.7) similarly shows that they are of second-order accuracy. We can write

$$u_{j-1} = u_j + \left(\frac{du}{dx}\right)_j (-\Delta x) + \left(\frac{d^2 u}{dx^2}\right)_j \frac{(-\Delta x)^2}{2!} + \left(\frac{d^3 u}{dx^3}\right)_j \frac{(-\Delta x)^3}{3!} + \dots \quad (2.12)$$

Subtracting (2.12) from (2.9) gives

$$u_{j+1} - u_{j-1} = 2 \left(\frac{du}{dx}\right)_j (\Delta x) + \frac{2}{3!} \left(\frac{d^3 u}{dx^3}\right)_j (\Delta x)^3 + \dots \text{ odd powers only}, \quad (2.13)$$

or

$$\left(\frac{du}{dx}\right)_j = \frac{u_{j+1} - u_{j-1}}{2\Delta x} - \left(\frac{d^3u}{dx^3}\right)_j \frac{(\Delta x)^2}{3!} + O[(\Delta x)^4]. \quad (2.14)$$

Similarly,

$$\left(\frac{du}{dx}\right)_{j+\frac{1}{2}} = \frac{u_{j+1} - u_j}{\Delta x} - \left(\frac{d^3u}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!} + O[(\Delta x)^4]. \quad (2.15)$$

In this way, we find that

$$\frac{\text{Error of } \left(\frac{du}{dx}\right)_j}{\text{Error of } \left(\frac{du}{dx}\right)_{j+\frac{1}{2}}} \cong \frac{\left(\frac{d^3u}{dx^3}\right)_j \frac{(\Delta x)^2}{3!}}{\left(\frac{d^3u}{dx^3}\right)_{j+\frac{1}{2}} \frac{(\Delta x/2)^2}{3!}} = \frac{4\left(\frac{d^3u}{dx^3}\right)_j}{\left(\frac{d^3u}{dx^3}\right)_{j+\frac{1}{2}}} \cong 4. \quad (2.16)$$

In other words, the error of $\left(\frac{du}{dx}\right)_j$ is four times as large as the error of $\left(\frac{du}{dx}\right)_{j+\frac{1}{2}}$, even

though both finite-difference quotients have second-order accuracy. This makes the point 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 term of the error has the form

$$\varepsilon \cong C(\Delta x)^p. \quad (2.17)$$

Then $\ln(\varepsilon) \cong p \ln(\Delta x) + \ln(C)$, and so

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

The implication is 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 it is necessary to compute the error of the finite-difference approximation, and that can only be done when the exact derivative is known. Therefore, this approach is usually implemented by using an analytical “test function.”

2.2 Difference quotients of higher accuracy

Suppose that we write

$$\frac{u_{j+2} - u_{j-2}}{4\Delta x} = \left(\frac{du}{dx}\right)_j + \frac{1}{3!} \left(\frac{d^3u}{dx^3}\right)_j (2\Delta x)^2 + \dots \text{ even powers only.} \quad (2.19)$$

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 Δx 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 $\left(\frac{du}{dx}\right)_j$:

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

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

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

With this choice, (2.20) reduces to

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

We have thus obtained a fourth-order scheme. In effect, this is a linear extrapolation of the value of the finite-difference expression to a smaller grid size, as schematically illustrated in Fig. 2.2.

Is there a more systematic way to construct schemes of any desired order of accuracy? The answer is “yes,” and one such approach is as follows.

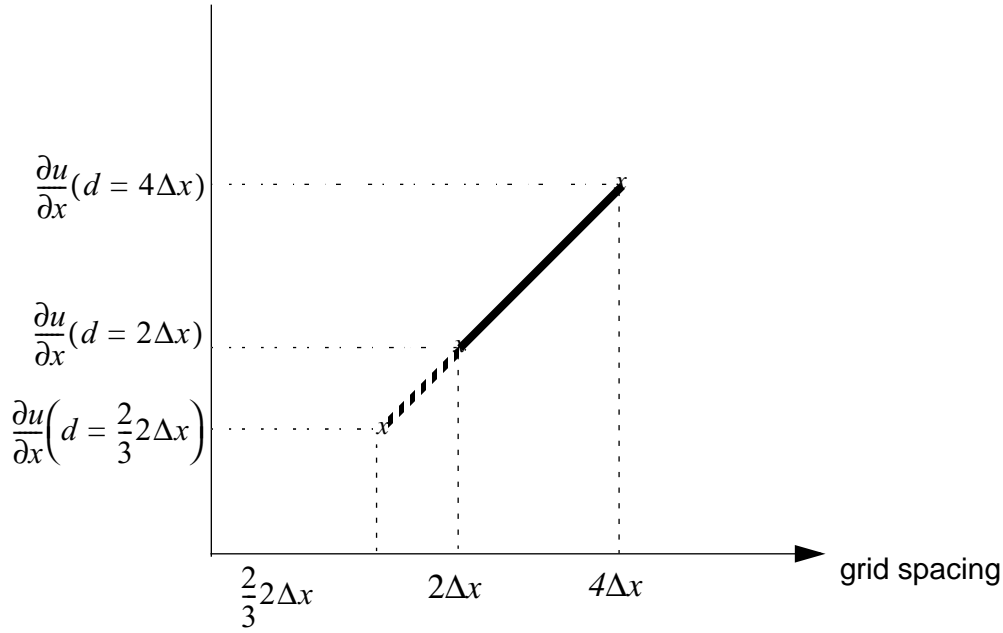


Figure 2.2: Schematic illustrating the interpretation of the fourth-order difference in terms of the extrapolation of the second-order difference based on a spacing of $4\Delta x$, and that based on a spacing of $2\Delta x$. The extrapolation reduces the effective grid size to $(2/3)2\Delta x$.

Suppose that we write a finite-difference approximation to $\left(\frac{df}{dx}\right)_j$ in the following somewhat generalized form:

$$\left(\frac{df}{dx}\right)_j \cong \frac{1}{\Delta x} \sum_{i=-\infty}^{\infty} a_i f(x_j + i\Delta x) . \quad (2.23)$$

Here the a_i are coefficients or “weights,” which are undetermined at this point. In most schemes, all but a few of the a_i will be zero, so that the sum in (2.23) will actually involve only a few non-zero terms. In writing (2.23), we have assumed for simplicity that Δx is a constant; this assumption can be relaxed, as will be shown below. The index i in (2.23) is a counter that is zero at our “home base” at grid point j . For $i < 0$ we count to the left, and for $i > 0$ we count to the right. According to (2.23), our finite-difference approximation to $\left(\frac{df}{dx}\right)_j$ has the form of a weighted sum of values of f 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

$$f(x_j + i\Delta x) = f_j^0 + f_j^1(i\Delta x) + f_j^2 \frac{(i\Delta x)^2}{2!} + f_j^3 \frac{(i\Delta x)^3}{3!} + \dots \quad (2.24)$$

Here f_j^n is the n th derivative of f , evaluated at the point j . Using (2.24), we can rewrite (2.23) as

$$\left(\frac{df}{dx}\right)_j \cong \frac{1}{\Delta x} \sum_{i=-\infty}^{\infty} a_i \left[f_j^0 + f_j^1(i\Delta x) + f_j^2 \frac{(i\Delta x)^2}{2!} + f_j^3 \frac{(i\Delta x)^3}{3!} + \dots \right]. \quad (2.25)$$

By inspection of (2.25), we see that in order to have at least first-order accuracy, we need

$$\sum_{i=-\infty}^{\infty} a_i = 0 \text{ and } \sum_{i=-\infty}^{\infty} i a_i = 1. \quad (2.26)$$

To have at least second-order accuracy, we must impose an additional requirement:

$$\sum_{i=-\infty}^{\infty} i^2 a_i = 0. \quad (2.27)$$

In general, to have at least n th-order accuracy, we must require that

$$\sum_{i=-\infty}^{\infty} i^m a_i = \delta_{m,1} \text{ for } 0 \leq m \leq n. \quad (2.28)$$

Here $\delta_{m,1}$ is the Kronecker delta. In order to satisfy (2.28), we must solve a system of $n+1$ linear equations for the $n+1$ unknown coefficients a_i .

According to (2.28), a scheme of n th-order accuracy can be constructed by satisfying $n+1$ equations. In particular, because (2.26) 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_i . This is pretty obvious. 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 second-order scheme must involve at least three grid points. A scheme that is parsimonious in its use of points is called “*compact*.”

Consider a simple example. Still assuming a uniform grid, a first order scheme for

$\left(\frac{df}{dx}\right)_j$ can be constructed using the points j and $j+1$ as follows. From (2.26), we get $a_0 + a_1 = 0$ and $a_1 = 1$. Obviously we must choose $a_0 = -1$. Substituting into (2.23), we find that the scheme is given by $\left(\frac{df}{dx}\right)_j \equiv \frac{f(x_j + \Delta x) - f(x_j)}{\Delta x}$, i.e., the one-sided difference discussed earlier. Obviously we can also construct a first-order scheme using the points j and $j-1$, with a similar one-sided result. If we choose the points $j+1$ and $j-1$, imposing the requirements for first-order accuracy, i.e., (2.26), will actually give us the centered second-order scheme, i.e., $\left(\frac{df}{dx}\right)_j \equiv \frac{f(x_j + \Delta x) - f(x_{j-1})}{2\Delta x}$, because (2.27) is satisfied “accidentally” or “automatically” -- 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.

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

$$\left(\frac{df}{dx}\right)_j \equiv \sum_{i=-\infty}^{\infty} b_i f(x_{j+i}) . \quad (2.29)$$

Note that, since Δx is no longer a constant, the factor of $\frac{1}{\Delta x}$ that appears in (2.23) has been omitted in (2.29), and in view of this, in order to avoid notational confusion, we have replaced the symbol a_i by b_i . Similarly, Eq. (2.24) is replaced by

$$f(x_{j+i}) = f_j^0 + f_j^1(x_{j+i} - x_j) + f_j^2 \frac{(x_{j+i} - x_j)^2}{2!} + f_j^3 \frac{(x_{j+i} - x_j)^3}{3!} + \dots \quad (2.30)$$

Substitution of (2.30) into (2.29) gives

$$\left(\frac{df}{dx}\right)_j \equiv \sum_{i=-\infty}^{\infty} b_i \left[f_j^0 + f_j^1(x_{j+i} - x_j) + f_j^2 \frac{(x_{j+i} - x_j)^2}{2!} + f_j^3 \frac{(x_{j+i} - x_j)^3}{3!} + \dots \right] . \quad (2.31)$$

To have first-order accuracy with (2.31), we must require that

$$\sum_{i=-\infty}^{\infty} b_i = 0 \text{ and } \sum_{i=-\infty}^{\infty} b_i (x_{j+i} - x_j) = 1 . \quad (2.32)$$

It may appear that when we require first-order accuracy by enforcing (2.32), the leading term of the error in (2.31), namely $\sum_{i=-\infty}^{\infty} b_i f_j^2 \frac{(x_{j+i} - x_j)^2}{2!}$, will be of order $(\Delta x)^2$, but this is not really true because, as shown below, $b_i \sim \frac{1}{\Delta x}$.

Similarly, to achieve second-order accuracy with (2.31), we must require, in addition to (2.32), that

$$\sum_{i=-\infty}^{\infty} b_i (x_{j+i} - x_j)^2 = 0. \quad (2.33)$$

In general, to have at least n th-order accuracy, we must require that

$$\sum_{i=-\infty}^{\infty} (x_{j+i} - x_j)^m b_i = \delta_{m,1} \quad \text{for } 0 \leq m \leq n. \quad (2.34)$$

As an example, the first-order accurate scheme using the points j and $j+1$ must satisfy the two equations obtained from (2.32), i.e., $b_0 + b_1 = 0$ and $b_1 = \frac{1}{x_{j+1} - x_j}$, so that $b_1 = \frac{-1}{x_{j+1} - x_j}$. From (2.29), the scheme is $\left(\frac{df}{dx}\right)_j \cong \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}$, which, clearly, is equivalent to the result that we obtained for the case of the uniform grid.

To obtain a second-order accurate approximation to $\left(\frac{df}{dx}\right)_j$ on an arbitrary grid, using the three points $j-1$, j and $j+1$, we must require, from (2.32) that

$$b_{-1} + b_0 + b_1 = 0, \text{ and } b_{-1}(x_{j-1} - x_j) + b_1(x_{j+1} - x_j) = 1, \quad (2.35)$$

and from (2.34) that

$$b_{-1}(x_{j-1} - x_j)^2 + b_1(x_{j+1} - x_j)^2 = 0. \quad (2.36)$$

The solution of this system of three equations is

$$b_{-1} = \left(\frac{-1}{x_{j+1} - x_{j-1}}\right) \left(\frac{x_{j+1} - x_j}{x_j - x_{j-1}}\right), \quad (2.37)$$

$$b_0 = \left(\frac{1}{x_{j+i} - x_{j-1}} \right) \left[\left(\frac{x_{j+i} - x_j}{x_j - x_{j-1}} \right) - \left(\frac{x_j - x_{j-1}}{x_{j+i} - x_j} \right) \right], \quad (2.38)$$

$$b_1 = \left(\frac{1}{x_{j+i} - x_{j-1}} \right) \left(\frac{x_j - x_{j-1}}{x_{j+i} - x_j} \right). \quad (2.39)$$

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

Here is a simple but very practical question: Suppose that we use a scheme that has second-order accuracy on a uniform grid, but we apply it on a non-uniform grid. What happens? As a concrete example, we use the scheme

$$\left(\frac{df}{dx} \right)_j \equiv \frac{f(x_{j+1}) - f(x_{j-1}))}{x_{j+1} - x_{j-1}}. \quad (2.40)$$

By inspection, we have

$$b_{-1} = \frac{-1}{x_{j+i} - x_{j-1}}, \quad (2.41)$$

$$b_0 = 0, \quad (2.42)$$

$$b_1 = \frac{1}{x_{j+i} - x_{j-1}}. \quad (2.43)$$

Eqs. (2.41)-(2.43) can be compared with (2.37)-(2.39). Obviously the scheme does not have second-order accuracy on a non-uniform grid, because (2.37)-(2.39) are not satisfied for a non-uniform grid. We note, however, that Eqs. (2.41)-(2.43) do satisfy both of the conditions in (2.35), even when the grid is non-uniform. *This means that the scheme does have first-order accuracy, even on the non-uniform grid.*

This example illustrates that a scheme that has been designed for use on a uniform grid, with second-order (or even better than second-order) accuracy, will reduce to a scheme of first-order accuracy when applied on a non-uniform grid. A special case has been used as an example here, but the conclusion is true quite generally.

Finally, we observe that a very similar approach can be used to derive approximations to higher-order derivatives of f . For example, to derive approximations to

$\left(\frac{d^2 f}{dx^2} \right)$, on a (possibly) non-uniform grid, we write

$$\left(\frac{d^2 f}{dx^2}\right)_j \cong \sum_{i=-\infty}^{\infty} c_i f(x_{j+i}). \quad (2.44)$$

Obviously, it is going to turn out that $c_i \sim \frac{1}{(\Delta x)^2}$. Substitution of (2.30) into (2.44) gives

$$\left(\frac{d^2 f}{dx^2}\right)_j \cong \sum_{i=-\infty}^{\infty} c_i \left[f_j^0 + f_j^1 (x_{j+i} - x_j) + f_j^2 \frac{(x_{j+i} - x_j)^2}{2!} + f_j^3 \frac{(x_{j+i} - x_j)^3}{3!} + \dots \right]. \quad (2.45)$$

Keeping in mind that $c_i \sim \frac{1}{(\Delta x)^2}$, we see that a first-order accurate approximation is ensured if we enforce the three conditions

$$\sum_{i=-\infty}^{\infty} c_i = 0, \text{ and } \sum_{i=-\infty}^{\infty} c_i (x_{j+i} - x_j)^2 = 2! \text{ and } \sum_{i=-\infty}^{\infty} c_i (x_{j+i} - x_j) = 0. \quad (2.46)$$

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

$$\sum_{i=-\infty}^{\infty} c_i (x_{j+i} - x_j)^3 = 0. \quad (2.47)$$

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

$$\sum_{i=-\infty}^{\infty} (x_{j+i} - x_j)^m c_i = (2!) \delta_{m,2} \text{ for } 0 \leq m \leq n+1. \quad (2.48)$$

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., (2.35) and (2.36)]. Now we see that a second-order approximation to the second derivative must involve a minimum of four grid points, because four conditions must be satisfied, i.e., (2.46) and (2.47). In the special case of a uniform grid, three points suffice. With a non-uniform grid, five points may be preferable to four, from the point of view of symmetry.

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

$$\left(\frac{d^l f}{dx^l}\right)_j \cong \sum_{i=-\infty}^{\infty} d_i f(x_{j+i}) \quad (2.49)$$

where

$$\sum_{i=-\infty}^{\infty} (x_{j+i} - x_j)^m d_i = (l!) \delta_{m,l} \quad \text{for } 0 \leq m \leq n + l - 1. \quad (2.50)$$

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 .

What happens for $l = 0$?

2.3 Extension to two dimensions

The approach presented above can be generalized to multi-dimensional problems. We will use the two-dimensional Laplacian operator as an example. Consider a finite-difference approximation to the Laplacian, of the form

$$(\nabla^2 f)_j \cong \sum_{k=-\infty}^{\infty} c_k f(x_{j+k}, y_{j+k}) \quad (2.51)$$

Here we use one-dimensional indices even though we are on a two-dimensional grid. The subscript j denotes a particular grid point (“home base” for this calculation), whose coordinates are (x_j, y_j) . Similarly, the subscript $j + k$ denotes a different grid point whose coordinates are (x_{j+k}, y_{j+k}) .

The two-dimensional Taylor series expansion is

$$\begin{aligned}
f(x_{j+k}, y_{j+k}) &= f(x_j, y_j) + [(\delta x)_k f_x + (\delta y)_k f_y] \\
&+ \frac{1}{2!} [(\delta x)_k^2 f_{xx} + 2(\delta x)_k (\delta y)_k f_{xy} + (\delta y)_k^2 f_{yy}] + \dots, \\
&+ \frac{1}{3!} [(\delta x)_k^3 f_{xxx} + 3(\delta x)_k^2 (\delta y)_k f_{xxy} + 3(\delta x)_k (\delta y)_k^2 f_{xyy} + (\delta y)_k^3 f_{yyy}] \\
&+ \frac{1}{4!} [(\delta x)_k^4 f_{xxxx} + 4(\delta x)_k^3 (\delta y)_k f_{xxx} + 6(\delta x)_k^2 (\delta y)_k^2 f_{xxyy} \\
&+ 4(\delta x)_k (\delta y)_k^3 f_{xyyy} + (\delta y)_k^4 f_{yyyy}] + \dots \} ,
\end{aligned} \tag{2.52}$$

where

$$(\delta x)_k \equiv x_{j+k} - x_j \text{ and } (\delta y)_k \equiv y_{j+k} - y_j, \tag{2.53}$$

and it is understood that all of the derivatives are evaluated at the point (x_j, y_j) .

Substituting from (2.52) into (2.51), we find that

$$\begin{aligned}
(\nabla^2 f)_j &\equiv \sum_{k=-\infty}^{\infty} c_k \{ f(x_j, y_j) + [(\delta x)_k f_x + (\delta y)_k f_y] \\
&+ \frac{1}{2!} [(\delta x)_k^2 f_{xx} + 2(\delta x)_k (\delta y)_k f_{xy} + (\delta y)_k^2 f_{yy}] + \dots, \\
&+ \frac{1}{3!} [(\delta x)_k^3 f_{xxx} + 3(\delta x)_k^2 (\delta y)_k f_{xxy} + 3(\delta x)_k (\delta y)_k^2 f_{xyy} + (\delta y)_k^3 f_{yyy}] \\
&+ \frac{1}{4!} [(\delta x)_k^4 f_{xxxx} + 4(\delta x)_k^3 (\delta y)_k f_{xxx} + 6(\delta x)_k^2 (\delta y)_k^2 f_{xxyy} \\
&+ 4(\delta x)_k (\delta y)_k^3 f_{xyyy} + (\delta y)_k^4 f_{yyyy}] + \dots \} .
\end{aligned} \tag{2.54}$$

To have first-order accuracy, we need

$$\sum_{k=-\infty}^{\infty} c_k = 0, \tag{2.55}$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta x)_k^2 = 2! , \quad (2.56)$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta y)_k^2 = 2! , \quad (2.57)$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta x)_k = 0 \quad (2.58)$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta y)_k = 0 , \text{ and} \quad (2.59)$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta x)_k(\delta y)_k = 0 . \quad (2.60)$$

From (2.56) and (2.57), it is clear that c_k is of order δ^2 , where δ denotes δx or δy .

Therefore, the quantities inside the sums in (2.58) and (2.59) are of order δ^{-1} , and the quantities inside the sum of (2.60) or of order one. This is why (2.58)-(2.60) are required, in addition to (2.55)-(2.57), to obtain first-order accuracy.

So far (2.55) - (2.60) involve only six equations, and so six grid points are needed. To get second-order (or higher) accuracy, we will need to add more points, unless we are fortunate enough to use a highly symmetrical grid that permits the conditions for higher-order accuracy to be satisfied automatically. For example, if we satisfy (2.55) - (2.60) on a square grid, we will get second-order accuracy “for free.” More generally, with a non-uniform grid, we need the following four additional conditions to achieve second-order accuracy:

$$\sum_{k=-\infty}^{\infty} c_k(\delta x)_k^3 = 0 , \quad (2.61)$$

$$\sum_{k=-\infty}^{\infty} c_k(\delta x)_k^2(\delta y)_k = 0 , \quad (2.62)$$

$$\sum_{k=-\infty}^{\infty} c_k (\delta x)_k (\delta y)_k^2 = 0 , \quad (2.63)$$

$$\sum_{k=-\infty}^{\infty} c_k (\delta y)_k^3 = 0 . \quad (2.64)$$

Therefore, in general a total of ten conditions must be satisfied to ensure second-order accuracy on a non-uniform grid.

For the continuous Laplacian on a closed or periodic domain, we can prove the following:

$$\int_A (\nabla^2 f) dA = 0 , \quad (2.65)$$

$$\int_A f(\nabla^2 f) dA \leq 0 . \quad (2.66)$$

Here the integrals are with respect to area, over the entire domain. The corresponding finite-difference requirements are

$$\sum_{\text{all } j} (\nabla^2 f)_j A_j \cong \sum_{\text{all } j} \left[\sum_{k=-\infty}^{\infty} c_k f(x_{j+k}, y_{j+k}) \right] A_j = 0 , \text{ and} \quad (2.67)$$

$$\sum_{\text{all } j} f_j (\nabla^2 f)_j A_j \cong \sum_{\text{all } j} f_j \left[\sum_{k=-\infty}^{\infty} c_k f(x_{j+k}, y_{j+k}) \right] A_j \leq 0 , \quad (2.68)$$

where A_j is the area of grid-cell j . These requirements must hold for an arbitrary spatial distribution of f , so they actually represent conditions on the c_k . Very similar (but more complicated) requirements were discussed by Arakawa (1966), in the context of the Jacobian operator. Further discussion is given later.

2.4 An example of a finite difference–approximation to a differential equation

With these definitions and concepts, we proceed directly to a simple example of a partial differential equation. Consider the one-dimensional “advection” equation,

$$\left(\frac{\partial u}{\partial t}\right)_x + c\left(\frac{\partial u}{\partial x}\right)_t = 0, \quad (2.69)$$

where c is a constant, and $u = u(x, t)$. This is a first-order linear partial differential equation with a constant coefficient, namely c . Eq. (2.69) looks harmless, but it causes no end of trouble, as we will see.

Suppose that

$$u(x, 0) = F(x) \text{ for } -\infty < x < \infty. \quad (2.70)$$

This is an “initial condition.” What is $u(x, t)$? This is a simple example of an initial value problem. We first work out the analytic solution, for later comparison with our numerical solution. Define

$$\xi \equiv x - ct. \quad (2.71)$$

We can write

$$\begin{aligned} \left(\frac{\partial u}{\partial x}\right)_\xi &= \left(\frac{\partial u}{\partial x}\right)_t + \left(\frac{\partial u}{\partial t}\right)_x \left(\frac{\partial t}{\partial x}\right)_\xi \\ &= \left(\frac{\partial u}{\partial x}\right)_t + \left(\frac{\partial u}{\partial t}\right)_x \frac{1}{c} \\ &= 0. \end{aligned} \quad (2.72)$$

The first line of (2.72) can be understood by reference to Fig. 2.4. Similarly,

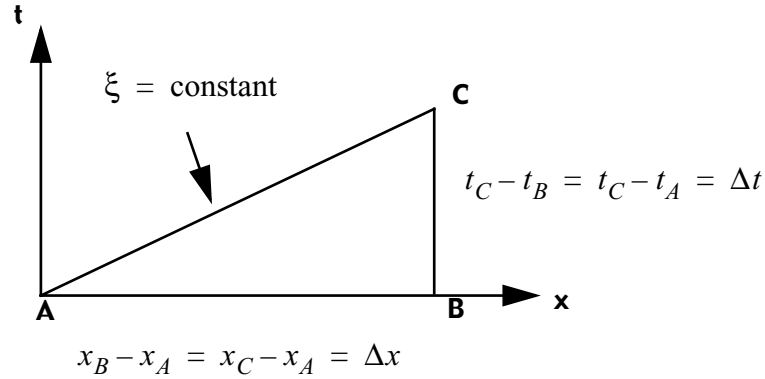


Figure 2.3: Figure used in the derivation of the first line of (2.72).

$$\begin{aligned}
\left(\frac{\partial u}{\partial t}\right)_{\xi} &= \left(\frac{\partial u}{\partial t}\right)_x + \left(\frac{\partial u}{\partial x}\right)_t \left(\frac{\partial x}{\partial t}\right)_{\xi} \\
&= \left(\frac{\partial u}{\partial x}\right)_t + \left(\frac{\partial u}{\partial t}\right)_x c \\
&= 0 .
\end{aligned} \tag{2.73}$$

It follows that

$$u = f(\xi) \tag{2.74}$$

is the general solution to (2.69). At $t = 0$, $\xi \equiv x$ and $u(x) = f(x)$, i.e. the shape of f is determined by the initial condition. Eq. (2.74) means that u is constant along the line $\xi \equiv x - ct = \text{constant}$. In order to satisfy the initial condition (2.70), we chose $f \equiv F$. Thus $u(\xi) = F(\xi) = F(x - ct)$ is the solution to the differential equation (2.69) which satisfies the initial condition. We see that an initial value simply “moves along” the lines of constant ξ . The initial “shape” of $u(x)$, namely $F(x)$, is just carried along by the wind. From a physical point of view this is obvious.

Keeping in mind this exact solution, we now investigate one possible numerical solution of (2.69). We construct a grid, as in Fig. 2.4. An example of a finite difference approximation to (2.69) is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c \left(\frac{u_j^n - u_{j-1}^n}{\Delta x} \right) = 0 . \tag{2.75}$$

Here we have used the forward difference quotient in time and the backward difference quotient in space. If $c > 0$, (2.75) is called the “upstream” difference scheme. Because

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} \rightarrow \frac{\partial u}{\partial t} \text{ as } \Delta t \rightarrow 0 , \tag{2.76}$$

and

$$\frac{u_j^n - u_{j-1}^n}{\Delta x} \rightarrow \frac{\partial u}{\partial x} \text{ as } \Delta x \rightarrow 0 . \tag{2.77}$$

we conclude that (2.75) does approach (2.69) as Δt and Δx both approach zero. The upstream scheme has some serious weaknesses, but it also has some very useful properties.

If we know u_j^n at some time level n for all j , then we can compute u_j^{n+1} at the

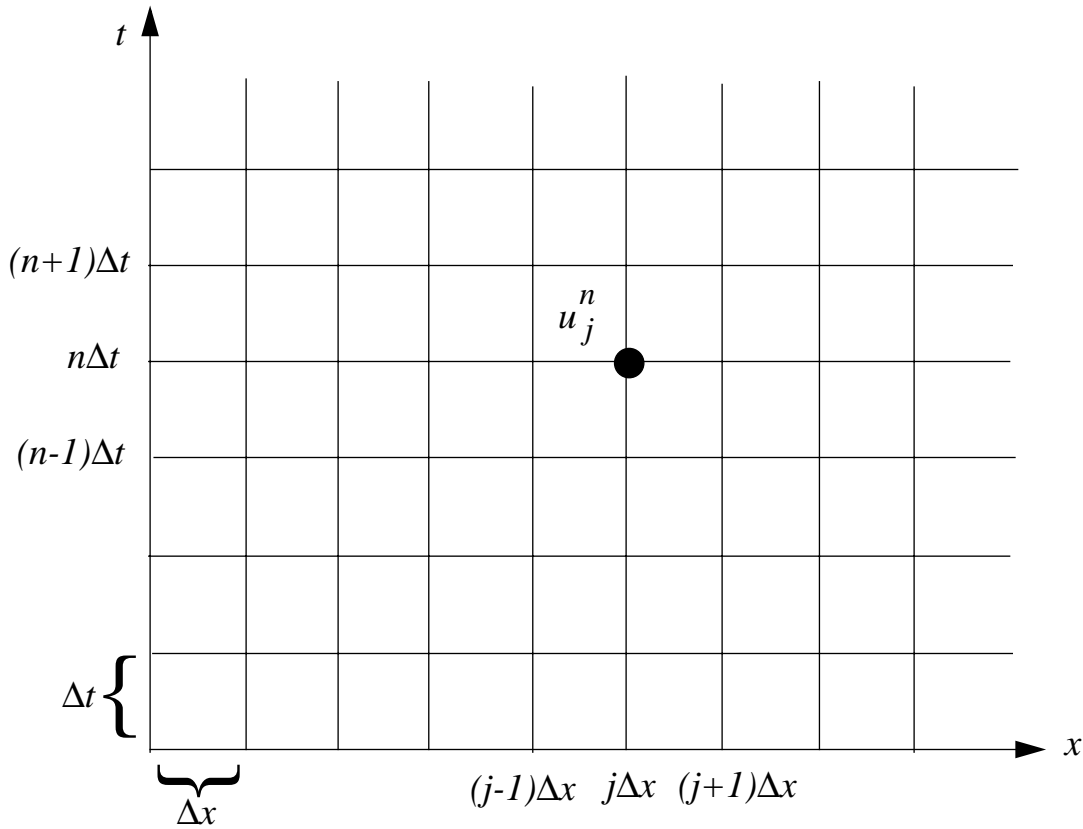


Figure 2.4: A grid for the solution of the one-dimensional advection equation.

next time level $n + 1$. Note that this scheme 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.

In view of (2.76) and (2.77), it may seem obvious that the *solution* of (2.75) approaches the *solution* of (2.69) as $\Delta x \rightarrow 0$ and $\Delta t \rightarrow 0$. This “obvious” conclusion is not necessarily true, as we shall see.

2.5 Accuracy and truncation error of a finite-difference scheme.

We have already defined accuracy and truncation error for finite difference quotients. Now we define truncation error and accuracy for a finite-difference scheme. Here we define a finite-difference scheme as a finite-difference equation which approximates, term-by-term, a differential equation.

It is easy to find an approximation 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.* Now 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 be a useful approximation to the solution of D . Wouldn't that be nice? Unfortunately, it isn't necessarily true.

Letting $u(x, t)$ denote the (exact) solution of the differential equation, we see that $u(j\Delta x, n\Delta t)$ is the value of this exact solution at the discrete point $(j\Delta x, n\Delta t)$ on the grid shown in Fig. 2.3, while u_j^n is the “exact” solution of a finite-difference equation, at the same point. In general, $u_j^n \neq u(j\Delta x, n\Delta t)$. We wish that they were equal!

A measure of the accuracy of the finite-difference *scheme* can be obtained by substituting the solution of the differential equation into the finite-difference equation. For the upstream scheme given by (2.75), we get

$$\frac{u[j\Delta x, (n+1)\Delta t] - u(j\Delta x, n\Delta t)}{\Delta t} + c \left\{ \frac{u(j\Delta x, n\Delta t) - u[(j-1)\Delta x, n\Delta t]}{\Delta x} \right\} = \varepsilon, \quad (2.78)$$

where ε is called “truncation error” of the scheme. The truncation error of the scheme is a measure of how accurately the solution $u(x, t)$ of the original differential equation (2.69) satisfies the finite-difference equation (2.75).

Note that, since u_j^n is defined only at discrete points, it is not differentiable, and so we cannot substitute u_j^n into the differential equation. Because of this, we cannot measure how accurately u_j^n satisfies the differential equation.

If we obtain the terms in (2.78) from Taylor Series expansion of $u(x, t)$ about the point $(j\Delta x, n\Delta t)$, and use the fact that $u(x, t)$ satisfies (2.69), we find that

$$\varepsilon = \left(\frac{1}{2!} \Delta t \frac{\partial^2 u}{\partial t^2} + \dots \right) + c \left(-\frac{1}{2!} \Delta x \frac{\partial^2 u}{\partial x^2} + \dots \right). \quad (2.79)$$

We say this is a “first-order scheme” because the lowest powers of Δt and Δx in (2.79) are 1. The notations $O(\Delta t, \Delta x)$ or $O(\Delta t) + O(\Delta x)$ can be used to express this. We say that a scheme is consistent with the differential equation if the truncation error of the scheme approaches zero as Δt and Δx approach zero. The upstream scheme under consideration here is, therefore, consistent.

2.6 Discretization error and convergence

There are two sources of error in a numerical solution. One is the *round-off error*, which is the difference of a numerical solution from the “exact” solution of the finite difference equation, u_j^n , and is a property of the machine being used (and to some extent the details of the program). The other source of error is the *discretization error* defined by

$u_j^n - u(j\Delta x, n\Delta t)$. Round-off error can sometimes be a problem but usually it is not, and we will not consider it in this course.

The truncation error, discussed in the last section, can be made as small as desired by making Δx and Δt smaller and smaller, so long as the scheme is consistent and $u(x, t)$ is a smooth function. *A decrease in the truncation error does not necessarily guarantee that the discretization error will also become small, however.* This is demonstrated below.

We now analyze how the discretization error changes as the grid is refined (i.e., as Δt and $\Delta x \rightarrow 0$). If the discretization error approaches zero, then we say that the solution *converges*. Fig. 2.5 gives an example of a situation in which accuracy is increased but the

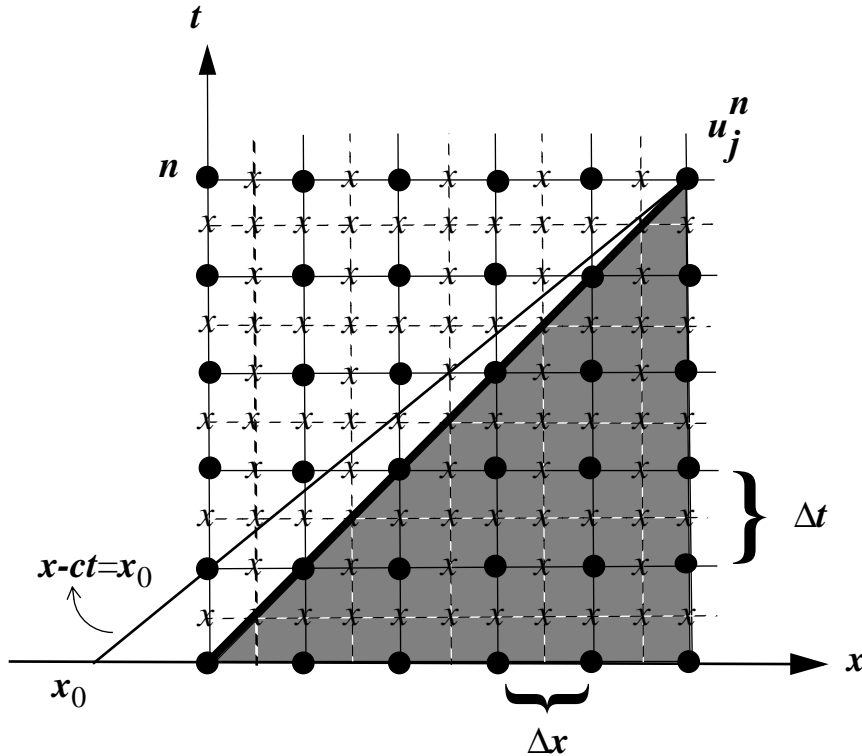


Figure 2.5: The shaded area represents the “domain of dependence” of the solution of the upstream scheme at the point $x = j\Delta x, t = n\Delta t$.

solution nevertheless does not converge. The thin diagonal line in the figure shows the characteristic along with u is “carried,” i.e. u is constant along the line. This is the exact solution. To work out the numerical approximation to this solution, we first choose Δx and Δt such that the grid points are the dots in the figure. The domain consisting of the grid points carrying values of u on which u_j^n depends is called the “domain of dependence.” The shaded area in the figure shows this domain for the upstream scheme (2.75).

We could increase the accuracy of the scheme by cutting Δx and Δt in half, that is, by adding the points denoted by small x 's forming a denser grid. Notice that the domain of dependence does not change, no matter how refined or dense the grid is, so long as the ratio $\frac{c\Delta t}{\Delta x}$ remains the same. This is a clue that $\frac{c\Delta t}{\Delta x}$ is an important quantity.

Suppose that the line through the point $(j\Delta x, n\Delta t)$, $x - ct = x_0$, where x_0 is a constant, does not lie in the domain of dependence. This is the situation shown in the figure. In general, there is no hope of obtaining smaller discretization error, no matter how small Δx and Δt become, so long as $\frac{c\Delta t}{\Delta x}$ is unchanged, because the true solution $u(j\Delta x, n\Delta t)$ depends only on the initial value of u at the single point $(x_0, 0)$ which cannot influence u_j^n . You could change $u(x_0, 0)$ [and hence $u(j\Delta x, n\Delta t)$], but the computed solution u_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.

A finite-difference scheme for which the discretization error can be made small *for any initial condition* is called a *convergent* finite difference scheme. Therefore,

$$0 \leq \frac{c\Delta t}{\Delta x} \leq 1, \quad (2.80)$$

is a *necessary* condition for convergence when the upstream scheme is used. Notice that if c is negative (giving what we might call a “downstream” scheme), it is impossible to satisfy (2.80). Of course, for $c < 0$ we can use

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c \left(\frac{u_{j+1}^n - u_j^n}{\Delta x} \right) = 0, \quad (2.81)$$

in place of (2.75). So our computer program can have an “if-test” that checks the sign of c , and uses (2.75) if $c \geq 0$, and (2.81) if $c < 0$. This is bad, though, because if-tests can cause slow execution on certain types of computers, and besides, if-tests are ugly. If we define

$$c_+ \equiv \frac{c + |c|}{2} \geq 0, \text{ and } c_- \equiv \frac{c - |c|}{2} \leq 0, \quad (2.82)$$

then the upstream scheme can be written as

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c_+ \left(\frac{u_j^n - u_{j-1}^n}{\Delta x} \right) + c_- \left(\frac{u_{j+1}^n - u_j^n}{\Delta x} \right) = 0. \quad (2.83)$$

This form avoids the use of if-tests and is also convenient for use in pencil-and-paper analysis, 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 a differential equation has been approximated.

Discretization error measures the accuracy with which the solution of the differential equation has been approximated.

Minimizing the truncation error is usually easy. Minimizing the discretization error can be much harder.

2.7 Interpolation and extrapolation

Referring to (2.75), we can rewrite the upstream scheme as

$$u_j^{n+1} = u_j^n (1 - \mu) + u_{j-1}^n \mu \quad (2.84)$$

where

$$\mu \equiv \frac{c \Delta t}{\Delta x}. \quad (2.85)$$

This scheme has the form of either an *interpolation* or an *extrapolation*, depending on the value of μ . To see this, refer to Fig. 2.6. Along the line plotted in the figure,

$$u = u_{j-1}^n + \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \right) (u_j^n - u_{j-1}^n) = \left[1 - \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \right) u_{j-1}^n + \left(\frac{x - x_{j-1}}{x_j - x_{j-1}} \right) u_j^n \right] \quad (2.86)$$

which has the same form as our scheme if we identify

$$\mu \equiv \frac{x - x_{j-1}}{x_j - x_{j-1}} \quad (2.87)$$

For $0 \leq \mu \leq 1$ we have *interpolation*. For $\mu < 0$ or $\mu > 1$ we have *extrapolation*. Note that for the case of interpolation, u_j^{n+1} will be intermediate in value between u_{j-1}^n and u_j^n . For instance, if u_{j-1}^n and u_j^n are both ≥ 0 , then u_j^n will also be ≥ 0 . For the case

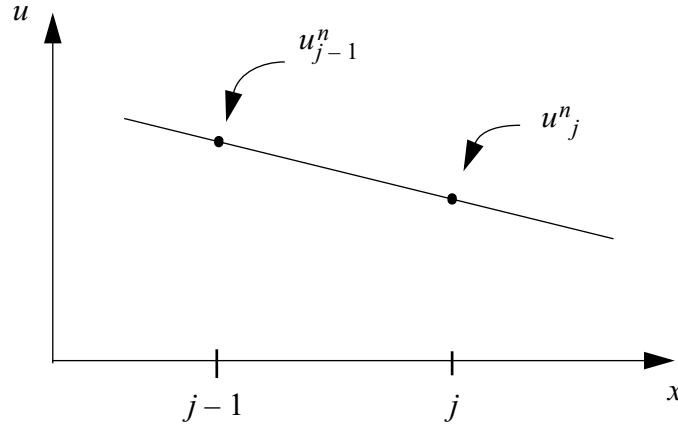


Figure 2.6: Diagram illustrating the concepts of interpolation and extrapolation. See text for details.

extrapolation, u_j^{n+1} will lie outside the range of u_{j-1}^n and u_j^n .

We use both interpolation and extrapolation extensively in this course.

2.8 Stability

We now investigate the behavior of the discretization error $|u_j^n - u(j\Delta x, n\Delta t)|$ as n increases, for fixed Δx and Δt . Does the error remain bounded for any initial condition? If so the scheme is said to be stable; otherwise it is unstable.

In most physical problems the true solution is bounded, at least for finite t , so that the solution of the scheme is bounded if the scheme is stable.

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*.

As an illustration of the direct method, consider the upstream scheme, as given by (2.84). Note that u_j^{n+1} is a weighted mean of u_j^n and u_{j-1}^n . If $0 \leq \mu \leq 1$ [the necessary condition for convergence according to (2.80)], we may write

$$|u_j^{n+1}| \leq |u_j^n|(1 - \mu) + |u_{j-1}^n|\mu. \quad (2.88)$$

Therefore,

$$\max_{(j)} |u_j^{n+1}| \leq \max_{(j)} |u_j^n|(1 - \mu) + \max_{(j)} |u_{j-1}^n|\mu, \quad (2.89)$$

or if we assume that $\max_{(j)} |u_j^n| = \max_{(j)} |u_{j-1}^n|$, then

$$\max_{(j)} |u_j^{n+1}| \leq \max_{(j)} |u_j^n|, \text{ provided that } 0 \leq \mu \leq 1. \quad (2.90)$$

We have shown that for $0 \leq \mu \leq 1$ the solution u_j^n remains bounded for all time. Therefore, $0 \leq \mu \leq 1$ is a *sufficient* condition for stability. For this scheme, the condition for stability has turned out to be the same as the condition for convergence. In other words, if the scheme is convergent it is stable, and vice versa.

This conclusion is actually obvious from (2.84), because when $0 \leq \mu \leq 1$, u_j^{n+1} is obtained by linear interpolation *in space*, from the available u_j^n to the point $x = j\Delta x - c\Delta t$. This is reasonable, since in advection the time rate of change at a point is closely related to the spatial variations in the neighborhood of that point.

Note that in the true solution of the differential equation for advection, the maxima and minima of u never change. They are just carried along to different spatial locations. So, for the exact solution, the equality in (2.90) would hold.

The direct method is not very widely applicable. It becomes intractable for complex schemes.

The second method, the energy method, is more widely applicable, even for some nonlinear equations. We illustrate it here by means of application to the scheme (2.75). With this method we ask: “Is $\sum_j (u_j^n)^2$ bounded after an arbitrary number of time steps?” Here the summation obviously ^j must be taken over a finite number of grid points. This is not an important limitation because in practice we are always dealing with a finite number of grid points. If the sum is bounded, then each u_j^n must also be bounded. Whereas in the direct method we checked $\max_{(j)} |u_j^{n+1}|$, here in the energy method we check $\sum_j (u_j^n)^2$. The two approaches are therefore somewhat similar.

Returning then to (2.84), squaring both sides, and summing over the domain, we obtain

$$\begin{aligned} \sum_j (u_j^{n+1})^2 &= \sum_j [(u_j^n)^2 (1-\mu)^2 + 2\mu(1-\mu)u_j^n u_{j-1}^n + \mu^2 (u_{j-1}^n)^2] \\ &= (1-\mu)^2 \sum_j (u_j^n)^2 + 2\mu(1-\mu) \sum_j u_j^n u_{j-1}^n + \mu^2 \sum_j (u_{j-1}^n)^2. \end{aligned} \quad (2.91)$$

For simplicity, suppose that u is periodic in x , and consider a summation over one complete cycle. Then

$$\sum_j (u_{j-1}^n)^2 = \sum_j (u_j^n)^2. \quad (2.92)$$

We note that

$$\text{if } \sum_j u_j^n u_{j-1}^n < 0 \text{ then } \sum_j u_j^n u_{j-1}^n < \sum_j (u_j^n)^2; \text{ and} \quad (2.93)$$

$$\text{if } \sum_j u_j^n u_{j-1}^n > 0 \text{ then } \sum_j u_j^n u_{j-1}^n \leq \sum_j (u_j^n)^2. \quad (2.94)$$

To derive (2.94) we have used Schwartz's inequality, i.e.,

$$\left(\sum_j a_j\right)^2 \left(\sum_j b_j\right)^2 \leq \left(\sum_j a_j^2\right) \left(\sum_j b_j^2\right) \quad (2.95)$$

for any sets of a s and b s, and (2.92), i.e.

$$\left[\sum_j u_j^n u_{j-1}^n\right]^2 \leq \sum_j (u_j^n)^2 \sum_j (u_{j-1}^n)^2 = \left[\sum_j (u_j^n)^2\right]^2. \quad (2.96)$$

Use of (2.92), (2.93) and (2.94) in (2.91) gives

$$\sum_j (u_j^{n+1})^2 \leq [(1-\mu)^2 + 2\mu(1-\mu) + \mu^2] \sum_j (u_j^n)^2, \quad (2.97)$$

provided that $\mu(1-\mu) \geq 0$, which is equivalent to $0 \leq \mu \leq 1$. The quantity in square brackets in (2.97) is equal to 1. We conclude that

$$\sum_j (u_j^{n+1})^2 \leq \sum_j (u_j^n)^2, \text{ provided that } 0 \leq \mu \leq 1. \quad (2.98)$$

As with the direct method, we conclude that $0 \leq \mu \leq 1$ is a sufficient condition for the scheme to be stable.

A very powerful tool for testing the stability of linear partial difference equations with constant coefficients is von Neumann's method. It is the method that will be used most often in this course. Solutions to linear partial differential equations can be expressed as superposition of waves, by means of Fourier series. Von Neuman's method simply tests the stability of each Fourier component.

To illustrate von Neumann's method, we return first to the advection equation, (2.69):

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0 . \quad (2.99)$$

We assume for simplicity that the domain is periodic. First, we look for a solution with the wave form

$$u(x, t) = \text{Re}[\hat{u}(t)e^{ikx}] , \quad (2.100)$$

where $|\hat{u}(t)|$ is the amplitude of the wave. Here we consider a single wave number, for simplicity, but in general we could replace the right-hand side of (2.100) by a sum over all relevant wave numbers. Using (2.100), (2.69) becomes

$$\frac{d\hat{u}}{dt} + ikc \hat{u} = 0 . \quad (2.101)$$

By Fourier expansion we have converted the partial differential equation (2.69) into an ordinary differential equation, (2.101), whose solution is

$$\hat{u}(t) = \hat{u}(0)e^{-ikct} , \quad (2.102)$$

where $u(0)$ is the initial value of u . The solution to (2.69) is, from (2.100),

$$u(x, t) = \text{Re}[\hat{u}(0)e^{ik(x-ct)}] . \quad (2.103)$$

Note that (2.103) is a valid solution only for $c = \text{constant}$.

For a finite difference equation, we use in place of (2.100)

$$u_j^n = \text{Re}[\hat{u}^{(n)} e^{ikj\Delta x}] . \quad (2.104)$$

Then $|\hat{u}^{(n)}|$ is the amplitude of the wave at time-level n . Note that the shortest resolvable wave, with $L = 2\Delta x$, has $k\Delta x = \pi$, while longer waves have $k\Delta x < \pi$, so there is never any need to consider $k\Delta x > \pi$. Define λ , which may be complex, by

$$\hat{u}^{(n+1)} \equiv \lambda \hat{u}^{(n)} . \quad (2.105)$$

Then $|\hat{u}^{(n+1)}| = |\lambda| |\hat{u}^{(n)}|$. We call λ the “amplification factor.” As shown below, we can work out the form of λ for a particular finite-difference scheme. In general λ depends on k , so we could write λ_k , but usually we suppress that urge for the sake of keeping the notation simple. Note that λ can also be defined for the exact solution to the differential equation; from (2.102), we simply have $\hat{u}(t + \Delta t) \equiv e^{-ikc\Delta t} \hat{u}(t)$, so that for

the differential equation $\lambda = e^{-ikc\Delta t}$. For a particular scheme, we can compare the “exact” λ with the approximate λ defined by (2.105). Note that for the exact advection equation $|\lambda|$ is 1. For other problems, the exact $|\lambda|$ can differ from 1.

From (2.105) we see that after n time steps (starting from $n = 0$) the solution will be

$$\hat{u}^{(n)} = \hat{u}^{(0)} \lambda^n . \quad (2.106)$$

If we require that the solution remains bounded after arbitrarily many time steps, then we need

$$|\lambda| \leq 1 . \quad (2.107)$$

This is the condition for the stability of mode k .

To check the stability of a finite-difference scheme, using von Neumann’s method, we need to work out $|\lambda|$ for that scheme. We now illustrate the computation of $|\lambda|$ for the upstream scheme, which is given by (2.75). Substituting (2.104) into (2.75) gives

$$\frac{\hat{u}^{(n+1)} - \hat{u}^{(n)}}{\Delta t} + c \left(\frac{1 - e^{-ik\Delta x}}{\Delta x} \right) \hat{u}^{(n)} = 0 . \quad (2.108)$$

Notice that the true advection speed, c , is multiplied, in (2.108), by the factor $\left(\frac{1 - e^{-ik\Delta x}}{\Delta x} \right)$. Comparing with (2.101), we see that this factor is taking the place of ik in the exact solution. As $\Delta x \rightarrow 0$, the factor reduces to ik . This is a hint that the scheme gives an error in the advection speed. We return to this point later.

For now, we use the definition of λ , i.e. (2.105), together with (2.108), to infer that

$$\lambda = 1 - \mu(1 - \cos k\Delta x + i \sin k\Delta x) . \quad (2.109)$$

Note that λ is complex. Taking the modulus of (2.109), we obtain

$$|\lambda|^2 = 1 + 2\mu(\mu - 1)(1 - \cos k\Delta x) . \quad (2.110)$$

At $\mu = \frac{1}{2}$, for example, (2.110) reduces to

$$|\lambda|^2 = \frac{1}{2}(1 + \cos k\Delta x) . \quad (2.111)$$

According to (2.111), the amplification factor $|\lambda|$ depends on the wave number, k . Using $k \equiv \frac{2\pi}{2\Delta x}$ for $L = 2\Delta x$, $k \equiv \frac{\pi}{2\Delta x}$ for $L = 4\Delta x$, etc., the various curves shown in Fig. 2.7 can be constructed. We see clearly that this scheme damps for $0 < \mu < 1$ and is unstable for $\mu < 0$ and $\mu > 1$.

Although λ depends on k , it does not depend on x (i.e., on j) or on t (i.e., on n). Why not? The reason is that our “coefficient” c , has been assumed to be independent of x and t . Of course, in realistic problems the advecting current varies in both space and time. We normally apply von Neumann’s method to idealized versions of our equations, in which the various coefficients, such as c , are treated as constants. As a result, von Neumann’s method can “miss” instabilities that arise from variations of the coefficients. The energy method does not suffer from this limitation. It is very important to understand that von Neumann’s method can only analyze the stability of a linearized version of the equation. In fact, the equation has to be linear and with constant coefficients. This is an important weakness 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 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 kind of instability will be discussed in Chapter 6. If von Neumann’s method tells us that a scheme is *unstable*, then it is unstable.

As mentioned above, in general, the solution u_j^n can be expressed as a Fourier series. For simplicity, let us assume that the solution is periodic in x with period L_0 . Then u_j^n can be written as

$$u_j^n = Re \left[\sum_{m=-\infty}^{\infty} \hat{u}_m^{(n)} e^{imk_0j\Delta x} \right] = Re \left[\sum_{m=-\infty}^{\infty} \hat{u}_m^{(0)} e^{imk_0j\Delta x} (\lambda_m)^n \right], \quad (2.112)$$

where $k \equiv mk_0$

$$k_0 \equiv \frac{2\pi}{L_0}, \quad (2.113)$$

and m is an integer, which is analogous to what we call the “zonal wave number” in large-scale dynamics. In (2.112), the summation has been formally taken over all integers, although of course only a finite number of m 's could be used in a real application. Note that $|\lambda_m|$ is the amplification factor for mode m . We have

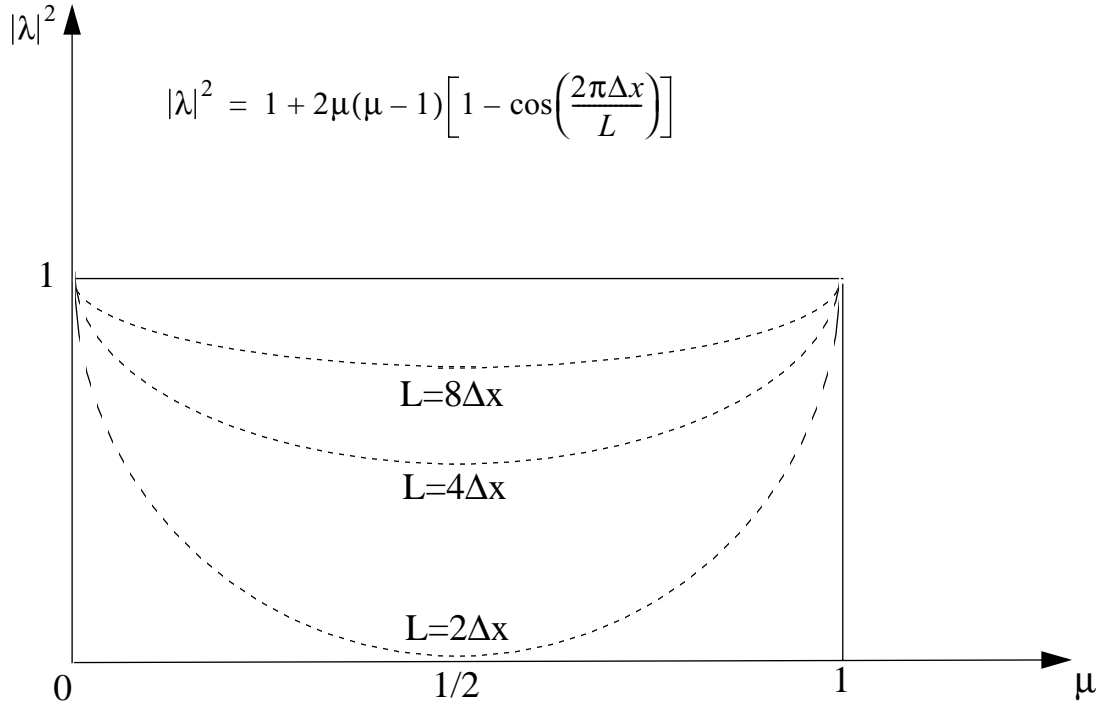


Figure 2.7: The amplification factor for the upstream scheme, plotted for three different wave lengths.

$$\begin{aligned}
 |u_j^n| &\leq \left| \sum_{m=-\infty}^{\infty} \hat{u}_m^{(0)} e^{imk_0j\Delta x} (\lambda_m)^n \right| \\
 &\leq \sum_{m=-\infty}^{\infty} \left| \hat{u}_m^{(0)} e^{imk_0j\Delta x} (\lambda_m)^n \right| \\
 &= \sum_{m=-\infty}^{\infty} |\hat{u}_m^{(0)}| |\lambda_m|^n.
 \end{aligned} \tag{2.114}$$

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

$$|u_j^n| \leq \sum_{m=-\infty}^{\infty} |\hat{u}_m^{(0)}|. \tag{2.115}$$

Therefore, $|u_j^n|$ will be bounded provided that $\sum_{m=-\infty}^{\infty} \hat{u}_m^{(0)} e^{imk_0j\Delta x}$, which gives the initial condition, is an absolutely convergent Fourier series. *This shows that $|\lambda_m| \leq 1$ for all m is sufficient for stability.* It is also necessary, because if $|\lambda_m| > 1$ 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 (2.109), λ_m for the upstream scheme is given by

$$\lambda_m = 1 - \mu(1 - \cos mk_0\Delta x + i \sin mk_0\Delta x). \quad (2.116)$$

The amplification factor is

$$|\lambda_m| = [1 + 2\mu(1 - \cos mk_0\Delta x)(\mu - 1)]^{\frac{1}{2}}. \quad (2.117)$$

From (2.117) we can show that $|\lambda_m| \leq 1 \leq$ holds for all m , if and only if $\mu(\mu - 1) \leq 0$, or $0 \leq \mu \leq 1$. This is the necessary and sufficient condition for the stability of the scheme.

Finally, we can test stability using the “matrix method¹,” which is really just von Neumann’s method with more general boundary conditions. The upstream scheme given by (2.75) (or by (2.112)) can be written in matrix form as

$$\begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \dots \\ u_{j-1}^{n+1} \\ u_j^{n+1} \\ u_{j+1}^{n+1} \\ \dots \\ u_{j-1}^{n+1} \\ u_j^{n+1} \end{bmatrix} = \begin{bmatrix} 1-\mu & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \mu \\ \mu & 1-\mu & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \mu & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1-\mu & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & \mu & 1-\mu & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \mu & 1-\mu & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \mu & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1-\mu & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \mu & 1-\mu \end{bmatrix} \cdot \begin{bmatrix} u_1^n \\ u_2^n \\ \dots \\ u_{j-1}^n \\ u_j^n \\ u_{j+1}^n \\ \dots \\ u_{j-1}^n \\ u_j^n \end{bmatrix}, \quad (2.118)$$

or

¹. Developed by K. Reeves.

$$[u_j^{n+1}] = [A][u_j^n], \quad (2.119)$$

where $[A]$ is the matrix written out on the right-hand side of (2.118). In writing (2.118), the cyclic boundary condition

$$u_1^{n+1} = (1 - \mu)u_1^n + \mu u_J^n \quad (2.120)$$

has been assumed. Recall from the definition of λ that $u_j^{n+1} = \lambda u_j^n$. This can be written in matrix form as

$$\begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \dots \\ u_{j-1}^{n+1} \\ u_j^{n+1} \\ u_{j+1}^{n+1} \\ \dots \\ u_{J-1}^{n+1} \\ u_J^{n+1} \end{bmatrix} = \begin{bmatrix} \lambda & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \lambda & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \lambda & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix} \cdot \begin{bmatrix} u_1^n \\ u_2^n \\ \dots \\ u_{j-1}^n \\ u_j^n \\ u_{j+1}^n \\ \dots \\ u_{J-1}^n \\ u_J^n \end{bmatrix}, \quad (2.121)$$

or

$$[u_j^{n+1}] = \lambda[I][u_j^n], \quad (2.122)$$

where $[I]$ is the identity matrix. Comparing (2.119) and (2.122), we see that

$$([A] - \lambda[I])[u_j^n] = 0, \quad (2.123)$$

and this equation must hold regardless of the values of the u_j^n . It follows that the amplification factors, λ , are the *eigenvalues* of $[A]$, obtained by solving

$$|[A] - \lambda[I]| = 0. \quad (2.124)$$

For the current example, we can show that

$$\lambda = 1 - \mu \left(1 - e^{i \frac{2m\pi}{J}} \right), \quad m = 0, 1, 2, \dots, J-1. \quad (2.125)$$

This has essentially the same form as (2.109), and so we find that $0 \leq \mu \leq 1$ is the stability condition. *The advantage of the matrix method is that the boundary conditions can be directly included in the stability analysis, as in the example above.*

2.9 The effects of increasing the number of grid points

Recall that the upstream scheme can be written as

$$u_j^{n+1} = u_j^n(1 - \mu) + u_{j-1}^n \mu. \quad (2.126)$$

Also recall that for this scheme the amplification factor, λ , is

$$\lambda = 1 - \mu(1 - \cos k\Delta x + i \sin k\Delta x), \quad (2.127)$$

so that

$$|\lambda|^2 = 1 + 2\mu(\mu - 1)(1 - \cos k\Delta x). \quad (2.128)$$

The stability criterion is

$$0 \leq \mu \leq 1. \quad (2.129)$$

When (2.129) is satisfied, we have

$$|\lambda| \leq 1. \quad (2.130)$$

Consider what happens when we increase the number of grid points, while keeping the domain size, D , the wind speed, c , and the wave number, k , of the advected signal the same. We consider grid spacing Δx , such that

$$D = J\Delta x. \quad (2.131)$$

As we decrease Δx , we increase J correspondingly, so that D remains the same, and

$$k\Delta x = \frac{kD}{J}. \quad (2.132)$$

Substituting this into (2.128), we obtain

$$|\lambda|^2 = 1 + 2\mu(\mu - 1) \left[1 - \cos\left(\frac{kD}{J}\right) \right]. \quad (2.133)$$

In order to maintain computational stability, we keep μ fixed as Δx decreases, so

that

$$\begin{aligned}\Delta t &= \frac{\mu \Delta x}{c} \\ &= \frac{\mu D}{cJ}.\end{aligned}\tag{2.134}$$

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

$$T = \frac{D}{c}.\tag{2.135}$$

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

$$\begin{aligned}N &= \frac{T}{\Delta t} \\ &= \frac{D}{c\Delta t} \\ &= \frac{D}{\mu \Delta x} \\ &= \frac{J}{\mu}.\end{aligned}\tag{2.136}$$

The total amount of damping that “accumulates” as the air moves across the domain is measured by

$$\begin{aligned}|\lambda|^N &= (|\lambda|^2)^{N/2} \\ &= \left\{ 1 - 2\mu(1 - \mu) \left[1 - \cos\left(\frac{kD}{J}\right) \right] \right\}^{\frac{J}{2\mu}}.\end{aligned}\tag{2.137}$$

Here we have used (2.133) and (2.136).

As we increase the resolution, J increases. This causes the cosine factor in (2.137) to approach 1, which weakens the damping associated with $|\lambda| < 1$; but on the other hand it also causes the exponent in (2.137) to increase, which strengthens the damping. Which effect dominates is not obvious. The answer can be seen in Fig. 2.8. Increasing the resolution leads to less total damping, even though the number of time steps needed to cross the domain increases.

2.10 Summary

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

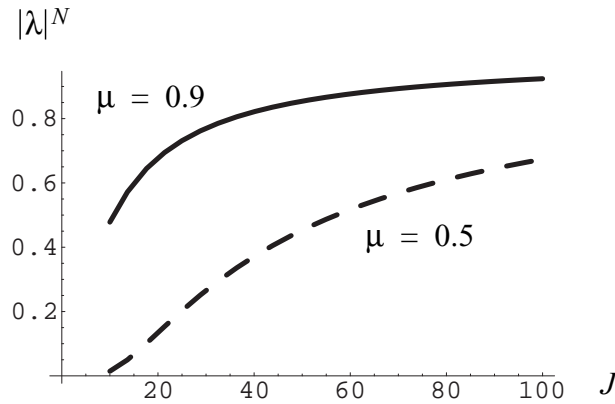


Figure 2.8: “Total” damping experienced by a disturbance crossing the domain, as a function J , the number of grid points across the domain. Here we have assumed that $D/L = 2$.

follows:

- **Check truncation error.** Normally this is done by means of a Taylor series expansion. We are concerned with the lowest powers of the space and time grid-interval in the expansion of the independent variables.
- **Check linear stability** for a simplified (linearized, constant coefficients) version of the equation. Von Neumann's method is often used here.
- **Check nonlinear stability**, if possible. This can be accomplished, in some cases, by using the energy method. Otherwise, empirical tests are needed.

Increased accuracy does not always give 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.

In general, “good” schemes have the following properties, among others:

- High accuracy.
- Stability.
- Simplicity.
- Computational economy.

Later we will extend this list to include additional desirable properties.

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.

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.
2. Choose a simple differentiable function $f(x)$ that is *not* a polynomial. Find the exact numerical value of $\frac{df}{dx}$ at a particular value of x , say x_1 . Then choose
 - a) a first-order scheme, and
 - b) a second-order scheme
 to approximate $\left(\frac{df}{dx}\right)_{x=x_1}$. Plot the log of the absolute value of the *error* of these approximations as a function of the log of Δx . By inspection of the plot, verify that the errors of the schemes decrease, at the expected rates, as Δx decreases.
3. 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. Discuss your results.
4. **Using the energy method**, determine the stability of the **forward** time scheme as applied to the following pair of equations:

$$\frac{du}{dt} = fv, \quad (2.138)$$

$$\frac{dv}{dt} = -fu. \quad (2.139)$$

Note: Solution following the energy method as required does not involve the use of complex variables.

5. Work out the form of the most compact second-order accurate approximation for $\left(\frac{d^2f}{dx^2}\right)_j$ on a non-uniform grid. Also find the simpler form that applies when the grid is uniform.

CHAPTER 3***A Survey of Time-Differencing Schemes
for the Oscillation and Decay Equations***

Copyright 2004 David A. Randall

3.1 Introduction

In atmospheric dynamics, the governing equations are usually non-linear partial differential equations. Some knowledge of finite-difference approximations to ordinary differential equations (especially first order) is needed, however. In fact, if we linearize a governing partial differential equation and assume a wave form for the solution, the equation simply reduces to an ordinary differential equation. An example of this was given in Chapter 2. The stability of a finite difference approximation to such an ordinary differential equation can be examined using von Neumann's method, as explained in Chapter 2. In this Chapter, we deliberately side-step the complexities of space differencing and consider the problem of time differencing in isolation.

Consider an arbitrary first-order ordinary differential equation of the form:

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

Both q and $f(q, t)$ may be complex variables. In the following two subsections, we do not specify $f(q, t)$. Later we will consider particular cases. Keep in mind that $f(q, t)$ could be very complicated.

3.2 Non-iterative schemes.

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. We also assume that $n \geq m$, which may not be true close to the initial condition; this point is considered later. Then we have

$$q[(n + 1)\Delta t] - q[(n - m)\Delta t] = \int_{(n - m)\Delta t}^{(n + 1)\Delta t} f(q, t) dt. \quad (3.2)$$

Equation (3.2) is still “exact;” no finite-difference approximations have been introduced. With a finite difference-scheme, q and, therefore, f are defined only at discrete time levels. Suppose that we approximate the integral on the right-hand side of (3.2) using the values of f at the discrete time levels. We use symbol q^{n+1} in place of $q[(n + 1)\Delta t]$, f^{n+1} in place of

$f\{q[(n+1)\Delta t], (n+1)\Delta t\}$, etc. Equation (3.2), divided by $(1+m)\Delta t$, can be approximated by

$$\frac{q^{n+1} - q^{n-m}}{(1+m)\Delta t} \cong \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1} + \alpha_{n-2} f^{n-2} + \dots + \alpha_{n-l} f^{n-l}, \quad (3.3)$$

where l can be minus one (for $\beta \neq 0$ only), zero, or a positive integer. Take a minute to look carefully at the form of (3.3), 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 $(m+1)\Delta t$, 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 . We get to choose l and m when we design a scheme. A *family of schemes* is defined by (3.3). It is possible to have $l > m$ or $l < m$ or $l = m$. Viable schemes can be constructed with all three possibilities, and examples will be given below.

If $\beta \neq 0$ the scheme is called “implicit,” and if $\beta = 0$ it is called “explicit.” Implicit schemes have nice properties, as discussed later, but they 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 the argument of f^{n+1} . Examples will be shown later.

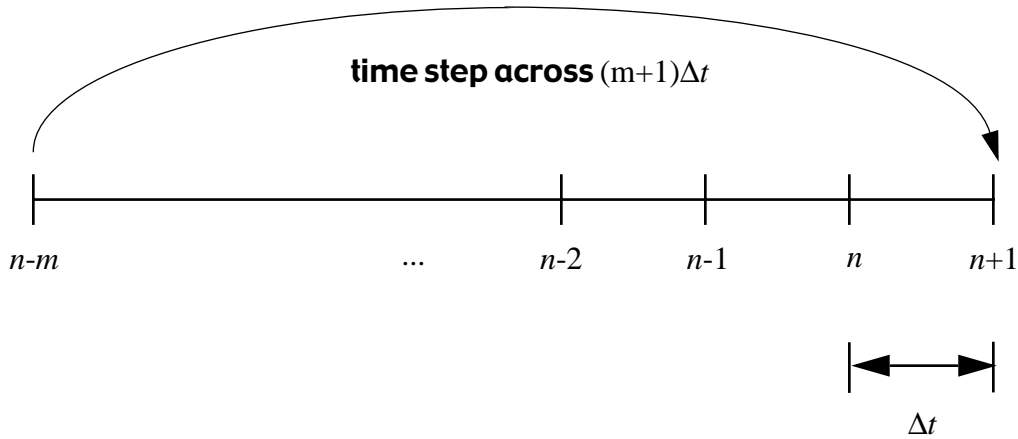


Figure 3.1: In Eq. (3.3), we use a weighted combination of $f^{n+1}, f^n, f^{n-1}, \dots, f^{n-l}$ to compute an “average” value of $f \equiv \frac{dq}{dt}$ over the time interval $(m+1)\Delta t$.

The smallest possible value of l is -1, in which case only time level $n+1$ appears on

the right-hand side of (3.3); this is the case with the backward-implicit scheme, for which $\beta \neq 0$. The Euler forward scheme uses time level n , so that $l = 0$. Note, however, that the Euler forward scheme omits time level $n + 1$; it is an explicit scheme with $\beta = 0$. There are many (in principle, infinitely many) other possibilities, as will be discussed later in this Chapter.

Now substitute the *true solution* $q(t)$ and corresponding $f[q(t), t]$ into (3.3), and expand into a Taylor series around $t = n\Delta t$. We get

$$\begin{aligned}
 & \frac{1}{(1+m)\Delta t} \left\{ \left(q + \Delta t q' + \frac{\Delta t^2}{2!} q'' + \frac{\Delta t^3}{3!} q''' + \frac{\Delta t^4}{4!} q'''' + \dots \right) \right. \\
 & \quad \left. - \left[q - (m\Delta t)q' + \frac{(m\Delta t)^2}{2!} q'' - \frac{(m\Delta t)^3}{3!} q''' + \frac{(m\Delta t)^4}{4!} q'''' - \dots \right] \right\} \\
 &= \beta \left[f + \Delta t f' + \frac{\Delta t^2}{2!} f'' + \frac{\Delta t^3}{3!} f''' + \dots \right] \\
 &+ \alpha_n f \\
 &+ \alpha_{n-1} \left[f - \Delta t f' + \frac{\Delta t^2}{2!} f'' - \frac{\Delta t^3}{3!} f''' + \dots \right] \\
 &+ \alpha_{n-2} \left[f - 2\Delta t f' + \frac{(2\Delta t)^2}{2!} f'' - \frac{(2\Delta t)^3}{3!} f''' + \dots \right] \\
 &+ \alpha_{n-3} \left[f - 3\Delta t f' + \frac{(3\Delta t)^2}{2!} f'' - \frac{(3\Delta t)^3}{3!} f''' + \dots \right] \\
 &+ \dots \\
 &+ \alpha_{n-l} \left[f - l\Delta t f' + \frac{(l\Delta t)^2}{2!} f'' - \frac{(l\Delta t)^3}{3!} f''' + \dots \right] \\
 &+ \varepsilon, \tag{3.4}
 \end{aligned}$$

where ε is the truncation error and a prime denotes a time derivative. Collecting powers of Δt , and using $q' = f$, $q'' = f'$, etc., we obtain

$$\begin{aligned}
& q'[1-(\beta + \alpha_n + \alpha_{n-1} + \alpha_{n-2} + \alpha_{n-3} + \dots + \alpha_{n-l})] \\
& + \Delta t q'' \left\{ \frac{1}{2} \frac{1-m^2}{1+m} - \beta + \alpha_{n-1} + 2\alpha_{n-2} + 3\alpha_{n-3} + \dots + l\alpha_{n-l} \right\} \\
& + \frac{(\Delta t)^2}{2!} q''' \left\{ \frac{1}{3} \frac{1+m^3}{1+m} - \beta - \alpha_{n-1} - 4\alpha_{n-2} - 9\alpha_{n-3} - \dots - l^2\alpha_{n-l} \right\} \\
& + \frac{(\Delta t)^3}{3!} q'''' \left\{ \frac{1}{4} \frac{1-m^4}{1+m} - \beta + \alpha_{n-1} + 8\alpha_{n-2} + 27\alpha_{n-3} + \dots + l^3\alpha_{n-l} \right\} \\
& + \dots \\
& = \epsilon .
\end{aligned} \tag{3.5}$$

Each line on the left-hand side of (3.5) goes to zero “automatically” as $\Delta t \rightarrow 0$, except for the first line, which does not involve Δt at all. We have to force the first line to be zero, because otherwise the error, ϵ , will not go to zero as $\Delta t \rightarrow 0$. In order to force the first line to be zero, we have to choose β and the various α ’s in such a way that

$$1 = \beta + \alpha_n + \alpha_{n-1} + \alpha_{n-2} + \alpha_{n-3} + \dots + \alpha_{n-l} . \tag{3.6}$$

Note that (3.6) simply means that the sum of the coefficients on the right-hand side of (3.3) is equal to one, so that the right-hand side is a kind of “average f .” We refer to (3.6) as the “*consistency condition*.” When (3.6) is satisfied, the expression for the truncation error reduces to

$$\epsilon = \Delta t q'' \left\{ \frac{1}{2} \left(\frac{1-m^2}{1+m} \right) - \beta + \alpha_{n-1} + 2\alpha_{n-2} + 3\alpha_{n-3} + \dots + l\alpha_{n-l} \right\} + O[(\Delta t)^2] . \tag{3.7}$$

This shows that when (3.6) is satisfied the scheme has at least first order accuracy, i.e. the scheme is consistent, and the error, ϵ , goes to zero at least as fast as Δt . *Note, however, that we are still free to choose $l+1$ coefficients.* Moreover, the value of l itself is under our control; we can choose it when we design the scheme. If $l \geq 0$, then we can choose the coefficients in such a way that the coefficient of Δt in (3.5) is also zero. This will give us a second-order scheme, i.e. one in which the error, ϵ , goes to zero like $(\Delta t)^2$. Obviously this process can be continued, giving higher and higher accuracy, so 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 β . Recall that $\beta = 0$ for explicit schemes. Generally, then, the accuracy of an explicit

scheme can be made at least as high as $l + 1$. Later we refer back to these rules of thumb.

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. This will be explained later.

We now survey a number of time-differencing schemes, without specifying f . In this analysis, we can determine the order of accuracy of each scheme. We cannot decide whether a scheme is stable or unstable, however, unless f is specified. Later in this Chapter we investigate what happens with two particular choices of f , and later in the course we will consider additional choices for f .

3.2.1 Explicit schemes ($\beta = 0$)

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

For this case we have $\alpha_n \neq 0$, and all of the other α 's are zero. The consistency condition, (3.6), immediately forces us to choose $\alpha_n = 1$. The scheme (3.3) then reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = f^n \quad (3.8)$$

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

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

Better accuracy can be obtained by proper choice of the α 's, if we use $l > 0$. For example, consider the case $l = 1$. The scheme reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = \alpha_n f^n + \alpha_{n-1} f^{n-1}, \quad (3.9)$$

the consistency condition, (3.6), reduces to

$$\alpha_n + \alpha_{n-1} = 1, \quad (3.10)$$

and the truncation error is

$$\varepsilon = \Delta t q'' \left(\alpha_{n-1} + \frac{1}{2} \right) + O(\Delta t^2). \quad (3.11)$$

If we choose $\alpha_{n-1} = -\frac{1}{2}$, the scheme has second-order accuracy. Of course, this means that $\alpha_n = \frac{3}{2}$. This is the second-order Adams-Bashforth scheme.

Although the right-hand side of (3.9) 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 later use on the next time step. We have to allocate additional memory in the computer 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 only, since when $n = 0$ time level $n - 1$ is “before the beginning” of the computation.

In a similar way, we can obtain Adams - Bashforth schemes with higher accuracy by using larger l , and choosing the α 's accordingly. The table below 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 $l = 0$.

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

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

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

The leap-frog scheme is

$$\frac{1}{2\Delta t}(q^{n+1} - q^{n-1}) = f^n \quad (3.12)$$

From (3.5) we can immediately see that the truncation error is $\frac{\Delta t^2}{6}q''' + O(\Delta t^4)$.

Note that 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.

$$m = 1, l = 1$$

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

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

We can increase the order of accuracy by choosing appropriate α 's if $l > 1$.

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

3.2.2 Implicit schemes ($\beta \neq 0$)

Here we should be able to achieve accuracy at least as high as $l + 2$. Note that with implicit schemes it is possible to have $l = -1$, whereas with the explicit schemes the smallest allowed value of l is 0.

$$m = 0, l = 0$$

Eq. (3.3) reduces to

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta f^{n+1} + \alpha_n f^n. \quad (3.13)$$

The consistency condition reduces to $\alpha_n + \beta = 1$. The truncation error is $\Delta t q'' \left(\frac{1}{2} - \beta \right) + O(\Delta t^2)$. When $\beta = 1, \alpha_n = 0$, 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.

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

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

summarizes the properties of these schemes, for $l = 1, 2$, and 3 .

Table 3.2: Adams–Moulton schemes.

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

$m = 1, l = 0$

Highest accuracy (2nd order) is obtained for $\beta = 0$, which gives the leapfrog scheme.

$m = 1, l = 1$ (Milne corrector¹)

Eq. (3.3) reduces to

$$\frac{q^{n+1} - q^{n-1}}{2\Delta t} = \beta f^{n+1} + \alpha_n f^n + \alpha_{n-1} f^{n-1}, \quad (3.14)$$

where

$$\beta + \alpha_n + \alpha_{n-1} = 1. \quad (3.15)$$

The truncation error is

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

From this we can see that $\beta = \frac{1}{6}$, $\alpha_n = \frac{4}{6}$, $\alpha_{n-1} = \frac{1}{6}$ gives fourth-order accuracy. This is again more than would be expected from the general rule.

¹. If there is a “Milne corrector,” then there must be a “Milne predictor.” (See subsection 3.3 for an explanation of this terminology.) In fact, the Milne predictor is an explicit scheme with $m = 3$, $l = 3$, and $\alpha_n = \frac{2}{3}$, $\alpha_{n-1} = -\frac{1}{3}$, $\alpha_{n-2} = \frac{2}{3}$, $\alpha_{n-3} = 0$.

$$m = 1, l = 2$$

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.

3.3 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.”

The advantage of iterative schemes is that we can gain higher accuracy. The disadvantage is computational expense, because each evaluation of f involves doing a certain amount of arithmetic. In contrast, non-iterative schemes such as those discussed in the preceding subsection involve only a single evaluation of f for each time step. For this reason, iterative schemes tend to be computationally more expensive than non-iterative schemes, for a given order of accuracy.

Consider (3.13) as an example. Replace $f^{n+1} \equiv f[q^{n+1}, (n+1)\Delta t]$ by $f^{n+1*} \equiv f[q^{n+1*}, (n+1)\Delta t]$, where q^{n+1*} is obtained by the Euler scheme,

$$\frac{q^{n+1*} - q^n}{\Delta t} = f^n. \quad (3.16)$$

Then

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta^* f^{n+1*} + \alpha f^n. \quad (3.17)$$

When $\beta^* = 1$, $\alpha = 0$, Eq. (3.17) is an imitation of the backward (implicit) difference scheme, and is called the Euler-backward scheme or the Matsuno scheme (Matsuno, 1966).

When $\beta^* = \frac{1}{2}$, $\alpha = \frac{1}{2}$, Eq. (3.17) is an imitation of the trapezoidal (implicit) scheme and is called the Heun scheme or the second-order Runge-Kutta scheme. The Matsuno scheme has first-order accuracy, and the Heun scheme has second-order accuracy.

Note that (3.17) cannot be “fit” into the framework of (3.3), because f^{n+1*} does not appear on the right-hand side of (3.3), and in general f^{n+1*} cannot be written as a combination of f^{n-l} ’s.

Also note that the Heun scheme is explicit, and does not require the past history (does

not require $l > 0$). Still, 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. This is an excellent scheme when f has a simple form, but it is not economically practical when f is complicated. The scheme is given by:

$$\begin{aligned}
 q^{n+1} &= q^n + \Delta t \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4), \\
 k_1 &= f(q^n, n\Delta t), & k_2 &= f\left[q^n + \frac{k_1\Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right], \\
 k_3 &= f\left[q^n + \frac{k_2\Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right], & k_4 &= f(q^n + k_3\Delta t, (n+1)\Delta t).
 \end{aligned} \tag{3.18}$$

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.

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 has fourth-order accuracy.

3.4 Finite-difference schemes applied to the oscillation equation

In order to test the stability of the finite-difference schemes discussed in the previous section, we must specify the form of the function $f(q, t)$. As a first example, consider the oscillation equation:

$$\frac{dq}{dt} = i\omega q, \quad q \text{ complex, } \omega \text{ real.} \tag{3.19}$$

The exact solution is $q = \hat{q}(t)e^{i\omega t}$, where \hat{q} is the "initial" value of q , at $t = 0$. The amplitude \hat{q} is an invariant of this system, i.e.,

$$\frac{d\hat{q}}{dt} = 0. \tag{3.20}$$

The following are examples of more familiar equations which are reducible to (3.19).

- **Advection:** The governing equation is $\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$. If $u = \hat{u}(t)e^{ikx}$, then


```

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 example to illustrate how the fourth-order Runge-Kutta scheme works. Note the four calls to subroutine "dot." This makes the scheme expensive.

$\frac{d\hat{u}}{dt} = -ikc\hat{u} = i\omega\hat{u}$, with $\omega \equiv -kc$. An observer at a point watching a *single Fourier mode* advect by will see an oscillation.

- **Rotation:** Pure inertial motion is described by

$$\frac{du}{dt} - fv = 0, \quad (3.21)$$

$$\frac{dv}{dt} + fu = 0. \quad (3.22)$$

Multiplying (3.22) by i and adding it to (3.21), we obtain

$$\frac{d}{dt}(u + iv) + f(-v + iu) = 0. \quad (3.23)$$

With $q = u + iv$, we get

$$\frac{dq}{dt} + ifq = 0, \quad (3.24)$$

which is identical to (3.21) with $\omega = -f$. Note that, although u and v are real, q is complex, and $|q|^2 = u^2 + v^2$ is twice the kinetic energy per unit mass. We can obtain $\frac{d}{dt}|q|^2 = 0$, i.e., kinetic energy conservation, directly from (3.21) and (3.22). By differentiating (3.21) with respect to time, and substituting from (3.22) for $\frac{dv}{dt}$, we obtain $\frac{d^2u}{dt^2} = -f^2u$, perhaps a more familiar form of the oscillation equation, which in this case describes a pure inertial oscillation.

In principle, any of the schemes described earlier in this chapter can be applied to (3.19). Each scheme has its own properties, as discussed below.

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

Write a finite difference analog of (3.19) as follows

$$q^{n+1} - q^n = i\omega\Delta t(\alpha q^n + \beta q^{n+1}). \quad (3.25)$$

We require $\alpha + \beta = 1$ in order to guarantee consistency. We obtain the Euler scheme with $\alpha = 1$, $\beta = 0$; the backward scheme with $\alpha = 0$, $\beta = 1$; and the trapezoidal-implicit scheme with $\alpha = \beta = \frac{1}{2}$. Eq. (3.25) can easily be solved for q^{n+1} :

$$(1 - i\Omega\beta)q^{n+1} = (1 + i\Omega\alpha)q^n, \quad (3.26)$$

or

$$q^{n+1} = \left(\frac{1 + i\Omega\alpha}{1 - i\Omega\beta} \right) q^n \equiv \lambda q^n, \quad (3.27)$$

where we introduce the shorthand notation $\Omega \equiv \omega\Delta t$. In (3.27), λ is the amplification factor.

We want to know how the amplitude $|q|$ behaves in time. Recall that $|q|$ is invariant for the differential equation. This means that, for the true solution, $|\lambda| = 1$. If the computed $|\lambda|$ is not equal to one, we have “amplitude errors.”

Since λ is complex, we write

$$\lambda = \lambda_r + i\lambda_i = |\lambda|e^{i\theta}, \quad \text{where} \quad \tan\theta = \frac{\lambda_i}{\lambda_r}, \lambda_r = |\lambda|\cos\theta, \lambda_i = |\lambda|\sin\theta. \quad (3.28)$$

We can interpret θ as the phase change of the oscillation per time step. Positive θ denotes counterclockwise 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 λ is pure imaginary. The discrete numerical solution may look as shown schematically in Fig. 3.3 for the case of $\theta = \frac{\pi}{2}$; the

ordinate represents the imaginary part of q^n . Note that for the exact solution the phase change over the time interval Δt is $\Omega \equiv \omega\Delta t$. Generally θ (the computed phase change) and Ω (the true or exact phase change) will differ because of discretization errors. If the computed θ is not equal to Ω , we have “phase errors.”

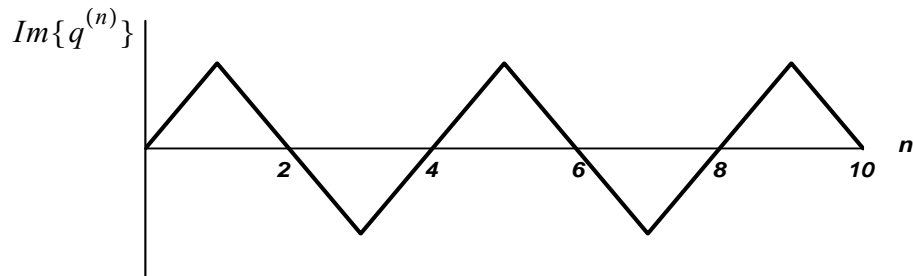


Figure 3.3: Schematic illustration of the solution of the oscillation equation for the case in which λ is pure imaginary and the phase changes by $\pi/2$ on each time step.

The true solution to (3.19) is $q \propto e^{i\omega t}$, so that

$$q[(n+1)\Delta t] = e^{i\omega\Delta t} q(n\Delta t) = e^{i\Omega} q(n\Delta t). \quad (3.29)$$

Now $|e^{i\Omega}| = 1$, and so $\lambda = |\lambda|e^{i\theta}$ corresponds to (“should be”) $e^{i\Omega}$; in other words, $|\lambda|$ should be equal to 1 and θ should be equal to Ω . We want to compare the computed amplification factor $|\lambda|$ with the true one, which is $|e^{i\Omega}| = 1$, and the computed phase change per time step, θ , with the true one, which is Ω . We will examine both the phase error and the amplitude error.

For the forward (Euler) scheme, $\alpha = 1$, $\beta = 0$, and so from (3.27) we find that

$$\lambda = 1 + i\Omega. \quad (3.30)$$

Therefore,

$$|\lambda| = \sqrt{1 + \Omega^2} > 1. \quad (3.31)$$

We conclude that, for the oscillation equation, the forward scheme is *unconditionally unstable*. We have reason to *suspect*, therefore, that forward time differencing is not a good choice for the advection or coriolis terms of a dynamical model. In reality, whether or not the forward scheme is a good choice for the advection terms depends on the space-differencing scheme used, as will be discussed in Chapter 4. From (3.30) we see that the phase change per time step, θ , satisfies $\tan \theta = \Omega$, so that $\theta \cong \Omega$ for small Δt , as expected.

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

$$\lambda = \frac{1}{1 - i\Omega} = \frac{1 + i\Omega}{1 + \Omega^2}, \quad (3.32)$$

so that

$$|\lambda| = \frac{\sqrt{1 + \Omega^2}}{1 + \Omega^2} = \frac{1}{\sqrt{1 + \Omega^2}} < 1. \quad (3.33)$$

This scheme is, therefore, unconditionally stable, and in fact the amplitude of the oscillation decreases with time. The real part of λ is always positive, which means that $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. This scheme can be used for the coriolis terms in a model, but because of the damping it is not a very good choice. The phase change per time step again satisfies $\tan \theta = \Omega$, so again the phase error is small for small Δt .

For the trapezoidal implicit scheme, $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}$, we find that

$$|\lambda| = \frac{\left|1 + \frac{i\Omega}{2}\right|}{\left|1 - \frac{i\Omega}{2}\right|} = 1. \quad (3.34)$$

This scheme is *unconditionally stable*; in fact, it has no amplitude error at all. Its phase error per time step is small. It is a nice scheme.

3.4.2 Iterative two-level schemes for the oscillation equation

Now consider a finite-difference analogue of (3.19) given by

$$q^{n+1*} - q^n = i\Omega q^n, \quad (3.35)$$

$$q^{n+1} - q^n = i\Omega(\alpha q^n + \beta^* q^{n+1*}). \quad (3.36)$$

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 (3.35) and (3.36) for the Matsuno scheme, we find that

$$\lambda = (1 - \Omega^2) + i\Omega,$$

$$|\lambda| = \sqrt{1 - \Omega^2 + \Omega^4} \begin{cases} > 1 \text{ for } \Omega > 1 \\ = 1 \text{ for } \Omega = 1 \\ < 1 \text{ for } \Omega < 1 \end{cases} \quad (3.37)$$

This is, therefore, a *conditionally stable* scheme (the condition is $\Omega \leq 1$).

For the Heun scheme, we obtain

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

$$|\lambda| = \sqrt{\left(1 - \frac{\Omega^2}{2}\right)^2 + \Omega^2} = \sqrt{1 + \frac{\Omega^4}{2}} > 1. \quad (3.38)$$

This scheme is *unconditionally unstable*, but notice that for small Ω it is not as unstable as the forward scheme. In fact, it can be used with some success if very long-term integrations are not required.

The results discussed above are summarized in Fig. 3.4 and Fig. 3.5.

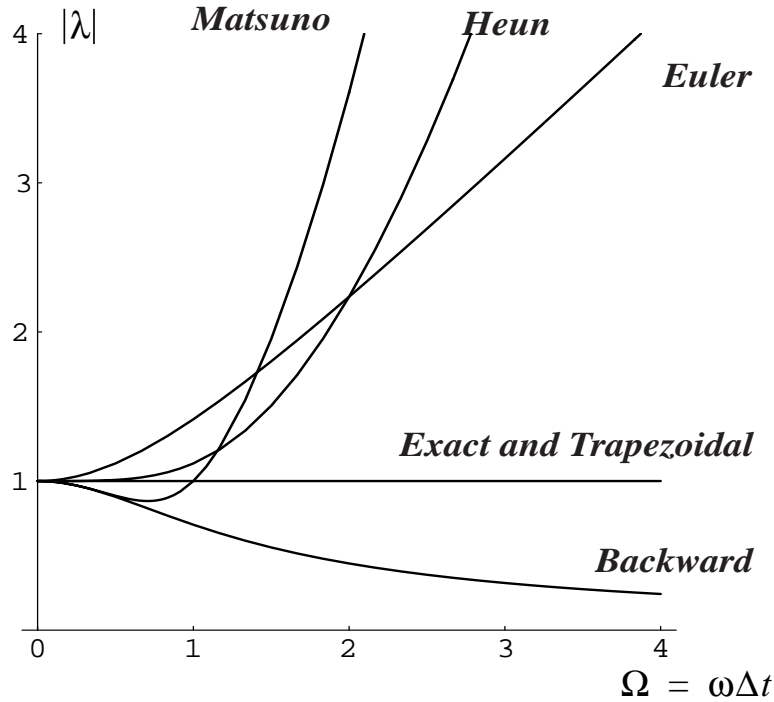


Figure 3.4: This figure shows the magnitude of the amplification factor as a function of $\Omega \equiv \omega\Delta t$ for various difference schemes. The Euler, backward, trapezoidal, Euler-backward, and Heun schemes are shown by curves I, II, III, IV, and V respectively. The magnitude of the amplification factor for the trapezoidal scheme coincides with that of the true solution for all values of $\omega\Delta t$. Caution: This does not mean that the trapezoidal scheme gives the exact solution!

3.4.3 The leapfrog scheme for the oscillation equation

The leapfrog scheme, which is illustrated in Fig. 3.6, is not “self-starting” at $n = 0$, because we do not know q at $n = 1$. A special procedure must, therefore, be used to start the solution, i.e. we must somehow determine the value of q at $n = 1$. 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 that we need for the differential equation. The other arises because of the form of the finite-difference scheme itself, and has nothing to do with the physics. It is usually referred to as the “computational” initial condition.

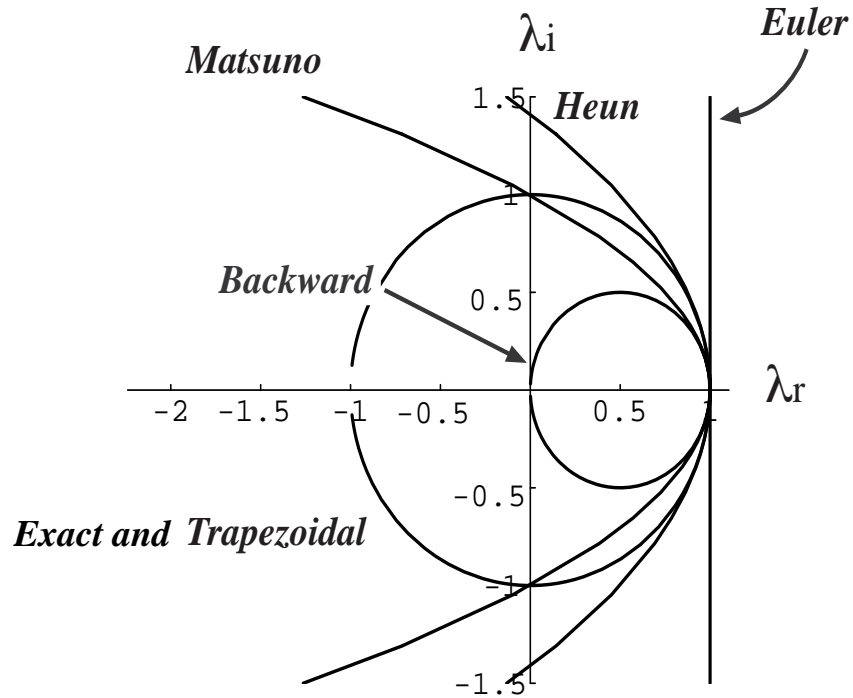


Figure 3.5: This figure shows the behavior of the imaginary (λ_i) and real (λ_r) components of the amplitude, as Ω varies. Recall that $\tan \theta$ is given by $\frac{\lambda_i}{\lambda_r}$ for each scheme.

From this plot we can also see the behavior of $|\lambda|$ as θ varies, for each scheme.

Consider the leapfrog analogue to (3.19):

$$q^{n+1} - q^{n-1} = 2\Delta t i \omega q^n. \quad (3.39)$$

For the simple case $\omega=0$, we obtain

$$q^{n+1} - q^{n-1} = 0. \quad (3.40)$$

Of course, $q = \text{constant}$ is the true solution of $\frac{dq}{dt} = 0$, but according to (3.39) the solution will depend on the initial conditions given at both the levels $n = 0$ and $n = 1$. If these two values are different, an oscillation will occur, as shown in Fig. 3.7. We have $q^2=q^0$, $q^3=q^1$, $q^4=q^2=q^0$, etc. If we assign $q^1 = q^0$, the solution will be constant. This example illustrates that judicious selection of $q^{(1)}$ is essential for schemes with more than two time levels.

Now rewrite (3.39) as

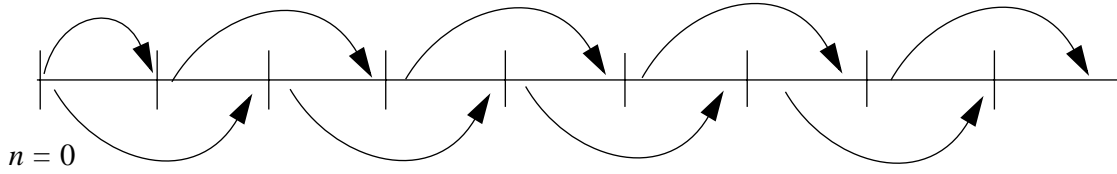


Figure 3.6: The leapfrog scheme.

$$q^{n+1} - 2i\Omega q^n - q^{n-1} = 0, \quad (3.41)$$

where, as before, $\Omega \equiv \omega \Delta t$. We look for a solution of the form $q^{n+1} = \lambda q^n$, for all n . Then (3.41) reduces to

$$\lambda^2 - 2i\Omega\lambda - 1 = 0. \quad (3.42)$$

The solutions of (3.42) are

$$\lambda_1 = i\Omega + \sqrt{1 - \Omega^2}, \quad \lambda_2 = i\Omega - \sqrt{1 - \Omega^2}, \quad (3.43)$$

giving two solutions or “modes”,

$$q_1^{n+1} = \lambda_1 q_1^n, \quad q_2^{n+1} = \lambda_2 q_2^n. \quad (3.44)$$

The differential equation only has one solution, so getting two solutions to the finite-difference equation is bad. Consider the limits of λ_1 and λ_2 as $\Omega \rightarrow 0$ or $\Delta t \rightarrow 0$. Notice that $\lambda_1 \rightarrow 1$, while $\lambda_2 \rightarrow -1$. We know that for the true solution $\lambda = 1$, and so we can identify q_1 as the “physical” mode, and q_2 as the computational mode. Notice that q_2^{n+1} generally does not approach q_2^n as $\Delta t \rightarrow 0$! This illustrates that simply reducing the time step does not reduce problems associated with computational modes.

The computational mode arises from the three-level scheme. Two-level schemes do not have computational modes. Schemes with more than three time levels have multiple computational modes. *The existence of computational modes is a major disadvantage of all schemes that involve more than two time levels.* The current discussion is about computational modes in time. Later we will see that computational modes can also occur in space, in at least two distinct ways.

From (3.44) we have

$$q_1^n = \lambda_1^n q_1^0, \quad q_2^n = \lambda_2^n q_2^0. \quad (3.45)$$

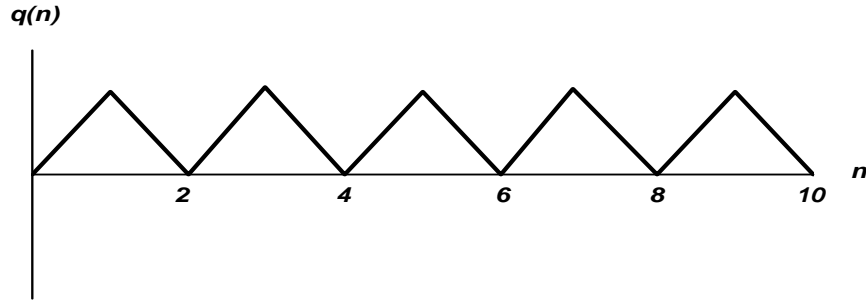


Figure 3.7: An oscillatory solution that arises with the leapfrog scheme for $dq/dt = 0$, for the case in which the two initial values of q are not the same.

The general solution is a linear combination of these two modes

$$q^n = a\lambda_1^n q_1^0 + b\lambda_2^n q_2^0. \quad (3.46)$$

Applying initial conditions, we have

$$(n = 0) \quad aq_1^0 + bq_2^0 = q^0, \quad (3.47)$$

and

$$(n = 1) \quad \lambda_1 a q_1^0 + \lambda_2 b q_2^0 = q^1. \quad (3.48)$$

If we solve (3.47) and (3.48) for aq_1^0 and bq_2^0 in terms of q^0 and q^1 , and substitute the results into (3.46), we obtain

$$q^n = \left(\frac{1}{\lambda_1 - \lambda_2} \right) [(q^1 - \lambda_2 q^0) \lambda_1^n - (q^1 - \lambda_1 q^0) \lambda_2^n]. \quad (3.49)$$

This shows that $\left[\frac{q^1 - \lambda_2 q^0}{\lambda_1 - \lambda_2} \right]$ and $-\left[\frac{q^1 - \lambda_1 q^0}{\lambda_1 - \lambda_2} \right]$ are the initial values of the physical and computational modes, respectively. Which predominates in the numerical solution will, therefore, depend on how we specify q^0 . If we give q^1 such that $q^1 = \lambda_1 q^0$, we will have the physical mode only. In real cases of interest, this is usually impossible to arrange. Notice that (3.49) applies for any choice of $f(q, t)$; it is not specific to the oscillation equation.

A simple way to give q^1 is by the forward (Euler) difference scheme. A more

sophisticated procedure is use of the forward scheme to $n = \frac{1}{2}$, followed by computation of q^1 using the leapfrog scheme. The second method gives a smaller amplitude for the computational mode.

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

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

In this case $\sqrt{1 - \Omega^2}$ in (3.43) is real, and we obtain $|\lambda_1| = |\lambda_2| = 1$. This means that both modes -- the physical and the computational -- are stable and neutral. Let the phase changes per time step be denoted by θ_1 and θ_2 for the physical and computational modes, respectively. Then

$$\lambda_1 = e^{i\theta_1}, \quad \lambda_2 = e^{i\theta_2}. \quad (3.50)$$

Comparing (3.50) with (3.43), we find that

$$\begin{aligned} \cos \theta_1 &= \sqrt{1 - \Omega^2}, & \cos \theta_2 &= -\sqrt{1 - \Omega^2}, \\ \sin \theta_1 &= \Omega, & \sin \theta_2 &= \Omega. \end{aligned} \quad (3.51)$$

Note that we can put $\theta_2 = \pi - \theta_1$. For simplicity of notation, let $\theta \equiv \theta_1$, so that $\theta_2 = \pi - \theta$. When Ω is small, $\theta_1 \cong \Omega$, and $\theta_2 \cong \pi$. The apparent frequency of the physical mode is $\frac{\theta_1}{\Delta t}$, which is approximately equal to ω . Then we can write

$$q_1^{n+1} = e^{i\theta} q_1^n \quad (3.52)$$

for the physical mode, and

$$q_2^{n+1} = e^{i(\pi - \theta)} q_2^n \quad (3.53)$$

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

$$q[(n+1)\Delta t] = e^{i\Omega} q(n\Delta t). \quad (3.54)$$

In the limit as $\Delta t \rightarrow 0$ (i.e. $\Omega \rightarrow 0$), we have $\theta \rightarrow \Omega$. Panels a and b of Fig. 3.8 respectively show λ_1 and λ_2 in the complex λ -plane. The figures have been drawn for the case of

$\theta = \frac{\pi}{8}$. The absolute value of λ is, of course, always equal to 1. Panel c of Fig. 3.8 shows the

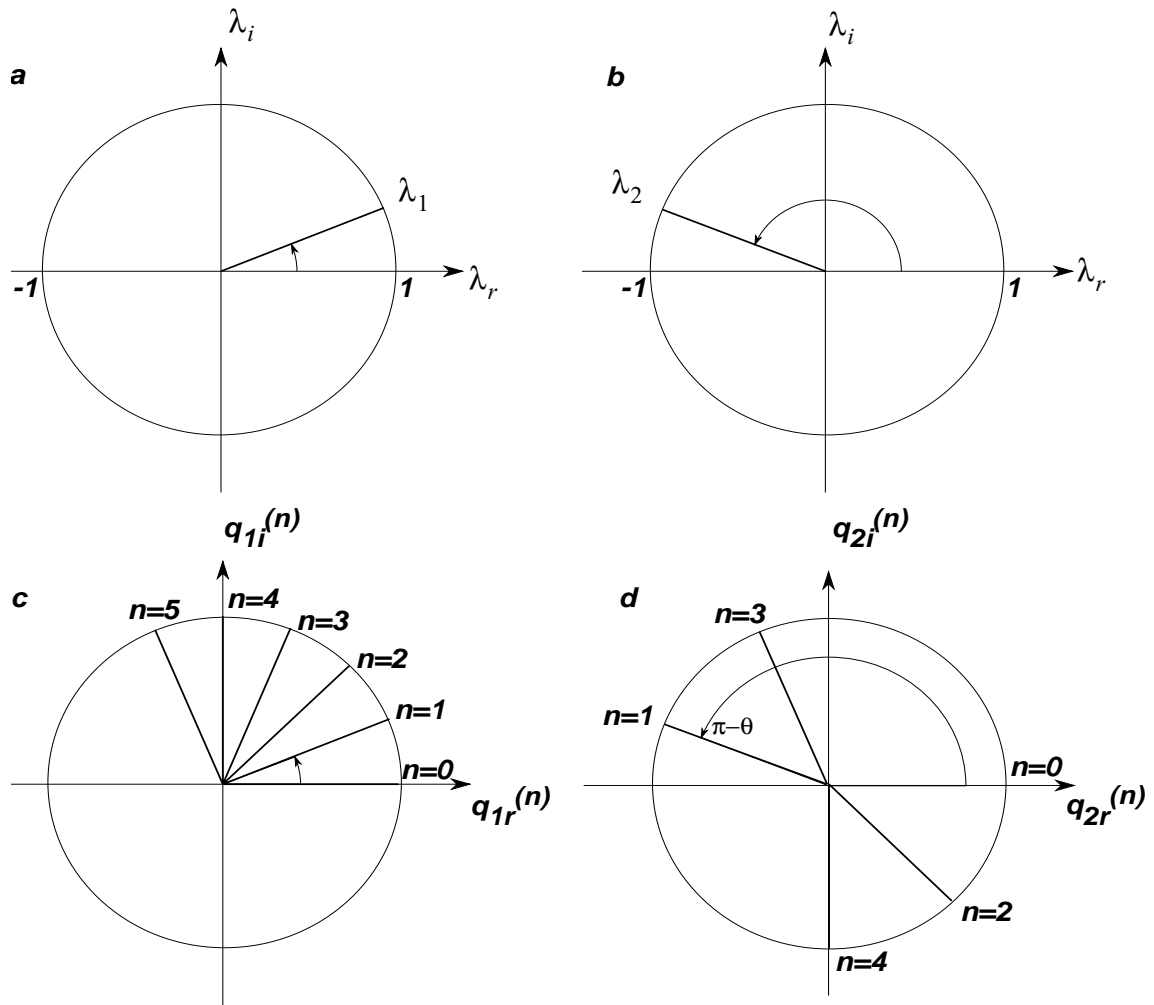


Figure 3.8: Panels a and b: Amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| < 1$. Panels c and d: Solutions of the oscillation equation as obtained with the leapfrog scheme for $|\Omega| < 1$. In making these figures it has been assumed that $\theta = \frac{\pi}{8}$.

graph of the real part of q_1^n versus its imaginary part. Recall that $q_1^n = \lambda_1^n q_1^0 = e^{in\theta} q_1^0$. The

graph is drawn for the case of $q_{1i}^0 = 0$ and $\theta = \frac{\pi}{8}$. Panel d of Fig. 3.8 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 Fig. 3.9. The physical mode looks nice. The computational mode is ugly.

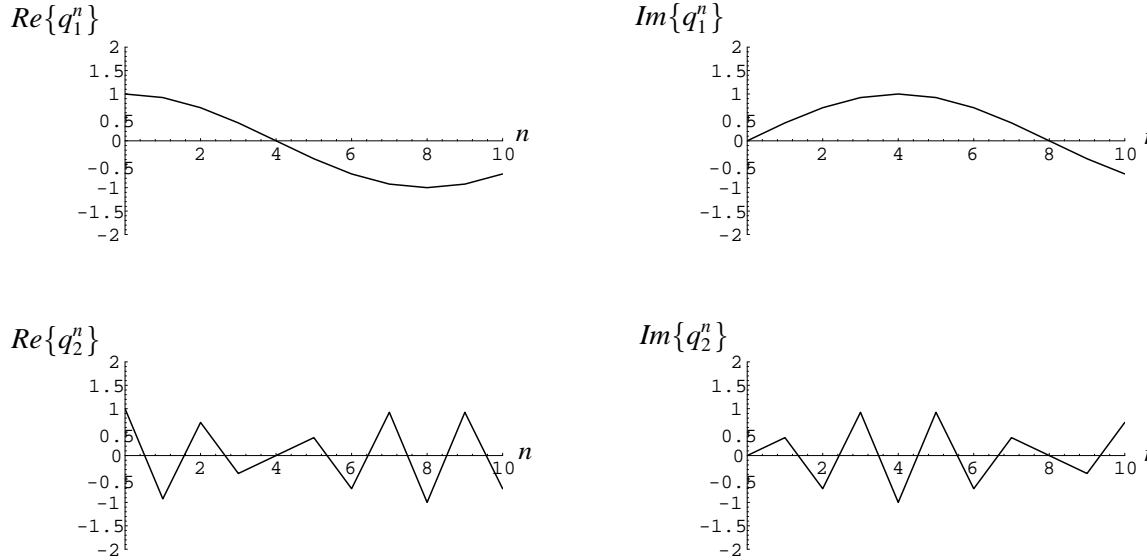


Figure 3.9: Graphs of the real and imaginary parts of the physical and computational modes of the solution of the oscillation equation as obtained with the leapfrog scheme for $|\Omega| < 1$.

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

Here $\lambda_1 = \lambda_2 = i\Omega$ [see (3.43)], i.e. both λ 's are pure imaginary, as shown in Fig. 3.10. This means that, as shown in Fig. 3.10, both solutions rotate through $\frac{\pi}{2}$ on each time step, so that the period is $4\Delta t$, regardless of the true value of ω ! Obviously, $|\lambda_1| = |\lambda_2| = 1$, so both modes are neutral. The phase errors are very large, however. This illustrates a simple fact that should be remembered: A scheme that is stable but on the verge of instability is usually subject to large truncation 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!

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

Here again both λ_1 and λ_2 are pure imaginary, so again both solutions rotate by $\frac{\pi}{2}$ on each time step. We find that

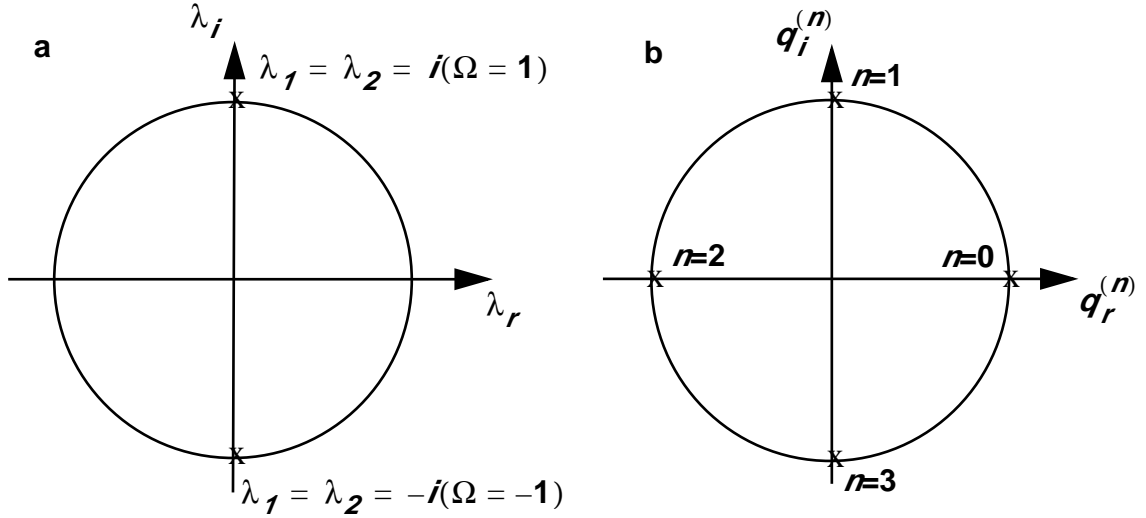


Figure 3.10: Panel a shows the amplification factors for the leapfrog scheme as applied to the oscillation equation with $|\Omega| = 1$. Panel b shows the real and imaginary parts of the corresponding solution, for $n=0, 1, 2$, and 3.

$$\lambda_1 = i(\Omega + \sqrt{\Omega^2 - 1}) \quad \text{and} \quad \lambda_2 = i(\Omega - \sqrt{\Omega^2 - 1}). \quad (3.55)$$

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

A graphical representation of λ in the complex plane, for $|\Omega| > 1$, is shown in panels a and b of Fig. 3.11. Note that $\lambda_1 = \left| \Omega + \sqrt{\Omega^2 - 1} \right| e^{i\frac{\pi}{2}}$ and $\lambda_2 = \left| \Omega - \sqrt{\Omega^2 - 1} \right| e^{i\frac{\pi}{2}}$.

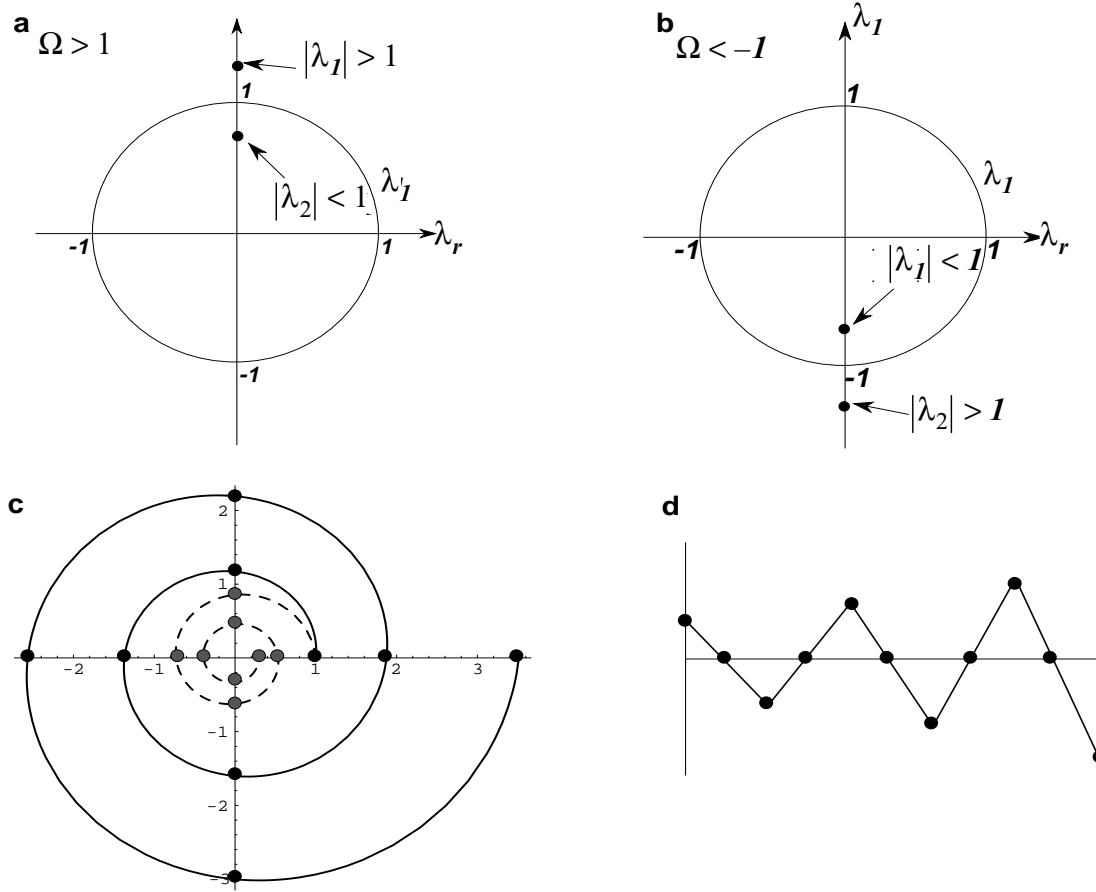


Figure 3.11: Panels a and b show the amplification factors for the oscillation equation with the leapfrog scheme, with $|\Omega| > 1$. Panel c shows the corresponding solution. The solid curve shows the unstable mode, which is actually defined only at the black dots. The dashed curve shows the damped 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.

Panel c of Fig. 3.11 shows a plot of $q^{(n)}$ in the complex plane for the modes corresponding to λ_I and λ_2 for $\Omega > 1$. The phase changes by $\frac{\pi}{2}$ on each step, because λ is pure imaginary,

and so the period is $4\Delta t$. Panel d of Fig. 3.11 schematically shows $q^{(n)}$ as a function of n for the amplifying mode corresponding to λ_I , i.e., q_1 is unstable and q_2 is damped. A growing oscillation of this period is a telltale sign of instability with this type of scheme.

In summary, the centered or leapfrog scheme is a second-order scheme that gives a neutral solution for (3.35) when $|\Omega| \leq 1$. For $|\Omega| > 1$, or large Δt , the scheme is unstable. In other words, the leapfrog scheme is conditionally stable when applied to the oscillation

equation.

We have identified another neutral scheme -- the trapezoidal implicit scheme -- but to use such an implicit scheme in more complicated nonlinear problems is relatively difficult, in comparison with an explicit scheme. 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.

3.4.4 The second-order Adams Bashforth Scheme ($m=0, l=1$) for the oscillation equation

The second-order Adams-Bashforth scheme and its third-order cousin (Durrant, 1991) have some very nice properties. The second-order Adams-Bashforth finite-difference approximation to the oscillation equation is

$$q^{n+1} - q^n = i\Omega \left(\frac{3}{2}q^n - \frac{1}{2}q^{n-1} \right). \quad (3.56)$$

Like the leapfrog scheme, this is a three-level scheme. Since $m = 0$, however, the time interval on the left-hand side is Δt , rather than $2\Delta t$ as in the leapfrog scheme. The right-hand side represents a linear extrapolation (in time) of q from $q^{(n-1)}$ and $q^{(n)}$ to $n + \frac{1}{2}$. It essentially represents a scheme centered around time level $n + \frac{1}{2}$, and it does have second-order accuracy. The amplification factor for this scheme is given by

$$\lambda^2 - \lambda \left(1 + \frac{3}{2}i\Omega \right) + i\frac{1}{2}\Omega = 0. \quad (3.57)$$

Since this is a three-time-level scheme, we have two modes, given by

$$\lambda_1 = \frac{1}{2} \left(1 + i\frac{3}{2}\Omega + \sqrt{1 - \frac{9}{4}\Omega^2 + i\Omega} \right), \quad (3.58)$$

and

$$\lambda_2 = \frac{1}{2} \left(1 + i\frac{3}{2}\Omega - \sqrt{1 - \frac{9}{4}\Omega^2 + i\Omega} \right). \quad (3.59)$$

Since $\lambda_1 \rightarrow 1$ as $\Omega \rightarrow 0$, the first mode is the physical mode and corresponds to the true solution as $\Delta t \rightarrow 0$. Note, however, that $\lambda_2 \rightarrow 0$ as $\Omega \rightarrow 0$. This means that the “computational” mode tends to damp. It is not neutral, as in the leapfrog scheme.

In order to examine $|\lambda_1|$, we just consider $\Omega \ll 1$, since the expression in (3.58) is complicated and in practice Ω is usually small. Using the binomial theorem, we can approximate λ_1 by

$$\lambda_1 \cong 1 + i\Omega - \frac{9}{16}\Omega^2 \cong 1 + i\Omega - \frac{1}{2}\Omega^2, \quad (3.60)$$

so that

$$|\lambda_1| \cong \sqrt{1 + \frac{\Omega^4}{4}} \cong 1 + \frac{\Omega^4}{8}. \quad (3.61)$$

which is always greater than 1. The physical mode is, therefore, unconditionally unstable, but so far as Δt or Ω is sufficiently small, and because the deviation of $|\lambda_1|$ from 1 is of order Ω^4 , the solution is only weakly unstable. If physical damping is included in the problem, the instability may be suppressed.

3.4.5 A survey of time differencing schemes for the oscillation equation

Fig. 3.12 and Fig. 3.13 are taken from the work of Baer and Simons (1970). They summarize the properties of seven explicit and seven implicit schemes, which are listed in Table 3.3. Properties of these schemes are shown in Fig. 3.12 and Fig. 3.13.

Table 3.3: List of 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. (3.3) and Fig. 3.1; and l , which controls the number of values of f used, again as defined in (3.3).

| <i>Scheme identifier (m,l)</i> | <i>Name</i> | β | α_n | α_{n-1} | α_{n-2} | α_{n-3} | α_{n-4} | <i>Order of accuracy</i> |
|--------------------------------|-----------------------------|---------|------------|----------------|----------------|----------------|----------------|--------------------------|
| <i>E01</i> | <i>Adams-Bashforth</i> | | $3/2$ | $-1/2$ | | | | $(\Delta t)^2$ |
| <i>E02</i> | <i>Adams-Bashforth</i> | | $23/12$ | $-4/3$ | $5/12$ | | | $(\Delta t)^3$ |
| <i>E03</i> | | | $55/24$ | $-59/24$ | $37/24$ | $-9/24$ | | $(\Delta t)^4$ |
| <i>E04</i> | | | $1901/720$ | $-2774/720$ | $2616/720$ | $-1274/720$ | $251/720$ | $(\Delta t)^5$ |
| <i>E11</i> | <i>Leapfrog</i> | | 1 | | | | | $(\Delta t)^2$ |
| <i>E12</i> | | | $7/6$ | $-2/6$ | $1/6$ | | | $(\Delta t)^3$ |
| <i>E33</i> | <i>Milne Predictor</i> | | $2/3$ | $-1/3$ | $2/3$ | | | $(\Delta t)^4$ |
| <i>I00</i> | <i>Trapezoidal Implicit</i> | $1/2$ | $1/2$ | | | | | $(\Delta t)^2$ |

Table 3.3: List of 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. (3.3) and Fig. 3.1; and l , which controls the number of values of f used, again as defined in (3.3).

| Scheme identifier (m, l) | Name | β | α_n | α_{n-1} | α_{n-2} | α_{n-3} | α_{n-4} | Order of accuracy |
|---------------------------------|--------------------|---------|------------|----------------|----------------|----------------|----------------|-------------------|
| I01 | | 5/12 | 8/12 | -1/12 | | | | $(\Delta t)^3$ |
| I02 | Moulton Corrector | 9/24 | 19/24 | -5/24 | 1/24 | | | $(\Delta t)^4$ |
| I03 | | 251/720 | 646/720 | -264/720 | 106/720 | -19/720 | | $(\Delta t)^5$ |
| I12 | Milne Corrector | 1/6 | 4/6 | 1/6 | | | | $(\Delta t)^4$ |
| I13 | | 29/180 | 124/180 | 24/180 | 4/180 | -1/180 | | $(\Delta t)^5$ |
| I34 | Milne II Corrector | 14/180 | 64/180 | 24/180 | 64/180 | 14/180 | | $(\Delta t)^6$ |

There are many other schemes of higher-order accuracy. Books on numerical analysis discuss such schemes. Since our meteorological interest mainly leads us to partial differential equations, the solutions to which will also suffer from truncation error due to space differencing, we cannot hope to gain much by increasing the accuracy of the time differencing only.

3.5 Finite-difference schemes for the decay equation

We now turn our attention to the decay equation,

$$\frac{dq}{dt} = -\kappa q, \quad \kappa > 0, \quad (3.62)$$

which is relevant to many physical parameterizations, including those of the boundary layer, radiative transfer, cloud physics, and convection. The exact solution is

$$q(t) = q(0)e^{-\kappa t}. \quad (3.63)$$

This describes a simple exponential decay with time. For large time, $q \rightarrow 0$. A good scheme should give $q^{n+1} \rightarrow 0$ as $\kappa \Delta t \rightarrow \infty$.

For the Euler (forward) scheme, the finite-difference analogue of (3.62) is

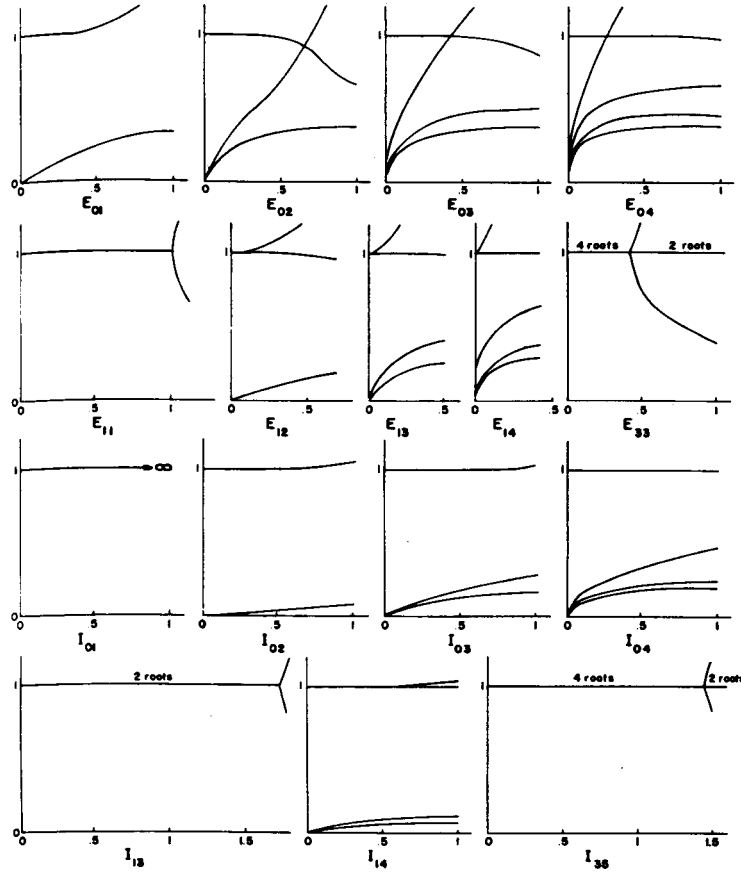


Figure 3.12: Amplification factor of various schemes for the oscillation equation (from Baer and Simons, 1970). The horizontal axis in each panel is Ω . See Table 3.3.

$$q^{n+1} - q^n = -Kq^n, \quad (3.64)$$

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

$$q^{n+1} = (1 - K)q^n. \quad (3.65)$$

Note that $\lambda = 1 - K$ is real. For $|1 - K| < 1$, i.e. κ or Δt small enough, so that $K \leq 2$, the scheme is stable. This is, therefore, a conditionally stable scheme

By the same technique, it is easy to show that, when applied to the decay equation,

- the backward implicit scheme is *unconditionally stable*;

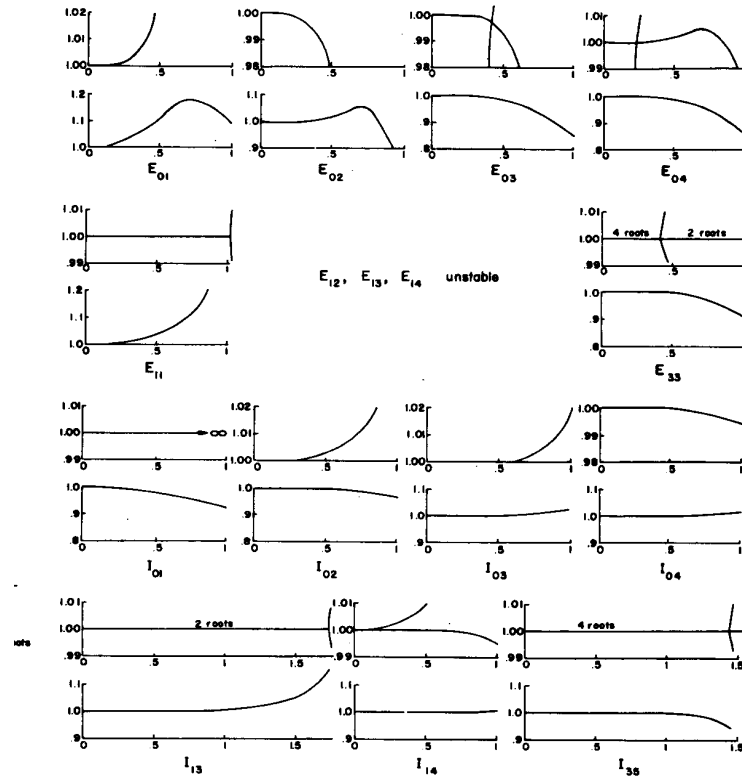


Figure 3.13: $|\lambda|$ and $\frac{\theta}{\Omega}$ (for the physical mode), plotted as functions of Ω , for the oscillation equation (from Baer and Simons, 1970). See Table 3.3.

- the trapezoidal implicit scheme is *unconditionally stable*;
- the Matsuno (Euler-Backward) scheme is *conditionally stable*;
- the Heun scheme is *conditionally stable*; and
- the second-order Adams-Bashforth scheme is *conditionally stable*.

Finally, the leapfrog scheme for the decay equation is

$$q^{n+1} - q^{n-1} = -2Kq^n, \quad (3.66)$$

and so λ satisfies

$$\lambda^2 + 2K\lambda - 1 = 0. \quad (3.67)$$

The two roots are

$$\lambda_1 = -K + \sqrt{K^2 + 1}, \quad \lambda_2 = -K - \sqrt{K^2 + 1}. \quad (3.68)$$

Since $0 \leq \lambda_1 \leq 1$, and $\lambda_1 \rightarrow 1$ as $K \rightarrow 0$, we see that λ_1 corresponds to the physical mode. However, $|\lambda_2| > 1$ always. Actually $\lambda_2 \leq -1$ ($\lambda_2 \rightarrow -1$ as $\Delta t \rightarrow 0$), so the computational mode oscillates in sign from one time level to the next, and amplifies.

Therefore, the leapfrog scheme is *unconditionally unstable* for this type of equation. *We cannot use the leapfrog scheme whenever we have any damping in the problem.* A simple interpretation of this can be given. Suppose we have $q = 0$ at $n = 0$ and $q > 0$ at $n = 1$, as shown schematically in Fig. 3.14. From (3.66) we see that the restoring effect computed at $n = 1$ is added to q^0 , resulting in a negative deviation at $n = 2$. The restoring effect computed at $n = 2$ is added to q^1 , resulting in the amplified positive deviation at $n = 3$, as graphically illustrated in Fig. 3.14. This shows why the leapfrog scheme is a very bad choice for this type of problem.

3.6 Damped oscillations

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

$$\frac{dq}{dt} = i\omega q - \kappa q = (-\kappa + i\omega)q? \quad (3.69)$$

As an example, we can mix the leapfrog and forward (or backward) schemes in the following manner. We write the finite difference analogue of (3.69) as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n-1}, \quad (3.70)$$

(decay term forward differenced), or as

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1}. \quad (3.71)$$

(decay term backward differenced). The oscillation term on the right-hand sides of both (3.70) and (3.71) is in “centered” form, whereas the damping term is an uncentered form. These schemes are conditionally stable.

3.7 Nonlinear damping

In real applications, it is quite typical that κ depends on q , so that (3.62) becomes nonlinear. Kalnay and Kanamitsu (1988) studied the behavior of ten time-differencing schemes for a nonlinear version of (3.62) given by

$$\frac{dq}{dt} = -(\kappa q^P)q + S. \quad (3.72)$$

$$q^0 = 0$$

$$q^1 > 0$$

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

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

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

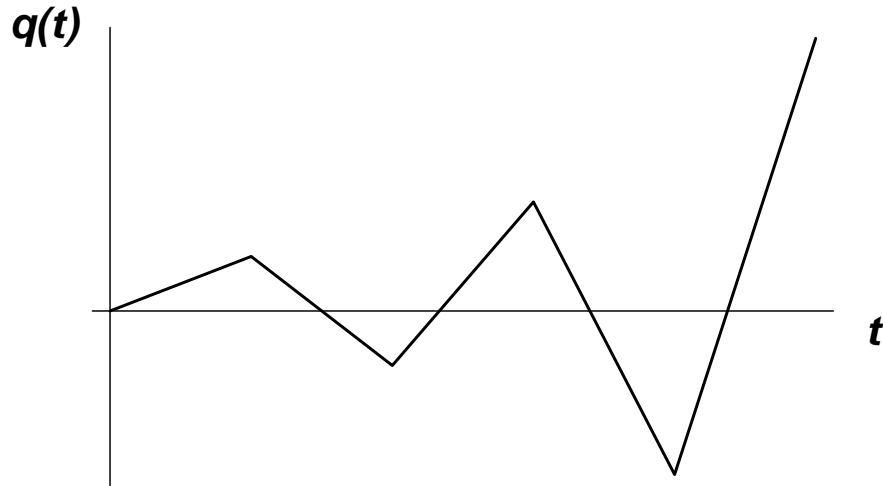


Figure 3.14: An example illustrating how the leapfrog scheme leads to instability with the decay equation. The solution shown here represents the computational mode only and would be superimposed on the physical mode.

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 $S = 0$, (3.72) reduces to (3.62).

An example of a real application that gives rise to an equation of the form (3.72) is boundary-layer parameterization. The soil temperature, T_g , satisfies an equation roughly of the form

$$C \frac{dT_g}{dt} = -\rho_a c_T (T_g - T_a) V (T_g - T_a) + S_g, \quad (3.73)$$

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

The air temperature is governed by a similar equation:

$$\rho_a D c_P \frac{dT_a}{dt} = \rho_a c_T (T_g - T_a) V (T_g - T_a) + S_a \quad (3.74)$$

Here c_P is the heat capacity of air at constant pressure, and D is the depth of the layer of air whose temperature is represented by T_a .

Comparing (3.73) with (3.74), we find that

$$\frac{d(T_g - T_a)}{dt} = -\rho_a c_T (T_g - T_a) V (T_g - T_a) \left(\frac{1}{C} + \frac{1}{\rho_a D c_P} \right) + \left(\frac{S_g}{C} - \frac{S_a}{\rho_a D c_P} \right) \quad (3.75)$$

The correspondence between (3.75) and (3.72) is obvious. The two equations are essentially the same if the transfer coefficient $c_T(T_g - T_a)$ has a power-law dependence on $T_g - T_a$. Virtually all realistic atmospheric models involve equations something like (3.73) and (3.74), so this is a very practical example.

From what we have already discussed, it should be clear that an implicit scheme would be a good choice for (3.72), i.e.

$$\frac{q^{n+1} - q^n}{\Delta t} = -\kappa (q^{n+1})^{P+1} + S. \quad (3.76)$$

Such a scheme is in fact unconditionally stable, but for arbitrary P it must be solved iteratively, which can be quite expensive. For this practical reason, (3.76) would not be considered a viable choice, except where P is a small integer, in which case, (3.76) can be solved analytically.

Let q denote an equilibrium solution of (3.72). Then q satisfies

$$\kappa q^{P+1} = S. \quad (3.77)$$

Let q' denote a departure from this equilibrium, so that $q = \bar{q} + q'$. Then (3.72) can be linearized as follows:

$$\frac{d}{dt}(\bar{q} + q') = -\kappa \bar{q}^{P+1} - \kappa(P+1)\bar{q}^P q' + S, \quad (3.78)$$

which reduces to

$$\frac{dq'}{dt} = -\kappa(P+1)\bar{q}^P q'. \quad (3.79)$$

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

As an example, the forward time-differencing scheme, applied to (3.79), gives

$$q^{n+1} - q^n = -\alpha(P+1)q^n, \quad (3.80)$$

where

$$\alpha \equiv \kappa \bar{q}^P \Delta t, \quad (3.81)$$

and we have dropped the “prime” notation for simplicity. We can rearrange (3.80) to

$$q^{n+1} = [1 - \alpha(P+1)]q^n, \quad (3.82)$$

from which we see that

$$\lambda = 1 - \alpha(P+1). \quad (3.83)$$

Table 3.4 summarizes the ten schemes that Kalnay and Kanamitsu analyzed, and

Table 3.4: Schemes for the nonlinear decay equation, as studied by Kalnay and Kanamitsu (1988).

| Name of scheme | Form of scheme | Amplification factor | Linear stability criterion |
|--------------------------|---|-----------------------------|----------------------------|
| <i>Forward explicit</i> | $\frac{q^{n+1} - q^n}{\Delta t} = -\kappa(q^n)^{P+1} + S$ | $1 - \alpha(P+1)$ | $\alpha(P+1) < 2$ |
| <i>Backward implicit</i> | $\frac{q^{n+1} - q^n}{\Delta t} = -\kappa(q^{n+1})^{P+1} + S$ | $\frac{1}{1 + \alpha(P+1)}$ | Unconditionally stable |

Table 3.4: Schemes for the nonlinear decay equation, as studied by Kalnay and Kanamitsu (1988).

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

gives the amplification factors and stability criteria for each. 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. As a recommendation, the “Double time step, explicit coefficient, implicit temperature with time average filter” scheme (the second-last scheme in the table) is quite nice.

3.8 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, especially using symbolic algebra programs, 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.

More accurate schemes 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.

For the advection and oscillation equations, truncation errors can be separated into amplitude errors and phase errors. Neutral schemes, like the leapfrog scheme, have only phase errors.

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

- 1) Choose a scheme that involves only two time levels;
- 2) Choose the computational initial condition well, and periodically re-start;
- 3) Choose the computational initial condition well, and use a time filter (e.g. Asselin, 1972) to suppress the computational mode;
- 4) 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).

Appendix to Chapter 3

A Proof that the Fourth-Order Runge-Kutta Scheme has Fourth-Order Accuracy

We wish to obtain an approximate numerical solution of the ordinary differential equation

$$\frac{dq}{dt} = f(q, t) . \quad (3.84)$$

As discussed earlier, the fourth-order Runge-Kutta scheme is given by:

$$q^{(n+1)} - q^{(n)} = \Delta t \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4) , \quad (3.85)$$

where

$$\begin{aligned} k_1 &= f(q^{(n)}, n\Delta t) , & k_2 &= f\left[q^{(n)} + \frac{k_1\Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right] , \\ k_3 &= f\left[q^{(n)} + \frac{k_2\Delta t}{2}, \left(n + \frac{1}{2}\right)\Delta t\right] , & k_4 &= f(q^{(n)} + k_3\Delta t, (n+1)\Delta t) . \end{aligned} \quad (3.86)$$

To prove that this scheme has fourth-order accuracy, we construct the right-hand side of (3.85), and then to show that it is equal to the left-hand side, with an error of order $(\Delta t)^5$.

Define an operator δ by

$$\delta \equiv k_1 \frac{\partial}{\partial q} + \frac{\partial}{\partial t} = f \frac{\partial}{\partial q} + \frac{\partial}{\partial t} , \quad (3.87)$$

where we drop the notation that indicates the time level, for convenience. If we apply the operator δ to a function (such as f) which depends on both q and t , then δ returns the total time rate of change of the function, taking into account the part of this tendency that comes from the change in q and also the part that comes from the change in t

$$\delta(f) \equiv \frac{df}{dt} . \quad (3.88)$$

We now write

$$k_1 = f , \quad (3.89)$$

$$\begin{aligned} k_2 &= f \left[q^{(n)} + \frac{k_1 \Delta t}{2}, \left(n + \frac{1}{2} \right) \Delta t \right] \\ &= f + \frac{\Delta t}{2} \delta(f) + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \delta^2(f) + \frac{1}{3!} \left(\frac{\Delta t}{2} \right)^3 \delta^3(f) + O[(\Delta t)^4] \\ &= \left[1 + \frac{\Delta t}{2} \delta + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \delta^2 + \frac{1}{3!} \left(\frac{\Delta t}{2} \right)^3 \delta^3 \right] f + O[(\Delta t)^4] , \end{aligned} \quad (3.90)$$

and

$$\begin{aligned}
k_3 &= f \left[q^{(n)} + \frac{k_2 \Delta t}{2}, \left(n + \frac{1}{2} \right) \Delta t \right] \\
&= f + \left(\frac{\Delta t}{2} k_2 \frac{\partial}{\partial q} + \frac{\Delta t}{2} \frac{\partial}{\partial t} \right) f + \frac{1}{2!} \left(\frac{\Delta t}{2} k_2 \frac{\partial}{\partial q} + \frac{\Delta t}{2} \frac{\partial}{\partial t} \right)^2 f + \frac{1}{3!} \left(\frac{\Delta t}{2} k_2 \frac{\partial}{\partial q} + \frac{\Delta t}{2} \frac{\partial}{\partial t} \right)^3 f + O[(\Delta t)^4] \\
&= f + \frac{\Delta t}{2} \left\{ \left[1 + \frac{\Delta t}{2} \delta + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \delta^2 \right] f \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right\} f \\
&\quad + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \left[\left(1 + \frac{\Delta t}{2} \delta \right) f \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right]^2 f \\
&\quad + \frac{1}{3!} \left(\frac{\Delta t}{2} \right)^3 \left(f \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right)^3 f + O[(\Delta t)^4] \\
&= f + \frac{\Delta t}{2} \left[\delta(f) + \frac{\Delta t}{2} \delta(f) f_q + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \delta^2(f) f_q \right] \\
&\quad + \frac{1}{2!} \left(\frac{\Delta t}{2} \right)^2 \left[\delta^2(f) + \frac{\Delta t}{2} \delta(f) f_q \right] \\
&\quad + \frac{1}{3!} \left(\frac{\Delta t}{2} \right)^3 \delta^3(f) + O[(\Delta t)^4] \\
&= f + \Delta t \frac{\delta(f)}{2} + \frac{(\Delta t)^2}{2!} \left[\frac{\delta^2(f)}{4} + \frac{\delta(f) f_q}{2} \right] \\
&\quad + \frac{(\Delta t)^3}{3!} \left[\frac{\delta^3(f)}{8} + \frac{3}{4} \delta(f) f_q + \frac{3}{8} \delta^2(f) f_q \right] + O[(\Delta t)^4] .
\end{aligned} \tag{3.91}$$

In (3.91), we have simplified along the way by suppressing higher-order terms whenever possible, and we have used the notation $f_q \equiv \frac{\partial f}{\partial q}$. It remains to assemble k_4 :

$$\begin{aligned}
k_4 &= (q^{(n)} + k_3 \Delta t, (n+1)\Delta t) \\
&= f + \left(k_3 \Delta t \frac{\partial}{\partial q} + \Delta t \frac{\partial}{\partial t} \right) f + \frac{1}{2!} \left(k_3 \Delta t \frac{\partial}{\partial q} + \Delta t \frac{\partial}{\partial t} \right)^2 f + \frac{1}{3!} \left(k_3 \Delta t \frac{\partial}{\partial q} + \Delta t \frac{\partial}{\partial t} \right)^3 f \\
&= f + \Delta t \left\{ \left[f + \Delta t \frac{\delta(f)}{2} + \frac{(\Delta t)^2}{2!} \left[\frac{\delta^2(f)}{4} + \frac{\delta(f)f_q}{2} \right] \right] \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right\} f \\
&\quad + \frac{(\Delta t)^2}{2!} \left\{ \left[f + \Delta t \frac{\delta(f)}{2} \right] \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right\}^2 f + \frac{(\Delta t)^3}{3!} \left(f \frac{\partial}{\partial q} + \frac{\partial}{\partial t} \right)^3 f + O[(\Delta t)^4] \\
&= f + \Delta t \left\{ \delta(f) + \Delta t \frac{\delta(f)}{2} f_q + \frac{(\Delta t)^2}{2!} \left[\frac{\delta^2(f)}{4} + \frac{\delta(f)f_q}{2} \right] f_q \right\} \\
&\quad + \frac{(\Delta t)^2}{2!} [\delta^2(f) + \Delta t \delta(f) \delta(f_q)] + \frac{(\Delta t)^3}{3!} \delta^3(f) + O[(\Delta t)^4] \\
&= f + \Delta t \delta(f) + \frac{(\Delta t)^2}{2!} [\delta^2(f) + \delta(f)f_q] \\
&\quad + \frac{(\Delta t)^3}{3!} \left[\delta^3(f) + 3\delta(f)\delta(f_q) + \frac{3}{4}\delta^2(f)f_q + \frac{3}{2}\delta(f)(f_q)^2 \right] + O[(\Delta t)^4] .
\end{aligned} \tag{3.92}$$

Finally, we combine terms to obtain

$$\begin{aligned}
\Delta t \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4) &= f \Delta t + \frac{(\Delta t)^2}{2!} \delta(f) + \frac{(\Delta t)^3}{3!} [\delta^2(f) + \delta(f)f_q] \\
&\quad + \frac{(\Delta t)^4}{4!} [\delta^3(f) + 3\delta(f)\delta(f_q) + \delta^2(f)f_q + \delta(f)(f_q)^2] + O[(\Delta t)^5] .
\end{aligned} \tag{3.93}$$

This can be simplified using

$$f = \frac{dq}{dt} \tag{3.94}$$

$$\delta(f) = \frac{df}{dt} = \frac{d^2 q}{dt^2} , \tag{3.95}$$

$$\delta^2(f) + \delta(f)f_q = \frac{d^2 f}{dt^2} = \frac{d^3 q}{dt^3} , \tag{3.96}$$

$$\delta^3(f) + 3\delta(f)\delta(f_q) + \delta^2(f)f_q + \delta(f)(f_q)^2 = \frac{d^3 f}{dt^3} = \frac{d^4 q}{dt^4} \quad . \quad (3.97)$$

By substituting from (3.94)-(3.97), we can rewrite (3.93) as

$$\Delta t \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4) = \frac{dq}{dt} \Delta t + \frac{(\Delta t)^2}{2!} \frac{d^2 q}{dt^2} + \frac{(\Delta t)^3}{3!} \frac{d^3 q}{dt^3} + \frac{(\Delta t)^4}{4!} \frac{d^4 q}{dt^4} + O[(\Delta t)^5] \quad . \quad (3.98)$$

Problems

1. a) Find the exact solution of

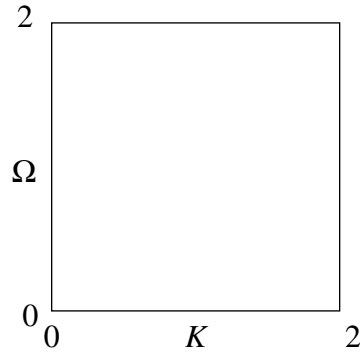
$$\frac{dq}{dt} = i\omega q - \kappa q. \quad (3.99)$$

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

$$q^{n+1} - q^{n-1} = 2i\Omega q^n - 2Kq^{n+1}. \quad (3.100)$$

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



- c) 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:

- a) $\frac{\omega}{2\pi} = 0.1$, $\kappa = 0$
- b) $\frac{\omega}{2\pi} = 0$, $\kappa = 0.1$
- c) $\frac{\omega}{2\pi} = 0.1$, $\kappa = 0.1$

For each case, plot $Re\{q\}$ 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).

d) Derive an equation satisfied by the amplification factor for the second-order Adams-Bashforth scheme applied to Eq. (3.99). (The result is quite complicated.) Contour plot $|\lambda|$ as a function of both ω and κ . Find an approximate solution valid for sufficiently small Δt .

2. For the oscillation equation, compare the phase change per time step of:

a) the leapfrog scheme's physical mode;

b) the trapezoidal implicit scheme.

Plot the phase change per time step as a function of ω , for both schemes. Discuss the phase errors of the two schemes, as functions of ω .

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

$$q^{(n+1)} - q^{(n)} = \frac{i\omega\Delta t}{2}(q^{(n)} + q^{(n+1)}). \quad (3.101)$$

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

b) Find the phase change per time step, and compare with the phase change per Δt in the exact solution.

4. Determine the order of accuracy of the Matsuno scheme.

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

CHAPTER 4***A closer look at the advection equation***

Copyright 2004 David A. Randall

4.1 Introduction

Most of this chapter is devoted to a discussion of the one-dimensional advection equation,

$$\frac{\partial A}{\partial t} + c \frac{\partial A}{\partial x} = 0. \quad (4.1)$$

Here A is the advected quantity, and c is the advecting current. This is a linear, first-order, partial differential equation with a constant coefficient, namely c . Both space and time differencing are discussed in this chapter, but more emphasis is placed on space differencing.

We have already presented the exact solution of (4.1). Before proceeding, however, it is useful 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

$$\frac{DA}{Dt} = 0. \quad (4.2)$$

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 (4.2), 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 characteristic of advection.

Here is another way of describing this characteristic: If we worked out the probability density function (pdf) for 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 for 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. A third example, and a particularly troublesome one, is the case in which A is a component of the advecting velocity field itself; here A is a component of the momentum per unit mass.

Of course, in general these various quantities are not really conserved following particles; 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 replace (4.2) by

$$\frac{DA}{Dt} = S, \quad (4.3)$$

where S is the source of A per unit time. (A negative value of S represents a sink.) We still refer to (4.3) 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 can be written as

$$\frac{\partial \rho}{\partial t} = -\nabla \bullet (\rho \mathbf{V}), \quad (4.4)$$

where ρ is the density (mass per unit volume) and \mathbf{V} is the velocity vector. Using the velocity vector, we can expand (4.3) into the Eulerian advective form of the conservation equation for A :

$$\frac{\partial A}{\partial t} = -(\mathbf{V} \bullet \nabla)A + S. \quad (4.5)$$

Multiply (4.4) by A , and (4.5) by ρ and add the results to obtain

$$\frac{\partial}{\partial t}(\rho A) = -\nabla \bullet (\rho \mathbf{V} A) + \rho S. \quad (4.6)$$

This is called the flux form of the conservation equation for A . Notice that if we put $A \equiv 1$ and $S \equiv 0$ then (4.6) reduces to (4.4). This is an important point that can and should be used in the design of advection schemes.

If we integrate (4.4) over a closed domain R (“closed” meaning that R experiences no sources or sinks of mass) then we find, using Gauss’s Theorem, that

$$\frac{d}{dt} \int_R \rho \, dR = 0. \quad (4.7)$$

This simply states that mass is conserved within the domain. Similarly, we can integrate (4.6) over R to obtain

$$\frac{d}{dt} \int_R \rho A \, dR = \int_R \rho S \, dR. \quad (4.8)$$

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 say that (4.6) and (4.8) are 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 advected about by the wind. In such a Lagrangian model, the properties associated with each particle would include its spatial coordinates, e.g. its longitude, latitude, and height. These would change in response to the predicted velocity field. Such a Lagrangian approach will be discussed later in this chapter.

At the present time, virtually all models in atmospheric science are based on Eulerian methods, although the Eulerian coordinates are sometimes permitted to “move” as the circulation evolves (e.g. Phillips, 1957; Hsu and Arakawa, 1990).

When we design finite-difference schemes to represent advection, we strive for

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

$$\sum_j \rho_j^{n+1} dR_j = \sum_j \rho_j^n dR_j \quad (4.9)$$

and

$$\sum_j (\rho A)_j^{n+1} dR_j = \sum_j (\rho A)_j^n dR_j + \Delta t \sum_j (\rho S)_j^n dR_j. \quad (4.10)$$

These are finite-difference analogs to the integral forms (4.7) and (4.8), respectively. In (4.10) 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.

Finally, we may wish to require that a non-negative variable, such as the water vapor mixing ratio, remain non-negative under advection. An advection scheme with this property is often called “positive-definite” or “sign-preserving” positive. Definite 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.

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. Note that in a model based on Lagrangian methods, advection does not alter the pdf of the advected quantity.

4.2 Conservative finite-difference methods

Let A be a “conservative” variable, satisfying the following one-dimensional conservation law:

$$\frac{\partial}{\partial t}(mA) + \frac{\partial}{\partial x}(muA) = 0. \quad (4.11)$$

Here m is a mass variable, which might be the density of the air, or might be the depth of shallow water, and mu is a mass flux. Putting $A \equiv 1$ in (4.11) gives mass conservation:

$$\frac{\partial m}{\partial t} + \frac{\partial}{\partial x}(mu) = 0. \quad (4.12)$$

Approximate (4.11) and (4.12) with:

$$\frac{d}{dt}(m_j A_j) + \frac{(mu)_{j+\frac{1}{2}} A_{j+\frac{1}{2}} - (mu)_{j-\frac{1}{2}} A_{j-\frac{1}{2}}}{\Delta x_j} = 0, \quad (4.13)$$

$$\frac{dm_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} - (mu)_{j-\frac{1}{2}}}{\Delta x_j} = 0. \quad (4.14)$$

These are called differential-difference equations (or sometimes semi-discrete equations), because the time-change terms are in differential form, while the spatial derivatives have been approximated using a finite-difference quotient. The variables m and A are defined at integer points, while u and mu are defined at half-integer points. See Fig. 4.1. This is an example of

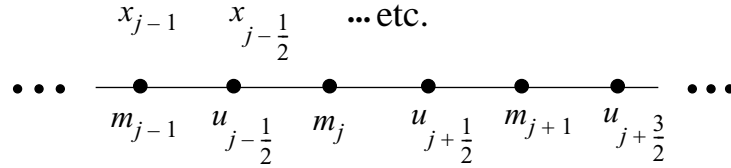


Figure 4.1: The staggered grid used in (4.13) and (4.14).

a “staggered” grid. $A_{j+\frac{1}{2}}$ and $A_{j-\frac{1}{2}}$ must be interpolated somehow from the predicted values

of A . Note that if we put $A \equiv 1$, (4.13) reduces to (4.14). This is an important point that will be discussed further later.

Multiply (4.13) and (4.14) through by Δx_j , and sum over the domain:

$$\frac{d}{dt} \sum_{j=0}^J (m_j A_j \Delta x_j) + (mu)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} - (mu)_{-\frac{1}{2}} A_{-\frac{1}{2}} = 0, \quad (4.15)$$

$$\frac{d}{dt} \sum_{j=0}^J (m_j \Delta x_j) - (mu)_{J+\frac{1}{2}} + (mu)_{-\frac{1}{2}} = 0. \quad (4.16)$$

If

$$(mu)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} = (mu)_{-\frac{1}{2}} A_{-\frac{1}{2}}, \quad (4.17)$$

and

$$(mu)_{j+\frac{1}{2}} = (mu)_{j-\frac{1}{2}}, \quad (4.18)$$

then we obtain

$$\frac{d}{dt} \sum_{j=0}^J (m_j A_j \Delta x_j) = 0, \quad (4.19)$$

$$\frac{d}{dt} \sum_{j=0}^J (m_j \Delta x_j) = 0, \quad (4.20)$$

which express conservation of mass [Eq. (4.20)] and of the mass-weighted value of A [Eq. (4.19)]. Compare with (4.9) and (4.10). Note that (4.19) holds regardless of the form of the interpolation used for $A_{j+\frac{1}{2}}$.

By combining (4.11) and (4.12), we obtain the advective form of our conservation law:

$$m \frac{\partial A}{\partial t} + mu \frac{\partial A}{\partial x} = 0. \quad (4.21)$$

From (4.13) and (4.14), we can derive a finite-difference “advective form,” analogous to (4.21):

$$m_j \frac{dA_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} \left(A_{j+\frac{1}{2}} - A_j \right) + (mu)_{j-\frac{1}{2}} \left(A_j - A_{j-\frac{1}{2}} \right)}{\Delta x_j} = 0. \quad (4.22)$$

Since (4.22) is consistent with (4.13) and (4.14), use of (4.22) and (4.14) 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 (4.22) gives $\frac{dA_j}{dt} = 0$, which is “the right answer.” This is ensured **because** (4.13) reduces to (4.14) when A is uniform over the grid. *If the flux-form advection equation does not reduce to the flux-form continuity equation when A is uniform over the grid, then a uniform tracer field will not remain uniform under advection.*

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^n , $\ln(A)$, etc. This is most easily seen from the Lagrangian form of (4.21):

$$\frac{DA}{Dt} = 0 . \quad (4.23)$$

According to (4.23), A is conserved “following a particle.” As discussed earlier, this implies that

$$\frac{D}{Dt}[F(A)] = 0 , \quad (4.24)$$

where $F(A)$ is an arbitrary function of A only. We can derive (4.24) by multiplying (4.23) by dF/dA .

In a finite difference system, we can force conservation of at most one nontrivial function of A , in addition to A itself. Let F_j denote $F(A_j)$, where F is an arbitrary function, and let F'_j denote $\frac{d[F(A_j)]}{dA_j}$. Multiplying (4.22) by F'_j gives

$$m_j \frac{dF_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} F'_j \left(A_{j+\frac{1}{2}} - A_j \right) + (mu)_{j-\frac{1}{2}} F'_j \left(A_j - A_{j-\frac{1}{2}} \right)}{\Delta x_j} = 0 . \quad (4.25)$$

Now use (4.14) to rewrite (4.25) in “flux form”:

$$\frac{d}{dt}(m_j F_j) + \frac{1}{\Delta x_j} \left\{ (mu)_{j+\frac{1}{2}} \left[F'_j \left(A_{j+\frac{1}{2}} - A_j \right) + F_j \right] - (mu)_{j-\frac{1}{2}} \left[-F'_j \left(A_j - A_{j-\frac{1}{2}} \right) + F_j \right] \right\} = 0 . \quad (4.26)$$

Inspection of (4.26) shows that, to ensure conservation of $F(A)$, we must choose

$$F_{j+\frac{1}{2}} = F'_j \left(A_{j+\frac{1}{2}} - A_j \right) + F_j , \quad (4.27)$$

$$F_{j-\frac{1}{2}} = -F'_j \left(A_j - A_{j-\frac{1}{2}} \right) + F_j . \quad (4.28)$$

Let $j \rightarrow j+1$ in (4.28), giving

$$F_{j+\frac{1}{2}} = -F'_{j+1} \left(A_{j+1} - A_{j+\frac{1}{2}} \right) + F_{j+1} . \quad (4.29)$$

Eliminating $F_{j+\frac{1}{2}}$ between (4.27) and (4.29), we obtain

$$A_{j+\frac{1}{2}} = \frac{(F'_{j+1}A_{j+1} - F_{j+1}) - (F'_jA_j - F_j)}{F'_{j+1} - F'_j}. \quad (4.30)$$

By choosing $A_{j+\frac{1}{2}}$ accordingly to (4.30), 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

$$A_{j+\frac{1}{2}} = \frac{(2A_{j+1}^2 - A_{j+1}^2) - (2A_j^2 - A_j^2)}{2(A_{j+1} - A_j)} = \frac{1}{2}(A_{j+1} + A_j). \quad (4.31)$$

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

(4.31). 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, (4.31) gives second-order accuracy in space. If the grid spacing is nonuniform, however, the accuracy drops to first order. The strength of the first-order error depends on how rapidly the grid spacing changes. Substituting (4.31) back into (4.22) gives

$$m_j \frac{dA_j}{dt} + \frac{1}{2\Delta x_j} \left[(mu)_{j+\frac{1}{2}} (A_{j+1} - A_j) + (mu)_{j-\frac{1}{2}} (A_j - A_{j-1}) \right] = 0. \quad (4.32)$$

This is the advective form that allows conservation of A^2 (and of A).

One point to be noticed here is that there are infinitely many ways to interpolate a variable. We can spatially interpolate A itself in a linear fashion, e.g.

$$A_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} A_j + \left(1 - \alpha_{j+\frac{1}{2}}\right) A_{j+1}, \quad (4.33)$$

where $\alpha_{j+\frac{1}{2}}$ is a weighting factor that might be a constant, as in (4.31), or might be a function of x_j , x_{j+1} , and $x_{j+\frac{1}{2}}$. We can interpolate so as to conserve an arbitrary function of A , as in (4.30). We can compute an arbitrary function of A , interpolate the function using a

form such as (4.33), and then extract an interpolated value of A by applying the inverse of the function to the result. A practical example of this would be interpolation of the water vapor mixing ratio by computing the relative humidity, interpolating the relative humidity, and then converting back to mixing ratio. We can also make use of “averages” that are different from the simple and familiar arithmetic mean given by (4.31). Examples are the “*geometric mean*,”

$$A_{j+\frac{1}{2}} = \sqrt{A_j A_{j+1}}, \quad (4.34)$$

and the “*harmonic mean*,”

$$A_{j+\frac{1}{2}} = \frac{2A_j A_{j+1}}{A_j + A_{j+1}}. \quad (4.35)$$

Note that both (4.34) and (4.35) 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 “average.” They are both nonlinear averages. 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 (4.34) and (4.35) tend to make the interpolated value close to the smaller of the two input values.

Here is another interesting interpolation that has the opposite property, i.e., it makes the interpolated value close to the larger of the two input values:

$$A_{j+\frac{1}{2}} = \frac{A_j + A_{j+1} - 2A_j A_{j+1}}{2 - (A_j + A_{j+1})}. \quad (4.36)$$

In short, there are infinitely many ways to average and/or interpolate. This is good because it means that we have the opportunity to choose the *best* way for our particular application.

Fig. 4.2 shows four interpolations as functions of the two input values.

In this section, we have considered truncation errors only insofar as they affect conservation properties. We must also consider how they affect the various other aspects of the solution. This is taken up in the next section.

4.3 Examples of schemes with centered space differencing

A centered-difference quotient already discussed in Chapter 2 is

$$\frac{\partial u^A}{\partial x} \cong \frac{u_{j+1}^A - u_{j-1}^A}{2\Delta x} \text{ at } x_j, \quad (4.37)$$

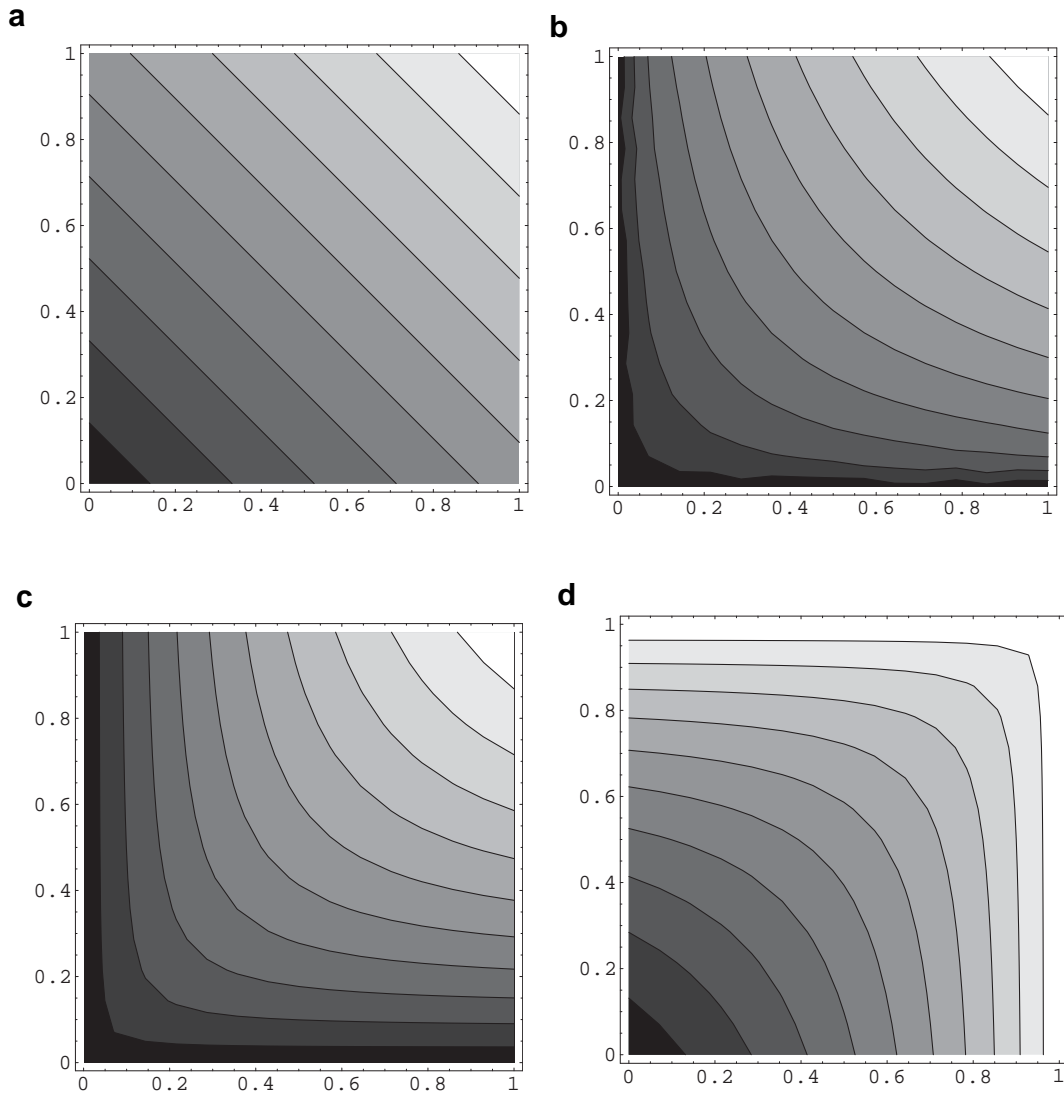


Figure 4.2: Four interpolations as functions of the input values. a) arithmetic mean, b) geometric mean, c) harmonic mean, d) Eq. (4.36), which makes the interpolated value close to the larger of the two input values. In all plots, black is close to zero, and white is close to one.

where j is the spatial index and n is the time index. If $u(x, t)$ has the wave form $u(x, t) = \hat{u}(t)e^{ikj\Delta x}$, where k is the wave number, then

$$\frac{u_{j+1} - u_{j-1}}{2\Delta x} = ik \frac{\sin k\Delta x}{k\Delta x} \hat{u}(t) e^{ikj\Delta x}. \quad (4.38)$$

Therefore, the advection equation becomes

$$\frac{d\hat{u}}{dt} + ikc \frac{\sin k\Delta x}{k\Delta x} \hat{u} = 0. \quad (4.39)$$

If we define $\omega \equiv -kc \frac{\sin k\Delta x}{k\Delta x}$, then (4.39) reduces to the oscillation equation, which was discussed at length in Chapter 3. Note that $\frac{\sin k\Delta x}{k\Delta x} \rightarrow 1$ as $k\Delta x \rightarrow 0$.

We can now study the properties of the various time-differencing schemes, as we did in Chapter 3, but we are now able to obtain an explicit relationship between Δx and Δt as a condition for stability, based on the use of (4.39). The forward time scheme is unstable when combined with the centered space scheme. You should prove this fact and remember it. It was mentioned already in Chapter 3. In the case of the leapfrog scheme, the finite-difference analogue of (4.1) is

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + c \left(\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \right) = 0. \quad (4.40)$$

If we assume that u_j^n has the wave form for which (4.38) holds, then (4.40) can be written as

$$\hat{u}^{n+1} - \hat{u}^{n-1} = 2i\Omega \hat{u}^n, \quad (4.41)$$

where

$$\Omega \equiv -kc \frac{\sin k\Delta x}{k\Delta x} \Delta t. \quad (4.42)$$

We recognize Eq. (4.41) as the leapfrog scheme for the oscillation equation. Recall from Chapter 3 that $|\Omega| \leq 1$ is necessary for (4.41) to be stable. Therefore,

$$\left| c \frac{\sin k\Delta x}{\Delta x} \Delta t \right| \leq 1 \quad (4.43)$$

must hold for stability, for any and all k . The “worst case” is $|\sin k\Delta x| = 1$. This occurs for wavelength $L = 4\Delta x$, so

$$|c| \frac{\Delta t}{\Delta x} \leq 1 \quad (4.44)$$

is the necessary condition for stability, i.e. stability for all modes. Note that the $2\Delta x$ wave is not the problem here. It is the $4\Delta x$ wave that is most likely to cause trouble. Eq. (4.44) is the famous “CFL” stability criterion associated with the names Courant, Friedrichs and Lewy.

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:

$$\hat{u}_1 = \lambda_1^n \hat{u}_1, \quad \hat{u}_2 = \lambda_2^n \hat{u}_2. \quad (4.45)$$

For $|\Omega| \leq 1$, we find, as discussed in Chapter 3, that

$$\lambda_1 = e^{i\theta}, \lambda_2 = e^{i(\pi-\theta)} = -e^{-i\theta}, \theta = \tan^{-1} \left(\frac{\Omega}{\sqrt{1-\Omega^2}} \right). \quad (4.46)$$

Both modes are neutral. For the physical mode,

$$(\hat{u}_j^n)_1 = \lambda_1^{(n)} \hat{u}_1^0 e^{ikj\Delta x} = \hat{u}_1^0 \exp \left[ik \left(j\Delta x + \frac{\theta}{k\Delta t} n\Delta t \right) \right]. \quad (4.47)$$

For the computational mode, similarly, we obtain

$$(\hat{u}_j^n)_2 = \hat{u}_2^0 (-1)^n \exp \left[ik \left(j\Delta x - \frac{\theta}{k\Delta t} n\Delta t \right) \right]. \quad (4.48)$$

Note the nasty factor of $(-1)^n$, which comes from the leading minus sign in (4.46). Comparing (4.47) and (4.48) with the expression $u(x, t) = \hat{u}(0)e^{ik(x-ct)}$, 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 $(\Delta x, \Delta t) \rightarrow 0$, $\theta \rightarrow \Omega \rightarrow -kc\Delta t$, and so the speed of the physical mode approaches c , while that of the computational mode approaches $-c$. The computational mode goes backwards!

Note that the finite-difference approximation to the phase speed depends on k , while the true phase speed, c , is independent of k . 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.

Further examples of schemes for the advection equation can be obtained by combining this centered space differencing with the two-level time-differencing schemes (see Chapter 3). In the case of the Matsuno scheme, the first approximation to u_j^{n+1} comes from

$$\frac{(u_j^{n+1})^* - u_j^{(n)}}{\Delta t} + c \left[\frac{u_{j+1}^{(n)} - u_{j-1}^{(n)}}{2\Delta x} \right] = 0, \quad (4.49)$$

and the final value from

$$\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} + c \frac{(u_{j+1}^{(n+1)})^* - (u_{j-1}^{(n+1)})^*}{2\Delta x} = 0. \quad (4.50)$$

Eliminating the terms with $()^*$ from (4.50) by using (4.49) twice (first with j replaced by $j+1$, then with j replaced by $j-1$), we obtain

$$\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} + c \frac{u_{j+1}^{(n)} - u_{j-1}^{(n)}}{2\Delta x} = \frac{c^2 \Delta t}{(2\Delta x)^2} (u_{j+2}^{(n)} - 2u_j^{(n)} + u_{j-2}^{(n)}). \quad (4.51)$$

The term on the right side of (4.51) approaches zero as $\Delta t \rightarrow 0$, and thus (4.51) is consistent with (4.1). If we let $\Delta x \rightarrow 0$ (and $\Delta t \rightarrow 0$ to keep stability), this term approaches $\frac{c^2 \Delta t}{4} \frac{\partial^2 u}{\partial x^2}$. In effect, it acts as a diffusion term that damps disturbances. The “diffusion coefficient” is $\frac{c^2 \Delta t}{4}$, so it goes to zero as $\Delta t \rightarrow 0$.

We now examine a similarly diffusive scheme, called the Lax-Wendroff scheme, which has second-order accuracy. Consider an explicit two-level scheme of the form:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{\Delta x} (\alpha u_{j-1}^n + \beta u_j^n + \gamma u_{j+1}^n) = 0. \quad (4.52)$$

Note that there are only two time levels, so this scheme does not have any computational modes. Replace the various u 's by the corresponding values of the true solution and then expand them into the Taylor series around the point $(j\Delta x, n\Delta t)$. The result is

$$\begin{aligned} & \left(u_t + \frac{\Delta t}{2!} u_{tt} + \frac{\Delta t^2}{3!} u_{ttt} \right)_j + \dots \\ & + \frac{c}{\Delta x} \left[\alpha \left(u - \Delta x u_x + \frac{\Delta x^2}{2!} u_{xx} - \frac{\Delta x^3}{3!} u_{xxx} + \dots \right)_j \right. \\ & \quad + \beta u_j \\ & \quad \left. + \gamma \left(u + \Delta x u_x + \frac{\Delta x^2}{2!} u_{xx} + \frac{\Delta x^3}{3!} u_{xxx} + \dots \right)_j \right] \\ & = \epsilon, \end{aligned} \quad (4.53)$$

where all quantities are evaluated at $(x, t) = (x_j, t^n)$, and ϵ is the truncation error. Make sure

that you understand where (4.53) comes from. From the consistency condition, we must require

$$\alpha + \beta + \gamma = 0, \text{ and } -\alpha + \gamma = 1. \quad (4.54)$$

These conditions ensure first-order accuracy. If we further require second-order accuracy *in both time and space*, we must require that

$$\frac{\Delta t}{2} c^2 + \frac{c \Delta x}{2} (\alpha + \gamma) = 0. \quad (4.55)$$

Here we have used

$$u_{tt} = c^2 u_{xx}, \quad (4.56)$$

which follows from the exact advection equation provided that c is constant. Solving, we get

$$\alpha = \frac{-1-\mu}{2}, \beta = \mu \text{ and } \gamma = \frac{1-\mu}{2}, \quad (4.57)$$

where, as usual, $\mu \equiv \frac{c \Delta t}{\Delta x}$.

The scheme can be written as

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{2 \Delta x} (u_{j+1}^n - u_{j-1}^n) = \frac{c^2 \Delta t}{2 \Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n) \quad (4.58)$$

Note that although (4.58) is second-order accurate in time, it involves only two time levels. On the other hand, the scheme achieves second-order accuracy in space through the use of three grid points. Compare (4.58) with (4.51).

In this example, we used three grid points $(j-1, j, j+1)$ to construct our approximation to $\frac{du}{dx}$, each with a coefficient (α, β, γ) . In a similar way, we could have used any number of grid points, each with a suitably chosen coefficient, to construct a scheme of arbitrary accuracy in both space and time. The result would still be a two-time-level scheme! This illustrates that *a non-iterative two-level scheme is not necessarily a first-order scheme*.

Eq. (4.58) was proposed by Lax and Wendroff (1960), and recommended by Richtmeyer (1963). The right-hand-side of (4.58) looks like a diffusion term. It tends to smooth out small-scale noise. The Lax-Wendroff scheme is equivalent to and can be

interpreted in terms of the following procedure: First calculate $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ and $u_{j-\frac{1}{2}}^{n+\frac{1}{2}}$ from

$$\frac{u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}(u_{j+1}^n + u_j^n)}{\frac{1}{2}\Delta t} = -c \left(\frac{u_{j+1}^n - u_j^n}{\Delta x} \right), \quad (4.59)$$

$$\frac{u_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}(u_j^n + u_{j-1}^n)}{\frac{1}{2}\Delta t} = -c \left(\frac{u_j^n - u_{j-1}^n}{\Delta x} \right), \quad (4.60)$$

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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c \left(\frac{u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \right). \quad (4.61)$$

Note that (4.62) is “centered in time.” If (4.59) and (4.60) are substituted into (4.61), we recover (4.58). This derivation helps us to rationalize why it is possible to obtain second-order accuracy in time with this two-time-level scheme.

To test the stability of the Lax-Wendroff scheme, we use von Neumann's method. Assuming $u_j = \hat{u} e^{ikj\Delta x}$ in (4.58), we get

$$\hat{u}^{n+1} - \hat{u}^n = -\mu \left[i \sin(k\Delta x) + 2\mu \sin^2\left(\frac{k\Delta x}{2}\right) \right] \hat{u}^n. \quad (4.62)$$

Here we have used the trigonometric identity $2\sin^2\left(\frac{\theta}{2}\right) = 1 - \cos\theta$. The amplification factor is

$$\lambda = 1 - 2\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) - i\mu \sin(k\Delta x), \quad (4.63)$$

so that

$$\begin{aligned}
|\lambda| &= \left\{ \left[1 - 4\mu^2 \sin^2\left(\frac{k\Delta x}{2}\right) + 4\mu^4 \sin^4\left(\frac{k\Delta x}{2}\right) \right] + \mu^2 \sin^2(k\Delta x) \right\}^{\frac{1}{2}} \\
&= \left[1 - 4\mu^2(1 - \mu^2) \sin^4\left(\frac{k\Delta x}{2}\right) \right]^{\frac{1}{2}}.
\end{aligned} \tag{4.64}$$

If $|\mu| < 1$, $|\lambda| < 1$ and the scheme is dissipative. Fig. 4.3 shows how $|\lambda|^2$ depends on μ and L . The scheme strongly but selectively damps the short waves. Compare with the

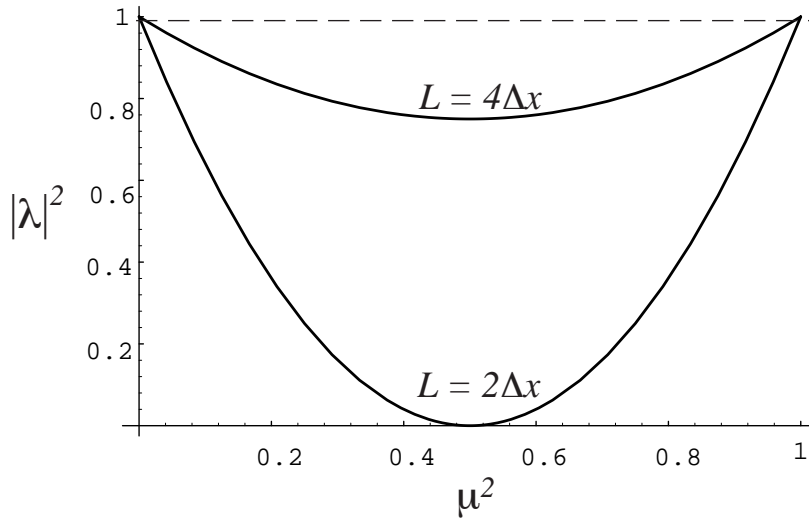


Figure 4.3: The amplification factor for the Lax-Wendroff scheme, for two different wavelengths, plotted as a function of μ^2 . Compare with Fig. 2.4.

corresponding plot for the upstream scheme, given earlier. The Lax-Wendroff scheme is comparably dissipative for the shortest wave, but less dissipative for the longer waves.

There are also various implicit schemes, such as the trapezoidal implicit scheme, which are neutral and unconditionally stable, so that in principle any Δt can be used if the error in phase can be tolerated. Such implicit schemes have the drawback that an iterative procedure may be 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 Δt .

4.4 Computational dispersion

Consider the differential-difference equation

$$\frac{du_j}{dt} + c \left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} \right) = 0. \tag{4.65}$$

Using $u_j = \hat{u} e^{ikj\Delta x}$, as before, we can write (4.65) as

$$\frac{d\hat{u}_j}{dt} + cik \frac{\sin(k\Delta x)}{k\Delta x} \hat{u}_j = 0. \quad (4.66)$$

If we had retained the differential form (4.1), we would have obtained $\frac{\partial \hat{u}}{\partial t} + cik \hat{u} = 0$.

Comparison with (4.66) shows that the phase speed is not simply c , as with (4.1), but c^* , given by

$$c^* \equiv c \frac{\sin(k\Delta x)}{k\Delta x}. \quad (4.67)$$

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. 4.4. (The second (lower) curve in the

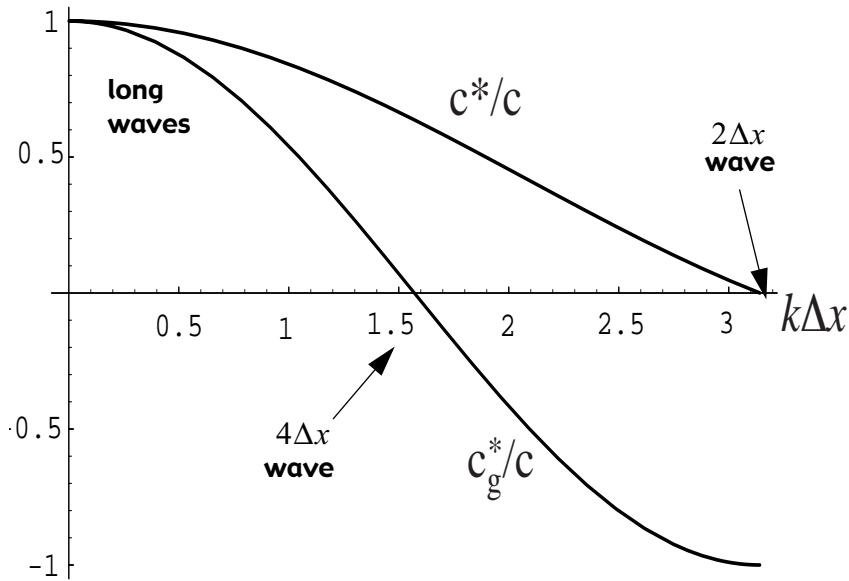


Figure 4.4: The ratio of the computational phase speed to the true phase speed, and also the ratio of the computational group speed to the true group speed, both plotted as functions of wave number.

figure, which illustrates the computational group velocity, is discussed later.)

If k_s is defined by $k_s \Delta x = \pi$, 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 $0 < k\Delta x < \pi$. 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 , 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.

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

$$k \equiv \frac{k_1 + k_2}{2}, \quad c \equiv \frac{c_1 + c_2}{2}, \quad \Delta k \equiv \frac{k_1 - k_2}{2}, \quad \Delta(kc) \equiv \frac{k_1 c_1 - k_2 c_2}{2}. \quad (4.68)$$

See Fig. 4.5. Note that $k_1 = k + \Delta k$, and $k_2 = k - \Delta k$. Similarly, $c_1 = c + \Delta c$.

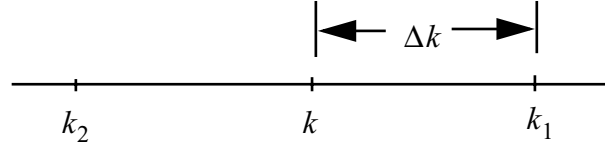


Figure 4.5: Sketch defining notation used in the discussion of the group velocity.

$c_2 = c - \Delta c$. You should be able to show that

$$k_1 c_1 \cong kc + \Delta(kc) \text{ and } k_2 c_2 \cong kc - \Delta(kc). \quad (4.69)$$

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 (4.69), we can write the sum of two waves, each with the same amplitude, as

$$\begin{aligned} & \exp[ik_1(x - c_1 t)] + \exp[ik_2(x - c_2 t)] \\ & \cong \exp(i\{(k + \Delta k)x - [kc + \Delta(kc)]t\}) + \exp(i\{(k - \Delta k)x - [kc - \Delta(kc)]t\}) \\ & = \exp[ik(x - ct)](\exp\{i[\Delta kx - \Delta(kc)]t\} + \exp\{-i[\Delta kx - \Delta(kc)]t\}) \\ & = 2 \cos[\Delta kx - \Delta(kc)t] \exp[ik(x - ct)] \\ & = 2 \cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\} \exp[ik(x - ct)] \quad . \end{aligned} \quad (4.70)$$

If Δk is small, the factor $\cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\}$ may appear schematically as the outer, slowly varying envelope in Fig. 4.6. 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 ,

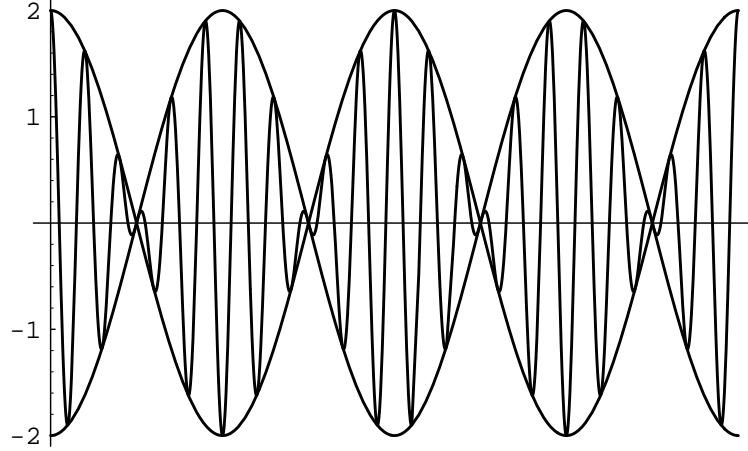


Figure 4.6: Sketch used to illustrate the concept of group velocity. The short waves are modulated by longer waves.

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 the problem at hand, i.e., advection, the “right answer” is $c_g = c$, i.e. the group velocity and phase velocity are the same. For this reason, we usually do not discuss the group velocity for advection.

With our finite-difference scheme, however, we have

$$c_g^* = \frac{d(kc^*)}{dk} = c \frac{d\left(\frac{\sin k\Delta x}{\Delta x}\right)}{dk} = c \cos k\Delta x. \quad (4.71)$$

A plot of c_g^* versus $k\Delta x$ is given in Fig. 4.4. 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. 4.7, which is a modified version of Fig. 4.6, illustrates this phenomenon in a different way, for the particular case $L = 2\Delta x$. Consider the upper solid curve and the thick

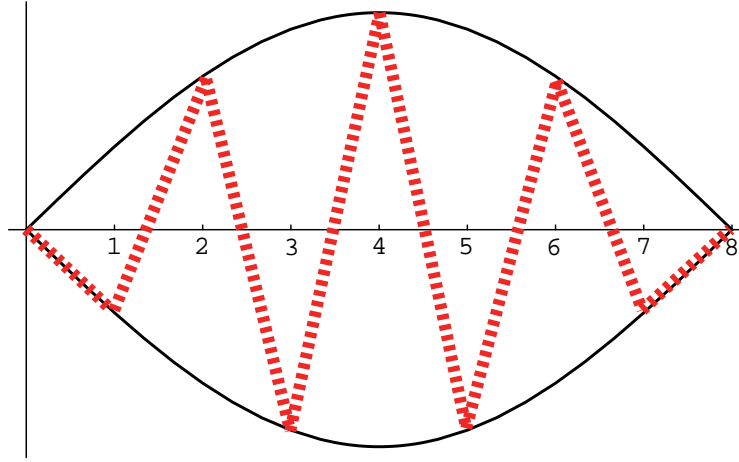


Figure 4.7: Yet another sketch used to illustrate the concept of group velocity. The short wave has wavelength $2\Delta x$.

dashed curve. If we denote points on the thick dashed curve (corresponding to our solution with $L = 2\Delta x$) by u_j , and points on the upper solid curve (the envelope of the thick dashed curve, moving with speed c_g^*) by U_j , we see that

$$U_j = (-1)^j u_j. \quad (4.72)$$

(This is true only for the particular case $L = 2\Delta x$.) Using (4.72), Eq. (4.65) can be rewritten as

$$\frac{\partial U_j}{\partial t} + (-c) \left(\frac{U_{j+1} - U_{j-1}}{2\Delta x} \right) = 0. \quad (4.73)$$

Eq.(4.73) shows that the upper solid curve will move with speed $-c$.

Recall that when we introduce time differencing, the computed phase change per time 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 $\Delta t \rightarrow 0$.

We now present an *analytical* solution of (4.65), which illustrates dispersion error in a very clear way, following an analysis by Matsuno (1966). If we write (4.65) in the form

$$2 \frac{du_j}{d\left(\frac{tc}{\Delta x}\right)} = u_{j-1} - u_{j+1}, \quad (4.74)$$

and define a non-dimensional time $\tau \equiv \frac{tc}{\Delta x}$, we obtain

$$2 \frac{du_j}{d\tau} = u_{j-1} - u_{j+1}. \quad (4.75)$$

This is a recursion formula satisfied by the Bessel functions of the first kind of order j , which are usually denoted by $J_j(\tau)$. (See any handbook of functions.) These functions have the property that $J_0(0) = 1$, and $J_j(0) = 0$ for $j \neq 0$. Because the $J_j(\tau)$ satisfy (4.75), each $J_j(\tau)$ represents the solution at a particular grid point, j , as a function of the nondimensional time, τ .

As an example, set $u_j = J_j(\tau)$, which is consistent with and in fact implies the initial conditions that $u_0(0) = 1$ and $u_j(0) = 0$ for all $j \neq 0$. This initial condition is an isolated “spike” at $j = 0$. The solution of (4.75) for the points $j = 0, 1$, and 2 is illustrated in Fig. 4.8. By using the identity

$$J_{(-j)} = (-1)^j J_j, \quad (4.76)$$

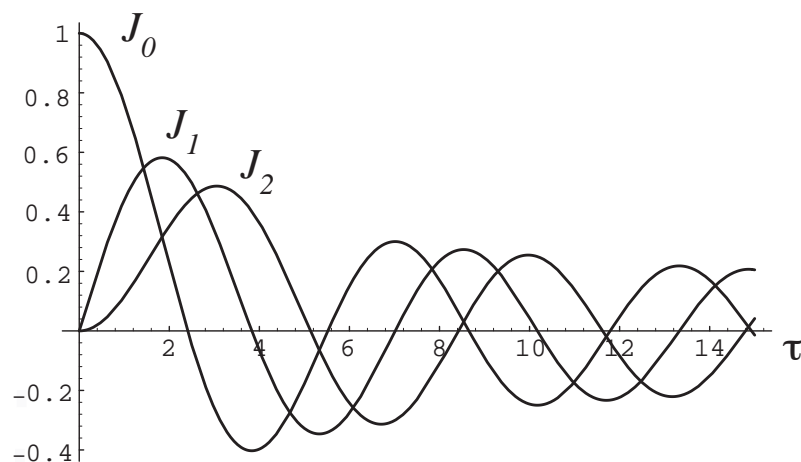


Figure 4.8: The time evolution of the solution of (4.75) at grid points $j = 0, 1$, and 2 .

we can obtain the solution at the points $j = -1, -2, -3$, etc.

The solution of (4.75) for $\tau = 5$ and $\tau = 10$, for $-15 \leq j \leq 15$, with these “spike” initial conditions, is shown in Fig. 4.9, which is taken from a paper by Matsuno (1966). Computational dispersion, schematically illustrated earlier in Fig. 4.4 and Fig. 4.7, is seen directly here. The figure also shows that c_g is negative for the shortest wave.

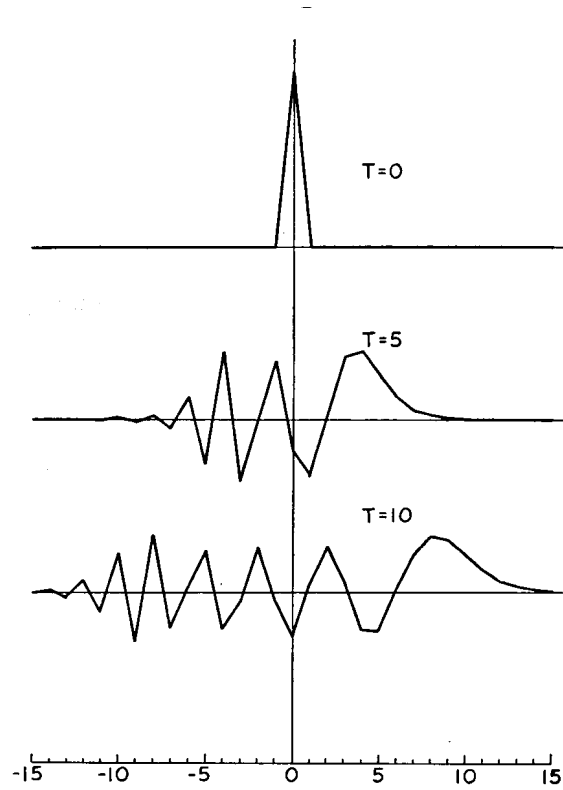


Figure 4.9: The solution of (4.75) for $t = 5$ and $t = 10$ for j in the range -15 to 15 , with “spike” initial conditions. From Matsuno (1966).

A similar type of solution is shown in Fig. 4.10, which is taken from a paper by Wurtele (1961). Here the initial conditions are slightly different, namely,

$$u_{-1} = 1, \quad u_0 = 1, \quad u_1 = 1 \quad \text{and} \quad u_j = 0 \quad \text{for} \quad j \leq -2, \quad j \geq 2.$$

This is a “top hat” or “box” initial condition. We can construct it by combining

$$J_{j-1}(0) = 1 \quad \text{for} \quad j = 1 \quad \text{and zero elsewhere,}$$

$$J_j(0) = 1 \quad \text{for} \quad j = 0 \quad \text{and zero elsewhere,}$$

$$J_{j+1}(0) = 1 \quad \text{for} \quad j = -1 \quad \text{and zero elsewhere,}$$

so that the full solution is given by

$$u_j(\tau) = J_{j-1}(\tau) + J_j(\tau) + J_{j+1}(\tau). \quad (4.77)$$

Dispersion is evident again in Fig. 4.10. The dashed curve is for centered space differencing, and the solid curve is for an uncentered scheme, which corresponds to the upstream scheme.

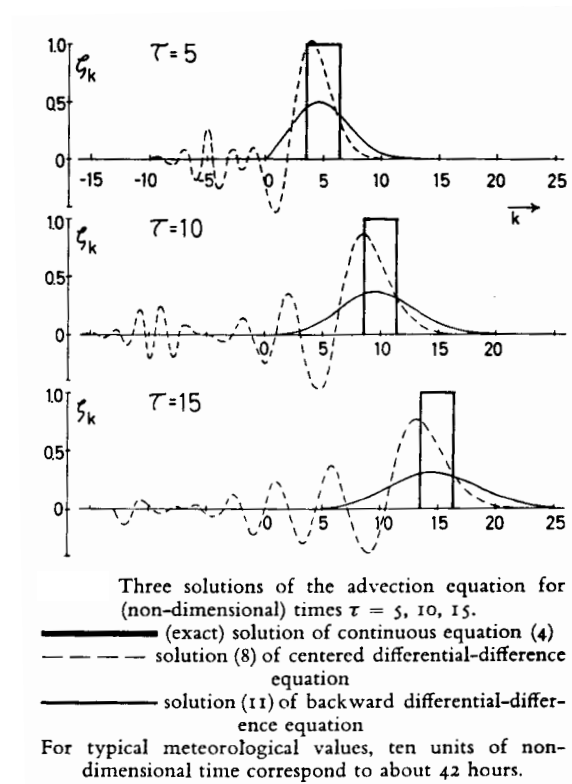


Figure 4.10: The solution of (4.72) with “box” initial conditions. From Wurtele (1961).

(The solution for the uncentered 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. A noisy solution is not necessarily unstable. At least initially, both dispersion and instability can lead to “noise.” In the case of dispersion, the waves are not growing in amplitude, but are becoming separated from one another (“dispersing”), each at its own speed.

4.5 The effects of fourth-order space differencing on the phase speed

As discussed in Chapter 2, the fourth-order difference quotient takes the form

$$\left(\frac{\partial u}{\partial x}\right)_j = \frac{4}{3} \left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} \right) - \frac{1}{3} \left(\frac{u_{j+2} - u_{j-2}}{4\Delta x} \right) + O[(\Delta x)^4] \quad (4.78)$$

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

$$c^* = c \left(\frac{\sin k\Delta x}{k\Delta x} \right). \quad (4.79)$$

Now we can also derive a similar equation for this fourth-order scheme. It is

$$c^* = c \left(\frac{4}{3} \frac{\sin k\Delta x}{k\Delta x} - \frac{1}{3} \frac{\sin 2k\Delta x}{2k\Delta x} \right). \quad (4.80)$$

Fig. 4.11 shows a graph of c^*/c versus $k\Delta x$ for each scheme. We see that 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.

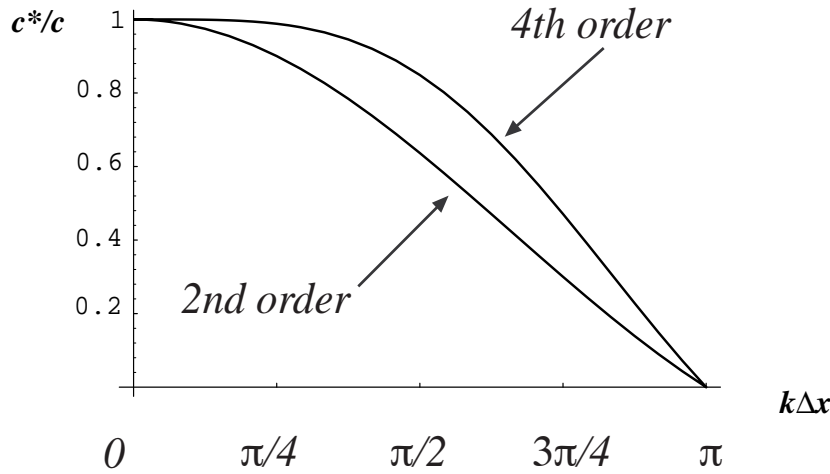


Figure 4.11: The ratio of the computational phase speed, c^* , to the true phase speed, c , plotted as a function of $k\Delta x$, for the second-order and fourth-order

4.6 Space-uncentered schemes

One way in which computational dispersion can be reduced in the numerical solution of (4.1) is to use uncentered space differencing, as, for example, in the upstream scheme. Recall that in Chapter 2 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. 4.12.

The “upstream scheme,” given by

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c \left(\frac{u_j^n - u_{j-1}^n}{\Delta x} \right) = 0 \text{ for } c > 0. \quad (4.81)$$

or

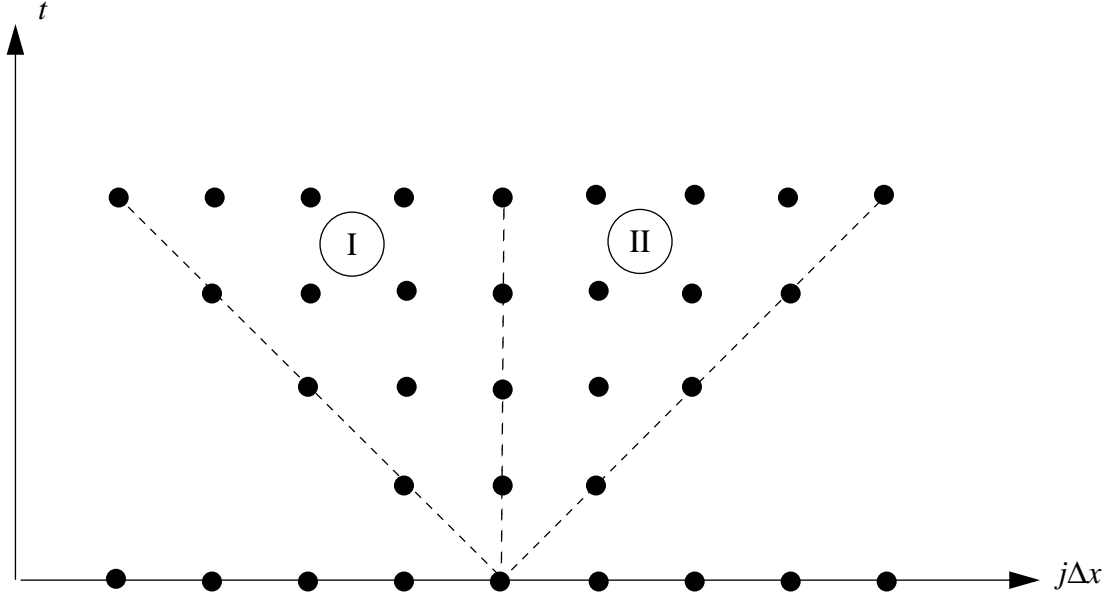


Figure 4.12: The domain of influence for explicit non-iterative space-centered schemes expands in time, as is shown by the union of Regions I and II.

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c \left(\frac{u_{j+1}^n - u_j^n}{\Delta x} \right) = 0 \text{ for } c < 0. \quad (4.82)$$

is an example of a space-uncentered scheme, and has already been discussed. As shown earlier, we can write (4.82) as

$$u_j^{n+1} = (1 - \mu)u_j^n + \mu u_{j-1}^n, \quad (4.83)$$

which has the form of an interpolation. 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. 4.10. 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.

As discussed earlier, the stability condition for the upstream scheme is $\mu \leq 1$. When this stability condition is met, (4.83) guarantees that

$$\text{Min}\{u_j^n, u_{j-1}^n\} \leq u_j^{n+1} \leq \text{Max}\{u_j^n, u_{j-1}^n\}. \quad (4.84)$$

This means that u_j^{n+1} cannot be smaller than the smallest value of u_j^n , or larger than the largest value of u_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 u if none are present initially, and neither can the upstream scheme, provided that $\mu \leq 1$. 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. This means that it cannot produce new maxima or minima, like those associated with the dispersive ripples seen in Fig. 4.10. This monotone property is expressed by (4.84).

It is easy to show that all monotone schemes are sign-preserving schemes. The converse is not true.

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 q , so that

$$\sum_i q_i^n = \sum_i q_i^0, \quad (4.85)$$

where the sum represents a sum over the whole spatial domain, the superscripts $n > 0$ and 0 represent two time levels. For simplicity we indicate only a single spatial dimension here, but the following argument holds for any number of dimensions. Suppose that the scheme that takes us from q^0 to q^n through n time steps is sign-preserving and conserves q . If q_i^0 is of one sign everywhere, it follows from (4.85) that

$$\sum_i |q_i^n| = \sum_i |q_i^0|. \quad (4.86)$$

Recall that for an arbitrary variable A , we have

$$\sum_i (A_i)^2 \geq \left(\sum_i |A_i| \right)^2. \quad (4.87)$$

Then from (4.86) and (4.87) we see that

$$\sum_i (q_i^n)^2 \leq \left(\sum_i |q_i^0| \right)^2. \quad (4.88)$$

Note that the right-hand side of (4.88) is a constant. Eq. (4.88) demonstrates that $(q_i^n)^2$ is bounded for all time, and so it proves stability by the energy method discussed in Chapter 2. The essence of (4.88) is that there is an upper bound on $\sum_i (q_i^n)^2$. This bound is rather weak, however; try some numbers to see for yourself. So, although (4.88) does demonstrate

absolute stability, it does not ensure good behavior!

In the preceding discussion we assumed that q_i^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 q into positive and negative parts, i.e.

$$q = q^+ + q^- . \quad (4.89)$$

We can consider that q^+ is positive where q is positive, and zero elsewhere; and similarly that q^- is negative where q is negative, and zero elsewhere. The total of q is then the sum of the two parts, as stated in (4.89). Advection of q is equivalent to advection of q^+ and q^- separately. If we apply a sign-preserving scheme to each part, then each of these two advectons is stable by the argument given above, and so the advection of q 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 (4.13) with either the geometric mean given by (4.34), or the harmonic mean given by (4.35). Both of these schemes have the property that as either A_j or A_{j+1} goes to zero, $A_{j+\frac{1}{2}}$ also 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, although they do tend to minimize it.

Better results can be obtained as follows: Replace (4.13) by

$$\frac{\frac{d}{dt}(m_j A_j) + \left[(mu)^+_{j+\frac{1}{2}} A^+_{j+\frac{1}{2}} + (mu)^-_{j+\frac{1}{2}} A^-_{j+\frac{1}{2}} \right] - \left[(mu)^+_{j-\frac{1}{2}} A^+_{j-\frac{1}{2}} + (mu)^-_{j-\frac{1}{2}} A^-_{j-\frac{1}{2}} \right]}{\Delta x_j} = 0 , \quad (4.90)$$

where

$$(mu)^+_{j+\frac{1}{2}} \equiv \frac{(mu)_{j+\frac{1}{2}} + |(mu)_{j+\frac{1}{2}}|}{2} \geq 0 , \quad (4.91)$$

$$(mu)^-_{j+\frac{1}{2}} \equiv \frac{(mu)_{j+\frac{1}{2}} - |(mu)_{j+\frac{1}{2}}|}{2} \leq 0, \quad (4.92)$$

$$A^+_{j+\frac{1}{2}} \equiv \frac{2A_j^{n+1}A_{j+1}^n}{(A_j^n + A_{j+1}^n)}, \quad (4.93)$$

$$A^-_{j+\frac{1}{2}} \equiv \frac{2A_j^nA_{j+1}^{n+1}}{(A_j^n + A_{j+1}^n)}. \quad (4.94)$$

Here we have introduced $(mu)^+_{j+\frac{1}{2}}$ as the mass flux in the “+x” direction, into cell $j + 1$ from cell j , and $(mu)^-_{j+\frac{1}{2}}$ as the mass flux in the “-x” direction, out of cell $j + 1$ and into cell j . We can associate different interpolated A s with the mass fluxes in the two directions.

Note that this scheme is implicit and must be solved as a coupled system over all grid points, but it remains linear. The “upstream” values of A are implicit in the numerators of (4.93) and (4.94), while the “downstream” values are explicit. Care must be taken to avoid division by zero when both A_{j+1}^n and A_j^n are zero; in such a case we simply set $A^+_{j+\frac{1}{2}} = A^-_{j+\frac{1}{2}} = 0$.

4.7 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 negative water on the mesh:

$$N \equiv \sum_{\text{where } q_j < 0} m_j q_j \leq 0 . \quad (4.95)$$

Here q_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 $m_j q_j$ is the mass of water in the cell. Note that m_j is *not* the density of dry air in the cell; rather it is the product of the density of dry air and the volume of the cell. The total amount of water on the mesh is given by

$$T \equiv \sum_{\text{all points}} m_j q_j . \quad (4.96)$$

Both T and N have the dimensions of mass. Define the nondimensional ratio

$$\Phi \equiv \frac{T + N}{T} \leq 1 ; \quad (4.97)$$

normally Φ is just very slightly less than one, because there are only a few holes and they are not very “deep”. We replace all negative values of q_j by zero, and then set

$$q_j^{\text{new}} = \Phi q_j . \quad (4.98)$$

In this way, we are ensured of the following:

- No negative values of q_j remain on the mesh.
- The total mass of water in the adjusted state is the same as that in the “holy” state.
- Water is borrowed most heavily from grid cells with large mixing ratios, and least from cells with small mixing ratios.

Hole-filling is ugly. Any hole-filling procedure is necessarily somewhat arbitrary, because we cannot mimic any natural process; nature has no holes to fill. In addition, hole-filling tends to be “quasi-diffusive” because it remove water from wet cells and adds it to dry cells, so that it reduces 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, like a Maytag repairman, the hole-filler will have very little work to do.

4.8 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, abstracted from the paper of Zalesak (1979), indicates how this is done.

Monotone and sign-preserving schemes can be derived by using the approach of “flux-corrected transport,” sometimes 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

$$\Psi_i^{n+1} = \Psi_i^n - \left(FH_{i+\frac{1}{2}} - FH_{i-\frac{1}{2}} \right). \quad (4.99)$$

Here FH represents the “high-order” fluxes associated with the scheme. Note that (4.99) is in “conservation” form, and that forward time-differencing has been used. Suppose that we have at our disposal a monotone or sign-preserving low-order scheme, whose fluxes are denoted by $FL_{i+\frac{1}{2}}$. 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

$$FH_{i+\frac{1}{2}} \equiv FL_{i+\frac{1}{2}} + A_{i+\frac{1}{2}}. \quad (4.100)$$

Here $A_{i+\frac{1}{2}}$ is a “residual” flux, sometimes called an “anti-diffusive” flux. Eq. (4.100) is

essentially the definition of $A_{i+\frac{1}{2}}$. According to (4.100), the high-order flux is the low-order

flux plus an anti-diffusive 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 it is monotone. 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 Ψ_i^{n+1} , i.e.

$$\Psi_i^{n+1} = \Psi_i^n - \left(FL_{i+\frac{1}{2}} - FL_{i-\frac{1}{2}} \right). \quad (4.101)$$

Since by assumption the low-order scheme is monotone, we know that

$$\Psi_i^{MAX} \geq \Psi_i^{n+1} \geq \Psi_i^{MIN}, \quad (4.102)$$

where Ψ_i^{MAX} and Ψ_i^{MIN} are suitably chosen upper and lower bounds, respectively, on the value of Ψ within the grid-box in question. For instance, Ψ_i^{MIN} might be zero, if Ψ is a non-negative scalar like the water vapor mixing ratio. Other possibilities will be discussed below.

There are two important points in connection with the inequalities in (4.102). 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 (4.99), (4.100), and (4.101) it is easy to see that

$$\psi_i^{n+1} = \Psi_i^{n+1} - \left(A_{i+\frac{1}{2}} - A_{i-\frac{1}{2}} \right). \quad (4.103)$$

This simply says that we can obtain the high-order solution from the low-order solution by adding the anti-diffusive fluxes.

We now define some coefficients, denoted by $C_{i+\frac{1}{2}}$, and “scaled-back” anti-diffusive fluxes, denoted by $\hat{A}_{i+\frac{1}{2}}$, such that

$$\hat{A}_{i+\frac{1}{2}} \equiv C_{i+\frac{1}{2}} A_{i+\frac{1}{2}}. \quad (4.104)$$

In place of (4.103), we use

$$\psi_i^{n+1} = \Psi_i^{n+1} - \left(\hat{A}_{i+\frac{1}{2}} - \hat{A}_{i-\frac{1}{2}} \right). \quad (4.105)$$

To see the idea, consider two limiting cases. If $C_{i+\frac{1}{2}} = 1$, then $\hat{A}_{i+\frac{1}{2}} = A_{i+\frac{1}{2}}$, and so (4.105) will reduce to (4.103) and so will simply give the high-order solution. If $C_{i+\frac{1}{2}} = 0$, then $\hat{A}_{i+\frac{1}{2}} = 0$, and so (4.105) will simply give the low-order solution. We enforce

$$0 \leq C_{i+\frac{1}{2}} \leq 1, \quad (4.106)$$

and try to make $C_{i+\frac{1}{2}}$ as close to one as possible, so that we get as much as possible of the high-order scheme and as little as possible of the low-order scheme, but we require that

$$\psi_i^{MAX} \geq \psi_i^{n+1} \geq \psi_i^{MIN} \quad (4.107)$$

be satisfied. Compare (4.107) with (4.102). We can always ensure that (4.107) will be satisfied by taking $C_{i+\frac{1}{2}} = 0$; this is the “worst case.” Quite often it may happen that (4.107) is

satisfied for $C_{i+\frac{1}{2}} = 1$; that is the “best case.”

It remains to specify the upper and lower bounds that appear in (4.107) and (4.102). Zalesak (1979) proposed limiting ψ_i^{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

$$\psi_i^{MAX} = \text{Max}\{\psi_{i-1}^n, \psi_i^n, \psi_{i+1}^n, \Psi_{i-1}^{n+1}, \Psi_i^{n+1}, \Psi_{i+1}^{n+1}\}, \quad (4.108)$$

and

$$\psi_i^{MIN} = \text{Min}\{\psi_{i-1}^n, \psi_i^n, \psi_{i+1}^n, \Psi_{i-1}^{n+1}, \Psi_i^{n+1}, \Psi_{i+1}^{n+1}\}. \quad (4.109)$$

Smolarkiewicz (1991) shows how (4.108) and (4.109) can be combined with (4.106) and (4.107) to obtain the largest feasible anti-diffusive fluxes.

The “limiter” denoted by (4.108) and (4.109) is not unique. Other possibilities are discussed by Smolarkiewicz (1991).

Our analysis of FCT schemes has been given in terms of one spatial dimension, but all of the discussion given above can very easily be extended to two or three dimensions, without time splitting. The literature on FCT schemes is very large and rapidly growing, although their application to atmospheric science is still fairly new.

FCT schemes are, philosophically, not that different from hole-fillers. The high-order scheme makes a hole, and the low-order scheme is used to fill it, immediately, before the end of the time step. Hole? What hole?

4.9 Lagrangian schemes

Lagrangian schemes, in which particles are tracked through space without the use of an Eulerian grid, have been used in the atmospheric sciences, as well as other fields including astrophysics and weapons physics (e.g. Mesinger, 1971; Trease, 1988; Monaghan, 1992; Norris, 1996; Haertel and Randall, 2001). 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 positive definite.
- 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.

On the other hand, Lagrangian schemes encounter a number of practical difficulties. Some of these problems 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 specify a way to compute a given field at any point in space, given the value of the field at a collection of particles that can be located anywhere in the domain. For an arbitrary field A , let

$$A(\mathbf{r}) = \int A(\mathbf{r}') W(\mathbf{r} - \mathbf{r}', h) d\mathbf{r}' , \quad (4.110)$$

where the integration is over the whole domain (e.g. the whole atmosphere), and W is an interpolating “kernel” such that

$$\int W(\mathbf{r} - \mathbf{r}', h) d\mathbf{r}' = 1 \quad (4.111)$$

and

$$\lim_{h \rightarrow 0} W(\mathbf{r} - \mathbf{r}', h) = \delta(\mathbf{r} - \mathbf{r}') , \quad (4.112)$$

where $\delta(\mathbf{r} - \mathbf{r}')$ is the Dirac delta function. In (4.110) - (4.112), h is a parameter, which is a measure of the “width” of W . We can interpret $W(\mathbf{r} - \mathbf{r}', h)$ 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

$$W(\mathbf{r} - \mathbf{r}', h) = \frac{e^{-[x(\mathbf{r} - \mathbf{r}')^2/h^2]}}{h\sqrt{\pi}} , \quad (4.113)$$

which can be shown to satisfy (4.112).

In a discrete model, we replace (4.110) by

$$A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} W(\mathbf{r} - \mathbf{r}_b, h) . \quad (4.114)$$

Here the index b denotes a particular particle, m_b is the mass of the particle, and ρ_b is the density of the particle. To see what is going on in (4.114), consider the case $A \equiv \rho$. Then

(4.114) reduces to

$$\rho(\mathbf{r}) = \sum_b m_b W(\mathbf{r} - \mathbf{r}_b, h), \quad (4.115)$$

which simply says that the density at a point \mathbf{r} is a weighted sum of the masses of particles in the vicinity of \mathbf{r} . In case there are no particles near the point \mathbf{r} , the density there will be small.

We can now perform spatial differentiation simply by taking the appropriate derivatives of $W(\mathbf{r} - \mathbf{r}_b, h)$, e.g.

$$\nabla A(\mathbf{r}) = \sum_b m_b \frac{A_b}{\rho_b} \nabla W(\mathbf{r} - \mathbf{r}_b, h). \quad (4.116)$$

This follows because m_b , A_b , and ρ_b are associated with particular particles and are, therefore, not functions of space.

Further discussion of SPH and related methods is given by Monaghan (1992) and the other references cited above.

4.10 Semi-Lagrangian schemes

Recently there has been considerable interest in a particular family of advection schemes called “semi-Lagrangian schemes” (e.g. Robert et al., 1985; Staniforth and Cote, 1991; Bates et al., 1993; Williamson and Olson, 1994). These schemes are of interest in part because they allow very long time steps, and in part 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 typically 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

$$(x_{\text{departure}})_j^n = x_j - c\Delta t. \quad (4.117)$$

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

$$x_{j-1} < (x_{\text{departure}})^n_j < x_j. \quad (4.118)$$

Then the simplest linear interpolation for q at the departure point is

$$\begin{aligned} (q_{\text{departure}})^n_j &= q^n_{j-1} + \left[\frac{(x_{\text{departure}})^n_j - x_{j-1}}{\Delta x} \right] (q^n_j - q^n_{j-1}) \\ &= q^n_{j-1} + \left(\frac{\Delta x - c\Delta t}{\Delta x} \right) (q^n_j - q^n_{j-1}) \\ &= q^n_{j-1} + (1 - \mu)(q^n_j - q^n_{j-1}) \\ &= \mu q^n_{j-1} + (1 - \mu)q^n_j \end{aligned} \quad (4.119)$$

Here we assume for simplicity that the mesh is spatially uniform, and $\mu \equiv \frac{c\Delta t}{\Delta x}$, as usual. The semi-Lagrangian scheme uses

$$q^{n+1}_j = (q_{\text{departure}})^n_j, \quad (4.120)$$

so we find that

$$q^{n+1}_j = \mu q^n_{j-1} + (1 - \mu)q^n_j. \quad (4.121)$$

This is simply the familiar upstream scheme. Note that (4.118), which was used in setting up the spatial interpolation, is equivalent to

$$0 < \mu < 1. \quad (4.122)$$

As shown earlier, this is the condition for stability of the upstream scheme.

What if (4.118) is not satisfied? This will be the case if the particle is moving quickly and/or the time step is long or, in other words, if $\mu > 1$. Then we might have, for example,

$$x_{j-a} < (x_{\text{departure}})^n_j < x_{j-a+1}, \quad (4.123)$$

where a is an *integer* greater than 1. For this case, we find in place of (4.119) that

$$(q_{\text{departure}})^n_j = \hat{\mu} q^n_{j-a} + (1 - \hat{\mu})q^n_{j-a+1}, \quad (4.124)$$

where

$$\hat{\mu} \equiv 1 - a + \mu . \quad (4.125)$$

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

$$0 \leq \hat{\mu} \leq 1 . \quad (4.126)$$

For $a = 1$, $\mu = \hat{\mu}$. Eq. (4.120) gives

$$q^{n+1}_j = \hat{\mu} q^n_{j-a} + (1 - \hat{\mu}) q^n_{j-a+1} . \quad (4.127)$$

It is easy to prove that we still have computational stability. This means that *the semi-Lagrangian scheme is computationally stable regardless of the size of the time step*. You should also be able to see that the scheme is monotone.

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, the strength of this computational diffusion can be reduced, although it cannot be eliminated completely.

Is the semi-Lagrangian 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 version of the advective form 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 advective form is used in their derivations, most semi-Lagrangian schemes are not conservative.

4.11 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 q be an arbitrary quantity advected, in two dimensions, by a constant basic current. The advection equation is

$$\frac{\partial q}{\partial t} + U \frac{\partial q}{\partial x} + V \frac{\partial q}{\partial y} = 0 , \quad (4.128)$$

where U and V are the x and y components of the current, respectively. Let i and j be the indices of grid points in the x and y directions. Replacing $\frac{\partial q}{\partial x}$ and $\frac{\partial q}{\partial y}$ by the corresponding centered difference quotients, we obtain

$$\frac{\partial q_{i,j}}{\partial t} + U \frac{1}{2\Delta x} (q_{i+1,j} - q_{i-1,j}) + V \frac{1}{2\Delta y} (q_{i,j+1} - q_{i,j-1}) = 0 . \quad (4.129)$$

Assume that q has the form

$$q_{ij} = R_e \left[Q(t) e^{i(kl\Delta x + lJ\Delta y)} \right], \quad (4.130)$$

where $i \equiv \sqrt{-1}$, and k and l are wave numbers in the x and y directions, respectively. Substitution gives the oscillation equation again:

$$\frac{dQ}{dt} = i\omega Q, \quad \omega \equiv -\left(U \frac{\sin k\Delta x}{\Delta x} + V \frac{\sin l\Delta y}{\Delta y} \right). \quad (4.131)$$

We have already analyzed the oscillation equation in detail, in Chapter 3. When we apply the leapfrog scheme, the stability criterion is $|\Omega| \leq 1$, where $\Omega \equiv \omega\Delta t$. Therefore, we must require

$$\left| U \frac{\sin k\Delta x}{\Delta x} + V \frac{\sin l\Delta y}{\Delta y} \right| \Delta t \leq 1. \quad (4.132)$$

Since

$$\left| U \frac{\sin k\Delta x}{\Delta x} + V \frac{\sin l\Delta y}{\Delta y} \right| \Delta t \leq \left(\left| U \frac{\sin k\Delta x}{\Delta x} \right| + \left| V \frac{\sin l\Delta y}{\Delta y} \right| \right) \Delta t \leq \left(\frac{|U|}{\Delta x} + \frac{|V|}{\Delta y} \right) \Delta t, \quad (4.133)$$

a condition sufficient to satisfy (4.132) is

$$\left(\frac{|U|}{\Delta x} + \frac{|V|}{\Delta y} \right) \Delta t \leq 1. \quad (4.134)$$

If we require the scheme to be stable for all possible k and l , and for all combinations of U and V , then (4.134) is also a necessary condition.

Put

$$|U| = C \cos \alpha, \quad |V| = C \sin \alpha, \quad (4.135)$$

where $0 \leq \alpha \leq \frac{\pi}{2}$. Note that with this definition C is the wind speed, and $C \geq 0$. For $\alpha = 0$, the flow is zonal, and for $\alpha = \pi/2$ it is meridional. Then (4.134) becomes

$$C \left(\frac{\cos \alpha}{\Delta x} + \frac{\sin \alpha}{\Delta y} \right) \Delta t \leq 1. \quad (4.136)$$

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

$$C \left(\frac{\cos \alpha_m}{\Delta x} + \frac{\sin \alpha_m}{\Delta y} \right) \Delta t \leq 1 , \quad (4.137)$$

where α_m is the “worst-case” α , which makes the left hand side of (4.136) a maximum. We can show that α_m satisfies

$$\tan \alpha_m = \frac{\Delta x}{\Delta y} . \quad (4.138)$$

As shown in Fig. 4.13, α_m measures the angle of the “diagonal” across a grid box. For

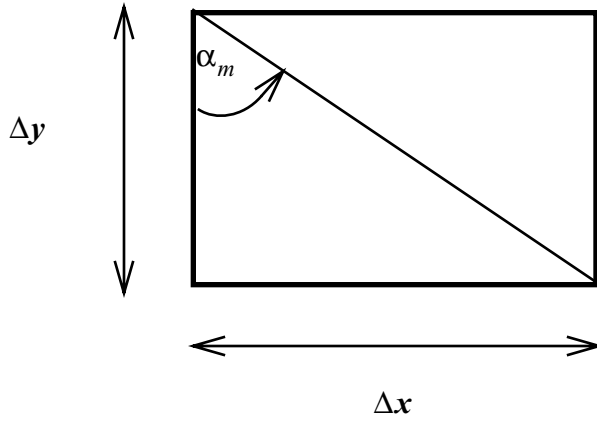


Figure 4.13: Sketch illustrating the angle α_m on a rectangular grid.

instance, when $\frac{\Delta y}{\Delta x} \ll 1$, α_m corresponds to a flow that is mainly meridional, because the worst case is the direction in which the grid cell is “narrowest.” As a second example, for $\Delta x = \Delta y$, we get $\alpha_m = \frac{\pi}{4}$. From (4.137) and (4.138) we see that the stability criterion can be written as

$$\frac{C \Delta t}{\sqrt{\Delta x^2 + \Delta y^2}} \left(\frac{\Delta y}{\Delta x} + \frac{\Delta x}{\Delta y} \right) \leq 1 . \quad (4.139)$$

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

$$\frac{C \Delta t}{d} \leq \frac{1}{\sqrt{2}} < 1 . \quad (4.140)$$

4.12 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 bugs 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 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 shortest 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.

Problems

1. Find a one-dimensional advection scheme that conserves both A and $\ln(A)$. Keep the time derivative continuous.
2. Adopt the continuity equation

$$\frac{\partial m_j}{\partial t} + \frac{(\hat{m}u)_{j+\frac{1}{2}} - (\hat{m}u)_{j-\frac{1}{2}}}{\Delta x} = 0 ,$$

and the advection equation

$$\frac{\partial A_j}{\partial t} + \frac{1}{2} \left(u_{j+\frac{1}{2}} + u_{j-\frac{1}{2}} \right) \left(\frac{A_{j+1} - A_{j-1}}{2\Delta x} \right) = 0 .$$

Determine whether or not this scheme conserves the mass-weighted average value of A .

3. Program the following one-dimensional model:

$$\frac{(hA)_j^{n+1} - (hA)_j^{n-1}}{2\Delta t} + \frac{(\hat{h}u)_{j+\frac{1}{2}}^n \hat{A}_{j+\frac{1}{2}}^n - (\hat{h}u)_{j-\frac{1}{2}}^n \hat{A}_{j-\frac{1}{2}}^n}{\Delta x} = 0 ,$$

$$\frac{h_j^{n+1} - h_j^{n-1}}{2\Delta t} + \frac{(\hat{h}u)_{j+\frac{1}{2}}^n - (\hat{h}u)_{j-\frac{1}{2}}^n}{\Delta x} = 0 ,$$

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^{n-1}}{2\Delta t} + \left(\frac{k_{j+1}^n - k_j^n}{\Delta x} \right) + g \left(\frac{h_{j+1}^n - h_j^n}{\Delta x} \right) = 0 .$$

Use a forward time step for the first step only. Take

$$\Delta x = 10^5 \text{ m} ,$$

$$g = 0.1 \text{ m s}^{-2} ,$$

$$\hat{h}_{j+\frac{1}{2}} = \frac{1}{2}(h_j + h_{j+1}) ,$$

$$k_j = \frac{1}{4} \left(u_{j+\frac{1}{2}}^2 + u_{j-\frac{1}{2}}^2 \right) .$$

Use 100 grid points, with periodic boundary conditions. Let the initial condition be

$$u_{j+\frac{1}{2}} = 0 \quad \text{for all } j ,$$

$$h_j = 1000 + 50 \cdot \sin\left(\frac{2\pi j}{20}\right) ,$$

$$A_j = 100 + 10 \cdot \cos\left(\frac{2\pi j}{4}\right) .$$

Use von Neuman's method to estimate the largest time step that is consistent with numerical stability. Experiment with time steps "close" (within a factor of 2) to

the predicted maximum stable Δt , in order to find a value that is stable in practice.

Run for the following two choices of $\hat{A}_{j+\frac{1}{2}}$:

$$\hat{A}_{j+\frac{1}{2}} = \frac{1}{2}(A_j + A_{j+1}) ,$$

$$\hat{A}_{j+\frac{1}{2}} = \sqrt{\text{Max}\{0, A_j A_{j+1}\}} .$$

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. Check conservation of A and A^2 for both cases. Explain how you do this.

4. Consider a domain $0 \leq j \leq 100$, with initial conditions

$$q_i = 100, 45 \leq i \leq 55,$$

$$q_i = 0 \text{ otherwise.}$$

Solve

$$\frac{\partial q}{\partial t} + c \frac{\partial q}{\partial x} = 0$$

using

1) Leapfrog in time, centered in space;

2) Lax Wendroff;

3) Upstream.

Choose $\mu = 0.7$ in each case. Compare the solutions.

5. The advective form of the finite-difference advection equation is:

$$m_j \frac{dA_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} \left(A_{j+\frac{1}{2}} - A_j \right) + (mu)_{j-\frac{1}{2}} \left(A_j - A_{j-\frac{1}{2}} \right)}{\Delta x} = 0 . \quad (4.141)$$

Here we have assumed that Δx is spatially constant. If the advecting mass flux is spatially constant, this reduces to

$$m_j \frac{dA_j}{dt} + mu \left(\frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x} \right) = 0 , \quad (4.142)$$

from which we see that we are using the approximation

$$\left(\frac{\partial A}{\partial x} \right)_j \cong \frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x} . \quad (4.143)$$

Suppose that we adopt

$$A_{j+\frac{1}{2}} = \frac{2A_j A_{j+1}}{(A_j + A_{j+1})} . \quad (4.144)$$

Determine the order of accuracy of the approximation (4.143), in case (4.144) is used for interpolation to the cell walls.

6. Discuss Eq. (4.30) for the case $F(A) = A$.

CHAPTER 5**Boundary-value problems**Copyright 2004 David A. Randall

5.1 Introduction

Boundary-value problems involve spatial derivatives and/or integrals, but no time derivatives and/or integrals. They can and do frequently arise in one, two, or three dimensions, in the atmospheric sciences. They can be solved by a wide variety of methods, which are discussed in standard texts on numerical analysis.

The solution of linear boundary-value problems is conceptually simple, but may nevertheless present challenges in practice. *The main issue in the numerical solution of boundary-value problems is how to minimize the amount of computational work that must be done to obtain the solution*, while at the same time minimizing amount of storage required. For the problems that arise in atmospheric science, and considering the characteristics of modern computers, maximizing computational speed is usually more of a concern than minimizing storage.

Two-dimensional linear boundary-value problems occur quite often in atmospheric science. A particularly ubiquitous example is the following. Consider a two-dimensional flow. Let ζ and δ be the vorticity and divergence, respectively. We can define a stream function, ψ , and a velocity potential, χ , by

$$\mathbf{V}_r = \mathbf{k} \times \nabla \psi, \quad (5.1)$$

and

$$\mathbf{V}_d = \nabla \chi, \quad (5.2)$$

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

$$\mathbf{V} = \mathbf{V}_r + \mathbf{V}_d. \quad (5.3)$$

The vorticity and divergence then satisfy

$$\zeta = \nabla^2 \psi \quad (5.4)$$

and

$$\delta = \nabla^2 \chi, \quad (5.5)$$

respectively.

Suppose that we are given the distributions of ζ and δ , and we need to determine the wind vector. This can be done by first solving the two boundary-value problems represented by (5.4)-(5.5), with suitable boundary conditions, then using (5.1)-(5.2) to obtain \mathbf{V}_r and \mathbf{V}_d , and finally using (5.3) to obtain the total horizontal wind vector.

A second example is the solution of the anelastic pressure equation.

Further examples arise from implicit time-differencing combined with space-differencing, e.g. for the diffusion equation or the shallow-water equations.

5.2 *Solution of one-dimensional boundary-value problems*

As a simple one-dimensional example, consider

$$\frac{d^2}{dx^2} q(x) = f(x), \quad (5.6)$$

on a periodic domain, where $f(x)$ is a given periodic function of x . Solution of (5.6) 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, e.g. the average of q over the domain may be prescribed.

The exact solution of (5.6) can be obtained by expanding $q(x)$ and $f(x)$ in infinite Fourier series. The individual Fourier modes will satisfy

$$-k^2 \hat{q}_k = \hat{f}_k, \quad (5.7)$$

which can readily be solved for the q_k , provided that k is not zero. The value of q_0 must be obtained directly from the second boundary condition mentioned above. The full solution for $q(x)$ can be obtained by Fourier summing the q_k .

This method to find the exact solution of (5.6) 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, it has both strengths and weaknesses. It will be discussed in a later chapter.

Suppose, however, that the problem posed by (5.6) 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 which are solved (approximately) through the use of spatial and temporal finite differences. In that case, the requirement of

consistency with the other equations of the model may dictate that the spatial derivatives in (5.6) be approximated by a finite-difference method, such as

$$\frac{q_{i+1} - 2q_i + q_{i-1}}{d^2} = f_i. \quad (5.8)$$

Here d is the grid spacing in the x -direction. We have used centered second-order spatial differences in (5.8). Assuming a periodic, wave-like solution for q_i , and correspondingly expanding f_i , we obtain, in the usual way,

$$-k^2 \hat{q}_k \left(\frac{\sin \frac{kd}{2}}{\frac{kd}{2}} \right)^2 = \hat{f}_k. \quad (5.9)$$

Note the similarity between (5.9) and (5.7). Clearly (5.9) can be solved to obtain each of the q_k , except q_0 , and the result will be consistent with the finite-difference approximation (5.8). This example illustrates that Fourier solution methods can be used even in combination with

finite-difference approximations. The factor of $-k^2 \left(\frac{\sin \frac{kd}{2}}{\frac{kd}{2}} \right)^2$ in (5.9) need only be evaluated

once and then stored, for each k , even if (5.9) must be solved on each of many time steps.

The 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 packages). It is easy to see that the method can be extended to two or three dimensions, provided only that the geometry of the problem is compatible with Fourier expansion.

There are other ways to solve (5.8). It can be regarded as a system of linear equations, in which the unknowns are the q_i . The matrix of coefficients is then “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:

$$\begin{bmatrix} d_1 & a_2 & 0 & 0 & 0 & b_6 \\ b_1 & d_2 & a_3 & 0 & 0 & 0 \\ 0 & b_2 & d_3 & a_4 & 0 & 0 \\ 0 & 0 & b_3 & d_4 & a_5 & 0 \\ 0 & 0 & 0 & b_4 & d_5 & a_6 \\ a_1 & 0 & 0 & 0 & b_5 & d_6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix}. \quad (5.10)$$

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. The solution of tri-diagonal linear systems is very fast and easy. For instance, the first of the six equations represented by (5.10) can be solved for x_1 as a function of x_2 , and x_6 , provided that $d_1 \neq 0$. This solution can be used to eliminate x_1 in the five remaining equations. The (modified version of the) second equation can then be solved for x_2 as a function of x_3 and x_6 , and this solution can be used to eliminate x_2 from the remaining four equations. Continuing in this way, we can ultimately obtain a single equation for the single unknown x_6 . Once the value of x_6 has been determined, we can obtain the other unknowns by back-substitution. It should be clear that the amount of arithmetic needed to implement this algorithm is simply proportional to the number of unknowns. This is good.

In case $d_1 = 0$ (assumed *not* to be true in the preceding discussion) we can immediately solve the first equation for x_2 in terms of x_6 , provided that a_2 is not also equal to zero.

Highly optimized versions of this simple tri-diagonal solver can be found in standard software libraries. Because tri-diagonal systems are easy to deal with, we are always happy when we can express a problem that we are working on as a tri-diagonal system. Naturally, tri-diagonal methods are not an option 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 methods work but they are very inefficient compared to the Fourier and tri-diagonal methods discussed above. The amount of arithmetic involved 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 (5.8) by a relaxation method. Here the idea is to make an “initial guess” for q_i , then refine the guess successively, until a “sufficiently good” approximation to the exact solution of (5.8) is obtained. Several relaxation methods are discussed later in this chapter.

5.3 *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).

As an example of a boundary-value problem, consider

$$\begin{aligned} -\nabla^2 u &= f \text{ in } \Omega, \\ u &= g \text{ on } \partial\Omega. \end{aligned} \quad (5.11)$$

Here Ω is a two-dimensional domain, and $\partial\Omega$ is the boundary of Ω . We consider f and g to be known.

We approximate (5.11) on a grid with uniform spacing $h = 1/N$ in both the x and y directions, with $N + 1$ grid points in each direction. Note that, with this definition, h is non-dimensional; in terms of dimensional quantities, h is the distance between grid points divided by the total width of the domain. Normally $h \ll 1$. Using second-order centered differences (for example), we write:

$$\begin{aligned} h^{-2}(4u_{j,k} - u_{j-1,k} - u_{j+1,k} - u_{j,k-1} - u_{j,k+1}) &= f_{j,k}, \quad 0 < (j,k) < N; \\ u_{j,k} &= g_{j,k}, \quad j = 0, j = N, k = 0, \text{ or } k = N. \end{aligned} \quad (5.12)$$

We now explore relaxation methods for the solution of (5.12). 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.”

We need a notation to distinguish approximate solutions from exact solutions. Here by “exact” solution we mean an exact solution to the finite-difference problem posed in (5.12).

We use a “hat” to denote the approximate solution, i.e. we let $\hat{u}_{j,k}$ denote an approximation to $u_{j,k}$.

The simplest relaxation method is called Jacobi relaxation or simultaneous relaxation. The Jacobi method defines the new value $\hat{u}_{j,k}^{new}$ by applying (5.12) with the new value at the point (j,k) and the “old” values at the neighboring points, i.e.

$$h^{-2}(4\hat{u}_{j,k}^{new} - \hat{u}_{j-1,k} - \hat{u}_{j+1,k} - \hat{u}_{j,k-1} - \hat{u}_{j,k+1}) = f_{j,k}, \quad (5.13)$$

or

$$\hat{u}_{j,k}^{new} = \frac{1}{4}(h^2 f_{j,k} + \hat{u}_{j-1,k} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1} + \hat{u}_{j,k+1}). \quad (5.14)$$

With this approach, we compute $\hat{u}_{j,k}^{new}$ at all interior points using (5.13), and then bodily replace the “old” approximate solution by the new one. This procedure is repeated until convergence is deemed adequate. Conditions for convergence are not discussed here.

Let the error of a given approximation be denoted by

$$v_{j,k} \equiv \hat{u}_{j,k} - u_{j,k}. \quad (5.15)$$

Here again $u_{j,k}$ is the exact solution of the finite-difference system. Consider one sweep of Jacobi relaxation. Substituting (5.15) into (5.14), we find that

$$\begin{aligned} v_{j,k}^{new} + u_{j,k} = \frac{1}{4} [& h^2 f_{j,k} + (v_{j-1,k} + v_{j+1,k} + v_{j,k-1} + v_{j,k+1}) \\ & + (u_{j-1,k} + u_{j+1,k} + u_{j,k-1} + u_{j,k+1})] \end{aligned} \quad (5.16)$$

Using (5.12), this can be simplified to

$$v_{j,k}^{new} = \frac{1}{4} (v_{j-1,k} + v_{j+1,k} + v_{j,k-1} + v_{j,k+1}). \quad (5.17)$$

This shows that the new error (after the sweep) is the average of the current errors (before the sweep) at the four surrounding points.

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., $v_{j,k} = 1$. For our assumed checkerboard error pattern, it follows that the errors at the neighboring points referenced on the right-hand side of (5.17) are all equal to -1. We conclude that $v_{j,k}^{new} = -1$. Then, on the next iteration, we will again obtain $v_{j,k} = 1$. You should be able to see that the checkerboard error pattern "flips sign" from one iteration to the next. The checkerboard error is never reduced to zero by Jacobi iteration.

A strategy to overcome this problem is to "under-relax." To understand this approach, we first re-write (5.14) as

$$\hat{u}_{j,k}^{new} = \hat{u}_{j,k} + \left[\frac{1}{4} (h^2 f_{j,k} + \hat{u}_{j-1,k} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1} + \hat{u}_{j,k+1}) - \hat{u}_{j,k} \right]. \quad (5.18)$$

This simply says that $\hat{u}_{j,k}^{new}$ is equal to $\hat{u}_{j,k}$ plus an "increment." For the checkerboard error the increment given by Jacobi relaxation is too large; this is why the sign flips from one iteration to the next. We can reduce the increment by multiplying it by a factor less than one, called ω , i.e., we replace (5.18) by

$$\hat{u}_{j,k}^{new} = \hat{u}_{j,k} + \omega \left[\frac{1}{4} (h^2 f_{j,k} + \hat{u}_{j-1,k} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1} + \hat{u}_{j,k+1}) - \hat{u}_{j,k} \right]. \quad (5.19)$$

where $0 < \omega < 1$. We can now rewrite (5.18) as

$$\hat{u}_{j,k}^{new} = \hat{u}_{j,k}(1 - \omega) + \frac{\omega}{4}(h^2 f_{j,k} + \hat{u}_{j-1,k} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1} + \hat{u}_{j,k+1}). \quad (5.20)$$

Substitution of (5.15) into (5.20), with the use of (5.12), gives

$$v_{j,k}^{new} = v_{j,k}(1 - \omega) + \frac{\omega}{4}[(v_{j-1,k} + v_{j+1,k} + v_{j,k-1} + v_{j,k+1})] \quad (5.21)$$

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.

Suppose that on a particular sweep the error is spatially uniform over the grid. Then, according to (5.17), the error will never change under Jacobi relaxation, and this is true even with under-relaxation, as can be seen from (5.21). This is not really a problem, however, because as discussed earlier when solving a problem of this type the average over the grid has to be determined by a boundary condition. For example, if the appropriate boundary condition can be applied at the time of formulating the first guess, then the domain-mean error will be zero even before the relaxation begins.

For errors of intermediate spatial scale, Jacobi relaxation works reasonably well.

5.4 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-most 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 u ’s that have already been updated, so that (5.14) is replaced by

$$\hat{u}_{j,k}^{new} = \frac{1}{4}(h^2 f_{j,k} + \hat{u}_{j-1,k}^{new} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1}^{new} + \hat{u}_{j,k+1}) . \quad (5.22)$$

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, relative to Jacobi relaxation.

Obviously (5.22) 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 (5.14). It is only for the second and later rows of points that (5.22) 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 (5.12) exactly, the final result will be independent of where the sweep starts.

For Gauss-Seidel relaxation, the error-reduction formula corresponding to (5.17) is

$$v_{j,k}^{new} = \frac{1}{4}(v_{j-1,k}^{new} + v_{j+1,k} + v_{j,k-1}^{new} + v_{j,k+1}) . \quad (5.23)$$

You should be able to see that with Gauss-Seidel iteration a checkerboard error is in fact reduced on each sweep. Consider the following simple example on a 6x6 mesh. Suppose that f is identically zero, so that the solution (with periodic boundary conditions) is that u is spatially constant. We make the rather ill-considered first guess that the solution is a checkerboard:

$$\hat{u}_{j,k}^0 = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} . \quad (5.24)$$

Here the superscript zero indicates 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:

$$\hat{u}_{j,k}^{1, \text{partial}} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ -0.5 & 0.25 & -0.281 & -1 & 1 & -1 \\ 1 & -0.5 & 0.625 & -0.593 & 0.602 & -0.60 \end{bmatrix} . \quad (5.25)$$

Although the solution is flipping sign as a result of the sweep, the amplitude of the checkerboard is decreasing significantly. You can finish the exercise for yourself.

5.5 Over-relaxation

The convergence of Gauss-Seidel relaxation can be accelerated by multiplying the increment by a factor *greater* than one. This is called “over-relaxation.” Corresponding to (5.18), we write

$$\hat{u}_{j,k}^{new} = \hat{u}_{j,k}(1 - \omega) + \frac{\omega}{4}(h^2 f_{j,k} + \hat{u}_{j-1,k}^{new} + \hat{u}_{j+1,k} + \hat{u}_{j,k-1}^{new} + \hat{u}_{j,k+1}) . \quad (5.26)$$

It can be shown that the convergence of (5.26) is optimized (i.e., made as rapid as possible) if we choose

$$\omega = \frac{2}{1 + \sin(\pi h)}. \quad (5.27)$$

The algorithm represented by (5.26) and (5.27) is called “successive over-relaxation,” or SOR. Choosing ω too large will cause the iteration to diverge. In practice, some experimentation may be needed to find the best value of ω .

5.6 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 (5.12) as

$$(-u_{j-1,k} + 2u_{j,k} - u_{j+1,k}) + (-u_{j,k-1} + 2u_{j,k} - u_{j,k+1}) = h^2 f_{j,k}. \quad (5.28)$$

The first quantity in parentheses on the left-hand side of (5.28) 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

$$[-\hat{u}_{j-1,k}^{int} + (2+r)\hat{u}_{j,k}^{int} - \hat{u}_{j+1,k}^{int}] + [-\hat{u}_{j,k-1}^{int} + (2-r)\hat{u}_{j,k}^{int} - \hat{u}_{j,k+1}^{int}] = h^2 f_{j,k} \quad (5.29)$$

for the values with superscript “*int*.” Here r is a parameter used to control convergence, as discussed below. Eq. (5.29) represents a set of *tri-diagonal* systems, each of which can easily be solved. The sweep is completed by solving

$$[-\hat{u}_{j,k-1}^{new} + (2-r)\hat{u}_{j,k}^{new} - \hat{u}_{j,k+1}^{new}] + [-\hat{u}_{j-1,k}^{int} + (2+r)\hat{u}_{j,k}^{int} - \hat{u}_{j+1,k}^{int}] = h^2 f_{j,k} \quad (5.30)$$

as a second set of N tridiagonal systems. It can be shown that the ADI method converges if r is positive and constant for all sweeps. The optimal value of r is

$$r = 2 \sin(\pi h). \quad (5.31)$$

5.7 Multigrid methods

Fulton et al. (1986) summarize the multi-grid approach to solving boundary-value problems. The basic idea is very simple, and comes from the fact that the largest-scale features in the error field (the difference between the approximate solution and the true solution) take the largest number of sweeps to eliminate.

As we have already discussed, with Gauss-Seidel relaxation the *small-scale errors are eliminated quickly, while the large-scale errors are removed more slowly*. As the iteration proceeds, the error field becomes smoother at the same time that it undergoes an overall decrease in amplitude.

Essentially by definition of “large-scale,” the large-scale errors can be represented on

a relatively coarse grid. On such a coarse grid, the large-scale errors actually appear to be of “smaller” scale, in the sense that they are represented by fewer grid points. The large-scale error can thus be removed quickly by relaxing on a coarse 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*, each coarse grid composed of a subset of the points used in the finer grids. The “multi” in the multi-grid method is quite important. We move back and forth between the grids, from coarse to fine by interpolation, and from fine to coarse by “injection” (copying) of the fine grid values onto the corresponding points of the coarse grid. A relaxation is done on each grid in turn. The relaxations on the coarser grids remove the large-scale part of the error, while the relaxations 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 on a single grid.

For further discussion of multi-grid methods, see the paper by Fulton et al. (1986).

5.8 Summary

Boundary-value problems occur quite frequently in atmospheric science. The main issue is not finding the solution, but rather finding it quickly. Fast solutions to one-dimensional problems are very easy to obtain, but two- and three-dimensional problems are more challenging, particularly when the geometry of the problem is complex. 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).

Table 5.1 summarizes the operations counts and storage requirements of some well

Table 5.1: Well known methods for solving boundary value problems, and the operation count and storage requirements for each, measured in terms of N^2 , the number of equations to be solved.

| Method | Operation Count | Storage Requirement |
|----------------------|-----------------|---------------------|
| Gaussian Elimination | N^4 | N^3 |
| Jacobi | N^4 | N^2 |
| Gauss-Seidel | N^4 | $< N^2$ |

Table 5.1: Well known methods for solving boundary value problems, and the operation count and storage requirements for each, measured in terms of N^2 , the number of equations to be solved.

| Method | Operation Count | Storage Requirement |
|--------------------------------------|-----------------|---------------------|
| Successive Over- Relaxation | N^3 | N^2 |
| Alternating Direction Implicit | $N^3 \ln N$ | N^2 |
| Multigrid | N^2 | N^2 |

known methods for solving boundary-value problems.

Problems

1. Consider a square domain, of width L , with periodic boundary conditions in both x and y directions. We wish to solve

$$\nabla^2 u = \left(\sin \frac{4\pi x}{L} \right) \left(\cos \frac{4\pi y}{L} \right) \quad (5.32)$$

for the unknown function $u(x, y)$, where

$$\begin{aligned} 0 &\leq x \leq L, \\ 0 &\leq y \leq L. \end{aligned} \quad (5.33)$$

Assume that the domain-average value of u is zero. For simplicity, use $L = 4\pi\sqrt{2}$. Use centered second-order differences to approximate $\nabla^2 u$. 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. For each of the relaxation methods, try the following two initial guesses:

$$\begin{aligned} 1) \quad &u_{j,k} = (-1)^{j+k}, \\ 2) \quad &u_{j,k} = 0 \text{ everywhere.} \end{aligned} \quad (5.34)$$

Jacobi relaxation;

Jacobi under-relaxation;

Gauss-Seidel relaxation

Gauss-Seidel over-relaxation.

- c) For Jacobi, Gauss-Seidel, and SOR, define the RMS error by

$$R^v \equiv \sqrt{\frac{1}{N^2} \sum_{j=1}^N \sum_{k=1}^N (\hat{u}_{j,k}^v - u_{j,k})^2} . \quad (5.35)$$

Here $\hat{u}_{j,k}^v$ is the approximate solution, and $u_{j,k}$ is the “exact” solution of the finite-difference problem. The parameter v is not an exponent; it is an “iteration counter,” i.e. $v = 0$ for the initial guess, $v = 1$ after one sweep, etc. Define the maximum absolute error by

$$M^v \equiv \text{Max} \forall (j, k) \{ |\hat{u}_{j,k}^v - u_{j,k}| \} . \quad (5.36)$$

Let the convergence criterion be

$$M^v < 10^{-2} \text{Max} \forall (j, k) \{ |u_{j,k}| \} . \quad (5.37)$$

How many iterations are needed to obtain convergence with Jacobi, Gauss-Seidel, and SOR? (Note that this convergence criterion refers to $u_{j,k}$, the exact solution of the finite-difference problem, and so could not be used in practice.)

d) Plot the error $u_{j,k}^v - u_{j,k}$, for all three methods, for $v = 0$ (all the same), $v = 10$, and $v = 20$.

e) Plot R^v as a function of v (or, if you prefer, as a function of $\ln v$) for all three methods.

CHAPTER 6**Diffusion**

Copyright 2004 David A. Randall

6.1 Introduction

Diffusion is a macroscopic interpretation 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, need be considered. The process of one-dimensional diffusion can be represented in simplified form by

$$\frac{\partial q}{\partial t} = -\frac{\partial F_q}{\partial x}. \quad (6.1)$$

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

$$F_q = -K \frac{\partial q}{\partial x}, \quad (6.2)$$

where K is a “diffusion coefficient,” which must be specified somehow. Physically meaningful applications of (6.2) are possible when

$$K \geq 0. \quad (6.3)$$

Substitution of (6.2) into (6.1) gives

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial x} \left(K \frac{\partial q}{\partial x} \right). \quad (6.4)$$

Because (6.4) involves second derivatives in space, it requires two boundary conditions. Here we simply assume periodicity of both q and $\frac{\partial q}{\partial x}$. It then follows immediately from (6.1) that the spatially averaged value of q does not change with time:

$$\frac{\partial}{\partial t} \int_{\text{spatial domain}} q \, dx = 0. \quad (6.5)$$

When (6.3) is satisfied, (6.4) 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

$$\frac{\partial}{\partial t} \int_{\text{spatial domain}} q^2 \, dx \leq 0. \quad (6.6)$$

To prove this, multiply both sides of (6.4) by q :

$$\begin{aligned} \frac{\partial}{\partial t} \left(\frac{q^2}{2} \right) &= q \frac{\partial}{\partial x} \left(K \frac{\partial q}{\partial x} \right) \\ &= \frac{\partial}{\partial x} \left(q K \frac{\partial q}{\partial x} \right) - K \left(\frac{\partial q}{\partial x} \right)^2. \end{aligned} \quad (6.7)$$

When we integrate the second line of (6.7) over a periodic domain, the first term vanishes and the second is negative (or possibly zero). The result (6.6) follows immediately.

With the assumed periodic boundary conditions, we can expand q in a Fourier series:

$$q(x, t) = \sum \hat{q}_k(t) e^{ikx}. \quad (6.8)$$

Substituting into (6.1), and assuming spatially constant K , we find that the amplitude of a particular Fourier mode satisfies

$$\frac{d\hat{q}_k}{dt} = -k^2 K \hat{q}_k, \quad (6.9)$$

which is the decay equation. This shows that there is a close connection between the diffusion equation and the decay equation. The solution of (6.9) is

$$\hat{q}_k(t) = \hat{q}_k(0) e^{-k^2 K t} \quad (6.10)$$

Note that higher wave numbers decay more rapidly, for a given value of K . Since

$$\hat{q}_k(t + \Delta t) = \hat{q}_k(0) e^{-k^2 K (t + \Delta t)} = \hat{q}_k(t) e^{-k^2 K \Delta t}. \quad (6.11)$$

This shows that, for the exact solution,

$$\lambda = e^{-k^2 K \Delta t} . \quad (6.12)$$

6.2 A simple explicit scheme

A finite-difference analog of (6.1) is

$$q_j^{n+1} - q_j^n = \kappa_{j+\frac{1}{2}}(q_{j+1}^n - q_j^n) - \kappa_{j-\frac{1}{2}}(q_j^n - q_{j-1}^n) , \quad (6.13)$$

where for convenience we define the nondimensional combination

$$\kappa_{j+\frac{1}{2}} \equiv \frac{K_{j+\frac{1}{2}} \Delta t}{(\Delta x)^2} \quad (6.14)$$

Here we have assumed for simplicity that Δx is a constant. The scheme given by (6.13) combines forward time differencing with centered space differencing. Recall that this combination is unconditionally unstable for the advection problem, but we will show that it is conditionally stable for diffusion. It should be obvious that, with periodic boundary conditions, (6.13) guarantees conservation of q in the sense that

$$\sum_j q_j^{n+1} \Delta x = \sum_j q_j^n \Delta x . \quad (6.15)$$

To analyze the stability of (6.13) using von Neumann's method, we assume that κ is a constant. Then (6.13) yields

$$(\lambda - 1) = \kappa[(e^{ik\Delta x} - 1) - (1 - e^{-ik\Delta x})] , \quad (6.16)$$

which is equivalent to

$$\lambda = 1 - 4\kappa \sin^2\left(\frac{k\Delta x}{2}\right) \leq 1 . \quad (6.17)$$

Note that λ is real.

Instability can occur if $\lambda < -1$, or

$$\kappa \sin^2\left(\frac{k\Delta x}{2}\right) > \frac{1}{2} . \quad (6.18)$$

The worst case is $\sin^2\left(\frac{k\Delta x}{2}\right) = 1$, which occurs for $\frac{k\Delta x}{2} = \frac{\pi}{2}$, or $k\Delta x = \pi$. This is the $2\Delta x$ wave. We conclude that with (6.13)

$$\kappa \leq \frac{1}{2} \text{ is required for stability.} \quad (6.19)$$

When the stability criterion derived above is satisfied, we can be sure that

$$\sum_i (q_j^{n+1})^2 - \sum_i (q_j^n)^2 < 0 ; \quad (6.20)$$

this is the condition for stability according to the energy method discussed in Chapter 2.

6.3 An implicit scheme

We can obtain unconditional stability through the use of an implicit scheme, but at the cost of some additional complexity. Replace (6.13) by

$$q_j^{n+1} - q_j^n = \left[\kappa_{j+\frac{1}{2}}(q_{j+1}^{n+1} - q_j^{n+1}) - \kappa_{j-\frac{1}{2}}(q_j^{n+1} - q_{j-1}^{n+1}) \right]. \quad (6.21)$$

We analyze the stability of (6.21), for the case of spatially variable but non-negative κ , using the energy method.

Multiplying (6.21) by q_j^{n+1} , we obtain:

$$\begin{aligned} (q_j^{n+1})^2 - q_j^{n+1} q_j^n = \\ \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \kappa_{j+\frac{1}{2}} (q_j^{n+1})^2 - \kappa_{j-\frac{1}{2}} (q_j^{n+1})^2 + \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} . \end{aligned} \quad (6.22)$$

Sum over the domain:

$$\begin{aligned}
& \sum_j (q_j^{n+1})^2 - \sum_j q_j^{n+1} q_j^n = \\
& \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} (q_j^{n+1})^2 - \sum_j \kappa_{j-\frac{1}{2}} (q_j^{n+1})^2 + \sum_j \kappa_{j-\frac{1}{2}} q_{j-1}^{n+1} q_j^{n+1} = \\
& \sum_j \kappa_{j+\frac{1}{2}} q_{j+1}^{n+1} q_j^{n+1} - \sum_j \kappa_{j+\frac{1}{2}} (q_j^{n+1})^2 - \sum_j \kappa_{j+\frac{1}{2}} (q_{j+1}^{n+1})^2 + \sum_j \kappa_{j+\frac{1}{2}} q_j^{n+1} q_{j+1}^{n+1} = \\
& - \sum_j \kappa_{j+\frac{1}{2}} (q_{j+1}^{n+1} - q_j^{n+1})^2 .
\end{aligned}$$

Rearranging, we find that

$$\sum_j q_j^{n+1} q_j^n = \sum_j \left[(q_j^{n+1})^2 + \kappa_{j+\frac{1}{2}} (q_{j+1}^{n+1} - q_j^{n+1})^2 \right]. \quad (6.23)$$

Next, note that

$$\sum_j (q_j^{n+1} - q_j^n)^2 = \sum_j [(q_j^{n+1})^2 + (q_j^n)^2 - 2q_j^{n+1} q_j^n] \geq 0. \quad (6.24)$$

Substitute (6.23) into (6.24), to obtain

$$\sum_j \left\{ (q_j^{n+1})^2 + (q_j^n)^2 - 2 \left[(q_j^{n+1})^2 + \kappa_{j+\frac{1}{2}} (q_{j+1}^{n+1} - q_j^{n+1})^2 \right] \right\} \geq 0. \quad (6.25)$$

This can be simplified and rearranged to

$$\sum_j [(q_j^{n+1})^2 - (q_j^n)^2] \geq -2 \sum_j \left\{ \left[\kappa_{j+\frac{1}{2}} (q_{j+1}^{n+1} - q_j^{n+1})^2 \right] \right\}. \quad (6.26)$$

This shows that $\sum_j [(q_j^{n+1})^2 - (q_j^n)^2]$ is less than a negative number. Therefore

$$\sum_j [(q_j^{n+1})^2 - (q_j^n)^2] < 0 \quad (6.27)$$

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.

Eq. (6.21) contains three unknowns, namely q_j^{n+1} , q_{j+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 too 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 often avoided in practice.

6.4 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

$$\frac{q_j^{n+1} - q_j^{n-1}}{2\Delta t} = \frac{1}{(\Delta x)^2} \left[\kappa_{j+\frac{1}{2}}(q_{j+1}^n - q_j^{n+1}) - \kappa_{j-\frac{1}{2}}(q_j^{n-1} - q_{j-1}^n) \right]. \quad (6.28)$$

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 i , so that the solution can be obtained without solving a system of simultaneous equations:

$$q_j^{n+1} = \frac{q_j^{n-1} + 2 \left[\kappa_{j+\frac{1}{2}} q_{j+1}^n - \kappa_{j-\frac{1}{2}} (q_j^{n-1} - q_{j-1}^n) \right]}{1 + 2\kappa_{j+\frac{1}{2}}}. \quad (6.29)$$

Consider spatially constant κ , and define

$$\alpha \equiv 2\kappa \quad (6.30)$$

The amplification factor satisfies

$$\lambda^2 - 1 = \alpha(\lambda e^{ik\Delta x} - \lambda^2 - 1 + \lambda e^{-ik\Delta x}) \quad (6.31)$$

which is equivalent to

$$\lambda^2(1 + \alpha) - \lambda 2\alpha \cos(k\Delta x) - (1 - \alpha) = 0. \quad (6.32)$$

The solutions are

$$\begin{aligned}
\lambda &= \frac{\alpha \cos(k\Delta x) \pm \sqrt{\alpha^2 \cos^2(k\Delta x) - (1 - \alpha^2)}}{1 + \alpha} \\
&= \frac{\alpha \cos(k\Delta x) \pm \sqrt{1 - \alpha^2 \sin^2(k\Delta x)}}{1 + \alpha} .
\end{aligned} \tag{6.33}$$

It should be clear that the plus sign corresponds to the physical mode, and the minus sign to the computational mode. Consider two cases. First, if $\alpha^2 \sin^2(k\Delta x) \leq 1$, then λ is real, and we find that

$$|\lambda| \leq \frac{1 + |\alpha \cos(k\Delta x)|}{1 + \alpha} \leq 1 . \tag{6.34}$$

Second, if $\alpha^2 \sin^2(k\Delta x) > 1$, which implies that $\alpha > 1$, then

$$|\lambda| = \frac{\sqrt{\alpha^2 \cos^2(k\Delta x) + \alpha^2 \sin^2(k\Delta x) - 1}}{1 + \alpha} = \frac{\sqrt{\alpha^2 - 1}}{1 + \alpha} = \sqrt{\frac{\alpha - 1}{\alpha + 1}} < 1 . \tag{6.35}$$

We conclude that the scheme is unconditionally stable.

It does not follow, however, that the scheme gives a good solution for large Δt . Consider the case of constant κ , and let $\alpha \rightarrow \infty$. Then (6.35) reduces to

$$|\lambda| \rightarrow 1 . \tag{6.36}$$

There is no damping, which is very wrong for the case of diffusion over a long time interval.

6.5 Summary

Diffusion is a relatively simple process which preferentially wipes out small-scale features. The most robust schemes for the diffusion equation are fully implicit schemes, but these give rise to systems of simultaneous equations. The DuFort-Frankel scheme is unconditionally stable and easy to implement, but behaves badly as the time step becomes large for fixed Δx .

Problems

1. Prove that the trapezoidal implicit scheme with centered second-order space differencing is unconditionally stable for the diffusion equation.
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

$$q_1=100, j = 1, 50, \text{ and } q_j = 110 \text{ for } j = 51, 100. \quad (6.37)$$

Compare the three solutions for different choices of the time step.

1. Use the energy method to evaluate the stability of (6.13).

CHAPTER 7**Making Waves**

Copyright 2004 David A. Randall

7.1 The shallow-water equations

In most of this chapter we will discuss the shallow-water equations, which can be written as

$$\frac{\partial \mathbf{v}}{\partial t} + (\zeta + f)\mathbf{k} \times \mathbf{v} = -\nabla[g(h + h_S) + K] \quad (7.1)$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (\mathbf{v}h) = 0. \quad (7.2)$$

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 (7.1), all frictional effects have been neglected, for simplicity.

A very useful idealized subset of the shallow-water system describes the special case of a one-dimensional, small-amplitude, external gravity wave for a shallow, non-rotating incompressible, homogeneous fluid (shallow water), with a resting basic state. Eqs. (7.1) and (7.2) become

$$\frac{\partial u}{\partial t} + g \frac{\partial h}{\partial x} = 0, \quad (7.3)$$

and

$$\frac{\partial h}{\partial t} + H \frac{\partial u}{\partial x} = 0, \quad (7.4)$$

respectively. Here H is the mean depth of the fluid. We refer to (7.3)-(7.4) as “the gravity wave equations.” Let

$$c^2 \equiv gH. \quad (7.5)$$

By combining (7.3)-(7.4) we can derive

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad (7.6)$$

and

$$\frac{\partial^2 h}{\partial t^2} = c^2 \frac{\partial^2 h}{\partial x^2}, \quad (7.7)$$

which are both examples of “the wave equation.”

Assuming solutions of the form $e^{i(kx - \omega t)}$, we obtain the dispersion equation

$$\omega^2 = gHk^2 \quad (7.8)$$

The exact phase speed of pure gravity waves (without the effects of rotation) is $\pm\sqrt{gH}$, regardless of wave length. There are two waves, one propagating in the positive x -direction, and the other in the negative x -direction.

7.2 The wave equation

The solutions of the wave equation, (7.6), are constant along space-time line (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 particular solution of the wave equation (7.6), but $g(x + ct)$ is a second particular solution. We can assume $c > 0$ without loss of generality. The general solution of (7.6) is given by

$$u(x, t) = f(x - ct) + g(x + ct), \quad (7.9)$$

where, as shown below, the forms of f and g are determined completely by the initial conditions

$$\begin{aligned} u_{t=0} &= F(x), \\ \left(\frac{\partial u}{\partial t}\right)_{t=0} &= G(x). \end{aligned} \quad (7.10)$$

Substituting (7.9) into (7.10), we find that

$$\begin{aligned} f(x) + g(x) &= F(x) , \\ -cf'(x) + cg'(x) &= G(x) . \end{aligned} \quad (7.11)$$

Here a prime denotes differentiation. Differentiating the first of (7.11) and then using the second of (7.11), we find that

$$\begin{aligned} f'(x) &= \frac{1}{2} \left[F'(x) - \frac{G(x)}{c} \right] , \\ g'(x) &= \frac{1}{2} \left[F'(x) + \frac{G(x)}{c} \right] . \end{aligned} \quad (7.12)$$

Integration of (7.12) gives

$$\begin{aligned} f(x) &= \frac{1}{2} \left[F(x) - \frac{1}{c} \int_0^x G(\xi) d\xi \right] + C_1 , \\ g(x) &= \frac{1}{2} \left[F(x) + \frac{1}{c} \int_0^x G(\xi) d\xi \right] + C_2 . \end{aligned} \quad (7.13)$$

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 f and g of (7.13), and then substituting into (7.9). This gives

$$u(x, t) = \frac{1}{2} \left[F(x - ct) + F(x + ct) + \frac{1}{c} \int_{x-ct}^{x+ct} G(\xi) d\xi \right] , \quad (7.14)$$

where we have required and used $C_1 + C_2 = 0$ in order to satisfy $u_{t=0} = F(x)$.

In order to further relate the wave equation to the advection equation that we have already studied, we reduce (7.6) to a pair of first-order equations by defining

$$p \equiv \frac{\partial u}{\partial t} \text{ and } q \equiv -c \frac{\partial u}{\partial x} . \quad (7.15)$$

Substitution of (7.15) into the wave equation (7.6) gives

$$\frac{\partial p}{\partial t} + c \frac{\partial q}{\partial x} = 0 , \quad (7.16)$$

and differentiation of the second of (7.15) with respect to t , with the use of the first of (7.15), gives

$$\frac{\partial q}{\partial t} + c \frac{\partial p}{\partial x} = 0. \quad (7.17)$$

If we alternately add (7.16) and (7.17) and subtract (7.17) from (7.16), we obtain

$$\frac{\partial P}{\partial t} + c \frac{\partial P}{\partial x} = 0, \text{ where } P \equiv p + q, \quad (7.18)$$

$$\frac{\partial Q}{\partial t} - c \frac{\partial Q}{\partial x} = 0, \text{ where } Q \equiv p - q. \quad (7.19)$$

Now we have a system of two first-order equations, each in the form of the advection equation. Note, however, that the “advectations” are in opposite directions! Assuming that $c > 0$, P is “advected” towards increasing x , while Q is “advected” towards decreasing x . From (7.18) and (7.19), it is clear that P is constant along the line $x - ct = \text{constant}$, and Q is constant along the line $x + ct = \text{constant}$. Equations (7.18) and (7.19) are called the *normal form* of (7.16) and (7.17).

These concepts are applicable, with minor adjustments, to any hyperbolic system of equations. The curves $x - ct = \text{constant}$ and $x + ct = \text{constant}$ are called “characteristics.” A hyperbolic equation is characterized, so to speak, by two such families of curves. In the present case they are straight, parallel lines, but in general they can have any shape so long as they do not intersect each other.

7.3 Staggered grids

Now we discuss the differential-difference equations

$$\frac{\partial u_j}{\partial t} + g \left(\frac{h_{j+1} - h_{j-1}}{2\Delta x} \right) = 0, \quad (7.20)$$

$$\frac{\partial h_j}{\partial t} + H \left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} \right) = 0, \quad (7.21)$$

which are, of course, differential-difference analogs of the one-dimensional shallow water equations, (7.3) and (7.4). Consider a distribution of the dependent variables on the grid as shown in Fig. 7.1. Notice that from (7.20) and (7.21) *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.

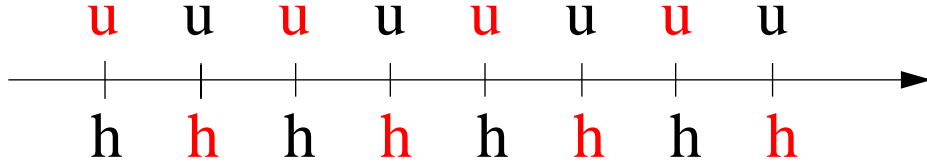


Figure 7.1: A grid for solution of the one-dimensional shallow water equations.

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.

Here is a mathematical way to draw the same conclusion. The wave solutions of (7.20) and (7.21) are

$$(u_j, h_j) \sim e^{i(kj\Delta x - \omega t)}, \quad (7.22)$$

giving

$$\begin{aligned} \omega u_j - gh_j \frac{\sin(k\Delta x)}{\Delta x} &= 0, \\ \omega h_j - Hu_j \frac{\sin(k\Delta x)}{\Delta x} &= 0. \end{aligned} \quad (7.23)$$

Provided that u_j and h_j are not both identically zero, we obtain the dispersion relation

$$\omega^2 = k^2 gH \left(\frac{\sin p}{p} \right)^2 \text{ where } p \equiv k\Delta x. \quad (7.24)$$

Suppose that ω is given. If $p = p_0$ satisfies (7.24), 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 ω , although physically there should only be two. The “extra” pair of modes come from the redundancy on the grid. *The extra modes are computational modes “in space.”* Earlier we encountered computational modes in time.

Without loss of generality, we suppose that $0 < p_0 < \frac{\pi}{2}$, so that $\sin(p_0) > 0$. Then the two solutions $p = p_0$ and $p = -p_0$ are approximations to the true solution, and therefore 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 significant 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 ω , the general solution for u_j is a linear combination of the four modes, and can be written as

$$u_j = [Ae^{ip_0j} + Be^{-ip_0j} + Ce^{i(\pi-p_0)j} + De^{-i(\pi-p_0)j}]e^{-i\omega t}. \quad (7.25)$$

Correspondingly, by substituting (7.25) into (7.21), we find that h_j satisfies

$$h_j = \frac{H \sin p_0}{\omega \Delta x} [Ae^{ip_0j} - Be^{-ip_0j} + Ce^{i(\pi-p_0)j} - De^{-i(\pi-p_0)j}]e^{-i\omega t} \quad (7.26)$$

If we assume $\omega > 0$, so that $\sin(p_0) = \frac{\omega \Delta x}{\sqrt{gH}}$ [see (7.24)], then (7.26) reduces to

$$h_j = \sqrt{\frac{H}{g}} [Ae^{ip_0j} - Be^{-ip_0j} + Ce^{i(\pi-p_0)j} - De^{-i(\pi-p_0)j}]e^{-i\omega t} \quad (7.27)$$

7.4 Numerical simulation of geostrophic adjustment. 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:

$$\frac{\partial u}{\partial t} - fv + g \frac{\partial h}{\partial x} = 0, \quad (7.28)$$

$$\frac{\partial v}{\partial t} + fu + g \frac{\partial h}{\partial y} = 0, \quad (7.29)$$

$$\frac{\partial h}{\partial t} + H\delta = 0. \quad (7.30)$$

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 (7.28)- (7.30), we can derive an equivalent set in terms of vorticity, $\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$, and divergence:

$$\frac{\partial \delta}{\partial t} - f\zeta + g \left(\frac{\partial^2}{\partial x^2} h + \frac{\partial^2}{\partial y^2} h \right) = 0, \quad (7.31)$$

$$\frac{\partial \zeta}{\partial t} + f\delta = 0, \quad (7.32)$$

$$\frac{\partial h}{\partial t} + H\delta = 0. \quad (7.33)$$

Of course, (7.33) is identical to (7.30). We can eliminate the vorticity and mass in (7.31) by using (7.32) and (7.33), respectively. Then by assuming wave solutions, we obtain the dispersion relation:

$$\left(\frac{\sigma}{f} \right)^2 = 1 + \lambda^2 (k^2 + l^2). \quad (7.34)$$

Here σ is the frequency, $\lambda \equiv \sqrt{gH}/f$ 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 (7.34) are important for the geostrophic adjustment process.

In their discussion of various numerical representations of (7.28)- (7.30), AL defined five grids denoted by “A” through “E,” as shown in Fig. 7.2. Fig. 7.2 also shows the Z grid, which will be discussed later. AL also gave the simplest centered finite-difference approximations to (7.28)- (7.30), for each of the five grids; these equations 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. 7.2. The table also gives a plot of the nondimensional frequency, (σ/f) , as a function of kd and ld , for the special case $\lambda/d = 2$. Here d is the grid size, assumed to be the same in the x and y directions. The significance of this particular choice of λ/d is discussed later. The plots 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. Approximation of the spatial derivatives in (1-3) inevitably involves averaging on the A grid, however. To illustrate this important point, consider the simplest centered approximation to $\partial h / \partial x$ at a u point, on the A grid. We must first obtain a value of h at $i + 1/2$ by averaging from i and $i + 1$, and a second value at $i - 1/2$ by averaging from i and $i - 1$. We can then subtract these two average values of h , and divide by Δx , to

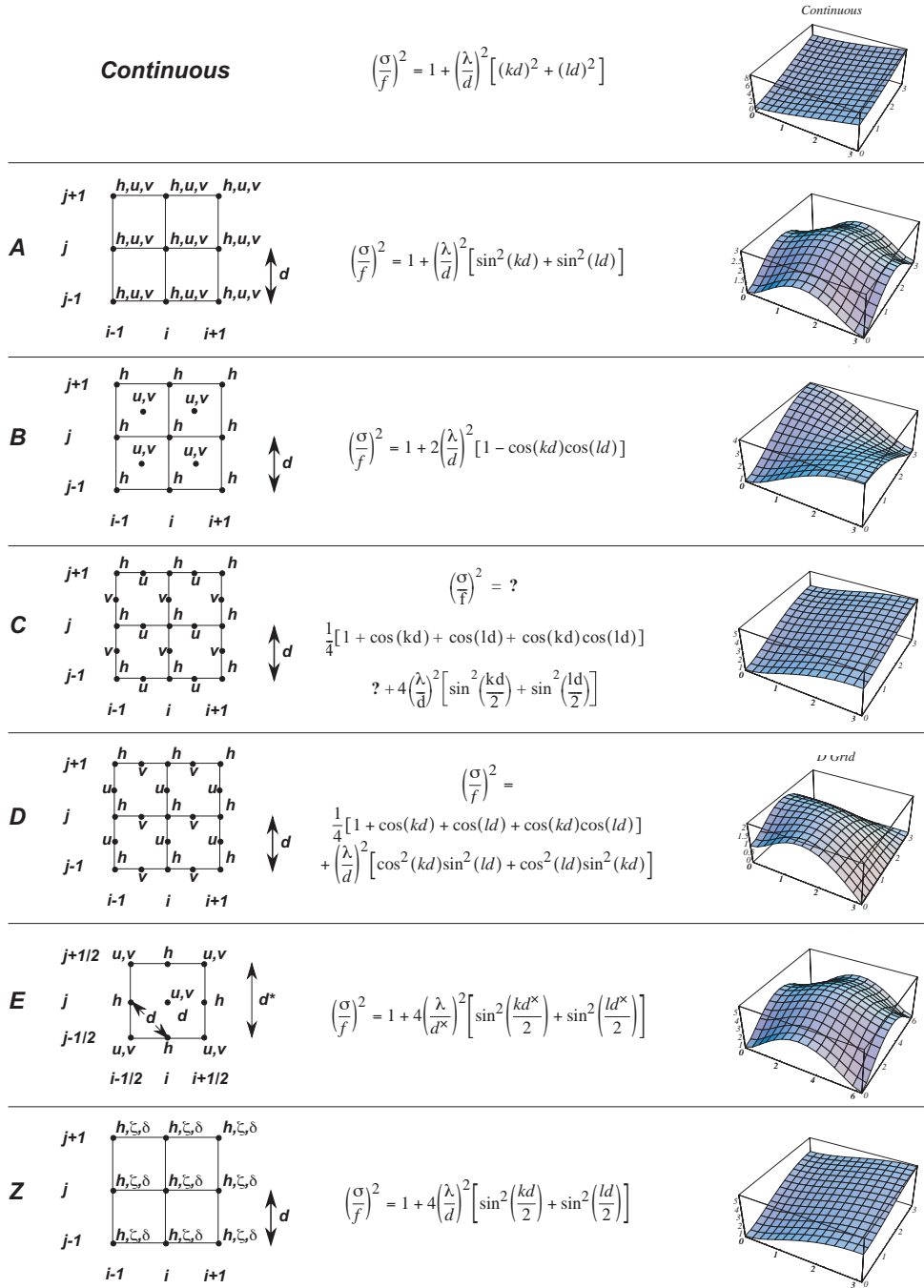


Figure 7.2: Grids, dispersion equations, and plots of dispersion equations for grids A–E and Z. The continuous dispersion equation and its plot are also shown for comparison. For plotting, it has been assumed that $\lambda/d = 2$.

obtain the desired approximation to $\partial h / \partial x$. Similarly, averaging is needed to define the mass convergence / divergence at h points.

The averaging described above inevitably “hides” noise at the smallest represented scales (e.g. a checkerboard pattern in h). Such dynamically “invisible” noise cannot participate in the dynamics of the model, e.g. by propagating and dispersing as in the process of geostrophic adjustment. A plot of the dispersion equation for the A grid, as shown in Fig. 7.2, indicates a maximum of the frequency (group speed equal to zero) for some combinations k and l . As a result, solutions on the A grid are extremely noisy in practice and must be smoothed, e.g. through filtering (e.g., Kalnay-Rivas et al., 1977). Because of this well known problem, the A grid is almost never used today.

The preceding discussion is an example which illustrates the general rule that it is desirable to avoid averaging in the design of a finite-difference scheme.

Next, consider the B grid. 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. 7.2. 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 λ , the radius of deformation. A plot of the dispersion equation, given in Fig. 7.2, 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.

Next, we turn to the D grid. Inspection of the stencil shown in Fig. 7.2 shows 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 is very badly behaved, giving zero phase speed for the shortest represented waves, and also giving a zero group speed for some modes.

Finally, the E grid is shown in Fig. 7.2. The grid spacing for the E grid is chosen to be $d^* \equiv \sqrt{2}d$, so that the “density” of h points is the same as in the other four grids. 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. Nevertheless there is a problem, which becomes apparent if we consider a solution that is uniform in one of the grid directions, say the y -direction. In that case, we effectively have a one-dimensional problem. In one dimension, the E grid “collapses” to the A grid, with a reduced grid spacing $d = d^* / \sqrt{2}$. For such one-dimensional motions, the E grid has all the problems of the A grid. These problems are apparent in the plot of the dispersion equation, given in Fig. 7.2. (For the E grid, the nondimensional frequency is plotted as a function of kd^* and ld^* , out to a value of 2π ; this corresponds to the shortest “one-dimensional” mode.) The group speed is zero for some combinations of k and l .

Now recall the conclusion of AL, described 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 so that λ/d can be less than one. Nevertheless, in practice modes for which $\lambda/d \ll 1$ 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 λ/d can be on the order of 0.1. Fig. 7.3 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. Janjic 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 λ/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.

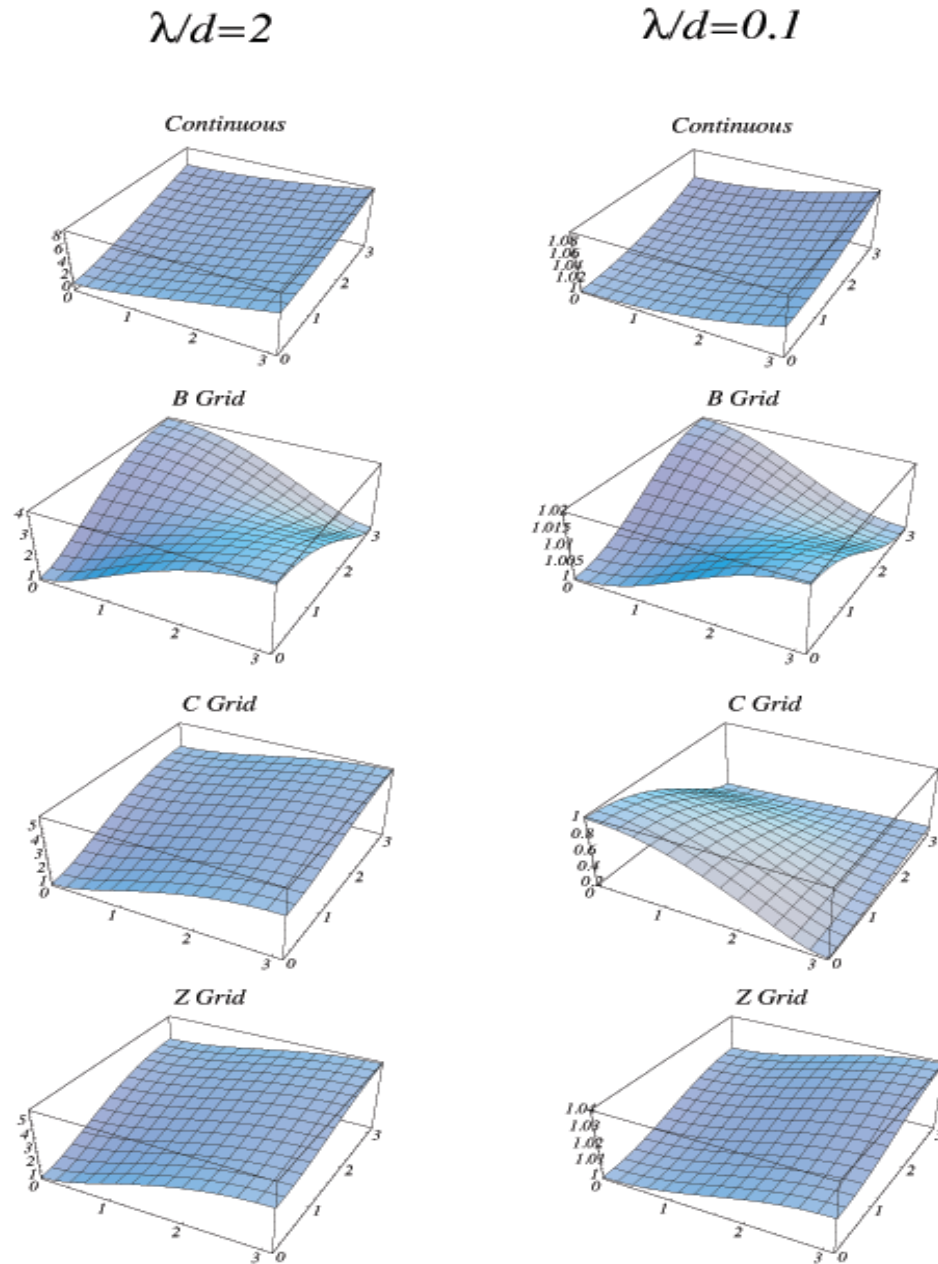


Figure 7.3: Dispersion relations for the continuous shallow water equations, and for finite-difference approximations based on the B, C, and Z grids. The horizontal coordinates in the plots are kd and ld , respectively, except for the E grid, for which kd^* and ld^* are used. The vertical coordinate is the normalized frequency, σ/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 $\lambda/d = 0.1$.

Now consider an unstaggered grid for the integration of (7.31) - (7.33), which was called the Z grid by Randall (1994). This grid is also illustrated in Fig. 7.2. 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 (7.31) - (7.33) 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 (7.31) - (7.33).

Fig. 7.3 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 λ/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. As discussed later, such solution procedures can be computationally expensive in finite-difference models, but are not expensive in spectral models. It may be appropriate to re-examine this point in the light of modern algorithms for solving linear systems, e.g., multi-grid methods.

7.5 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., $\frac{\partial u}{\partial x}$), but no vorticity. Linearizing about a state of rest, the continuous equations are (7.3) and (7.4).

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. The most obvious possible scheme is centered in both space and time:

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^{n-1}}{2\Delta t} + g \left(\frac{h_{j+1}^n - h_j^n}{\Delta x} \right) = 0, \quad (7.35)$$

$$\frac{h_j^{n+1} - h_j^{n-1}}{2\Delta t} + H \left(\frac{u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n}{\Delta x} \right) = 0. \quad (7.36)$$

Compare with (7.20)-(7.21). With assumed solutions of the form $u_j^n = \hat{u}^n \exp(ikj\Delta x)$, $h_j^n = \hat{h}^n \exp(ikj\Delta x)$, and the usual definition of the amplification factor, we find that

$$(\lambda^2 - 1)\hat{u}^n + \lambda \frac{g\Delta t}{\Delta x} 4i \sin\left(\frac{k\Delta x}{2}\right) \hat{h}^n = 0, \quad (7.37)$$

$$\lambda \frac{H\Delta t}{\Delta x} 4i \sin\left(\frac{k\Delta x}{2}\right) \hat{u}^n + (\lambda^2 - 1)\hat{h}^n = 0. \quad (7.38)$$

Non-trivial solutions occur for

$$(\lambda^2 - 1)^2 + \lambda^2 \left(\frac{4c_{\text{GW}}\Delta t}{\Delta x} \right)^2 \sin^2\left(\frac{k\Delta x}{2}\right) = 0, \quad (7.39)$$

where $c_{\text{GW}} \equiv \sqrt{gH}$. 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 (7.39) as a quadratic equation for λ^2 . As a first step, rewrite it as

$$(\lambda^2)^2 + \lambda^2(-2 + b) + 1 = 0, \quad (7.40)$$

where, for convenience, we define

$$b \equiv \left(\frac{4c_{\text{GW}}\Delta t}{\Delta x} \right)^2 \sin^2\left(\frac{k\Delta x}{2}\right) \geq 0. \quad (7.41)$$

Obviously, for $\Delta t \rightarrow 0$ with fixed Δx we get $b \rightarrow 0$. The solution of (7.40) is

$$\begin{aligned}\lambda^2 &= \frac{b-2 \pm \sqrt{(b-2)^2 - 4}}{2} \\ &= \frac{b-2 \pm \sqrt{b(b-4)}}{2}.\end{aligned}\quad (7.42)$$

Inspection of (7.42) shows that for $b \rightarrow 0$, we get $|\lambda| \rightarrow 1$, as expected. For $\lambda = |\lambda|e^{i\theta}$, we see that

$$|\lambda|^2 [\cos(2\theta) + i\sin(2\theta)] = \frac{b-2 \pm \sqrt{b(b-4)}}{2}. \quad (7.43)$$

It follows that

$$|\lambda|^2 \cos(2\theta) = \frac{b-2}{2}, \quad |\lambda|^2 \sin(2\theta) = \sqrt{b(4-b)}, \quad \text{for } b \leq 4, \quad (7.44)$$

from which we obtain

$$\tan(2\theta) = \frac{-2\sqrt{b(4-b)}}{2-b} \quad \text{for } b \leq 4. \quad (7.45)$$

A plot of (7.45) is given in Fig. 7.4. Eq. (7.43) also implies that

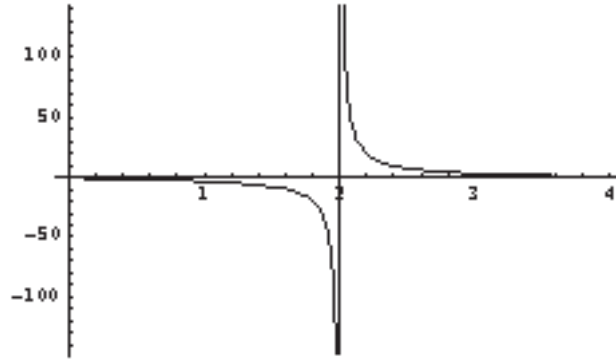


Figure 7.4: A plot of $\tan(2\theta)$ as a function of b , for $b \leq 4$.

$$|\lambda|^4 = \left(\frac{b-2}{2}\right)^2 + \frac{b(4-b)}{4} = 1 \quad \text{for } b \leq 4. \quad (7.46)$$

The scheme is thus neutral for $b \leq 4$, as could be anticipated from our earlier discussion of the oscillation equation.

Returning to (7.43), we find that

$$\sin(2\theta) = 0, \cos(2\theta) = \pm 1, \text{ and } \pm|\lambda|^2 = \frac{b-2 \pm \sqrt{b(b-4)}}{2} \text{ for } b > 4. \quad (7.47)$$

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 $\left(\frac{c_{\text{GW}}\Delta t}{\Delta x}\right)|\sin(k\Delta x)| < \frac{1}{2}$. The worst case occurs for $\left|\sin\left(\frac{k\Delta x}{2}\right)\right| = 1$, which corresponds to $k\Delta x = \pi$, i.e., the $2\Delta x$ -wave. It follows that

$$\frac{c_{\text{GW}}\Delta t}{\Delta x} < \frac{1}{2} \text{ is required for stability,} \quad (7.48)$$

and that the first wave to become unstable will be the $2\Delta x$ -wave.

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 also the speed of sound. The stability criterion for the leapfrog scheme as applied to the wave problem, i.e., (7.48), 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 (7.35)-(7.36) by

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n}{\Delta t} + g \left(\frac{h_{j+1}^{n+1} - h_{j-1}^{n+1}}{\Delta x} \right) = 0, \quad (7.49)$$

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + H \left(\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j-\frac{1}{2}}^{n+1}}{\Delta x} \right) = 0. \quad (7.50)$$

This leads to

$$(\lambda - 1)\hat{u}^n + \lambda \frac{g\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2}\right) \hat{h}^n = 0, \quad (7.51)$$

$$\lambda \frac{H\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2}\right) \hat{u}^n + (\lambda - 1) \hat{h}^n = 0. \quad (7.52)$$

The condition for non-trivial solutions is

$$(\lambda - 1)^2 + \lambda^2 4 \left(\frac{c_{\text{GW}} \Delta t}{\Delta x} \right)^2 \sin^2\left(\frac{k\Delta x}{2}\right) = 0, \quad (7.53)$$

which, using (7.41), is equivalent to

$$\lambda^2 \left(1 + \frac{b}{4} \right) - 2\lambda + 1 = 0. \quad (7.54)$$

This time there are no computational modes; the two physical modes satisfy

$$\begin{aligned} \lambda &= \frac{2 \pm \sqrt{4 - 4 \left(1 + \frac{b}{4} \right)}}{2 \left(1 + \frac{b}{4} \right)} \\ &= \frac{1 \pm i \sqrt{\frac{b}{4}}}{1 + \frac{b}{4}}. \end{aligned} \quad (7.55)$$

The solutions are always oscillatory, and

$$|\lambda|^2 = \frac{1 + \frac{b}{4}}{\left(1 + \frac{b}{4} \right)^2} = \frac{4}{4 + b} \leq 1, \quad (7.56)$$

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 7.49-50 by

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n}{\Delta t} + \frac{g}{\Delta x} \left[\left(\frac{h_{j+1}^n + h_{j+1}^{n+1}}{2} \right) - \left(\frac{h_j^n + h_j^{n+1}}{2} \right) \right] = 0, \quad (7.57)$$

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \frac{H}{\Delta x} \left[\left(\frac{u_{j+\frac{1}{2}}^n + u_{j+\frac{1}{2}}^{n+1}}{2} \right) - \left(\frac{u_{j-\frac{1}{2}}^n + u_{j-\frac{1}{2}}^{n+1}}{2} \right) \right] = 0. \quad (7.58)$$

This leads to

$$(\lambda - 1)\hat{u}^n + \left(\frac{1 + \lambda}{2} \right) \frac{g\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2} \right) \hat{h}^n = 0, \quad (7.59)$$

$$\left(\frac{1 + \lambda}{2} \right) \frac{H\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2} \right) \hat{u}^n + (\lambda - 1)\hat{h}^n = 0. \quad (7.60)$$

For non-trivial solutions, we need

$$(\lambda - 1)^2 + (1 + \lambda)^2 \left(\frac{c_{\text{GW}}\Delta t}{\Delta x} \right)^2 \sin^2\left(\frac{k\Delta x}{2} \right) = 0. \quad (7.61)$$

Using (7.40), we can show that this is equivalent to

$$\lambda^2 - 2\lambda \left(\frac{16 - b}{16 + b} \right) + 1 = 0 \quad (7.62)$$

The solutions are

$$\lambda = \left(\frac{16 - b}{16 + b} \right) \pm i \sqrt{1 - \left(\frac{16 - b}{16 + b} \right)^2}. \quad (7.63)$$

It follows that $|\lambda|^2 = 1$ for all modes, i.e., the 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.

Inspection of (7.3) and (7.4) suggests another approach, which is related to a scheme that is sometimes called “pressure averaging” (Haltiner and Williams, 1980). We can integrate the mass equation using an explicit method, then use the updated mass field in to compute the pressure-gradient force of the momentum equation. This approach is “partly implicit,” but does not give rise to a matrix problem. Here we present a version due to Higdon (2002), simplified for application to the 1D non-rotating shallow-water system that we have been discussing. The scheme is divided into “steps:”

Step 1: Predict the velocity using a forward time step:

$$\frac{u_{j+\frac{1}{2}}^{\text{pred}} - u_{j+\frac{1}{2}}^n}{\Delta t} + g \left(\frac{h_{j+1}^n - h_j^n}{\Delta x} \right) = 0. \quad (7.64)$$

Step 2: Update the mass field using an “imitation” trapezoidal step:

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \frac{H}{\Delta x} \left[\left(\frac{u_{j+\frac{1}{2}}^n + u_{j+\frac{1}{2}}^{\text{pred}}}{2} \right) - \left(\frac{u_{j-\frac{1}{2}}^n + u_{j-\frac{1}{2}}^{\text{pred}}}{2} \right) \right] = 0. \quad (7.65)$$

Step 3: Update the wind field again using a “true” trapezoidal step:

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n}{\Delta t} + \frac{g}{\Delta x} \left[\left(\frac{h_{j+1}^n + h_{j+1}^{n+1}}{2} \right) - \left(\frac{h_j^n + h_j^{n+1}}{2} \right) \right] = 0. \quad (7.66)$$

Because we update the wind field twice, this approach requires 1.5 times as much work (for the two equations considered here) as the schemes discussed above. As we shall see, however, the time step can be increased, relative to the other schemes, and it turns out that there is a net gain in computational speed.

We begin our analysis of Higdon’s scheme by using (7.64) to write

$$u_{j+\frac{1}{2}}^{\text{pred}} = u_{j+\frac{1}{2}}^n - \frac{g\Delta t}{\Delta x} (h_{j+1}^n - h_j^n), \quad u_{j-\frac{1}{2}}^{\text{pred}} = u_{j-\frac{1}{2}}^n - \frac{g\Delta t}{\Delta x} (h_j^n - h_{j-1}^n). \quad (7.67)$$

Substituting these equations into (7.65), we obtain

$$(h_j^{n+1} - h_j^n) + \frac{H\Delta t}{2\Delta x} \left[2 \left(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n \right) - \frac{g\Delta t}{\Delta x} (h_{j+1}^n - 2h_j^n + h_{j-1}^n) \right] = 0. \quad (7.68)$$

Eqs. (7.66) and (7.68) lead to

$$(\lambda - 1)\hat{u}^n + \left(\frac{\lambda + 1}{2} \right) \frac{g\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2} \right) \hat{h}^n = 0, \quad (7.69)$$

$$\frac{H\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2} \right) \hat{u}^n + \left\{ (\lambda - 1) - \frac{gH}{2} \left(\frac{\Delta t}{\Delta x} \right)^2 2[\cos(k\Delta x) - 1] \right\} \hat{h}^n = 0. \quad (7.70)$$

Using the trigonometric identity $\cos(k\Delta x) - 1 = -2\sin^2\left(\frac{k\Delta x}{2}\right)$, we can rewrite (7.70) as

$$\frac{H\Delta t}{\Delta x} 2i \sin\left(\frac{k\Delta x}{2}\right) \hat{u}^n + \left[(\lambda - 1) + 2gH\left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2\left(\frac{k\Delta x}{2}\right)\right] \hat{h}^n = 0. \quad (7.71)$$

Nontrivial solutions of (7.69) and (7.71) satisfy

$$(\lambda - 1) \left[(\lambda - 1) + 2gH\left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2\left(\frac{k\Delta x}{2}\right) \right] + 2(\lambda + 1)gH\left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2\left(\frac{k\Delta x}{2}\right) = 0, \quad (7.72)$$

which, using (7.41), can be simplified to

$$\lambda^2 + \left(\frac{b}{4} - 2\right)\lambda + 1 = 0.$$

The solutions are

$$\lambda = \frac{\left(2 - \frac{b}{4}\right) \pm \frac{1}{4}i\sqrt{b(16 - b)}}{2}. \quad (7.73)$$

For $b < 16$, find that

$$|\lambda|^2 = \frac{\left(2 - \frac{b}{4}\right)^2 + \frac{b(16 - b)}{16}}{4} = 1. \quad (7.74)$$

The scheme is thus neutral for $b < 16$. It follows that

$$\frac{c_{\text{GW}}\Delta t}{\Delta x} < 1 \text{ is required for stability.} \quad (7.75)$$

The allowed time step is thus double that of the leapfrog scheme.

Going to two dimensions and adding rotation does not change much. The Coriolis terms can easily be made implicit if desired, since they do not involve spatial derivatives.

7.6 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.

Problems

1. Derive the dispersion equation for the C-grid, as given in Fig. 7.2.
2. Analyze the stability of the following scheme, which is an alternative to the Higdon (2002) scheme:

Step 1: Predict the mass field using a forward time step:

$$\frac{h_j^{\text{pred}} - h_j^n}{\Delta t} + \frac{H}{\Delta x} \left(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n \right) = 0. \quad (7.76)$$

Step 2: Update the wind field using an “imitation” trapezoidal step:

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n}{\Delta t} + \frac{g}{\Delta x} \left[\left(\frac{h_{j+1}^{\text{pred}} + h_{j+1}^n}{2} \right) - \left(\frac{h_j^{\text{pred}} + h_j^n}{2} \right) \right] = 0. \quad (7.77)$$

Step 3: Update the mass field again using a “true” trapezoidal step:

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \frac{H}{\Delta x} \left[\left(\frac{u_{j+\frac{1}{2}}^n + u_{j+\frac{1}{2}}^{n+1}}{2} \right) - \left(\frac{u_{j-\frac{1}{2}}^n + u_{j-\frac{1}{2}}^{n+1}}{2} \right) \right] = 0. \quad (7.78)$$

CHAPTER 8***Schemes for the one-dimensional nonlinear shallow-water equations***

Copyright 2004 David A. Randall

8.1 Properties of the continuous equations

Consider the one-dimensional shallow-water equations, with bottom topography, without rotation and with $v = 0$. The prognostic variables are the water depth or mass, h , and the speed, u . The exact equations are

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \quad (8.1)$$

and

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}[K + g(h + h_S)] = 0. \quad (8.2)$$

Here

$$K \equiv \frac{1}{2}u^2 \quad (8.3)$$

is the kinetic energy per unit mass, g is the acceleration of gravity, and h_S is the height of the bottom topography. Note that in (8.2) the vorticity has been assumed to vanish, which is reasonable in the absence of rotation and in one dimension.

The design of the scheme is determined by a sequence of choices. It is always good to have 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. (8.2) is one possible choice for the continuous form of the momentum equation. An alternative choice is

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(huu) + gh \frac{\partial}{\partial x}(h + h_S) = 0, \quad (8.4)$$

i.e., the flux form of the momentum equation, which can be derived by combining (8.1) and (8.2).

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 (8.1) with respect to x , over a closed or periodic domain, we obtain

$$\frac{d}{dt} \left(\int_{\text{domain}} h dx \right) = 0, \quad (8.5)$$

which means that mass is conserved.

Using

$$h \frac{\partial h}{\partial x} = \frac{\partial}{\partial x} \frac{h^2}{2}, \quad (8.6)$$

we can rewrite (8.4) as

$$\frac{\partial}{\partial t} (hu) + \frac{\partial}{\partial x} \left(hu^2 + g \frac{h^2}{2} \right) = -gh \frac{\partial h_S}{\partial x}. \quad (8.7)$$

If we integrate with respect to x , over a periodic domain, we obtain

$$\frac{d}{dt} \left(\int_{\text{domain}} hu dx \right) = - \int_{\text{domain}} gh \frac{\partial h_S}{\partial x} dx. \quad (8.8)$$

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.

The flux form of the kinetic energy equation can be derived by multiplying (8.1) by K and (8.2) by hu , and adding the results, to obtain

$$\frac{\partial}{\partial t} (hK) + \frac{\partial}{\partial x} (huK) + hu \frac{\partial}{\partial x} [g(h + h_S)] = 0. \quad (8.9)$$

The last term of (8.9) represents conversion between potential and kinetic energy.

The potential energy equation can be derived by multiplying (8.1) by $g(h + h_S)$ to obtain

$$\frac{\partial}{\partial t} \left[hg \left(h_S + \frac{1}{2} h \right) \right] + g(h + h_S) \frac{\partial}{\partial x} (hu) = 0, \quad (8.10)$$

or

$$\frac{\partial}{\partial t} \left[h g \left(h_S + \frac{1}{2} h \right) \right] + \frac{\partial}{\partial x} [h u g(h + h_S)] - h u \frac{\partial}{\partial x} [g(h + h_S)] = 0 \quad (8.11)$$

The last term of (8.11) represents conversion between kinetic and potential energy; compare with (8.9). In deriving (8.10), we have assumed that h_S is independent of time.

When we add (8.9) and (8.10), the energy conversion terms cancel, and we obtain a statement of the conservation of total energy, i.e.

$$\frac{\partial}{\partial t} \left\{ h \left[K + g \left(h_S + \frac{1}{2} h \right) \right] \right\} + \frac{\partial}{\partial x} \{ h u [K + g(h + h_S)] \} = 0. \quad (8.12)$$

The integral of (8.10) over a closed or periodic domain gives

$$\frac{d}{dt} \int_{\text{domain}} h \left[K + g \left(h_S + \frac{1}{2} h \right) \right] dx = 0, \quad (8.13)$$

which shows that the domain-integrated total energy is invariant.

8.1 Space differencing

Now consider finite-difference approximations to (8.1) and (8.2). 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). The grid spacing, Δx , is assumed to be uniform. The grid is shown in Fig. 8.1. It can be viewed as a one-dimensional version of the C grid. Our



Figure 8.1: The staggered grid used in the one-dimensional case.

selection of this particular grid is a second *choice* made in the design of the space-differencing scheme.

The finite difference version of the mass conservation equation is

$$\frac{d}{dt} h_i + \frac{1}{\Delta x} \left[(hu)_{i+\frac{1}{2}} - (hu)_{i-\frac{1}{2}} \right] = 0. \quad (8.14)$$

It should be understood that

$$h_{i+\frac{1}{2}} u_{i+\frac{1}{2}} \equiv (hu)_{i+\frac{1}{2}}. \quad (8.15)$$

The “wind-point” masses, e.g., $h_{i+\frac{1}{2}}$, are undefined at this stage. The finite-difference approximation used in (8.14) 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” of (8.14) makes it possible for the model to conserve mass, regardless of how the mass fluxes are defined. i.e.,

$$\frac{d}{dt} \left(\sum_{\text{domain}} h_i \right) = 0. \quad (8.16)$$

This is analogous to (8.5).

The finite-difference momentum equation that mimics (8.2) is

$$\frac{d}{dt} u_{i+\frac{1}{2}} + \left(\frac{K_{i+1} - K_i}{\Delta x} \right) + g \left[\frac{(h + h_S)_{i+1} - (h + h_S)_i}{\Delta x} \right] = 0. \quad (8.17)$$

The kinetic energy per unit mass, K_i , is also undefined at this stage, but resides at mass points. The finite-difference approximations used in (8.17) 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.

Multiply (8.17) by $h_{i+\frac{1}{2}}$ to obtain

$$h_{i+\frac{1}{2}} \frac{d}{dt} u_{i+\frac{1}{2}} + h_{i+\frac{1}{2}} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) + g h_{i+\frac{1}{2}} \left[\frac{(h + h_S)_{i+1} - (h + h_S)_i}{\Delta x} \right] = 0. \quad (8.18)$$

In order to mimic the differential relationship (8.6) we must require that

$$h_{i+\frac{1}{2}} \left(\frac{h_{i+1} - h_i}{\Delta x} \right) = \left(\frac{h_{i+1}^2 - h_i^2}{2\Delta x} \right), \quad (8.19)$$

which leads to

$$h_{i+\frac{1}{2}} = \frac{h_{i+1} + h_i}{2}. \quad (8.20)$$

Use of (8.20) will make it possible for the finite-difference model to mimic (8.8). In view of (8.20), we can write

$$(hu)_{i+\frac{1}{2}} = \left(\frac{h_{i+1} + h_i}{2} \right) u_{i+\frac{1}{2}}. \quad (8.21)$$

Combining (8.20) with the continuity equation (8.14), we see that we can write a *continuity equation for the wind points*, as follows:

$$\frac{dh}{dt}_{i+\frac{1}{2}} + \frac{1}{2\Delta x} \left[(hu)_{i+\frac{3}{2}} - (hu)_{i-\frac{1}{2}} \right] = 0. \quad (8.22)$$

It should be clear from the form of (8.22) that the “wind-point mass” is actually conserved by the model. Of course, we do not actually use (8.22) when we integrate the model; instead we use (8.14). Nevertheless, (8.22) will be satisfied, because it can be derived from (8.14) and (8.20). An alternative form of (8.22) is

$$\frac{dh}{dt}_{i+\frac{1}{2}} + \frac{1}{\Delta x} [(hu)_{i+1} - (hu)_i] = 0, \quad (8.23)$$

where

$$(hu)_{i+1} \equiv \frac{1}{2} \left[(hu)_{i+\frac{3}{2}} + (hu)_{i+\frac{1}{2}} \right] \text{ and } (hu)_i \equiv \frac{1}{2} \left[(hu)_{i+\frac{1}{2}} + (hu)_{i-\frac{1}{2}} \right]. \quad (8.24)$$

Now add (8.18) and $u_{i+\frac{1}{2}} \cdot (8.23)$, and use (8.19), to obtain what “should be” the flux form of the momentum equation, analogous to (8.7):

$$\begin{aligned} \frac{d}{dt} \left(h_{i+\frac{1}{2}} u_{i+\frac{1}{2}} \right) + h_{i+\frac{1}{2}} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) + u_{i+\frac{1}{2}} \frac{1}{\Delta x} [(hu)_{i+1} - (hu)_i] \\ + g \left(\frac{h_{i+1}^2 - h_i^2}{2\Delta x} \right) = -gh_{i+\frac{1}{2}} \left[\frac{(h_S)_{i+1} - (h_S)_i}{\Delta x} \right]. \end{aligned} \quad (8.25)$$

Suppose that the kinetic energy is defined by

$$K_i \equiv \frac{1}{2} u_{i+\frac{1}{2}} u_{i-\frac{1}{2}}. \quad (8.26)$$

Then we can write

$$\begin{aligned}
& h_{i+\frac{1}{2}} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) + u_{i+\frac{1}{2}} \frac{1}{\Delta x} [(hu)_{i+1} - (hu)_i] \\
&= \frac{1}{2\Delta x} \left\{ h_{i+\frac{1}{2}} \left(u_{i+\frac{3}{2}} u_{i+\frac{1}{2}} - u_{i+\frac{1}{2}} u_{i-\frac{1}{2}} \right) + u_{i+\frac{1}{2}} \left[(hu)_{i+\frac{3}{2}} - (hu)_{i-\frac{1}{2}} \right] \right\} \\
&= \frac{1}{\Delta x} \left[\left(\frac{h_{i+\frac{1}{2}} + h_{i+\frac{3}{2}}}{2} \right) u_{i+\frac{3}{2}} u_{i+\frac{1}{2}} - \left(\frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2} \right) u_{i-\frac{1}{2}} u_{i+\frac{1}{2}} \right].
\end{aligned} \tag{8.27}$$

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 (8.27) is a flux

form, momentum will be conserved by the scheme if we define the kinetic energy by (8.26). Note, however, that (8.26) can give a negative value of K_i when u is dominated by the $2\Delta x$ -mode. Also, when u is dominated by the $2\Delta x$ -mode, the momentum flux is always negative, i.e., in the $-x$ direction, assuming that the interpolated masses that appear in the momentum fluxes are positive.

Next, we derive the kinetic energy equation. Recall that the kinetic energy is defined at mass points. To begin the derivation, multiply (8.17) by $(hu)_{i+\frac{1}{2}}$ to obtain

$$\begin{aligned}
& (hu)_{i+\frac{1}{2}} \frac{d}{dt} u_{i+\frac{1}{2}} + (hu)_{i+\frac{1}{2}} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) \\
&+ g(hu)_{i+\frac{1}{2}} \left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x} \right] = 0.
\end{aligned} \tag{8.28}$$

Rewrite (8.28) for grid point $i - \frac{1}{2}$, simply by subtracting one from each subscript:

$$(hu)_{i-\frac{1}{2}} \frac{d}{dt} u_{i-\frac{1}{2}} + (hu)_{i-\frac{1}{2}} \left(\frac{K_i - K_{i-1}}{\Delta x} \right) + g(hu)_{i-\frac{1}{2}} \left[\frac{(h+h_S)_i - (h+h_S)_{i-1}}{\Delta x} \right] = 0. \tag{8.29}$$

Now add (8.28) and (8.29), and multiply the result by $\frac{1}{2}$:

$$\begin{aligned}
 & \frac{1}{2} \left[(hu)_{i+\frac{1}{2}} \frac{d}{dt} u_{i+\frac{1}{2}} + (hu)_{i-\frac{1}{2}} \frac{d}{dt} u_{i-\frac{1}{2}} \right] \\
 & + \frac{1}{2} \left[(hu)_{i+\frac{1}{2}} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) + (hu)_{i-\frac{1}{2}} \left(\frac{K_i - K_{i-1}}{\Delta x} \right) \right] \\
 & + \frac{g}{2} \left\{ (hu)_{i+\frac{1}{2}} \left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x} \right] + (hu)_{i-\frac{1}{2}} \left[\frac{(h+h_S)_i - (h+h_S)_{i-1}}{\Delta x} \right] \right\} = 0 .
 \end{aligned} \tag{8.30}$$

This is an advective form of the kinetic energy equation.

Now we try to derive, from (8.30) and (8.14), a flux form of the kinetic energy equation. Begin by multiplying (8.14) by K_i :

$$K_i \left\{ \frac{dh_i}{dt} + \left[\frac{(hu)_{i+\frac{1}{2}} - (hu)_{i-\frac{1}{2}}}{\Delta x} \right] \right\} = 0 . \tag{8.31}$$

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 (8.31) and (8.30) to obtain

$$\begin{aligned}
 & K_i \frac{dh_i}{dt} + \frac{1}{2} \left[(hu)_{i+\frac{1}{2}} \frac{d}{dt} u_{i+\frac{1}{2}} + (hu)_{i-\frac{1}{2}} \frac{d}{dt} u_{i-\frac{1}{2}} \right] \\
 & + \left\{ (hu)_{i+\frac{1}{2}} \left[\frac{K_i}{\Delta x} + \frac{1}{2} \left(\frac{K_{i+1} - K_i}{\Delta x} \right) \right] - (hu)_{i-\frac{1}{2}} \left[\frac{K_i}{\Delta x} - \frac{1}{2} \left(\frac{K_i - K_{i-1}}{\Delta x} \right) \right] \right\} \\
 & + g \left\{ (hu)_{i+\frac{1}{2}} \frac{[(h+h_S)_{i+1} - (h+h_S)_i]}{2\Delta x} + (hu)_{i-\frac{1}{2}} \frac{[(h+h_S)_i - (h+h_S)_{i-1}]}{2\Delta x} \right\} = 0 .
 \end{aligned} \tag{8.32}$$

Eq. (8.32) “should” be a flux form of the kinetic energy equation.

The advection terms on the second line of (8.32) are very easy to deal with. They can be rearranged to

$$\frac{1}{\Delta x} \left[(hu)_{i+\frac{1}{2}} \frac{1}{2} (K_{i+1} + K_i) - (hu)_{i-\frac{1}{2}} \frac{1}{2} (K_i + K_{i-1}) \right]. \quad (8.33)$$

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 (8.32), i.e.

$$g \left\{ (hu)_{i+\frac{1}{2}} \left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{2\Delta x} \right] + (hu)_{i-\frac{1}{2}} \left[\frac{(h+h_S)_i - (h+h_S)_{i-1}}{2\Delta x} \right] \right\}. \quad (8.34)$$

We digress here to discuss the potential energy equation. The finite-difference form of the potential energy equation can be derived by multiplying (8.14) by $g(h+h_S)_i$:

$$\frac{d}{dt} \left[h_i g \left(h_S + \frac{1}{2} h \right)_i \right] + g(h+h_S)_i \left[\frac{(hu)_{i+\frac{1}{2}} - (hu)_{i-\frac{1}{2}}}{\Delta x} \right] = 0. \quad (8.35)$$

This is analogous to (8.10). We want to recast (8.35) so that we see advection of potential energy, as well as the energy conversion term corresponding to (8.34); compare with (8.11). We write

$$\begin{aligned} & \frac{d}{dt} \left[h_i g \left(h_S + \frac{1}{2} h \right)_i \right] + ADV_i \\ & - \frac{g}{2} \left\{ (hu)_{i+\frac{1}{2}} \left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x} \right] + (hu)_{i-\frac{1}{2}} \left[\frac{(h+h_S)_i - (h+h_S)_{i-1}}{\Delta x} \right] \right\} = 0, \end{aligned} \quad (8.36)$$

where “ ADV_i ” represents the advection of potential energy, in flux form. The second line of (8.36) is a copy of the energy conversion terms of (8.32), but with the sign reversed. We require that (8.36) be equivalent (8.35), and ask what form of ADV_i is implied by this requirement. The answer is:

$$\begin{aligned}
 ADV_i = (hu)_{i+\frac{1}{2}} & \left\{ \frac{g}{2\Delta x} [(h+h_S)_{i+1} + (h+h_S)_i] \right\} \\
 & - (hu)_{i-\frac{1}{2}} \left\{ \frac{g}{2\Delta x} [(h+h_S)_i + (h+h_S)_{i-1}] \right\}.
 \end{aligned} \tag{8.37}$$

This has the form of a finite-difference flux divergence, as desired. Substituting back, we find that the potential energy equation can be written as

$$\begin{aligned}
 & \frac{d}{dt} \left[h_i g \left(h_S + \frac{1}{2} h \right)_i \right] \\
 & + (hu)_{i+\frac{1}{2}} \left\{ \frac{g}{2\Delta x} [(h+h_S)_{i+1} + (h+h_S)_i] \right\} - (hu)_{i-\frac{1}{2}} \left\{ \frac{g}{2\Delta x} [(h+h_S)_i + (h+h_S)_{i-1}] \right\} \\
 & - \frac{g}{2} \left\{ (hu)_{i+\frac{1}{2}} \left[\frac{(h+h_S)_{i+1} - (h+h_S)_i}{\Delta x} \right] + (hu)_{i-\frac{1}{2}} \left[\frac{(h+h_S)_i - (h+h_S)_{i-1}}{\Delta x} \right] \right\} = 0.
 \end{aligned} \tag{8.38}$$

Now consider the time-rate-of-change terms of (8.32). Obviously, the first line of (8.32) must be analogous to $\frac{\partial}{\partial t}(hK)$. For convenience, we define

$$(\text{KE tendency})_i \equiv K_i \frac{dh_i}{dt} + \frac{1}{2} \left[(hu)_{i+\frac{1}{2}} \frac{d}{dt} u_{i+\frac{1}{2}} + (hu)_{i-\frac{1}{2}} \frac{d}{dt} u_{i-\frac{1}{2}} \right]. \tag{8.39}$$

Substituting for the mass fluxes from (8.21), we can write (8.39) as

$$(\text{KE tendency})_i = K_i \frac{dh_i}{dt} + \frac{1}{8} \left[(h_{i+1} + h_i) \frac{d}{dt} \left(u_{i+\frac{1}{2}}^2 \right) + (h_i + h_{i-1}) \frac{d}{dt} \left(u_{i-\frac{1}{2}}^2 \right) \right]. \tag{8.40}$$

The requirement for kinetic energy conservation is

$$\sum_{\text{domain}} (\text{KE tendency})_i = \sum_{\text{domain}} \frac{d}{dt} (h_i K_i). \tag{8.41}$$

Note that only the sums over i must agree; it is not necessary that $(\text{KE tendency})_i$ be equal to $\frac{d}{dt}(h_i K_i)$ for each i . To complete our check of kinetic energy conservation, we substitute

for K_i on the right-hand side of (8.41), and check to see whether the resulting equation is actually satisfied.

The bad news is that if we use (8.26), we find that (8.41) is not satisfied. This means that we cannot have both momentum conservation under advection and kinetic energy conservation, when we start from the continuous form of (8.2).

Two alternative definitions of the kinetic energy are

$$K_i \equiv \frac{1}{4} \left(u_{i+\frac{1}{2}}^2 + u_{i-\frac{1}{2}}^2 \right) \quad (8.42)$$

and

$$h_i K_i \equiv \frac{1}{4} \left(h_{i+\frac{1}{2}} u_{i+\frac{1}{2}}^2 + h_{i-\frac{1}{2}} u_{i-\frac{1}{2}}^2 \right). \quad (8.43)$$

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 (8.42) is equal to the sum over the domain of $h_i K_i$ given by (8.44). Either choice allows (8.41) to be satisfied, so both are consistent with kinetic energy conservation under advection, but neither is consistent with momentum conservation under advection.

In summary, when we start from the continuous form of (8.2), 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 (8.4). In that case, we can conserve both momentum under advection and kinetic energy under advection.

When we generalize to the two-dimensional shallow-water equations with rotation, the issues discussed here have to be revisited.

8.2 Summary

With the one-dimensional advection equation we had “propagation” in one direction only, and with a second-order space-centered scheme for this equation we had to contend with one physical mode and one computational mode. The one-dimensional wave equation is equivalent to a pair of one-dimensional advection equations, so that we have a pair of physical modes and a pair of computational modes -- one of each propagating in each of the two spatial directions.

The advection equation “prefers” uncentered space differencing, but centered space differencing is most natural with the wave equation.

An unstaggered grid permits computational modes in space, essentially allowing two distinct solutions that live separate lives and do not interact. A staggered grid helps to overcome the problem of computational modes in space. Many staggering schemes are

possible, especially in two or three dimensions. Some of the schemes give particularly realistic propagation characteristics, allowing realistic simulation of such important processes as geostrophic adjustment.

In a model that permits both quasi-geostrophic motions and inertia-gravity waves, such as a primitive-equation model, the time step is normally limited by the stability criterion for the inertia-gravity waves.

Finally, 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.

Problems

1. Repeat the analysis of Section 8.1 using the Z grid instead of the C grid.
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. 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), and using fourth-order accuracy for the spatial derivatives. (It is not necessary 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 A-grid and the C-grid, using a mesh of 101×101 mass points, with periodic boundary conditions in both directions. Use leapfrog time differencing. Set $g = 0.1 \text{ m s}^{-2}$, $H = 10^3 \text{ m}$, $f = 0.5 \times 10^{-4} \text{ s}^{-1}$, and $d = 10^5 \text{ m}$.

In the square region

$$\begin{aligned} 45 \leq i \leq 55, \\ 45 \leq j \leq 55, \end{aligned} \quad (8.44)$$

apply a forcing in the continuity equation, of the form

$$\left(\frac{\partial h}{\partial t} \right)_{\text{noise}} = (-1)^{i+j} N e^{i\omega_N t}, \quad (8.45)$$

and set $\left(\frac{\partial h}{\partial t} \right)_{\text{noise}} = 0$ at all other grid points. Adopt the values

$\omega_N = 2\pi \times 10^{-3} \text{ s}^{-1}$; and $N = 10^{-4} \text{ m s}^{-1}$. In addition, for the entire domain apply a forcing of the form

$$\left(\frac{\partial h}{\partial t} \right)_{\text{smooth}} = S \sin\left(\frac{2\pi x}{L}\right) \sin\left(\frac{2\pi y}{L}\right) e^{i\omega_s t} \quad (8.46)$$

with $\omega_s = \frac{2\pi\sqrt{gH}}{L} \text{ s}^{-1}$ and $S = 10^{-4} \text{ m s}^{-1}$. Here L is $101 \times d$, the width of the domain.

Finally, include damping in the momentum equations, of the form

$$\begin{aligned} \left(\frac{\partial u}{\partial t}\right)_{\text{fric}} &= -Ku, \\ \left(\frac{\partial v}{\partial t}\right)_{\text{fric}} &= -Kv, \end{aligned} \tag{8.47}$$

where $K = 2 \times 10^{-5} \text{ s}^{-1}$. Because the model has both forcing and damping, it is possible to obtain a statistically steady solution.

- a) Analyze the stability of the two models without the forcing or damping terms. Using your results, choose a suitable time step for each model. Note: The forcing and damping terms are not expected to limit the time step, so we simply omit them in our analysis of the stability criterion.
- 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 $f = 2 \times 10^{-4} \text{ s}^{-1}$. Discuss the changes in your results.

CHAPTER 9

Vertical Differencing for Quasi-Static Models

Copyright 2004 David A. Randall

9.1 Introduction

Vertical differencing is a very different problem from horizontal differencing. This may seem odd, but the explanation is very simple. There are three primary factors. First, gravitational effects are very powerful, and act only in the vertical. Second, the Earth's atmosphere is very shallow compared to its horizontal extent. Third, the atmosphere has a lower boundary.

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.

9.2 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 1 second. This may be palatable if the horizontal and vertical grid spacings are comparable. 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 some issues with this approach, depending on the intended applications of the model, but anelastic models are very widely used and the anelastic equations can be an excellent choice for some applications, e.g., cloud 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 of equations filters vertically propagating sound waves, while permitting Lamb waves, which are sound waves that propagate only in the horizontal. The quasi-static system is widely used in global models for both weather prediction and climate.

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.” This means that small time steps are used to integrate the terms of the equations that govern sound waves, while longer time steps are used for the remaining terms.

9.3 General vertical coordinate

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 “vertical motion” seen in the coordinate system (less 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.

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 a general vertical coordinate, ζ , the quasi-static equation can be expressed as

$$\begin{aligned}\frac{\partial \phi}{\partial \zeta} &= \left(\frac{\partial \phi}{\partial p} \right) \left(\frac{\partial p}{\partial \zeta} \right) \\ &= \alpha m_\zeta ,\end{aligned}\tag{9.1}$$

where $\phi \equiv gz$ is the geopotential, g is the acceleration of gravity, z is height, p is the pressure, and α is the specific volume. In deriving (9.1), we have used the hydrostatic equation in the form

$$\frac{\partial \phi}{\partial p} = -\alpha ,\tag{9.2}$$

and we define

$$m_\zeta \equiv - \left(\frac{\partial p}{\partial \zeta} \right)\tag{9.3}$$

as the pseudo-density, i.e., the amount of mass (as measured by the pressure difference) between two ζ -surfaces. The minus sign in (9.3) is arbitrary, and can be included or not according to taste, perhaps depending on the particular choice of ζ .

The equation expressing conservation of an arbitrary intensive scalar, ψ , can be written as

$$\left(\frac{\partial}{\partial t} m_\zeta \psi \right)_\zeta + \nabla_\zeta \cdot (m_\zeta \mathbf{V} \psi) + \frac{\partial}{\partial p} (m_\zeta \dot{\zeta} \psi) = m_\zeta S_\psi .\tag{9.4}$$

Here

$$\dot{\zeta} \equiv \frac{D\zeta}{Dt}\tag{9.5}$$

is the rate of change of ζ following a particle, and S_ψ is the source or sink of ψ , per unit mass. Eq. (9.4) can be derived by adding up the fluxes of ψ across the boundaries of a control volume. We can obtain the continuity equation in ζ -coordinates from (9.4), by putting $\psi \equiv 1$ and $S_\psi \equiv 0$:

$$\left(\frac{\partial m_\zeta}{\partial t} \right)_\zeta + \nabla_\zeta \cdot (m_\zeta \mathbf{V}) + \frac{\partial}{\partial p} (m_\zeta \dot{\zeta}) = 0 .\tag{9.6}$$

By combining (9.4) and (9.6), we can obtain the advective form of the conservation equation for ψ :

$$m_\zeta \frac{D\psi}{Dt} = S_\psi, \quad (9.7)$$

where the Lagrangian or material time derivative is expressed by

$$\frac{D}{Dt}(\) = \left(\frac{\partial}{\partial t} \right)_\zeta + \mathbf{V} \cdot \nabla_\zeta + \dot{\zeta} \frac{\partial}{\partial \zeta}. \quad (9.8)$$

For example, the vertical pressure velocity,

$$\omega \equiv \frac{Dp}{Dt}, \quad (9.9)$$

can be written as

$$\begin{aligned} \omega &= \left(\frac{\partial p}{\partial t} \right)_\zeta + \mathbf{V} \cdot \nabla_\zeta p + \dot{\zeta} \frac{\partial p}{\partial \zeta} \\ &= \left(\frac{\partial p}{\partial t} \right)_\zeta + \mathbf{V} \cdot \nabla_\zeta p - m_\zeta \dot{\zeta}. \end{aligned} \quad (9.10)$$

The lower boundary condition, i.e., that no mass crosses the Earth's surface, is expressed by requiring that a particle which is on the Earth's surface remain there:

$$\frac{\partial \zeta_S}{\partial t} + \mathbf{V}_S \cdot \nabla \zeta_S - \dot{\zeta}_S = 0. \quad (9.11)$$

In the special case for which ζ_S is independent of time and the horizontal coordinates, (9.11) reduces to $\dot{\zeta}_S = 0$. Eq. (9.11) can actually be derived by integration of (9.6) throughout the entire atmospheric column, which gives

$$\begin{aligned} &\frac{\partial}{\partial t} \int_{\zeta_S}^{\zeta_T} m_\zeta ds + \nabla \cdot \left(\int_{\zeta_S}^{\zeta_T} m_\zeta \mathbf{V} ds \right) \\ &+ \left(\frac{\partial \zeta_S}{\partial t} + \mathbf{V}_S \cdot \nabla \zeta_S - \dot{\zeta}_S \right) - \left(\frac{\partial \zeta_T}{\partial t} + \mathbf{V}_T \cdot \nabla \zeta_T - \dot{\zeta}_T \right) = 0. \end{aligned} \quad (9.12)$$

Here ζ_T is the value of ζ at the top of the model atmosphere. We allow the possibility that the top of the model is placed at a finite height. Even if the top of the model is at the “top of the atmosphere,” i.e., at $p = 0$, the value of ζ_T may or may not be finite, depending on the

definition of ζ . The quantity $\frac{\partial \zeta_T}{\partial t} + \mathbf{V}_T \cdot \nabla \zeta_T - \dot{\zeta}_T$ represents the mass flux across the top of the atmosphere, which we assume to be zero. Similarly, $\frac{\partial \zeta_S}{\partial t} + \mathbf{V}_S \cdot \nabla \zeta_S - \dot{\zeta}_S$, which is identical to the left-hand side of (9.11), represents the mass flux across the Earth's surface.

Substituting (9.11) into (9.12), we find that

$$\frac{\partial}{\partial t} \int_{\zeta_S}^{\zeta_T} m_\zeta d\zeta + \nabla \cdot \left(\int_{\zeta_S}^{\zeta_T} m_\zeta \mathbf{V} d\zeta \right) = 0. \quad (9.13)$$

In view of (9.3), this is equivalent to

$$\frac{\partial p_S}{\partial t} - \frac{\partial p_T}{\partial t} + \nabla \cdot \left(\int_{p_T}^{p_S} \mathbf{V} dp \right) = 0, \quad (9.14)$$

which is the surface pressure tendency equation. Depending on the definitions of ζ and ζ_T , it may or may not be appropriate to set $\frac{\partial p_T}{\partial t} = 0$. By analogy with the derivation of (9.14), we can show that the pressure tendency on an arbitrary ζ -surface satisfies

$$\left(\frac{\partial p}{\partial t} \right)_\zeta - \frac{\partial p_T}{\partial t} + \nabla \cdot \left(\int_\zeta^{\zeta_T} m_\zeta \mathbf{V} d\zeta \right) - (m_\zeta \dot{\zeta})_\zeta = 0. \quad (9.15)$$

The thermodynamic equation can be written as

$$c_p \left[\left(\frac{\partial T}{\partial t} \right)_\zeta + \mathbf{V} \cdot \nabla_\zeta T + \dot{\zeta} \frac{\partial T}{\partial \zeta} \right] = \omega \alpha + Q, \quad (9.16)$$

where c_p is the specific heat of air at constant pressure, α is the specific volume, and Q is the heating rate per unit mass. An alternative form of the thermodynamic equation is

$$\left(\frac{\partial \theta}{\partial t} \right)_\zeta + \mathbf{V} \cdot \nabla_\zeta \theta + \dot{\zeta} \frac{\partial \theta}{\partial \zeta} = \frac{Q}{\Pi}, \quad (9.17)$$

where

$$\Pi \equiv c_p \frac{T}{\theta} = c_p \left(\frac{p}{p_0} \right)^\kappa, \quad (9.18)$$

is the Exner function. In (9.18), θ is the potential temperature; p_0 is a constant reference pressure, usually taken to be 1000 hPa, and $\kappa \equiv \frac{R}{c_p}$, where R is the gas constant.

9.3.1 The equation of motion and the HPGF

The horizontal momentum equation can be written as

$$\left(\frac{\partial \mathbf{V}}{\partial t}\right)_\zeta + (\mathbf{V} \cdot \nabla_\zeta) \mathbf{V} + \zeta \frac{\partial \mathbf{V}}{\partial \zeta} = -\nabla_p \phi - f \mathbf{k} \times \mathbf{V} + \mathbf{F}. \quad (9.19)$$

Here $-\nabla_p \phi$ is the horizontal pressure-gradient force (HPGF), which is expressed as minus the gradient of the geopotential *along an isobaric surface*, and \mathbf{F} is the friction vector. Using the relation

$$\begin{aligned} \nabla_p &= \nabla_\zeta - (\nabla_\zeta p) \frac{\partial}{\partial p} \\ &= \nabla_\zeta + \frac{(\nabla_\zeta p)}{m_\zeta} \frac{\partial}{\partial \zeta}, \end{aligned} \quad (9.20)$$

we can rewrite the HPGF as

$$-\nabla_p \phi = -\nabla_\zeta \phi - \frac{1}{m_\zeta} \frac{\partial \phi}{\partial \zeta} (\nabla_\zeta p). \quad (9.21)$$

In view of (9.1), this can be expressed as

$$-\nabla_p \phi = -\nabla_\zeta \phi - \alpha \nabla_\zeta p. \quad (9.22)$$

This is a nice result. For the special case $\zeta \equiv z$, (9.23) reduces to $-\nabla_p \phi = -\alpha \nabla_z p$, and for the special case $\zeta = p$ it becomes $-\nabla_p \phi = -\nabla_p \phi$. These are both very familiar.

Another useful form of the HPGF is expressed in terms of the Montgomery potential, which is defined by

$$M \equiv c_p T + \phi. \quad (9.23)$$

For the special case in which $\zeta \equiv \theta$, which will be discussed in detail later, the hydrostatic equation (9.1) can be written as

$$\frac{\partial M}{\partial \theta} = \Pi. \quad (9.24)$$

With the use of (9.23) and (9.24), Eq. (9.21) can be expressed as

$$-\nabla_p \phi = -\nabla_\zeta M + \Pi \nabla_\zeta \theta. \quad (9.25)$$

This form of the HPGF will be discussed later.

When the HPGF is a gradient, it has no effect in the vorticity equation, since the curl of the gradient is always zero. It is apparent from (9.22) and (9.25), however, that in general the HPGF is not simply a gradient. When the HPGF is not a gradient, it can spin up or spin down a circulation on a ζ surface. From (9.22) we see that the HPGF is a pure gradient for $\zeta \equiv p$, and from (9.25) we see that the HPGF is a pure gradient for $\zeta \equiv \theta$. 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 (9.3), we can rewrite (9.21) as

$$-\nabla_p \phi = -\frac{1}{m_\zeta} \nabla_\zeta (m_\zeta \phi) - \frac{1}{m_\zeta} \frac{\partial}{\partial \zeta} (\phi \nabla_\zeta p). \quad (9.26)$$

Vertically integrating with respect to mass, we find that

$$-\int_{\zeta_T}^{\zeta_S} m_\zeta \nabla_p \phi d\zeta = -\nabla_\zeta \int_{\zeta_T}^{\zeta_S} m_\zeta \phi d\zeta - \phi_S \nabla p_S + \phi_T \nabla p_T. \quad (9.27)$$

Suppose that $\phi_T = \text{constant}$ or $\nabla p_T = 0$, and consider a line integral of the vertically integrated HPGF, i.e., $-\int_{\zeta_T}^{\zeta_S} m_\zeta \nabla_p \phi d\zeta$, along a closed path. It follows from (9.27) that the line integral must vanish if the surface geopotential is constant along the path of integration. In other words, *in the absence of topography along the bounding path there cannot be any net spin-up or spin-down of a circulation in the region enclosed by the path.* Later we will show how this important constraint can be mimicked in a vertically discrete model.

9.3.2 Vertical mass flux for a family of vertical coordinates

Konor and Arakawa (1997) derived a diagnostic equation that can be used to compute ζ 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.,

$$\zeta \equiv F(\theta, p, p_S). \quad (9.28)$$

While not completely general, Eq. (9.28) does cover a variety of interesting cases, which will be discussed below. By differentiating (9.28) with respect to time on a constant ζ surface, we find that

$$0 = \left[\frac{\partial}{\partial t} F(\theta, p, p_s) \right]_{\zeta}. \quad (9.29)$$

The chain rule tells us that this is equivalent to

$$\frac{\partial F}{\partial \theta} \left(\frac{\partial \theta}{\partial t} \right)_{\zeta} + \frac{\partial F}{\partial p} \left(\frac{\partial p}{\partial t} \right)_{\zeta} + \frac{\partial F}{\partial p_s} \frac{\partial p_s}{\partial t} = 0. \quad (9.30)$$

Substituting from (9.17), (9.15), and (9.14), we obtain

$$\begin{aligned} & \frac{\partial F}{\partial \theta} \left[- \left(\mathbf{V} \cdot \nabla_{\zeta} \theta + \zeta \frac{\partial \theta}{\partial \zeta} \right) + \frac{Q}{\Pi} \right] \\ & + \frac{\partial F}{\partial p} \left[\frac{\partial p_T}{\partial t} - \nabla \cdot \left(\int_{\zeta}^{\zeta_T} m_{\zeta} \mathbf{V} d\zeta \right) + (m_{\zeta} \dot{\zeta})_{\zeta} \right] \\ & + \frac{\partial F}{\partial p_s} \left[\frac{\partial p_T}{\partial t} - \nabla \cdot \left(\int_{p_T}^{p_s} m_{\zeta} \mathbf{V} dp \right) \right] = 0. \end{aligned} \quad (9.31)$$

This can be solved for the vertical velocity, $\dot{\zeta}$:

$$\begin{aligned} & \left\{ \frac{\partial \theta}{\partial \zeta} \frac{\partial F}{\partial \theta} - m_{\zeta} \frac{\partial F}{\partial p} \right\} \dot{\zeta} = \frac{\partial F}{\partial \theta} \left[- \mathbf{V} \cdot \nabla_{\zeta} \theta + \frac{Q}{\Pi} \right] \\ & + \frac{\partial F}{\partial p} \left[\frac{\partial p_T}{\partial t} - \nabla \cdot \left(\int_{\zeta}^{\zeta_T} m_{\zeta} \mathbf{V} d\zeta \right) \right] \\ & + \frac{\partial F}{\partial p_s} \left[\frac{\partial p_T}{\partial t} - \nabla \cdot \left(\int_{p_T}^{p_s} m_{\zeta} \mathbf{V} dp \right) \right]. \end{aligned} \quad (9.32)$$

Here we have assumed that the heating rate, Q , is not formulated as an explicit function of $\dot{\zeta}$. As a check, consider the special case $F \equiv p$, so that $m = 1$, and assume that $\frac{\partial p_T}{\partial t} = 0$, as would be natural for the case of pressure coordinates. Then (9.32) reduces to

$$\begin{aligned}
\dot{p} &= \nabla \cdot \left(\int_{\zeta}^{\zeta_T} m_p \mathbf{V} dp \right) \\
&= \nabla \cdot \left(\int_p^{p_T} \mathbf{V} dp \right) .
\end{aligned} \tag{9.33}$$

As a second special case, suppose that $F \equiv \theta$. Then (9.32) becomes

$$\dot{\theta} = \frac{Q}{\Pi}. \tag{9.34}$$

Both of these are the expected results.

We assume that the model top is a surface of constant ζ , i.e., $\zeta = \zeta_T$. Because (9.28) must apply at the model top, we can write

$$\frac{\partial F}{\partial \theta_T} \left(\frac{\partial \theta_T}{\partial t} \right)_{\zeta} + \frac{\partial F}{\partial p_T} \left(\frac{\partial p_T}{\partial t} \right)_{\zeta} + \frac{\partial F}{\partial p_S} \frac{\partial p_S}{\partial t} = 0. \tag{9.35}$$

9.4 Discussion of particular vertical coordinate systems

We now discuss the following nine particular choices of ζ :

- height, z
- pressure, p
- log-pressure, used in many theoretical studies
- sigma, defined by $\sigma \equiv \frac{p - p_T}{p_S - p_T}$, designed to simplify the lower boundary condition
- a “hybrid,” or “mix,” of sigma and pressure coordinates, used in numerous general circulation models including the forecast model of the European Centre for Medium Range Weather Forecasts
- eta, which is a modified sigma coordinate, defined by $\eta \equiv \left(\frac{p - p_T}{p_S - p_T} \right) \eta_S$, where η_S is a time-independent function of the horizontal coordinates
- potential temperature, θ , which has many attractive properties and is now being used

- entropy, $s = c_p \ln \theta$
- 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 (9.28).

9.4.1 Height

In height coordinates, the hydrostatic equation is

$$\frac{\partial p}{\partial z} = -\rho g, \quad (9.36)$$

where $\rho \equiv \frac{1}{\alpha}$ is the density. We can obtain (9.36) simply by flipping (9.2) over. For the case of the height coordinate, the pseudodensity reduces to ρg , which is proportional to the ordinary or “true” density.

The continuity equation in height coordinates is

$$\left(\frac{\partial \rho}{\partial t} \right)_z + \nabla_z \bullet (\rho \mathbf{V}) + \frac{\partial}{\partial z}(\rho w) = 0. \quad (9.37)$$

This equation is easy to understand, but it is mathematically complicated, in that 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

$$\frac{\partial z_S}{\partial t} + \mathbf{V}_S \bullet \nabla z_S - w_S = 0. \quad (9.38)$$

Normally we can assume that z_S is independent of time, but (9.38) 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.

Note that (9.37) and (9.38) are direct transcriptions of (9.6) and (9.11), respectively, with the appropriate changes in notation.

The thermodynamic energy equation is

$$c_p \rho \left(\frac{\partial T}{\partial t} \right)_z = -c_p \rho \left(\mathbf{V} \cdot \nabla_z T + w \frac{\partial T}{\partial z} \right) + \omega \alpha + Q. \quad (9.39)$$

Here Q is the diabatic heating per unit volume, and

$$\begin{aligned} \omega &= \left(\frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p + w \frac{\partial p}{\partial z} \\ &= \left(\frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p - \rho g w. \end{aligned} \quad (9.40)$$

By using (9.40) in (9.39), we find that

$$c_p \rho \left(\frac{\partial T}{\partial t} \right)_z = -c_p \rho \mathbf{V} \cdot \nabla_z T - \rho w c_p (\Gamma_d - \Gamma) + \left[\left(\frac{\partial p}{\partial t} \right)_z + \mathbf{V} \cdot \nabla_z p \right] + Q, \quad (9.41)$$

where the actual lapse rate and the dry-adiabatic lapse rate are given by

$$\Gamma \equiv -\frac{\partial T}{\partial z}, \quad (9.42)$$

and

$$\Gamma_d \equiv \frac{g}{c_p}, \quad (9.43)$$

respectively. This form of the thermodynamic equation 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 (9.15), which is

$$\left[\frac{\partial}{\partial t} p(z) \right]_z = -g \nabla_z \cdot \int_z^\infty (\rho \mathbf{V}) dz + g \rho(z) w(z) + \frac{\partial p_T}{\partial t}. \quad (9.44)$$

Substitution into (9.41) gives

$$\begin{aligned} c_p \rho \left(\frac{\partial T}{\partial t} \right)_z &= -c_p \rho \mathbf{V} \cdot \nabla_z T - \rho w c_p (\Gamma_d - \Gamma) \\ &+ \left[-g \nabla_z \cdot \int_z^\infty (\rho \mathbf{V}) dz + g \rho(z) w(z) + \frac{\partial p_T}{\partial t} \right] + \mathbf{V} \cdot \nabla_z p + Q. \end{aligned} \quad (9.45)$$

According to (9.45), the time rate of change of the temperature at a given height is influenced

by the motion field through a deep layer. An alternative, considerably simpler form of the thermodynamic energy equation is

$$\left(\frac{\partial \theta}{\partial t}\right)_z = -\left(\mathbf{V} \cdot \nabla_z \theta + w \frac{\partial \theta}{\partial z}\right) + \frac{Q}{\Pi}. \quad (9.46)$$

In quasi-static models using height coordinates, the equation of vertical motion is replaced by the hydrostatic equation, in which w does not even appear. How then can we compute w ? The height coordinate is not a member of the family of schemes defined by (9.28), and so (9.32), the formula for the vertical mass flux derived from (9.28), does not apply. Instead, w is 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 (9.32), but somewhat more complicated. The derivation of Richardson’s equation is also more complicated than the derivation of (9.32).

The equation of state is

$$p = \rho R T. \quad (9.47)$$

Logarithmic differentiation of (9.47) gives

$$\frac{1}{p} \left(\frac{\partial p}{\partial t}\right)_z = \frac{1}{\rho} \left(\frac{\partial \rho}{\partial t}\right)_z + \frac{1}{T} \left(\frac{\partial T}{\partial t}\right)_z. \quad (9.48)$$

The time derivatives can be eliminated by using (9.37), (9.44) and (9.45). After some manipulation, we find that

$$\begin{aligned} & c_p T \frac{\partial}{\partial z} (\rho w) + \rho w \left[g \frac{c_v}{R} + c_p (\Gamma_d - \Gamma) \right] \\ &= (-c_p \rho \mathbf{V} \cdot \nabla_z T + \mathbf{V} \cdot \nabla_z p) - c_p T \nabla_z \cdot (\rho \mathbf{V}) + g \frac{c_v}{R} \nabla_z \cdot \int_z^\infty (\rho \mathbf{V}) dz + Q, \end{aligned} \quad (9.49)$$

where

$$c_v \equiv c_p - R \quad (9.50)$$

is the specific heat of air at constant volume.

Eq. (9.49) can be simplified considerably as follows. Expand the vertical derivative term using the product rule:

$$\frac{c_p T}{\rho} \frac{\partial}{\partial z} (\rho w) = c_p T \frac{\partial w}{\partial z} + w \frac{c_p T}{\rho} \frac{\partial \rho}{\partial z}. \quad (9.51)$$

Logarithmic differentiation of the equation of state gives

$$\frac{1}{p} \frac{\partial p}{\partial z} = \frac{1}{\rho} \frac{\partial \rho}{\partial z} + \frac{1}{T} \frac{\partial T}{\partial z}, \quad (9.52)$$

which is equivalent to

$$\frac{1}{\rho} \frac{\partial \rho}{\partial z} = -\frac{\rho g}{p} + \frac{\Gamma}{T} = \frac{1}{T} \left(-\frac{g}{R} + \Gamma \right). \quad (9.53)$$

Substitute (9.53) into (9.51) to obtain

$$\frac{c_p T}{\rho} \frac{\partial}{\partial z} (\rho w) = c_p T \frac{\partial w}{\partial z} + w c_p \left(-\frac{g}{R} + \Gamma \right). \quad (9.54)$$

Finally, substitute (9.54) into (9.49), and combine terms, to obtain

$$\begin{aligned} \rho c_p T \frac{\partial w}{\partial z} &= (-c_p \rho \mathbf{V} \cdot \nabla_z T + \mathbf{V} \cdot \nabla_z p) - c_p T \nabla_z \cdot (\rho \mathbf{V}) \\ &\quad + g \frac{c_v}{R} \nabla_z \cdot \int_z^\infty (\rho \mathbf{V}) dz + Q. \end{aligned} \quad (9.55)$$

This is Richardson's equation. It can be solved as a linear first-order ordinary differential equation for $w(z)$, 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. A physical interpretation of (9.55) 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 (9.55) may tend to upset that balance.

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).

As an example to illustrate the implications of (9.55), suppose that we have horizontally uniform heating but no horizontal motion. Then (9.55) drastically simplifies to

$$\rho c_p T \frac{\partial w}{\partial z} = Q. \quad (9.56)$$

If the lower boundary is flat so that

$$w = 0 \text{ at } z = 0, \quad (9.57)$$

then we obtain

$$w(z) = \int_0^z \frac{Q}{\rho c_p T} dz, \quad (9.58)$$

i.e., heating (cooling) below a given level induces rising (sinking) motion at that level. The rising motion induced by heating below a given level can be interpreted as a manifestation of the upward movement of air particles as the air expands above the rigid lower boundary.

9.4.2 Pressure

The hydrostatic equation in pressure coordinates has already been stated; it is (9.2). The pseudo-density is simply unity, since (9.3) reduces to

$$m_p = 1 \quad (9.59)$$

As a result, the continuity equation in pressure coordinates is relatively simple; it is linear and does not involve a time derivative:

$$\nabla_p \bullet \mathbf{V} + \frac{\partial \omega}{\partial p} = 0. \quad (9.60)$$

On the other hand, the lower boundary condition is complicated:

$$\frac{\partial p_S}{\partial t} + \mathbf{V}_S \bullet \nabla p_S - \omega_S = 0. \quad (9.61)$$

Recall that p_S can be predicted using the surface pressure-tendency equation, (9.14). Substitution from (9.14) into (9.61) gives

$$\omega_S = \frac{\partial p_T}{\partial t} - \nabla \bullet \left(\int_{p_T}^{p_S} \mathbf{V} dp \right) + \mathbf{V}_S \bullet \nabla p_S, \quad (9.62)$$

which can be used to diagnose ω_S . Nevertheless, the fact that pressure surfaces intersect the ground at locations that change with time (unlike height coordinates), means that models that uses pressure coordinates are complicated. Largely for this reason pressure coordinates are hardly ever used in numerical models.

With the pressure coordinate, we can write

$$\frac{\partial}{\partial t} \left(\frac{\partial \phi}{\partial p} \right)_p = - \frac{R}{p} \left(\frac{\partial T}{\partial t} \right)_p. \quad (9.63)$$

This allows us to eliminate the temperature in favor of the geopotential, which is often done in theoretical studies.

9.4.3 Log-pressure

Let T_0 be a constant reference temperature. Define the “log-pressure coordinate” z^* by the differential relationship

$$dz^* = -\frac{RT_0}{g} d(\ln p) = -\frac{RT_0}{g} \frac{dp}{p}. \quad (9.64)$$

Note that z^* has the units of length (i.e., height), and that $dz^* = dz$ when $T(p) = T_0$. Although generally $z \neq z^*$, we can force $z(p = p_S) = z^*(p = p_S)$. From (9.64), we see that

$$\frac{\partial \phi^*}{\partial p} = -\frac{RT_0}{p}, \quad (9.65)$$

where

$$\phi^* \equiv gz^*. \quad (9.66)$$

We also have the hydrostatic equation in the form

$$\frac{\partial \phi}{\partial p} = -\frac{RT}{p}. \quad (9.67)$$

Subtracting (9.65) from (9.67), we obtain a useful form of the hydrostatic equation:

$$\frac{\partial}{\partial p}(\phi - \phi^*) = -\frac{R(T - T_0)}{p}. \quad (9.68)$$

Since ϕ^* and T_0 are independent of time, we see that

$$\frac{\partial}{\partial t} \left(\frac{\partial \phi}{\partial p} \right)_{z^*} = -\frac{R}{p} \left(\frac{\partial T}{\partial t} \right)_{z^*}. \quad (9.69)$$

9.4.4 The σ -coordinate

The σ -coordinate of Phillips (1957) is defined by

$$\sigma \equiv \frac{p - p_T}{\pi}, \quad (9.70)$$

where we define

$$\pi \equiv p_S - p_T, \quad (9.71)$$

which is independent of height. Obviously,

$$\sigma_S = 1 \text{ and } \sigma_T = 0. \quad (9.72)$$

Note that for a fixed value of σ

$$dp = \sigma d\pi, \quad (9.73)$$

where the differential can represent a fluctuation in either time or (horizontal) space, with a fixed value of σ . Also,

$$\frac{\partial}{\partial p}(\) = \frac{1}{\pi} \frac{\partial}{\partial \sigma}(\). \quad (9.74)$$

Here the differentials are evaluated at fixed horizontal position. The pseudodensity in σ -coordinates in σ -coordinates is

$$m_\sigma = \pi, \quad (9.75)$$

which is independent of height. The continuity equation in σ -coordinates can therefore be written as

$$\frac{\partial \pi}{\partial t} + \nabla_\sigma \cdot (\pi \mathbf{V}) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma}) = 0. \quad (9.76)$$

Although this equation does contain a time derivative, the differentiated quantity, π , is independent of height, which makes (9.76) considerably simpler than (9.6).

The lower boundary condition in σ -coordinates is very simple:

$$\dot{\sigma} = 0 \text{ at } \sigma = 1. \quad (9.77)$$

This simplicity was in fact Phillips' motivation for the invention of σ -coordinates. The upper boundary condition is similar:

$$\dot{\sigma} = 0 \text{ at } \sigma = 0. \quad (9.78)$$

The continuity equation in σ -coordinates plays a dual role. First, it is used to predict the surface pressure. This is done by integrating (9.76) through the depth of the vertical column and using the boundary conditions (9.77) and (9.78), to obtain the surface pressure-tendency equation in the form

$$\frac{\partial \pi}{\partial t} = -\nabla \cdot \left(\int_0^1 \pi \mathbf{V} d\sigma \right). \quad (9.79)$$

The continuity equation is also used to determine $\pi\dot{\sigma}$. Once $\frac{\partial\pi}{\partial t}$ has been evaluated using (9.79), which does not involve $\pi\dot{\sigma}$, we can substitute back into (9.76) to obtain

$$\frac{\partial}{\partial\sigma}(\pi\dot{\sigma}) = \nabla \cdot \left(\int_0^1 \pi \mathbf{V} d\sigma \right) - \nabla_\sigma \cdot (\pi \mathbf{V}). \quad (9.80)$$

This can be integrated vertically to obtain $\pi\dot{\sigma}$ as a function of σ , starting from either the Earth's surface or the top of the atmosphere, and using the appropriate boundary condition at the top or bottom. The same result is obtained regardless of the direction of integration.

The hydrostatic equation is simply

$$\frac{1}{\pi} \frac{\partial\phi}{\partial\sigma} = -\alpha. \quad (9.81)$$

Finally, the horizontal pressure-gradient force takes a relatively complicated form:

$$\text{HPGF} = -\sigma\alpha\nabla\pi - \nabla_\sigma\phi. \quad (9.82)$$

Using the hydrostatic equation, (9.81), we can rewrite this as

$$\text{HPGF} = \sigma \frac{1}{\pi} \frac{\partial\phi}{\partial\sigma} \nabla\pi - \nabla_\sigma\phi. \quad (9.83)$$

Rearranging, we find that

$$\begin{aligned} \pi(\text{HPGF}) &= \sigma \frac{\partial\phi}{\partial\sigma} \nabla\pi - \pi \nabla_\sigma\phi \\ &= \left[\frac{\partial}{\partial\sigma}(\sigma\phi) - \phi \right] \nabla\pi - \pi \nabla_\sigma\phi. \\ &= \frac{\partial}{\partial\sigma}(\sigma\phi) \nabla\pi - \nabla_\sigma(\pi\phi). \end{aligned} \quad (9.84)$$

Vertically integrating (9.84) through the entire vertical column, we obtain

$$\int_0^1 \pi(\text{HPGF}) d\sigma = \phi_s \nabla\pi - \nabla_\sigma \left(\int_0^1 \pi\phi d\sigma \right). \quad (9.85)$$

When we integrate around any closed path, the second term on the right-hand side of (9.84) vanishes because it is the integral of a gradient. The first term also vanishes, unless there is topography along the path of integration. In short, the vertically integrated HPGF vanishes

except in the presence of topography, in which case “mountain torque” may result. This conclusion is reached very easily when we start from (9.84).

Consider the two contributions to the HPGF when evaluated near a mountain, as illustrated in Fig. 9.1. Near steep topography, the spatial variations of p_S and the near-surface

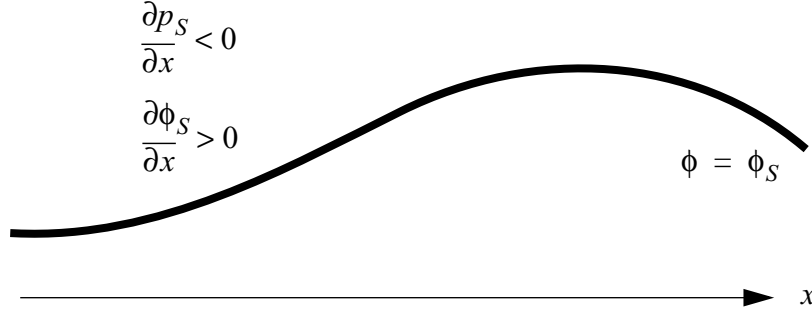


Figure 9.1: A mountain. As we move uphill in the x direction, the surface pressure decreases and the surface geopotential increases.

value of ϕ , along a σ -surface, are strong and of opposite sign. For example, moving uphill p_S decreases while ϕ_S increases. As a result, the two terms on the right-hand side of (9.82) are individually large and opposing, and the HPGF is the relatively small difference between them -- a dangerous situation. In numerical models based on the σ -coordinate, near steep mountains the relatively small discretization errors in the individual terms of the right-hand side of (9.82) can be as large as the HPGF. This will be discussed further below.

9.5 More on the HPGF in σ -coordinates

Consider Fig. 9.2. At the point O, we have $\sigma = \sigma^*$ and $p = p^*$. We can write

$$-\nabla_p \phi = -\nabla_\sigma \phi + (\nabla_\sigma \phi - \nabla_p \phi) = -\nabla_\sigma \phi + \nabla(\phi_{\sigma=\sigma^*} - \phi_{p=p^*}). \quad (9.86)$$

Compare with (9.82). Evidently

$$-\sigma \alpha \nabla \pi = \nabla + (\phi_{\sigma=\sigma^*} - \phi_{p=p^*}). \quad (9.87)$$

The right-hand-side of (9.86) involves the gradient of the difference between ϕ on a σ -surface and ϕ on a p -surface. Computation of this difference in a vertically discrete model amounts to vertical interpolation of ϕ from a σ -surface to a p -surface, and therefore should depend on the temperature through the hydrostatic equation. For a model that is discrete in both the horizontal and vertical, we must choose $\delta\sigma$ and δx so that

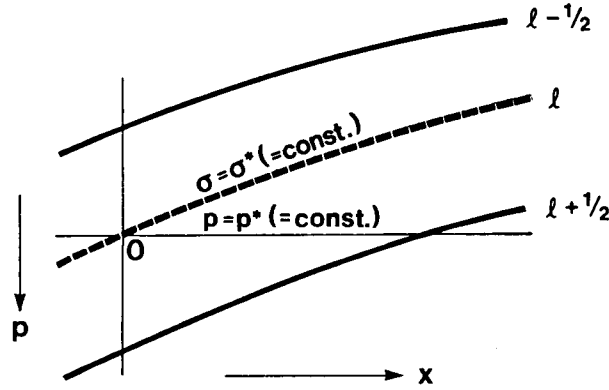


Figure 9.2: Evaluating the horizontal pressure gradient force.

$$\frac{\delta\sigma}{\delta x} \geq \frac{\left| \left(\frac{\delta\phi}{\delta x} \right)_\sigma \right|}{\left| \left(\frac{\delta\phi}{\delta\sigma} \right)_x \right|} \sim \frac{|\text{rate of change of } \phi \text{ along a } \sigma\text{-surface}|}{p_s \alpha}. \quad (9.88)$$

The numerator of the right-hand side of (9.88) increases when the terrain is steep. The denominator increases when T is warm, i.e. near the surface. The inequality (9.88) means that δx must be fine enough for a given $\delta\sigma$; this shows that increasing the vertical resolution of a model σ -coordinate can cause problems unless the horizontal resolution is correspondingly increased. One way to minimize problems is to artificially smooth the topography.

9.5.1 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 basic reference on this topic is the work of Simmons and Burridge (1981). They recommended the coordinate

$$\xi(p, p_s) \equiv \frac{p}{p_s} + \left(\frac{p}{p_s} - 1 \right) \left(\frac{p}{p_s} - \frac{p}{p_0} \right), \quad (9.89)$$

where p_0 is specified as 1013.2 hPa. Inspection of (9.89) shows that $\xi = 1$ for $p = p_s$, and $\xi = 0$ for $p = 0$. Eq. (9.89) can be expanded and simplified to yield

$$\xi = \frac{p}{p_0} + \left(\frac{p}{p_S}\right)^2 \left(1 - \frac{p_S}{p_0}\right). \quad (9.90)$$

Inspection of (9.90) shows that $\xi \rightarrow \frac{p}{p_0}$ as $\frac{p}{p_S} \rightarrow 0$. With (9.89), the pressure on an ξ -surface varies by less than one percent near the 10 mb level as the surface pressure varies in the range 1013 mb to 500 mb. When we evaluate the HPGF with the ξ -coordinate, there are still two terms, as with the σ -coordinate, but above the lower troposphere one of the terms is strongly dominant.

9.6 The η -coordinate

As a solution to the problem with the HPGF in σ -coordinates, Mesinger and Janjic (1985) proposed the η -coordinate, which is being used operationally at NCEP (the National Centers for Environmental Prediction):

$$\eta = \sigma \eta_S \quad (9.91)$$

where

$$\eta_S = \frac{p_{rf}(z_S) - p_T}{p_{rf}(0) - p_T}. \quad (9.92)$$

Whereas $\sigma = 1$ at the Earth's surface, Eq. (9.91) shows that $\eta = \eta_S$ at the Earth's surface. According to (9.92), $\eta_S = 1$ (just as $\sigma_S = 1$) 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 as a typical surface pressure for $z = z_S$. Because z_S depends on the horizontal coordinates, $p_{rf}(z_S)$ does too, and so, therefore, does η_S . In fact, after choosing $p_{rf}(z_S)$ and $z_S(x, y)$, one can make a map of $\eta_S(x, y)$, and of course *this map is independent of time*.

When we build a σ -coordinate model, we must specify (i.e., choose) fixed values of σ to serve as layer-edges and/or layer centers. Similarly, when we build an η -coordinate model, we must specify fixed values of η to serve as layer edges and/or layer centers. The values of η to be chosen include the possible values of η_S . This means that only a few discrete choices of η_S are permitted; the number increases as the vertical resolution of the model increases. Mountains must come in a few discrete sizes, like off-the-rack clothing! This is sometimes called the “step-mountain” approach. Fig. 9.3 shows how the η -coordinate works near mountains. Note that, unlike σ -surfaces, η -surfaces are nearly flat in the sense that the pressure is nearly uniform on them. The circled u -points have $u = 0$, as a boundary condition on the sides of the mountains.

In η -coordinates, the HPGF still consists of two terms:

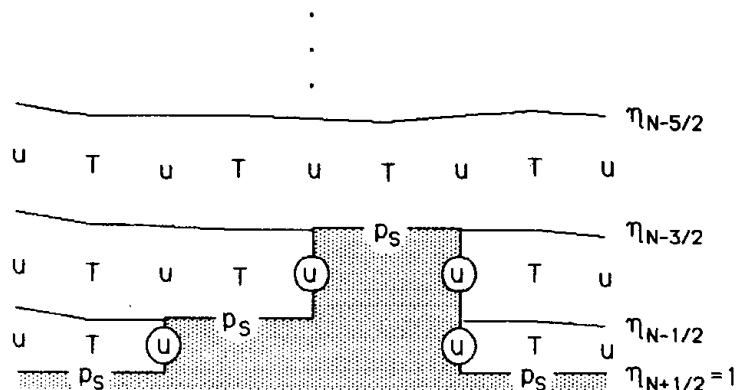


Figure 9.3: A schematic picture of the representation of mountains using the η coordinate.

$$-\nabla_p \phi = -\nabla_\eta \phi - \alpha \nabla_\eta p. \quad (9.93)$$

Because the η -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 does not occur.

9.6.1 Potential temperature

The potential temperature is defined by

$$\theta \equiv T \left(\frac{p_0}{p} \right)^\kappa. \quad (9.94)$$

The potential temperature increase upwards in a statically stable atmosphere, so that in such a case there is a monotonic relationship between θ and z . Note, however, that potential temperature cannot be used as a vertical coordinate when static instability occurs, and that the vertical resolution of a θ -coordinate model becomes very poor when the atmosphere is close to neutrally stable.

Potential temperature coordinates have particularly useful properties that have been recognized for many years, and have become more widely appreciated during the past decade or so. In the absence of heating, potential temperature is conserved following a particle. This means that the vertical motion in θ -coordinates is proportional to the heating rate:

$$\dot{\theta} = \frac{\theta}{c_p T} Q; \quad (9.95)$$

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. (9.95) is in fact an expression of the thermodynamic energy equation in θ -coordinates. In fact, θ -coordinates provide an especially simply pathway for the derivation of many important results, including the conservation equation for the Ertel potential vorticity. In addition, θ -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; Hsu and Arakawa, 1990).

The continuity equation in θ -coordinates is given by

$$\left(\frac{\partial m_\theta}{\partial t}\right)_\theta + \nabla_\theta \bullet (m_\theta \mathbf{V}) + \frac{\partial}{\partial p}(m_\theta \dot{\theta}) = 0, \quad (9.96)$$

which is a direct transcription of (9.6). Note, however, that $\dot{\theta} = 0$ in the absence of heating; in such case, (9.96) reduces to

$$\left(\frac{\partial m_\theta}{\partial t}\right)_\theta + \nabla_\theta \bullet (m_\theta \mathbf{V}) = 0, \quad (9.97)$$

which is analogous to the continuity equation of a shallow-water model.

The lower boundary condition in θ -coordinates is

$$\frac{\partial \theta_S}{\partial t} + \mathbf{V}_S \bullet \nabla \theta_S - \dot{\theta}_S = 0. \quad (9.98)$$

This equation must be used to predict θ_S . The complexity of the lower boundary condition in θ -coordinates is one of its chief drawbacks.

For the case of θ -coordinates, the hydrostatic equation, (9.1), reduces to

$$\frac{\partial \phi}{\partial \theta} = -\alpha \frac{\partial p}{\partial \theta}. \quad (9.99)$$

“Logarithmic differentiation” of (9.94) gives

$$\frac{d\theta}{\theta} = \frac{dT}{T} - \kappa \frac{dp}{p}. \quad (9.100)$$

It follows that

$$\alpha \frac{\partial p}{\partial \theta} = c_p \frac{\partial T}{\partial \theta} - c_p \frac{T}{\theta}. \quad (9.101)$$

Substitution of (9.101) into (9.99) gives

$$\frac{\partial M}{\partial \theta} = \Pi. \quad (9.102)$$

The HPGF in θ -coordinates is

$$\text{HPGF} = -\alpha \nabla_{\theta} p - \nabla_{\theta} \phi. \quad (9.103)$$

From (9.94), we see that

$$\frac{d\theta}{\theta} = \frac{dT}{T} - \kappa \frac{dp}{p}. \quad (9.104)$$

It follows that

$$\nabla_{\theta} p = c_p \left(\frac{p}{RT} \right) \nabla_{\theta} T. \quad (9.105)$$

Substitution of (9.105) into (9.103) gives

$$\text{HPGF} = -\nabla_{\theta} M. \quad (9.106)$$

Of course, θ -surfaces can intersect the Earth's surface, but we can consider these to follow the Earth's surface, by defining imaginary “massless layers,” as shown in Fig. 9.4. Since no mass resides between the θ surfaces in the portion of the domain where they “touch the Earth's surface,” no harm is done by this fantasy.

Obviously, a model that follows this approach has to be able to deal with massless layers. This practical difficulty has led most modelers to avoid θ -coordinates up to this time.

9.6.2 Entropy

The entropy coordinate is very similar to the θ -coordinate. We define the entropy by

$$s = c_p \ln \theta, \quad (9.107)$$

so that

$$ds = c_p \frac{d\theta}{\theta}. \quad (9.108)$$

The hydrostatic equation can then be written as

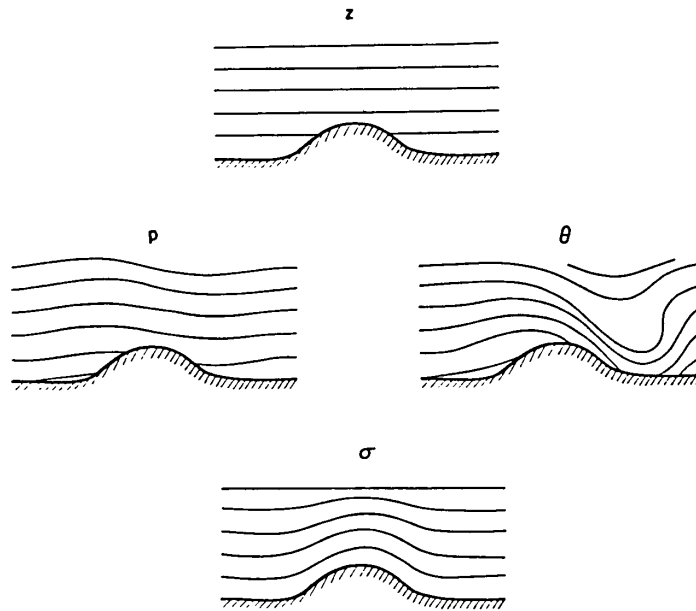


Figure 9.4: Four possible vertical coordinate systems.

$$\frac{\partial M}{\partial s} = T. \quad (9.109)$$

This is a particularly attractive form because the “thickness” is simply given by the temperature.

9.6.3 Hybrid σ - θ coordinates

Konor and Arakawa (1997) discuss a hybrid vertical coordinate that reduces to θ away from the surface, and to σ near the surface. This hybrid coordinate is designed to combine the strengths of θ and σ coordinates, while avoiding their weaknesses. Similar efforts have been reported by other authors, e.g. Johnson and Uccellini (1983) and Zhu et al. (1992). For further discussion, see the paper of Konor and Arakawa (1997).

9.6.4 Summary of vertical coordinate systems

The table on the following page 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. At present, there seems to be a movement away from σ coordinates and towards θ or hybrid θ - σ coordinates.

| Coordinate | Hydrostatic s | HPGF | Vertical Velocity | Continuity | LBC |
|--|---|--|---|---|--|
| z | $\frac{\partial p}{\partial z} = -\rho g$ | $-\alpha \nabla_z p$ | $w \equiv \frac{Dz}{Dt}$ | $\frac{\partial \rho}{\partial t} + \nabla_z \bullet (\rho \mathbf{V}_H) + \frac{\partial}{\partial z}(\rho w) = 0$ | $\mathbf{V}_S \bullet \nabla_{z_S} - w_S = 0$ |
| p | $\frac{\partial \phi}{\partial p} = -\alpha$ | $-\nabla_p \phi$ | $\omega \equiv \frac{Dp}{Dt}$ | $\nabla_p \bullet \mathbf{V}_H + \frac{\partial \omega}{\partial p} = 0$ | $\frac{\partial p_S}{\partial t} + \mathbf{V}_S \bullet \nabla p_S - \omega_S = 0$ |
| $z^* \equiv -\frac{RT_0}{g} \ln\left(\frac{p}{p_0}\right)$ | $\frac{\partial z}{\partial z^*} = -\frac{T}{T_0}$ | $-\nabla_{z^*} \phi$ | $w^* \equiv \frac{Dz^*}{Dt}$ $\equiv \frac{-H\omega}{p}$ | $\nabla_{z^*} \bullet \mathbf{V}_H + \frac{\partial w^*}{\partial z^*} - \frac{w^*}{H} = 0$ | $\frac{\partial z_S^*}{\partial t} + \mathbf{V}_S \bullet \nabla_{z_S^*} - w_S^* = 0$ |
| $\sigma \equiv \frac{p - p_T}{\pi}$ | $\frac{1}{\pi} \frac{\partial \phi}{\partial \sigma} = -\alpha$ | $-\nabla_\sigma \phi - \sigma \alpha \nabla \pi$ | $\dot{\sigma} \equiv \frac{D\sigma}{Dt}$ | $\frac{\partial \pi}{\partial t} + \nabla_\sigma \bullet (\pi \mathbf{V}_H) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma}) = 0$ | $-\dot{\sigma}_S = 0$ |
| q | $\frac{\partial \psi}{\partial \theta} = \Pi$ | $-\nabla_\theta \psi$ | $\dot{\theta} \equiv \frac{D\theta}{Dt}$ | $\frac{\partial m}{\partial t} + \nabla_\theta \bullet (m \mathbf{V}_H) + \frac{\partial}{\partial \theta}(m \dot{\theta}) = 0$ | $\frac{\partial \theta_S}{\partial t} + \mathbf{V}_S \bullet \nabla \theta_S - \dot{\theta}_S = 0$ |
| s | $\frac{\partial \psi}{\partial s} = T$ | $-\nabla_s \psi$ | $\dot{s} \equiv \frac{Ds}{Dt}$ | $\frac{\partial \mu}{\partial t} + \nabla_s \bullet (\mu \mathbf{V}_H) + \frac{\partial}{\partial s}(\mu \dot{s}) = 0$ | $\frac{\partial s_S}{\partial t} + \mathbf{V}_S \bullet \nabla s_S - \dot{s}_S = 0$ |

The θ -coordinate has many advantages. In the absence of heating, $\frac{D\theta}{Dt} \equiv \dot{\theta} = 0$, so there is “no vertical velocity.” This helps to minimize, for example, the problems associated with, e.g., the vertical advection of moisture. The HPGF has the simple form

$$-\nabla_p \phi = -\nabla_\theta (c_p T + gz), \quad (9.110)$$

i.e. it is a gradient. The quantity

$$M \equiv c_p T + \phi \quad (9.111)$$

is called the “Montgomery potential” or “Montgomery stream function.” It satisfies a form of the hydrostatic equation that is natural for use with θ -coordinates:

$$\frac{\partial M}{\partial \theta} = c_p \left(\frac{p}{p_0} \right)^\kappa, \quad (9.112)$$

where $\kappa \equiv R/c_p$.

The dynamically important isentropic potential vorticity, q , is easily constructed in θ -coordinates, since it involves the curl of \mathbf{V} on a θ -surface:

$$q \equiv (\mathbf{k} \cdot \nabla_\theta \times \mathbf{V} + f) \frac{\partial \theta}{\partial p}. \quad (9.113)$$

The available potential energy is also easily obtained, since it involves the distribution of p on θ -surfaces.

9.7 Vertical staggering

After the choice of vertical coordinate system, the next issue is the choice of vertical staggering. Two possibilities are discussed here, and are illustrated in Fig. 9.5. These are the “Lorenz” or “L” grid, and the “Charney-Phillips” or “C-P” grid. Suppose that both grids have N wind-levels. The L-grid also has N θ -levels, while the C-P grid has $N + 1$ θ -levels. On both grids, ϕ is hydrostatically determined on the wind-levels, and

$$\phi_l - \phi_{l+1} \sim \theta_{l+\frac{1}{2}}. \quad (9.114)$$

(Exercise: Show that $\partial \phi \Delta \Pi = -\theta$, where $\Pi \equiv c_p(p/p_0)^\kappa$.)

On the C-P grid, θ is located between ϕ -levels, so (9.114) is convenient. With the L-grid, θ must be interpolated, e.g.

$$\phi_l - \phi_{l+1} \sim \frac{1}{2}(\theta_l + \theta_{l+1}) . \quad (9.115)$$

Because (9.115) involves averaging, an oscillation in θ is not “felt” by ϕ , and so has no effect on the winds. This allows the possibility of a computational mode in the vertical. No such problem occurs with the C-P grid.

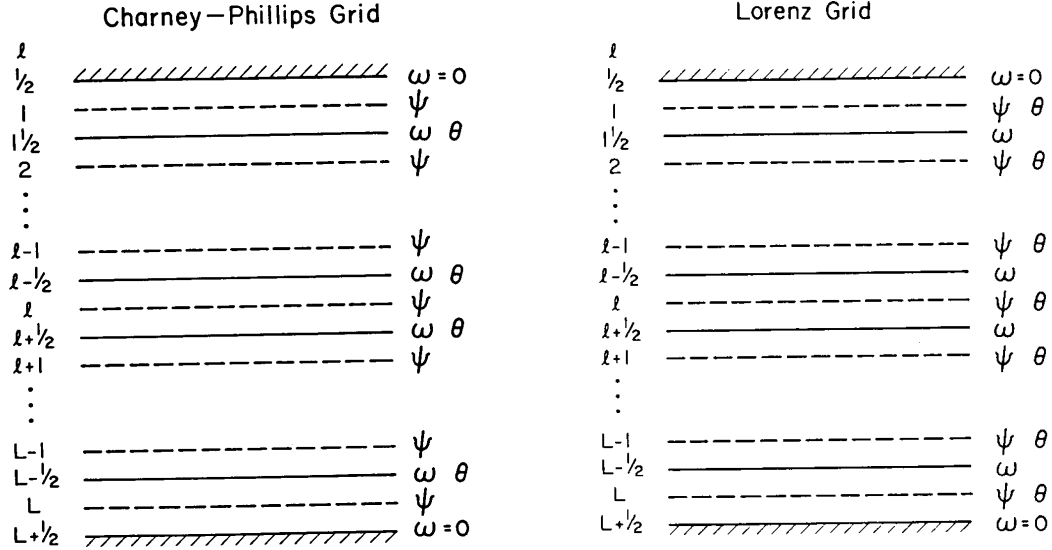


Figure 9.5: Schematic illustration of the Charney – Phillips grid and the Lorenz grid.

There is a second, less obvious problem with the L grid. The vertically discrete potential vorticity corresponding to (9.113) is

$$q_l \equiv (\mathbf{k} \cdot \nabla_{\theta} \times \mathbf{V}_l + f) \left(\frac{\partial \theta}{\partial p} \right)_l . \quad (9.116)$$

It is obvious that (9.116) “wants” the potential temperature to be defined at levels “in between” the wind levels, as they are on the C-P grid. In contrast, on the L grid the potential temperature and wind are defined at the same level. Suppose that we have N wind levels. Then with the C-P 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 kind of computational baroclinic instability (Arakawa and Moorthi, 1988). This is a drawback of the L grid.

As Lorenz (1955) pointed out, however, the L-grid is very convenient for maintaining total energy conservation, because the kinetic and thermodynamic energies are defined at the

same levels. Today, almost all models use the L-grid. This may change.

9.8 Conservation properties of vertically discrete models using σ -coordinates

We now investigate conservation properties of the vertically discretized equations, using σ -coordinates, and using the L-grid. The discussion follows Arakawa and Lamb (1977), although some of the ideas originated with Lorenz (1960). For simplicity, we consider only vertical discretization, and keep the temporal and horizontal derivatives in continuous form.

Conservation of mass is expressed, in the vertically discrete system, by

$$\frac{\partial \pi}{\partial t} + \nabla_{\sigma} \bullet (\pi \mathbf{V}_l) + \left[\frac{\delta(\pi \dot{\sigma})}{\delta \sigma} \right]_l = 0, \quad (9.117)$$

where

$$[\delta(\quad)]_l \equiv (\quad)_{l+\frac{1}{2}} - (\quad)_{l-\frac{1}{2}}. \quad (9.118)$$

Summing (9.117) over all levels, and using the boundary conditions

$$\dot{\sigma}_{\frac{1}{2}} = \dot{\sigma}_{L+\frac{1}{2}} = 0, \quad (9.119)$$

and

$$\sum_{l=1}^L \delta \sigma_l = 1, \quad (9.120)$$

we obtain

$$\frac{\partial \pi}{\partial t} + \nabla \bullet \sum_{l=1}^L [(\pi \mathbf{V}_l)(\delta \sigma_l)] = 0, \quad (9.121)$$

which is the vertically discrete form of the surface pressure tendency equation. From (9.121), 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 use

$$p_{l+\frac{1}{2}} = \pi \sigma_{l+\frac{1}{2}} + p_T, \quad (9.122)$$

where p_T is a constant, and the constant values of $\sigma_{l+\frac{1}{2}}$ are prescribed for each layer edge

when the model is started up. Eq. (9.122) tells how to compute layer-edge pressures. The method to discuss layer-center pressures will be discussed later.

Similarly, we can conserve an intensive scalar, such as the potential temperature θ , by using

$$\frac{\partial}{\partial t}(\pi\theta_l) + \nabla \cdot (\pi \mathbf{V}_l \theta_l) + \left[\frac{\delta(\pi \dot{\sigma} \theta)}{\delta \sigma} \right]_l = 0. \quad (9.123)$$

In order to use (9.123) it is necessary to define values of θ at the layer edges, via an interpolation. We have already discussed the interpolation issue in the context of advection, and that discussion applies to vertical advection as well as horizontal advection. In particular, the interpolation methods that allow conservation of an arbitrary function of the advected quantity can be used for vertical advection.

Now refer back to the discussion of the horizontal pressure-gradient force, in connection with (9.84) and (9.85). A finite-difference analog of (9.84) is

$$\pi(\text{HPGF})_l = \left[\frac{\delta(\sigma\phi)}{\delta \sigma} \right]_l \nabla \pi - \nabla(\pi\phi_l). \quad (9.124)$$

Multiplying (9.124) by $\delta\sigma_l$, and summing over all layers, we obtain

$$\begin{aligned} \sum_{l=1}^L \pi(\text{HPGF})_l (\delta\sigma)_l &= \sum_{l=1}^L [\delta(\sigma\phi)]_l \nabla \pi - \sum_{l=1}^L [\nabla(\pi\phi_l) (\delta\sigma)_l] \\ &= \phi_S \nabla \pi - \nabla \left\{ \sum_{l=1}^L [(\pi\phi_l) (\delta\sigma)_l] \right\}, \end{aligned} \quad (9.125)$$

which is a finite-difference analog of (9.85). This means that if we use the form of the HPGF given by (9.124), the vertically summed HPGF cannot generate a circulation inside a closed path, in the absence of topography (Arakawa and Lamb, 1977). *This “principle” provides a rational way to choose which of the many possible forms of the HPGF should be used in the model.* At this point, of course, the form is not fully determined, because we do not yet have a method to compute either ϕ_l or the layer-edge values of ϕ that appear in (9.124). A suitable method is derived below.

Eq. (9.124) is equivalent to

$$\pi(\text{HPGF})_l = \left\{ \left[\frac{\delta(\sigma\phi)}{\delta \sigma} \right]_l - \phi_l \right\} \nabla \pi - \pi \nabla \phi_l. \quad (9.126)$$

By comparison with (9.82), we identify

$$p_S(\sigma\alpha)_I = \phi_I - \left[\frac{\delta(\sigma\phi)}{\delta\sigma} \right]_I. \quad (9.127)$$

This will be used later.

Next consider total energy conservation. We begin by reviewing the continuous case. Potential temperature conservation is expressed by

$$\frac{\partial}{\partial t}(\pi\theta) + \nabla \cdot (\pi \mathbf{V}\theta) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma}\theta) = 0 \quad (9.128)$$

Here we assume no heating for simplicity. Using continuity this can be expressed in advective form:

$$\frac{\partial \theta}{\partial t} + \mathbf{V} \cdot \nabla \theta + \dot{\sigma} \frac{\partial \theta}{\partial \sigma} = 0. \quad (9.129)$$

With the use of the definition of θ , i.e.,

$$\theta = T \left(\frac{p_0}{p} \right)^\kappa, \quad (9.130)$$

and the equation of state, (9.129) can be used to derive the thermodynamic energy equation in the form

$$c_p \left(\frac{\partial T}{\partial t} + \mathbf{V} \cdot \nabla T + \dot{\sigma} \frac{\partial T}{\partial \sigma} \right) = \omega \alpha. \quad (9.131)$$

Here

$$\begin{aligned} \omega &\equiv \left(\frac{\partial p}{\partial t} \right)_\sigma + \mathbf{V} \cdot \nabla_\sigma p + \dot{\sigma} \frac{\partial p}{\partial \sigma} \\ &= \sigma \left(\frac{\partial \pi}{\partial t} + \mathbf{V} \cdot \nabla p_S \right) + \pi \dot{\sigma}. \end{aligned} \quad (9.132)$$

Continuity then allows us to transform (9.131) to the flux form:

$$\frac{\partial}{\partial t}(\pi c_p T) + \nabla \cdot (\pi \mathbf{V} c_p T) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma} c_p T) = \pi \omega \alpha. \quad (9.133)$$

The potential temperature equation, (9.128), is approximated by (9.123), which has already been discussed. Suppose that the model explicitly predicts θ_I . Adopting the

definition

$$\theta_l = \frac{T_l}{P_l}, \quad (9.134)$$

where for convenience we define

$$P_l \equiv \left(\frac{p_l}{p_0} \right)^\kappa, \quad (9.135)$$

we will now derive a finite-difference analog of (9.128), by starting from (9.123). Recall that the method to determine p_l has not been specified yet. The advective form corresponding to (9.123) is

$$\pi \left(\frac{\partial \theta_l}{\partial t} + \mathbf{V}_l \cdot \nabla \theta_l \right) + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\theta_{l+\frac{1}{2}} - \theta_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\theta_l - \theta_{l-\frac{1}{2}} \right) \right] = 0. \quad (9.136)$$

Substitute (9.134) into (9.136), to obtain the corresponding prediction equation for T_l :

$$\begin{aligned} & \pi \left(\frac{\partial T_l}{\partial t} + \mathbf{V}_l \cdot \nabla T_l \right) - \pi \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) \\ & + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(P_l \theta_{l+\frac{1}{2}} - T_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(T_l - P_l \theta_{l-\frac{1}{2}} \right) \right] = 0. \end{aligned} \quad (9.137)$$

The derivative $\frac{\partial P_l}{\partial \pi}$ cannot be evaluated until we specify the form of P_l . We now introduce the layer-edge temperatures, i.e., $T_{l+\frac{1}{2}}$ and $T_{l-\frac{1}{2}}$, although the method to determine them has not yet been specified. We rewrite (9.137) as

$$\begin{aligned}
 & \pi \left(\frac{\partial T_l}{\partial t} + \mathbf{V}_l \cdot \nabla T_l \right) + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(T_{l+\frac{1}{2}} - T_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(T_l - T_{l-\frac{1}{2}} \right) \right] \\
 & = \pi \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) \\
 & + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(T_{l+\frac{1}{2}} - P_l \theta_{l+\frac{1}{2}} \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(P_l \theta_{l-\frac{1}{2}} - T_{l-\frac{1}{2}} \right) \right] .
 \end{aligned} \tag{9.138}$$

Obviously the left-hand side of (9.138) can be rewritten in flux form through the use of the vertically discrete continuity equation:

$$\begin{aligned}
 & \frac{\partial}{\partial t} (\pi T_l) + \nabla \cdot (\pi \mathbf{V}_l T_l) + \left[\frac{\delta (\pi \dot{\sigma} T)}{\delta \sigma} \right]_l = \pi \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) \\
 & + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(T_{l+\frac{1}{2}} - P_l \theta_{l+\frac{1}{2}} \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(P_l \theta_{l-\frac{1}{2}} - T_{l-\frac{1}{2}} \right) \right] .
 \end{aligned} \tag{9.139}$$

By comparison of (9.133) with (9.139), we identify

$$\begin{aligned}
 & \frac{\pi (\omega \alpha)_l}{c_p} = \pi \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) \\
 & + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(T_{l+\frac{1}{2}} - P_l \theta_{l+\frac{1}{2}} \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(P_l \theta_{l-\frac{1}{2}} - T_{l-\frac{1}{2}} \right) \right] .
 \end{aligned} \tag{9.140}$$

This result will be used below.

Returning to the continuous case, we now derive the continuous mechanical energy equation, starting from the continuous momentum equation in the form

$$\left(\frac{\partial \mathbf{V}}{\partial t} \right)_\sigma + [f + \mathbf{k} \cdot (\nabla_\sigma \times \mathbf{V})] \mathbf{k} \times \mathbf{V} + \dot{\sigma} \frac{\partial \mathbf{V}}{\partial \sigma} + \nabla_\sigma K = -\nabla_\sigma \phi - \sigma \alpha \nabla \pi . \tag{9.141}$$

Here $K \equiv \frac{1}{2} \mathbf{V} \cdot \mathbf{V}$ is the kinetic energy per unit mass. Dotting (9.141) with \mathbf{V} gives the mechanical energy equation in the form

$$\left(\frac{\partial K}{\partial t} \right)_\sigma + \mathbf{V} \cdot \nabla_\sigma K + \dot{\sigma} \frac{\partial K}{\partial \sigma} = -\mathbf{V} \cdot (\nabla_\sigma \phi + \sigma \alpha \nabla \pi) . \tag{9.142}$$

The corresponding flux form is

$$\frac{\partial}{\partial t}(\pi K) + \nabla \cdot (\pi \mathbf{V} K) + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma} K) = -\pi \mathbf{V} \cdot (\nabla_{\sigma} \phi + \sigma \alpha \nabla \pi). \quad (9.143)$$

The pressure-work term on the right-hand side of (9.143) has to be manipulated to facilitate comparison with (9.133). Begin as follows:

$$\begin{aligned} -\pi \mathbf{V} \cdot (\nabla_{\sigma} \phi + \sigma \alpha \nabla \pi) &= -\nabla_{\sigma} \cdot (\pi \mathbf{V} \phi) + \phi \nabla_{\sigma} \cdot (\pi \mathbf{V}) - \pi \sigma \alpha \mathbf{V} \cdot \nabla \pi \\ &= -\nabla_{\sigma} \cdot (\pi \mathbf{V} \phi) - \phi \left[\frac{\partial \pi}{\partial t} + \frac{\partial}{\partial \sigma}(\pi \dot{\sigma}) \right] - \pi \sigma \alpha \mathbf{V} \cdot \nabla \pi \\ &= -\nabla_{\sigma} \cdot (\pi \mathbf{V} \phi) - \frac{\partial}{\partial \sigma}(\pi \dot{\sigma} \phi) + \pi \dot{\sigma} \frac{\partial \phi}{\partial \sigma} - \phi \frac{\partial \pi}{\partial t} - \pi \sigma \alpha \mathbf{V} \cdot \nabla \pi \\ &= -\nabla_{\sigma} \cdot (\pi \mathbf{V} \phi) - \frac{\partial}{\partial \sigma}(\pi \dot{\sigma} \phi) - \pi \dot{\sigma} \alpha p_s - \phi \frac{\partial \pi}{\partial t} - \pi \sigma \alpha \mathbf{V} \cdot \nabla \pi. \end{aligned} \quad (9.144)$$

In the final line of (9.144) we have used hydrostatics. Referring back to (9.132), we can write

$$\pi \dot{\sigma} \alpha \pi + \phi \frac{\partial \pi}{\partial t} + \pi \sigma \alpha \mathbf{V} \cdot \nabla \pi = \pi \omega \alpha + \frac{\partial}{\partial \sigma} \left(\phi \sigma \frac{\partial \pi}{\partial t} \right). \quad (9.145)$$

Substitution of (9.145) into (9.144) gives

$$-\pi \mathbf{V} \cdot (\nabla_{\sigma} \phi + \sigma \alpha \nabla \pi) = -\nabla_{\sigma} \cdot (\pi \mathbf{V} \phi) - \frac{\partial}{\partial \sigma} \left(\pi \dot{\sigma} \phi + \phi \sigma \frac{\partial \pi}{\partial t} \right) - \pi \omega \alpha. \quad (9.146)$$

Finally, plugging (9.146) back into (9.143), and collecting terms, gives the mechanical energy equation in the form

$$\frac{\partial}{\partial t}(\pi K) + \nabla \cdot [\pi \mathbf{V}(K + \phi)] + \frac{\partial}{\partial \sigma} \left[\pi \dot{\sigma}(K + \phi) + \phi \sigma \frac{\partial \pi}{\partial t} \right] = -\pi \omega \alpha. \quad (9.147)$$

Adding (9.133) and (9.147) gives a statement of the conservation of total energy:

$$\frac{\partial}{\partial t}[\pi(K + c_p T)] + \nabla \cdot [\mathbf{V} \pi(K + \phi + c_p T)] + \frac{\partial}{\partial \sigma} \left[\pi \dot{\sigma}(K + \phi + c_p T) + \phi \sigma \frac{\partial \pi}{\partial t} \right] = 0. \quad (9.148)$$

Integrating this through the depth of an atmospheric column, we find that

$$\frac{\partial}{\partial t} \left[\int_0^1 \pi(K + c_p T) d\sigma \right] + \nabla \cdot \left[\int_0^1 \mathbf{V} \pi(K + \phi + c_p T) d\sigma \right] + \phi_S \frac{\partial \pi}{\partial t} = 0, \quad (9.149)$$

which can also be written as

$$\frac{\partial}{\partial t} \left[\int_0^1 \pi(K + c_p T + \phi_S) d\sigma \right] + \nabla \cdot \left[\int_0^1 \mathbf{V} \pi(K + \phi + c_p T) d\sigma \right] = \pi \frac{\partial \phi_S}{\partial t}. \quad (9.150)$$

The right-hand side of (9.150) represents the work done on the atmosphere if the lower boundary is moving with time, e.g., in an earthquake.

We now carry out essentially the same derivation using the vertically discrete system. Taking the dot product of $\pi \mathbf{V}_l$ with the HPGF for layer l , we write, closely following (9.144)-(9.146),

$$\begin{aligned} -\pi \mathbf{V}_l \cdot [\nabla \phi_l + (\sigma \alpha)_l \nabla \pi] &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) + \phi_l \nabla \cdot (\pi \mathbf{V}_l) - \pi (\sigma \alpha)_l \mathbf{V}_l \cdot \nabla \pi \\ &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \phi_l \left\{ \frac{\partial \pi}{\partial t} + \left[\frac{\delta(\pi \dot{\sigma})}{\delta \sigma} \right]_l \right\} - \pi (\sigma \alpha)_l \mathbf{V}_l \cdot \nabla \pi \\ &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \left[\frac{\delta(\pi \dot{\sigma} \phi)}{\delta \sigma} \right]_l \\ &\quad + \frac{1}{(\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right) \right] \\ &\quad - \phi_l \frac{\partial \pi}{\partial t} - \pi (\sigma \alpha)_l \mathbf{V}_l \cdot \nabla \pi. \end{aligned} \quad (9.151)$$

Continuing down this path, we construct the terms that we need by adding and subtracting

$$\begin{aligned} -\pi \mathbf{V}_l \cdot [\nabla \phi_l + (\sigma \alpha)_l \nabla \pi] &= -\nabla \cdot (\pi \mathbf{V}_l \phi_l) - \left[\frac{\delta(\pi \dot{\sigma} \phi)}{\delta \sigma} \right]_l \\ &\quad + [\pi (\sigma \alpha)_l - \phi_l] \frac{\partial \pi}{\partial t} - \pi \left\{ (\sigma \alpha)_l \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \cdot \nabla \pi \right) \right. \\ &\quad \left. - \frac{1}{\pi (\delta \sigma)_l} \left[(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right) \right] \right\}. \end{aligned} \quad (9.152)$$

Using (9.127) in the form

$$\pi(\sigma\alpha)_l - \phi_l = -\left[\frac{\delta(\sigma\phi)}{\delta\sigma}\right]_l \quad (9.153)$$

we can rewrite this as

$$\begin{aligned} -\pi \mathbf{V}_l \bullet [\nabla \phi_l + (\sigma\alpha)_l \nabla \pi] &= -\nabla \bullet (\pi \mathbf{V}_l \phi_l) - \left\{ \frac{\delta \left[\left(\pi \dot{\sigma} + \sigma \frac{\partial \pi}{\partial t} \right) \phi \right]}{\delta \sigma} \right\}_l \\ &= -\pi \left\{ (\sigma\alpha)_l \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \bullet \nabla \pi \right) - \left[\frac{(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right)}{\pi(\delta\sigma)_l} \right] \right\}. \end{aligned} \quad (9.154)$$

By comparing with (9.146), we infer that

$$\begin{aligned} \pi(\omega\alpha)_l &= \pi(\sigma\alpha)_l \left(\frac{\partial \pi}{\partial t} + \mathbf{V}_l \bullet \nabla \pi \right) \\ &\quad - \left[\frac{(\pi \dot{\sigma})_{l+\frac{1}{2}} \left(\phi_{l+\frac{1}{2}} - \phi_l \right) + (\pi \dot{\sigma})_{l-\frac{1}{2}} \left(\phi_l - \phi_{l-\frac{1}{2}} \right)}{(\delta\sigma)_l} \right]. \end{aligned} \quad (9.155)$$

We have now reached the crux of the problem. In order to ensure total energy conservation, the form of $p_S(\omega\alpha)_l$ given by (9.155) must match that given by (9.140). In order for this to happen, we need the following conditions to be satisfied:

$$(\sigma\alpha)_l = c_p \frac{T_l \partial P_l}{P_l \partial \pi}, \quad (9.156)$$

$$\phi_l - \phi_{l+\frac{1}{2}} = c_p \left(T_{l+\frac{1}{2}} - P_l \theta_{l+\frac{1}{2}} \right), \quad (9.157)$$

$$\phi_{l-\frac{1}{2}} - \phi_l = c_p \left(P_l \theta_{l-\frac{1}{2}} - T_{l-\frac{1}{2}} \right). \quad (9.158)$$

Eq. (9.156) gives an expression for $(\sigma\alpha)_l$. We already had one, though, in Eq. (9.127). Requiring that these two formulae agree, we obtain

$$\phi_l - \left[\frac{\delta(\sigma\phi)}{\delta\sigma} \right]_l = c_p \pi \frac{T_l \partial P_l}{P_l \partial \pi} . \quad (9.159)$$

This is a finite-difference form of the hydrostatic equation.

With the use of Eq. (9.134), Eqs. (9.157)-(9.158) can be rewritten as

$$\left(c_p T_{l+\frac{1}{2}} + \phi_{l+\frac{1}{2}} \right) - (c_p T_l + \phi_l) = P_l c_p \left(\theta_{l+\frac{1}{2}} - \theta_l \right), \quad (9.160)$$

and

$$(c_p T_l + \phi_l) - \left(c_p T_{l-\frac{1}{2}} + \phi_{l-\frac{1}{2}} \right) = P_l c_p \left(\theta_l - \theta_{l-\frac{1}{2}} \right), \quad (9.161)$$

respectively. These are also finite-difference analogs of the hydrostatic equation. The subscripts in these equations are arbitrary. Add one to each subscript in (9.161), and add the result to (9.160). This yields

$$\phi_l - \phi_{l+1} = c_p (P_{l+1} - P_l) \theta_{l+\frac{1}{2}} . \quad (9.162)$$

If the forms of P_l and $\theta_{l+\frac{1}{2}}$ are specified, we can use (9.162) to integrate the hydrostatic equation upward from level $l+1$ to level l .

It is still necessary, however, to determine the value of ϕ_L , i.e., the layer-center geopotential for the lowest model layer. This can be done by first summing $(\delta\sigma)_l$ times (9.159) over all layers:

$$\sum_{l=1}^L \phi_l (\delta\sigma)_l - \phi_S = \sum_{l=1}^L \pi c_p \frac{T_l \partial P_l}{P_l \partial \pi} (\delta\sigma)_l . \quad (9.163)$$

But

$$\begin{aligned}
\sum_{l=1}^L \phi_l (\delta\sigma)_l &= \sum_{l=1}^L \phi_l \left(\sigma_{l+\frac{1}{2}} - \sigma_{l-\frac{1}{2}} \right) \\
&= \phi_L + \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} (\phi_l - \phi_{l+1}) \\
&= \phi_L + \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} c_p (P_{l+1} - P_l) \theta_{l+\frac{1}{2}} .
\end{aligned} \tag{9.164}$$

This is an identity. We can therefore write

$$\phi_L = \phi_S + \sum_{l=1}^L \pi c_p \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} (\delta\sigma)_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} c_p (P_{l+1} - P_l) \theta_{l+\frac{1}{2}} . \tag{9.165}$$

This is a bit odd, because it says that the thickness between the Earth's surface and the middle of the lowest model layer depends on θ at all levels in the entire column. From a mathematical point of view there is nothing wrong with that. In effect, all values of θ are being used to estimate the effective value of θ between the surface and level L . From a physical point of view, however, it is better for the thickness between the surface and level L to depend only on the lowest-level value of θ . Arakawa and Suarez (1983) showed that under some conditions the form (9.165) can lead to large errors in the horizontal pressure-gradient force. We return to this point below.

It remains to specify the forms of P_l and $\theta_{l+\frac{1}{2}}$. Phillips (1974) suggested

$$P_l = \left(\frac{1}{1 + \kappa} \right) \left[\frac{\delta(Pp)}{\delta p} \right]_l , \tag{9.166}$$

on the grounds that this helps to give a good simulation of vertical wave propagation. The form of $\theta_{l+\frac{1}{2}}$ can be chosen to permit conservation of some function of θ .

Arakawa and Suarez (1983) proposed a modified version of the scheme, in which (9.165) is replaced by

$$\phi_L = \phi_S + A_{L+\frac{1}{2}} c_p \theta_L , \tag{9.167}$$

where $A_{L+\frac{1}{2}}$ is a nondimensional parameter discussed below. The point of (9.167) is that only θ_L influences the thickness between the surface and the middle of the bottom layer; the

remaining values of θ do not enter. This makes the hydrostatic equation “local.” Arakawa and Suarez showed how this can be done with only minimal modifications to the derivation given above. The starting point is to replace (9.162) by

$$\phi_l - \phi_{l+1} = c_p \left(A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1} \right), \quad (9.168)$$

and

where, again, $A_{l+\frac{1}{2}}$ and $B_{l+\frac{1}{2}}$ are to be determined. Substituting (9.168) and (9.159) into the identity

$$\phi_L - \phi_S = \sum_{l=1}^L [\phi \delta \sigma - \delta(\sigma \phi)]_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} (\phi_l - \phi_{l+1}), \quad (9.169)$$

we obtain

$$\phi_L - \phi_S = \sum_{l=1}^L c_p \pi \frac{T_l}{P_l} \frac{\partial P_l}{\partial \pi} (\delta \sigma)_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} c_p \left(A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1} \right). \quad (9.170)$$

With the use of (9.134), we can write this as

$$\phi_L - \phi_S = \sum_{l=1}^L c_p \pi \theta_l \frac{\partial P_l}{\partial \pi} (\delta \sigma)_l - \sum_{l=1}^{L-1} \sigma_{l+\frac{1}{2}} c_p \left(A_{l+\frac{1}{2}} \theta_l + B_{l+\frac{1}{2}} \theta_{l+1} \right). \quad (9.171)$$

Every term on the right-hand-side of (9.171) involves a layer-center value of θ . We “collect terms” around individual values of θ_l and force the coefficients to vanish for $l < L$. This gives

$$\pi \frac{\partial P_l}{\partial \pi} (\delta \sigma)_l = \sigma_{l+\frac{1}{2}} A_{l+\frac{1}{2}} + \sigma_{l-\frac{1}{2}} B_{l+\frac{1}{2}}. \quad (9.172)$$

With the use of (9.172), (9.171) reduces to

$$\phi_L - \phi_S = \left[\pi \frac{\partial P_L}{\partial \pi} (\delta \sigma)_L - \sigma_{L-\frac{1}{2}} B_{L-\frac{1}{2}} \right] c_p \theta_L, \quad (9.173)$$

which has the form of (9.167).

9.9 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 quasi-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, while vertical resolutions are increasing; the era of hundred-layer models appears to be upon us.

CHAPTER 10**Aliasing instability**

Copyright 2004 David A. Randall

10.1 Aliasing error

Suppose that we have a wave given by the solid line in Fig. 10.1. The wave is plotted

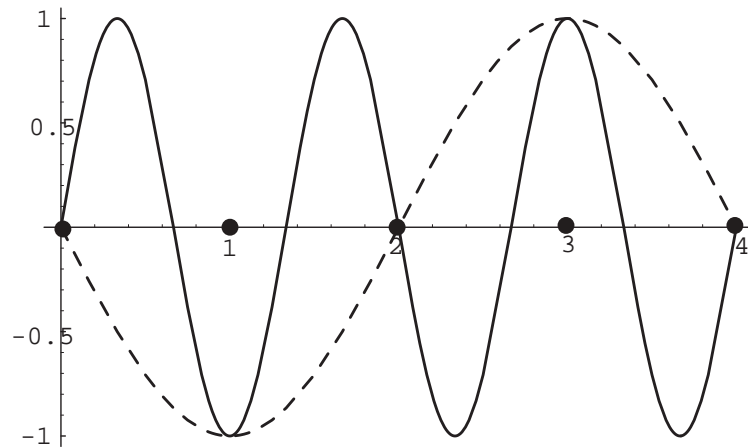


Figure 10.1: An example of aliasing error. Distance along the horizontal axis is measured in units of Δx . The wave given by the solid line has a wave length of $(4/3)\Delta x$. This is shorter than $2\Delta x$, and so the wave cannot be represented on the grid. Instead, the grid “sees” a wave of wavelength $4\Delta x$, as indicated by the dashed line. Note that the $4\Delta x$ -wave is “upside-down.”

as a continuous function of x , but at the same time we suppose that 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 $(4/3)\Delta x$. 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. 10.1, 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.*

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. 10.2 is an example, from real life. The blue curve in

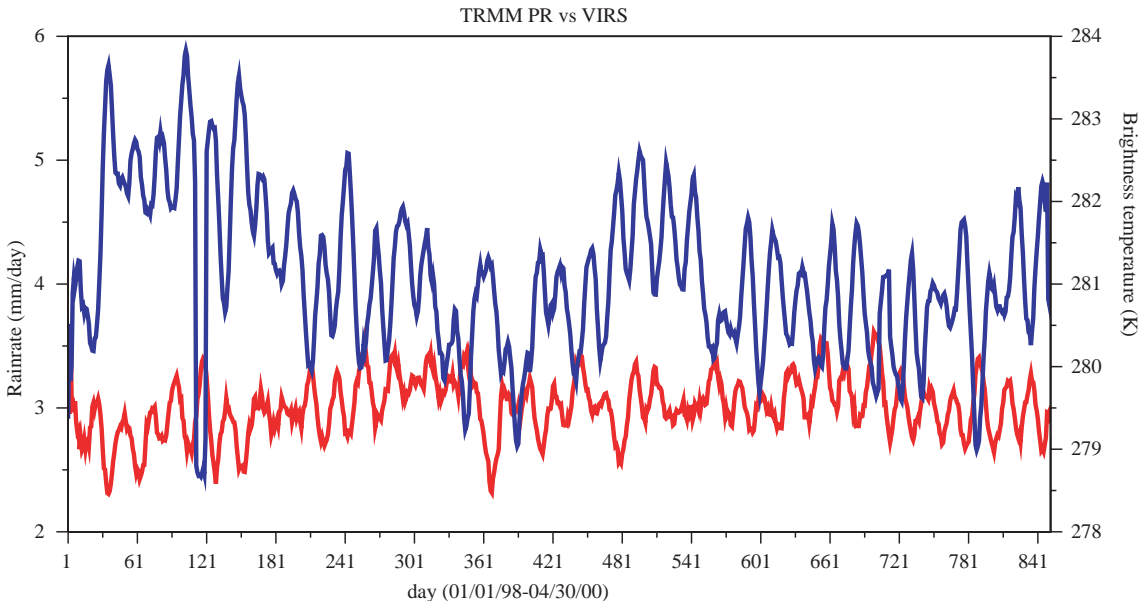


Figure 10.2: An example of aliasing in the analysis of observations. The blue curve shows the precipitation rate, averaged over the global tropics (20 S to 20 N), and the red curve shows a the thermal radiation in the $10.8\text{ }\mu\text{m}$ band, averaged over the same region. The horizontal axis is time, and the period covered is slightly more than two years. The data were obtained from the TRMM (Tropical Rain Mapping Mission) satellite. The obvious oscillation in both curves, with a period close to 23 days, is an artifact due to aliasing. See text for further explanation.

the figure makes it appear that the precipitation rate averaged over the global tropics fluctuates with a period of 23 days and an amplitude approaching 1 mm day^{-1} . If this tropical precipitation oscillation (TPO) were real it would be one of the most amazing phenomena in atmospheric science, and its discoverer would be on the cover of *Rolling Stone*. But alas, the TPO is bogus, even though you can see it with your own eyes in Fig. 10.2, and even though the figure is based on real data, for goodness sake. 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 inadequate* to resolve the diurnal cycle.

Aliasing error is also important in modeling, when we try to solve either non-linear equations or linear equations with variable coefficients. The reason is that the product terms (or other nonlinear terms) in such equations can produce, or try to produce, waves shorter

than the grid can represent. For example, suppose that we have two modes on a one-dimensional grid, given by

$$A(x_j) = \hat{A}e^{ikj\Delta x} \text{ and } B(x_j) = \hat{B}e^{ilj\Delta x}, \quad (10.1)$$

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

$$\alpha A + \beta B, \quad (10.2)$$

where α and β are spatially constant coefficients, then no “new” waves are generated; k and l continue to be the only wave numbers present. In contrast, if we multiply A and B together, then we generate the new wave number, $k + l$:

$$AB = \hat{A}\hat{B}e^{i(k+l)j\Delta x}. \quad (10.3)$$

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.*

Because the shortest wavelength that the grid can represent is $L = 2\Delta x$, the maximum representable wave number is $k_{max} \equiv \frac{\pi}{\Delta x}$. What happens when a wave with $k > k_{max}$ is produced, e.g. through nonlinear interactions? Since $2k_{max}\Delta x = 2\pi$, we can assume that $2k_{max} > k > k_{max}$. (A wave with $k > 2k_{max}$ “folds back.”) We can write the expression $\sin(kj\Delta x)$ as

$$\begin{aligned} \sin(kj\Delta x) &= \sin[(2k_{max} - 2k_{max} + k)j\Delta x] \\ &= \sin[2\pi j - (2k_{max} - k)j\Delta x] \\ &= \sin[-(2k_{max} - k)j\Delta x] \\ &= \sin(k^*j\Delta x). \end{aligned} \quad (10.4)$$

where $k^* \equiv -(2k_{max} - k)$. (Question: What happens if $k < k_{max}$?) Similarly,

$$\cos[k(j\Delta x)] = \cos[k^*(j\Delta x)]. \quad (10.5)$$

This shows that the wave of wave number $k > k_{max}$ is interpreted (or misinterpreted) by the

grid as a wave of wave number $k^* \equiv -(2k_{max} - k)$. The minus sign means that the phase change per Δx is reversed, or “backwards”. For $k > k_{max}$ we get $k^* < k_{max}$. For $k = k_{max}$, we get $k^* = k$.

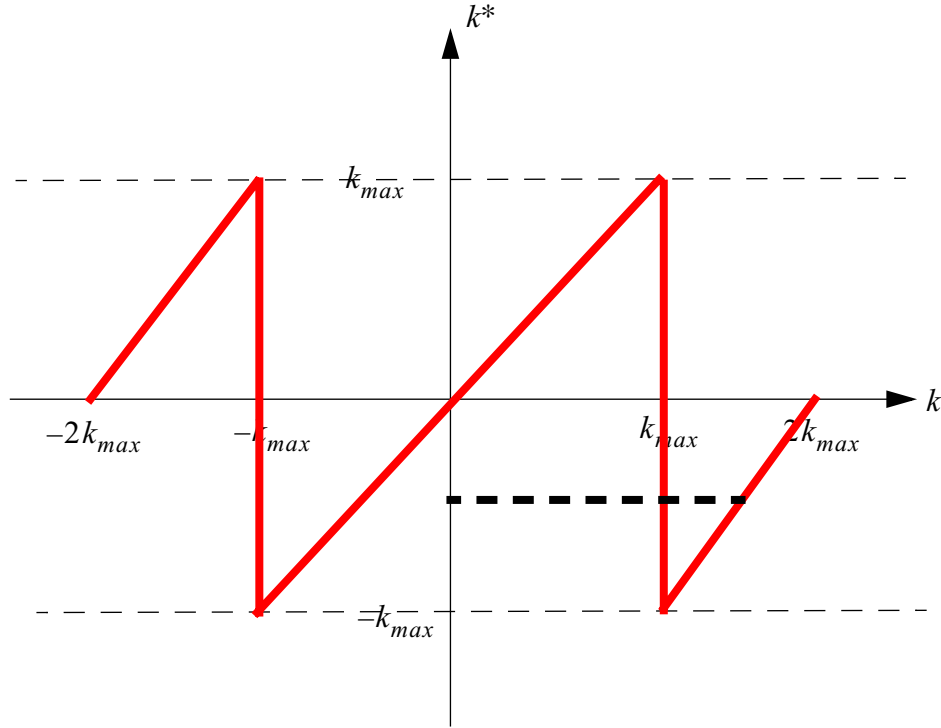


Figure 10.3: The red line is a plot of k^* versus k . The dashed black line connects $k = \frac{3}{2} \frac{\pi}{\Delta x}$ with $k^* = \frac{\pi}{2\Delta x}$, corresponding to the example of Fig. 10.1.

In the example of Fig. 10.1, $L = \frac{4}{3}\Delta x$ so $k = \frac{2\pi}{L} = \frac{2\pi}{4\Delta x} = \frac{1}{2} \frac{\pi}{\Delta x}$. Therefore $k^* \equiv 2k_{max} - k = \frac{2\pi}{\Delta x} - \frac{1}{2} \frac{\pi}{\Delta x} = \frac{3}{2} \frac{\pi}{\Delta x}$, which means that $L^* = 4\Delta x$, as we have already surmised by inspection of Fig. 10.1.

For $k < k_{max}$, the phase change, as j increases by one, is less than π . This is shown in Fig. 10.4 a. For $k > k_{max}$, the phase change as j increases by one is greater than π . This is shown in Fig. 10.4 b. The dot in the figure appears to move clockwise, i.e. “backwards.” This is a manifestation of aliasing that is familiar from the movies. It also explains why the minus sign appears in Eq. (10.4).

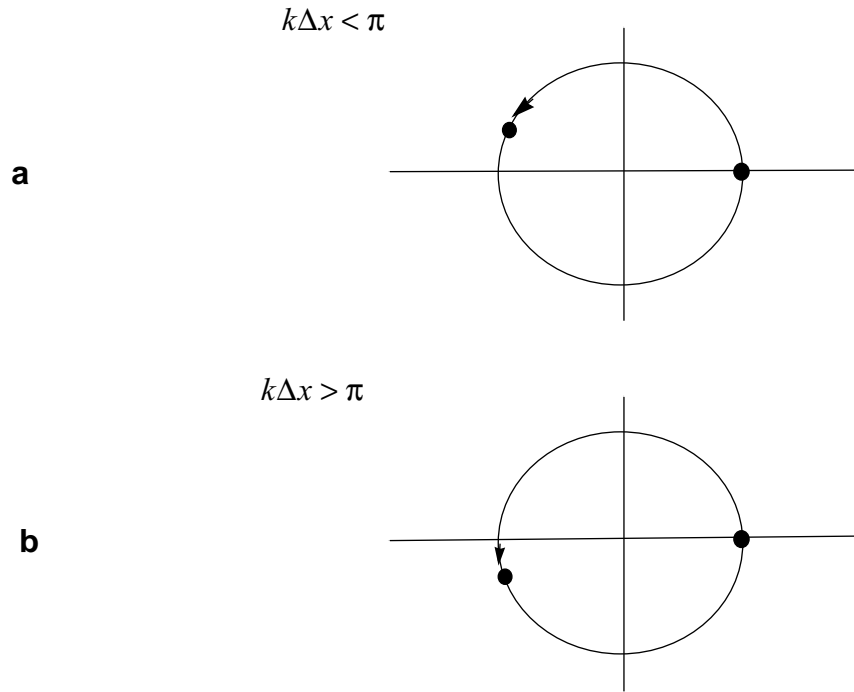


Figure 10.4: The phase change per grid point for: a) $k\Delta x < \pi$, and b) $k\Delta x > \pi$.

Aliasing error is important in part because it is the root cause of what is often called “nonlinear computational instability.” This instability occurs with nonlinear equations, but as explained below it can also occur with linear equations that have spatially variable coefficients. A better name for the instability would be “aliasing instability.” An example is presented in the next section.

10.2 Advection by a variable, non-divergent current

Suppose that an arbitrary variable q is advected in two dimensions on a plane, so that

$$\frac{\partial q}{\partial t} + \mathbf{v} \cdot \nabla q = 0, \quad (10.6)$$

where the flow is assumed to be non-divergent, i.e.

$$\nabla \cdot \mathbf{v} = 0. \quad (10.7)$$

Two-dimensional non-divergent flow is a not-too-drastic idealization of the large-scale circulation of the atmosphere. In view of (10.7), we can describe \mathbf{v} in terms of a stream function ψ , such that

$$\mathbf{v} = \mathbf{k} \times \nabla \psi. \quad (10.8)$$

Substituting (10.6) into (10.6), we get

$$\frac{\partial q}{\partial t} + (\mathbf{k} \times \nabla \psi) \cdot \nabla q = 0. \quad (10.9)$$

Using the vector identity

$$(\mathbf{V}_1 \times \mathbf{V}_2) \cdot \mathbf{V}_3 = \mathbf{V}_2 \cdot (\mathbf{V}_3 \times \mathbf{V}_1), \quad (10.10)$$

which holds for any three vectors $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$, we set $\mathbf{V}_2 \equiv \mathbf{k}$, $\mathbf{V}_1 \equiv \nabla \psi$, and $\mathbf{V}_3 \equiv \nabla q$, to obtain

$$(\nabla \psi \times \mathbf{k}) \cdot \nabla q = \mathbf{k} \cdot (\nabla q \times \nabla \psi). \quad (10.11)$$

With the use of (10.11), we can re-write (10.6) as

$$\frac{\partial q}{\partial t} + J(\psi, q) = 0, \quad (10.12)$$

or alternatively as

$$\frac{\partial q}{\partial t} = J(q, \psi). \quad (10.13)$$

Here J is the Jacobian operator, which is defined by

$$\begin{aligned} J(p, q) &\equiv \mathbf{k} \cdot (\nabla p \times \nabla q) \\ &= -\mathbf{k} \cdot \nabla \times (q \nabla p) \quad , \\ &= \mathbf{k} \cdot \nabla \times (p \nabla q) \quad , \end{aligned} \quad (10.14)$$

for arbitrary p and q . Note that

$$J(p, q) = -J(q, p) \quad , \quad (10.15)$$

which can be deduced from (10.14), and this has been used to pass from (10.12) to (10.13).

In Cartesian coordinates, we can write $J(q, \psi)$ in the following three alternative forms:

$$J(p, \psi) = \frac{\partial p}{\partial x} \frac{\partial q}{\partial y} - \frac{\partial p}{\partial y} \frac{\partial q}{\partial x} \quad (10.16)$$

$$= \frac{\partial}{\partial y} \left(q \frac{\partial p}{\partial x} \right) - \frac{\partial}{\partial x} \left(q \frac{\partial p}{\partial y} \right) \quad (10.17)$$

$$= \frac{\partial}{\partial x} \left(p \frac{\partial q}{\partial y} \right) - \frac{\partial}{\partial y} \left(p \frac{\partial q}{\partial x} \right). \quad (10.18)$$

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 p or q is constant. You should be able to prove the following:

$$\overline{J(p, q)} = 0, \quad (10.19)$$

$$\overline{pJ(p, q)} = 0, \quad (10.20)$$

$$\overline{qJ(p, q)} = 0. \quad (10.21)$$

Multiplying both sides of (10.13) by q , we obtain

$$\frac{1}{2} \frac{\partial}{\partial t} q^2 = qJ(q, \psi) = J\left(\frac{1}{2} q^2, \psi\right). \quad (10.22)$$

Integrating over the entire area, we see that

$$\int J\left(\frac{1}{2} q^2, \psi\right) ds = -\int \mathbf{v} \cdot \nabla \frac{1}{2} q^2 ds = -\int \nabla \cdot \left(\mathbf{v} \frac{1}{2} q^2 \right) ds = 0, \quad (10.23)$$

if the domain is surrounded by a rigid boundary where the normal component of \mathbf{v} is zero, or if the domain is periodic.

When the stream function ψ is a prescribed function of the spatial coordinates, (10.13) is linear, although it has variable coefficients. As already mentioned, what is often called “non-linear” instability is actually a type of instability that can occur in the numerical integration of a linear equation with variable coefficients, as well as in the numerical integration of nonlinear equations. What this instability really amounts to is a spurious growth of waves due in part to the aliasing error arising from the multiplication of the finite difference analogs of *any* two spatially varying quantities.

To illustrate the problem, we begin by writing down a differential-difference version of (10.13), 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

$$\frac{dq_{i,j}}{dt} = [J_1(q, \psi)]_{i,j} \quad (10.24)$$

where

$$[J_1(q, \psi)]_{i,j} \equiv \frac{1}{4d^2} [(q_{i+1,j} - q_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) - (q_{i,j+1} - q_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j})] . \quad (10.25)$$

You should confirm for yourself that (10.25) really is a finite-difference approximation to (10.13). In fact, it is modeled after (10.16). Later we are going to discuss several other finite-difference approximations for the Jacobian. The particular approximation given in (10.25) is called J_1 . It will come up again in the later discussion.

Now we work through a simple example of aliasing instability, which was invented by Phillips (1959; also see Lilly, 1965). We combine (10.24) and (10.25) to obtain

$$\frac{dq_{i,j}}{dt} \equiv \frac{1}{4d^2} [(q_{i+1,j} - q_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) - (q_{i,j+1} - q_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j})] . \quad (10.26)$$

Assume that the solution $q_{i,j}(t)$ is of the form

$$q_{i,j}(t) = \left[C(t) \cos \frac{\pi i}{2} + S(t) \sin \frac{\pi i}{2} \right] \sin \frac{2\pi j}{3} . \quad (10.27)$$

The use of such assumption may appear strange; it will be justified later. For all t , let $\psi_{i,j}$ be prescribed as

$$\psi_{i,j} = U \cos(\pi i) \sin\left(\frac{2\pi j}{3}\right) . \quad (10.28)$$

In (10.28), we are prescribing a time-independent but *spatially variable* advecting current. Earlier in the course we often prescribed the advecting current, but it was always spatially uniform. Because $\psi_{i,j}$ is prescribed, the model that we are considering here is linear. The forms of $q_{i,j}$ and $\psi_{i,j}$ given by (10.27) and (10.28) are plotted in Fig. 10.5.

Because (10.28) specifies $\psi_{i,j}$ to have a wavelength of $2d$ in the x -direction, we can simplify (10.25) to

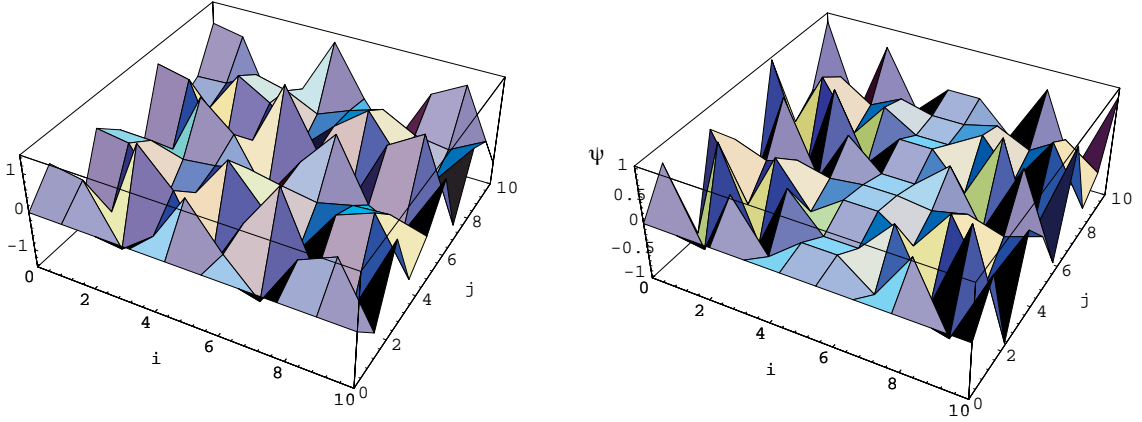


Figure 10.5: Plots of the functions q and ψ given by (10.27) and (10.28), respectively. For plotting purposes, we have used $C = S = U = 1$. The functions have been evaluated only for integer values of i and j , which gives them a jagged appearance. Nevertheless it is fair to say that they are rather ugly. This is the sort of thing that can appear in your simulations as a result of aliasing instability.

$$\frac{\partial q_{i,j}}{\partial t} = \frac{1}{4d^2} (q_{i+1,j} - q_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}). \quad (10.29)$$

From (10.27), we see that

$$\begin{aligned} & q_{i+1,j} - q_{i-1,j} \\ &= \left\{ C \left[\cos \frac{\pi(i+1)}{2} - \cos \frac{\pi(i-1)}{2} \right] + S \left[\sin \frac{\pi(i+1)}{2} - \sin \frac{\pi(i-1)}{2} \right] \right\} \sin \frac{2\pi j}{3} \\ &= 2 \left(-C \sin \frac{\pi i}{2} + S \cos \frac{\pi i}{2} \right) \sin \frac{2\pi j}{3}. \end{aligned} \quad (10.30)$$

Here we have used some trigonometric identities. Similarly, we can show that

$$\psi_{i,j+1} - \psi_{i,j-1} = U \cos \pi i \left(2 \cos \frac{2\pi j}{3} \sin \frac{2\pi}{3} \right) = \sqrt{3} U \cos \pi i \left(\cos \frac{2\pi j}{3} \right). \quad (10.31)$$

As already mentioned, (10.31) holds for all t .

The product of (10.30) and (10.31) gives the right-hand side of (10.29), which can be written (again with the use of trigonometric identities) as

$$\begin{aligned}
\frac{dq_{i,j}}{dt} &= \frac{\sqrt{3}}{4d^2} U \left[-C \left(\sin \frac{3\pi i}{2} - \sin \frac{\pi i}{2} \right) + S \left(\cos \frac{3\pi i}{2} + \cos \frac{\pi i}{2} \right) \right] \sin \frac{4\pi j}{3} \\
&= \frac{\sqrt{3}}{4d^2} \left(C \sin \frac{\pi i}{2} + S \cos \frac{\pi i}{2} \right) \sin \frac{4\pi j}{3}.
\end{aligned} \tag{10.32}$$

Note that

$$\frac{4\pi j}{3} = ly = l(jd). \tag{10.33}$$

Therefore,

$$l = \frac{4\pi}{3d}. \tag{10.34}$$

This shows that the product on the right-hand side of (10.29) has produced a wave number in the y -direction of $l = \frac{4\pi}{3d} > l_{max} = \frac{\pi}{d}$, i.e., a wave too short to be represented on the grid. This wave will, according to our previous results, be interpreted by the grid as having wave number $-(2l_{max} - l) = -\left(\frac{2\pi}{3d}\right)$. Therefore (10.32) can be re-written as

$$\frac{dq_{i,j}}{dt} = -\left(\frac{\sqrt{3}U}{4d^2}\right) \left(C \sin \frac{\pi i}{2} + S \cos \frac{\pi i}{2} \right) \sin \frac{2\pi j}{3}. \tag{10.35}$$

Rewriting (10.32) as (10.35) is obviously a key step in this discussion, because it is where aliasing enters. In doing the problem algebraically, we have to put in the aliasing “by hand.”

According to (10.35) the spatial form of $\frac{dq_{i,j}}{dt}$ agrees with the assumed form of $q_{i,j}$, given by (10.27). In other words, the spatial shape of $q_{i,j}$ does not change with time. *This justifies our assumption (10.27).* In order to recognize that the spatial shape of $q_{i,j}$ does not change with time, we had to take into account that there will be aliasing.

If we now simply differentiate (10.27) with respect to time, and substitute the result into the left-hand side of (10.35), we get

$$\frac{dC}{dt} \cos \frac{\pi i}{2} \sin \frac{2\pi j}{3} + \frac{dS}{dt} \sin \frac{\pi i}{2} \sin \frac{2\pi j}{3} = -\frac{\sqrt{3}}{4d^2} U \left(C \sin \frac{\pi i}{2} + S \cos \frac{\pi i}{2} \right) \sin \frac{2\pi j}{3}. \tag{10.36}$$

Note that time derivatives of C and S appear on the left-hand side of (10.36). Using the

linear independence of the sine and cosine functions, we see from (10.36) that

$$\frac{dC}{dt} = -\frac{\sqrt{3}}{4d^2}US, \quad \frac{dS}{dt} = -\frac{\sqrt{3}}{4d^2}UC. \quad (10.37)$$

From (10.37), it follows that

$$\frac{d^2C}{dt^2} = \sigma^2 C, \quad \text{and} \quad \frac{d^2S}{dt^2} = \sigma^2 S, \quad (10.38)$$

where $\sigma \equiv \frac{\sqrt{3}U}{4d^2}$. According to (10.38), C and S will grow exponentially. This demonstrates that the finite-difference scheme is unstable. The unstable modes will have the form given by (10.27).

Fig. 10.6 summarizes the mechanism of this 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, etc. Note, however, that *if the numerical scheme conserved energy, the total amount of energy*

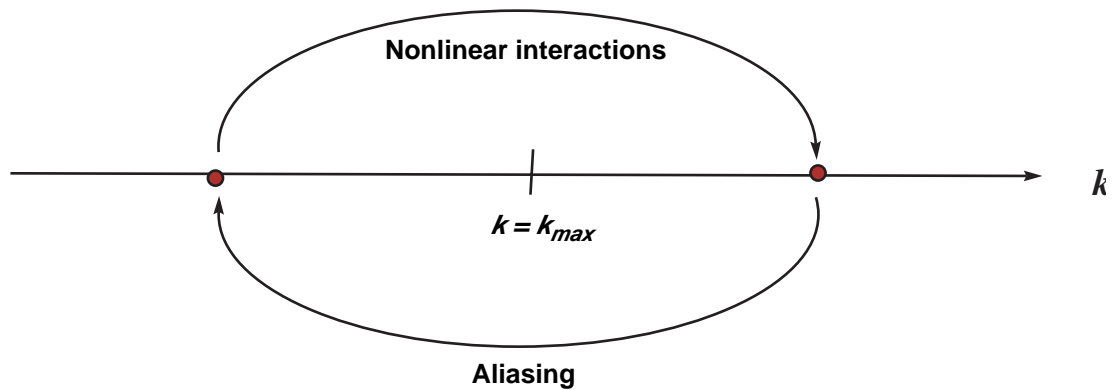


Figure 10.6: Schematic illustration of the mechanism of aliasing instability. Nonlinear interactions feed energy into scales too small to be represented on the grid, and this energy folds back through aliasing into scales that *can* be represented. The process feeds on itself. This can cause the total amount of energy to increase, unless the scheme is energy conserving.

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, we used J_1 . Later we demonstrate that J_1 does not conserve kinetic energy. Some other finite-difference Jacobians do conserve kinetic energy. Further discussion is given later.

Some further general insight into this type of instability (the above example being a

contrived special case) can be obtained by investigating the truncation error of the expression on the right side of (10.25). This can be expressed as

$$\frac{\partial q}{\partial t} = J_1(q, \psi) = J(q, \psi) + \frac{d^2}{6} \left[\frac{\partial q \partial^3 \psi}{\partial x \partial y^3} - \frac{\partial q \partial^3 \psi}{\partial y \partial x^3} + \frac{\partial^3 q \partial \psi}{\partial x^3 \partial y} - \frac{\partial^3 q \partial \psi}{\partial y^3 \partial x} \right] + O(d^4). \quad (10.39)$$

Note the second-order accuracy. Through repeated integration by parts, it can be shown (after a page or so of algebra) that the second-order part of the truncation error in (10.39) is given by

$$\frac{d^2}{6} \int q \left[\frac{\partial q \partial^3 \psi}{\partial x \partial y^3} - \frac{\partial q \partial^3 \psi}{\partial y \partial x^3} + \frac{\partial^3 q \partial \psi}{\partial x^3 \partial y} - \frac{\partial^3 q \partial \psi}{\partial y^3 \partial x} \right] ds = \frac{d^2}{4} \int \frac{\partial^2 \psi}{\partial x \partial y} \left[\left(\frac{\partial q}{\partial x} \right)^2 - \left(\frac{\partial q}{\partial y} \right)^2 \right] ds, \quad (10.40)$$

if q and its derivatives vanish at the boundary or if the domain is periodic. Multiplying (10.39) by q , integrating over the whole domain, and making use of (10.22), (10.23) and (10.40), we find that when we use J_1 ,

$$\frac{1}{2} \frac{\partial}{\partial t} \int q^2 ds = \frac{d^2}{4} \int \frac{\partial^2 \psi}{\partial x \partial y} \left[\left(\frac{\partial q}{\partial x} \right)^2 - \left(\frac{\partial q}{\partial y} \right)^2 \right] ds + \int O(d^4) ds. \quad (10.41)$$

This means that, if $\frac{\partial^2 \psi}{\partial x \partial y} > 0$, 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. Now look at Fig. 10.7. In the figure, the streamlines are given

such that $\psi_1 < \psi_2 < \psi_3$, so that $\partial \psi / \partial y < 0$, and $\frac{\partial^2 \psi}{\partial x \partial y} > 0$. This resembles the “exit” region of the jet stream. [Note: The stream function sketched in Fig. 10.7 does *not* correspond to (10.27).]

In fact, the solution of the differential-difference equation tends to prefer a positive value of the integrand of the right-hand side of (10.41), as illustrated schematically in Fig. 10.7.

Notice that at t_2 , $\frac{\partial q}{\partial x}$ becomes greater than it was at t_1 , and the reverse is true for $\frac{\partial q}{\partial y}$.

Therefore, although at t_1 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_2 it has

become positive. From (10.41), it can be seen that the area-integrated q^2 tends to increase with time, whereas it is invariant in the differential case.

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.

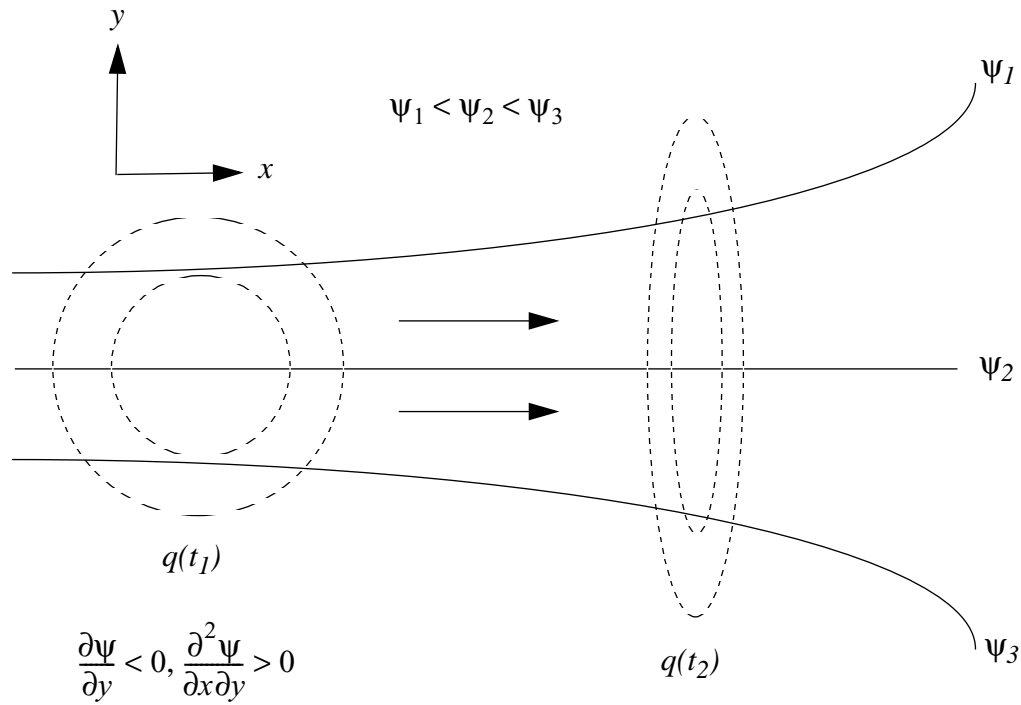


Figure 10.7: Sketch illustrating the mechanism of aliasing instability.

A number of methods have been proposed to prevent or control aliasing instability. One method is to prevent aliasing. Aliasing error can actually be eliminated in a spectral model, at least for models that contain only “quadratic” aliasing, i.e., aliasing that arises from the multiplication of two spatially varying fields; this will be discussed later.

Phillips (1959) suggested that aliasing instability can be prevented if a Fourier analysis of the predicted fields is made after each time step, and all waves of wave number $k > \frac{k_{max}}{2}$ 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 $\frac{k_{max}}{2}$ (his maximum wave number) and thus any wave generated by quadratic nonlinearities would have a wave number of at most k_{max} . This method is strongly dissipative, however.

Others have suggested that use of a dissipative 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 $\frac{c\Delta t}{\Delta x} \rightarrow 0$.

A third approach is to use a sign-preserving advection scheme, as discussed in Chapter 4, and advocated by Smolarkeiwicz (1991).

A fourth approach is to use space-differencing schemes for the advection terms that designed to conserve the square of the advected quantity. The “energy approach” to checking stability, discussed in Chapter 2, ensures that such schemes are computationally stable. This approach has the advantage that stability is ensured simply by mimicking a property of the exact equations.

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 the nature of two-dimensional nondivergent flows.

10.3 Fjortoft's Theorem

In the absence of viscosity, *vorticity and enstrophy are both conserved in two-dimensional nondivergent flows*. The frictionless momentum equation for shallow water,

$$\frac{\partial \mathbf{v}}{\partial t} = -(\mathbf{v} \bullet \nabla) \mathbf{v} - f \mathbf{k} \times \mathbf{v} , \quad (10.42)$$

where

$$f \equiv 2\Omega \sin \varphi \quad (10.43)$$

is the coriolis parameter, Ω is the angular velocity of the Earth's rotation, and φ is latitude. By taking the curl of (10.45) we can obtain the vorticity equation

$$\frac{\partial \zeta}{\partial t} = -\mathbf{v} \bullet \nabla (\zeta + f) - (\zeta + f) \nabla \bullet \mathbf{v} . \quad (10.44)$$

When the flow is nondivergent, so that (10.7) is satisfied, the vorticity equation reduces to

$$\frac{\partial \zeta}{\partial t} = -\mathbf{v} \bullet \nabla (\zeta + f) , \quad (10.45)$$

Since f is independent of time, we can write (10.45) as

$$\frac{\partial}{\partial t} (\zeta + f) = -\mathbf{v} \bullet \nabla (\zeta + f) . \quad (10.46)$$

The says that the absolute vorticity is simply advected by the mean flow. We also see that only the sum $(\zeta + f)$ matters for the vorticity equation; henceforth we just replace $(\zeta + f)$

by ζ , for simplicity.

Using Eq. (10.8), we can show that the vorticity and the stream function are related by

$$\zeta \equiv \mathbf{k} \bullet (\nabla \times \mathbf{v}) = \nabla^2 \psi. \quad (10.47)$$

Eq. (10.45) can be rewritten as

$$\frac{\partial \zeta}{\partial t} = -\nabla \bullet (\mathbf{v} \zeta), \quad (10.48)$$

or, alternatively, as

$$\frac{\partial \zeta}{\partial t} = J(\zeta, \psi). \quad (10.49)$$

From (10.48) we see that the domain-averaged vorticity is conserved:

$$\frac{d}{dt} \bar{\zeta} = \overline{\frac{\partial \zeta}{\partial t}} = 0. \quad (10.50)$$

By combining (10.49) and (10.21), we obtain

$$\overline{\zeta \frac{\partial \zeta}{\partial t}} = 0, \quad (10.51)$$

from which it follows that the domain-average of the enstrophy is also conserved:

$$\frac{d}{dt} \left(\frac{1}{2} \bar{\zeta}^2 \right) = 0. \quad (10.52)$$

Similarly, from (10.49) and (10.20) we find that

$$\overline{\psi \frac{\partial \zeta}{\partial t}} = 0. \quad (10.53)$$

To see what this implies, substitute (10.47) into (10.53), to obtain

$$\overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} = 0. \quad (10.54)$$

This is equivalent to

$$\begin{aligned}
0 &= \overline{\psi \frac{\partial}{\partial t} \nabla^2 \psi} \\
&= \overline{\psi \frac{\partial}{\partial t} [\nabla \cdot (\nabla \psi)]} \\
&= \overline{\psi \nabla \cdot \frac{\partial}{\partial t} \nabla \psi} \\
&= \overline{\nabla \cdot \left(\psi \frac{\partial}{\partial t} \nabla \psi \right)} - \overline{\nabla \psi \cdot \frac{\partial}{\partial t} \nabla \psi} \\
&= - \overline{\frac{\partial}{\partial t} \left(\frac{1}{2} |\nabla \psi|^2 \right)}.
\end{aligned} \tag{10.55}$$

Eq. (10.55) is a statement of kinetic energy conservation. We conclude that (10.53) implies kinetic energy conservation. In fact, we can show that

$$\bar{K} = \bar{\psi \zeta}. \tag{10.56}$$

Since both kinetic energy and enstrophy are conserved in frictionless two-dimensional flows, their ratio is also conserved, and has the dimensions of a length squared:

$$\frac{\text{energy}}{\text{enstrophy}} \sim \frac{L^2 t^{-2}}{t^{-2}} = L^2. \tag{10.57}$$

This length can be interpreted as the typical scale of energy-containing eddies, and (10.57) states that it is invariant. The implication is that energy does not cascade in frictionless two-dimensional flows; it “stays where it is” in wave number space.

The exchanges of energy and enstrophy among different scales in two-dimensional turbulence were studied by Fjortoft (1953), who obtained some very fundamental and famous results, which can be summarized in a simplified way as follows. Consider three equally

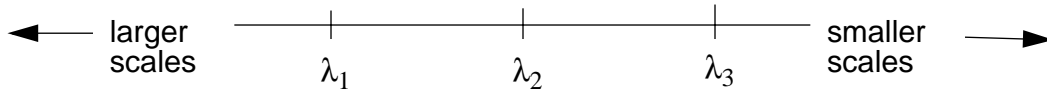


Figure 10.8: Diagram used in the explanation of Fjortoft’s (1953) analysis of the exchanges of energy and enstrophy among differing scales in two-dimensional motion.

spaced wave numbers, as shown in Fig. 10.8. By “equally spaced” we mean that

$$\lambda_2 - \lambda_1 = \lambda_3 - \lambda_2 = \Delta\lambda. \tag{10.58}$$

The enstrophy, E , is

$$E = E_1 + E_2 + E_3 , \quad (10.59)$$

and the kinetic energy is

$$K = K_1 + K_2 + K_3 . \quad (10.60)$$

It can be shown that

$$E_n = \lambda_n^2 K_n , \quad (10.61)$$

where λ_n is a wave number, and the subscript n denotes a particular Fourier component. Suppose that kinetic energy is redistributed, i.e.

$$K_n \rightarrow K_n + \delta K_n , \quad (10.62)$$

such that

$$\sum \delta K_n = 0 , \quad (10.63)$$

$$\sum \delta E_n = 0 , \quad (10.64)$$

and note from (10.61) that

$$\delta E_n = \lambda_n^2 \delta K_n . \quad (10.65)$$

It follows that

$$\delta K_1 + \delta K_3 = -\delta K_2 , \quad (10.66)$$

$$\begin{aligned} \lambda_1^2 \delta K_1 + \lambda_3^2 \delta K_3 &= -\lambda_2^2 \delta K_2 \\ &= \lambda_2^2 (\delta K_1 + \delta K_3) . \end{aligned} \quad (10.67)$$

Collecting terms, we find that

$$\frac{\delta K_3}{\delta K_1} = \frac{\lambda_2^2 - \lambda_1^2}{\lambda_3^2 - \lambda_2^2} . \quad (10.68)$$

Using (10.58), we get

$$\frac{\delta K_3}{\delta K_1} = \frac{\lambda_2 + \lambda_1}{\lambda_3 + \lambda_2} < 1 \quad (10.69)$$

Eq. (10.69) shows that the energy transferred to higher wave numbers (δK_3) is less than the energy transferred to lower wave numbers (δK_1). This conclusion rests on both (10.63) and (10.64), 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 (10.65) and (10.69) to write

$$\begin{aligned} \frac{\delta E_3}{\delta E_1} &= \frac{\lambda_3^2}{\lambda_1^2} \left(\frac{\lambda_2 + \lambda_1}{\lambda_3 + \lambda_2} \right) \\ &= \frac{(\lambda_2 + \Delta\lambda)^2 \left(\lambda_2 - \frac{1}{2}\Delta\lambda \right)}{(\lambda_2 - \Delta\lambda)^2 \left(\lambda_2 + \frac{1}{2}\Delta\lambda \right)} > 1. \end{aligned} \quad (10.70)$$

To show that this ratio is greater than one, 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:

$$a = \frac{\lambda_2 + \Delta\lambda}{\lambda_2 - \Delta\lambda} > 1, \quad (10.71)$$

$$b = \frac{\lambda_2 - \frac{1}{2}\Delta\lambda}{\lambda_2 - \Delta\lambda} > 1, \quad (10.72)$$

$$c = \frac{\lambda_2 + \Delta\lambda}{\lambda_2 + \frac{1}{2}\Delta\lambda} > 1. \quad (10.73)$$

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.

The conclusion is that in two-dimensional turbulence, enstrophy is dissipated but kinetic energy is not.

In three-dimensions, vorticity is not conserved (nor is enstrophy) because of

stretching and twisting. Vortex stretching, in particular, 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 the viscous force.

In summary, vorticity and enstrophy are conserved in two-dimensional flow but not in three-dimensional flow. Kinetic energy is conserved under inertial processes in both two-dimensional and three-dimensional flows. Because two-dimensional flows are obliged to conserve both energy and enstrophy, they “have fewer options” than do 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 flow, 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.

10.4 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. His method has been very widely used. The following is a summary of Arakawa’s approach.

We begin by writing down a spatially discrete version of (10.49), keeping the time derivative in continuous form:

$$\begin{aligned}\sigma_i \frac{d\zeta_i}{dt} &= \sigma_i J_i(\zeta, \psi) \\ &= \sum_{\mathbf{i}'} \sum_{\mathbf{i}''} c_{\mathbf{i}, \mathbf{i}', \mathbf{i}''} \zeta_{\mathbf{i} + \mathbf{i}'} \psi_{\mathbf{i} + \mathbf{i}''} .\end{aligned}\tag{10.74}$$

Here the area of grid cell \mathbf{i} is denoted by $\sigma_{\mathbf{i}}$. 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. For instance, we could start counting from the lower left corner of the grid, with $\mathbf{i} = 1$, and take the next grid point to the right as $\mathbf{i} = 2$, and so on, until we came to the end of the bottom row with $\mathbf{i} = I$, and then pick up at the left-most grid point of the second row from the bottom with $\mathbf{i} = I + 1$, etc. These two-dimensional counters are used to minimize the number of subscripts; we could use double subscripts (i, j) in the usual way, but choose not to just to keep the notation a little easier on the eyes. Just for reference, with double subscripts (10.74) would become

$$\sigma_{i,j} \frac{d\zeta_{i,j}}{dt} = \sum_{j'} \sum_{i'} \sum_{j''} \sum_{i''} c_{i,j;i',j';i'',j''} \zeta_{i',j'} \psi_{i'',j''} .\tag{10.75}$$

The second line of (10.74) looks a little mysterious and requires some explanation. As can be seen by inspection of (10.16), the Jacobian operator $J(\zeta, \psi)$ involves derivatives of the vorticity, multiplied by derivatives of the stream function. We can anticipate, therefore,

that the form we choose for the finite-difference Jacobian at the point 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 (10.25). Such products appear in (10.74). The neighboring grid points can be specified in (10.74) by assigning appropriate values to i' and i'' . The $c_{i,i',i''}$ are suitable “interaction coefficients” involving the grid distances, etc., and their form will be chosen later. It is by appropriate choices of the $c_{i,i',i''}$ that we will construct an approximation to the Jacobian. The double sum in (10.74) essentially picks out the combinations of ζ and ψ , at neighboring grid points, that we wish to bring into our finite-difference operator.

Of course, there is nothing about the form of (10.74) that shows that it is actually a consistent finite-difference approximation to the Jacobian operator; all we can say at this point is that (10.74) has the *potential* to be a consistent finite-difference approximation to the Jacobian operator, if we choose the interaction coefficients properly. We return to this point later. We note, however, that in order to ensure conservation of vorticity under advection, we must require that our finite-difference Jacobian satisfy

$$\begin{aligned} 0 &= \sum_i \sigma_i J_i(\zeta, \psi) \\ &= \sum_i \sum_{i'} \sum_{i''} c_{i,i',i''} \zeta_{i+i'} \psi_{i+i''} . \end{aligned} \quad (10.76)$$

Another interesting and important point is that the form of (10.74) is so general that it is impossible to tell what kind of grid we are on. 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 later).

Before going on, let's consider an example. Recall that in (10.25) we have already introduced a finite-difference Jacobian called J_1 , which is defined on a square grid. We can write

$$\begin{aligned} [J_1(\zeta, \psi)]_{i,j} &\equiv \\ \frac{1}{4d^2} [(\zeta_{i+1,j} - \zeta_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) - (\zeta_{i,j+1} - \zeta_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j})] . \end{aligned} \quad (10.77)$$

Expanding, and using $\sigma_{i,j} = d^2$, we find that

$$\begin{aligned} \sigma_{i,j} [J_1(\zeta, \psi)]_{i,j} &\equiv \\ \frac{1}{4} (\zeta_{i+1,j} \psi_{i,j+1} - \zeta_{i+1,j} \psi_{i,j-1} - \zeta_{i-1,j} \psi_{i,j+1} + \zeta_{i-1,j} \psi_{i,j-1} \\ - \zeta_{i,j+1} \psi_{i+1,j} + \zeta_{i,j+1} \psi_{i-1,j} + \zeta_{i,j-1} \psi_{i+1,j} - \zeta_{i,j-1} \psi_{i-1,j}) . \end{aligned} \quad (10.78)$$

Comparing (10.78) with (10.74), we identify

$$\begin{aligned}
c_{i,j;i+1,j;i,j+1} &= \frac{1}{4}, \\
c_{i,j;i+1,j;i,j-1} &= -\frac{1}{4}, \\
c_{i,j;i-1,j;i,j+1} &= -\frac{1}{4}, \\
c_{i,j;i-1,j;i,j-1} &= \frac{1}{4}, \\
c_{i,j;i,j+1;i+1,j} &= -\frac{1}{4}, \\
c_{i,j;i,j+1;i-1,j} &= \frac{1}{4}, \\
c_{i,j;i,j-1;i+1,j} &= \frac{1}{4}, \\
c_{i,j;i,j-1;i-1,j} &= -\frac{1}{4}.
\end{aligned} \tag{10.79}$$

Look carefully at the subscripts. It should be clear that $c_{i,j;i+1,j;i,j+1}$ specifies contributions of the vorticity east of the point (i, j) and the stream function north of the point (i, j) to the time rate of change of the vorticity at the point i, j . In this simple case of a uniform square grid, the coefficients are very simple. Exactly the same formalism can be applied to much more complicated cases, however, such as nonuniform grids on a sphere.

Is (10.76) satisfied for J_1 ? Each term of the triple sum in (10.76) involves the product of a vorticity and a stream function. Each product will appear exactly twice when we form the sum. In order for (10.76) to be satisfied for arbitrary values of the vorticity and the stream function, we need the two contributions from each product to cancel in the sum, i.e., their coefficients must be equal and opposite. 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 . 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 $i+1, j+1$. See Fig. 10.9. Cancellation will occur if

$$c_{i,j;i+1,j;i,j+1} = -c_{i+1,j+1;i+1,j;i,j+1}. \tag{10.80}$$

To see whether or not this is the case, note that the value of $c_{i+1,j+1;i+1,j;i,j+1}$ must remain unchanged if we subtract one from each i subscript and one from each j subscript. In other words,

$$c_{i+1,j+1;i+1,j;i,j+1} = c_{i,j;i,j-1;i-1,j}. \tag{10.81}$$

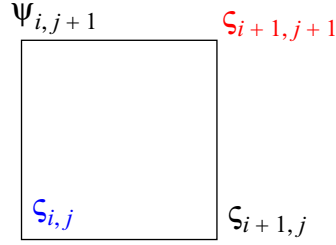


Figure 10.9: Stencil used in the discussion of vorticity conservation for J_1 . See text for details.

Therefore the requirement (10.80) is equivalent to

$$c_{i,j;i+1,j;i,j+1} = -c_{i,j;i,j-1;i-1,j}. \quad (10.82)$$

What has been accomplished by subtracting one from each subscript is that in the result, i.e., (10.82), both of the c s are associated with the time-rate of change at the point (i, j) , and so both of them are explicitly listed in (10.79). Inspection of (10.79) shows that (10.82) is indeed satisfied. Similar results apply for the remaining terms. In this way, we can prove 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 (10.51) and (10.53):

$$0 = \sum_i \sigma_i \zeta_i J_i(\zeta, \psi) = \sum_i \zeta_i \left(\sum_{\vec{i}} \sum_{\vec{i}''} c_{i,\vec{i},\vec{i}''} \zeta_{i+\vec{i}} \psi_{i+\vec{i}''} \right), \quad (10.83)$$

$$0 = \sum_i \sigma_i \psi_i J_i(\zeta, \psi) = \sum_i \psi_i \left(\sum_{\vec{i}} \sum_{\vec{i}''} c_{i,\vec{i},\vec{i}''} \zeta_{i+\vec{i}} \psi_{i+\vec{i}''} \right). \quad (10.84)$$

By enforcing these two requirements we can ensure conservation of enstrophy and kinetic energy in the finite-difference model. The requirements can be met, as we will see, by suitable choice of the interaction coefficients. The requirements look daunting, though, because they involve triple sums. How in the world are we ever going to work this out?

Inspection of (10.83) shows that the individual terms of the sum are going to involve products of vorticities at pairs of grid points. With this in mind, we go back to (10.74) and rewrite it as

$$\begin{aligned} \sigma_i J_i(\zeta, \psi) &= \sum_{\vec{i}} \sum_{\vec{i}''} c_{i,\vec{i},\vec{i}''} \zeta_{i+\vec{i}} \psi_{i+\vec{i}''} \\ &= \sum_{\vec{i}} a_{i,i+\vec{i}} \zeta_{i+\vec{i}}, \end{aligned} \quad (10.85)$$

where, by definition,

$$a_{i, i+i'} \equiv \sum_{i''} c_{i, i', i''} \psi_{i+i''} . \quad (10.86)$$

Multiply (10.85) by ζ_i to obtain

$$\sigma_i \zeta_i J_i(\zeta, \psi) = \sum_{i'} a_{i, i+i'} \zeta_i \zeta_{i+i'} . \quad (10.87)$$

Here we have simply taken ζ_i inside the sum, which we can do because the sum is over i' , not i . From this point it is straightforward to enforce (10.83), which can be rewritten as

$$0 = \sum_i \left(\sum_{i'} a_{i, i+i'} \zeta_i \zeta_{i+i'} \right) . \quad (10.88)$$

Think of the outer sum in (10.88) as a “DO” loop. As we sweep over the grid, each product $\zeta_i \zeta_{i+i'}$ will enter the sum exactly twice. We can specify the vorticities any way we want, e.g., when we set up the initial conditions, so the only way to make sure that (10.88) is satisfied is to force these two contributions to the sum to be equal and opposite, i.e. we must take

$$a_{i, i+i'} = -a_{i+i', i}, \text{ for all } i \text{ and } i' . \quad (10.89)$$

By enforcing (10.89), we can ensure enstrophy conservation.

Kinetic energy conservation can be guaranteed by a very similar approach. We rewrite (10.74) as

$$\begin{aligned} \sigma_i J_i(\zeta, \psi) &= \sum_{i'} \sum_{i''} c_{i, i', i''} \zeta_{i+i'} \psi_{i+i''} \\ &= \sum_{i''} b_{i, i+i''} \psi_{i+i''} , \end{aligned} \quad (10.90)$$

where

$$b_{i, i+i''} \equiv \sum_{i'} c_{i, i', i''} \zeta_{i+i'} . \quad (10.91)$$

By an argument similar to that given above, we find that

$$b_{i, i+i''} = -b_{i+i'', i} \text{ for all } i \text{ and } i'' \quad (10.92)$$

is necessary to ensure kinetic energy conservation.

As already mentioned, we have not assumed anything about the shape of the domain.

It could be a doubly periodic plane, or it could be a sphere. Also, we have not assumed anything about the grid. The individual cells have not been assumed to have any particular shape, e.g. quadrilaterals; the argument starting from (10.74) would apply equally well on a grid of hexagons.

At this point we acknowledge that (10.92) is not really sufficient to ensure kinetic energy conservation. We must also make sure that the finite-difference analog of (10.55) holds true, i.e.

$$\sum_i \left(\sigma_i \psi_i \frac{d\zeta_i}{dt} \right) = - \sum_i \left[\sigma_i \frac{d}{dt} \left(\frac{1}{2} |\nabla \psi|_i^2 \right) \right], \quad (10.93)$$

so that we can mimic with the finite-difference equations the derivation that we did with the continuous equations. In order to pursue this objective, we have to define a finite-difference Laplacian, and we do so now by choosing the simplest possibility, assuming a square grid with grid spacing d :

$$\zeta_{i,j} = (\nabla^2 \psi)_{i,j} \equiv \frac{1}{d^2} (\psi_{i+1,j} + \psi_{i-1,j} + \psi_{i,j+1} + \psi_{i,j-1} - 4\psi_{i,j}) . \quad (10.94)$$

Here we have reverted to a conventional double-subscripting scheme, for clarity. We also define a finite-difference kinetic energy by

$$\begin{aligned} K_{i,j} &= \frac{1}{2} |\nabla \psi|_{i,j}^2 \\ &\equiv \frac{1}{4d^2} [(\psi_{i+1,j} - \psi_{i,j})^2 + (\psi_{i,j+1} - \psi_{i,j})^2 + (\psi_{i,j} - \psi_{i-1,j})^2 + (\psi_{i,j} - \psi_{i,j-1})^2] \end{aligned} \quad (10.95)$$

Because the right-hand side of (10.95) is a sum of squares, we are guaranteed that the kinetic energy is non-negative. By substitution, and after a little algebra, we can show that (10.94) and (10.95) do, in fact, satisfy (10.93).

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 (10.74) 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 (10.74). 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:

$$\begin{aligned} (J_1)_{i,j} &= \frac{1}{4d^2} [(\zeta_{i+1,j} - \zeta_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) \\ &\quad - (\zeta_{i,j+1} - \zeta_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j})], \end{aligned} \quad (10.96)$$

$$(J_2)_{i,j} = \frac{1}{4d^2} [-(\zeta_{i+1,j+1} - \zeta_{i+1,j-1})\psi_{i+1,j} + (\zeta_{i-1,j+1} - \zeta_{i-1,j-1})\psi_{i-1,j} \\ + (\zeta_{i+1,j+1} - \zeta_{i-1,j+1})\psi_{i,j+1} - (\zeta_{i+1,j-1} - \zeta_{i-1,j-1})\psi_{i,j-1}], \quad (10.97)$$

$$(J_3)_{i,j} = \frac{1}{4d^2} [\zeta_{i+1,j}(\psi_{i+1,j+1} - \psi_{i+1,j-1}) - \zeta_{i-1,j}(\psi_{i-1,j+1} - \psi_{i-1,j-1}) \\ - \zeta_{i,j+1}(\psi_{i+1,j+1} - \psi_{i-1,j+1}) + \zeta_{i,j-1}(\psi_{i+1,j-1} - \psi_{i-1,j-1})]. \quad (10.98)$$

These can be interpreted, respectively, as finite-difference analogs to the right-hand sides of (10.16) - (10.18). We can show that all three of these finite-difference Jacobians are consistent with vorticity conservation under advection, i.e., they all satisfy (10.76).

What we need to do next is identify the coefficients a and b for each of (10.96), (10.97), and (10.98), and then check to see whether the requirements (10.89) and (10.92) are satisfied for any of them. In order to understand more clearly what these requirements actually mean, look at Fig. 10.10. The Jacobians J_1 , J_2 , and J_3 are represented in the top row of the

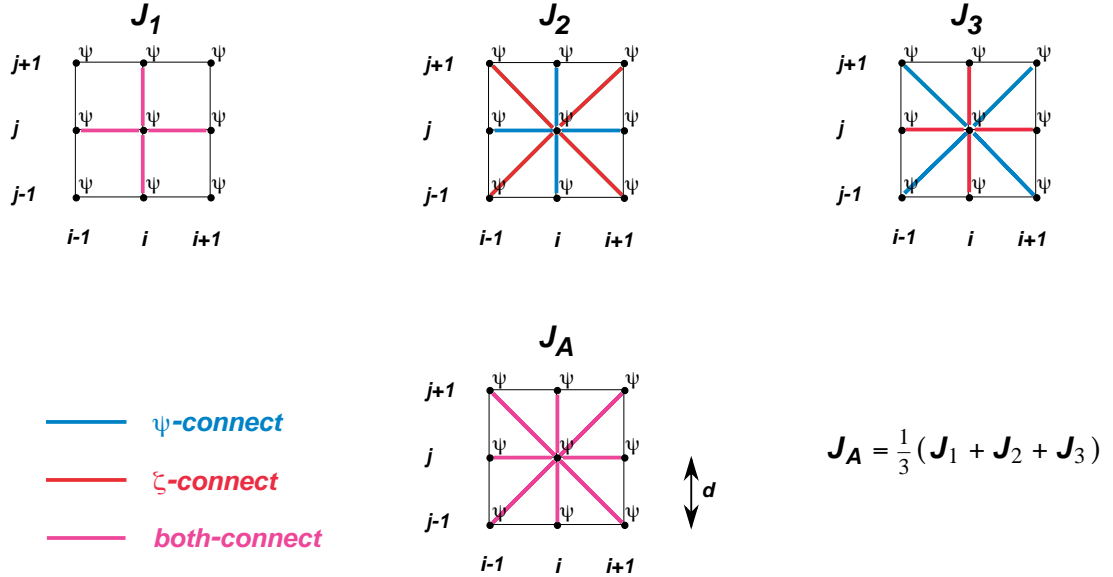


Figure 10.10: The central point in each figure is (i, j) . Stream function and vorticity are both defined at each of the mesh points indicated by the black dots. The colored lines represent contributions to $J_{i,j}$ from ψ , ζ , or both, from the various neighboring points.

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 begin by rewriting (10.85) using the conventional double-subscript notation and equating it to $(J_1)_{i,j}$:

$$\begin{aligned}
 \sigma_{i,j}(J_1)_{i,j}(\zeta, \psi) &= \sum_{i'} \sum_{j'} \sum_{i''} \sum_{j''} c_{i',j';i'',j''} \zeta_{i,j;i'+i'',j'+j''} \psi_{i+i'',j+j''} \\
 &= \sum_{i'} \sum_{j'} a_{i,j;i'+i'',j'+j''} \zeta_{i,j;i'+i'',j'+j''} \\
 &= \frac{1}{4} [(\zeta_{i+1,j} - \zeta_{i-1,j})(\psi_{i,j+1} - \psi_{i,j-1}) - (\zeta_{i,j+1} - \zeta_{i,j-1})(\psi_{i+1,j} - \psi_{i-1,j})] \quad (10.99) \\
 &= \frac{1}{4} [\zeta_{i+1,j}(\psi_{i,j+1} - \psi_{i,j-1}) - \zeta_{i-1,j}(\psi_{i,j+1} - \psi_{i,j-1}) \\
 &\quad - \zeta_{i,j+1}(\psi_{i+1,j} - \psi_{i-1,j}) + \zeta_{i,j-1}(\psi_{i+1,j} - \psi_{i-1,j})].
 \end{aligned}$$

Here we have used

$$\sigma_{i,j} = d^2. \quad (10.100)$$

By inspection of (10.99) and comparison with (10.85), we can read off the definitions of the a coefficients for J_1 :

$$a_{i,j;i+1,j} = \frac{1}{4}(\psi_{i,j+1} - \psi_{i,j-1}), \quad (10.101)$$

$$a_{i,j;i-1,j} = -\frac{1}{4}(\psi_{i,j+1} - \psi_{i,j-1}), \quad (10.102)$$

$$a_{i,j;i,j+1} = -\frac{1}{4}(\psi_{i+1,j} - \psi_{i-1,j}), \quad (10.103)$$

$$a_{i,j;i,j-1} = \frac{1}{4}(\psi_{i+1,j} - \psi_{i-1,j}). \quad (10.104)$$

Are these consistent with (10.89)? To find out, replace i by $i+1$ in (10.102); this gives:

$$a_{i+1,j;i,j} = -\frac{1}{4}(\psi_{i+1,j+1} - \psi_{i+1,j-1}). \quad (10.105)$$

Now simply compare (10.105) with (10.101), to see that (10.89) is *not* satisfied by J_1 . We have thus deduced that J_1 does not conserve enstrophy.

We can interpret that $a_{i, i+i'}$ denotes ζ -interactions of point i with point $i+i'$, while $a_{i+i', i}$ denotes ζ -interactions of point $i+i'$ with point i . When we compare $a_{i, i+i'}$ with $a_{i+i', i}$, it is like peering along one of the red lines in Fig. 10.10, first outward from the point (i, j) , to one of the other points, and then back towards the point (i, j) . The condition (10.89) on the a 's essentially means that all such interactions are “equal and opposite,” allowing suitable algebraic cancellations to occur when we sum over all points. The condition (10.92) on the b 's has a similar interpretation.

By proceeding as illustrated above, 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:

$$J_A = \alpha J_1 + \beta J_2 + \gamma J_3, \quad (10.106)$$

such that

$$\alpha + \beta + \gamma = 1. \quad (10.107)$$

With three unknown coefficients, and only one constraint, (10.107), we are free to satisfy two additional constraints; and we take these to be (10.89) and (10.92). In this way, we find that J_A will conserve both enstrophy and kinetic energy if we choose

$$\alpha = \beta = \gamma = 1/3. \quad (10.108)$$

The composite Jacobian J_A is often called the “Arakawa Jacobian.”

Fig. 10.11 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, so long as we do not violate the criteria for linear computational instability. 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 conservative second-order accurate (in space) finite-difference approximation to the Jacobian. Arakawa (1966) also showed how to obtain the corresponding conservative Jacobian with fourth-order accuracy.

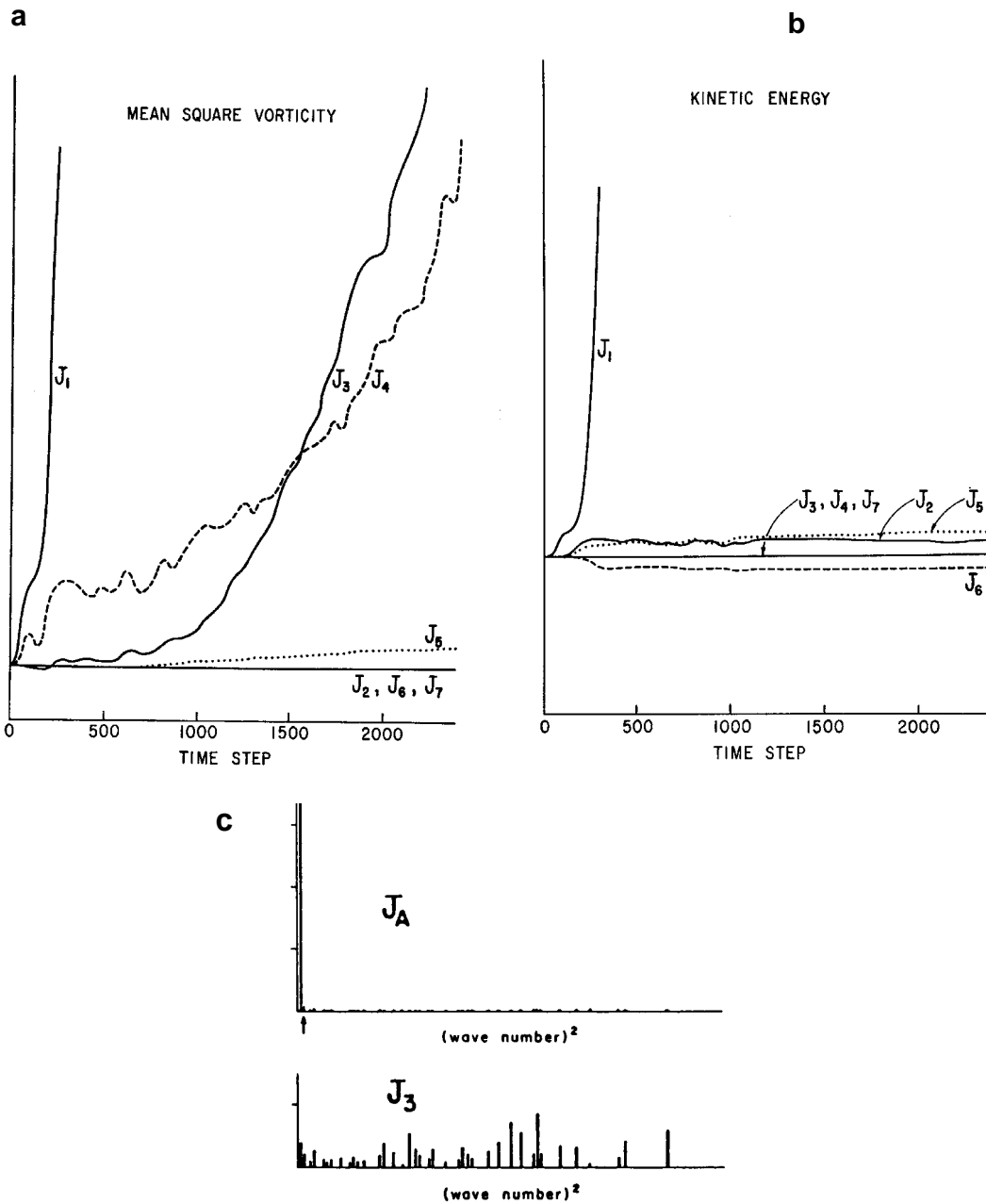


Figure 10.11: Results of tests with the various finite-difference Jacobians. Panel c shows the initial kinetic energy is at a low wave number.

The preceding analysis shows 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.

In Chapter 4, 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 was denoted 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. What is going on?

The key difference with the vorticity equation is that we can choose not only how to interpolate the vorticity (so as to conserve the enstrophy), but also *the actual finite-difference expression for the advecting wind itself*, in terms of the stream function, because that expression is implicit in the form of the Jacobian that we use. In choosing the form of the advecting current, we have a second “freedom,” which allows us to conserve a second quantity, namely the kinetic energy.

As discussed earlier, the constraint of enstrophy conservation is needed to ensure that kinetic energy does not cascade in two-dimensional nondivergent 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.

In contrast, a scheme that conserves kinetic energy but not enstrophy will permit a kinetic energy cascade. The resulting noisy flow will lead to large errors in enstrophy conservation. Such a scheme will not behave well.

These considerations suggest that formal enstrophy conservation is “more important” than formal kinetic energy conservation.

10.5 Angular momentum conservation

Finally, for completeness, define the relative angular momentum per unit mass, M , by

$$M_{\text{rel}} \equiv ua \cos \varphi. \quad (10.109)$$

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,

$$\frac{\partial M}{\partial t} = -(\mathbf{v} \cdot \nabla)M, \quad (10.110)$$

where λ is longitude, and

$$M \equiv M_{\text{rel}} + \Omega a^2 \cos \varphi \quad (10.111)$$

is the component of the absolute angular momentum vector in the direction of the axis of the Earth's rotation. From (10.110) it follows that the absolute angular momentum is conserved under advection.

Using integration by parts, it can be demonstrated that

$$\overline{M_{\text{rel}}} = a^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (\sin \varphi) \zeta \cos \varphi \, d\lambda d\varphi. \quad (10.112)$$

We can also prove that

$$\frac{d}{dt} \overline{M_{\text{rel}}} = a^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (\sin \varphi) \frac{\partial \zeta}{\partial t} \cos \varphi \, d\lambda d\varphi = 0. \quad (10.113)$$

This means that angular momentum is conserved.

10.6 *Conservative schemes for the two-dimensional shallow water equations with rotation*

To study the two-dimensional shallow-water equations, we use cartesian coordinates x and y , with velocity vector \mathbf{V} , such that

$$\mathbf{V} \equiv u\mathbf{i} + v\mathbf{j}, \quad (10.114)$$

where \mathbf{i} and \mathbf{j} are the unit vectors in the x and y directions, respectively. The shallow-water equations can be written as

$$\frac{\partial h}{\partial t} + \nabla \bullet (h\mathbf{V}) = 0, \quad (10.115)$$

and

$$\frac{\partial \mathbf{V}}{\partial t} + \left(\frac{\zeta + f}{h} \right) \mathbf{k} \times (h\mathbf{V}) + \nabla [K + g(h + h_S)] = 0, \quad (10.116)$$

where

$$\zeta \equiv \mathbf{k} \bullet (\nabla \times \mathbf{V}) \quad (10.117)$$

is the relative vorticity, and

$$f \equiv 2\Omega \sin \varphi \quad (10.118)$$

is the Coriolis parameter. In (10.116), we have multiplied and divided the vorticity term by h , for two good reasons, to be explained later. The combination $\left(\frac{\zeta+f}{h}\right)$ is the potential vorticity.

The corresponding equations for the zonal and meridional wind components are

$$\frac{\partial u}{\partial t} - \left(\frac{\zeta+f}{h}\right)(hv) + \frac{\partial}{\partial x}[K + g(h + h_S)] = 0, \quad (10.119)$$

and

$$\frac{\partial v}{\partial t} + \left(\frac{\zeta+f}{h}\right)(hu) + \frac{\partial}{\partial y}[K + g(h + h_S)] = 0, \quad (10.120)$$

respectively.

When we take the dot product of (10.116) with $h\mathbf{V}$, the vorticity term of (10.116) contributes nothing *because of the vector identity*

$$(h\mathbf{V}) \bullet [\mathbf{k} \times (h\mathbf{V})] = 0, \quad (10.121)$$

and so we obtain very directly the advective form of the kinetic energy equation, i.e.

$$h \frac{\partial K}{\partial t} + (h\mathbf{V}) \bullet \nabla[K + g(h + h_S)] = 0, \quad (10.122)$$

where $K \equiv \frac{1}{2} \mathbf{V} \bullet \mathbf{V}$ is the kinetic energy per unit mass. By use of the continuity equation (10.115), we can rewrite (10.122) in flux form:

$$\frac{\partial}{\partial t}(hK) + \nabla \bullet (h\mathbf{V}K) + (h\mathbf{V}) \bullet \nabla[g(h + h_S)] = 0. \quad (10.123)$$

Similarly, the flux form of the potential energy equation is

$$\frac{\partial}{\partial t} \left[h \left(gh_S + \frac{1}{2} h \right) \right] + \nabla \bullet [(h\mathbf{V})g(h + h_S)] - (h\mathbf{V}) \bullet \nabla[g(h + h_S)] = 0. \quad (10.124)$$

The method presented earlier to ensure conservation of kinetic energy for the one-dimensional finite-difference shallow-water equations carries through to two dimensions in very straightforward fashion, and so the details will not be given here.

The important new physical ingredient that must be considered in the two-dimensional finite-difference system is rotation, including both Earth-rotation, f , and the

relative vorticity, ζ , associated with the wind field. We have already considered the effects of rotation for the case of two-dimensional nondivergent flow. Now we have divergence, so in place of vorticity conservation and enstrophy conservation we must generalize to potential vorticity conservation and potential enstrophy conservation. One obvious and important question is: *Can we find a finite-difference scheme that allows us to “mimic” the identity (10.121)?*

The approach outlined below follows Arakawa and Lamb (1981). We adopt the C-grid, as shown in Fig. 10.12. Recall that on the C-grid the zonal winds are east and west of

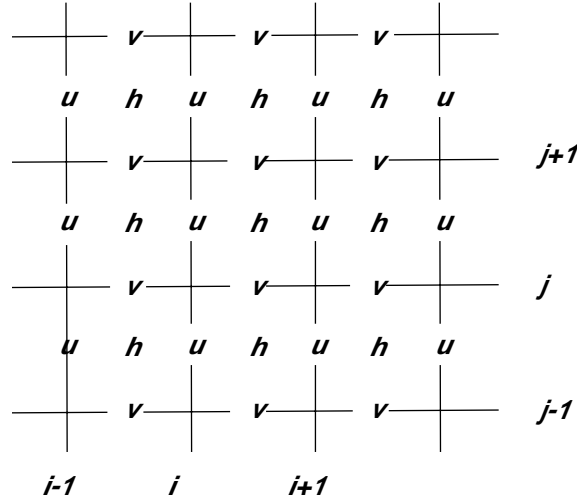


Figure 10.12: The arrangement of the mass, zonal wind, and meridional wind on the C grid.

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 (i, j) ; and the vorticity “wants” to be defined along the diagonal lines connecting mass points, e.g., at the point $(i + 1/2, j - 1/2)$.

The finite-difference form of the continuity equation is

$$\frac{dh}{dt}_{i+\frac{1}{2}, j+\frac{1}{2}} = \frac{(hu)_{i, j+\frac{1}{2}} - (hu)_{i+1, j+\frac{1}{2}}}{\Delta x} + \frac{(hv)_{i+\frac{1}{2}, j} - (hv)_{i+\frac{1}{2}, j+1}}{\Delta y}. \quad (10.125)$$

The various mass fluxes that appear in (10.125) 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

$$\begin{aligned} & \frac{d}{dt} u_{i,j+\frac{1}{2}} - \left[\left(\frac{\zeta+f}{h} \right) (hv) \right]_{i,j+\frac{1}{2}} \\ & + \frac{1}{\Delta x} \left(K_{i+\frac{1}{2},j+\frac{1}{2}} - K_{i-\frac{1}{2},j+\frac{1}{2}} \right) + \frac{g}{\Delta x} \left[(h+h_S)_{i+\frac{1}{2},j+\frac{1}{2}} - (h+h_S)_{i-\frac{1}{2},j+\frac{1}{2}} \right] = 0, \end{aligned} \quad (10.126)$$

and

$$\begin{aligned} & \frac{d}{dt} v_{i+\frac{1}{2},j} + \left[\left(\frac{\zeta+f}{h} \right) (hu) \right]_{i+\frac{1}{2},j} \\ & + \frac{1}{\Delta y} \left(K_{i+\frac{1}{2},j+\frac{1}{2}} - K_{i+\frac{1}{2},j-\frac{1}{2}} \right) + \frac{g}{\Delta y} \left[(h+h_S)_{i+\frac{1}{2},j+\frac{1}{2}} - (h+h_S)_{i+\frac{1}{2},j-\frac{1}{2}} \right] = 0, \end{aligned} \quad (10.127)$$

respectively. As in the one-dimensional case, 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 on the C-grid the mass fluxes that appear in (10.130) and (10.126) 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 (10.126) and (10.127) 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 (10.126) and (10.127). Note, however, that it is actually *products* of mass fluxes and vorticities that are needed.

Arakawa and Lamb constructed the finite-difference vorticity terms in such a way that a finite-difference analog to (10.121) is satisfied, regardless of the specific forms of the mass fluxes and potential vorticities that are chosen. They constructed the vorticity terms as follows:

$$\begin{aligned} \left[\left(\frac{\zeta+f}{h} \right) (hv) \right]_{i,j+\frac{1}{2}} &= \alpha_{i,j+\frac{1}{2};i+\frac{1}{2},j+1} (hv)_{i+\frac{1}{2},j+1} + \beta_{i,j+\frac{1}{2};i-\frac{1}{2},j+1} (hv)_{i-\frac{1}{2},j+1} \\ &+ \gamma_{i,j+\frac{1}{2};i-\frac{1}{2},j} (hv)_{i-\frac{1}{2},j} + \delta_{i,j+\frac{1}{2};i+\frac{1}{2},j} (hv)_{i+\frac{1}{2},j}, \end{aligned} \quad (10.128)$$

and

$$\begin{aligned} \left[\left(\frac{\zeta+f}{h} \right) (hu) \right]_{i+\frac{1}{2},j} &= \gamma_{i+\frac{1}{2},j;i+1,j+\frac{1}{2}} (hu)_{i+1,j+\frac{1}{2}} + \delta_{i+\frac{1}{2},j;i,j+\frac{1}{2}} (hu)_{i,j+\frac{1}{2}} \\ &+ \alpha_{i+\frac{1}{2},j;i,j-\frac{1}{2}} (hu)_{i,j-\frac{1}{2}} + \beta_{i+\frac{1}{2},j;i+1,j-\frac{1}{2}} (hu)_{i+1,j-\frac{1}{2}}. \end{aligned} \quad (10.129)$$

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 (10.128) and (10.129), the α 's, β 's, γ 's, and δ '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 α 's, β 's, γ 's, and δ '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 (10.128), the interpolated potential vorticities multiply the mass fluxes $h\nu$ at the four v -wind points surrounding the u -wind point $i, j + \frac{1}{2}$, and similarly in (10.129) they multiply the mass fluxes hu at the four u -wind points surrounding the v -wind point $i + \frac{1}{2}, j$.

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 (10.126) by $(hu)_{i,j+\frac{1}{2}}$ and (10.126) by $(h\nu)_{i+\frac{1}{2},j}$, and add the results. With the forms given by (10.128) and (10.129), the vorticity terms will sum to

$$\begin{aligned} &-(hu)_{i,j+\frac{1}{2}} \left[\left(\frac{\zeta+f}{h} \right) (h\nu) \right]_{i,j+\frac{1}{2}} + (h\nu)_{i+\frac{1}{2},j} \left[\left(\frac{\zeta+f}{h} \right) (hu) \right]_{i+\frac{1}{2},j} \\ &= -(hu)_{i,j+\frac{1}{2}} \left[\alpha_{i,j+\frac{1}{2};i+\frac{1}{2},j+1} (h\nu)_{i+\frac{1}{2},j+1} + \beta_{i,j+\frac{1}{2};i-\frac{1}{2},j+1} (h\nu)_{i-\frac{1}{2},j+1} \right. \\ &\quad \left. + \gamma_{i,j+\frac{1}{2};i-\frac{1}{2},j} (h\nu)_{i-\frac{1}{2},j} + \delta_{i,j+\frac{1}{2};i+\frac{1}{2},j} (h\nu)_{i+\frac{1}{2},j} \right] \\ &\quad + (h\nu)_{i+\frac{1}{2},j} \left[\gamma_{i+\frac{1}{2},j;i+1,j+\frac{1}{2}} (hu)_{i+1,j+\frac{1}{2}} + \delta_{i+\frac{1}{2},j;i,j+\frac{1}{2}} (hu)_{i,j+\frac{1}{2}} \right. \\ &\quad \left. + \alpha_{i+\frac{1}{2},j;i,j-\frac{1}{2}} (hu)_{i,j-\frac{1}{2}} + \beta_{i+\frac{1}{2},j;i+1,j-\frac{1}{2}} (hu)_{i+1,j-\frac{1}{2}} \right]. \end{aligned} \quad (10.130)$$

Inspection of (10.130) makes it clear that cancellation will occur when we sum over the 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 for the mass fluxes, and regardless of the expressions that we choose for the α 's, β 's, γ 's, and δ 's. The cancellation arises purely from the forms of (10.128) and (10.129), and is analogous to the cancellation that makes (10.121) work, i.e.

$$A\mathbf{V} \bullet (\mathbf{k} \times \mathbf{V}) = A(u\mathbf{i} + v\mathbf{j}) \bullet (-v\mathbf{i} + u\mathbf{j}) = A(-uv + uv) = 0, \quad (10.131)$$

regardless of the input quantities A and \mathbf{V} . This is yet another example of “mimetic discretization.”

The above discussion shows that the finite-difference momentum equations represented by (10.126) and (10.126) with the use of (10.128) and (10.129) will guarantee kinetic energy conservation under advection, regardless of the forms chosen for the mass fluxes and the interpolated potential vorticities α , β , γ , and δ . 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.

10.7 The effects of time differencing on energy conservation

A family of finite-difference schemes for (10.13) can be written in the generic form

$$\frac{q_{i,j}^{n+1} - q_{i,j}^n}{\Delta t} = J_{i,j}(q^*, \psi), \quad (10.132)$$

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. Examples are given in Table 10.1.

Table 10.1: Examples of time differencing schemes obtained through various choices of q^* .
The subscripts i and j have been omitted for simplicity.

| Name of Scheme | Form of Scheme |
|-----------------------------|------------------------------------|
| <i>Euler (forward)</i> | $q^* = q^n$ |
| <i>Backward implicit</i> | $q^* = q^{n+1}$ |
| <i>Trapezoidal implicit</i> | $q^* = \frac{1}{2}(q^n + q^{n+1})$ |

Table 10.1: Examples of time differencing schemes obtained through various choices of q^* . The subscripts i and j have been omitted for simplicity.

| Name of Scheme | Form of Scheme |
|---|---|
| <i>Leapfrog (here the time interval is $\Delta t/2$)</i> | $q^* = q^{n+\frac{1}{2}}$ |
| <i>Second-order Adams Bashforth</i> | $q^* = \frac{3}{2}q^n - \frac{1}{2}q^{n-1}$ |
| <i>Heun</i> | $q^* = q^n + \frac{\Delta t}{2}J(q^n, \psi)$ |
| <i>Lax-Wendroff</i> (here S is a smoothing operator) | $q^* = Sq^n + \frac{\Delta t}{2}J(q^n, \psi)$ |
| <i>Matsuno</i> | $q^* = q^n + \Delta t J(q^n, \psi)$ |

Multiplying (10.132) by q^* , we get

$$q^*(q^{n+1} - q^n) = \Delta t q^* J(q^*, \psi), \quad (10.133)$$

or, after some algebraic sleight-of-hand,

$$(q^{n+1})^2 - (q^n)^2 = 2\left(\frac{q^{n+1} + q^n}{2} - q^*\right)(q^{n+1} - q^n) + 2\Delta t q^* J(q^*, \psi). \quad (10.134)$$

The left-hand side of (10.134) represents the change of q^2 in one time step. Consider the summation of q^2 over all grid points, divided by the number of grid points, and let this mean be denoted by an overbar. We find that

$$\overline{(q^{n+1})^2} - \overline{(q^n)^2} = 2\overline{\left(\frac{q^{n+1} + q^n}{2} - q^*\right)(q^{n+1} - q^n)} + 2\Delta t \overline{q^* J(q^*, \psi)}, \quad (10.135)$$

which shows that the change of the mean-square of q depends on two terms. The first term involves the choice of q^* . For $q^* = q^n$, the contribution of this term is positive and so tends to increase $\overline{q^2}$, while for $q^* = q^{n+1}$, it is negative and so tends to decrease $\overline{q^2}$. If we use the trapezoidal scheme, 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 scheme is consistent with (allows) exact energy conservation. Of course, the form of the finite-difference Jacobian must also be consistent with energy conservation.

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 Neuman's method.

10.8 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 Neuman's method. It can be detected by the energy method (see Chapter 2), 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.

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 r frames per second.
 - a) What inequality must r satisfy to avoid aliasing?
 - b) What rotation rate does a person watching the film see, for given values of r and R ?
 - c) Under what conditions does the wheel appear to turn “backwards?”
2. Prove that J_3 gives kinetic energy conservation for the case of two-dimensional nondivergent flow.
3. Prove that J_2 gives exact vorticity conservation (ignoring time truncation error). Assume periodic boundary conditions.
4. Work out the *continuous* form of the Jacobian for the case of spherical coordinates (longitude λ and latitude ϕ).

CHAPTER 11

Finite Differences on the Sphere

Copyright 2004 David A. Randall

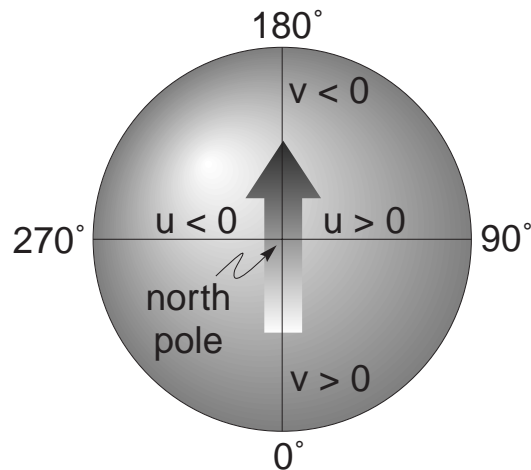
11.1 Introduction

There are a number of problems associated with trying to solve differential equations numerically in spherical geometry. Difficulties can arise from the use of spherical coordinate systems, and from trying to discretize the surface of the sphere itself. These difficulties include what is called the “pole problem.”

Even before any numerical considerations are confronted, we run into difficulties simply in specifying a coordinate system. It is important to distinguish between true scalar-valued and vector-valued functions. The value of a scalar function, say temperature, is independent of coordinate system. That is, the temperature at some point in space is the same regardless of how we define the position of the point. On the other hand, the individual components of a vector-valued function, such as the wind velocity, obviously differ depending on the coordinate system. In a spherical coordinate system, the lines of constant longitude converge at the poles, so longitude is multivalued at the poles. This means that the *components* of the wind vector are discontinuous at the poles, although the wind vector itself is perfectly well behaved at the pole.

As a simple example, consider a jet directed over the North Pole. This is represented by the shaded arrow in Fig. 11.1. Measured at points along the prime meridian, the wind will

Figure 11.1: 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° longitude have a strong positive zonal component, while points near 270° longitude have a strong negative zonal component.



have a positive v component. Measured along the international date line, however, the wind

will have 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 90° longitude, and is negative measured along 270° longitude. This problem does not occur in a Cartesian coordinate system centered on the pole. At each point along a great circle which includes the pole, the components measured in Cartesian coordinates are well defined and vary continuously.

11.2 Coordinate systems and map projections

We now express the shallow water equations in the spherical coordinate system (λ, ϕ) . In three-dimensional spherical coordinates, the gradient, divergence, and curl operators take the following forms:

$$\nabla A = \left(\frac{1}{r \cos \phi} \frac{\partial A}{\partial \lambda}, \frac{1}{r} \frac{\partial A}{\partial \phi}, \frac{\partial A}{\partial r} \right), \quad (11.1)$$

$$\nabla \cdot \mathbf{V} = \frac{1}{r \cos \phi} \frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{r \cos \phi} \frac{\partial}{\partial \phi} (V_\phi \cos \phi) + \frac{1}{r^2} \frac{\partial}{\partial r} (V_r r^2), \quad (11.2)$$

$$\begin{aligned} \nabla \times \mathbf{V} = & \left\{ \frac{1}{r} \left[\frac{\partial V_r}{\partial \phi} - \frac{\partial}{\partial r} (r V_\phi) \right], \right. \\ & \frac{1}{r} \frac{\partial}{\partial r} (r V_\lambda) - \frac{1}{r \cos \phi} \frac{\partial V_r}{\partial \lambda}, \\ & \left. \frac{1}{r \cos \phi} \left[\frac{\partial V_\phi}{\partial \lambda} - \frac{\partial}{\partial \phi} (V_\lambda \cos \phi) \right] \right\}, \end{aligned} \quad (11.3)$$

For use with the two-dimensional shallow-water equations, we can simplify these to

$$\nabla A = \left(\frac{1}{a \cos \phi} \frac{\partial A}{\partial \lambda}, \frac{1}{a} \frac{\partial A}{\partial \phi} \right), \quad (11.4)$$

$$\nabla \cdot \mathbf{V} = \frac{1}{a \cos \phi} \frac{\partial V_\lambda}{\partial \lambda} + \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} (V_\phi \cos \phi), \quad (11.5)$$

$$\mathbf{k} \cdot (\nabla \times \mathbf{V}) = \frac{1}{a \cos \phi} \left[\frac{\partial V_\phi}{\partial \lambda} - \frac{\partial}{\partial \phi} (V_\lambda \cos \phi) \right]. \quad (11.6)$$

Here a is the radius of the spherical planet.

The shallow water equations in spherical coordinates can be expressed as

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} - \left(f + \frac{u}{a} \tan \phi \right) v + \frac{g}{a \cos \phi} \frac{\partial}{\partial \lambda} (h + h_S) = 0, \quad (11.7)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \left(f + \frac{u}{a} \tan \phi \right) u + \frac{g}{a \sin \phi} \frac{\partial}{\partial \phi} (h + h_S) = 0, \quad (11.8)$$

$$\frac{\partial h}{\partial t} + u \frac{\partial h}{\partial x} + v \frac{\partial h}{\partial y} + \frac{h}{a \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right] = 0. \quad (11.9)$$

Here h is the depth of the fluid, and h_S is the height of the bottom topography.

An early approach to numerically solving the shallow water equations on the sphere was to project the equations from the sphere to 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. In other words, there does not exist a one-to-one mapping g such that for every point on the sphere $(\lambda, \phi) \in S$ there exists $(x, y) \in P \equiv \{(x, y) | -\infty < x, y < \infty\}$ satisfying $g(\lambda, \phi) = (x, y)$. There are mappings which map almost all of S onto P ; examples are given below. Unfortunately, these mappings, or projections, tend to badly distort distances and areas near the singular points of the transformation. Nevertheless, we can use a projection to map the piece of the sphere where the transformation is well behaved onto a finite region of the plane. 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) . Define the metric coefficients to be h_x and h_y so that the distance increment dl satisfies

$$dl^2 = h_x^2 dx^2 + h_y^2 dy^2. \quad (11.10)$$

Note that, as a matter of notation, the metric coefficients h_x and h_y are distinguished from depth of the fluid, h , by a subscript. In the (x, y) coordinate system, the horizontal velocity components are given by

$$U = h_x \frac{dx}{dt}, \quad (11.11)$$

$$V = h_y \frac{dy}{dt}. \quad (11.12)$$

Williamson (1979) gives the equations of motion for the general velocity components:

$$\frac{dU}{dt} - \left[f + \frac{1}{h_x h_y} \left(V \frac{\partial h_y}{\partial x} - U \frac{\partial h_x}{\partial y} \right) \right] V + \frac{g}{h_x} \frac{\partial}{\partial x} (h + h_S) = 0, \quad (11.13)$$

$$\frac{dV}{dt} + \left[f + \frac{1}{h_x h_y} \left(V \frac{\partial h_y}{\partial x} - U \frac{\partial h_x}{\partial y} \right) \right] U + \frac{g}{h_y} \frac{\partial}{\partial y} (h + h_S) = 0, \quad (11.14)$$

where

$$\frac{d}{dt}(\quad) = \frac{\partial}{\partial t}(\quad) + \frac{U}{h_x} \frac{\partial}{\partial x}(\quad) + \frac{V}{h_y} \frac{\partial}{\partial y}(\quad). \quad (11.15)$$

The continuity equation is given by

$$\frac{dh}{dt} + \frac{h}{h_x h_y} \left[\frac{\partial}{\partial x} (h_x U) + \frac{\partial}{\partial y} (h_y V) \right] = 0. \quad (11.16)$$

For example, if we set

$$x = \lambda \text{ and } y = \phi, \quad (11.17)$$

and correspondingly set the metric coefficients to

$$h_x = a \cos \phi \text{ and } h_y = a, \quad (11.18)$$

then by (11.11) and (11.12) we have

$$U = u \equiv a \cos \phi \frac{d\lambda}{dt} \text{ and } V = v \equiv a \frac{d\phi}{dt}. \quad (11.19)$$

Substituting (11.18) and (11.19) into (11.13), (11.14) and (11.16) gives (11.7), (11.8) and (11.9), the shallow water equations in spherical coordinates.

Two map projections are commonly used in numerical modeling of the atmospheric circulation -- Polar Stereographic and Mercator. 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., $h_x = h_y$. The effects of these projections on the outlines of the continents are shown in Fig. 11.2.

The polar stereographic projection can be visualized in terms of a plane tangent to the Earth at the North Pole. A line drawn from the South Pole that intersects the Earth will also intersect the plane. This line establishes a one-to-one correspondence between all points on the plane and all points on the sphere except for the South Pole itself. In the plane, we can

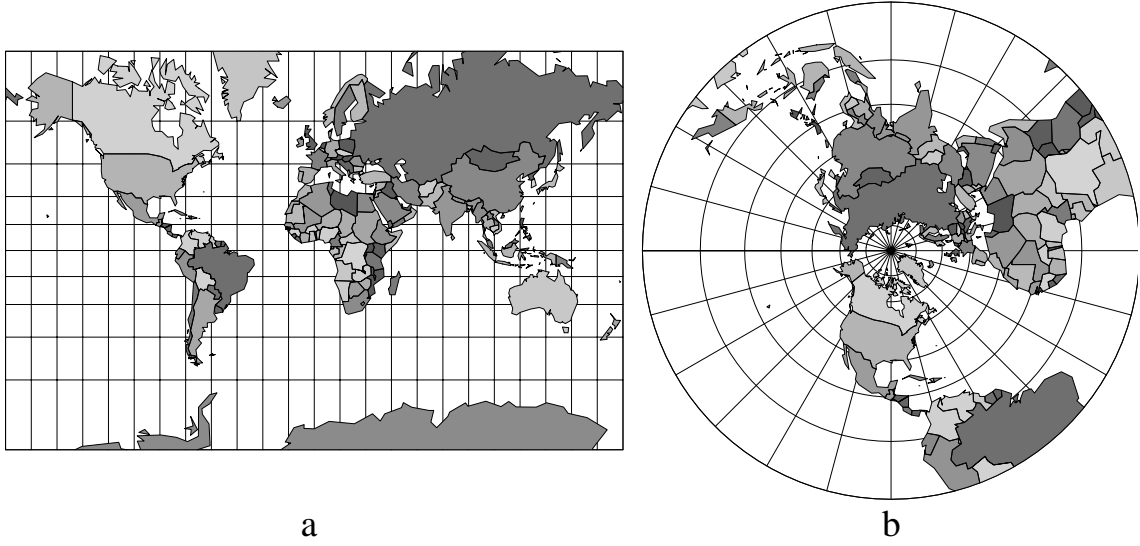


Figure 11.2: Map projections of the continents: a.) Mercator projection. b.) North polar stereographic projection.

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 (1984) give the equations relating the projection coordinates (X, Y) and the spherical coordinates (λ, φ) :

$$X = \frac{2a \cos \varphi \cos \lambda}{1 + \sin \varphi}, \quad (11.20)$$

$$Y = \frac{2a \cos \varphi \sin \lambda}{1 + \sin \varphi}. \quad (11.21)$$

Note that there is a problem at the South Pole. Taking differentials of (11.20) and (11.21) gives

$$\begin{bmatrix} dX \\ dY \end{bmatrix} = \frac{2a}{1 + \sin \varphi} \begin{bmatrix} -\cos \varphi \sin \lambda & -\cos \lambda \\ \cos \varphi \cos \lambda & -\sin \lambda \end{bmatrix} \begin{bmatrix} d\lambda \\ d\varphi \end{bmatrix}. \quad (11.22)$$

Now we determine the metrics of the polar stereographic map projection. Substituting $x = \lambda$, $y = \varphi$ and the metrics for spherical coordinates into $dl^2 = h_x^2 dx^2 + h_y^2 dy^2$ gives

$$dl^2 = (a \cos \varphi)^2 d\lambda^2 + a^2 d\varphi^2. \quad (11.23)$$

Solving (11.22) for $d\phi$, and $d\lambda$, and substituting the results into (11.23), we obtain

$$dl^2 = \left(\frac{1 + \sin\phi}{2} \right)^2 dX^2 + \left(\frac{1 + \sin\phi}{2} \right)^2 dY^2. \quad (11.24)$$

Comparing (11.24) with (11.10), we see that metric coefficient for the polar stereographic projection is given by

$$h_x = h_y = \frac{1 + \sin\phi}{2}. \quad (11.25)$$

We define the map factor, $m(\phi)$, as the inverse of the metric coefficient, i.e., $m(\phi) = 2/(1 + \sin\phi)$. Using (11.13), (11.14) and (11.16), we can write the shallow water equations in north polar stereographic coordinates:

$$\frac{dU}{dt} - \left[f + \frac{UY - VX}{2a^2} \right] V + mg \frac{\partial}{\partial x} (h + h_S) = 0, \quad (11.26)$$

$$\frac{dV}{dt} + \left[f + \frac{UY - VX}{2a^2} \right] U + mg \frac{\partial}{\partial x} (h + h_S) = 0, \quad (11.27)$$

$$\frac{dh}{dt} + m^2 h \left[\frac{\partial}{\partial X} \left(\frac{U}{m} \right) + \frac{\partial}{\partial Y} \left(\frac{V}{m} \right) \right] = 0. \quad (11.28)$$

The total derivative is given by (11.15).

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. 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.

More recently, Browning (1989) discussed a different 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 (11.26), (11.27) and (11.28). 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. 11.3. Values for

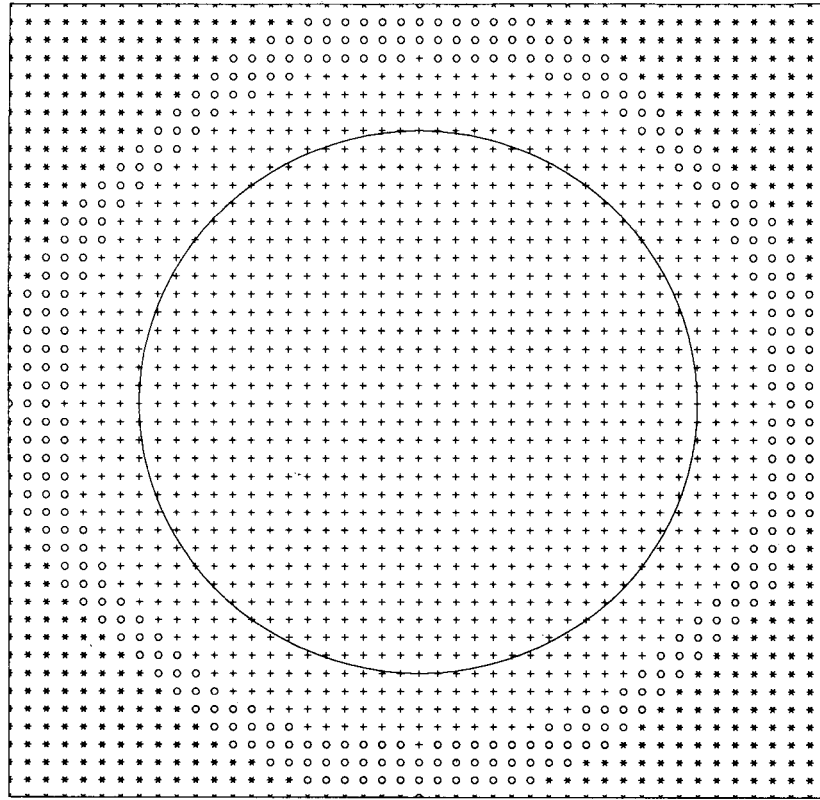


Figure 11.3: Composite grid method grid. Two such grids are used to cover the sphere. Points labeled with \circ are the boundary conditions for the points labeled with $+$. Values at the \circ points are obtained by interpolation from the other grid. The big circle is the image of the Equator. Points labeled $*$ are not used.

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 makes this scheme more stable than in Phillips’ model, in which the coordinate systems were simply butted together at a certain latitude. This model is also easier to write computer code for because the equations are only expressed in the polar stereographic coordinate systems. Browning tested the model and reported good results.

11.3 Latitude-longitude grids and the “pole problem”

An obvious way to discretize the shallow water equations expressed in spherical coordinates is to use a regular latitude-longitude grid in which the grid intervals ($\Delta\lambda$, $\Delta\phi$) are constants. A discretization scheme is straight forward except for the row of grid points next to the pole, where special considerations are necessary.

A portion (one eighth) of a uniform latitude-longitude grid is shown in Fig. 11.4. 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

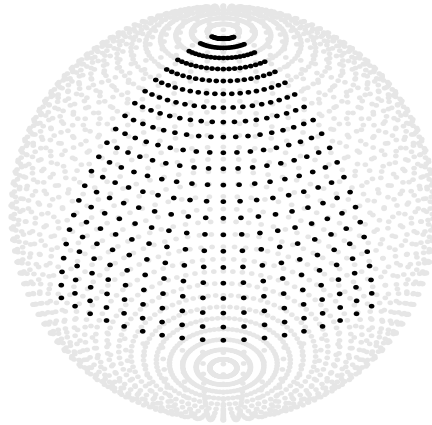


Figure 11.4: One octant of the latitude-longitude grid used by Arakawa and Lamb (1981). In the example shown, there are 72 grid points around a latitude circle and 44 latitude bands from pole to pole. The longitudinal grid spacing is globally uniform, and in this example is 5° . The latitudinal grid 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° except that the pizza slices are 6° tall.

trapezoidal in shape. There are other ways to deal with the polar regions, e.g. by defining local cartesian coordinates at the poles.

A regular latitude-longitude grid has some drawbacks. The scales of meteorological action do not vary dramatically from place to place, nor do the meridional and zonal scales of the circulations of interest differ very much. 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. 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 4 (for advection) and 5 (for wave propagation). This is often referred to as “the pole problem.”

To derive the stability criterion for the shallow water equations on the sphere, following Arakawa and Lamb (1977), we begin by linearizing (11.7), (11.8), and (11.9) about a state of rest, as follows:

$$\frac{\partial u}{\partial t} + \frac{g}{a \cos \phi} \frac{\partial h}{\partial \lambda} = 0, \quad (11.29)$$

$$\frac{\partial v}{\partial t} + \frac{g}{a} \frac{\partial h}{\partial \phi} = 0, \quad (11.30)$$

$$\frac{\partial h}{\partial t} + \frac{H}{a \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi} (v \cos \phi) \right] = 0. \quad (11.31)$$

Here we have neglected rotation and bottom topography, for simplicity, and H denotes the mean depth of the fluid. We spatially discretize these as follows:

$$\frac{\partial u}{\partial t} \Big|_{i+\frac{1}{2},j} + \frac{g(h_{i+1,j} - h_{i,j})}{a \cos \phi \Delta \lambda} = 0, \quad (11.32)$$

$$\frac{\partial v}{\partial t} \Big|_{i,j+\frac{1}{2}} + \frac{g(h_{i,j+1} - h_{i,j})}{a \Delta \phi} = 0, \quad (11.33)$$

$$\frac{\partial h_{i,j}}{\partial t} + H \left\{ \frac{\left(u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j} \right)}{a \cos \phi_j \Delta \lambda} + \left[\frac{(v \cos \phi)_{i,j+\frac{1}{2}} - (v \cos \phi)_{i,j-\frac{1}{2}}}{a \cos \phi_j \Delta \phi} \right] \right\} = 0. \quad (11.34)$$

Note that the C grid has been used here. We look for solutions of the form

$$u_{i+\frac{1}{2},j} = \text{Re} \left\{ \hat{u}_j \exp \left[i s \left(i + \frac{1}{2} \right) \Delta \lambda + i \sigma t \right] \right\}, \quad (11.35)$$

$$v_{i,j+\frac{1}{2}} = \text{Re} \left\{ \hat{v}_{j+\frac{1}{2}} \exp [i s i \Delta \lambda + i \sigma t] \right\}, \quad (11.36)$$

$$h_{i,j} = \text{Re} \{ \hat{h}_j \exp [i s i \Delta \lambda + i \sigma t] \}, \quad (11.37)$$

where $i \equiv \sqrt{-1}$. Note that the zonal wave number, s , is defined with respect to longitude rather than distance. After substitution of these solutions into our finite-difference equations, we obtain

$$i \sigma \hat{u}_j + \frac{i s}{a \cos \phi_j} \frac{\sin(s \Delta \lambda / 2)}{s \Delta \lambda / 2} g [S_j(s) \hat{h}_j] = 0, \quad (11.38)$$

$$i \sigma \hat{v}_{j+\frac{1}{2}} + \frac{g(\hat{h}_{j+1} - \hat{h}_j)}{a \Delta \phi} = 0, \quad (11.39)$$

$$i\sigma\hat{h}_j + H \left\{ \frac{is}{a \cos \phi_j} \frac{\sin(s\Delta\lambda/2)}{s\Delta\lambda/2} [S_j(s)\hat{u}_j] + \left[\frac{(\hat{v} \cos \phi)_{j+\frac{1}{2}} - (\hat{v} \cos \phi)_{j-\frac{1}{2}}}{a \cos \phi_j \Delta \phi} \right] \right\} = 0, \quad (11.40)$$

where $S_j(s)$ is a “smoothing parameter” which depends on wave number. The smoothing parameter appears in the term of (11.38) corresponding to the zonal pressure gradient force, and in the term of (11.40) corresponding to the zonal mass flux divergence. These are the key terms for zonally propagating gravity waves. We have introduced $S_j(s)$ artificially in (11.39) and (11.40); later in this discussion it will be set to values less than unity in order to allow computational stability with a “large” time step near the pole. For now, just consider it to be equal to one.

By eliminating u and v in (11.38)-(11.40), we can obtain the “meridional structure equation” for \hat{h} . It is

$$\begin{aligned} & c^2 \left[\frac{s}{a \cos \phi_j} \frac{\sin(s\Delta\lambda/2)}{s\Delta\lambda/2} S_j(s) \right]^2 \hat{h}_j \\ & + \frac{c^2}{(a\Delta\phi)^2} \left[(\hat{h}_j - \hat{h}_{j-1}) \frac{\cos \phi_{j-\frac{1}{2}}}{\cos \phi_j} - (\hat{h}_{j+1} - \hat{h}_j) \frac{\cos \phi_{j+\frac{1}{2}}}{\cos \phi_j} \right] = \sigma^2 \hat{h}_j, \end{aligned} \quad (11.41)$$

where $c^2 \equiv gH$. By using the boundary condition $\hat{h}_j = 0$ at the poles, this equation can be solved as an eigenvalue problem for the frequency, σ . For high values of the zonal wave number s , the first term on the left-hand side of (11.41) dominates the second, and we obtain

$$\begin{aligned} \sigma &= |c| \left[\frac{s}{a \cos \phi_j} \frac{\sin(s\Delta\lambda/2)}{s\Delta\lambda/2} S_j(s) \right] \\ &= 2|c| \frac{S_j(s) \sin\left(\frac{s\Delta\lambda}{2}\right)}{a \cos \phi_j \Delta\lambda}. \end{aligned} \quad (11.42)$$

Although we have not introduced a specific time differencing scheme here, we know that the CFL criterion takes the form

$$\sigma \Delta t < \varepsilon, \quad (11.43)$$

where ε is a constant of order one. In view of (11.42) and (11.43), the CFL criterion will place more stringent conditions on Δt as $\cos \phi_j$ decreases, i.e. near the poles. In addition, the criterion becomes more stringent as s increases, at a given latitude. Putting $S_j(s) = 1$ temporarily, and assuming $\varepsilon = 1$, we can write the stability condition for zonal wave number s as

$$\frac{|c|\Delta t}{a \cos \phi_j \Delta \lambda} \sin\left(\frac{s\Delta \lambda}{2}\right) < \frac{1}{2} \quad (11.44)$$

The worst case is $\sin\left(\frac{s\Delta \lambda}{2}\right) = 1$, for which (11.44) reduces to

$$\frac{|c|\Delta t}{\Delta x_j} < \frac{1}{2} \quad (11.45)$$

where we define $\Delta x_j \equiv a \cos \phi_j \Delta \lambda$. For the grid shown in Fig. 11.4, with a longitudinal grid spacing of $\Delta \lambda = 5^\circ$ and a latitudinal grid spacing of $\Delta \phi = 4^\circ$ (which are the values used to draw the figure), the northernmost row of grid points where the u component of velocity is defined is at latitude 86°N . The zonal distance between grid points on the northernmost row is then $\Delta x = 39 \text{ km}$. The fast, external gravity wave has a phase speed of approximately 300 m s^{-1} . Substituting into (11.45), we find that the largest permissible time step near the pole is about 70 seconds. This is roughly one tenth of the largest permissible time step at the Equator.

It would be nice if the CFL criterion was 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. One 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.

Inspection of (11.42) shows that this can be accomplished by choosing the smoothing parameter so that

$$\frac{S_j(s) \sin\left(\frac{s\Delta \lambda}{2}\right)}{a \cos \phi_j \Delta \lambda} = \frac{1}{d^*}, \quad (11.46)$$

where d^* is a suitably chosen length, comparable to the zonal grid spacing at the Equator. With the use of (11.46), (11.42) reduces to

$$\sigma = \frac{2|c|}{d^*}, \quad (11.47)$$

and the CFL condition reduces to

$$\frac{2|c|}{d^*} \Delta t < \varepsilon, \quad (11.48)$$

so that the time step required is independent of latitude, as desired. If we choose

$$d^* \equiv a \Delta \varphi, \quad (11.49)$$

i.e. the distance between grid points in the meridional direction, then, referring back to (11.46), we see that $S_j(s)$ must be chosen so that

$$S_j(s) = \left(\frac{\Delta \lambda}{\Delta \varphi} \right) \frac{\cos \varphi_j}{\sin\left(\frac{s \Delta \lambda}{2}\right)}. \quad (11.50)$$

Of course, at low latitudes (11.50) can give values of $S_j(s)$ which are greater than one; these should be replaced by one, so that we actually use

$$S_j(s) = \text{Min} \left\{ \left(\frac{\Delta \lambda}{\Delta \varphi} \right) \frac{\cos \varphi_j}{\sin\left(\frac{s \Delta \lambda}{2}\right)}, 1 \right\}. \quad (11.51)$$

A plot of (11.51) is given in Fig. 11.5, for the case of the shortest zonal mode. The plot shows that some smoothing is needed all the way down into the subtropics.

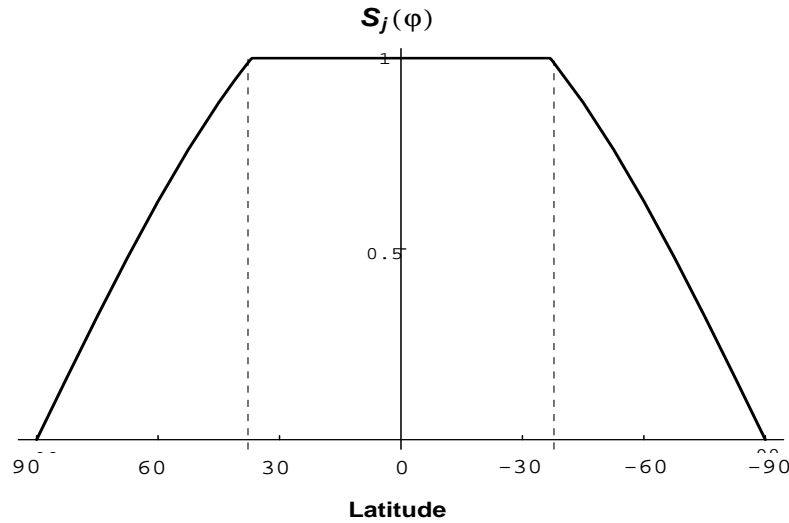


Figure 11.5: A plot of the smoothing parameter as given by (11.51), for the “worst case” of the shortest zonal mode. The dashed vertical lines demarcate the belt of latitude near the Equator for which no smoothing is needed. It has been assumed that the longitudinal grid spacing is 5/4 times the latitudinal grid spacing, as it is for the grid shown in Fig. 11.4.

11.4 Kurihara's 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 0° 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. 11.6; compare with Fig. 11.4. 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, $a \cos \phi_j \Delta \lambda_j$, and the meridional distance $a \Delta \phi$, for a grid point at latitude ϕ_j . Here,

$\Delta \phi \equiv \frac{\pi}{2N}$ and $\Delta \lambda_j \equiv \frac{\pi}{2j-1}$. At $j = N + 1$, the Equator, the ratio is one, and near the pole

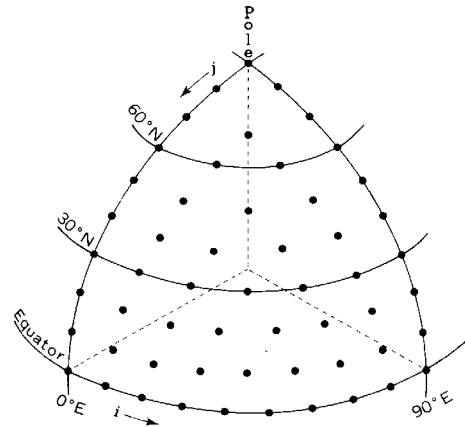


Figure 11.6: Kurihara grid on one octant of the sphere.

the ratio approaches $\pi/2 \cong 1.57$.

Kurihara built a model using this grid, based on the shallow water equations. He tested the model with the Rossby-Haurwitz wave, with wave number 4 as the initial condition. This set of initial conditions was used by Phillips (1959), and 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 wave should move from west to east, without distortion. In several of Kurihara's runs, however, the wave degenerated to higher wave numbers.

11.5 The Wandering Electron Grid

An approach to 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, it will move to maximize the distance between it and its closest neighbors. In this way, the electrons will distribute themselves as evenly as possible over the sphere. We associate a grid point with each electron. It seems advantageous to constrain the grid so that it is symmetric across the Equator. An Equator can be defined by restricting the movement of a subset of the electrons to 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 we fix the positions of some of the electrons, their positions tend to wander indefinitely. Fig. 11.7 shows a grid constructed using the wandering electron algorithm. Most cells have six walls, but some have five or seven walls. While this approach more or less homogeneously covers the sphere, it is not very satisfactory.

11.6 Spherical geodesic grids

Grids based on icosahedra offer an attractive framework for simulation of the global 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 (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

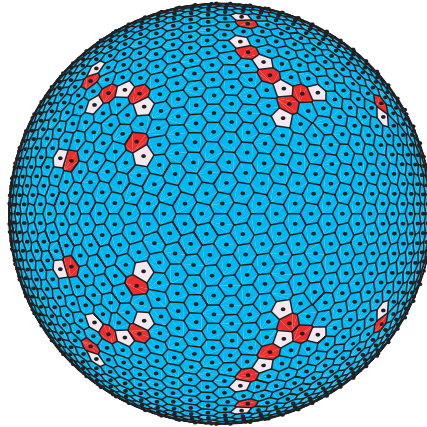


Figure 11.7: Wandering electron grid. White cells have five walls, light gray cells have six walls, and dark gray cells have seven walls.

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 (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. 11.8. One can

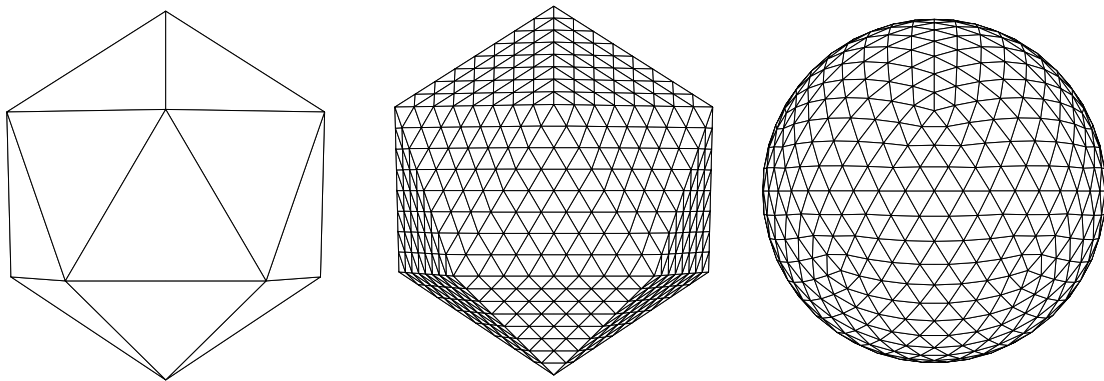


Figure 11.8: a.) Icosahedron. b.) Partition each face into 64 smaller triangles. c.) Project onto the sphere.

construct a more homogeneous grid by partitioning the spherical equilateral triangles instead. Williamson (1968) and Sadourny (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. Hexagonal grids are

quasi-isotropic. As is well known, only three regular polygons tile the plane: equilateral triangles, squares, and hexagons. Fig. 11.9 shows planar grids made up of each of these three

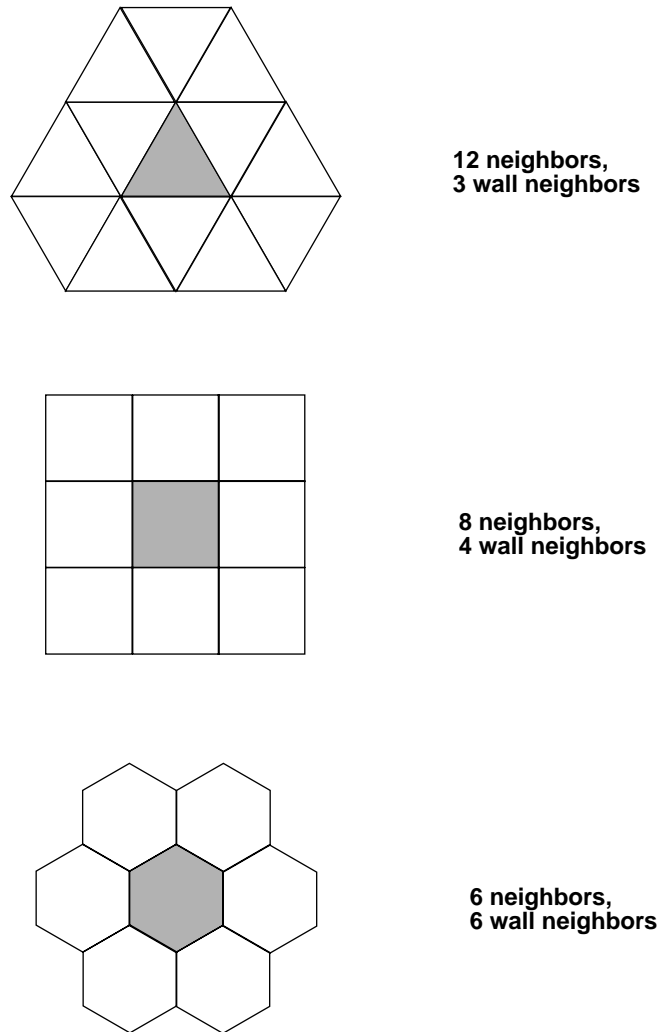


Figure 11.9: Cells neighboring a given cell (shaded) on triangular, square, and hexagonal grids. A “wall neighbor” is a neighbor which lies directly across a cell wall.

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 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 neighboring cells is used [e.g. the Arakawa Jacobian, as discussed by Arakawa (1966)], the essential anisotropies of these grids remain, and are unavoidably manifested in the forms of the finite-difference operators. Hexagonal grids, in contrast, have the property that all neighbors of a

given cell lie across cell walls; there are no “vertex neighbors.” As a result, finite-difference operators constructed on hexagonal grids treat all neighboring cells in the same way; in this sense, the operators are as symmetrical and isotropic as possible. The twelve pentagonal cells also have only wall neighbors; there are no vertex neighbors anywhere on the sphere.

Williamson (1968) chose the nondivergent shallow water equations to test the new grid, i.e. he used

$$\frac{\partial \zeta}{\partial t} = -J(\psi, \eta), \quad (11.52)$$

where ζ is relative vorticity, $\eta = \zeta + f$ is absolute vorticity and ψ is the stream function, such that

$$\nabla^2 \psi = \zeta. \quad (11.53)$$

Recall that, for arbitrary functions α and β , the Jacobian in differential form has the property that

$$\int_A J(\alpha, \beta) dA = \oint_S \alpha \frac{\partial \beta}{\partial s} ds. \quad (11.54)$$

So, integrating (11.52) over the area A , we get

$$\frac{\partial}{\partial t} \int_A \zeta dA = - \oint_S u_n \eta ds, \quad (11.55)$$

where $u_n = -\frac{\partial \psi}{\partial s}$ is the velocity normal to the boundary S .

Consider K triangles surrounding the grid point P_0 in Fig. 11.10. We approximate the line integral along the polygon defined by the path $P_1, P_2, \dots, P_5, P_1$. Let ζ_0 be the relative vorticity defined at the point P_0 . First, we make the approximation $\int_A \zeta dA \approx \zeta_0 A$. Let η_i and η_{i+1} be the absolute vorticity defined at points P_i and P_{i+1} , respectively. We approximate the absolute vorticity along the edge between P_i and P_{i+1} by $(\eta_i + \eta_{i+1})/2$. Similarly, $\partial \psi / \partial s \approx (\psi_{i+1} - \psi_i) / \Delta s$, where Δs is the distance from P_i to P_{i+1} . Hence, we can approximate (11.55) by

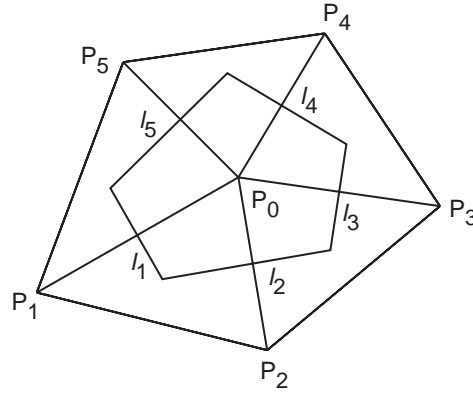


Figure 11.10: Configuration of grid triangles for the case $K = 5$.

$$\frac{\partial \zeta_0}{\partial t} = \frac{1}{A} \sum_{i=1}^K \left(\frac{\psi_{i+1} - \psi_i}{\Delta s} \right) \left(\frac{\eta_{i+1} + \eta_i}{2} \right) \Delta s. \quad (11.56)$$

To solve equation (11.53), we must discretize the Laplacian operator. Consider the smaller, inner polygon in Fig. 11.10. The walls of the polygon are formed from the perpendicular bisectors of the line segments $\overline{P_0P_i}$. For any arbitrary scalar function α , we can use Gauss' Theorem to write

$$\int_a \nabla^2 \alpha da = \oint_{s'} \frac{\partial \alpha}{\partial n} ds, \quad (11.57)$$

where a is the area of the small polygon, and s' is its boundary. Using (11.53) and (11.57), we get

$$\Gamma = \oint_{s'} u_s ds, \quad (11.58)$$

where Γ is the circulation around the boundary and u_s is the counterclockwise tangential velocity. We now set $\Gamma \equiv a \zeta_0$. We assume that the tangential velocity $u_s = \partial \psi / \partial n$ is approximately constant along each wall of s' , and can be approximated by $(\psi_i - \psi_0) / |P_0P_i|$, where $|P_0P_i|$ is the distance from P_0 to P_i . Define $\omega_i \equiv l_i / |P_0P_i|$, where l_i is the length of wall i . The resulting finite-difference approximation is

$$\zeta_0 = \sum_{i=1}^K \omega_i (\psi_i - \psi_0). \quad (11.59)$$

This is a finite-difference approximation to (11.53). It can be solved for the ψ_i by relaxation, as discussed in earlier.

Williamson showed that his scheme conserves kinetic energy and enstrophy as the exact equations do. When applied to an equilateral triangular 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. In a later study, Williamson (1971) described a finite-difference scheme that is second-order accurate on the spherical geodesic grid, but lacks certain desirable conservation properties.

Sadourny (1968) discussed a nondivergent model very similar to Williamson's. Also, Sadourny (1969) developed a geodesic-grid model based on the shallow water equations.

Masuda (1986) developed an elegant algorithm for solving the shallow water equations on the sphere. He used the Z grid (see Chapter 6). Like Williamson, Masuda chose the Rossby-Haurwitz wave with wave number 4 as his initial condition. Fig. 11.11 shows the

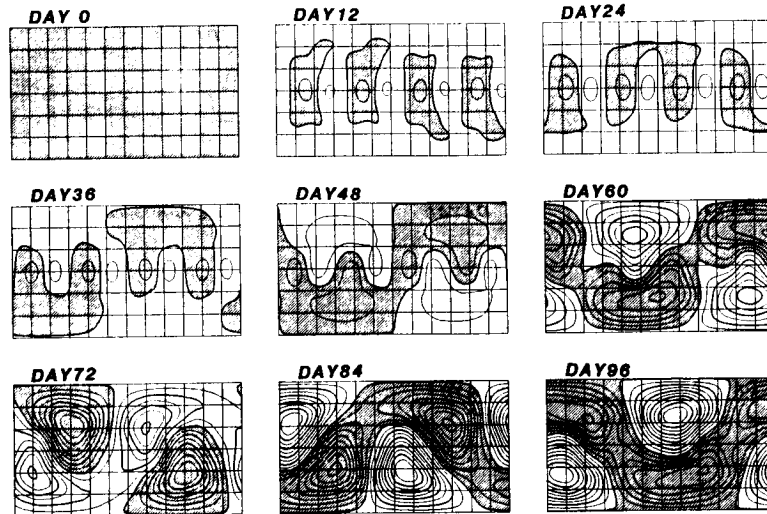


Figure 11.11: Masuda's velocity potential field.

evolution of the velocity potential field in a 96-simulated-day run using Masuda's model. The initial conditions are nondivergent, so initially the velocity potential is zero. As time progresses, a wave number 4 pattern develops. As time progresses further, the pattern at higher latitudes begins to break down, forming a wave number 1 pattern. Significantly, the wave number 1 pattern is antisymmetric across the Equator, even though the initial condition is symmetric across the Equator. Masuda suggested that this is due to the antisymmetry of the grid across the Equator.

Heikes and Randall (1995 a, b) extended Masuda's work, through the use of a "twisted icosahedral grid" that has symmetry across the equator. They used a multi-grid method to compute the stream function and velocity potential from the vorticity and divergence, respectively. Heikes and Randall (1995 b) also showed that the grid (whether twisted or not) has to be slightly altered to permit consistent finite-difference approximation 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. (1999) have constructed a full-physics global atmospheric model using this approach.

There have also been recent attempts to use grids based on cubes and octahedrons (e.g. McGregor, 1996; Purser and Rancic, 1998).

11.7 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 can come close.

A third approach, discussed in the next chapter, is to use the spectral method, with spherical harmonics as the basis functions.

CHAPTER 12 | Spectral Methods

Copyright 2004 David A. Randall

12.1 Introduction

Assume that $u(x, t)$ is real and integrable. If the domain is periodic, with period L , we can express $u(x, t)$ exactly by a Fourier series expansion:

$$u(x, t) = \sum_{k=-\infty}^{\infty} \hat{u}_k(t) e^{ikx}. \quad (12.1)$$

The complex coefficients $u_k(t)$ can be evaluated using

$$\hat{u}_k(t) = \frac{1}{L} \int_{x-L/2}^{x+L/2} u(x, t) e^{-ikx} dx. \quad (12.2)$$

Recall that the proof of (12.1) and (12.2) involves use of the orthogonality condition

$$\frac{1}{L} \int_{x-L/2}^{x+L/2} e^{-ikx} e^{ilx} dx = \delta_{kl}, \quad (12.3)$$

where

$$\delta_{kl} \equiv \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases} \quad (12.4)$$

is the Kronecker delta.

From (12.1), we see that the x -derivative of u satisfies

$$\frac{\partial u}{\partial x}(x, t) = \sum_{k=-\infty}^{\infty} i k \hat{u}_k(t) e^{ikx}. \quad (12.5)$$

Inspection of (12.5) shows that $\frac{\partial u}{\partial x}$ does not have a contribution from u_0 ; the reason for this should be clear.

A spectral model uses equations similar to (12.1), (12.2), and (12.5), but with a finite set of wave numbers, and with x defined on a finite mesh:

$$u(x_j, t) \cong \sum_{k=-n}^n \hat{u}_k(t) e^{ikx_j}, \quad (12.6)$$

$$\hat{u}_k(t) \cong \frac{1}{M} \sum_{j=1}^M u(x_j, t) e^{-ikx_j}, \quad -n \leq k \leq n, \quad (12.7)$$

$$\frac{\partial u}{\partial x}(x_j, t) \cong \sum_{k=-n}^n i k \hat{u}_k(t) e^{ikx_j}. \quad (12.8)$$

Note that we have used “approximately equal signs” in (12.6) - (12.8). In (12.7) we sum over a grid with M points. In the following discussion, we assume that the value of n is chosen by the user. The value of M , corresponding to a given value of n , is discussed below.

Substitution of (12.6) into (12.7) gives

$$\hat{u}_k(t) = \frac{1}{M} \sum_{j=1}^M \left\{ \left[\sum_{l=-n}^n \hat{u}_l(t) e^{ilx_j} \right] e^{-ikx_j} \right\}, \quad -n \leq k \leq n. \quad (12.9)$$

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 (12.9) 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 (12.9); the range for wave number k is understood because, as indicated, we wish to evaluate the left-hand side of (12.9) for k in the range $-n$ to n . Because each term on the right-hand side of (12.9) 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 $u_k(t)$.

Because u is real, it must be true that $\hat{u}_{-k} = \hat{u}_k^*$, where the $*$ denotes the conjugate. To see why this is so, consider the $+k$ and $-k$ contributions to the sum in (12.6):

$$\begin{aligned} T_k(x_j) &\equiv \hat{u}_k(t)e^{ikx_j} + \hat{u}_{-k}(t)e^{-ikx_j} \\ &\equiv R_k e^{i\theta} e^{ikx_j} + R_{-k} e^{i\mu} e^{-ikx_j}. \end{aligned} \quad (12.10)$$

where $R_k e^{i\theta} \equiv \hat{u}_k(t)$ and $R_{-k} e^{i\mu} \equiv \hat{u}_{-k}(t)$, and R_k and R_{-k} are real and non-negative. By linear independence, our assumption that $u(x_j, t)$ for all x_j is real implies that the imaginary part of $T_k x_j$ must be zero, for all x_j . It follows that

$$R_k \sin(\theta + kx_j) + R_{-k} \sin(\mu - kx_j) = 0 \text{ for all } x_j. \quad (12.11)$$

The only way to satisfy this for all x_j is to set

$$\theta + kx_j = -(\mu - kx_j) = -\mu + kx_j, \text{ which means that } \theta = -\mu, \quad (12.12)$$

and

$$R_k = R_{-k}. \quad (12.13)$$

Eqs. (12.12) and (12.13) imply that

$$\hat{u}_{-k} = \hat{u}_k^*, \quad (12.14)$$

as was to be demonstrated.

Eq. (12.14) implies that u_k and u_{-k} together involve only two distinct real numbers. In addition, it follows from (12.14) that u_0 is real. Therefore, the $2n + 1$ complex values of u_k embody the equivalent of only $2n + 1$ distinct real numbers. The Fourier representation up to wave number n is thus equivalent to representing the real function $u(x, t)$ on $2n + 1$ 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 more than $2n + 1$ points, but not fewer.

As a very simple example, a highly truncated Fourier representation of u including just wave numbers zero and one is equivalent to a grid-point representation of u using 3 grid points. The real values of u assigned at the three grid points suffice to compute the coefficient of wave number zero (the mean value of u) and the phase and amplitude (or “sine and cosine coefficients”) of wave number one.

Substituting (12.7) into (12.8) gives

$$\frac{\partial u}{\partial x}(x_l, t) \equiv \sum_{k=-n}^n \left[\frac{ik}{M} \sum_{j=1}^M u(x_j, t) e^{-ikx_j} \right] e^{ikx_l} . \quad (12.15)$$

Reversing the order of summation leads to

$$\frac{\partial u}{\partial x}(x_l, t) \equiv \sum_{j=1}^M \alpha_j^l u(x_j, t) , \quad (12.16)$$

where

$$\alpha_j^l \equiv \frac{i}{M} \sum_{k=-n}^n k e^{ik(x_l - x_j)} . \quad (12.17)$$

The point of this little derivation is that (12.16) can be interpreted as a finite-difference approximation - a special one involving *all* grid points in the domain. From this point of view, spectral models can be regarded as a class of finite-difference models.

Now consider the one-dimensional advection equation with a constant current, c :

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0 . \quad (12.18)$$

Substituting (12.6) and (12.8) into (12.18) gives

$$\sum_{k=-n}^n \frac{d\hat{u}_k}{dt} e^{ikx} + c \sum_{k=-n}^n ik \hat{u}_k e^{ikx} = 0 . \quad (12.19)$$

By linear independence, we obtain

$$\frac{d\hat{u}_k}{dt} + ikc \hat{u}_k = 0 \text{ for } -n \leq k \leq n . \quad (12.20)$$

Note that $\frac{\hat{du}_0}{dt}$ will be equal to zero; the interpretation of this should be clear. We can use (12.20) to predict $u_k(t)$. When we need to know $u(x, t)$, we can get it from (12.6).

Compare (12.20) with

$$\frac{d\hat{u}_k}{dt} + ikc\left(\frac{\sin k\Delta x}{k\Delta x}\right)\hat{u}_k = 0, \quad (12.21)$$

which, as we have seen, 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. Similarly, spectral methods give the *exact* phase speeds for linear waves, 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.

To evaluate the horizontal pressure gradient force, it is necessary to take horizontal derivatives of the terrain height. Suppose that we have continents and oceans, as schematically shown in Fig. 12.1. In a spectral model the terrain heights will have to be expanded and

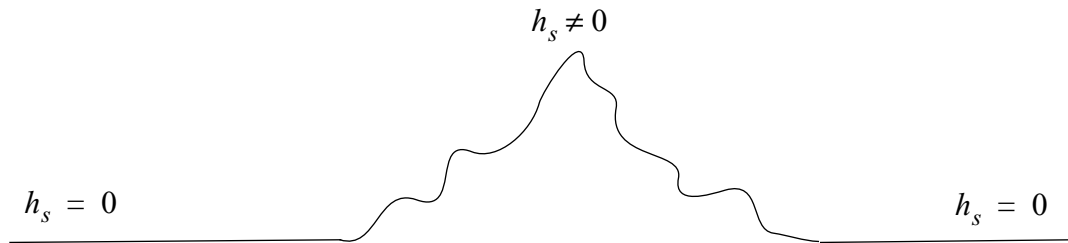


Figure 12.1: The Earth is bumpy.

truncated, like all of the other variables. Truncation leads to “bumpy” oceans. Recently some new approaches have been suggested to alleviate this problem (Boutelou, 1995; Holzer, 1996; Lindberg and Broccoli, 1996).

Another strength of spectral methods is that they make it very easy to solve boundary value problems. As an example, consider

$$\nabla^2 u = f(x, y), \quad (12.22)$$

as a problem to determine u for given $f(x, y)$. In one dimension, (12.22) becomes

$$\frac{d^2 u}{dx^2} = f(x). \quad (12.23)$$

We assume periodic boundary conditions and expand both u and f as Fourier series in x , following (12.1). Then (12.23) becomes

$$\sum_{k=-n}^n (-k^2) \hat{u}_k e^{ikx} = \sum_{k=-n}^n \hat{f}_k e^{ikx}. \quad (12.24)$$

Equating coefficients of e^{ikx} , we find that

$$\hat{u}_k = \frac{\hat{f}_k}{k^2} \quad \text{for } -n \leq k \leq n \quad (\text{unless } k = 0). \quad (12.25)$$

Eq. (12.25) can be used to obtain \hat{u}_k , for $k = 1, n$. Then $u(x)$ can be constructed using (12.1). This completes the solution of (12.23), apart from the application of an additional boundary condition to determine u_0 . The solution is exact *for the modes that are included*; it is approximate because not all modes are included.

Now consider a nonlinear problem, such as

$$\frac{\partial u}{\partial t} = -u \frac{\partial u}{\partial x}, \quad (12.26)$$

again with a periodic domain. Substitution gives

$$\sum_{k=-n}^n \frac{d\hat{u}_k}{dt} e^{ikx} = - \left(\sum_{l=-n}^n \hat{u}_l e^{ilx} \right) \left(\sum_{m=-n}^n i m \hat{u}_m e^{imx} \right). \quad (12.27)$$

Our goal is to predict $u_k(t)$ for k in the range $-n$ to n . The right-hand-side of (12.27) involves products of the form

$$e^{ilx} e^{imx}, \quad (12.28)$$

where l and m are 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 and make use of the left-hand side of (12.27).

For a given Fourier mode, (12.27) implies that

$$\frac{d\hat{u}_k}{dt} = \sum_{l=-\alpha}^{\alpha} \sum_{m=-\alpha}^{\alpha} [\hat{u}_l \hat{u}_m e^{i(l+m)x}] e^{-ikx}, \quad -u \leq k \leq n \quad (12.29)$$

Here we must choose α large enough so that we pick up all possible combinations of l and m that lie in the range $-n$ to n . See Fig. 12.2. The circled X's in the figure denote

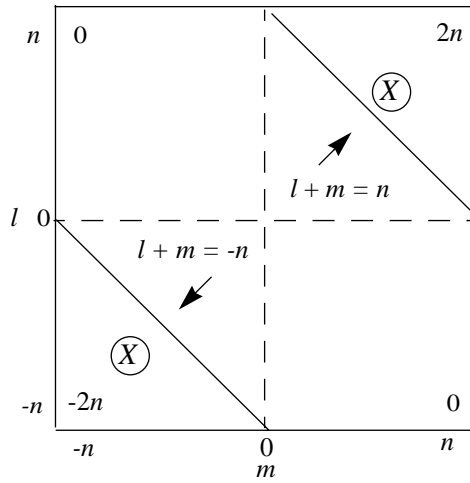


Figure 12.2: Table of $l + m$, showing which (l, m) pairs can contribute to wave numbers in the range $-n$ to n . The pairs in the triangular regions marked by X's do not contribute.

excluded triangular regions. The number of points in each region is

$$1 + 2 + 3 + \dots + (n-1) = \frac{n(n-1)}{2}. \quad (12.30)$$

The number of points *retained* is

$$\begin{aligned} & (2n+1)^2 - 2 \left[\frac{n(n-1)}{2} \right] \\ &= (4n^2 + 4n + 1) - (n^2 - n) \\ &= 3n^2 + 5n + 1 \end{aligned} \quad (12.31)$$

This is the number of terms that must be evaluated in (12.27). The number of terms in the product of sums on the right-hand-side of (12.27) is of order n^2 , i.e. it grows very rapidly as n increases. 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.

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 (12.1) and (12.5) are used to evaluate u and $\frac{\partial u}{\partial x}$ on a grid. Sufficient grid points are used 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 (12.7), we can write

$$\left(\hat{u} \frac{\partial u}{\partial x} \right)_k = \frac{1}{M} \sum_{j=1}^M \left[u(x_j) \frac{\partial u}{\partial x}(x_j) e^{-ikx_j} \right], -n \leq k \leq n. \quad (12.32)$$

Here the hat on $\left(u \frac{\partial u}{\partial x} \right)_k$ indicates that the entire quantity is represented in wave-number space rather than grid space. Now use (12.6) and (12.8) to express $u(x_j)$ and $\frac{\partial u}{\partial x}(x_j)$ in terms of Fourier series:

$$\left(\hat{u} \frac{\partial u}{\partial x} \right)_k = \frac{1}{M} \sum_{j=1}^M \left[\left(\sum_{l=-n}^n \hat{u}_l e^{ilx_j} \right) \left(\sum_{m=-n}^n i m \hat{u}_m e^{imx_j} \right) e^{-ikx_j} \right], -n \leq k \leq n. \quad (12.33)$$

Eq. (12.33) is analogous to (12.9). When expanded, each term on the right-hand side of (12.33) 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 $\pm 3n$, and because $u \frac{\partial u}{\partial x}$ is real these $3n + 1$ complex coefficients actually correspond to $3n + 1$ independent real numbers. We therefore need

$$M \geq 3n + 1 \quad (12.34)$$

grid points to represent $u \frac{\partial u}{\partial x}$ exactly, up to wave number n .

In summary, the transform method to solve (12.26) works as follows:

- 1) Initialize the spectral coefficients u_k , for $-n \leq k \leq n$.
- 2) Evaluate both u and $\frac{\partial u}{\partial x}$ on a grid with M points, where $M \geq 3n + 1$. Here $\frac{\partial u}{\partial x}$ is computed using the spectral method, i.e. Eq. (12.8).
- 3) Form $u \frac{\partial u}{\partial x}$ on the grid, by multiplication.
- 4) Transform $u \frac{\partial u}{\partial x}$ back into wave-number space, for $-n \leq k \leq n$.
- 5) Predict new values of the u_k , using

$$\frac{d\hat{u}_k}{dt} = -\left(u \frac{\partial u}{\partial x}\right)_k.$$

- 1) 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 problem 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. It 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 aliasing due to quadratic nonlinearities.

12.2 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

$$F(\lambda, \phi) = \sum_{m=-\infty}^{\infty} \sum_{n=|m|}^{\infty} F_n^m Y_n^m(\lambda, \phi), \quad (12.35)$$

where the

$$Y_n^m(\lambda, \phi) = e^{im\lambda} P_n^m(\sin \phi) \quad (12.36)$$

are spherical harmonics, and the $P_n^m(\sin \phi)$ are the associated Legendre functions of the first kind, which happen to be polynomials, satisfying

$$\begin{aligned} P_n^m(x) = & \frac{(2n)!}{2^n n! (n-m)!} (1-x^2)^{\frac{m}{2}} \left[x^{n-m} - \frac{(n-m)(n-m-1)}{2(2n-1)} x^{n-m-2} \right. \\ & \left. + \frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2 \cdot 4(2n-1)(2n-3)} x^{n-m-4} - \dots \right], \end{aligned} \quad (12.37)$$

Here m is the zonal wave number and n is the “meridional nodal number.” As discussed in the Appendix on spherical harmonics, we must require that $n \geq m$. The spherical harmonics Y_n^m are the eigenfunctions of the Laplacian on the sphere:

$$\nabla^2 Y_n^m = \frac{-n(n+1)}{a^2} Y_n^m. \quad (12.38)$$

Here a is the radius of the sphere. See the Appendix for further discussion.

We can approximate F by a truncated sum:

$$\bar{F} = \sum_{m=-M}^M \sum_{n=|m|}^{N(m)} F_n^m Y_n^m. \quad (12.39)$$

Here the overbar indicates that \bar{F} is an approximation to F . In (12.39), 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. Only a few terms are needed to obtain a good representation.

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. What is so special about the Laplacian operator? There are infinitely many differential operators, so why choose the Laplacian? A justification is that:

- the Laplacian consumes scalars and returns scalars, unlike, for example, the gradient, the curl, or the divergence;
- the Laplacian can be defined without reference to any coordinate system;
- 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 the problem of truncation. The two best-known possibilities are *triangular truncation* and *rhomboidal truncation*:

$$\text{Rhomboidal: } N - |m| = M = \text{constant} \quad (12.40)$$

$$\text{Triangular: } N = M = \text{constant} , \text{ or } N - |m| = M - |m| . \quad (12.41)$$

These are illustrated in Fig. 12.3. As shown in Fig. 12.4, triangular truncation represents the *observed* kinetic energy spectrum more accurately, with a small number of terms, than does rhomboidal truncation (Baer, 1972). The thin lines in Fig. 12.4 show the modes kept with triangular truncation. With rhomboidal truncation the thin lines would be horizontal. The thick lines show the observed kinetic energy percentage in each component. For example, we might want to truncate so that we keep all modes with $\geq 0.01\%$ of the kinetic energy, and discard all others. Triangular truncation can do that.

In addition, triangular truncation has the beautiful property that it is not tied to a coordinate system. Here is what this means: In order to actually perform a spherical harmonic transform, it is necessary to adopt a spherical coordinate system (λ, φ) . There are of course infinitely many such systems, which differ in the orientations of their poles. There is no reason in principle that the coordinates have to be chosen in the conventional way, so that the poles of the coordinate system coincide with the Earth's poles of rotation. The choice of a particular spherical coordinate system is, therefore, somewhat arbitrary. Suppose that we choose two different spherical coordinate systems (tilted with respect to one another in an arbitrary way), perform a triangularly truncated expansion in both, then plot the results. It can be shown that the two maps will be identical, i.e.

$$F(\lambda_1, \varphi_1) = F(\lambda_2, \varphi_2) , \quad (12.42)$$

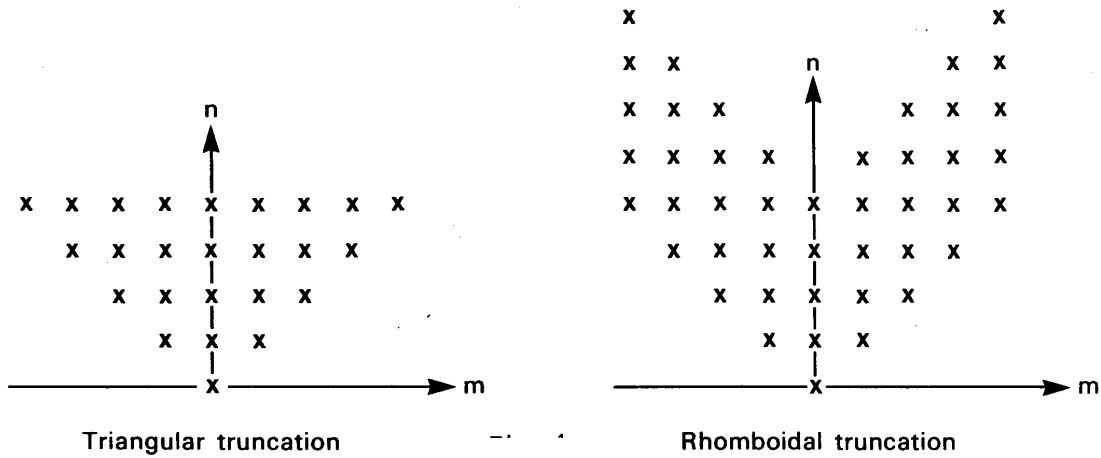


Figure 12.3: Rhomboidal and triangular truncation. From Jarraud and Simmons (1983).

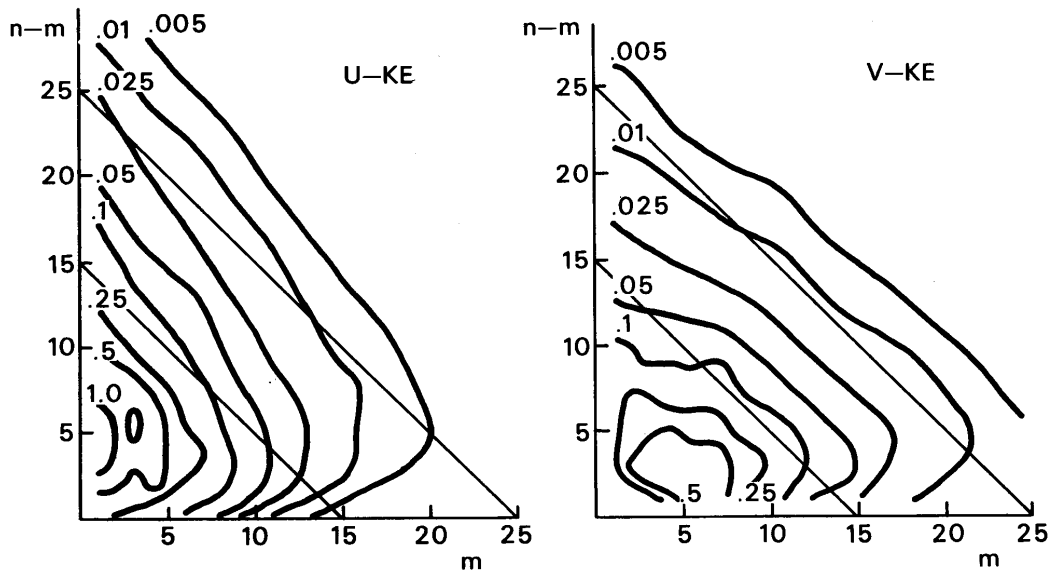


Figure 12.4: Percentage of total kinetic energy in each spectral component. From Jarraud and Simmons (1983) based on Baer (1972).

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.

In order to use (12.39), we need a “spherical harmonic transform,” analogous to a Fourier transform. From (12.36), 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 method called “Gaussian quadrature.” The idea is as follows. Suppose that we are given a function $f(x)$ defined on the interval $-1 \leq x \leq 1$, and we wish to evaluate

$$I = \int_{-1}^1 f(x) dx, \quad (12.43)$$

by a numerical method. If f is defined at a finite number of x , denoted by x_i , then

$$I \cong \sum_{i=1}^N f(x_i) w_i, \quad (12.44)$$

where the w_i are “weights.” Now suppose that $f(x)$ is a weighted sum of Legendre polynomials. Gauss showed that in that case (12.44) gives the *exact* value of I , provided that the x_i are chosen to be the roots of the highest Legendre polynomial used. In other words, we can use (12.44) to evaluate the integral (12.43) 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. The Gaussian quadrature algorithm is used to perform the Legendre transform.

With *either* triangular *or* rhomboidal truncation, choosing M fixes the expansion; hence the expressions “R15” or “T106.” The numeral is the value of M . The numbers of complex coefficients needed are

$$f_R = (M+1)^2 + M^2 + M, \quad (12.45)$$

and

$$f_T = (M+1)^2, \quad (12.46)$$

respectively.

With the transform method described earlier, the number of grid points needed to avoid aliasing of quadratic nonlinearities exceeds the number of degrees of freedom in the spectral representation. The number of grid points around a latitude circle must be

$\geq 3M + 1$. The number of latitude circles must be $\geq \frac{(3M+1)}{2}$ for triangular truncation, and so the total number of grid points needed is $\geq \frac{(3M+1)^2}{2}$. Referring back to (12.46), 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. Computing the physics on the fine grid is standard procedure, but wasteful.

In summary, the spectral transform method as it is 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 same as 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 $O(N^3)$, where N is the number of spherical harmonics used. Finite difference methods are, in effect, of $O(N^2)$. This means that spectral models become increasingly expensive, relative to grid-point models, at high resolution.

12.3 The “equivalent grid resolution” of spectral models

Laprise (1992) distinguishes four possible ways to answer the following obvious 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 $L_1 \cong 425\text{km}$. 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, $L_2 \cong 650\text{km}$.
- 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 this is the measure that spectral modelers almost always use when specifying the equivalent grid spacing of their models.

12.4 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 8. 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.

12.5 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, the models do not exactly conserve anything -- 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.

12.6 Moisture advection

The mixing ratio of water vapor is non-negative. We have already discussed the possibility of spurious negative mixing ratios caused by dispersion errors in finite-difference schemes, and we have 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 occur not only locally on individual time steps, but even in zonal averages that have been time-averaged over a month.

Because of this disastrous situation, many spectral models are now using non-spectral

methods for advection (e.g. Williamson and Olson, 1994). This means that they are only “partly spectral.” 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. This leads to a major savings in computational cost.

12.7 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 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 adjust convectively unstable columns so as to remove the instability. Suppose 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 can be solved by doing the physics on a “linear” grid that has the same number of degrees of freedom as the spectral representation.

12.8 Summary

In summary, the spectral method has both strengths and weaknesses:

Strengths:

- Especially with triangular truncation, it eliminates the “pole problem.”
- It gives the exact phase speeds for linear waves and advection by a constant current such as solid-body rotation.
- It converges very rapidly, and gives 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.

Problems

1. Write subroutines to compute Fourier transforms and inverse transforms, for arbitrary complex $u(x_j)$. 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.

a) Let

$$u(x_j) = 14 \cos(k_0 x_j) + 6i \cos(k_1 x_j) + 5 \quad (12.47)$$

where

$$\begin{aligned} k_0 &= \frac{2\pi}{L_o}, \quad L_o = \frac{X}{4}, \\ k_1 &= \frac{2\pi}{L_1}, \quad L_1 = \frac{X}{8}. \end{aligned} \quad (12.48)$$

Compute the Fourier coefficients starting from values of x_j on a grid of M points, for each value of M in the angle.

$$2 \leq M \leq 20 \quad (12.49)$$

Tabulate u_k for $0 \leq M \leq 8$, and $2 \leq M \leq 20$. Discuss your results.

b) Repeat for

$$u(x_j) = 5 \cos k_0 x_j + 7 \sin k_1 x_j + 2. \quad (12.50)$$

Use the same values of k_0 and k_1 as given above.

CHAPTER 13**Boundary conditions and nested grids**Copyright 2004 David A. Randall

13.1 Introduction

Boundary conditions can be real or fictitious. At 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. Many models also have "fictitious walls" at their tops.

With some vertical coordinate systems (such as height), topography imposes a lateral boundary condition in an atmospheric model. Ocean models describe flows in basins, and so (depending again on the vertical coordinate system used) have "real" lateral boundary conditions.

Limited area models have artificial lateral boundaries. Even global models are limited-domain models in the sense that they have physical lower boundaries and artificial "lids."

Models in which the grid spacing changes rapidly (e.g. nested grid models) effectively apply boundary conditions where the dissimilar grids meet.

This chapter deals with the effects of boundary conditions, with emphasis on what can go wrong.

13.2 Inflow Boundaries

Consider advection in an artificially bounded domain. Instead of prescribing initial values of the advected quantity u for all x , we prescribe u at $x = 0$ for all time. In particular, we assume $u(t)$ at $x = 0$ as a simple harmonic function, with frequency ω . Note that we can choose ω as we please. Referring again to the advection equation, i.e.

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0, \quad (13.1)$$

we assume $c > 0$ and write

$$u(x, t) = \text{Re}[U(x)e^{-i\omega t}], \quad \omega \neq 0. \quad (13.2)$$

We prescribe $U(0)$ as a real constant. Then

$$u(0, t) = U(0)Re(e^{-i\omega t}) = U(0)\cos\omega t. \quad (13.3)$$

Since $c > 0$, we have effectively prescribed an “inflow” or “upstream” boundary condition.

Use of (13.3) in (13.1) gives

$$-i\omega U + c \frac{dU}{dx} = 0, \quad (13.4)$$

which has the solution

$$U(x) = U(0)e^{ikx}. \quad (13.5)$$

The dispersion equation is obtained by substituting (13.5) into (13.4):

$$\omega = ck. \quad (13.6)$$

The full solution is thus

$$\begin{aligned} u(x, t) &= U_0 Re\{exp[i(kx - \omega t)]\} \\ &= U_0 Re\{exp[ik(x - ct)]\}. \end{aligned} \quad (13.7)$$

Now consider the same problem again, this time as represented through the differential-difference equation

$$\frac{du_j}{dt} + c \left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} \right) = 0. \quad (13.8)$$

We assume a solution of the form

$$u_j = Re[U_j e^{-i\omega t}], \quad (13.9)$$

we obtain the now-familiar dispersion relation

$$\omega = ck \frac{\sin(k\Delta x)}{k\Delta x}. \quad (13.10)$$

This should be compared with (13.6). Fig. 13.1 gives a schematic plot, with $\omega\Delta x/c$ and $k\Delta x$ as coordinates, for the true dispersion equation (13.6) and the approximate dispersion equation (13.10). The results are plotted only out to $k\Delta x = \pi$, which corresponds to $L = 2\Delta x$, the shortest resolvable wave. *For a given ω* 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. Note that

$$k_2\Delta x = \pi - k_1\Delta x. \quad (13.11)$$

Also note that the group velocity is positive for $k\Delta x < \pi/2$, and negative for $k\Delta x > \pi/2$.

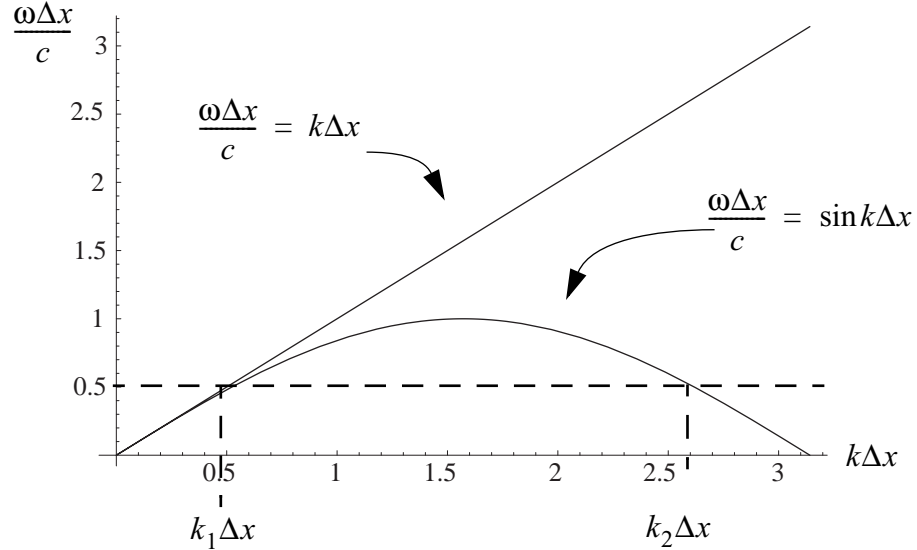


Figure 13.1: A schematic plot, with $\frac{\omega\Delta x}{c}$ and $k\Delta x$ as coordinates, for the true solution (13.3) and the approximate solution (13.10). The dashed line illustrates that for a given ω the approximate solution allows two different wave numbers.

When we studied the leapfrog time-differencing scheme in Chapter 3, we had a somewhat similar result. In Chapter 3 we had two solutions with different frequencies for a given wavelength. Here we have a single given frequency but two solutions of differing wavelength associated with that frequency. We can call the solution corresponding to k_1 a “physical mode” in space, and the solution corresponding to k_2 a “computational” mode in space. Notice that 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 $k_1\Delta x > \frac{\pi}{2}$, there really is no physical mode. In

other words, *the physical mode exists only for $L \geq 4\Delta x$* . In view of (13.12), $k_1\Delta x < \frac{\pi}{2}$, which is the requirement that a physical mode exists, corresponds to $\sin^{-1}\left(\frac{\omega\Delta x}{c}\right) < \frac{\pi}{2}$. This condition can be satisfied by choosing Δx small enough, for given values of ω and c .

Referring back to (13.10), we see that the two modes can be written as

$$\text{physical mode: } u_j \sim \text{Re} \left\{ \exp \left[i k_1 \left(j \Delta x - \frac{\omega}{k_1} t \right) \right] \right\}, \quad (13.12)$$

$$\begin{aligned} \text{computational mode: } u_j &\sim \text{Re} \left\{ \exp \left[i k_2 \left(j \Delta x - \frac{\omega}{k_2} t \right) \right] \right\} \\ &= \text{Re} \{ \exp [i(j\pi - k_1 j \Delta x - \omega t)] \} \\ &= (-1)^j \text{Re} \left\{ \exp \left[-i k_1 \left(j \Delta x + \frac{\omega}{k_1} t \right) \right] \right\}, \end{aligned} \quad (13.13)$$

where $k_2 \Delta x = \pi - k_1 \Delta x$ and $e^{ij\pi} = (-1)^j$ have been used, note that the two modes will have different amplitudes. The phase velocity of the computational mode is opposite to that of the physical mode, and it oscillates in space with wave length $2\Delta x$, due to the factor of $(-1)^j$.

In general, the solution is a superposition of the physical and computational modes. In case $c > 0$ and 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. 13.2 shows this schematically for some arbitrary time. The dashed line for $j < 0$ represents

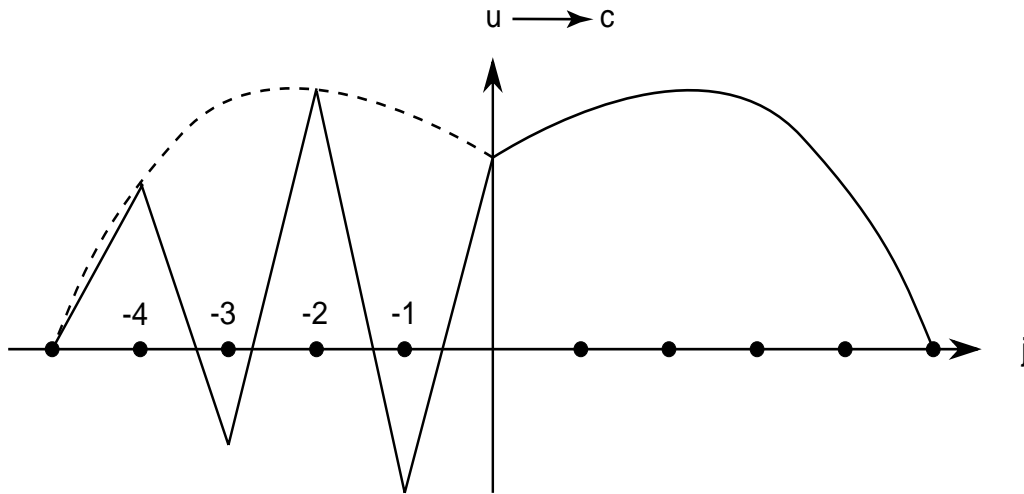


Figure 13.2: Schematic illustration of a computational mode that is restricted to the domain $j < 0$, and a physical mode that is restricted to the domain $j > 0$.

(13.13), 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 Δx), the computational mode will appear as an oscillation from point to point, i.e. a wave of length $2\Delta x$.

According to (13.10), the apparent phase change per grid interval, Ω , is related to the wave number by

$$\Omega \equiv \frac{\omega \Delta x}{c} = \sin(k\Delta x) \cong k\Delta x \text{ for } k\Delta x \ll 1. \quad (13.14)$$

With the exact equations, $\Omega = k\Delta x$. Since we control ω , Δx , and c , we effectively control Ω . Suppose that we give $\Omega > 1$, by choosing a large value of Δx . In that case, in order to satisfy (13.14), k must be complex:

$$k = k_r + ik_i. \quad (13.15)$$

To see what this means, we write

$$\begin{aligned} \sin(k\Delta x) &= -\frac{i}{2}(e^{ik\Delta x} - e^{-ik\Delta x}) \\ &= -\frac{i}{2}[e^{ik_r\Delta x - k_i\Delta x} - e^{-ik_r\Delta x + k_i\Delta x}] \\ &= -\frac{i}{2}e^{-k_i\Delta x}[\cos(k_r\Delta x) + i\sin(k_r\Delta x)] - e^{k_i\Delta x}[\cos(k_r\Delta x) - i\sin(k_r\Delta x)] \quad (13.16) \\ &= \frac{1}{2}\{[e^{-k_i\Delta x}\sin(k_r\Delta x) + e^{k_i\Delta x}\sin(k_r\Delta x)] + i[-e^{-k_i\Delta x}\cos(k_r\Delta x) + e^{k_i\Delta x}\cos(k_r\Delta x)]\} \\ &= \sin(k_r\Delta x)\cosh(k_i\Delta x) + i\cos(k_r\Delta x)\sinh(k_i\Delta x). \end{aligned}$$

Substituting back into (13.14), and equating real and imaginary parts, we conclude that

$$\begin{aligned} \Omega &= \sin(k_r\Delta x)\cosh(k_i\Delta x), \\ 0 &= \cos(k_r\Delta x)\sinh(k_i\Delta x). \end{aligned} \quad (13.17)$$

We cannot have $\sinh(k_i\Delta x) = 0$, because this would imply $k_i\Delta x = 0$. Hence

$$\cos(k_r\Delta x) = 0, \text{ which implies that } k_r\Delta x = \frac{\pi}{2}. \quad (13.18)$$

This is the $4 - \Delta x$ wave, for which

$$\sin(k_r\Delta x) = 1, \quad (13.19)$$

and so from (13.17) we find that

$$k_i \Delta x = \cosh^{-1}(\Omega) > 0. \quad (13.20)$$

The inequality follows because $\Omega > 1$ by hypothesis. We can now write (13.9) as

$$u_j = U_0 e^{-i\omega t} e^{ik_r j \Delta x} e^{-k_i j \Delta x}. \quad (13.21)$$

Since $k_i > 0$, u_j damps as $j \rightarrow \infty$.

Suppose that we use an uncentered scheme in place of (13.8), e.g.

$$\frac{\partial u_j}{\partial t} + \frac{c}{\Delta x} (u_j - u_{j-1}) = 0, \quad (13.22)$$

with $c > 0$. we will show that the uncentered scheme damps regardless of the values of w , Δx , and c . Let

$$u_j = U e^{-i\omega t} e^{ik j \Delta x}, \quad (13.23)$$

$$u_{j-1} = U e^{-i\omega t} e^{ik(j-1)\Delta x}. \quad (13.24)$$

Then we obtain

$$-i\omega + \frac{c}{\Delta x} (1 - e^{-ik\Delta x}) = 0. \quad (13.25)$$

First, suppose that k is real. Setting the real and imaginary parts of (13.25) to zero gives

$$\begin{aligned} \cos k \Delta x &= 1, \\ -\omega + ck \left[\frac{\sin(k\Delta x)}{k\Delta x} \right] &= 0. \end{aligned} \quad (13.26)$$

Since $\cos k \Delta x = 1$ implies that $k \Delta x = 0$, this solution is not acceptable. We conclude that k must be complex.

Accordingly, use (13.15) to obtain

$$k = k_r + ik_i. \quad (13.27)$$

Then we obtain

$$-i\omega + \frac{c}{\Delta x}(1 - e^{-ik_r\Delta x}e^{k_i\Delta x}) = 0. \quad (13.28)$$

Setting the real part to zero gives:

$$1 - e^{k_i\Delta x}\cos(k_r\Delta x) = 0. \quad (13.29)$$

Setting the imaginary part to zero gives:

$$\Omega + e^{k_i\Delta x}\sin(k_r\Delta x) = 0. \quad (13.30)$$

These two equations can be solved for the two unknowns k_r and k_i . Let

$$\begin{aligned} X &\equiv e^{k_i\Delta x}, \\ Y &\equiv k_r\Delta x, \end{aligned} \quad (13.31)$$

Then

$$\begin{aligned} 1 - X\cos(Y) &= 0, \\ -\Omega &= X\sin(Y) = 0, \end{aligned} \quad (13.32)$$

which implies that

$$\begin{aligned} X &= \sec Y, \\ \tan Y &= \Omega, \end{aligned} \quad (13.33)$$

from which it follows that

$$X = \sec[\tan^{-1}(\Omega)] > 1. \quad (13.34)$$

From (13.34) and (13.31), we see that $k_i > 0$. Substituting back, we obtain

$$u_j = Ue^{-i\omega t}e^{ik_r j\Delta x}e^{-k_i j\Delta x}. \quad (13.35)$$

This shows that, as $j \rightarrow \infty$, the signal weakens.

Exercise: Repeat the analysis above, starting from $\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + c\left(\frac{u_{j+1} - u_{j-1}}{2\Delta x}\right) = 0$.

13.3 Outflow boundaries

So far we have assumed that the initial condition is given everywhere on the space axis, but now we consider a case in which it is given only in a certain limited domain. To illustrate, in Fig. 13.3, lines of constant $x - ct$ are shown. If the initial condition is specified

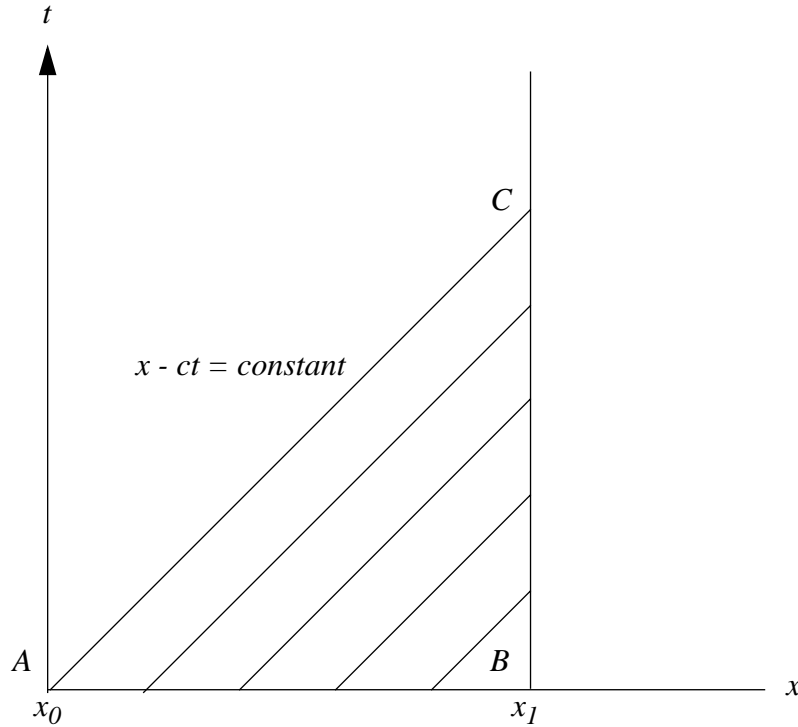


Figure 13.3: An initial condition that is specified between $x = x_0$ and $x = x_1$ determines the solution only along the characteristics shown.

at $t = 0$, between the points A ($x = x_0$, $t = 0$), and B ($x = x_1$, $t = 0$), then $u(x, t)$ is determined in the triangular domain ABC . To determine $u(x, t)$ above the line AC , we need a boundary condition for $t > 0$ 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.

Suppose that we are carrying out our numerical solution of (13.1) over the region between $j = 0$ and $j = J$, as shown in Fig. 13.4, using leapfrog time-differencing and centered space-differencing, i.e.

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + c \left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} \right) = 0, \quad (13.36)$$

and that we are given u at $j = 0$ as a function of time. At $j = 1$ we can write, using centered space differencing,

$$\frac{\partial u_1}{\partial t} + c \left(\frac{u_2 - u_0}{2\Delta x} \right) = 0, \quad (13.37)$$

At $j = J - 1$ we have

$$\frac{\partial u_{J-1}}{\partial t} + c \left(\frac{u_J - u_{J-2}}{2\Delta x} \right) = 0. \quad (13.38)$$

Eq. (13.38) shows that in order to predict u_{J-1} , we need to know u_J . We need to give a condition on u_J as a function of time, or in other words, a “*computational boundary condition*.” Unless we are integrating over the entire globe or are dealing with some other specific problem in which spatial periodicity of the solution can be assumed, we must specify an artificial boundary condition at the point J ; see Fig. 13.4. Ideally, this artificial boundary condition should not affect the interior in any way, since its only purpose is to limit (for computational purposes) the size of the domain.



Figure 13.4: An outflow boundary condition must be specified at $j = J$, in this finite and non-periodic domain.

For the case of the continuous advection equation, we 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.

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 13.1 summarizes the boundary conditions or “Methods” that Nitta considered. The results that he obtained are

shown in Fig. 13.5.

Table 13.1: A summary of the computational boundary conditions studied by Nitta.

| | |
|----------|---|
| Method 1 | $u_J = \text{constant in time}$ |
| Method 2 | $u_J^n = u_{J-1}^n$ |
| Method 3 | $\left(\frac{du}{dt}\right)_J = \left(\frac{du}{dt}\right)_{J-1}$ |
| Method 4 | $u_J^n = u_{J-2}^n$ |
| Method 5 | $u_J^n = 2u_{J-1}^n - u_{J-2}^n$ |
| Method 6 | $\left(\frac{du}{dt}\right)_J = 2\left(\frac{du}{dt}\right)_{J-1} - \left(\frac{du}{dt}\right)_{J-2}$ |
| Method 7 | $\left(\frac{du}{dt}\right)_J = -c \left(\frac{u_J^n - u_{J-1}^n}{\Delta x} \right)$ |
| Method 8 | $\left(\frac{du}{dt}\right)_J = -\frac{c}{2\Delta x} (3u_J - 4u_{J-1} + u_{J-2})$ |

With Method 1, u_J is constant in time. With Method 2, $u_J^{(n)} = u_{J-1}^{(n)}$, i.e. the first derivative of u vanishes at the wall. With Method 4, $u_J^{(n)} = u_{J-2}^{(n)}$, which is similar to Method 2. Method 5, on the other hand, sets $u_J^{(n)} = 2u_{J-1}^{(n)} - u_{J-2}^{(n)}$, a linear extrapolation of the two interior points to $u_J^{(n)}$. This is equivalent to setting the second derivative to zero at the wall. Method 7 predicts u_J by means of

$$\frac{u_J^{(n+1)} - u_J^{(n-1)}}{2\Delta t} + \frac{u_J^{(n)} - u_{J-1}^n}{\Delta x} = 0, \quad (13.39)$$

which uses uncentered differencing in space. This is equivalent to using $u_{J+1} = 2u_{J-1}$, 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.

Recall that, with respect to the spatial dimension, the “physical” mode is given by $u_j \propto \exp\left[ik_1\left(j\Delta x - \frac{\omega}{k_1}t\right)\right]$ and the “computational” mode is given by $u_j \propto (-1)^j \exp\left[-ik_1\left(j\Delta x + \frac{\omega}{k_1}t\right)\right]$. Since the computational mode propagates “upstream,” we wish to examine the solution at the outflow boundary in order to determine the “initial” amplitude of the computational mode excited there. We assume that the domain extends far upstream towards decreasing j . In general, our solution can be expressed as a linear combination of the two modes. Referring to (13.12) and (13.13), we can write

$$u_j = U_0 \operatorname{Re} \left\{ \exp\left[ik\left(j\Delta x - \frac{\omega}{k}t\right)\right] + r(-1)^j \exp\left[-ik\left(j\Delta x + \frac{\omega}{k}t\right)\right] \right\}, \quad (13.40)$$

where k is the wave number of the physical mode and r is the virtual reflection rate at the boundary for the computational mode, so that $|r|$ is the ratio of the amplitude of the computational mode to that of the physical mode. We would like to make $r = 0$.

We now do an analysis to try to understand why Nitta obtained these results.

In Method 1, u_J is kept constant. Assume $u_J = 0$, for simplicity, and let J be even (“without loss of generality”). We then can write from (13.40)

$$u_J = U_0 \{ \exp[ikJ\Delta x] + r \exp[-ikJ\Delta x] \} \exp[-i\omega t] = 0, \quad (13.41)$$

or, since $e^{-i\omega t} \neq 0$, we conclude that

$$r = \frac{-\exp[ikJ\Delta x]}{\exp[-ikJ\Delta x]} = -\exp[2ikJ\Delta x], \quad (13.42)$$

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. 13.5.

With Method 2, and still assuming that J is even, we put $u_J = u_{J-1}$. Then we obtain

$$\exp[ikJ\Delta x] + r \exp[-ikJ\Delta x] = \exp[ik(J-1)\Delta x] - r \exp[-ik(J-1)\Delta x], \quad (13.43)$$

which reduces 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 $|r| \rightarrow 0$, i.e. very long waves are not falsely reflected. In Fig. 13.5, the incident mode is relatively long.

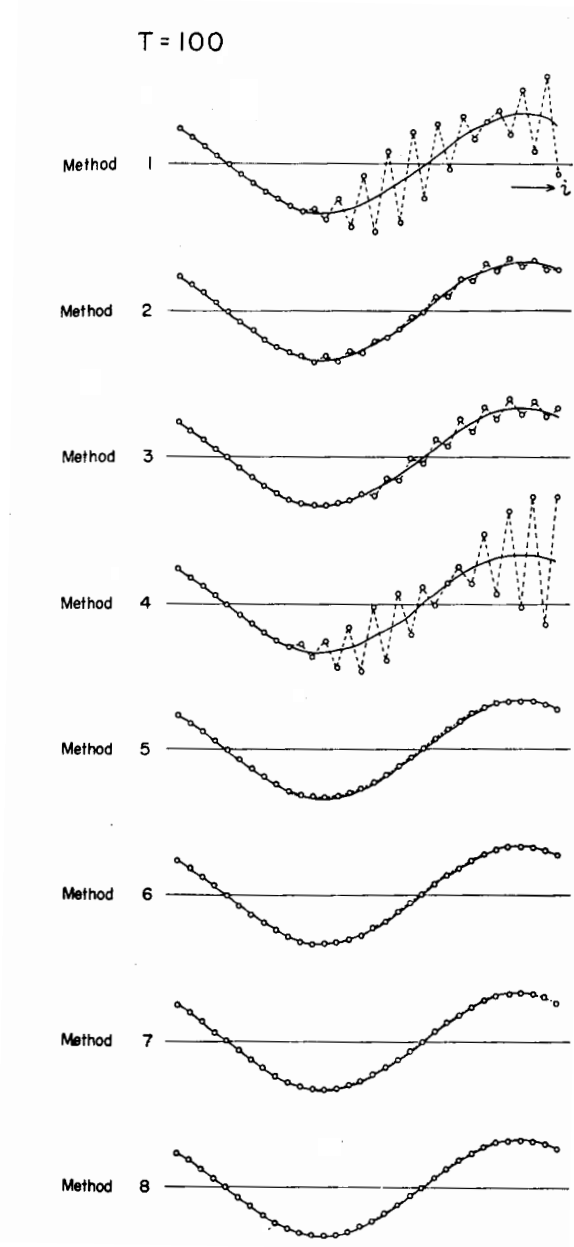


Figure 13.5: A summary of Nitta's numerical results, with various computational boundary conditions. Here leapfrog time differencing was used.

With Method 5, it turns out that $|r| = \tan^2\left(\frac{k\Delta x}{2}\right)$. Fig. 13.6 is a graph of $|r|$ versus $k\Delta x$ for Methods 1, 2, and 5. Because k is the wave number of the physical mode, we only

consider $k\Delta x$ between 0 and $\frac{\pi}{2}$; it is only in this region that we have a “physical” mode corresponding to the real solution. Higher-order extrapolations give even better results for the lower wave numbers, but there is little motivation for doing this.

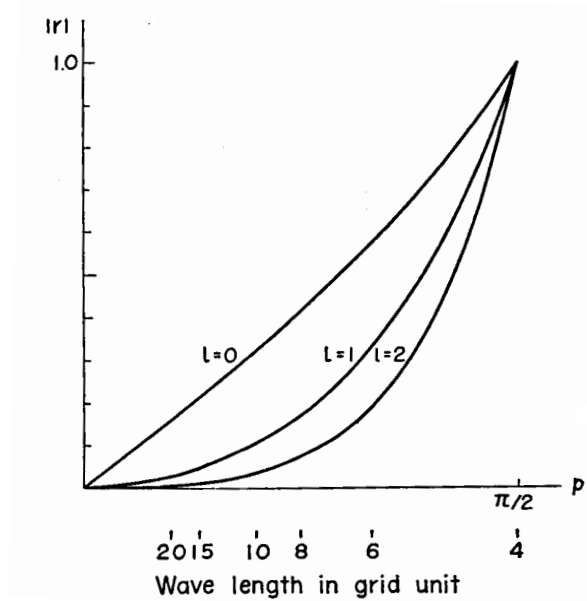


Figure 13.6: A graph of $|r|$ versus $k\Delta x$ for Methods 1, 2, and 5. From Matsuno (1966).

In actual computations one must also have 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? Platzman’s (1954) conclusion, *for the leapfrog scheme*, was that specification of u_j constant on the outflow boundary - and not an extrapolation from nearby points - is necessary for stability.

Therefore, we have a rather paradoxical situation, at least with the leapfrog scheme. If we use Method 1, the domain is quickly filled with small scale “noise,” but this “noise” remains stable. If we use Methods 5 or 7, the domain will only 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 noticeable, it will continue to grow and the model will blow up.

This situation is analogous to using the leapfrog scheme with a friction term specified at the central time level. There is an energy loss in the solution domain through the boundaries when using Method 5 or 7. In Method 1, all of the energy is held, whereas in Methods 5 and 7 some of it is lost due to incomplete reflection at the outflow boundary.

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. The schemes that permit such loss, namely 5 and 7, are therefore more realistic, but nevertheless they can cause an instability. This could be avoided by use of a different time differencing scheme.

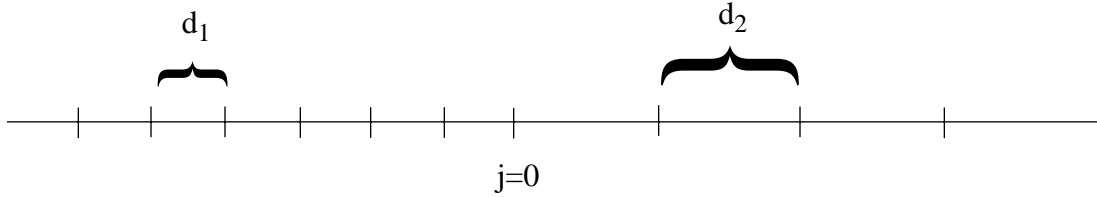
13.4 Advection on nested grids

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 use nested grids. The problem can be minimized by various techniques, but it cannot be eliminated.

Consider the one-dimensional advection equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0, \quad c > 0. \quad (13.44)$$

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 , as shown below:



The sketch shows $d_2 > d_1$, but we will also consider the opposite case. We use the following differential-difference equations:

$$\frac{du_j}{dt} + c \left(\frac{u_{j+1} - u_{j-1}}{2d_1} \right) = 0, \quad \text{for } j < 0$$

$$\frac{du_0}{dt} + c \left[\alpha \left(\frac{u_0 - u_{-1}}{d_1} \right) + \beta \left(\frac{u_1 - u_0}{d_2} \right) \right] = 0, \quad \alpha + \beta = 1$$

$$\frac{du_j}{dt} + c \left(\frac{u_{j+1} - u_{j-1}}{2d_2} \right) = 0, \quad \text{for } j > 0 \quad (13.45)$$

Let $J = 0$ without loss of generality. Define

$$p_1 = k_0 d_1, \quad p_2 = k d_2, \quad (13.46)$$

where k_0 is the incident wave number, and k is the wave number for $j \geq 0$. The solution for $j \leq 0$ is given by:

$$u_j = e^{i(jp_1 - \omega t)} + r(-1)^j e^{-i(jp_1 + \omega t)} , \quad (13.47)$$

where

$$\omega = c \frac{\sin p_1}{d_1} . \quad (13.48)$$

The solution for $j \geq 0$ is

$$u_j = R e^{i(jp_2 - \omega t)} , \quad (13.49)$$

where

$$\omega = \frac{c \sin p_2}{d_2} . \quad (13.50)$$

The frequency, ω , must be the same in both parts of the domain. Equating (13.48) and (13.50) gives

$$\sin p_2 = \frac{d_2}{d_1} \sin p_1 . \quad (13.51)$$

This relates p_2 to p_1 , or k to k_0 .

Since the incident wave must have $c_g^* > 0$ (this is what is meant by “incident”), we know that

$$0 < k_0 d_1 < \frac{\pi}{2} . \quad (13.52)$$

It follows that the wavelength of the incident wave is longer than $4d_1$.

Now consider several cases:

1. $d_2/d_1 > 1$. This means that the wave travels from a relatively fine grid to a relatively coarse grid. Let

$$\sin p_2 = \frac{d_2}{d_1} \sin p_1 \equiv a . \quad (13.53)$$

This implies that

$$e^{ip_2} = \pm \sqrt{1 - a^2} + ia . \quad (13.54)$$

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 ≤ 1 . We consider these two possibilities separately.

a) $a \equiv \sin p_2 > 1$. In this case, p_2 has to be complex. From (13.54),

$$e^{ip_2} = i(a \pm \sqrt{a^2 - 1}) = e^{i\frac{\pi}{2}}(a \pm \sqrt{a^2 - 1}) . \quad (13.55)$$

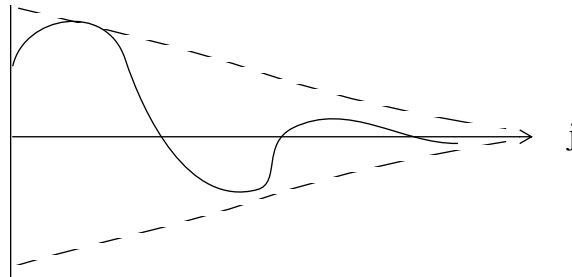
Here we have used $i = e^{i\pi/2}$. The solution for $j \geq 0$ is then

$$u_j = R(a \pm \sqrt{a^2 - 1})^j e^{i\left(\frac{\pi}{2}j - \omega t\right)} . \quad (13.56)$$

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 u_j remains bounded as $j \rightarrow \infty$, we must choose the minus sign. Then

$$u_j = R(a - \sqrt{a^2 - 1})^j e^{i\left(\frac{\pi}{2}j - \omega t\right)} \quad (13.57)$$

The wavelength is $4d_2$, and the amplitude decays as j increases, as shown in the sketch.



b) $a \equiv \sin p_2 \leq 1$. In this case p_2 is real. Since $\sin p_2 \equiv a < 1$, we see that

$$|p_2| < \frac{\pi}{2}. \quad (13.58)$$

This implies that $L = \frac{2\pi}{k} > 4d_2$. The solution is

$$u_j = R e^{i(j \sin^{-1} a - \omega t)}. \quad (13.59)$$

From (13.51), since we are presently considering $d_2/d_1 > 1$, we see that $p_2 > p_1$. We also have from (13.51) that

$$\frac{k}{k_0} = \frac{\left(\frac{\sin p_1}{p_1} \right)}{\left(\frac{\sin p_2}{p_2} \right)}. \quad (13.60)$$

We know that $\sin x/x$ is a decreasing function of x for $0 < x < \pi/2$. We conclude, then, that $k/k_0 > 1$. This means that the wavelength of the transmitted wave is less than that of the incident wave.

2. $d_2/d_1 < 1$. This means that the wave travels from a relatively coarse grid to a relatively fine grid. In this case, p_2 is always real. The analysis is similar to (I, b) above. It turns out that the wavelength of the transmitted wave is longer than that of the incident wave. Since $p_2 \leq \sin^{-1}(d_2/d_1)$, we can show that the maximum wavelength of the transmitted wave is

$$L_{max} = \frac{2\pi d_2}{\sin^{-1}(d_2/d_1)}. \quad (13.61)$$

When $d_2/d_1 = 1/2$, the maximum wavelength is $12d_2$.

Next, we find R and r . At $j = 0$, (13.47) and (13.49) must agree. Then

$$1 + r = R. \quad (13.62)$$

We are also given that

$$\frac{\partial u_0}{\partial t} + c \left[\alpha \left(\frac{u_0 - u_1}{d_1} \right) + \beta \left(\frac{u_1 - u_0}{d_2} \right) \right] = 0 . \quad (13.63)$$

We can substitute (13.47) and (13.49) into (13.63):

$$-i\omega R + c \left\{ \frac{\alpha}{d_1} [1 - e^{-ip_1} + r(1 + e^{ip_1})] + \frac{\beta}{d_2} R(e^{ip_2} - 1) \right\} = 0 . \quad (13.64)$$

Use (13.62) to eliminate r in (13.64), and solve for R :

$$R = \frac{-c \frac{\alpha}{d_1} (1 - e^{-ip_1} - 1 - e^{-ip_1})}{-i\omega + c \left[\frac{\alpha}{d_1} (1 + e^{ip_1}) + \frac{\beta}{d_2} (e^{ip_2} - 1) \right]} . \quad (13.65)$$

Now use (13.48) to eliminate ω . Also use $\alpha + \beta = 1$. The result is

$$R = \frac{2 \cos p_1}{1 + \cos p_1 - \gamma(1 - \cos p_2)} , \quad (13.66)$$

where we have defined

$$\gamma \equiv \left(\frac{\beta}{\alpha} \right) \left(\frac{d_1}{d_2} \right) . \quad (13.67)$$

Substituting (13.66) back into (13.62) gives the reflection coefficient as

$$r = - \left[\frac{1 - \cos p_1 - \gamma(1 - \cos p_2)}{1 + \cos p_1 - \gamma(1 - \cos p_2)} \right] . \quad (13.68)$$

This is the basic result sought.

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 (13.51), we see that for this case $k = k_0$ and $p_1 = p_2$. Then (13.66) and (13.68) give $R = 1$, $r = 0$, i.e. complete transmission and no reflection, as expected.

For $\alpha \rightarrow 0$ with finite d_1/d_2 , we get $\gamma \rightarrow \infty$, $R \rightarrow 0$, and $|r| \rightarrow 1$, 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 $\beta \rightarrow 0$ with finite d_1/d_2 , $\gamma \rightarrow 0$ so that

$$\begin{aligned} R &\rightarrow \frac{2 \cos p_1}{1 + \cos p_1} = 1 - \tan^2 \left(\frac{p_1}{2} \right), \\ r &\rightarrow -\left(\frac{1 - \cos p_1}{1 + \cos p_1} \right) = -\tan^2 \left(\frac{p_1}{2} \right). \end{aligned} \quad (13.69)$$

This is the result obtained with Nitta's "Method 5."

As $p_1, p_2 \rightarrow 0$, $R \rightarrow 1$ and $r \rightarrow 0$, regardless of the value of γ . When p_1 and p_2 are small but not zero,

$$\cos p_1 \cong 1 - \frac{p_1^2}{2}, \quad (13.70)$$

$$\cos p_2 \cong 1 - \frac{p_2^2}{2}, \quad (13.71)$$

and

$$p_2 \cong \left(\frac{d_2}{d_1} \right) p_1. \quad (13.72)$$

Then we find that

$$\begin{aligned} R &\cong \frac{2 \left(1 - \frac{p_1^2}{2} \right)}{2 - \frac{p_1^2}{2} - \gamma \frac{p_2^2}{2}} \cong \left(1 - \frac{p_1^2}{2} \right) \left(1 + \frac{p_1^2}{4} + \gamma \frac{p_2^2}{4} \right) \\ &\cong 1 - \frac{p_1^2}{4} + \gamma \frac{p_2^2}{4} \\ &\cong 1 - \frac{p_1^2}{4} \left[1 - \gamma \left(\frac{d_2}{d_1} \right)^2 \right] \end{aligned} \quad (13.73)$$

Choosing

$$\gamma = (d_1/d_2)^2 \quad (13.74)$$

gives $R = 1 + O(p_2^4)$. Referring back to (13.67), we see that this choice of γ corresponds to

$$\frac{\beta}{\alpha} = \frac{d_1}{d_2}. \quad (13.75)$$

This is a good choice of β/α , because it gives R close to one and $|r|$ close to zero. We can re-write (13.75) as

$$-\alpha d_1 + \beta/d_2 = 0. \quad (13.76)$$

It can be shown that (13.76) is the requirement for second-order accuracy at the “seam” between the two grids. Since the given equations have second-order accuracy elsewhere, (13.75) [and (13.76)] essentially express the requirement that the order of accuracy be spatially homogeneous.

13.5 Analysis of boundary conditions for the advection equation using the energy method

Consider the one-dimensional advection equation:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0. \quad (13.77)$$

Multiplying (13.77) by $2u$, we obtain

$$\frac{\partial}{\partial t} u^2 + \frac{\partial}{\partial x} c u^2 = 0. \quad (13.78)$$

This shows that u^2 is also advected by the current. Defining u^2 as the “energy”, we see that $c u^2$ is the energy flux and $\frac{\partial}{\partial x} c u^2$ is the energy flux divergence.

Suppose that (13.77) is approximated by the differential difference equation:

$$\frac{\partial u_j}{\partial t} + c \frac{u_{j+1} - u_{j-1}}{2\Delta x} = 0. \quad (13.79)$$

Multiplying (13.79) by $2u_j$, we obtain

$$\frac{\partial}{\partial t} u_j^2 + \frac{cu_j u_{j+1} - cu_{j-1} u_j}{\Delta x} = 0, \quad (13.80)$$

Comparing (13.80) with (13.78), we see that $cu_j u_{j+1}$ and $cu_{j-1} u_j$ are the energy fluxes from grid point j to grid point $j+1$, and from grid point $j-1$ to grid point j , respectively. Applying (13.80) to the grid point $j+1$ gives

$$\frac{\partial}{\partial t} u_{j+1}^2 + \frac{cu_{j+1} u_{j+2} - cu_j u_{j+1}}{\Delta x} = 0. \quad (13.81)$$

By comparing (13.80) and (13.81), we see that $cu_j u_{j+1}$ represents the energy transferred from the grid point j to grid point $j+1$. This is illustrated in Fig. 13.7

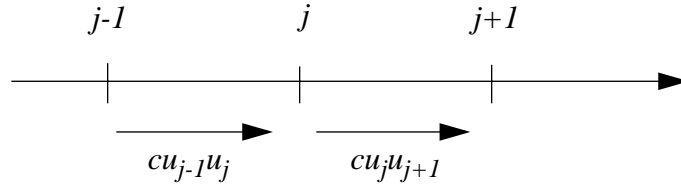


Figure 13.7: Sketch illustrating the energy fluxes that appear in (13.80).

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 $u_j u_{j+1}$ is not necessarily positive. When $u_j u_{j+1}$ is negative, as when u 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 $\frac{\pi}{2} < k\Delta x < \pi$. The implication is that for short waves, for which $u_j u_{j-1} < 0$, energy flows in the $-x$ direction, i.e. “backward.” This is consistent with our earlier analysis of the group speed.

When we put an *artificial* boundary at $j = J$, and if we let $u_J = 0$ 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.

For Nitta’s Method 2, $u_J = u_{J-1}$. This gives

$$cu_J u_{J-1} = cu_J^2 > 0. \quad (13.82)$$

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

u_J .

For Nitta's Method 4, $u_J = u_{J-2}$. Then for short waves

$$cu_{J-1}u_J = cu_{J-1}u_{J-2} < 0. \quad (13.83)$$

This is bad.

For Nitta's Method 5,

$$u_J = 2u_{J-1} - u_{J-2}, \quad (13.84)$$

so

$$cu_Ju_{J-1} = cu_{J-1}(2u_{J-1} - u_{J-2}) = c(2u_{J-1}^2 - u_{J-1}u_{J-2}). \quad (13.85)$$

For very short waves,

$$u_{J-1}u_{J-2} < 0, \quad (13.86)$$

so that the flux given by (13.85) is positive, as it should be. For very long waves,

$$u_{J-1}u_{J-2} \cong u_{J-1}u_{J-1}, \quad (13.87)$$

so the flux is approximately

$$cu_Ju_{J-1} \cong cu_{J-1}^2 > 0. \quad (13.88)$$

For Nitta's Method 7,

$$\frac{\partial u_J}{\partial t} + c \frac{u_J - u_{J-1}}{\Delta x} = 0, \quad (13.89)$$

so we find that

$$\frac{\partial u_J^2}{\partial t} + 2c \frac{u_J^2 - u_J u_{J-1}}{\Delta x} = 0. \quad (13.90)$$

The energy flux "into J " is $u_J u_{J-1}$, while that "out of J " is $u_J^2 > 0$. Applying (13.80) at J ,

$$\frac{\partial}{\partial t}(u_{J-1}^2) + c \frac{u_{J-1}u_J - u_{J-2}u_{J-1}}{\Delta x} = 0. \quad (13.91)$$

This shows that the energy flux out of $J - 1$ is the same as the flux into J , which is good.

13.6 Physical and computational reflection of gravity waves at a wall

Now we discuss

$$\frac{\partial u_j}{\partial t} + g \frac{h_{j+1} - h_{j-1}}{2\Delta x} = 0, \quad (13.92)$$

$$\frac{\partial h_j}{\partial t} + H \frac{u_{j+1} - u_{j-1}}{2\Delta x} = 0, \quad (13.93)$$

which are, of course, differential-difference analogs of the one-dimensional shallow water equations. Consider a distribution of the dependent variables on the grid as shown in Fig. 13.8.

The wave solutions of (13.92) and (13.93) are

$$(u_j, h_j) \propto e^{i(kj\Delta x - \omega t)}, \quad (13.94)$$

giving

$$\begin{aligned} \omega u_j - g h_j \frac{\sin(k\Delta x)}{\Delta x} &= 0, \\ \omega h_j - H u_j \frac{\sin(k\Delta x)}{\Delta x} &= 0. \end{aligned} \quad (13.95)$$

Since u_j and h_j are not both identically zero, we obtain the familiar dispersion relation

$$\omega^2 = k^2 g H \left(\frac{\sin p}{p} \right)^2 \text{ where } p \equiv k\Delta x. \quad (13.96)$$

As discussed in Chapter 5, there are four solutions for a given value of ω , i.e. $p = p_0$, $p = -p_0$, $p = \pi - p_0$ and $p = -(\pi - p_0)$. In general, for a given ω , the solution for u_j is a linear combination of the four modes, and can be written as

$$u_j = [A e^{ip_0 j} + B e^{-ip_0 j} + C e^{i(\pi - p_0)j} + D e^{-i(\pi - p_0)j}] e^{-i\omega t}. \quad (13.97)$$

By substituting (13.97) into (13.93), we find that h_j satisfies

$$h_j = \frac{H \sin p_0}{\omega \Delta x} [A e^{ip_0 j} - B e^{-ip_0 j} + C e^{i(\pi - p_0)j} - D e^{-i(\pi - p_0)j}] e^{-i\omega t}. \quad (13.98)$$

If we assume $\omega > 0$, so that $\sin p_0 = \frac{\omega \Delta x}{\sqrt{gH}}$ [see (13.96)], then (13.98) reduces to

$$h_j = \sqrt{\frac{H}{g}} [a e^{ip_0 j} - b e^{-ip_0 j} + c e^{i(\pi - p_0)j} - d e^{i(\pi - p_0)j}] e^{-i\omega t} . \quad (13.99)$$

Consider an incident wave traveling toward the right with a certain wave number k_0 , such that $0 < p_0 (= k_0 \Delta x) < \pi$. Since we are assuming $\omega > 0$, $e^{ip_0 j} e^{-i\omega t}$ represents such a wave.

Two additional waves can be produced by reflection at the boundary. We assume that the amplitude of the incident wave with $p = p_0$ is 1, and that of the reflected wave with $p = -p_0$ is R , and that of the reflected wave with $p = \pi - p_0$ is r . In other words, we take $A = 1$, $B = R$, $C = r$, and $D = 0$. Then (13.97) and (13.99) can be written as

$$u_j = [e^{ip_0 j} + R e^{-ip_0 j} + r e^{i(\pi - p_0)j}] e^{-i\omega t} , \quad (13.100)$$

$$h_j = \sqrt{\frac{H}{g}} [e^{ip_0 j} - R e^{-ip_0 j} + r e^{i(\pi - p_0)j}] e^{-i\omega t} . \quad (13.101)$$

Now suppose that at $j = J$ we have a rigid wall (a real, physical wall). Since there is no flow through the wall, we know that $u_J = 0$, for all time. This is a physical boundary condition. Also, $\left(\frac{\partial h}{\partial x}\right)_J = 0$ is required, because otherwise there would be a pressure gradient which would cause u_J to change with time. Consider two possible methods for approximating this:

$$\text{Method I: } u_J = 0, \quad h_J - h_{J-1} = 0 , \quad (13.102)$$

$$\text{Method II: } u_J + u_{J-1} = 0, \quad h_J - h_{J-1} = 0 . \quad (13.103)$$

Method II essentially corresponds to placing the wall at $J - 1/2$ rather than at J . Then u is assumed to be antisymmetric and h is assumed to be symmetric about the wall.

A third method is to predict h_j using uncentered differencing:

$$\text{Method III: } u_J = 0, \quad \frac{\partial h_J}{\partial t} + H \left(\frac{0 - u_{J-1}}{\Delta x} \right) = 0 . \quad (13.104)$$

This is equivalent to assuming $u_{J+1} = -u_{J-1}$, and then applying (13.93) to the point J .

In a straightforward manner, the R 's and the r 's can be determined for each of Methods I, II, and III. Table 5.1 gives the expressions for R and r for each method, and Fig. 13.8

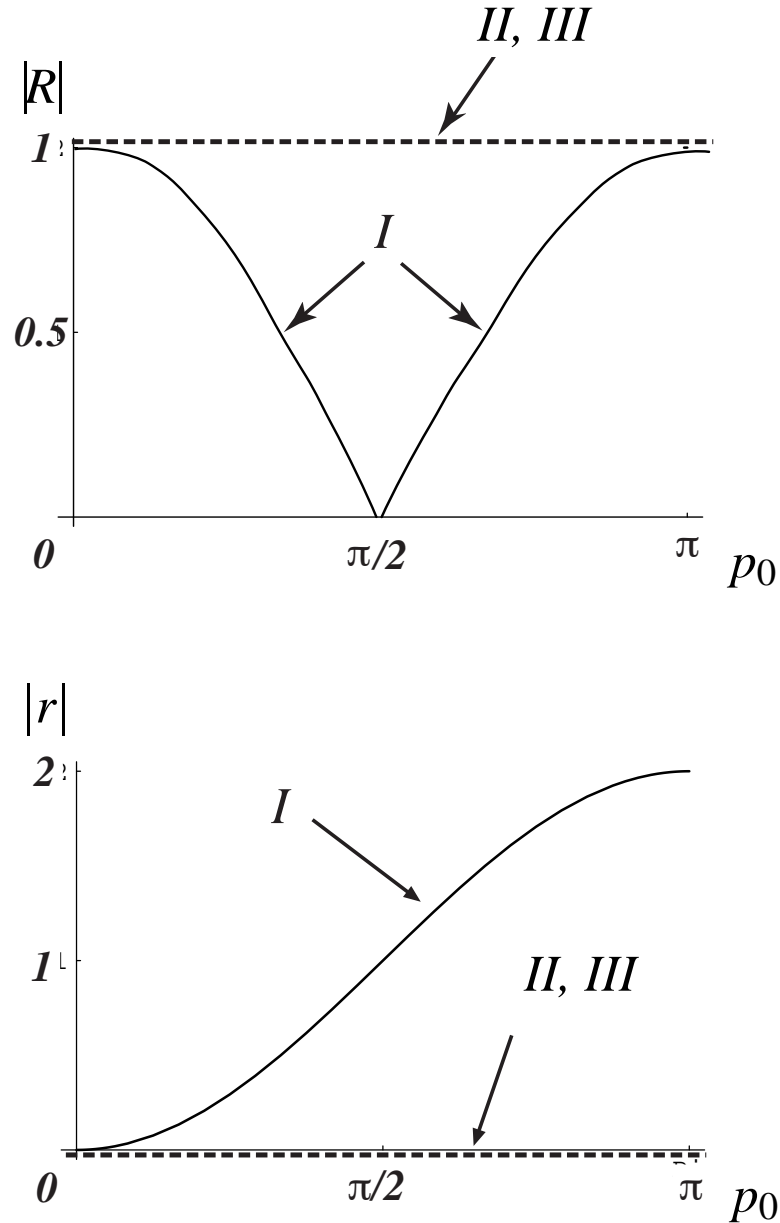


Figure 13.8: The upper panel shows the variation of $|R|$ with p_0 for Methods I, II, and III. The lower panel shows the corresponding results for $|r|$.

shows $|R|$ and $|r|$ plotted as functions of p_0 . Method I is obviously bad. False energy is

Table 13.2: Expressions for R and r , for Methods I, II, and III.

| | Method I | Method II | Method III |
|-----|-----------------|---|------------|
| R | $-\cos p_0$ | $-\cos p_0 + i \sin p_0$ $(R = 1)$ | -1 |
| r | $-1 + \cos p_0$ | 0 | 0 |

produced when p_0 is close to π . Even when p_0 is small, part of the incident energy is reflected back with $p = \pi - p_0$, and the solution will become “noisy”. Methods II and III are better. Method III is best.

A way to bypass most of the problems associated with the existence of too many modes is to use a grid with a “staggered” spatial distribution of the dependent variables, as shown in Fig. 13.9. Note that u is defined “at the wall,” where $j = J$. If we are dealing with a rigid wall, the boundary condition $u_J = 0$ is sufficient because h is not defined at the boundary. Use of this staggered grid means use of either only circled or only boxed quantities in Fig. 13.8.

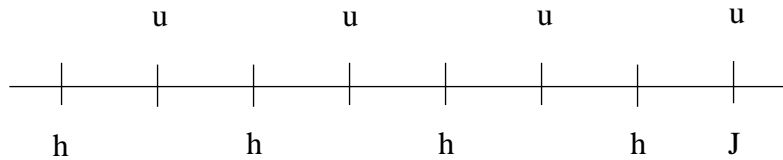


Figure 13.9: A one-dimensional staggered grid for solution of the shallow water equations, near a wall where $j = J$.

13.7 Boundary conditions for the gravity wave equations with an advection term

We now generalize our system of equations to include advection by a mean flow U , in the following manner:

$$\left(\frac{\partial}{\partial t} + U \frac{\partial}{\partial x} \right) u + g \frac{\partial h}{\partial x} = 0, \quad (13.105)$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)v = 0, \quad (13.106)$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)h + H\frac{\partial u}{\partial x} = 0. \quad (13.107)$$

We have also added a velocity component, v , in the y -direction. The system is still linear, but it is getting more realistic! 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 (13.105) through (13.107) are hyperbolic, we can write them in normal form:

$$\left[\frac{\partial}{\partial t} + (U + c)\frac{\partial}{\partial x}\right]\left(u + \sqrt{\frac{g}{H}}h\right) = 0, \quad (13.108)$$

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)v = 0, \quad (13.109)$$

$$\left[\frac{\partial}{\partial t} + (U - c)\frac{\partial}{\partial x}\right]\left(u - \sqrt{\frac{g}{H}}h\right) = 0. \quad (13.110)$$

Here $c \equiv \sqrt{gH}$. We assume $c > |U|$, which is often true in the atmosphere. For (13.108) and (13.109), the lines $x - (U + c)t = \text{constant}$ and $x - (U - c)t = \text{constant}$ are the characteristics, and are shown as the solid lines in Fig. 13.10. 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 (13.109). This, of course, is an advection equation, and so v is a constant along the lines $x - Ut = \text{constant}$, which are shown schematically by the broken lines in Fig. 13.10. 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.

13.8 The energy method as a guide in choosing boundary conditions for gravity waves

From the gravity wave equations we can derive the total energy equation

$$\frac{\partial}{\partial t}\left(\frac{1}{2}Hu^2 + \frac{1}{2}gh^2\right) + gH\frac{\partial}{\partial x}(hu) = 0. \quad (13.111)$$

Suppose that the gravity-wave equations are approximated by

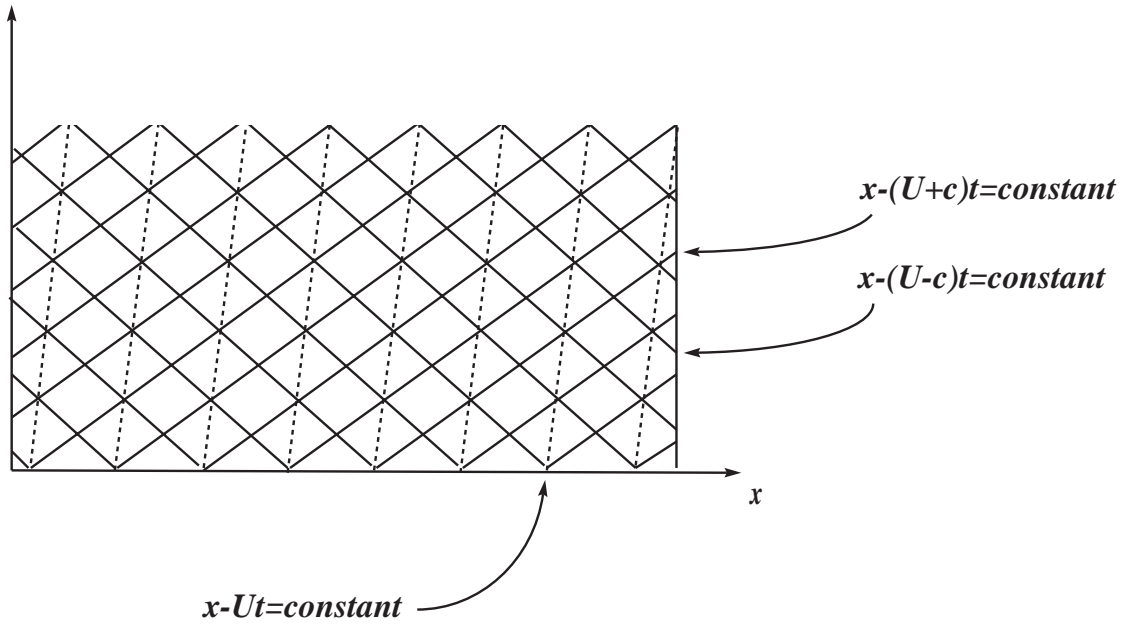


Figure 13.10: Characteristics for the case of shallow water wave propagation with an advecting current U .

$$\frac{\partial u_j}{\partial t} + g \frac{h_{j+1} - h_{j-1}}{2\Delta x} = 0, \quad (13.112)$$

$$\frac{\partial h_j}{\partial t} + H \frac{u_{j+1} - u_{j-1}}{2\Delta x} = 0. \quad (13.113)$$

The corresponding discretized total energy equation is

$$\frac{\partial}{\partial t} \left(\frac{1}{2} H u_j^2 + \frac{1}{2} g h_j^2 \right) + g H \left[\frac{\frac{1}{2} (u_j h_{j+1} + u_{j+1} h_j) - \frac{1}{2} (u_{j-1} h_j + u_j h_{j-1})}{\Delta x} \right] = 0. \quad (13.114)$$

Earlier, $u = \sqrt{\frac{g}{H}} h$ was recommended as a computational boundary condition at the artificial boundary BB'. Correspondingly, if we let

$$h_J = \sqrt{\frac{H}{g}} u_{J-1}, \quad u_J = \sqrt{\frac{g}{H}} h_{J-1}, \quad (13.115)$$

then the energy flux from the point $J-1$ to the point J becomes

$$gH\frac{1}{2}(u_{J-1}h_J + u_Jh_{J-1}) = \sqrt{gH}\left(\frac{1}{2}Hu_{J-1}^2 + \frac{1}{2}gh_{J-1}^2\right) > 0 \quad (13.116)$$

which is guaranteed to be positive (outward energy flow).

When the boundary is a real, rigid wall, we may put it between the points $J-1$ and J , and let $u_J + u_{J-1} = 0$ and $h_J - h_{J-1} = 0$ (Method II). Then the energy flux is $gH\frac{1}{2}(u_{J-1}h_J + u_Jh_{J-1}) = 0$. Alternatively, we may put the wall at the point J ($u_J = 0$) and let

$$\frac{\partial h_J}{\partial t} + H\left(\frac{0 - u_{J-1}}{\Delta x}\right) = 0 \text{ (Method III)}. \quad (13.117)$$

The total energy equation at the point J is then

$$\frac{\partial}{\partial t}\frac{1}{2}gh_J^2 + gH\left(\frac{0 - h_Ju_{J-1}}{\Delta x}\right) = 0. \quad (13.118)$$

There is no kinetic energy term since $u_J = 0$. In (13.118), gHh_Ju_{J-1} is the energy flux from the point $J-1$ to the point J . Note, however, that $u_Jh_{J-1} = 0$. There is no energy flux beyond the point J .

For the total energy equation (13.111), the boundary condition $u - \sqrt{\frac{g}{H}}h = 0$ gives

$$uh = \sqrt{\frac{g}{H}}h^2 > 0, \quad (13.119)$$

so that there will be *outward* energy flux at the right boundary. From (13.114), we see that with a rigid wall, the energy flux at $J - \frac{1}{2}$ is $u_{J-1}h_J + u_Jh_{J-1}$. We have already discussed three methods for giving boundary conditions on the wave equations.

For Method I, we find that

$$u_J = 0, \quad h_J - h_{J-1} = 0, \quad (13.120)$$

the energy flux is

$$u_{J-1}h_{J-1} \quad (13.121)$$

which is generally different from zero. This means that total reflection does not occur.

Method II is

$$u_J + u_{J-1} = 0, \quad h_J - h_{J-1} = 0. \quad (13.122)$$

The energy flux is then

$$u_{J-1}h_J + u_Jh_{J-1} = u_{J-1}h_{J-1} + (-u_{J-1})h_{J-1} = 0, \quad (13.123)$$

so we have *complete reflection*.

13.9 Summary

Computational modes in space can be generated at real and / or artificial walls, and arise from space differencing schemes. In problems with boundaries, these modes can necessitate the introduction of “computational boundary conditions” at outflow boundaries. The computational modes have short wavelengths and move “backwards.” Various methods can be used to minimize problems at boundaries, as discussed in the paper by Matsuno. None of these methods completely eliminates the problems caused by artificial boundaries, however.

Problem

1. Derive the form of $|r|$ for Nitta’s Method 7. Assume that J is even.

| | |
|--|-----|
| Table 3.1 Adams-Bashforth Schemes ($b = m = 0, l > 0$) | 48 |
| Table 3.2 Adams-Moulton schemes. | 50 |
| Table 3.3 List of 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. (3.3) and Fig. 3.1; and l , which controls the number of values of f used, again as defined in (3.3). | 68 |
| Table 3.4 Schemes for the nonlinear decay equation, as studied by Kalnay and Kanamitsu (1988). | 75 |
| Table 5.1 Well known methods for solving boundary value problems, and the operation count and storage requirements for each, measured in terms of, the number of equations to be solved. | 136 |
| Table 10.1 Examples of time differencing schemes obtained through various choices of. The subscripts and have been omitted for simplicity. | 257 |
| Table 13.1 A summary of the computational boundary conditions studied by Nitta. | 310 |
| Table 13.1 Expressions for and , for Methods I, II, and III. | 326 |

- FIG. 2.1:** An example of a grid, with uniform grid spacing Δx . The grid points are denoted by the integer index j . Half-integer points can also be defined. - - - - 17
- FIG. 2.2:** Schematic illustrating the interpretation of the fourth-order difference in terms of the extrapolation of the second-order difference based on a spacing of $4\Delta x$, and that based on a spacing of $2\Delta x$. The extrapolation reduces the effective grid size to $(2/3)2\Delta x$. - - - - - 12
- FIG. 2.3:** Figure used in the derivation of the first line of (2.72). - - - - - 22
- FIG. 2.4:** A grid for the solution of the one-dimensional advection equation. - - 24
- FIG. 2.5:** The shaded area represents the “domain of dependence” of the solution of the upstream scheme at the point j . - - - - - 26
- FIG. 2.6:** Diagram illustrating the concepts of interpolation and extrapolation. See text for details. - - - - - 29
- FIG. 2.7:** The amplification factor for the upstream scheme, plotted for three different wave lengths. - - - - - 35
- FIG. 2.8:** “Total” damping experienced by a disturbance crossing the domain, as a function J , the number of grid points across the domain. Here we have assumed that λ is pure imaginary. - - - - - 40
- FIG. 3.1:** In Eq. (3.3), we use a weighted combination of \bar{u} to compute an “average” value of \bar{u} over the time interval Δt . - - - - - 44
- FIG. 3.2:** A simple fortran example to illustrate how the fourth-order Runge-Kutta scheme works. Note the four calls to subroutine “dot.” This makes the scheme expensive. - - - - - 53
- FIG. 3.3:** Schematic illustration of the solution of the oscillation equation for the case in which λ is pure imaginary and the phase changes by θ on each time step. - 55
- FIG. 3.4:** This figure shows the magnitude of the amplification factor as a function of $\omega\Delta t$ for various difference schemes. The Euler, backward, trapezoidal, Euler-backward, and Heun schemes are shown by curves I, II, III, IV, and V respectively. The magnitude of the amplification factor for the trapezoidal scheme coincides with that of the true solution for all values of $\omega\Delta t$. Caution: This does not mean that the trapezoidal scheme gives the exact solution! - - - - - 58
- FIG. 3.5:** This figure shows the behavior of the imaginary (λ_i) and real (λ_r) components of the amplitude, as Ω varies. Recall that $\tan \theta$ is given by λ_i/λ_r for each scheme. From this plot we can also see the behavior of θ as θ varies, for each scheme. - - - - 59
- FIG. 3.6:** The leapfrog scheme. - - - - - 60

FIG. 3.7: An oscillatory solution that arises with the leapfrog scheme for Δt , for the case in which the two initial values of ϕ are not the same. ----- 61

FIG. 3.8: Panels a and b: Amplification factors for the leapfrog scheme as applied to the oscillation equation with Δt . Panels c and d: Solutions of the oscillation equation as obtained with the leapfrog scheme for Δt . In making these figures it has been assumed that $\Delta t = 0.1$. ----- 63

FIG. 3.9: Graphs of the real and imaginary parts of the physical and computational modes of the solution of the oscillation equation as obtained with the leapfrog scheme for Δt . ----- 64

FIG. 3.10: Panel a shows the amplification factors for the leapfrog scheme as applied to the oscillation equation with Δt . Panel b shows the real and imaginary parts of the corresponding solution, for $n=0, 1, 2$, and 3 . ----- 65

FIG. 3.11: Panels a and b show the amplification factors for the oscillation equation with the leapfrog scheme, with Δt . 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 ϕ , which is characteristic of this type of instability. -66

FIG. 3.12: Amplification factor of various schemes for the oscillation equation (from Baer and Simons, 1970). The horizontal axis in each panel is Ω . See Table 3.3. ----- 70

FIG. 3.13: ϕ and ψ (for the physical mode), plotted as functions of Ω , for the oscillation equation (from Baer and Simons, 1970). See Table 3.3. ----- 71

FIG. 3.14: An example illustrating how the leapfrog scheme leads to instability with the decay equation. The solution shown here represents the computational mode only and would be superimposed on the physical mode. ----- 73

FIG. 4.1: The staggered grid used in (4.13) and (4.14). -----89

FIG. 4.2: Four interpolations as functions of the input values. a) arithmetic mean, b) geometric mean, c) harmonic mean, d) Eq. (4.36), 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. -----94

FIG. 4.3: The amplification factor for the Lax-Wendroff scheme, for two different wavelengths, plotted as a function of Ω . Compare with Fig. 2.4. -----100

FIG. 4.4: 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. - - - - - | 101 |
| FIG. 4.5: Sketch defining notation used in the discussion of the group velocity. - - - - - | 102 |
| FIG. 4.6: Sketch used to illustrate the concept of group velocity. The short waves are modulated by longer waves. - - - - - | 103 |
| FIG. 4.7: Yet another sketch used to illustrate the concept of group velocity. The short wave has wavelength $2\Delta x$. - - - - - | 104 |
| FIG. 4.8: The time evolution of the solution of (4.75) at grid points $j = 0, 1$, and 2 . - - - - - | 105 |
| FIG. 4.9: The solution of (4.75) for $t = 5$ and $t = 10$ for j in the range -15 to 15 , with “spike” initial conditions. From Matsuno (1966). - - - - - | 106 |
| FIG. 4.10: The solution of (4.72) with “box” initial conditions. From Wurtele (1961). - - - - - | 107 |
| FIG. 4.11: The ratio of the computational phase speed, c_p , to the true phase speed, c , plotted as a function of $k\Delta x$, for the second-order and fourth-order schemes. - - - - - | 108 |
| FIG. 4.12: 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. - - - - - | 109 |
| FIG. 4.13: Sketch illustrating the angle α_m on a rectangular grid. - - - - - | 122 |
| FIG. 7.1: A grid for solution of the one-dimensional shallow water equations. - - - - - | 153 |
| FIG. 7.2: 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 $\Delta x = \Delta y$. - - - - - | 156 |
| FIG. 7.3: 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 k and l , respectively, except for the E grid, for which k and l are used. The vertical coordinate is the normalized frequency, ω . For the E grid, the results are meaningful only in the triangular region for which $\omega < 1$. The left column shows results for ω , and the right column for ω . - - - - - | 159 |
| FIG. 7.4: A plot of ω as a function of k , for $l = 0$. - - - - - | 162 |
| FIG. 8.1: The staggered grid used in the one-dimensional case. - - - - - | 171 |
| FIG. 9.1: A mountain. As we move uphill in the x direction, the surface pressure decreases and the surface geopotential increases. - - - - - | 200 |
| FIG. 9.2: Evaluating the horizontal pressure gradient force. - - - - - | 201 |

- FIG. 9.3:** A schematic picture of the representation of mountains using the h coordinate. -----203
- FIG. 9.4:** Four possible vertical coordinate systems. -----206
- FIG. 9.5:** Schematic illustration of the Charney - Phillips grid and the Lorenz grid. 209
- FIG. 10.1:** An example of aliasing error. Distance along the horizontal axis is measured in units of Δx . The wave given by the solid line has a wave length of Δx . This is shorter than Δx , and so the wave cannot be represented on the grid. Instead, the grid “sees” a wave of wavelength $2\Delta x$, as indicated by the dashed line. Note that the $2\Delta x$ -wave is “upside-down.” -----223
- FIG. 10.2:** 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 10.8 mm 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 Mapping Mission) satellite. The obvious oscillation in both curves, with a period close to 23 days, is an artifact due to aliasing. See text for further explanation. -----224
- FIG. 10.3:** The red line is a plot of \bar{u} versus t . The dashed black line connects \bar{u} with \bar{u} , corresponding to the example of Fig. 10.1. -----226
- FIG. 10.4:** The phase change per grid point for: a) \bar{u} , and b) \bar{u} . -----227
- FIG. 10.5:** Plots of the functions \bar{u} and \bar{u} given by (10.27) and (10.28), respectively. For plotting purposes, we have used Δx . The functions have been evaluated only for integer values of n and m , 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. -----231
- FIG. 10.6:** Schematic illustration of the mechanism of aliasing instability. Nonlinear interactions feed energy into scales too small to be represented on the grid, and this energy folds back through aliasing into scales that can be represented. The process feeds on itself. This can cause the total amount of energy to increase, unless the scheme is energy conserving. -----233
- FIG. 10.7:** Sketch illustrating the mechanism of aliasing instability.-----235
- FIG. 10.8:** Diagram used in the explanation of Fjortoft’s (1953) analysis of the exchanges of energy and enstrophy among differing scales in two-dimensional motion. -----238
- FIG. 10.9:** Stencil used in the discussion of vorticity conservation for \bar{u} . See text for details. -----244

FIG. 10.10: The central point in each figure is . Stream function and vorticity are both defined at each of the mesh points indicated by the black dots. The colored lines represent contributions to ψ from ψ , or both, from the various neighboring points. 247

FIG. 10.11: -----
Results of tests with the various finite-difference Jacobians. Panel c shows the initial kinetic energy is at a low wave number. ----- 250

FIG. 10.12: The arrangement of the mass, zonal wind, and meridional wind on the C grid. ----- 254

FIG. 11.1: 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° longitude have a strong positive zonal component, while points near 270° longitude have a strong negative zonal component. ----- 261

FIG. 11.2: Map projections of the continents: a.) Mercator projection. b.) North polar stereographic projection. ----- 265

FIG. 11.3: 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 0 points are obtained by interpolation from the other grid. The big circle is the image of the Equator. Points labeled $*$ are not used. ----- 267

FIG. 11.4: 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° . 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° except that the pizza slices are 6° tall. ----- 268

FIG. 11.5: A plot of the smoothing parameter as given by (11.51), for the “worst case” of the shortest zonal mode. The dashed vertical lines demarcate the belt of latitude near the Equator for which no smoothing is needed. It has been assumed that the longitudinal grid spacing is $5/4$ times the latitudinal grid spacing, as it is for the grid shown in Fig. 11.4. ----- 273

FIG. 11.6: Kurihara grid on one octant of the sphere. ----- 274

FIG. 11.7: Wandering electron grid. White cells have five walls, light gray cells have six walls, and dark gray cells have seven walls. ----- 275

FIG. 11.8: a.) Icosahedron. b.) Partition each face into 64 smaller triangles. c.) Project

| | |
|--|-----|
| onto the sphere. - - - - - | 275 |
| FIG. 11.9: Cells neighboring a given cell (shaded) on triangular, square, and hexagonal grids. A “wall neighbor” is a neighbor which lies directly across a cell wall. 276 | |
| FIG. 11.10: Configuration of grid triangles for the case $K = 5$. - - - - - | 278 |
| FIG. 11.11: Masuda’s velocity potential field. 279 | |
| FIG. 12.1: The Earth is bumpy. - - - - - | 285 |
| FIG. 12.2: Table of ϵ , showing which pairs can contribute to wave numbers in the range ϵ to ϵ . The pairs in the triangular regions marked by X’s do not contribute. 287 | |
| FIG. 12.3: Rhomboidal and triangular truncation. From Jarraud and Simmons (1983). - - - - - | 292 |
| FIG. 12.4: Percentage of total kinetic energy in each spectral component. From Jarraud and Simmons (1983) based on Baer (1972). - - - - - | 292 |
| FIG. 13.1: A schematic plot, with α and β as coordinates, for the true solution (13.3) and the approximate solution (13.10). The dashed line illustrates that for a given α the approximate solution allows two different wave numbers. - - - - - | 303 |
| FIG. 13.2: Schematic illustration of a computational mode that is restricted to the domain α , and a physical mode that is restricted to the domain β . - - - - - | 304 |
| FIG. 13.3: An initial condition that is specified between $x = x_0$ and $x = x_1$ determines the solution only along the characteristics shown. - - - - - | 308 |
| FIG. 13.4: An outflow boundary condition must be specified at $j = J$, in this finite and non-periodic domain. - - - - - | 309 |
| FIG. 13.5: A summary of Nitta’s numerical results, with various computational boundary conditions. Here leapfrog time differencing was used. 312 | |
| FIG. 13.6: A graph of ϵ versus $k\Delta x$ for Methods 1, 2, and 5. From Matsuno (1966). 313 | |
| FIG. 13.7: Sketch illustrating the energy fluxes that appear in (13.80). - - - - - | 321 |
| FIG. 13.8: The upper panel shows the variation of ϵ with p_0 for Methods I, II, and III. The lower panel shows the corresponding results for ϵ . - - - - - | 325 |
| FIG. 13.9: A one-dimensional staggered grid for solution of the shallow water equations, near a wall where ϵ . - - - - - | 326 |
| FIG. 13.10: Characteristics for the case of shallow water wave propagation with an | |

advecting current U. ----- 328

References and Bibliography

A

- Arakawa, A., 1966: Computational design for long-term numerical integration of the equations of fluid motion. Two-dimensional incompressible flow. Part I. *J. Comp. Phys.*, **1**, 119-143
- Arakawa, A., 1968: Numerical simulation of large-scale atmospheric motions. *Proceedings of a Symposium in Applied Mathematics*, Durham, N.C., 1968. 24-40.
- Arakawa, A., 1988: Finite-difference methods in climate modeling. *Physically-Based Modelling and Simulation of Climate and Climatic Change - Part I*, M. E. Schlesinger (ed.), Kluwer Academic Press, 79-168.
- Arakawa, A. and Y.-J. Hsu, 1990: Energy conserving and potential-entropy dissipating schemes for the shallow water equations. *Mon. Wea. Rev.*, **118**, 1960-1969.
- Arakawa A., and S. Moorthi, 1988: Baroclinic instability in vertically discrete systems. *J. Atmos. Sci.*, **45**, 1688-1707.
- Arakawa, A., 1966: Computational design for long-term numerical integration of the equations of fluid motion. Two-dimensional incompressible flow. Part I. *J. Comp. Phys.*, **1**, 119-143.
- 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**, Academic Press, New York, pp. 173-265.
- Arakawa, A., and V. R. Lamb, 1981: A potential entropy and energy conserving scheme for the shallow water equations. *Mon. Wea. Rev.*, **109**, 18-36.
- Arfken, G., 1985: *Mathematical methods for physicists*. Academic Press, San Diego, 985 pp.
- Arpé, K., and E. Klinker, 1986: Systematic errors of the ECMWF operational forecasting model in mid-latitudes. *Quart. J. Roy. Meteor. Soc.*, **112**, 181-202.
- Asselin, R., 1972: Frequency filter for time integrations. *Mon. Wea. Rev.*, **100**, 487-490.

B

- Baer, F., and T. J. Simons, 1970: Computational stability and time truncation of coupled nonlinear equations with exact solutions. *Mon. Wea. Rev.*, **98**, 665-679. (This paper first appeared as: Baer, F. and T. J. Simons, 1968: Computational stability and time truncation of coupled nonlinear equations with exact solutions. *Colorado State University, Atmospheric Science Paper No. 131*.)

- Baer, F., 1972: An alternate scale representation of atmospheric energy spectra. *J. Atmos. Sci.*, **29**, 649-664.
- Bates, J. R., S. Moorthi, and R. W. Higgins, 1993: A global multilevel atmospheric model using a vector semi-Lagrangian finite-difference scheme. Part I: Adiabatic formulation. *Mon. Wea. Rev.*, **121**, 244-263.
- Bleck, R., 1973: Numerical forecasting experiments based on the conservations of potential vorticity on isentropic surfaces. *J. Appl. Meteor.*, **12**, 737-752.
- Bleck, R., and D. B. Boudra, 1981: Initial testing of a numerical ocean circulation model using a hybrid (quasi-isopycnic) vertical coordinate. *J. Phys. Oceanogr.*, **11**, 755-770.
- Bleck, R., 1984: Vertical coordinate transformation of vertically discretized atmospheric fields. *Mon. Wea. Rev.*, **112**, 2535-2539.
- Bleck, R., H. P. Hanson, D. Hu, and E. B. Kraus, 1989: Mixed layer-thermocline interaction in a three-dimensional isopycnic coordinate model. *J. Phys. Oceanogr.*, **19**, 1417-1439.
- Bleck, R., and S. G. Benjamin, 1993: Regional weather prediction with a model combining terrain-following and isentropic coordinates. Part I: Model description. *Mon. Wea. Rev.*, **121**, 1770-1785.
- Boris, J. P., and D. L. Book, 1973: Flux-corrected transport I: SHASTA -- a fluid transport algorithm that works. *J. Comp. Phys.*, **11**, 38-69.

C-D

- Durran, D. R., 1991: The third-order Adams-Bashforth method: An attractive alternative to leapfrog time differencing. *Mon. Wea. Rev.*, **119**, 702-720.

E

- Eliassen, A., B. Machenhauer, and E. Rasmussen, 1970: On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields. *Rept. No. 2, Institute for Theoretical Meteorology*, Copenhagen University, Copenhagen, 35 pp.
- Eliassen, A., and E. Raustein, 1968: A numerical integration experiment with a model atmosphere based on isentropic coordinates. *Meteorologiske Annaler*, **5**, 45-63.
- Eliassen, A., and E. Raustein, 1970: A numerical integration experiment with a six-level atmospheric model with isentropic information surface. *Meteorologiske Annaler*,

5, 429-449.

F

Fulton, S. R., P. E. Ciesielski, and W. H. Schubert, 1986: Multigrid methods for elliptic problems: A review. *Mon. Wea. Rev.*, **114**, 943-959.

G

Gerald, C. F., 1970: *Applied Numerical Analysis*, Addison-Wesley, 340 pp.

Green, J. S. A., 1970: Transfer properties of the large-scale eddies and the general circulation of the atmosphere. *Quart. J. Roy. Meteor. Soc.*, **96**, 157-185.

H

Haertel, P. T., and D. A. Randall, 2001: Could a pile of slippery sacks behave like an ocean? Undergoing revisions for *Mon. Wea. Rev.*

Haltiner, G. J., Williams, R. T., 1984: *Numerical Prediction and Dynamic Meteorology*. Wiley.

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. *Mon. Wea. Rev.*, **111**, 609-662.

Heckley, W. A., 1985: Systematic errors of the ECMWF operational forecasting model in tropical regions. *Quart. J. Roy. Meteor. Soc.*, **111**, 709-738.

Heikes, R. P., and D. A. Randall, 1995: 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. P., and D. A. Randall, 1995: 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.

Holzer, M., 1996: Optimal spectral topography and its effect on model climate. *J. Climate*, **9**, 2443-2463.

Hoskins, B. J., M. E. McIntyre, and A. W. Robertson, 1985: On the use and significance of isentropic potential vorticity maps. *Quart. J. Roy. Meteor. Soc.*, **111**, 877-946.

Hsu, Y.-J., and A. Arakawa, 1990: Numerical modeling of the atmosphere with an isentropic vertical coordinate. *Mon. Wea. Rev.*, **118**, 1933-1959.

J

- Janjic, Z. I., and F. Mesinger, 1989: Response to small-scale forcing on two staggered grids used in finite-difference models of the atmosphere. *Q. J. Roy. Meteor. Soc.*, **115**, 1167-1176.
- .Janjic, Z. I., 1990: The step-mountain coordinate: Physical package. *Mon. Wea. Rev.*, **118**, 1429-1443.
- Jarraud, M., and A. J. Simmons, 1983: The spectral technique. *Seminar on Numerical Methods for Weather Prediction*, European Centre for Medium Range Weather Prediction, Reading, England, 99. 1-59.
- 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. *Mon. Wea. Rev.*, **111**, 870-886.

K

- Kalnay-Rivas, E., A. Bayliss, and J. Storch, 1977: The 4th order GISS model of the global atmosphere. *Contrib. Atmos. Phys.*, **50**, 306-311.
- Kalnay, E. and M. Kanamitsu, 1988: Time schemes for strongly nonlinear damping equations. *Mon. Wea. Rev.*, **116**, 1945-1958.
- Kasahara, A., 1974: Various vertical coordinate systems used for numerical weather prediction. *Mon. Wea. Rev.*, **102**, 509-522.
- Kasahara, A., and W. M. Washington, 1967: NCAR global general circulation model of the atmosphere. *Mon. Wea. Rev.*, **95**, 7, 389-402.
- Konor, C. S., and A. Arakawa, 1997: Design of an atmospheric model based on a generalized vertical coordinate. *Mon Wea. Rev.*, **125**, 1649-1673.
- Krueger, S. K., 1988: Numerical simulation of tropical cumulus clouds and their interaction with the subcloud layer. *J. Atmos. Sci.*, **45**, 2221-2250.
- Kurihara, Y., 1965: Numerical Integration of the Primitive Equations on a Spherical Grid. *Mon. Wea. Rev.*, **93**, 399-415.

L

- Laprise, R., 1992: The resolution of global spectral models. *Bull. Amer. Meteor. Soc.*, **73**, 1453-1454.
- Lax, P. D. and B. Wendroff, 1960: Systems of conservation laws. *Communications on*

pure and applied mathematics, **XIII**, pp. 217-237.

Lilly, D. K., 1965: On the computational stability of numerical solutions of time-dependent nonlinear geophysical fluid dynamics problems. *Mon. Wea. Rev.*, **93**, 11-26.

Lindberg, C., and A. Broccoli, 1996: Representation of topography in spectral climate models and its effect on simulated precipitation. *J. Climate*, **9**, 2641-2659.

Lindzen, R. S., and H.-L. Kuo, 1969: A reliable method for the numerical integration of a large class of ordinary and partial differential equations. *Mon. Wea. Rev.*, **97**, 732-734.

Lorenz, E. N., 1955: Available potential energy and the maintenance of the general circulation. *Tellus*, **7**, 157-167.

Lorenz, E. N., 1960: Energy and numerical weather prediction. *Tellus*, **12**, 364-373.

Lorenz, E. N., 1969: Three approaches to atmospheric predictability. *Bull. Amer. Meteor. Soc.*, **50**, 345-349.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505-513.

M

Masuda, Y., and H. Ohnishi, 1986: An Integration Scheme of the Primitive Equations Model with an Icosahedral-Hexagonal Grid System and its Application to the Shallow Water Equations. *Short- and Medium-Range Numerical Weather Prediction*. Japan Meteorological Society, Tokyo, 317-326.

Matsuno, T., 1966: Numerical integrations of the primitive equations by a simulated backward difference method. *J. Meteor. Soc. Japan*, **Ser. 2 44**, 76-84.

Matsuno, T., 1966: False reflection of waves at the boundary due to the use of finite differences. *J. Meteor. Soc. Japan*, **Ser. 2 44**, 145-157.

McKee, T. B., and S. K. Cox, 1974: Scattering of visible radiation by finite clouds. *J. Atmos. Sci.*, **31**, 1885-1892.

McGregor, J. L., 1996: Semi-Lagrangian advection on conformal-cubic grids. *Mon Wea. Rev.*, **124**, 1311-1322.

Mellor, G. L., and T. Yamada, 1974: A hierarchy of turbulence closure models for the planetary boundary layer. *J. Atmos. Sci.*, **31**, 1791-1806.

- Mesinger, F., 1971: Numerical integration of the primitive equations with a floating set of computation points: Experiments with a barotropic global model. *Mon. Wea. Rev.*, **99**, 15-29.
- Mesinger, F., and Z. I. Janjic, 1985: Problems and numerical methods of the incorporation of mountains in atmospheric models. *Large-Scale Computations in Fluid Mechanics*. Part 2. Lec. Appl. Math., **22**, Amer. Math. Soc., 81-120.
- Monaghan, J. J., 1992: Smoothed particle hydrodynamics. *Ann. Rev. Astron. Astrophys.*, **30**, 543-574.
- Moeng, C.-H., 1984: A large-eddy simulation model for the study of planetary boundary-layer turbulence. *J. Atmos. Sci.*, **41**, 2052-2062.
- Moeng, C.-H., and J. C. Wyngaard, 1986: An analysis of closures for pressure-scalar covariances in the convective boundary layer. *J. Atmos. Sci.*, **43**, 2499-2513.
- Monaghan, J. J., 1992: Smoothed particle hydrodynamics. *Ann. Rev. Astron. Astrophys.*, **30**, 543-574.
- Morse, P. M., and H. Feshbach, 1953: *Methods of theoretical physics, Part I*. McGraw-Hill, 997 pp.

N

- Nitta, T., 1964: On the reflective computational wave caused by the outflow boundary condition. *J. Met. Soc. Japan*, **42**, 274-276.
- Norris, P. M., 1996: *Radiatively-driven convection in marine stratocumulus clouds*. Ph.D. thesis, University of California, San Diego, 175 pp.
- North, G. R., 1975: Theory of energy balance climate models. *J. Atmos. Sci.*, **32**, 2033 - 2043.

O

- Orszag, S. A., 1970: Transform method for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. *J. Atmos. Sci.*, **27**, 890-902.

P

- Phillips, N. A., 1957: A coordinate system having some special advantages for numerical forecasting. *J. Meteor.*, **14**, 184-185.

- Phillips, N. A., 1957: A map projection system suitable for large-scale numerical weather prediction. *J. Meteor. Soc. Japan.*, **75**, 262-267.
- Phillips, N. A., 1959: An example of non-linear computational instability. In *The Atmosphere and Sea in Motion*, (Bert Bolin, ed.), Rockefeller Inst. Press, New York, 501-504.
- Phillips, N. A., 1959: Numerical Integration of the Primitive Equations on the Hemisphere. *Mon. Wea. Rev.*, **87**, 333-345.
- Platzman, G. W., 1954: The computational stability of boundary conditions in numerical integration of the vorticity equation. *Arch. Meteor. Geophys. u. Bioklimatol., Ser. A7*, 29-40.
- Popper, K. R., 1959: *The logic of scientific discovery*. Hutchinson Education, 479 pp.
- Purser, R. J., and M. Rancic, 1998: Smooth quasi-homogeneous gridding of the sphere. *Quart. J. Roy. Meteor. Soc.*, **124**, 637-647.

Q - R

- Randall, D. A., 1994: Geostrophic adjustment and the finite-difference shallow-water equations. *Mon. Wea. Rev.*, **122**, 1371-1377.
- Richtmeyer, 1963: A survey of difference methods for nonsteady fluid dynamics. *NCAR Technical Note 63-2*, NCAR, Boulder, Colo.
- Ringler, T. D., R. P. Heikes, and D. A. Randall, 1999: Modeling the atmospheric general circulation using a spherical geodesic grid: A new class of dynamical cores. *Mon. Wea. Rev.* (in press).
- Robert, A., T. L. Yee, and H. Ritchee, 1985: Semi-Lagrangian and semi-implicit numerical integration scheme for multilevel atmospheric models. *Mon. Wea. Rev.*, **113**, 388-394.

S

- Sadourny, R., A. Arakawa, and Y. Mintz, 1968: Integration of the Nondivergent Barotropic Vorticity Equation with an Icosahedral-Hexagonal Grid for the Sphere. *Mon. Wea. Rev.*, **96**, 351-356.
- Sadourny, R., and P. Morel, 1969: A Finite Difference Approximation of the Primitive Equations for a Hexagonal Grid on a Plane. *Mon. Wea. Rev.*, **97**, 439-445.
- Semtner, A. J., Jr., and R. M. Chervin, 1992: Ocean general circulation from a global eddy-

resolving model. *J. Geophys. Res.*, **97C**, 5493-5551.

Shewchuk, J. R., 1994: An introduction to the conjugate gradient method without the agonizing pain. Available on the web at <http://www.cs.berkeley.edu/~jrs/>.

Simmons, A. J., and D. M. Burridge, 1981: Energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. *Mon. Wea. Rev.*, **109**, 758-7661.

Smagorinski, J., 1963: General circulation experiments with the primitive equations. *Mon. Wea. Rev.*, **91**, 99-164.

Smolarkiewicz, P. K., 1991: Nonoscillatory advection schemes. In *Proceedings of the Seminar on Numerical Methods in Atmospheric Models*, ECMWF, November 1991.

Staniforth, A., and J. Cote, 1991: Semi-lagrangian integration schemes for atmospheric models -- A review. *Mon. Wea. Rev.*, **119**, 2206-2223.

T

Takacs, L. L., 1985: A two-step scheme for the advection equation with minimized dissipation and dispersion errors. *Mon. Wea. Rev.*, **113**, 1050-1065.

Trease, H. E., 1988: Three-dimensional free-Lagrange hydrodynamics. *Computer Physics Communications*, **48**, 39-50.

U - V - W

Williamson, D. L., 1968: Integration of the Barotropic Vorticity Equation on a Spherical Geodesic Grid. *Tellus*, **20**, 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., 1971: Numerical Methods Used in Atmospheric Models, GARP Pub. Ser. No. 17 (JOC, WMO, Geneva, 1979), Chap. 2, 51-120.

Williamson, D. L., and J. G. Olson, 1994: Climate simulations with a semi-Lagrangian version of the NCAR Community Climate Model. *Mon. Wea. Rev.*, **122**, 1594-1610.

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. *J. Comp. Phys.*, **102**, 221-224.

-
- Williamson, D. L., and P. J. Rasch, 1994: Water vapor transport in the NCAR CCM2. *Tellus*, **46A**, 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, 1961: On the problem of truncation error. *Tellus*, **13**, 379-391.

Y-Z

- Zalesak, S. T., 1979: Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comp. Phys.*, **31**, 335-362.
- Zapotocny, T. H., D. R. Johnson, and F. M. Reames, 1994: Development and initial test of the University of Wisconsin global isentropic-sigma model. *Mon. Wea. Rev.*, **122**, 2160-2178.
- Zhu, Z., J. Thuburn, B. J. Hoskins, and P. H. Haynes, 1992: A vertical finite-difference scheme based on a hybrid σ - θ -p coordinate. *Mon. Wea. Rev.*, **120**, 851-862.

